DISSERTATION

INTERDISCIPLINARY TECHNIQUES IN PROTEIN BINDING PREDICTION AND CRYSTAL ENGINEERING

Submitted by

Jacob Benjamin DeRoo

School of Biomedical Engineering

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Fall 2024

Doctoral Committee:

Advisor: Melissa Reynolds
Co-Advisor: Christopher D. Snow

Ken Reardon
Mark Zabel

ABSTRACT

INTERDISCIPLINARY TECHNIQUES IN PROTEIN BINDING PREDICTION AND CRYSTAL ENGINEERING

This dissertation explores the integration of interdisciplinary methods such as advanced robotic automation, machine learning, and hybrid materials synthesis to dual protein engineering challenges: predicting protein-peptide binding specificity and the preparation of crystalline protein materials. The first chapter introduces a computational pipeline, PAbFold, based on AlphaFold2, designed to predict linear antibody epitopes from a given antigen sequence. This method provides a rapid and cost-effective alternative to traditional experimental techniques for epitope mapping, significantly lowering the financial barrier for laboratories. By accurately identifying binding sites on target proteins, PAbFold enhances the understanding of antibody-antigen interactions, facilitating the development of diagnostic and therapeutic antibodies in a more accessible manner.

The second chapter presents an innovative approach to protein crystallization scale-up utilizing the Opentrons 2 liquid handling robot. This automation not only reduces manual labor and variability in crystallization experiments but also makes high-throughput crystallization more accessible to a broader range of laboratories by decreasing costs. Traditional high throughput protein crystallization liquid handling robots are priced around $75,000; the Opentrons 2 costs around $15,000. By employing Python scripts for precise control of the Opentrons 2, the study demonstrates successful crystallization of model and non-model proteins, highlighting the potential of automated systems in structural biochemistry to democratize access to high-quality protein crystals.

The third chapter delves into the creation of hybrid materials by combining metal-organic frameworks (MOFs) with porous protein crystals. The research demonstrates the feasibility of embedding MOF domains within protein crystals, potentially opening new avenues for applications in catalysis, gas storage, and chemical warfare agent detoxification. By developing a new class of hybrid materials, this work contributes to making advanced structural biochemical research.

Together, these chapters illustrate a modern interdisciplinary approach that embraces machine learning and automation in service of the engineering of peptide-binding proteins and crystalline protein materials. The integration of automation, computational predictions, and hybrid materials offers a promising path toward more efficient and innovative solutions in biochemical research, while significantly lowering the cost barriers, thereby increasing accessibility for researchers worldwide.

ACKNOWLEDGEMENTS

I would like to thank all members of the Reynolds lab and of the Snow lab for their scientific input, as well as their camaraderie. I would also like to thank the members of the interdisciplinary group TagTeam, comprised of the Snow lab, Geiss lab, and Stasevich lab. I would of course like to thank Dr. Melissa Reynolds and Dr. Christopher Snow, not only for their scientific support, but for their humanity, patience, and emotional support when life turned almost as ugly as it possible could. I would like to thank Dr. Brian Geiss for taking me on as a mentee, even when he was under no obligation to.

A warm thank you to Will, Joe, and James for being chosen brothers.

A warmer thanks to Isaac, for being the first brother. And to Jessica and Raymond for their support throughout life.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1 – Introduction

This work spans the interactions between antibodies and peptide antigens, high throughput protein crystallization experiments, and the creation and assessment of crystalline protein materials doped with metal-organic frameworks. Throughout my journey here at Colorado State University, I have had the privilege and honor to collaborate with myriad scientists with a diverse set of backgrounds, including computational biologists, synthetic chemists, microbiologists, virologists, crystallographers, analytical chemists, and enzymologists. I was in a unique position to be either the only or one of a few computational scientists on the projects I worked on, allowing me to be involved in numerous – perhaps too many – projects. Participating in numerous collaborations provided first-hand exposure to many different disciplines and let me find my place in the scientific community that allowed me to combine my interests, passions, and strengths.

1.1 Computational Modelling of Biomolecular Structures

Computational modelling is a vast area of mathematics that reaches every part of the scientific community at large. Modelling lets us do many things that otherwise we wouldn't be able to do: predict/describe trends, extract additional information or parameters, make an inference, extrapolate beyond what we've observed, and more. Modelling is especially important when we have datasets that are extremely large or when the subject of the data is particularly large or small. In the case of structural chemistry and biology often, we cannot directly measure or observe how changing something may affect the behavior or shape of the molecule – thus, we must be exceptionally careful and crafty when we are interested in how the structure of our molecule of interest changes after some sort of perturbation. A core competency of any scientific

modeler is collecting and processing data that is meaningful to the system we're trying to observe, else the model becomes worthless. The intersection of math and science is key to us as humans as we navigate through life on the beautiful blue planet we call home – and indeed is important as we question that which we cannot directly observe, both small (like molecules) and (large like interplanetary travel) as we inevitably begin to explore that which we cannot yet touch.

What is a model? A model is any set of mathematical equations or relationships that describe a physical phenomenon, and they can vary greatly in complexity depending on the nature of the observed phenomenon. For example, a very simple model could be timing a car as it travels down a road to get a function or description of where it is in respect to time (**Figure 1.1**, top left). With a model obtained that describes how far away the car is from its origin, we can take the derivative of this function to calculate a velocity – this is an example of getting extra information, or parameter extraction, from a model that we may otherwise have not been able to get. There are many types of models, like nonlinear simulations that depend on the previous time step for a solution. For example, a ball dropped from a height of 50 m is going to behave differently at 3 seconds than a ball that is dropped from a height of 25 m (a single nonlinear equation) (**Figure 1.1**, top right), or the contents of a reactor as a series of chemical reactions proceeds (multiple coupled nonlinear ODEs), (**Figure 1.1**, bottom left). All of these models are examples of phenomena that can have their behavior derived or predicted, but this is not true for all phenomena. On the frontier of model generation lies machine learning (ML), neural networks (NN), and artificial intelligence (AI). These types of models need a significant number of data points but can be used to make predictions that otherwise would be impossible with more standard processes. For example, a tumor can be malignant or benign depending on how

negatively it is affecting the body. It may be possible to predict how dangerous the tumor is based on its size and surface roughness with a support vector machine (**Figure 1.1**, bottom right).

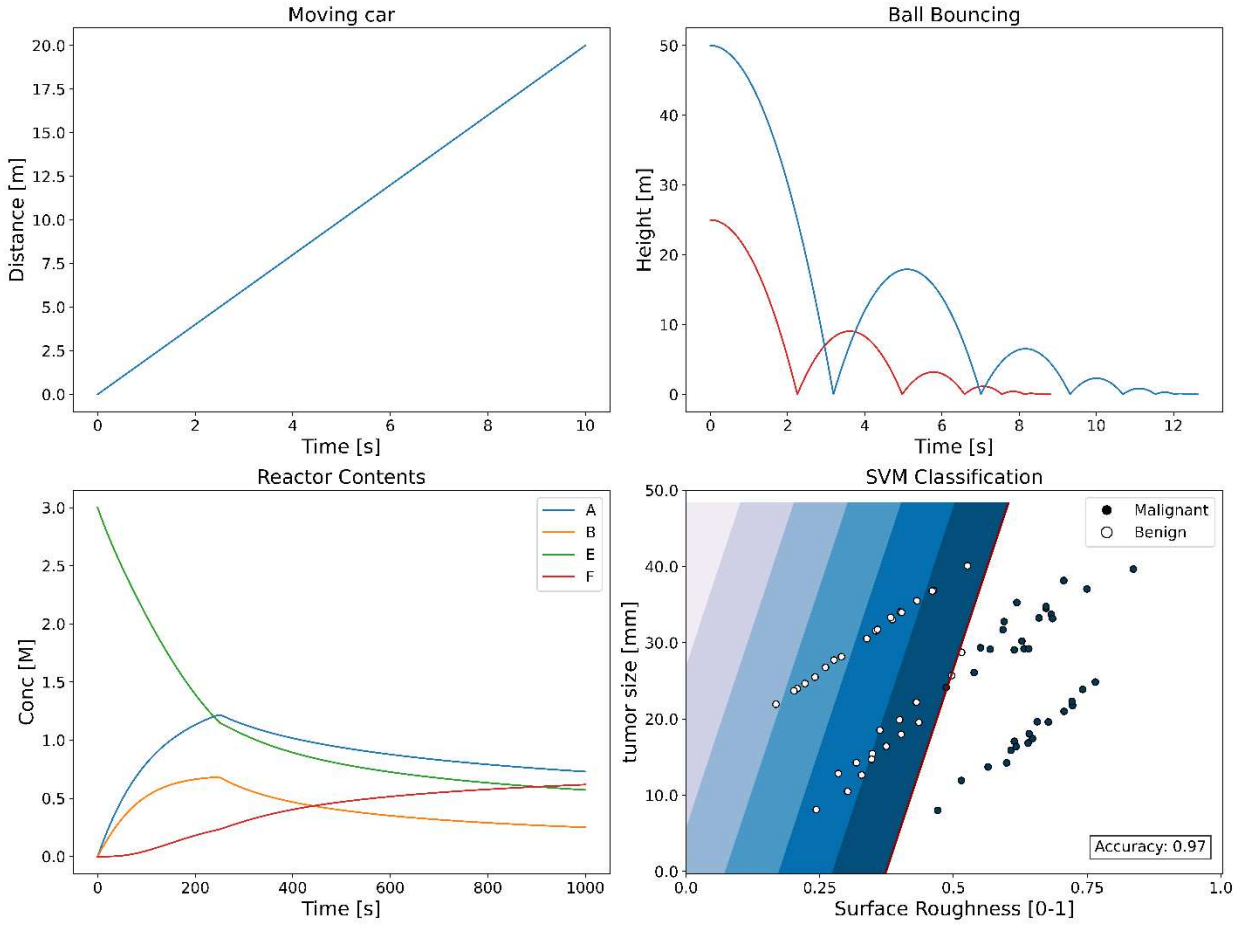


Figure 1.1: Examples of toy models. A) The behavior of a car traveling at 2 m/s – a linear model. B) How a ball's height might be observed after dropping it from a height of 50 m – a simulation of a ball being dropped. C) The contents of a membrane reactor as a reaction proceeds and the membrane begins to foul, allowing less product to pass through – a system of nonlinear ODEs. D) The size and surface roughness of tumors, and whether those tumors are benign or malignant – data characterized by a Support Vector Machine, a supervised machine learning technique. These data are a fabricated random sampling, used only for illustration purposes. It should not be used for anything other than to demonstrate how an SVM works.

Ultimately, any model can and will differ from reality, particular when attempting to predict complex relationships. It is critical to verify the output of the model for two reasons: the first is that models still only make predictions, and the second is that models can only make

predictions based on the data it has access to. The reliance on data leads to the commonly used phrase, “garbage in, garbage out.” Additionally, data that is similar to the experiment is valuable in getting a prediction close to the experimental observation, reducing the need for the model to extrapolate.

Given recent rapid progress in machine learning, so many tools have been developed for biomolecular structural modelling that a plethora of twitter accounts, githubs, and blogs are not sufficient to keep up with all of the emerging technologies. However, one extremely impactful tool is certainly worth specific discussion; AlphaFold2[1] . Every 2 years, the Critical Assessment of Techniques for Protein Structure Prediction (CASP) organizers hold a competition to find which research group and model is the best at predicting a protein’s structure. For a long time, physics-based approaches dominated the competition – methods that rely to varying extents on describing the relationships between amino acids and the atoms therein using forcefields. Then, in 2020 during CASP 14, Google’s DeepMind group showcased their deep learning model called AlphaFold2. AlphaFold2 (AF2) outperformed every other model at CASP 14 by such a significant margin (**Figure 1.2**) that the structure prediction community was in disbelief – for many cases AF2 effectively “solved” the protein structure prediction problem by producing structural models with quality comparable to experimental structure determination.

In addition to AF2 predicting a structure from a sequence, it also provides us with several metrics with which to determine how well we can trust the output coming from AF2. One of these metrics is the predicted local distance difference test (pLDDT). This is a scale from 0-100 that gives us a confidence score for each residue. The way this is calculated is by iterating through each atom at a time, and then finding all the contacts of that atom within a certain inclusion radius (default is 15Å). We then calculate the number of contacts in the inclusion

radius that exist within 4Å, 2Å, 1Å, and 0.5Å. All of the contacts in each of these thresholds are then averaged to create a percentage of contacts, or amount of contacts being made across the four different distances.

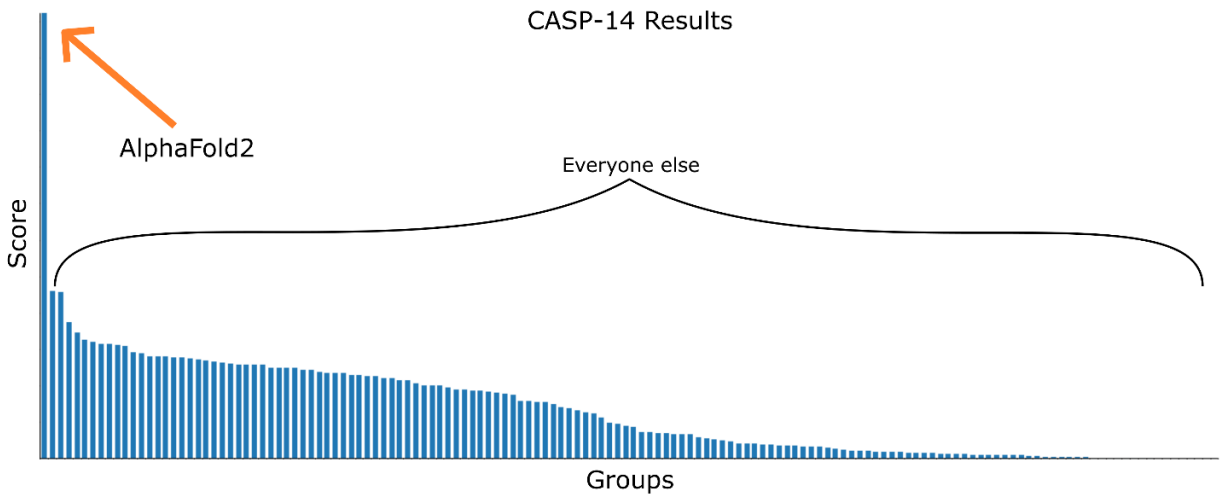


Figure 1.2: The results of the CASP 14 competition (https://www.predictioncenter.org/casp14/zscores_final.cgi). The custom score that the CASP organization has developed was used to assess how well each group did at predicting the structures of some protein structures that hadn't been deposited in the Protein Data Bank. AF2 outperformed all the other groups significantly.

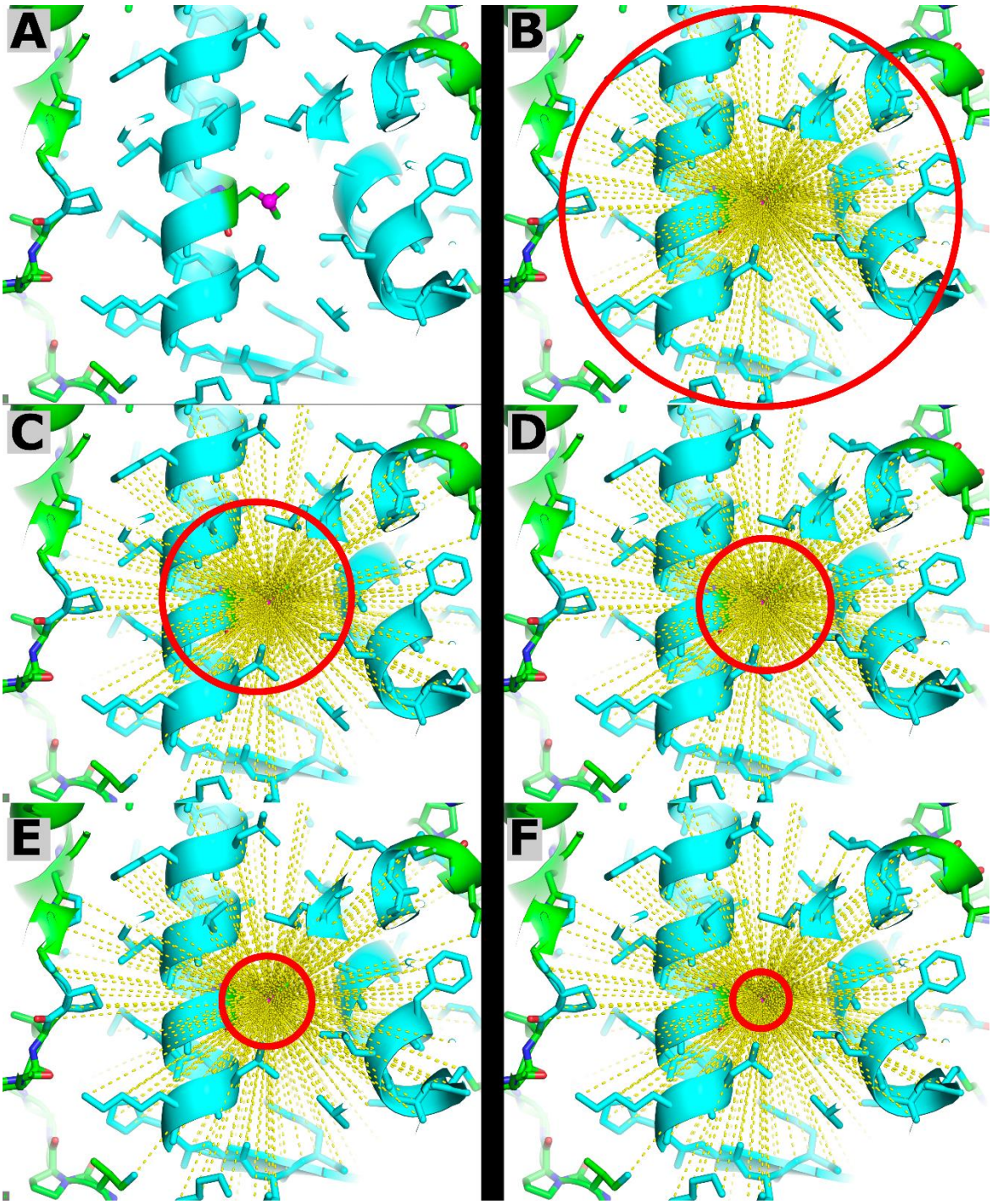


Figure 1.3: Explanation of pLDDT calculation. **A)** The example atom is highlighted in a small purple sphere. **B)** All the contacts within a 15Å sphere are found. **C)** The contacts that exist within 4Å. **D)** The contacts that exist within 2Å. **E)** The contacts that exist within 1Å. **F)** The contacts that exist within 0.5Å.

Within a year, DeepMind published their model and made it open access. With their method published, the scientific community was off to the races to create models that could compete with AF2. Noteworthy competitors include ESMFold[2], RosettaFold[3], and a faithful Python reimplementation of AF2, OpenFold[4]. Each of these models have their pros and cons: speed, accuracy, reliance on a multiple sequence alignment (MSA), and memory usage are the biggest considerations when choosing a model to use to predict the structures of proteins. Throughout my time at Colorado State University, I have largely stuck with AF2 because it provides results that have the closest structure predictions to experiment-validated models at the cost of longer run time. Additionally, an outstanding effort was made to make AF2 run locally and easily with LocalColabFold[5].

We applied AF2 as a means to detect epitope subsequences from an antibody:antigen pair, with nothing other than the sequence of both proteins. This is discussed further in Chapter 2.

1.2 High Throughput Crystallization

To train machine learning structure prediction algorithms like AF2, ample amounts of data are needed. However, experimental determination of the structure of a protein is not easy. This is most commonly achieved in one of three ways: crystallization followed by single crystal X-ray diffraction (scXRD), electron microscopy (particularly CryoEM), and nuclear magnetic resonance (NMR) spectroscopy. Each of these techniques has its own set of benefits and drawbacks.

Protein crystallization and scXRD accounts for a vast majority of the crystal structures that are solved and deposited into the Protein Data Bank (PDB). At this writing, approximately 84% of the structures deposited into the PDB were solved this way. After a protein of interest has

been expressed and purified, it is mixed with a few key components to induce a self-assembling, well-ordered crystal lattice. Crystallization conditions are highly specific to each individual protein, but often include a precipitant (e.g. ammonium sulfate or poly-ethylene glycol) to reduce the solubility of the protein thereby encouraging crystal nucleation and growth, a buffer (e.g. HEPES) to establish a particular pH for protein stability and interface strength, and other additives to further tune protein nucleation and growth rates. After crystallization, these crystals are commonly subjected to the addition of a cryo-preservative like glycerol, flash frozen in liquid nitrogen, and then shipped to an X-ray beamline, where a high-powered X-ray is shot at the protein crystal and the resulting diffraction pattern is collected. From this diffraction data, the protein's structure can be solved[6–8].

A competing method, electron microscopy, has provided about 9% of the structures deposited into the PDB. The most recently developed and emerging technique is CryoEM, where protein samples are flash-frozen in a thin layer of vitreous ice, then imaged with an electron microscope. CryoEM is particularly useful for studying large macromolecular complexes and membrane proteins (proteins notoriously difficult to crystalize in their native state). CryoEM rapidly increased in effectiveness due to high quality direct electron detectors[9]. Accordingly, an increasingly diverse of a biomolecular targets can now be observed without crystallization (thereby bypassing an extremely technically challenging step). Flash-freezing proteins and complexes in place allows for the preservation of native conformations, key to elucidating true biological functions[10–13].

NMR spectroscopy is responsible for a further 6% of the structures deposited in the PDB. NMR spectroscopy can be used to determine the structure of the protein while in solution, which can be more representative of their natural state in a cell compared to crystalline structures. NMR

spectroscopy is typically used to solve the structures of smaller proteins due to its limitations in handling larger molecules. This is important because there are minute differences in the shape a protein adopts while crystalline and while free floating in solution – for some circumstances this change is quite important. Two key examples of this are how a protein folds and significant conformational changes a protein may undergo during binding[14,15].

Finding the conditions that induce protein crystallization is a time-consuming and critically skilled-labor intensive process. This complexity arises from the necessity to systematically explore a vast array of experimental conditions, such as varying concentrations of salts, buffers, precipitants, and pH levels. Each protein behaves uniquely under different conditions, and determining the precise set of parameters that will result in high-quality crystals often requires extensive trial and error. Scientists must meticulously prepare numerous samples, each with slightly different compositions, and monitor them over time, which demands significant expertise, precision, and patience. The intricate nature of this process means that even minor deviations can lead to failures, making the role of experienced personnel indispensable.

Given the laborious nature of protein crystallization, automating this process with a liquid handling robot presents a compelling solution. Liquid handling robots can significantly increase throughput by rapidly and accurately preparing hundreds or thousands of crystallization trials simultaneously, far beyond the capacity of manual methods. These robots can precisely control the dispensing of tiny volumes of liquids, ensuring consistent and reproducible sample preparation. Automation reduces the likelihood of human error and frees up skilled researchers to focus on analyzing results and designing new experiments rather than performing repetitive manual tasks. Additionally, robots can work continuously without fatigue, enabling around-the-clock experimentation and accelerating the overall pace of research. By integrating liquid

handling robots into the crystallization workflow, laboratories can enhance efficiency, reduce costs, and improve the likelihood of discovering optimal crystallization conditions, ultimately advancing the field of structural biology more rapidly and effectively.

Several liquid handling machines exist to help accelerate this brute force process, such as the Crystal Gryphon from Art Robbins Instruments and the Mosquito from SPT Labtech. However, these machines are not perfect; as standalone instruments, they carry maintenance costs and may not be sufficiently economical for all laboratories and are expensive; in the neighborhood of $75,000 to fully set up. The Gryphon protein crystallization machine is recognized for its high-throughput capabilities, allowing for the screening of 96 conditions simultaneously. It supports various crystallization plate sizes and offers a range of screening blocks, enhancing accessibility for high throughput screening. The Gryphon's design minimizes evaporation and human error, providing fast and consistent liquid drop distribution. This efficiency is particularly beneficial when setting up multiple plates. However, there are notable drawbacks. For example, the Gryphon’s design was finalized in 2016, and learning a scratch-like language is required to appropriately build protocols and loop with the Gryphon. Additionally, stringent cleaning is required due to the reuse of needles, which can lead to the buildup of salts and PEGs, potentially affecting accuracy and reproducibility, especially when working with microliter volumes.

Furthermore, liquid handling robots like the Gryphon can struggle with highly viscous protein solutions, as viscous protein drops tend to accumulate on the pipettes, hindering successful transfer into the wells. Although there are settings to mitigate this, the problem occasionally persists. Machines dedicated to specific tasks cannot easily be repurposed for other uses. While the Gryphon streamlines the process of setting up crystallization plates, a major part

of the time savings comes from using pre-made reservoir solution screens (which do not exist for all proteins, and are expensive in their own right to set up for custom screens). Alternative liquid handling instruments, such as the Rigaku Alchemist (further increasing the cost by tens of thousands of dollars), are needed to automate the creation of such 96-well plates. In contrast, more general-purpose liquid handling systems like those from Opentrons (priced around $15,000) offer versatility and fine control over all liquid handling steps through programmable Python code, accommodating diverse protein crystallization schemes. These systems, while offering flexibility, also come with their own costs and maintenance considerations. Therefore, while automation with liquid handling robots can significantly enhance the efficiency of protein crystallization experiments, laboratories must carefully consider the upfront costs, ongoing maintenance, and specific needs of their research workflows when choosing the appropriate equipment.

We sought to provide open access, 3D-printable files and python scripts to make the Opentrons adaptable for protein crystallography and crystallization. This is discussed more thoroughly in chapter 3.

1.3 Metal-Organic Frameworks

While protein crystals can be catalytic (e.g. enzyme crystals) the palette of canonical amino acids can limit the scope of available reactions. Indeed, many enzymes rely on metals and other prosthetic groups. We sought to literally and figuratively combine the favorable properties of stabilized protein crystals (e.g. biocompatibility, precise nanostructure) with the dramatically different catalytic capabilities of non-biological catalysts. Specifically, metal-organic frameworks (MOFs) are a rich and diverse class of porous, crystalline materials with diverse applications.

MOFs are a hybrid organic-inorganic material, where small organic molecules act as struts that connect metal ions or clusters, which can be tiled outwards infinitely to create the MOF crystal (**Figure 1.4**). These materials are relatively new in the past 25 years, with an explosion in new structures and chemistries being developed at an exponential rate. The metal atoms or clusters often have uncoordinated sites readily available for use, making them an attractive material for catalysis. MOFs are also highly porous and, much like proteins, can also be rationally designed.

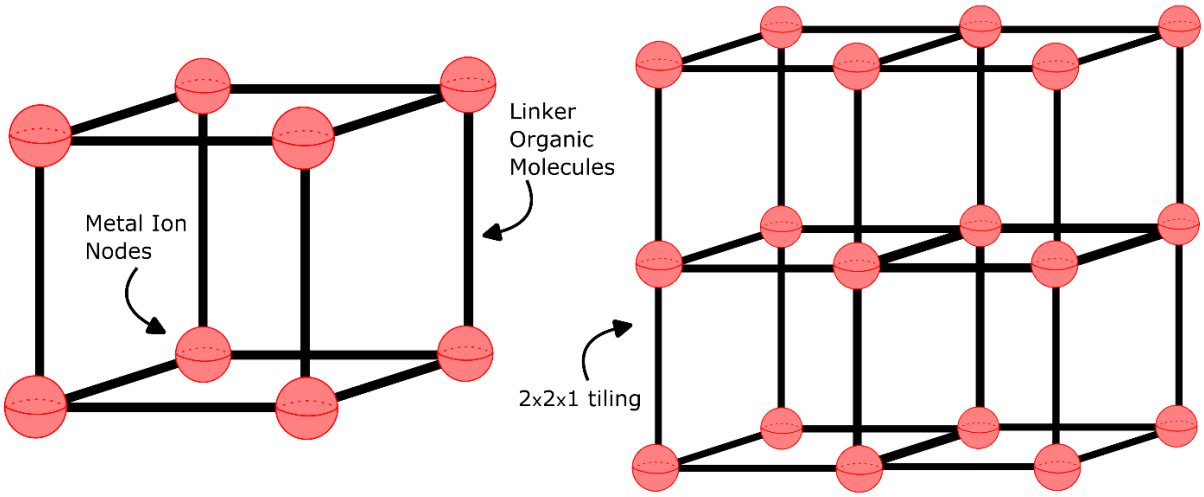


Figure 1.4: A cartoon depiction of a metal organic framework. The metal ions or clusters are depicted as red spheres, and the organic linker molecules are shown as sticks, connecting the metals together. On the right is an extended tiling of the MOF unit cell, exemplifying how a large crystal of the material would be structured at the atomic level.

Due to their high density of functional sites and other favorable physical properties, MOFs are an ideal material for catalytic activity. Catalysis on MOFs often proceeds via vacant metal coordination sites, via organic linker molecules as catalytic sites, and as a host for the

encapsulation of additional catalysts such as nanoparticles and enzymes, and any combination of these catalyst methods – among other creative approaches[16–18].

While there are thousands of MOFs that exist, two are especially pertinent for this discussion. These MOFs are CuBTC (aka HKUST-1 and Basolite C300) and UiO-67. CuBTC is an exceptionally well-studied[19–23] and characterized MOF, and readily grows under a myriad of conditions. While CuBTC has been used in many interesting applications ranging from CO_2 capture to nitric oxide regeneration, we primarily used it because it grows readily, grows bright blue, and has a previously published stable precursor solution. UiO-67 is more water resistant than CuBTC[24], and is very catalytically active to organophosphate bond cleavage[25]; this is particularly interesting because some chemical warfare threats (specifically G-class nerve agents) have an organophosphate bond. This makes UiO-67 an appealing MOF for the catalytic degradation of these nerve agents.

Because of these capabilities, and the modularity of MOFs in general, we sought to install both CuBTC and UiO-67 as proof of concept into a host protein crystal with unusually large solvent channels (13-nm diameter). This is explored and discussed more thoroughly in Chapter 4.

These concepts of predicting protein structure, automating the process of protein crystal production, and exploring the space of MOF installation into porous protein crystals all feed and reinforce into one another (**Figure 1.5**). The mass production of crystals allows for the further exploration of MOF growth into the crystals, and lightening the load of solving crystal crystals via protein automation of protein crystallization. The use of AI could completely close the loop on protein crystallization – attaching a camera with vision to the liquid handling robot would allow the robot to investigate the wells to see if protein crystals have been made, then make

decisions on whether or not they have been. This would effectively boil the problem of protein crystallization down to "supply robot with purified protein and crystallization reagents", then wait a week. The production of MOF protein crystals with their structure solved would allow for the machine learning algorithms to better learn how proteins (and protein crystals) interact with non-proteinaceous molecules – and thus would let us predict new and better combinations of MOFs loaded into protein crystals. While this dissertation doesn't necessarily explore all these opportunities, it is very easy to see where these directions could go in the near future.

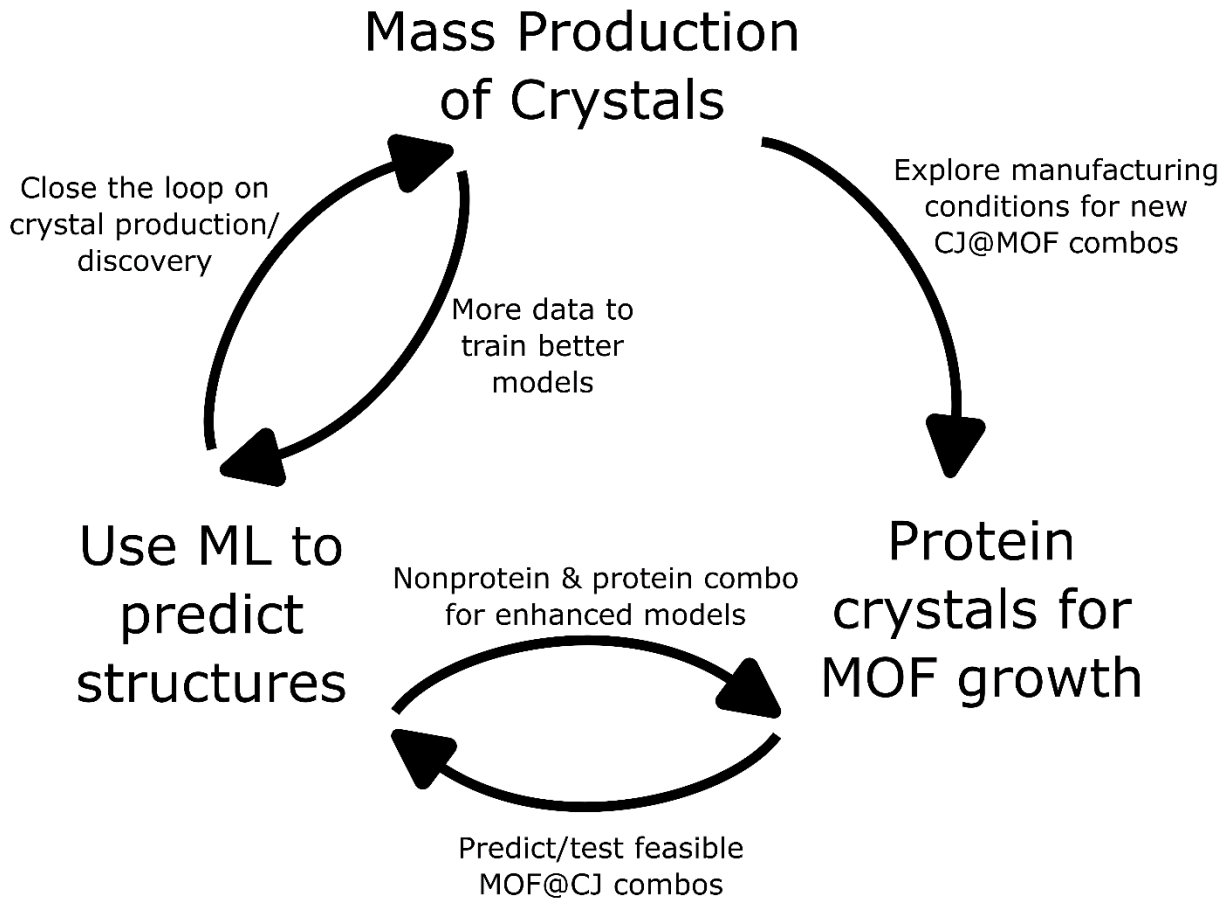


Figure 1.5: Chapter relationship diagram. Relationship between ideas/chapters presented in this dissertation, and what is explored and not explored.

Chapter 2 – PAbFold: Linear Antibody Epitope Prediction using AlphaFold2

2.1 Individual Contributions

My major contributions to this project [1]were developing the Python code necessary for the pipeline, running the scripts, the multiple sequence alignment generations, and the data analysis upon the conclusion of AlphaFold2 running. I conceived the idea at a TagTeam meeting (the recurring meeting for the collaboration between the Snow lab, Geiss lab, and Stasevich lab). Specifically, a discussion came up about identifying the epitope sequence in an antigen, and the time expense of experimentally determining that sequence. AF2 had been around for a little less than a year at this point, and I saw an opportunity to dive into AF2 and explore its potential as a tool. While predicting structures is what it does, we can gain a significant amount of information from those structures – and herein lies the significant efforts of Peptide Antibody Fold (PAbFold). AlphaFold2 has five different sets of weights it can use (ten if you include the multimer sets). PAbFold allows for the analysis of the data via two different methods – a consensus mode, where all the data is used, and a top model mode, where only the data generated via the top scoring model is used. From this, subsequences of the antigen are rank sorted based off their pLDDT values; these are the potential epitope sequences used.

[1] The authors on this paper are Jacob DeRoo, James Terry, Ning Zhao, Christopher Snow, and Brian Geiss. This paper has been submitted to eLife, and we are currently drafting a response to the reviewer's comments and feedback regarding the publication.

2.2 Summary

Defining the binding epitopes of antibodies is essential for understanding how they bind to their antigens and perform their molecular functions. However, while determining linear epitopes of monoclonal antibodies can be accomplished utilizing well-established empirical procedures, these approaches are generally labor- and time-intensive and costly. To take advantage of the recent advances in protein structure prediction algorithms available to the scientific community, we developed a calculation pipeline based on the localColabFold implementation of AlphaFold2 that can predict linear antibody epitopes by predicting the structure of the complex between antibody heavy and light chains and target peptide sequences derived from antigens. We found that this AlphaFold2 pipeline, which we call PAbFold, was able to accurately flag known epitope sequences for several well-known antibody targets (HA / Myc) when the target sequence was broken into small overlapping linear peptides and antibody complementarity determining regions (CDRs) were grafted onto several different antibody framework regions in the single-chain antibody fragment (scFv) format. To determine if this pipeline was able to identify the epitope of a novel antibody with no structural information publicly available, we determined the epitope of a novel anti-SARS-CoV-2 nucleocapsid targeted antibody using our method and then experimentally validated our computational results using peptide competition ELISA assays. These results indicate that the AlphaFold2-based PAbFold pipeline we developed is capable of accurately identifying linear antibody epitopes in a short time using just antibody and target protein sequences. This emergent capability of the method is sensitive to methodological details such as peptide length, AlphaFold2 neural network versions, and multiple-sequence alignment database. PAbFold is available at https://github.com/jbderoo/PAbFold.

2.3 Introduction

Understanding where and how an antibody binds to its target protein is important for understanding how the antibody performs its function, whether that function is neutralizing a pathogen during an immune response, binding an epitope in immunoassays, or labeling a target molecule in a live-cell imaging experiment. However, determining the binding epitope of an antibody can be a time and labor-intensive endeavor with significant expense. Traditionally, antibody epitopes on target proteins have been identified by performing deletion analysis on the target protein to determine if the antibody loses reactivity for the deletion mutants in various immunoassays, which provides the general region of the target protein the antibody binds to. With the advent of widely available chemical peptide synthesis, sequence-specific synthetic peptides can be used for competitive immunoassays (such as enzyme-linked immunosorbent assays (ELISA)) to establish sequences that can effectively compete with the antigen for antibody binding. Peptide mapping experiments are a powerful method for determining the fine sequence of linear antibody epitopes, but these experiments can be relatively expensive and the time between experimental design and data acquisition can be weeks to months due to the need to design and chemically synthesize peptides. Once a peptide has been identified that binds with high affinity and specificity to an antibody antigen binding fragment (Fab), crystal structures can be determined that demonstrate intermolecular interactions between the peptide and antibody. These can then provide a molecular-level explanation for an antibody's binding mode. Finally, with the advent of rapid single B-cell sequencing technologies to analyze humoral immune responses towards vaccination or infection, determining where specific antibody clones bind on an antigen becomes even more challenging due to the need to isolate or synthesize specific antibody genes, produce antibodies, and then perform deletion or epitope mapping experiments

described above to fully understand how and where antibodies bind. These challenges make determining antibody epitopes expensive and time-consuming, and limit the number of antibodies that are characterized in detail.

Antibodies that bind to linear epitopes represent an important subset to molecular biology, as they can be added to recombinant proteins for use in various types of immunoassays. A number of linear antibodies have been developed for use in various immunoassays (ELISA, western blot, immunofluorescence, etc.). The development of computational methods for linear epitope determination could increase the number and quality of new linear epitopes available to the field. Most epitope prediction tools (such as BepiPred[26], ElliPro[27], and ABCpred[28]) are generally designed to predict regions of an antigen that could be recognized by any antibody rather than a specific antibody. These programs also provide no insight into the structural match of the epitope and antibody, potentially making decisions without key structural information that otherwise may be relevant. The challenge in predicting epitopes for a *specific* antibody lies in the complexity of protein-protein interaction dynamics, which includes conformational changes, binding affinities, and thermodynamic stability. Structure based approaches including HADDOCK[29,30] and ZDOCK[29,31] can be used to dock peptides into antibody structures, but these require known peptides for binding. Significant progress has been made to address this problem via deep learning: some of the new and exciting tools are GearBind[32], PALM and A2binder[33], and DSMBind[34]. We point the reader to this review for an excellent overview of some of the tools that have existed for some time, along with a comparison of these tools[35].

Determining antibody-epitope interactions is, at its most basic level, a structural biology problem. Determining what molecular interactions are present between an antibody and its antigen can define the epitope, determine what portions of the epitope and CDR sequences are

responsible for molecular interactions, and provide clues to antibody specificity and affinity. With the advent of highly accurate structural predictions, including the AlphaFold2 (AF2) neural networks[1,36], the ability to accurately predict protein structures, and potential protein-protein interactions, has dramatically increased. AlphaFold2 was trained on existing protein structures and can effectively model new protein structures. Numerous antibodies, antibody Fab regions, and other related constructs with bound target peptides or proteins have been crystalized and deposited into the Protein Data Bank (PDB) (for example[37–40]). These PDB entries represent a valuable training set that may increase the likelihood that AlphaFold2 can successfully predict the structure for antibody-epitope complexes[36,41–43]. The authors of AlphaFold2 multimer[36] comment on the difficulty of predicting antibody-epitope complexes, and results for this are indeed mixed at best[41–43]. One way in which this current report is distinct is our focus on linear epitopes. We hypothesize that the lack of strong competing structure within the short peptide may boost AF2 prediction of scFv-epitope binding predictions relative to conformational epitopes. This problem has precedent, as AlphaFold2 has previously been used to study the interactions between proteins and peptides[41,42]. AlphaFold2's ability to correctly dock independent protein chains can be repurposed to predict how strongly two proteins interact together and extends to predicting the interaction between an antibody and short flexible peptides (linear epitopes) drawn from a larger protein antigen.

To maximize compute efficiency, it is helpful to minimize the size of the system subject to structure prediction. The computational expense of AlphaFold2 scales with the square of the length of the concatenated sequences involved. Fortunately, not all portions of the antibody are critical. Antigen binding by antibodies is primarily dictated by the antigen binding fragment (Fab) containing the variable light (V_L) and variable heavy (V_H) fragments. Conversion of full

antibody sequences into single chain variable fragments (scFv) can significantly reduce structure prediction complexity and compute time. A wildtype scFv sequence can easily be generated directly from translated antibody heavy and light chain DNA sequences. Briefly, the sequences are first divided into framework and complementarity determining regions (CDRs) using Kabat[44] or IMGT[45] nomenclature. A flexible linker sequence (GGGGSGGGGSGGGGS, 15 a.a.) is then added between the new C-terminus of the truncated light chain and the original N-terminus of the shortened heavy chain to generate a single protein sequence that incorporates both antigen-binding chains. The resulting fusion protein often functions in a similar fashion to the original antibody. Another well-known protein engineering strategy for antibodies is "loop grafting", where the CDR loops from one antibody are grafted onto a different framework region. We have recently used this approach to develop scFvs with improved *in vivo* performance[46]. The structures of the novel scFv chimeras can be rapidly and confidently be predicted by AlphaFold2 due to their small size and the extensive immunoglobin representation within sequence databases and the PDB. Excluding the time needed to obtain a multiple sequence alignment (MSA), predicting the structure for a single scFv in complex with a 10-a.a. peptide requires only 1.5 minutes on an NVIDIA A5000 graphics processing unit (GPU). This modest compute time allows a GPU-laden server or workstation to handle large-scale structure prediction of hundreds of related systems. As for the MSA input, a high quality MSA can quickly be obtained via ColabFold[47], which relies on the MMseqs2 MSA server. In our workflow, we repeatedly predict the structure for a fixed single scFv sequence in complex with varying peptide partners. In this case, we do not expect the peptide portion of the MSA to be useful. Therefore, to avoid sending hundreds of nearly identical MSA requests to MMseqs2 MSA server, and to avoid varying information in the MSA, we slightly modified the LocalColabFold code to include the option to

cache the MSA (install available on the GitHub). We generate one cached MSA per epitope scan, where each residue in the query peptide is a glycine.

Several recent papers have attempted to use AlphaFold2 to identify antibody epitopes[48–50], but have primarily focused on computational identification and have not verified their results using new antibodies that are not within the PDB training set. While there are many other structure prediction models other than AlphaFold2[2,51], including some specifically dedicated to predicting antibodies or antibody-like structures[52–55], we chose AlphaFold2 to directly test its ability to correctly identify and place epitopes into an antibody binding cleft. We selected AlphaFold2 due to its widespread use throughout the literature, as well as its ease of installation and modification via the LocalColabFold implementation[47]. In this project we test a method we call PAbFold, a LocalColabFold-based pipeline to identify epitopes for several well-known linear-epitope antibodies from sequence information only. There was a strong correlation between AlphaFold2's confidence in the peptide structure (pLDDT)[56] and the experimentally verified epitope binding sequence. Additionally, we found that AlphaFold2 very accurately predicted the linear epitope of a novel SARS-CoV-2 nucleocapsid-specific antibody (mBG17) with minimal prior epitope information. The molecular interactions predicted by AlphaFold2 were experimentally validated using peptide mapping ELISA experiments. Overall, this work demonstrates that AlphaFold2 has compelling promise for linear antibody epitope discovery from sequence information alone. We also have observed that this emergent linear epitope prediction ability is sensitive to the peptide length and that the performance was optimal when using AlphaFold2-multimer version 2 and older MSAs generated by MMSEQS version 2202 server, rather than the more recent AlphaFold2-multimer version 3 models and MMSEQS version 2302 server.

2.4 Materials and Methods

2.4.1 Software

All structure predictions were completed on a single AMD EPYC 7443 server with two NVIDIA RTX A5000 GPU cards. PAbFold code was written in Python 3.7 and Bash. The only extra Python dependencies are NumPy and Matplotlib. AlphaFold2 calculations were run using an installation of LocalColabFold[47]. Briefly, PAbFold contains 3 stages. In the first stage, a python script 'A_PeptideMapping_prep_submission_files.py' writes FASTA input files for ColabFold. Each FASTA file contains the entire sequence of the subject scFv, a colon ":", and then the candidate linear epitope which represents a small section of the target antigen protein that changes dependent upon both the epitope length (default 10 a.a.) and a sliding window (default 1 a.a.).

After completion of the ColabFold jobs, two different analysis methods are presented in this paper, and both are accessible via the 'B_PeptideMapping_plddt_perres_analysis.py' python script. The first is the 'Simple Max' method, which assesses each peptide window with only the output model that is top ranked by ColabFold (on the basis of ipTM). The AlphaFold2 confidence pLDDT[56] is recorded for each residue within the peptide. Other than the N- and C-terminal residues, each residue is observed within multiple windows. We proceed to calculate (and plot) the maximum pLDDT observed for each residue across the set of sliding window peptides that contain that residue. Thus, in the 'Simple Max' method each residue is considered independently. To obtain aggregate scores for each peptide window, we sum the maximum pLDDT associated with each member residue. This method is sensitive in that any isolated high-confidence residue placements in the top ranked AlphaFold2 peptide prediction can increase the

score, but a high aggregate peptide score could arise from multiple, mutually inconsistent peptide binding poses. Our second, complementary analysis method instead focuses on recognizing full peptide poses of elevated AlphaFold2 confidence. We refer to the second method as the 'Consensus method' because it begins by averaging the per-residue pLDDT across the five AlphaFold2 models. We then compute the average pLDDT for each peptide. For visual inspection, scripts output a heat map for the average per-residue pLDDT and a bar-chart that for the subsequent per-peptide average pLDDT. In this case, we simply rank top peptides based on the per-peptide average pLDDT. Scripts are available at https://github.com/jbderoo/PAbFold.

2.4.2 Antibody sequences

Sequences and references for antibodies, scFvs, and antigens can be found in **Table 2.1**. To create an scFv, the complementarity determining regions or loops of an antibody are identified via the Kabat numbering scheme. The loops are then spliced onto the scFv backbones of the 15F11 and 2E2 as previously described by our group[46]. The scFv sequences are aligned with their CDR loops and flexible linkers highlighted in **Table 2.2**.

Table 2.1 Raw sequences for every protein used in this study.

>mBG17 scFv MAEVKLEESGGGLVQPGGSMKFSCVASGFTFSDYWMNWVRQSPDKGLEWVAEIRLKSNNYATHYAASVK GRFTISRDDSKSSVYLQMNNLRAEDSGIYYCTRSAMDYWGQGTSVTVSSGGGGSGGGGSGGGGSDIVMSQ SPSSLAVSVGEKITMSCKSSQSLLYTSDQKNYLAWFQQKPGQSPKLLIFWASTRDSGVPDRFTGSGSGTDFTL TISSVKAEDLAVYYCQQFYNYPRTFGGGTKLEI

>mBG17-15F11 MAEVKLVESGGGLVKPGGSLKLSCAASGFTFSDYWMNWVRQTPEKRLEWVAEIRLKSNNYATHYAASVK GRFTISRDNAKNTLYLQMSSLRSEDTAIYYCARSAMDYWGQGTTLTVSSGGGGSGGGGSGGGGSDIVLTQS PASLTVSLGQRATISCKSSQSLLYTSDQKNYLAWYQQKPGQPPKLLIYWASTRDSGIPARFSGSGSGTDFTLNI HPVEEEDAATYYCQQFYNYPRTFGAGTKLEI
>mBG17-2E2 MAEVQLVESGGDLVKPGGSLKLSCAASGFTFSDYWMNWVRQTPDKRLEWVAEIRLKSNNYATHYAASVK GRFTISRDNAKNTLYLQMSSLKSEDTAMYYCARSAMDYWGQGTSVTVSSGGGGSGGGGSGGGGSDIVLTQ SPASLAVSLGQRATISCKSSQSLLYTSDQKNYLAWYQQKPGQPPKLLIYWASTRDSGIPARFSGSGSGTDFTL NIHPVEEEDAATYYCQQFYNYPRTFGGGTKLEI
>mBG17 Fab VH:VL MYLGLNCVFIVFLLKGVQSEVKLEESGGGLVQPGGSMKFSCVASGFTFSDYWMNWVRQSPDKGLEWVAEI RLKSNNYATHYAASVKGRFTISRDDSKSSVYLQMNNLRAEDSGIYYCTRSAMDYWGQGTSVTVSS:MDSQ AQVLMLLLLWVSGTCGDIVMSQSPSSLAVSVGEKITMSCKSSQSLLYTSDQKNYLAWFQQKPGQSPKLLIF WASTRDSGVPDRFTGSGS
>mBG17 experimentally derived epitope DDFSKQLQQS
>mBG17 target protein sequence – SARS CoV-2 Nucleocapsid protein MSDNGPQNQRNAPRITFGGPSDSTGSNQNGERSGARSKQRRPQGLPNNTASWFTALTQHGKEDLKFPRGQ GVPINTNSSPDDQIGYYRRATRRIRGGDGKMKDLSPRWYFYYLGTGPEAGLPYGANKDGIIWVATEGALNT PKDHIGTRNPANNAAIVLQLPQGTTLPKGFYAEGSRGGSQASSRSSSRSRNSSRNSTPGSSRGTSPARMAGN GGDAALALLLLDRLNQLESKMSGKGQQQQGQTVTKKSAAEASKKPRQKRTATKAYNVTQAFGRRGPEQT QGNFGDQELIRQGTDYKHWPQIAQFAPSASAFFGMSRIGMEVTPSGTWLTYTGAIKLDDKDPNFKDQVILL NKHIDAYKTFPPTEPKKDKKKKADETQALPQRQKKQQTVTLLPAADLDDFSKQLQQSMSSADSTQA
>HA scFv

MAEVKLVESGGDLVKPGGSLKLSCAASGFTFSSYGMSWVRQTPDKRLEWVATISRGGSYTYYPDSVKGRF

TISRDNAKNTLYLQMSSLKSEDTAMYYCARRETYDEKGFAYWGQGTTVTVSSGGGGSGGGGSGGGGSDIE

LTQSPSSLTVTAGEKVTMSCKSSQSLLNSGNQKNYLTWYQQKPGQPPKLLIYWASTRESGVPDRFTGSGSGR

DFTLTISSVQAEDLAVYYCQNDNSHPLTFGAGTKLEL

>HA-15F11

MEVKLVESGGGLVKPGGSLKLSCAASGFTFSSYGMSWVRQTPEKRLEWVATISRGGSYTYYPDSVKGRFTI

SRDNAKNTLYLQMSSLRSEDTAIYYCARRETYDEKGFAYWGQGTTLTVSSGGGGSGGGGSGGGGSDIVLTQ

SPASLTVSLGQRATISCKSSQSLLNSGNQKNYLTWYQQKPGQPPKLLIYWASTRESGIPARFSGSGSGTDFTL

NIHPVEEEDAATYYCQNDNSHPLTFGAGTKLEI

>HA-2E2

MAEVQLVESGGDLVKPGGSLKLSCAASGFTFSSYGMSWVRQTPDKRLEWVATISRGGSYTYYPDSVKGRF

TISRDNAKNTLYLQMSSLKSEDTAMYYCARRETYDEKGFAYWGQGTSVTVSSGGGGSGGGGSGGGGSDIV

LTQSPASLAVSLGQRATISCKSSQSLLNSGNQKNYLTWYQQKPGQPPKLLIYWASTRESGIPARFSGSGSGTD

FTLNIHPVEEEDAATYYCQNDNSHPLTFGGGTKLEI

>HA epitope

YPYDVPDYA

>HA target protein sequence – influenza hemmaglutanin A

MKTIIALSYILCLVSAQKLPGSENRTATLCLGHHAVQNGTLVKTITNDQIEVTNATELVQSSSTGRICDNPHRV
LDGRDCTLIDALLGDPHCDSFQNKEWDLFIERSKAYSNCYPYDVPDYASLRSLVASSGTLEFTTEGFDWTGV
TQNGTSYSCKRGSANSFFSRLNWLHKLNYKYPAQNVTMPNDDKFDKLYIWGVHHPSTDNDQTSLYVQTSG
RVTVSTKRSQQTVVPDIGSRPWVRGISSRISIHWTIVKPGDILLINSTGNLIAPRGYFKIRNGKSSIMKSDALIG
NCNSECITPNGSIPNDKPFQNVNRITYGDCPRYVKQSTLKLATGMRNVPEKQTRGIFGAIAGFIENGWEGMV
DGWYGFRHRNSEGTGQAADLKSTQAAIDQINGKLNRLIKKTNEKFHQIEKEFSEVEGRIQDLEKYVEDTKV
DLWSYNAELLVALENQHTIDLTDSEMNKLFERTRKQLRENAEDMGNGCFKIYHRCDNACIGSIRNGTYNHN
VYRDEALNNRFKIKGVELKSGYKDWILWISFAISCFLLCVGLMGLIMWTCQKGNIRCIRCNICH

>Myc scFv

MEVKLVESGGDLVKPGGSLKLSCAASGFTFSHYGMSWVRQTPDKRLEWVATIGSRGTYTHYPDSVKGRFTI
SRDNDKNALYLQMNSLKSEDTAMYYCARRSEFYYYGNTYYYSAMDYWGQGASVTVSSGGGGSGGGGSG
GGGSDIVLTQSPASLAVSLGQRATISCRASESVDNYGFSFMNWFQQKPGQPPKLLIYAISNRGSGVPARFSGS
GSGTDFSLNIHPVEEDDPAMYFCQQTKEVPWTFGGGTKLEI

>Myc-15F11

MEVKLVESGGGLVKPGGSLKLSCAASGFTFSHYGMSWVRQTPEKRLEWVATIGSRGTYTHYPDSVKGRFTI
SRDNAKNTLYLQMSSLRSEDTAIYYCARRSEFYYYGNTYYYSAMDYWGQGTTLTVSSGGGGSGGGGSGG
GGSDIVLTQSPASLTVSLGQRATISCRASESVDNYGFSFMNWYQQKPGQPPKLLIYAISNRGSGIPARFSGSGS
GTDFTLNIHPVEEEDAATYYCQQTKEVPWTFGAGTKLEI

>Myc-2E2

MAEVQLVESGGDLVKPGGSLKLSCAASGFTFSHYGMSWVRQTPDKRLEWVATIGSRGTYTHYPDSVKGRF
TISRDNAKNTLYLQMSSLKSEDTAMYYCARRSEFYYYGNTYYYSAMDYWGQGTSVTVSSGGGGSGGGGS
GGGGSDIVLTQSPASLAVSLGQRATISCRASESVDNYGFSFMNWYQQKPGQPPKLLIYAISNRGSGIPARFSGS
GSGTDFTLNIHPVEEEDAATYYCQQTKEVPWTFGGGTKLEI

>Myc target protein sequence

MDFFRVVENQPPATMPLNVSFTNRNYDLDYDSVQPYFYCDEEENFYQQQQQSELQPPAPSEDIWKKFELLP

TPPLSPSRRSGLCSPSYVAVTPFSLRGDNDGGGGSFSTADQLEMVTELLGGDMVNQSFICDPDDETFIKNIIIQ DCMWSGFSAAAKLVSEKLASYQAARKDSGSPNPARGHSVCSTSSLYLQDLSAAASECIDPSVVFPYPLNDSS SPKSCASQDSSAFSPSSDSLLSSTESSPQGSPEPLVLHEETPPTTSSDSEEEQEDEEEIDVVSVEKRQAPGKRSE SGSPSAGGHSKPPHSPLVLKRCHVSTHQHNYAAPPSTRKDYPAAKRVKLDSVRVLRQISNNRKCTSPRSSDT EENVKRRTHNVLERQRRNELKRSFFALRDQIPELENNEKAPKVVILKKATAYILSVQAEEQKLISEEDLLRKR REQLKHKLEQLRNSCA
>Myc epitope EQKLISEEDL

Table 2.2 Visually appealing sequence alignment with the CDR loops and linker regions.

```
Kabat numbering    1--------10--------20--------30--------40--------50-------
MYC               -MEVKLVESGGDLVKPGGSLKLSCAASGFTFSHYGMSWVRQTPDKRLEWVATIG--SRGT
MYC-2E2           MAEVQLVESGGDLVKPGGSLKLSCAASGFTFSHYGMSWVRQTPDKRLEWVATIG--SRGT
MYC-15F11         -MEVKLVESGGGLVKPGGSLKLSCAASGFTFSHYGMSWVRQTPEKRLEWVATIG--SRGT
mBG17             MAEVKLEESGGGLVQPGGSMKFSCVASGFTFSDYWMNWVRQSPDKGLEWVAEIRLKSNNY
mBG17-2E2         MAEVQLVESGGDLVKPGGSLKLSCAASGFTFSDYWMNWVRQTPDKRLEWVAEIRLKSNNY
mBG17-15F11       MAEVKLVESGGGLVKPGGSLKLSCAASGFTFSDYWMNWVRQTPEKRLEWVAEIRLKSNNY
HA-scFv           MAEVKLVESGGDLVKPGGSLKLSCAASGFTFSSYGMSWVRQTPDKRLEWVATISRG--GS
HA-2E2            MAEVQLVESGGDLVKPGGSLKLSCAASGFTFSSYGMSWVRQTPDKRLEWVATISRG--GS
HA-15F11          -MEVKLVESGGGLVKPGGSLKLSCAASGFTFSSYGMSWVRQTPEKRLEWVATISRG--GS

Kabat numbering   ---------65---70--------80-----------90---95---------------102
MYC               YTHYPDSVKGRFTISRDNDKNALYLQMNSLKSEDTAMYYCARRSEFYYYGNTYYYSAMDY
MYC-2E2           YTHYPDSVKGRFTISRDNAKNTLYLQMSSLKSEDTAMYYCARRSEFYYYGNTYYYSAMDY
MYC-15F11         YTHYPDSVKGRFTISRDNAKNTLYLQMSSLRSEDTAIYYCARRSEFYYYGNTYYYSAMDY
mBG17             ATHYAASVKGRFTISRDDSKSSVYLQMNNLRAEDSGIYYCTRS-------------AMDY
mBG17-2E2         ATHYAASVKGRFTISRDNAKNTLYLQMSSLKSEDTAMYYCARS-------------AMDY
mBG17-15F11       ATHYAASVKGRFTISRDNAKNTLYLQMSSLRSEDTAIYYCARS-------------AMDY
HA-scFv           YTYYPDSVKGRFTISRDNAKNTLYLQMSSLKSEDTAMYYCARRET-------YDEKGFAY
HA-2E2            YTYYPDSVKGRFTISRDNAKNTLYLQMSSLKSEDTAMYYCARRET-------YDEKGFAY
HA-15F11          YTYYPDSVKGRFTISRDNAKNTLYLQMSSLRSEDTAIYYCARRET-------YDEKGFAY

                  -------110-               1--------10------------24---------
MYC               WGQGASVTVSSGGGGSGGGGSGGGGSDIVLTQSPASLAVSLGQRATISCRASESVDNYG-
MYC-2E2           WGQGTSVTVSSGGGGSGGGGSGGGGSDIVLTQSPASLAVSLGQRATISCRASESVDNYG-
MYC-15F11         WGQGTTLTVSSGGGGSGGGGSGGGGSDIVLTQSPASLTVSLGQRATISCRASESVDNYG-
mBG17             WGQGTSVTVSSGGGGSGGGGSGGGGSDIVMSQSPSSLAVSVGEKITMSCKSSQSLLYTSD
mBG17-2E2         WGQGTSVTVSSGGGGSGGGGSGGGGSDIVLTQSPASLAVSLGQRATISCKSSQSLLYTSD
mBG17-15F11       WGQGTTLTVSSGGGGSGGGGSGGGGSDIVLTQSPASLTVSLGQRATISCKSSQSLLYTSD
HA-scFv           WGQGTTVTVSSGGGGSGGGGSGGGGSDIELTQSPSSLTVTAGEKVTMSCKSSQSLLNSGN
HA-2E2            WGQGTSVTVSSGGGGSGGGGSGGGGSDIVLTQSPASLAVSLGQRATISCKSSQSLLNSGN
HA-15F11          WGQGTTLTVSSGGGGSGGGGSGGGGSDIVLTQSPASLTVSLGQRATISCKSSQSLLNSGN

                  -----34----40--------50--------60--------70--------80-------
MYC               -FSFMNWFQQKPGQPPKLLIYAISNRGSGVPARFSGSGSGTDFSLNIHPVEEDDPAMYFC
MYC-2E2           -FSFMNWYQQKPGQPPKLLIYAISNRGSGIPARFSGSGSGTDFTLNIHPVEEEDAATYYC
MYC-15F11         -FSFMNWYQQKPGQPPKLLIYAISNRGSGIPARFSGSGSGTDFTLNIHPVEEEDAATYYC
mBG17             QKNYLAWFQQKPGQSPKLLIFWASTRDSGVPDRFTGSGSGTDFTLTISSVKAEDLAVYYC
mBG17-2E2         QKNYLAWYQQKPGQPPKLLIYWASTRDSGIPARFSGSGSGTDFTLNIHPVEEEDAATYYC
mBG17-15F11       QKNYLAWYQQKPGQPPKLLIYWASTRDSGIPARFSGSGSGTDFTLNIHPVEEEDAATYYC
HA-scFv           QKNYLTWYQQKPGQPPKLLIYWASTRESGVPDRFTGSGSGRDFTLTISSVQAEDLAVYYC
HA-2E2            QKNYLTWYQQKPGQPPKLLIYWASTRESGIPARFSGSGSGTDFTLNIHPVEEEDAATYYC
HA-15F11          QKNYLTWYQQKPGQPPKLLIYWASTRESGIPARFSGSGSGTDFTLNIHPVEEEDAATYYC

                  -90--------100----
MYC               QQTKEVPWTFGGGTKLEI
MYC-2E2           QQTKEVPWTFGGGTKLEI
MYC-15F11         QQTKEVPWTFGAGTKLEI
mBG17             QQFYNYPRTFGGGTKLEI
mBG17-2E2         QQFYNYPRTFGGGTKLEI
mBG17-15F11       QQFYNYPRTFGAGTKLEI
HA-scFv           QNDNSHPLTFGAGTKLEL
HA-2E2            QNDNSHPLTFGGGTKLEI
HA-15F11          QNDNSHPLTFGAGTKLEI
```

Legend: Heavy chain loops | linker | Light chain loops

2.4.3 Monoclonal Antibody Production

Anti-SARS-CoV-2 nucleocapsid protein (NP) monoclonal mouse antibody mBG17 was previously developed and characterized [57]. Briefly, two BALB/c mice immunized with recombinant NP were sacrificed and primary splenocytes isolated. Splenocytes were fused with Sp2/0 Ag14 myeloma cells and individual hybridoma clones were isolated after eleven days. Hybridoma clones were tested for antibody production against NP via enzyme-linked immunosorbent assay (ELISA) and western blot. Clones were further tested for isotype and cross-reactivity, and V_H (variable heavy chain) and V_L (variable light chain) sequences were determined. The hybridoma clone mBG17 was identified as a SARS-CoV-2 nucleocapsid-specific antibody targeting linear epitope via ELISA and western blot[57]. Generation of recombinant mBG17 and production of recombinant antibody in 293F cells was previously described[57]. The approximate epitope region for mBG17 was determined via western blot with modified recombinant NP proteins containing 40 to 50 amino acid deletions. The epitope location was determined to reside between SARS-CoV-2 nucleocapsid residues a.a. 381-419 based on loss of western blot signal with the a.a. 381-419 deletion[57].

2.4.4 Peptide Competition ELISA

The anti-SARS-CoV-2 nucleocapsid protein mBG17 antibody epitope was experimentally identified using competition enzyme-linked immunosorbent assay (ELISA). Using the previously determined 39 nucleocapsid protein amino acid range for the mBG17 epitope as a starting point, seven overlapping peptides were synthesized (Thermo Scientific) spanning the 39 amino acid region with overlaps of 5 amino. These peptides were termed

Fragment 1 through 7 (**Table 1**). A 96-well ELISA plate was coated with 0.1ug/ml of recombinant SARS-CoV-2 NP [57] overnight at 4ºC. The plate was blocked with 4% (w/v) dry non-fat milk in 1X PBS with 0.1% (v/v) Tween-20 for 1 h shaking at room temperature. While blocking, inhibited recombinant mBG17 antibody samples were produced by incubating 40 µL of antibody with 40 µg (approximately 30 nMol) of a single peptide fragment for one hour at room temperature. Following this, peptide-incubated mBG17 was applied to the blocked nucleocapsid protein coated plate in triplicate and allowed to incubate for 1 h at room temperature while shaking. The plate was rinsed with 0.1% (v/v) Tween-20 in 1X PBS and washed three more times for 5 minutes shaking at room temperature. The plate was then incubated with HRP-conjugated goat anti-mouse polyclonal antibody solution diluted at 1:20,000 in 1X PBS for 1 h shaking at room temperature. After another rinse and three more washes the plate was developed with 1-Step™ Ultra TMB-ELISA Solution (ThermoFisher) before stopping the reaction with an equal volume of 2M H_2SO_4. Solution absorbance at 450 nm was measured using a PerkinElmer Victor X5 multilabel plate reader. Absorbances were averaged within fragment-inhibited sample groups and corrected with the average value of the negative control. These absorbances were then normalized against the absorbance from the group with the highest value before multiplying by 100 to obtain percentage of potential signal.

The effect of single alanine substitutions on fragment 5 (DDFSKQLQQS) peptide binding was determined by competition ELISA using a series of ten alanine-substituted peptides (**Table 2.3**) at a range of concentrations to determine relative competition activity. A modified version of the previously described inhibition ELISA was performed using the unmodified Fragment 5 peptide and the ten alanine-substituted peptides. During the mBG17 inhibition step, the mBG17 antibody solution was incubated with a 4-fold serial dilution of peptides beginning at

40 µg and continuing to ~2.5 ng before being applied to the NP coated plates in triplicate. The remainder of the competition ELISA was carried out as described above.

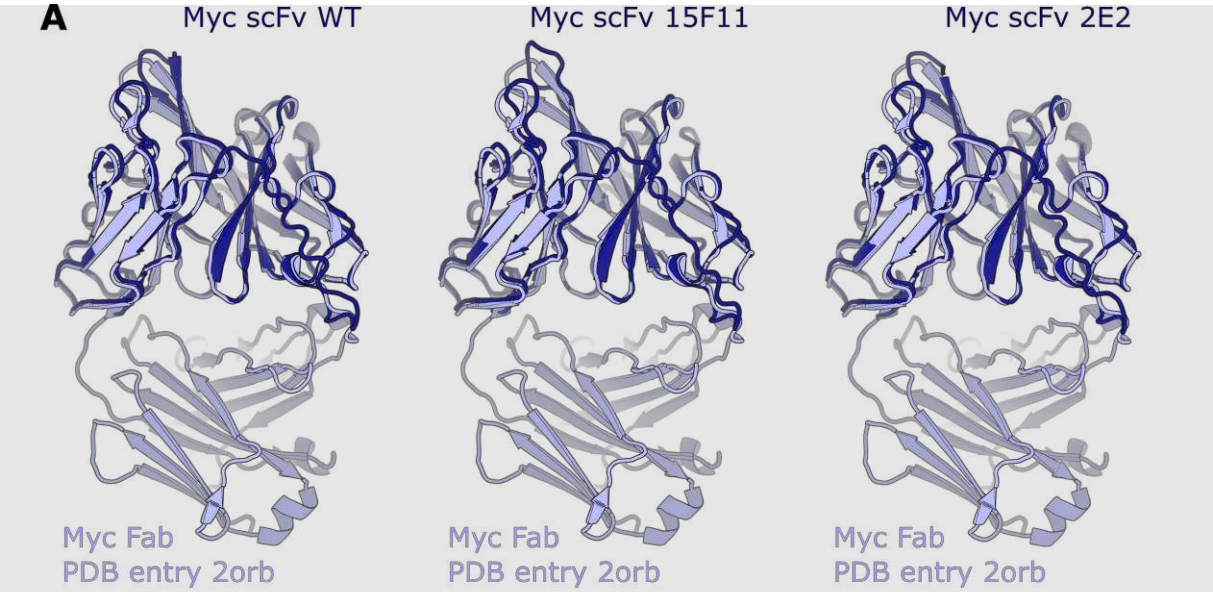
Table 2.3

Peptide Name	Peptide Sequence
Nucleocapsid a. a. 381-390 (Frag 1)	ALPQRQKKQQ
Nucleocapsid a. a. 386-395 (Frag 2)	QKKQQTVTLL
Nucleocapsid a. a. 391-400 (Frag 3)	TVTLLPAADL
Nucleocapsid a. a. 396-405 (Frag 4)	PAADLDDFSK
Nucleocapsid a. a. 401-410 (Frag 5)	DDFSKQLQQS
Nucleocapsid a. a. 406-415 (Frag 6)	QLQQSMSSAD
Nucleocapsid a. a. 411-419 (Frag 7)	MSSADSTQA
Nucleocapsid D401A	ADFSKQLQQS
Nucleocapsid D402A	DAFSKQLQQS
Nucleocapsid F403A	DDASKQLQQS
Nucleocapsid S404A	DDFAKQLQQS
Nucleocapsid K405A	DDFSAQLQQS
Nucleocapsid Q406A	DDFSKALQQS
Nucleocapsid L407A	DDFSKQAQQS
Nucleocapsid Q408A	DDFSKQLAQS
Nucleocapsid Q409A	DDFSKQLQAS
Nucleocapsid S410A	DDFSKQLQQA

2.4.5 Assessment of AlphaFold2 generated scFv structures

We first verified that AlphaFold2 could generate scFv structures that have similar structures to their parent monoclonal antibodies. We chose the 9E10 clone of the anti-Myc antibody as an initial test system, as the scFv sequence is available[58] and has a well-known linear epitope (EQKLSEEDL)[59]. We predicted the wild-type Myc scFv structure and aligned this model to the corresponding Fab crystal structure (PDB entry 2orb) via the align command in PyMOL (**Figure 2.1 A**). The AlphaFold2 predicted scFv was very similar (RMSD value of 0.42Å) to the anti-Myc Fab structure, suggesting that the predicted scFv structure was a suitable starting point

for epitope prediction. We also examined the structures of the Myc CDRs loop grafted onto the 15F11 [60] and 2E2 [46] frameworks, as we have previously observed that loop grafting onto these frameworks can enhance protein folding and solubility[46]. The loop-grafted Myc-2E2 and Myc-15F11 and structures were also similar to the Myc Fab structure (PDB 2ORB)[59] with similar RMSD values of 0.45Å (**Figure 2.1 B**), indicating that they are also reasonable starting points for epitope prediction.



B

Structure	RMSD $_{Cα}$ [Å] vs 2orb
Myc scFv WT	0.42
Myc scFv 15F11	0.45
Myc scFv 2E2	0.45

Figure 2.1 Alignment of AlphaFold2 predicted scFv structures to an anti-c-Myc Fab crystal structure. A) Alignments of AlphaFold2-derived wild-type Myc scFv, Myc-2E2 scFv, and Myc-15F11 scFv structures with a Myc Fab crystal structure (PDB: 2orb). Predicted scFv structures are shown in dark blue, 2orb Myc Fab structures are shown in light blue. **B)** RMSD values comparing structural similarities between the wild-type Myc scFv, Myc-2E2 scFv, and Myc-15F11 scFv structures with a Myc Fab crystal structure (PDB: 2orb) were computed by the PyMOL align command.

2.4.6 Development of Python-based scripts for automated scFv:peptide structure prediction

We developed a series of Python scripts that automate the process of epitope prediction and analysis with AF2. A_Peptide_Mapping_prep_submission_files.py accepts a linear scFv sequence and a linear full-length antigen sequence, and processes the antigen sequence into a series of short peptides with custom peptide length and sliding window sizes (default parameters are 10 amino acid peptides with a 1 amino acid sliding window). It then adds lines for each scFv:peptide pair to a FASTA file. Structures are then predicted via LocalColabFold for each scFv:peptide pair with AlphaFold2 in parallel on two NVIDIA RTX A5000 GPUs. The python script B_PeptideMapping_plddt_perres_analysis.py parses the AlphaFold2 output structures to extract per-residue pLDDT for the peptide residues in each scFv:peptide pair. Conf_plot_and_top10.py will plot the maximum pLDDT (across all host peptides) scores as a function of amino acid position within the antigen sequence and ranks predicted peptides based on ΣpLDDT scores for the 'Simple max' method. To use the 'Consensus' method, include the –all-models flag when running B_PeptideMapping_plddt_perres_analysis.py. We also supply a python script that replicates how we present the data called all_model_analysis.py for use. An overview of the method is shown in **Figure 2.2**. AF2's failure to predict whole antigen structure coupled with the scFv is highlighted in **Figure 2.3.**

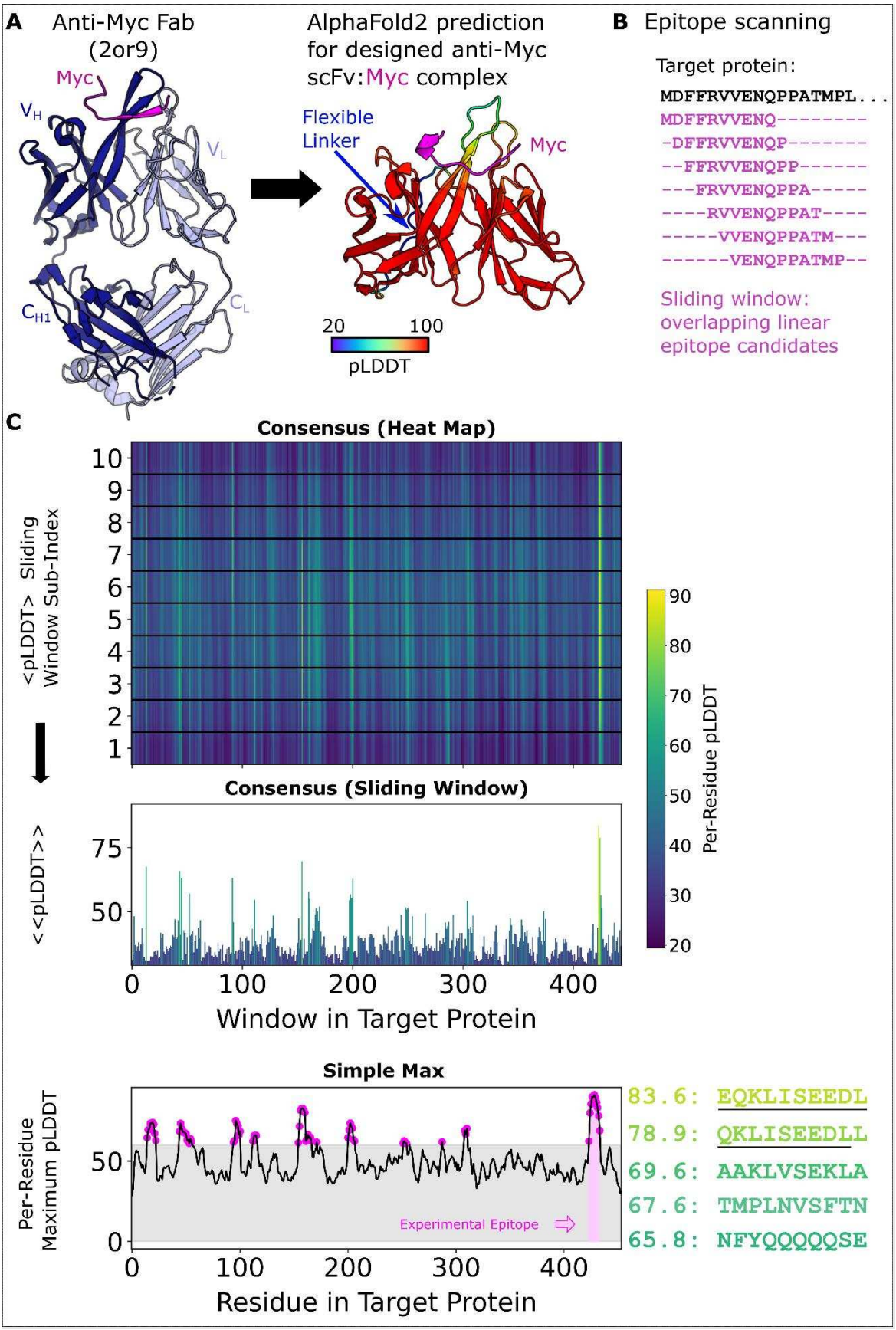

A
Anti-Myc Fab
(2or9)
Myc
V_H
V_L
C_{H1}
C_L
AlphaFold2 prediction
for designed anti-Myc
scFv:Myc complex
Flexible
Linker
Myc
20
100
pLDDT
B
Epitope scanning
Target protein:
MDFFRVVENQPPATMPL...
MDFFRVVENQ--------
-DFFRVVENQP--------
--FFRVVENQPP------
---FRVVENQPPA-----
----RVVENQPPAT----
-----VVENQPPATM---
------VENQPPATMP--
Sliding window:
overlapping linear
epitope candidates
C
Consensus (Heat Map)
<pLDDT> Sliding
Window Sub-Index
10
9
8
7
6
5
4
3
2
1
90
80
70
60
50
40
30
20
Per-Residue pLDDT
Consensus (Sliding Window)
<<pLDDT>>
75
50
0
100
200
300
400
Window in Target Protein
Simple Max
Per-Residue
Maximum pLDDT
50
0
Experimental Epitope
0
100
200
300
400
Residue in Target Protein
83.6: EQKLISEEDL
78.9: QKLISEEDLL
69.6: AAKLVSEKLA
67.6: TMPLNVSFTN
65.8: NFYQQQQQSE

Figure 2.2 PAbFold pipeline for linear epitope prediction. A) Antibody V_H and V_L protein sequences are used to generate scFv sequences, either based on the native antibody sequences or loop grafting complementarity determining regions (CDRs) onto either the 2E2 or15F11 antibody framework regions (2E2 shown). **B)** The target antigen sequence is parsed into a list of small overlapping peptide sequences, with peptide step and window size parameters adjusted as needed. Rank ordered peptides are output, and partial epitope sequences are underlined manually to highlight the identification of the correct sequence. **C)** The scFv sequences from Panel A are co-folded with each of the peptide sequences derived from the target antigen in parallel batch mode on a GPU server. pLDDT scores from each structure prediction experiment are collected and scores are presented in their sliding window, both as a heat map organized along the length of the target antigen sequence and a bar chart that shows the per-peptide average pLDDT (Consensus Method). Additionally, the Simple Max data is presented in the third and final panel.

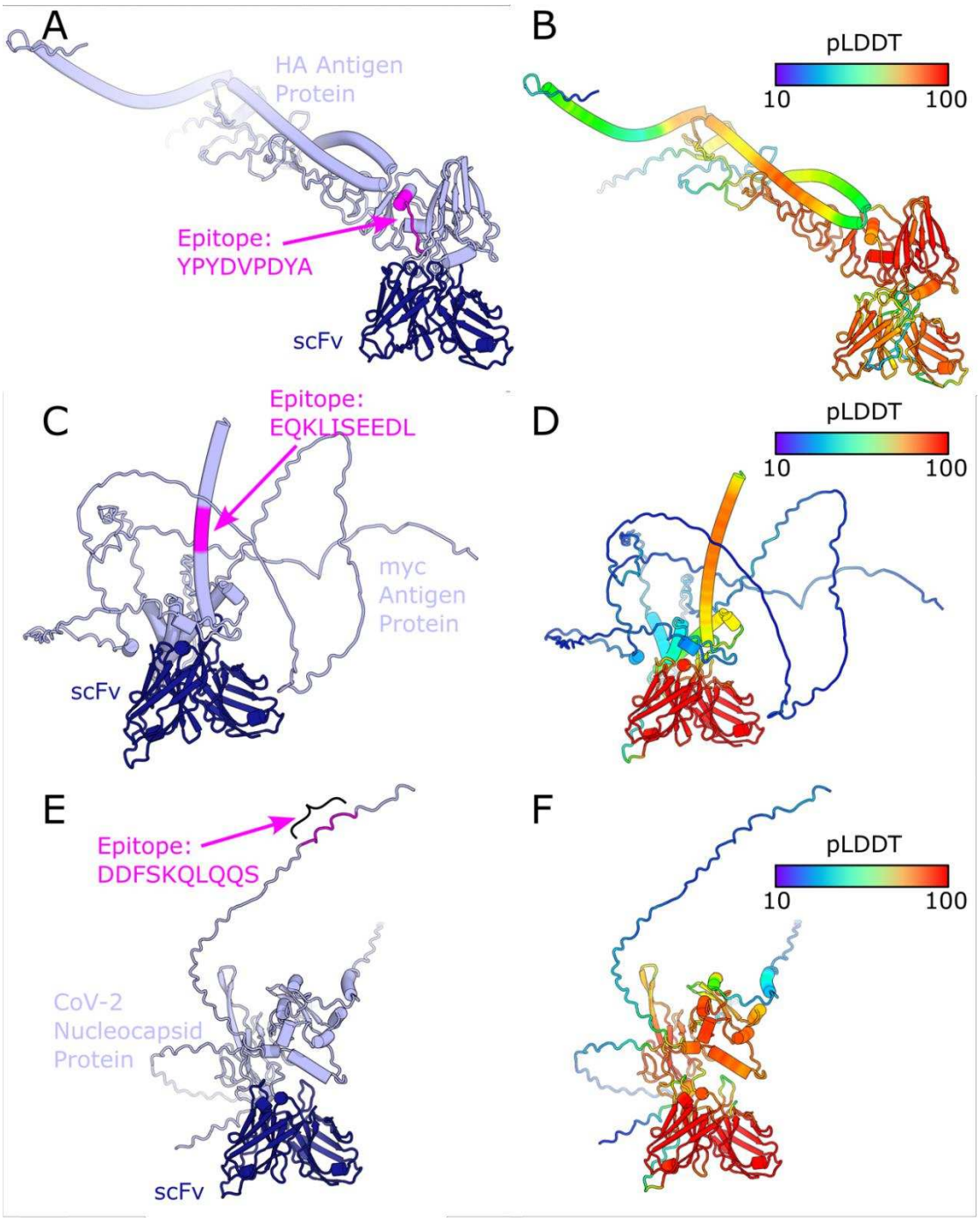


Figure 2.3 Alphafold2's best attempt to dock whole sequences with the respective sequence's scFv. **A)** The whole HA protein structure and scFv complex as predicted by AF2, with the correct epitope sequence highlighted in magenta. **B)** Shows the same structure by highlighted by confidence (pLDDT) of the structure with AF2. Similarly, the entire Myc protein-scFv complex are shown with **C)** the correct epitope highlighted in magenta and **D)** the confidence of the structure shown, and again for the mBG17 N-protein-scFv complex in **E)** and **F)**.

2.5 Testing of scFv:peptide structure prediction method using the Myc Epitope

We first tested the PAbFold method with the anti-Myc-scFv described in [61], using the full-length human Myc proto-oncogene protein sequence as the antigen. We initially used an antigen peptide length of 10 and a 1 amino acid sliding window. Given these parameters, the 9 a.a. Myc epitope motif (EQKLISEEDL) appeared intact within one of the 10-mer peptides, with subsets of the 8, 9, 11, and 12 a.a. appearing in neighboring sliding peptide windows. PAbFold generated predicted structures, each of which took an average of ~200 seconds to process. The entire process took approximately 12 hours on our GPU server. AlphaFold2 placed all peptides into or near the traditional antigen binding site between the CDR loops (**Figure 2.4**). The average confidence (mean pLDDT across residues) for these peptides ranged from 20 to 90. When we inspect the consensus confidence for each residue in each sliding window (**Figure 2.5**), the expected Myc peptide epitope (EQKLISEEDL) was one of several peptides with high average pLDDT. The second highest ranked peptide in this analysis (QKLISEEDLL) was a near perfect match for the expected epitope. We consider this window to be a successful prediction. Perhaps surprisingly, the peptide window with the exact match (EQKLISEEDL) did not score particularly well due to its average pLDDT of 51.0. In this instance, the expected epitope sequence did not stand out when plotting the maximum observed per-residue pLDDT for each residue (**Figure 2.5 A bottom, E**).

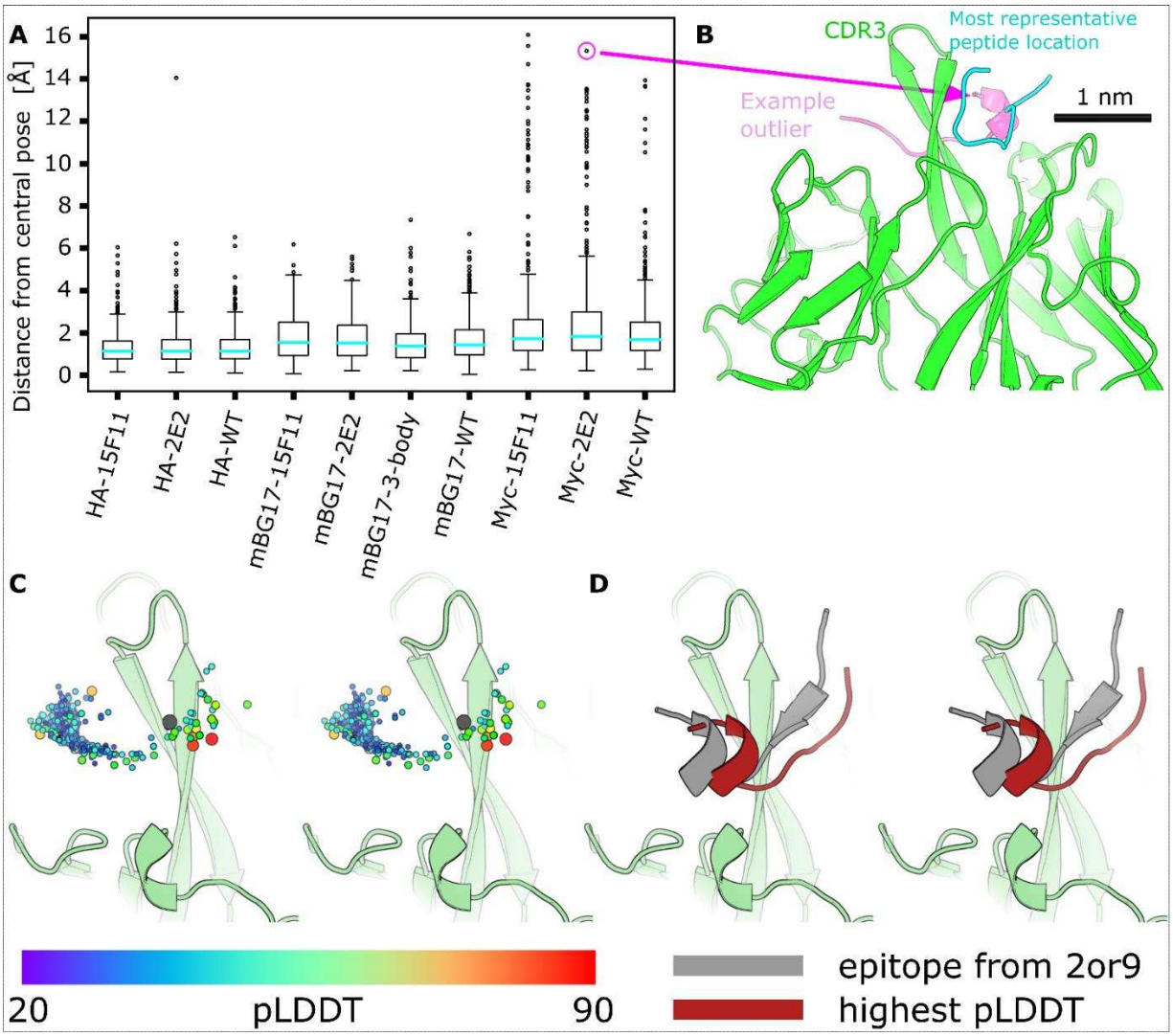


Figure 2.4 AlphaFold2 places all peptides near the CDR loops. The predicted Cα coordinates for all scFv (excluding the flexible linker) were extracted, and all were aligned together using the Kabsch algorithm (48, 49). With the scFvs structurally aligned, an all-against-all RMSD was calculated for the epitope peptides. To visually represent each peptide as a single point, the coordinates for all epitope atoms were averaged. The "central" exemplar epitope (cyan) is the peptide with the smallest sum of RMSD to all other peptides. **A)** The average and quartile for peptide placement relative to the central peptide via Box-and-Whisker plot reveals that AlphaFold2 largely places all epitopes in the same area. The Myc CDRH3 runs through the middle of a traditional paratope pocket, it isn't a "cradle" for the epitope to sit on. AlphaFold2 places peptides on both sides of the CDRH3, causing significant spread in the peptide placement. **B)** An example of an exemplar, most-central predicted peptide structure (cyan) for the peptide PKSCASQDSS (cyan) bound to the Myc-2E2 scFv (green) that is distant from an example outlier peptide (magenta, peptide PHSPLVLKRC, center-to-center distance 14.8 Å). All peptide placements are still in contact with CDRH3, consistent with a strong AlphaFold2 bias to place

peptides in a typical antibody binding site. **C)** The Myc-2E2 scFv (pale-green) and the average epitope placement (cyan) peptide alongside the crystal structure solution of the Myc epitope (grey). Remaining peptide placements are represented as a cloud of spheres at the mean peptide position. Each peptide sphere is colored and sized by epitope pLDDT (ranging from 20 to 90). Although AlphaFold2 frequently placed peptides on the opposite side of the CDRH3 from the Myc epitope (grey), it was not confident in these peptide placements (low, small, blue pLDDT spheres). In contrast, some of the peptides placed around the CDRH3, and in positions similar to the native epitope (grey) were placed with higher pLDDT confidence (increasingly large spheres trending from green to yellow to orange and red). **D)** The top ranked peptide as predicted by PAbFold with sequence QKLISEEDLL (red) and the crystal structure solution of the Myc epitope (grey).

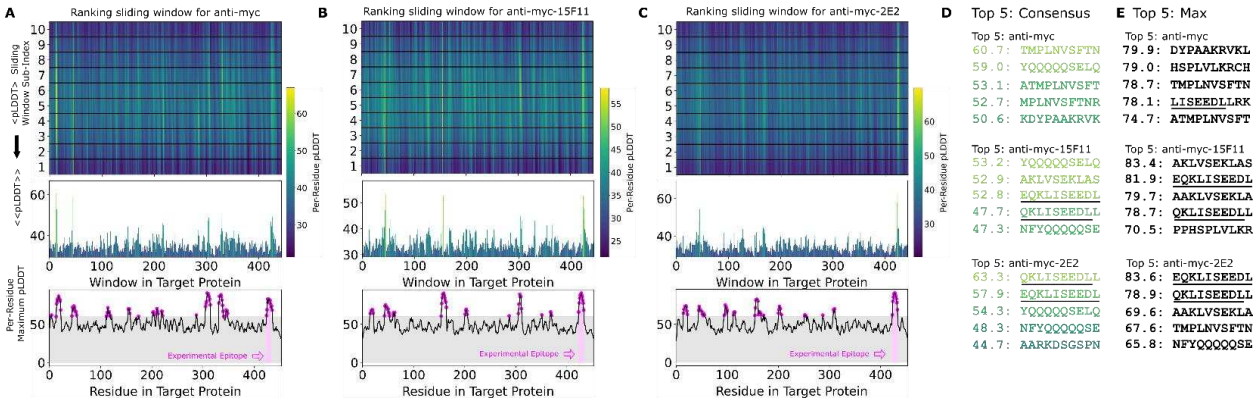


Figure 2.5 The Alphafold2-based PAbFold method predicted the Myc linear epitope in different scFv backbones. The anti-Myc V_H and V_L antibody sequences were used to generate either A) wild-type Myc scFv or loop grafted chimeric B) Myc-15F11 or C) Myc-2E2 2E2 scFv variants. The Myc proto-oncogene protein sequence (Genbank NP_001341799.1) was used as the target antigen and processed into 10 amino acid overlapping peptides with a 1 amino acid sliding window. The structure for each scFv:peptide pair was predicted with AlphaFold2 in batch mode on two NVIDIA A5000 GPUs. Average consensus pLDDT values for each scFv:peptide window are illustrated, as well as the maximum pLDDT observed for each residue in any window (bottom). **D)** Top ranking binding peptides based on average consensus pLDDT. **E)** Top ranked binding peptides based on summing per-residue maximum pLDDT. For D and E, underlining represents overlap with the reported Myc epitope (EQKLISEEDL).

We proceeded to test predictions with two engineered scFv chimeras where loop grafting was used to place the Myc recognition CDRs onto two antibody framework regions with high *in*

vivo performance, generating Myc-15F11 and Myc-2E2 scFv sequences. Epitope prediction performance was markedly improved with the chimeric scFvs (**Figure 2.5 B, C**). Specifically, the QKLSEEDLL peptide window became the top ranked peptide on the basis of average consensus pLDDT. In the case of Myc-2E2 (**Figure 2.5 C**), the average confidence for the correctly predicted epitope was particularly high compared to alternate peptide windows, and another close match to the expected epitope (EEQKLSEED) was ranked within the top 5 peptides (**Figure 2.5 D**). Ranking epitopes using the Simple Max analysis was similar; the region containing the correct epitope was nearly top ranked for Myc-15F11 and was top ranked for Myc-2E2 (**Figure 2.5 E**). Thus, AlphaFold2 was able to more clearly detect authentic Myc antibody epitope using CDRs loop grafted onto the 2E2 or 15F11 frameworks, relative to the native Myc scFv framework.

To investigate the superior epitope recognition performance of the chimeric Myc scFvs, we aligned the Cα coordinates for the predicted scFv structures (predicted with and without the target epitope) to the reference crystal structure and calculated the RMSD for all backbone positions (N, Cα, C, O) and the loops (**Figure 2.6**). Notably, regardless of the Myc scFv variant, the CDR loop RMSD improved by more than 1Å when the epitope was present. Secondly, consistent with the improved epitope prediction performance for the chimeric scFvs (15F11 and 2E2), the epitope peptide QKLISEEDL was placed more accurately for those predicted structures than in the WT scFv (**Figure 2.6**). We could not discern an obvious structural difference between the WT and chimeric scFvs that explains the structure prediction performance gap.

scFv	Apo			Docked		
	BB Ca RMSD	Loop all backbone RMSD	Epitope all atom RMSD	BB Ca RMSD	Loop all backbone RMSD	Epitope all atom RMSD
Myc	0.65	2.87	NA	0.47	1.75	6.69
Myc-15F11	0.62	3.06	NA	0.51	1.51	2.45
Myc-2E2	0.61	2.96	NA	0.51	1.61	2.68

scFv	Apo			Docked		
	BB Ca RMSD	Loop all backbone RMSD	Epitope all atom RMSD	BB Ca RMSD	Loop all backbone RMSD	Epitope all atom RMSD
HA	0.56	1.39	NA	0.58	1.25	3.2
HA-15F11	0.56	1.32	NA	0.6	1.26	3.1
HA-2E2	0.58	1.21	NA	0.6	1.27	3.1

Figure 2.6 RMSD comparison (all numbers have units of Å) for AlphaFold2 predicted scFv structures compared to reference crystal structures, **A)** 2or9 (Myc) and **B)** 1frg (HA), respectively. The loops of the scFv more closely mimic the crystal structure when the epitope peptide is present. The backbone also undergoes subtle changes during docking that make it slightly more similar to the crystal structure. These structures were aligned by identifying the framework residues in all structures, then aligning the framework region Cα with the Kabsch algorithm [62,63]. Specifically excluded from this process were the heavy and light CDR loops of the structures, as well as the flexible linker structure that connects the heavy and light chains due to the inherent floppy, unstructured nature of this region. After aligning the framework regions of the AlphaFold2 predicted structures and the crystal structures (2or9 and 1frg respectively), an RMSD of these Cα was calculated and is reported as the first column 'BB Cα RMSD'. Without further alignment, loop placement was analyzed with an all backbone RMSD by calculating the RMSD between the C, Cα, N, and O along the backbone of all residues in the scFv that were not used for the framework superimposition. This RMSD is reported in the second column as 'Loop all backbone RMSD'. Finally, to investigate peptide predicted placement and potential scFv:epitope interactions, an all-atom RMSD was calculated between the crystal structure and the AF2 predicted peptide structure (no additional alignment). Because the apo structure lacks a peptide position, this is only reported in the 'Docked' category and is in the 3rd column labeled 'Epitope all atom RMSD'. One script was written for each scFv (Myc and HA), and can be found in the Zenodo deposition of our data (https://zenodo.org/records/10884181) because this analysis is not a key part of PAbFold. Briefly this analysis reveals that all three HA scFv variants have predicted framework regions and loop regions in the apo structures that closely match the reference structure (0.56-0.58 Å and 1.21-1.39 Å). Accordingly, when the cognate epitope peptide is present, it can be placed with relatively high accuracy for all three scFvs (3.1-3.2 Å), with only small changes in the loops (1.39 Å to 1.25 Å, 1.32 Å to 1.26 Å, and 1.21 Å to 1.27 Å). In contrast, the apo structures for the three Myc scFvs have a much higher deviation in the loop regions (2.87 to 3.06 Å). When the epitope peptide is added, there is significant motion in the

loops consistent with an "induced fit" description. In the two chimeric Myc scFvs (Myc-15F11 and Myc-2E2) the final loop RMSD is reduced to 1.51-1.61 Å, and the epitope peptide is successfully predicted (2.45-2.68 Å). However, despite a lower apo-state loop RMSD (2.87 Å), the loop RMSD for the wild-type Myc scFv only drops to 1.75 Å, and the epitope peptide placement does not match the experimental structure (6.69 Å). This is consistent with the failure of the wild-type Myc scFv AlphaFold2 predictions in **Figure 2.5**.

2.5.1 Assessment of peptide length, sliding window size, and position on AlphaFold2 scFv:peptide structure prediction

Our initial selection of the 10 a.a. window was intended to match or slightly exceed the size of known epitopes such as Myc and HA. We next assessed how different peptide sizes and sliding window lengths would affect epitope prediction accuracy and run time. We re-ran the Myc-2E2-scFv:peptide complex prediction calculations varying peptide size between 8, 9, 10, and 11 (with a fixed sliding window size of 2) or varying the sliding window size to 1 or 5 (with a fixed peptide size of 10). We observed that using a sliding window of 2 a.a. provided nearly the same level of accuracy and resolution as the 1 a.a. Ultimately, we determined that our original peptide size of 10 amino acids and sliding window of 1 a.a. provided highest resolution data possible (**Figure 2.7**), and therefore maintained a peptide size of 10 and a sliding window length of 1 for our remaining experiments.

We then predicted the complex structure for Myc-2E2 with various negative control peptides: A_{10}, $(GS)_5$, $(GGGGS)_2$, and G_{10} to determine how non-binding peptides are docked and scored (**Figure 2.7 I, J**). We again observed that AlphaFold2 placed all peptides into the traditional antigen binding between the CDR loops, but the reported peptide scores for the negative controls were particularly low (29 – 41). These results indicate that AlphaFold2 "knows" where antigens bind in antibody or scFv structures and attempts to model any peptide

partner into this region, but the low pLDDT scores indicate confidence in the interactions are quite low.

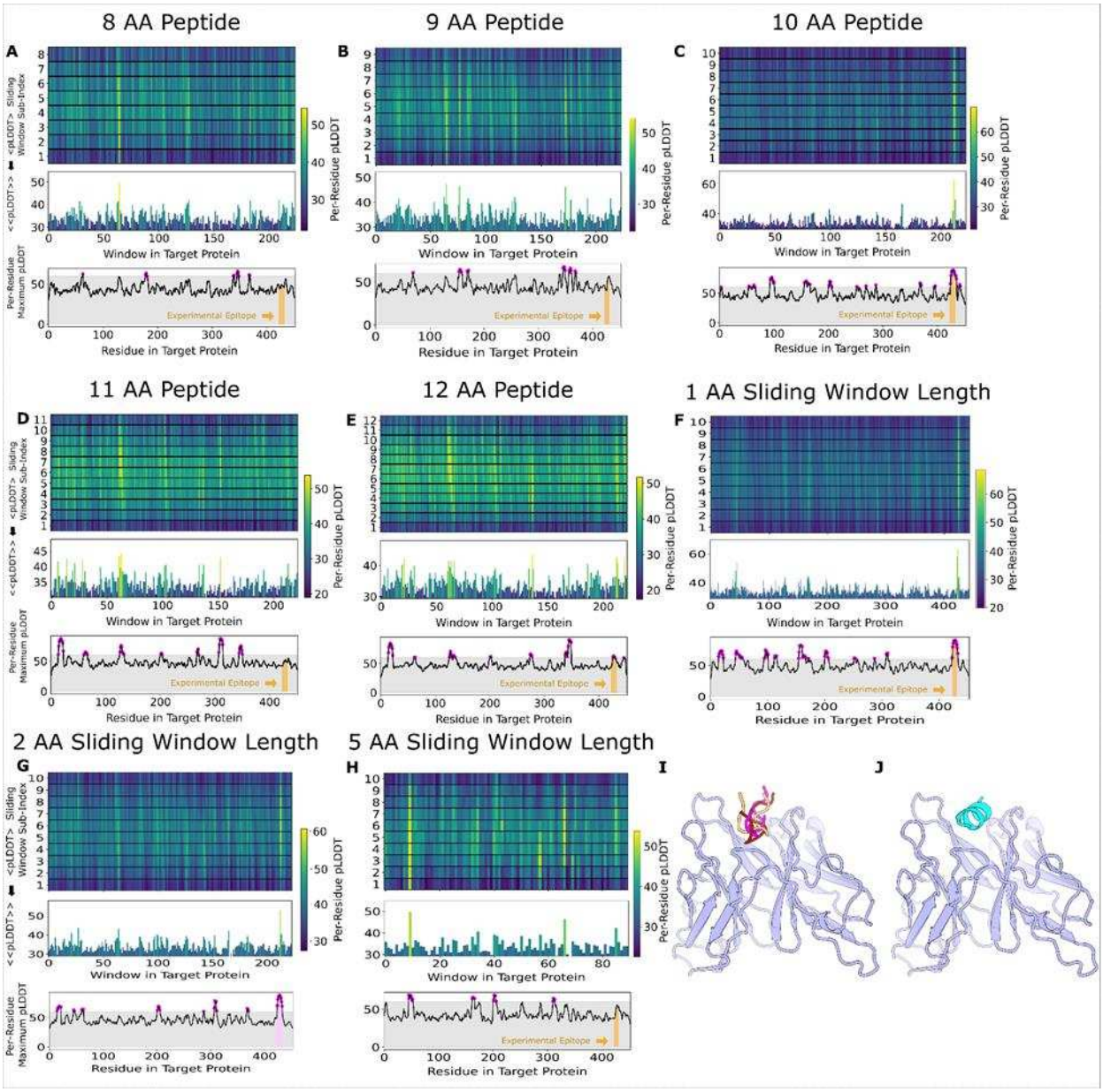


Figure 2.7 Assessment of peptide size and sliding window sizes on epitope prediction efficacy. Myc-2E2 scFv:peptide structures were predicted with peptides of 8 (**A**), 9 (**B**), 10 (**C**), 11 (**D**), and 12 (**E**) amino acid lengths derived from the Myc protein with a sliding window of 2 amino acids, and pLDDT scores from each predicted structure were plotted against the Myc amino acid position and sliding window length target. **F)** Negative control peptides bind to antibody binding sites, but with poor pLDDT scores. Similarly, with a fixed peptide length of 10 and a sliding

window step size of 1 (**F**), 2 (**G**), and 5 (**H**), we can see the practical epitope detection outcome was similar for a sliding window of 1 and 2, but resolution and accuracy were reduced for a sliding window step size of 5. To more fully illustrate the strong learned bias that AlphaFold2 has for placing any peptides among the CDR loops, we predicted the structure of Myc-2E2 in complex with several control peptides. These negative control peptides bind to the generally expected antibody binding site, but with poor pLDDT. **I)** GSx5 in magenta (GSGSGSGSGS) had a score (mean peptide from Simple Max method pLDDT) of 29.5. $(GGGGS)_2$ in orange (GGGGSGGGGS) had a score of 31.9. G_{10} in red (GGGGGGGGGG) had a score of 33. Lastly, **J)** A_{10} in cyan (AAAAAAAAAA) had a score of 41 and is the only negative control peptide to have an alpha-helical secondary structure (presumably due to the increased alpha helical propensity of alanine).

We also tested if AlphaFold2 could detect the Myc epitope if it was inserted as an epitope tag within different positions of a heterologous protein. We created a synthetic antigen by adding the Myc epitope within the 99-a.a. unrelated HIV-1 Gag protease protein sequence at either the N- or C-terminus or in the middle of the protein sequence, and used PAbFold to detect the Myc peptide (**Figure 2.8**). In each case, the average consensus pLDDT was highest for the inserted epitope, such that the authentic epitope would be top ranked and prioritized for testing. Thus, as expected for a sliding window analysis, the epitope position within the antigen was no barrier to detection.

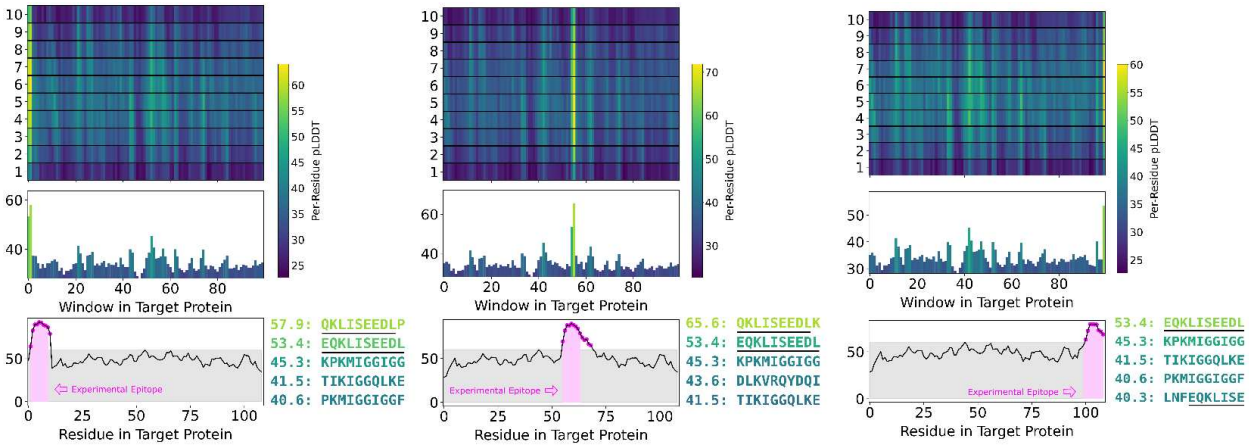


Figure 2.8 PAbFold epitope detection is independent of position within target sequence. The Myc epitope (EQKLISEEDL) was added into the beginning, middle, or end of the 99-a.a. HIV protease sequence (Genbank Accession: NP_705926.1) prior to epitope scanning structure prediction. Positions of the Myc epitope sequence added to in the **A)** N-terminus **B)** middle and **C)** C-terminus of the HIV protease sequence. **D)** Highlights the ranked sequences recovered from each experiment in A, B, and C.

2.5.2 Testing of the PAbFold method using the HA Epitope

Based on our success detecting the Myc epitope, we sought to determine if our method could detect a different well-known linear peptide, HA, derived from positions 114-126 within the Influenza A virus hemagglutinin protein (YDVPDYASLR). Using an anti-HA scFv sequence that had been previously generated[46,61], we generated new HA-15F11 and HA-2E2 scFvs loop grafted sequences. We used the same procedure described above to predict structures for influenza A virus HA derived peptides on HA-scFv (**Figure 2.9 A**), HA-15F11-scFv (**Figure 2.9 B**) and HA-2E2-scFv (**Figure 2.9 C**). In the HA case, the expected epitope was ranked highly for all three scFv variants, but when assessing entire peptides by average consensus pLDDT was only ranked in the top 5 for the HA-15F11-scFv. These results, in combination with the Myc results described above, indicate that AlphaFold2 can accurately detect linear antibody epitopes

in antigen sequences, and that grafting CDR loops onto alternative scFv backbones may increase the noise-to-signal ratio, making the identification of correct epitopes more accurate.

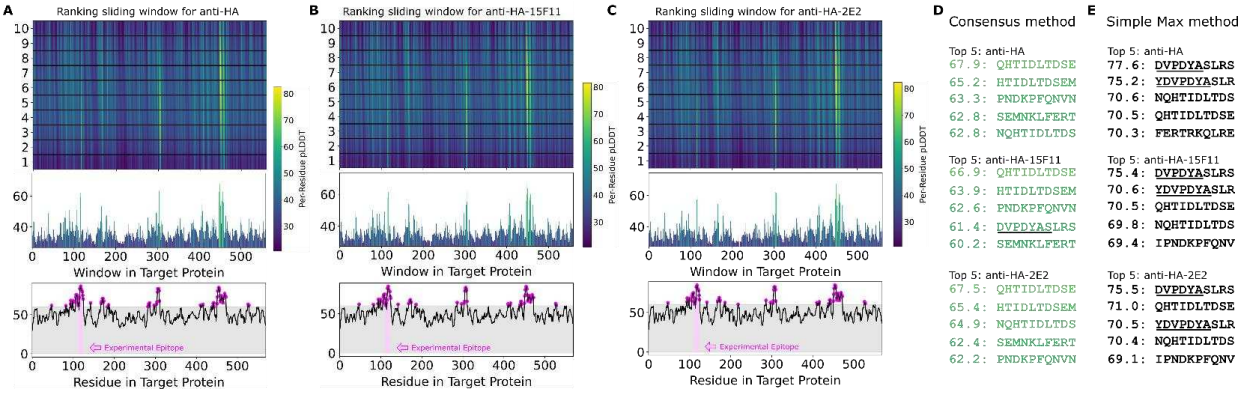


Figure 2.9 Alphafold2 can accurately predict the HA linear epitope in different scFv backbones. The anti-HA VH and VL antibody sequences were used to generate either **A)** wild-type scFv or CDR loop grafted onto the **B)** 15F11 or **C)** 2E2 antibody backbones. The Influenza A virus hemagglutinin protein sequence (Genbank AUT17530.1) was used as the target antigen and processed into 10 amino acid overlapping peptides with a 1 amino acid sliding window. The structures for each scFv:peptide pair were predicted with Alphafold2, and pLDDT values for each scFv:peptide pair are shown. **D)** The top-ranking epitope sequences via pLDDT scores are reported via the consensus method. Sequence underlining represents overlap with the known HA epitope (HA a.a. 114-125: YDVPDYASL). **E)** The top-ranking epitope sequences via pLDDT scores are reported via the simple max method.

Like the Myc system, trends are observed with the HA system regarding loop placement. Although not as extreme, the loops for all HA scFvs undergo movement that make it more closely match the crystal structure (PDB entry 1frg). Again, the epitope placement of predicted structures of the chimeric scFvs more closely mimicked the deposited crystal structure than the WT scFv (**Figure 2.6**).

2.5.3 Determination and experimental validation of a novel linear antibody epitope

The Myc and HA monoclonal antibodies are well known and several crystal structures (Myc PDB: 2or9, peptide bound (2009) | HA PDB:1frg, peptide bound (1994)) have been solved[46,59,61,64], raising the possibility that AlphaFold2 has incorporated these antibody or epitope structures into its training set. The AlphaFold2 training set was reported to exclude chains of less than 10, which would eliminate the myc and HA epitope peptides. Nonetheless, to guard against the possibility that the AlphaFold2 models have incorporated specific knowledge into the training set thereby directly probing if PAbFold epitope scanning can predict a linear antibody epitope without *a priori* knowledge of the antibody or antigen sequence, we tested if PAbFold can predict the epitope sequence of a recently developed antibody lacking structural information available in the Protein Data Bank. The mBG17 mouse monoclonal antibody was generated in response to the COVID-19 pandemic, the antibody V_H and V_L sequences were determined, and the epitope was localized to a. a. 381-419 via Western blot analysis of deletion mutants of the nucleocapsid protein[57]. mBG17 was not included in AlphaFold2's training or test set, making it an ideal test case for *de novo* epitope prediction.

The mBG17 monoclonal antibody was converted to wild-type scFv, 15F11-scFv, and 2E2-scFv using the same procedures used for Myc and HA scFv. As an additional control calculation (labeled "3-body"), we used AlphaFold2 to predict the structure for a 3-protein complex (the peptide, and the disconnected nontruncated mBG17 V_H and V_L variable domain sequences). All 4 Fab variants (WT scFv mBG17, 15F11-mBG17 scFv, 2E2-mBG17 scFv, and 3-body mBG17) were screened against all 10 a.a. peptides with a 1 a.a. sliding window, as with Myc and HA. In all 4 cases, AlphaFold2 predicted that the top ranked peptides were located in

the a.a. 381-419 region of the SARS-CoV-2 nucleocapsid protein, and more specifically residues a.a. 400-415 (**Figure 2.10 A, B, C, and D**). The top scoring peptide for all three scFv variants was the 402-411 window (DFSKQLQQSM) (**Figure 2.10 E, F**). The strong AF2 preference for peptides from this C-terminal segment was particularly evident in the average consensus pLDDT analysis.

We next sought to experimentally verify the minimal linear epitope for mBG17 to determine how closely the AlphaFold2 prediction corresponded to our experimental data. Seven 10 a.a. peptides that overlapped by 5 a.a. each were synthesized and used in competition ELISAs with mBG17 monoclonal antibody and recombinant SARS-CoV-2 nucleocapsid protein (**Figure 2.10 G, H**). The peptide corresponding to a.a. 401-410 showed almost complete competition of mBG17 binding to the SARS-CoV-2 nucleocapsid protein in the ELISA, whereas none of the other peptides were able to compete for mBG17 binding to nucleocapsid. Peptides a.a. 296-405 and a.a. 406-415 overlap a.a. 401-410 at the N- and C-terminus, respectively, but neither was able to compete, indicating that mBG17 binds a.a. 401-410 on both sides of a.a. 405 and a.a. 406**.** An alignment of all the peptides used in the overlapping peptide competition ELISA experiments showed that peptide sequence DDFSKQLQQS represents the experimentally determined epitope for mBG17, nearly identical to the epitope predicted by AlphaFold2 (**Figure 2.10 H:** DDFSKQLQQS). These results demonstrate that the PAbFold pipeline was able to very accurately predict the region that an antibody binds to a novel linear epitope that is not present in AlphaFold2's training set.

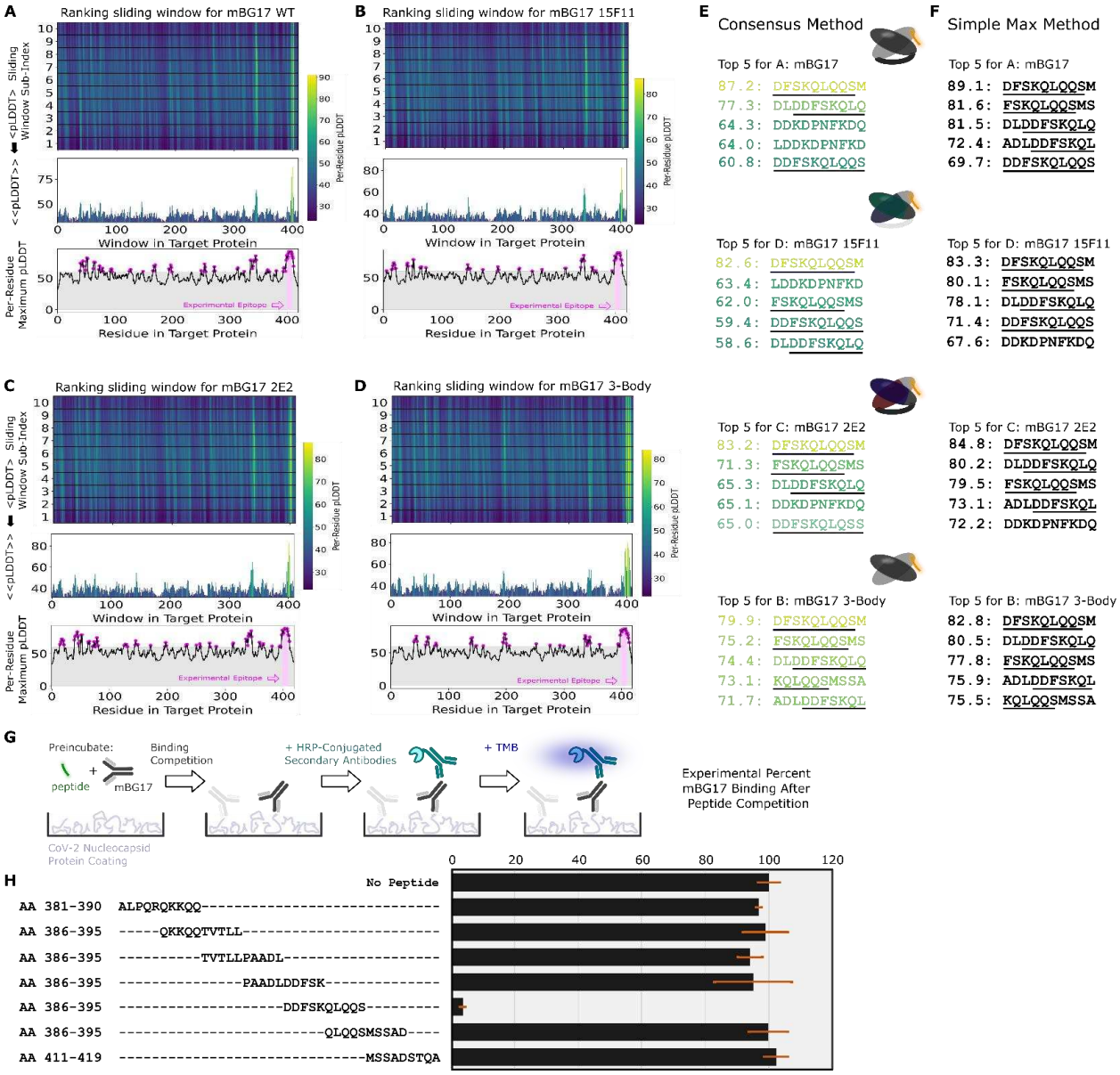


Figure 2.10 The AlphaFold2-driven PAbFold epitope scan method can accurately identify a linear epitope for a novel SARS-CoV-2 antibody. Antibody VH and VL sequences for SARS-CoV-2 nucleocapsid protein targeted antibody were used to generate scFv sequences **A)** WT, **B)** 15F11, **C)** 2E2 or native VH and VL sequences **D)** 3 body). Variant scFv sequence in complex with peptide windows from the SARS-CoV-2 nucleocapsid protein (Genbank Accession: YP_009724397) were subjected to AlphaFold2 structure prediction. The top 5 peptides ranked by either the **E)** Consensus method or the **F)** Simple Max method, with the underlined sequence highlighting the experimentally verified sequences and a cartoon schematic for each system shown. **G)** Competition ELISA schematic for assessing the ability of synthetic peptides derived from the SARS-CoV-2 nucleocapsid protein. **H)** Amino acid windows showing binding interference, with mBG17 binding to SARS-CoV-2 nucleocapsid protein (n = 3). Percentage of

binding values were calculated from the no-peptide control. Alignment of synthetic peptides corresponding to SARS-CoV-2 nucleocapsid a. a. 381-419. Peptide a. a. 401-410, which demonstrated mBG17 competition.

2.5.4 Fine-characterization of the mBG17 epitope and comparison to the predicted AlphaFold2 model

To further experimentally characterize the binding of the mBG17 to the a.a. 401-410 (DDFSKQLQQS) peptide and compare experimental data with the predicted AlphaFold2 model, we designed and synthesized ten additional peptides, each containing an alanine point mutation at one position in the a.a. 401-410 peptide. The peptides are labeled D1A, D2A, F3A, S4A, K5A, Q6A, L7A, Q8A, Q9A, and S10A. Competition ELISAs were performed using increasing concentrations of each peptide to better assess differential binding (**Figure 2.11A**). As expected, WT (a.a. 401-410) peptide showed strong competition, although Q9A showed slightly better competition. This could be attributed to alanine's propensity to be in an alpha-helical coil ($Prop_{A,\ AHC} = 0$) vs glutamine's propensity to escape it ($Prop_{Q,\ AHC} = 0.39$) [65], thus further stabilizing the Q9A alpha helix. D1A showed no change in competition, indicating that D1 was not involved in binding. Peptides with substitutions K5A, Q6A, and S10A showed minor reductions in competition, S4A showed a moderate reduction on competition, whereas resides D2A, F3A, L7A, and Q8A all showed strong reductions in competition. These data indicate that the key interactions between mBG17 and the a.a. 401-410 peptide are residues D2, F3, L7, and Q8, with S4 playing a moderate role and D1, K5, Q6, Q9 and S10 playing negligible roles in binding.

Finally, we compared the experimental data shown above with the best scoring mBG17:DDFSKQLQQ model generated by AlphaFold2 (**Figure 2.11 B, C**). The AlphaFold2

model suggests that residue D2 forms a hydrogen bond with mBG17 a.a. Y34, residue F3 forms a hydrophobic interaction with mBG17 a.a. L185, residue S4 lacks a hydrogen bond partner, residue L7 forms a hydrophobic interaction at the base of the binding cleft with mBG17 a.a. A104, and residue Q8 hydrogen bonds with the backbone carbonyl of Y34 and the backbone amide of W35. Residues that experimentally showed no or minimal effects on competition (D1, K5, Q6, Q9) are all predicted to interact primarily with the solvent and lacked visible interactions between the peptide and scFv sequence. In summary, the AlphaFold2-driven PAbFold prediction was remarkably consistent with the experimental alanine scanning data, suggesting that the prediction of the mBG17 linear epitope location was accurate due to the correct prediction of the structural details for how that linear epitope binds to the antibody.

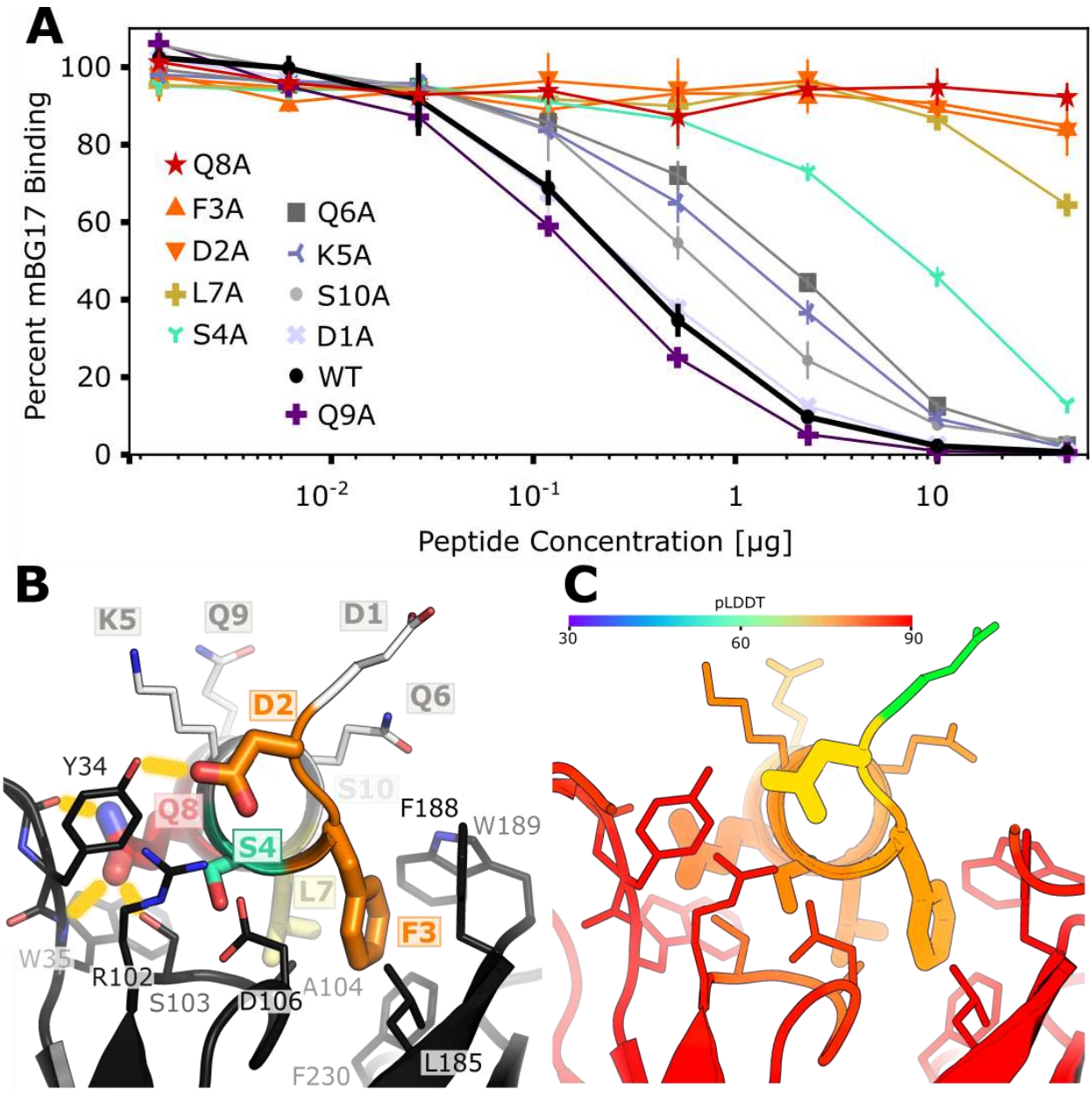


Figure 2.11 The Alphafold2-Driven PAbFold method accurately predicts molecular interactions between a linear epitope and a scFv A) Competition ELISA assessing the ability of synthetic alanine mutant peptides derived from the SARS-CoV-2 nucleocapsid nucleocapsid protein (a. a. 401-410: DDFSKQLQQS) to interfere with mBG17 binding to SARS-CoV-2 nucleocapsid protein (n = 3). Percentage of binding values were calculated from the no-peptide control. **B)** AlphaFold2 model for mBG17-15F11 scFv bound to a. a. 401-410 peptide (the average peptide pLDDT was 83.5). Residues that display sharply reduced binding to mBG17 upon mutation to alanine in competition ELISAs (D2, F3, S4, L7, Q8) are shown as warm-colored thick sticks. Predicted hydrogen bonds between the peptide and the scFv are depicted by yellow bars. Sites where mutation to alanine was less disruptive to binding (Q6A, K5A, S10A, D1A, and Q9A) are

depicted as thin sticks with cool colors. The carbon atoms of residues in panel B are colored according to the corresponding data in panel A. **C)** The same AlphaFold2 model for the mBG17-15F11 scFv bound to a.a. 401-410 colored with confidence (pLDDT) as predicted by AF2.

2.6 Discussion

In this project we assessed the ability of an AlphaFold2-based linear epitope scan pipeline we call PAbFold (Peptide:Antibody Fold) to predict linear antibody epitopes using just antibody and antigen sequences. We first assessed the quality of scFv models produced by AlphaFold2. We then developed a series of Python scripts that accept scFv and whole antigen protein sequences as inputs, parse the antigen protein sequences into short overlapping peptides, run batch predictions for each scFv:peptide pair, and output two peptide scoring schemes based on the peptide per-residue pLDDT scores as a metric for AlphaFold2 model confidence.

Binding of the expected epitope to the WT-Myc scFv could only be detected via the consensus method, but either analysis method could readily detect the expected epitope bound to the chimeric Myc scFvs. Conversely, the alternate analysis method (Simple Max) performed better with respect to ranking the expected HA epitope binding to the WT and chimeric anti-HA scFv variants. In the HA case, performance was comparable for both the WT and chimeric scFv variants.

It is important to note that binding of scFv variants to sequences other than the expected epitopes may be statistically unlikely but not impossible. For example, consider the peptide ATMPLNVSFT near the N-terminus of the Myc proto-oncogene protein sequence. In the context of the WT anti-Myc scFv this peptide had slightly higher average consensus pLDDT (52.4 rather than 51.0) than a peptide (QKLISEEDL) that closely matched the expected epitope. In the

absence of direct experimental evidence, predicted affinity for this unexpected sequence is not necessarily incorrect, though the lack of comparable predicted binding to the 15F11 and 2E2 chimeric scFv variants further decreases the likelihood. In the future, it might be useful to assess peptide binding via consensus across scFv variants.

Lastly, we tested this process on a novel antibody generated by our group targeting the SARS-CoV-2 nucleocapsid protein (mBG17) and found the method performed significantly better than with Myc and HA. Either analysis method could very easily flag peptide windows containing the authentic experimentally validated epitope. This worked for the WT scFv, the chimeric scFv variants, and even a structure with disconnected heavy and light chain domains. Experimentally, we cleanly validated the AlphaFold2 prediction using a peptide competition ELISA assay to experimentally determine the mBG17 epitope. Confidence in the AlphaFold2 prediction was further buoyed via alanine scanning peptide competition ELISAs that verified the importance of the key binding interactions predicted by AlphaFold2.

Identification of antibody V_H and V_L sequences from monoclonal B-cells has become a routine task, with sequence information obtainable via various sequencing technologies such as next generation sequencing and nanopore sequencing for a relatively low cost. As a result, the determination of the epitope in service of a deeper understanding of how antibodies bind their antigen is an increasingly notable bottleneck. An experimental epitope determination campaign can take weeks or months of work, but with the advent of AlphaFold2 and the epitope prediction method we describe here, an antibody and its antigen could be sequenced in a few days (often through contract research organizations for low cost) and accurate linear epitope predictions generated within less than a day, dramatically epitope validation throughput as well as providing detailed predictions for the molecular features of antibody-epitope interaction.

Conformational epitopes are structured antigens that are found during many immune responses, and prediction of these epitopes from antibody and antigen sequences would be a significant boon to the field of biology. For example, conformational epitope prediction coupled with single-cell B-cell sequencing would allow for detailed analysis of antibody maturation during immune responses to vaccines or pathogen infection, helping better define how the immune response to infection evolves over time and how evolution of antigen sequences affects the antibody response. In this work we did not focus on using AlphaFold2 to predict conformational epitopes primarily because of the complex structures that conformational epitopes possess. Literature reports suggest that prediction of the complexes between antibodies and both whole antigens and conformational epitope proteins has proven to be very difficult for AlphaFold2, and indeed the authors themselves make this observation[36,66,67]. Notably, the structures that proved most difficult to predict for AF2 and other tools in the CASP15-CAPRI154 challenges were antibody-antigen complexes[68]. Reports suggest that a mix of both statistics-based approaches (neural networks like AF2) and physics-based approaches (such as Rosetta) predict optimal antibody-antigen complexes[69]. Indeed, if we attempt to predict binding of our scFvs to intact antigen proteins (**Figure 2.3**), we find no predictive capability. When predicting scFv:peptide complexes, it may be the case that AlphaFold2 is able to thoroughly evaluate an induced fit for the peptide due to both its length (small sample space) and its propensity to not adopt a strong competing structure. In contrast, a larger and complicated structure may be more challenging to move during the AlphaFold2 structure prediction or recycle steps. Additional complexities may arise in extreme induced conformational changes during docking. Recent reports indicate that progress is being made in predicting the binding locations of conformational epitopes[70,71].

We observed that the ability of AlphaFold2 to successfully predict the epitope peptide binding is quite delicate. First, epitope prediction was highly sensitive to the peptide length (**Figure 2.7**) with minimal predictive power for peptide length other than 10 a.a. Further investigation of this sensitivity would be a useful avenue for future research. Perhaps with enhanced sampling, epitopes can be detected within longer peptides (e.g. 11 a.a., 12 a.a., etc.). Methodological tuning of this type could ultimately help illuminate the path to increasingly difficult protein-protein binding prediction problems. Similarly, we have likewise determined that epitope scanning performance was sensitive to changes in the underlying AlphaFold2 neural networks and the MSA. Specifically, unless otherwise noted, all data in this report was obtained using ColabFold version 1.5.2 and the 5 neural networks that comprise AlphaFold2 multimer version 2 (mm2). Likewise, the MSAs we use were obtained from the MMSEQS server (and cached) when the default sequence databases were UniRef30 2202 and PDB70 220313. They have since been updated to PDB30 2302 and PDB100 230517. For a complete description, see the change logs on the github for ColabFold (https://github.com/sokrypton/ColabFold#colabfold--v152).

Insofar as protein-peptide prediction is an emergent "off-label" capability for AlphaFold2 that is not part of the training sets, further training of the models or other changes can degrade performance. Benchmarking performance can be difficult when there are multiple moving targets. The most recent calculations we have analyzed were using ColabFold version 1.5.2 which was current as of February 19, 2023. The changes from ColabFold 1.5.2 to 1.5.5 (current as of this writing) are limited to version control and ensuring ColabFold still works on Google Colab, and therefore will not change the calculation performance. Relative to ColabFold 1.3 (the current method at the outset of this project), ColabFold 1.5.2 embodied two substantial changes.

First, ColabFold 1.5.2 used the updated AlphaFold multimer (mm) version 3 by default. Second, the backend server MMSEQS ([72] and (https://github.com/soedinglab/MMseqs2)) that supplies MSAs also underwent updates, namely the database updates. Upon evaluation, we found that the recent default methods (ColabFold 1.5.2) still predicted the epitope successfully for the mBG17 system (**Figure 2.12**). However, the ColabFold 1.5.2 default methods had a pronounced decline in PAbFold performance for the HA and Myc systems. Specifically, the combination of mm3 and the revamped ColabFold MSA server tended to be less discriminating compared to the default settings for ColabFold 1.3 (ColabFold 1.3 was the most up to date version when this project was initialized). The updated configuration flagged diverse peptide sequences with elevated pLDDT values (**Figure 2.13 and Figure 2.14**) resulting in the loss of successful epitope predictive power. While testing ColabFold 1.5.2 with the most recent MSA server, but reverting the AlphaFold2 models to mm2, the outcome improved, with experimentally validated sequences rising to the top more frequently than when using mm3, but still falling short in ranking the experimentally validated epitope sequence embedded within the antigen. However, when previously cached MSAs were paired with mm2 (using ColabFold 1.5.2), performance was maximized. Furthermore, we attempted to recreate the MSA databases locally with similar but not identical results to queueing the server with databases UniRef30 2202 and PDB70 220313 (**Figure 2.15**). Additionally, the MMSEQS team ([72] and (https://github.com/soedinglab/MMseqs2)) graciously rebuilt a server we could query using LocalColabFold that mimicked the original UniRef30 2202 and PDB70 220313 database set up as closely as possible on their end. The MSA that was generated from these databases was used, and still did not perform as well as the original MSAs that were generated upon first retrieval and generation (**Figure 2.16**). As a negative control, we repeated all calculations without using any

MSAs, and only relying upon the sequence to make a structural prediction. As expected, all epitopes were scored very poorly (**Figure 2.17**). Despite our significant efforts, it is unclear why our initial results cannot be perfectly recapitulated, but the difference has been traced to detailed MSA contents (**Figure 2.18**), resulting in differences in correct epitope identification. These results are summarized in (**Figure 2.19**).

One key lesson of this research effort is that caching the MSAs proved to be very useful as a method to guard against changes in the performance of 3rd party tools. We recommend that future methods development work using LocalColabFold adopt the strategy of caching MSAs when feasible. It is also our hope that by describing the latent ability of AlphaFold2 to predict scFv-binding epitopes that this ability will be preserved and enhanced in future iterations.

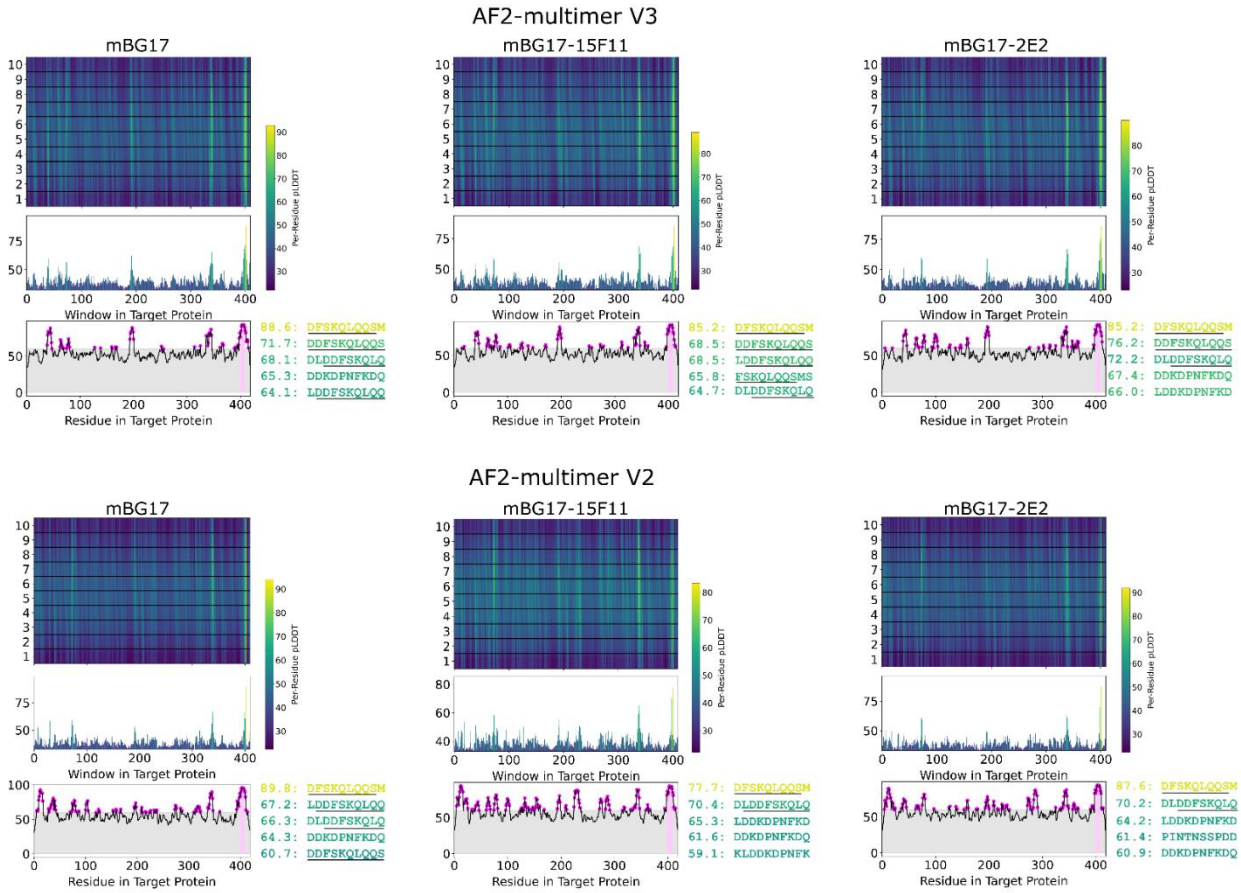


Figure 2.12 A comparison of Alphafold2 multimer version 3 and multimer version 2 applied to the mBG17 system. The experimental epitope, DDFSKQLQQS, is still easily identified with all three scFv backbones (wildtype, 15F11, and 2E2).

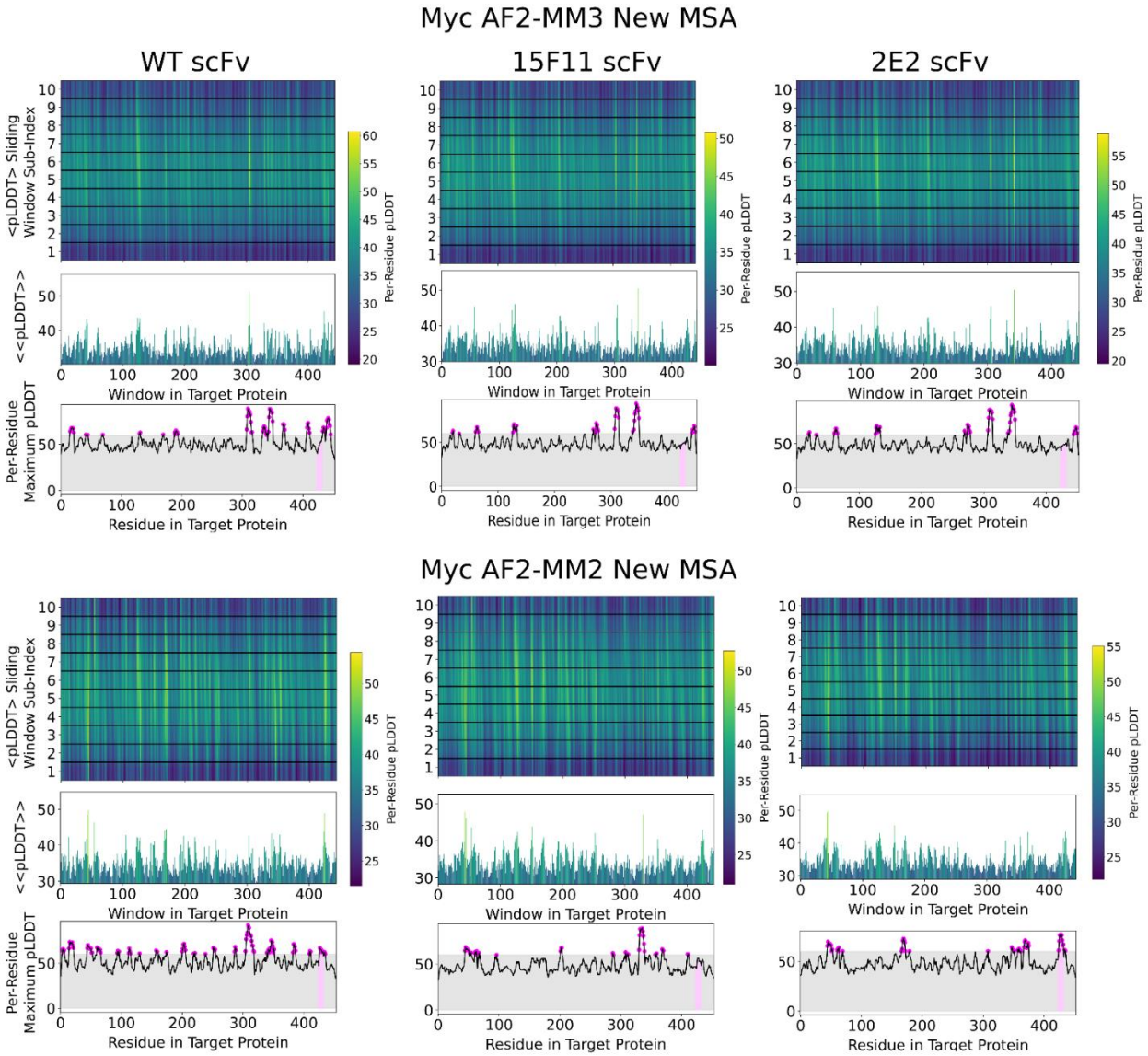


Figure 2.13 Myc comparison of epitope identification accuracy, comparing model types. Performance variation with AlphaFold2 model (multiple versions 2 and 3) and MSA versions (most up to date version of the ColabFold MSA server uses UniRef30 (2302) and PDB100 (220517)) vs the old MSA server (when this data was initially generated, ColabFold MSA server used UniRef30 (2202) and PDB70 (220313)). The left column is the WT scFv, the middle column is the CDR loops spliced onto the 15F11 backbone, and the right column is the CDR loops spliced onto the 2E2 backbone. Performance was ablated when using MM3 and the new MSA, and significantly degraded when using MM2 with the new MSA. For AF2-MM2 Old MSA, see Figure 2.

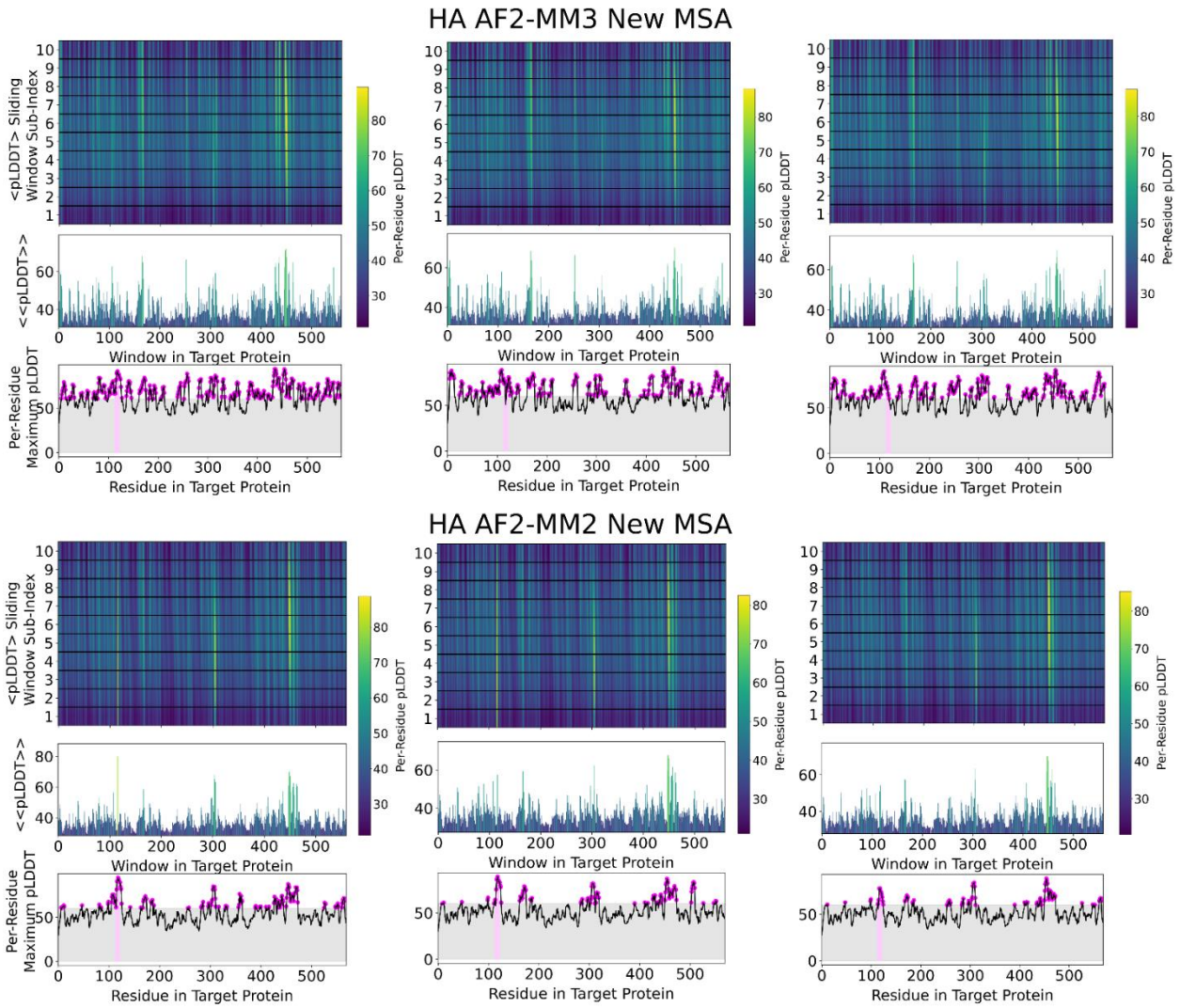


Figure 2.14 HA comparison of epitope identification accuracy, comparing model types. A comparison of the differing AlphaFold2 models with the Myc system (multimer version 3 and 2) along with a comparison of the new MSA (most up to date version of the ColabFold MSA server uses UniRef30 (2302) amd PDB100 (220517)) vs the old MSA server (when this data was initially generated, ColabFold MSA server used UniRef30 (2202) and PDB70 (220313)). The left column is the WT scFv, the middle column is the CDR loops spliced onto the 15F11 backbone, and the right column is the CDR loops spliced onto the 2E2 backbone. For AF2-MM2 Old MSA, see **Figure 2.9**.

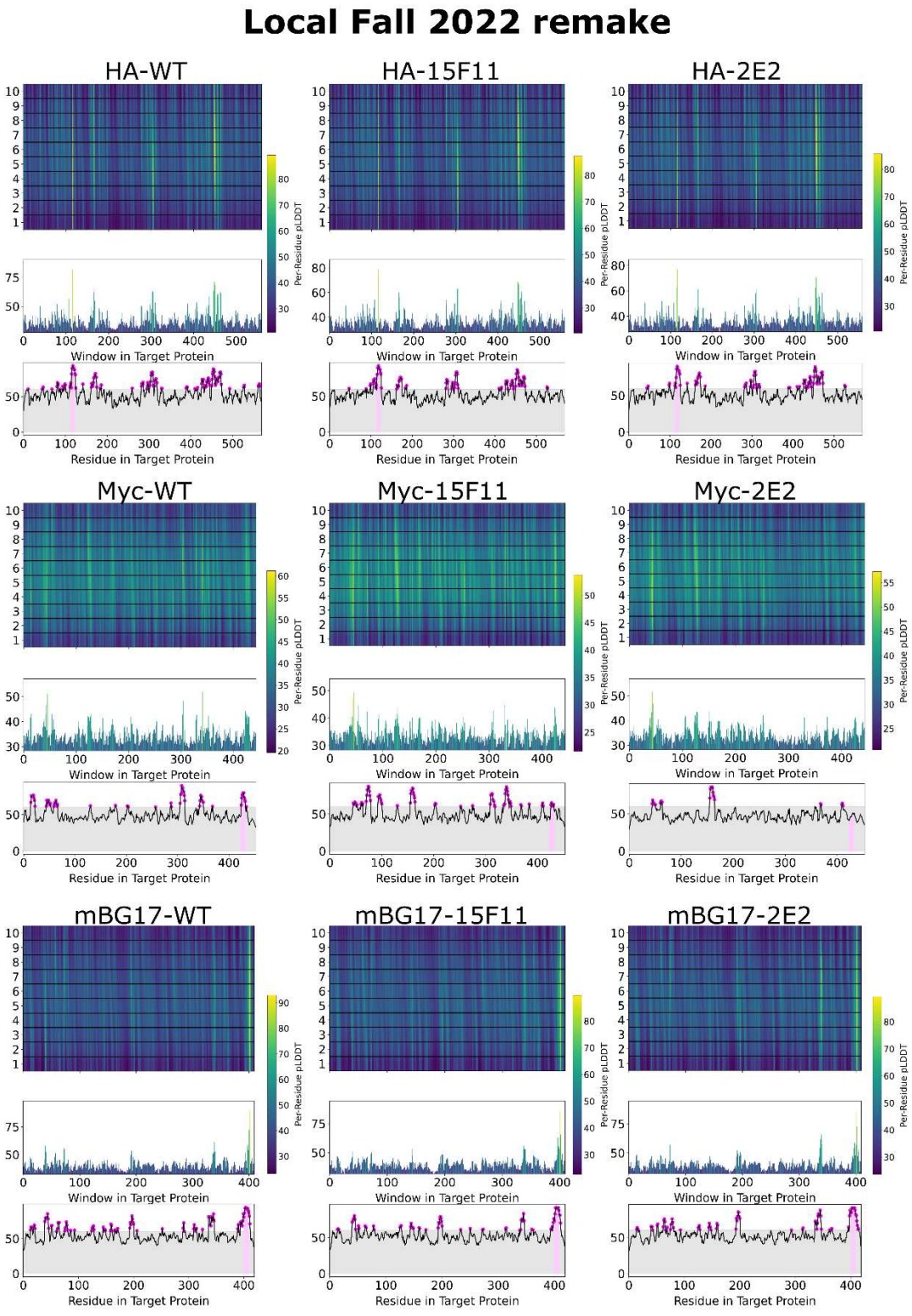


Figure 2.15 Local remake of the databases used by the MMSEQS server. Databases were downloaded (UniRef30 (2202) and PDB70 (220313)) and were queried locally to produced MSA's for testing. These runs all were done with the multimer version 2 model of Alphafold 2.

The left column is the WT scFv, the middle column is the CDR loops spliced onto the 15F11 backbone, and the right column is the CDR loops spliced onto the 2E2 backbone. The first row is the HA system, the second row is the Myc system, and the final row is the mBG17 system.

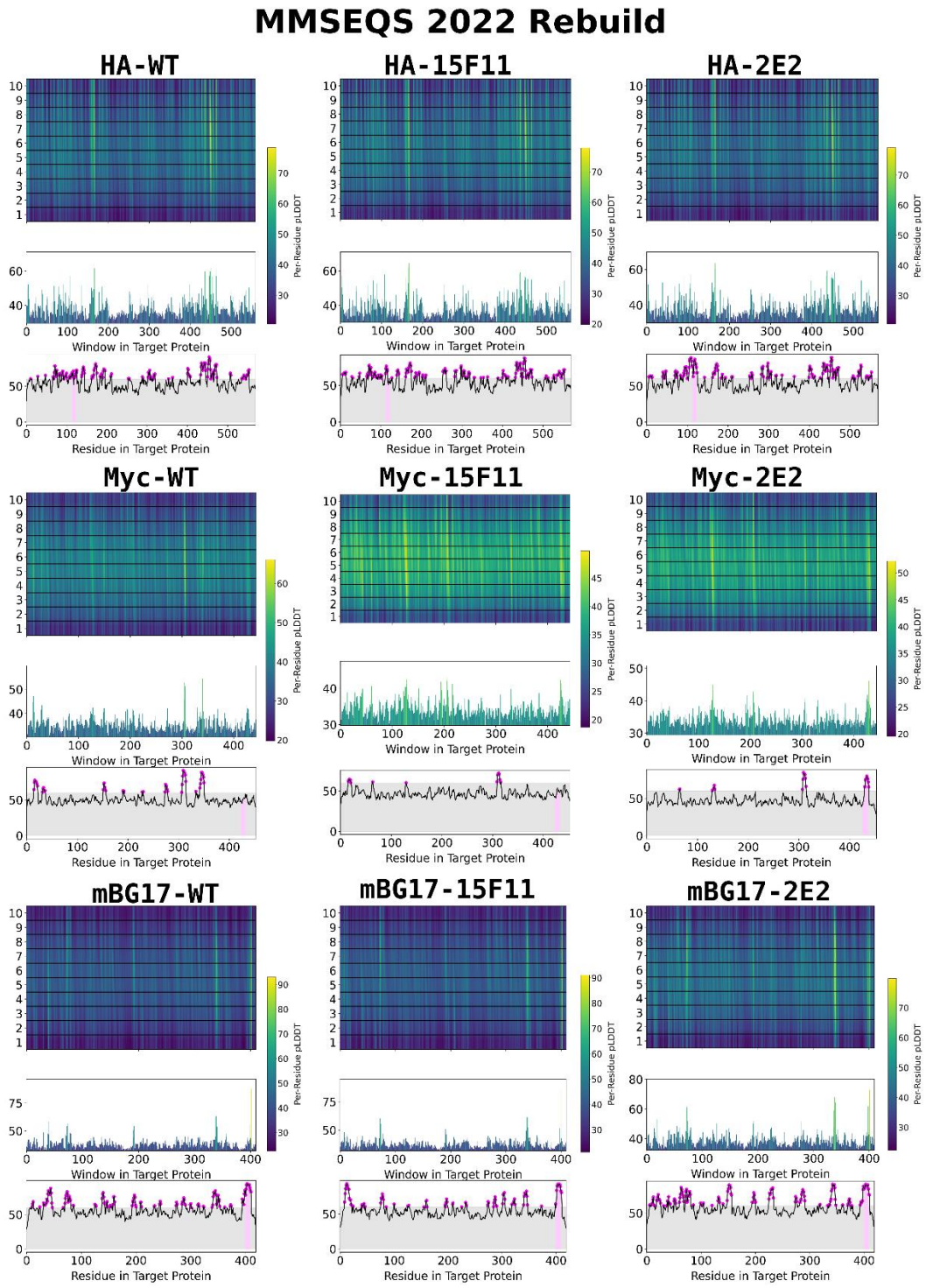


Figure 2.16 Server remake of the MMSEQS databases. The databases were rebuilt by the MMSEQS team UniRef30 (2202) and PDB70 (220313)) on the Colabfold MSA server and were queried produced MSA's for testing. These runs all were done with the multimer version 2 model of Alphafold 2. The left column is the WT scFv, the middle column is the CDR loops spliced

onto the 15F11 backbone, and the right column is the CDR loops spliced onto the 2E2 backbone. The first row is the HA system, the second row is the Myc system, and the final row is the mBG17 system.

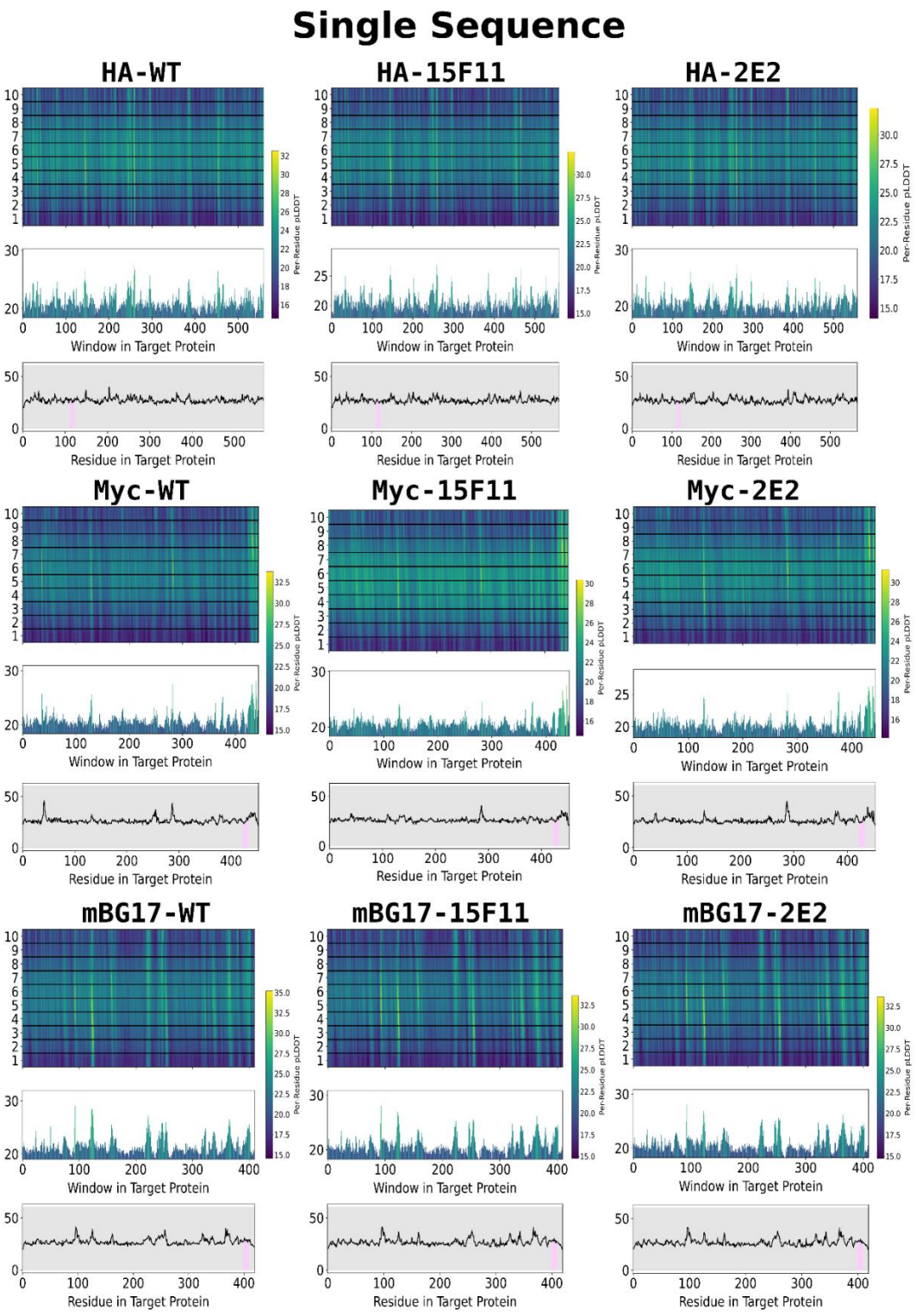


Figure 2.17 Single Sequence mode (no MSA's) of epitope prediction with AF2. These runs all were done with the multimer version 2 model of Alphafold 2 in single sequence mode (i.e. no MSA was used) as a negative control, to highlight the importance of a quality MSA. The left

column is the WT scFv, the middle column is the CDR loops spliced onto the 15F11 backbone, and the right column is the CDR loops spliced onto the 2E2 backbone. The first row is the HA system, the second row is the Myc system, and the final row is the mBG17 system.

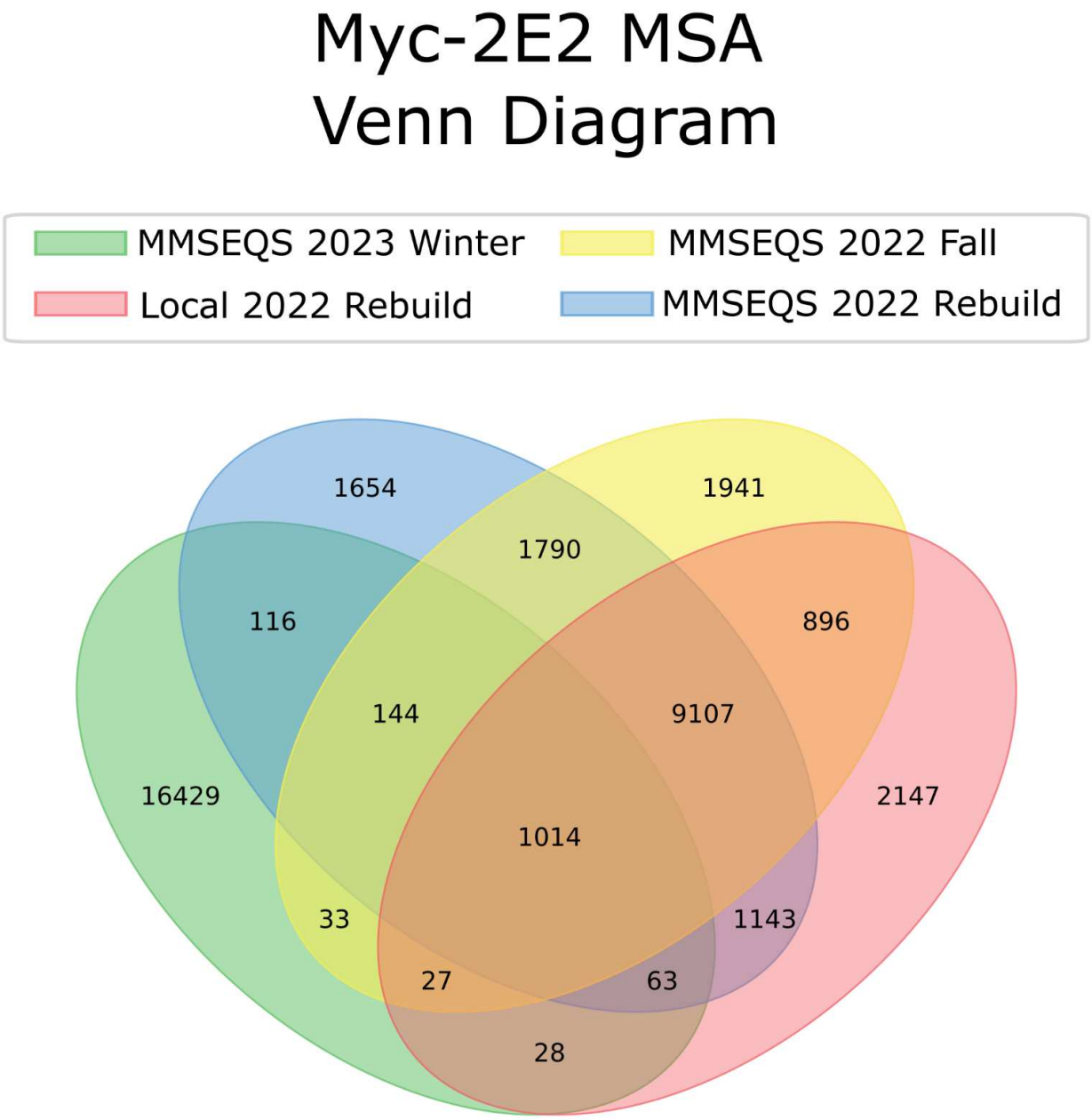


Figure 2.18 MSA overlap between the 4 generation methods. **.** Here we highlight the number of unique entries that are shared amongst all of the MSA methods, those being: **1)** using the databases right now via colabfold (PDB30 2302 and PDB100 230517) (green) **2)** the databases after they had been accessed via colabfold and cached for repeated use (UniRef30 (2202) and PDB70 (220313)) (yellow), **3)** downloading the databases locally (UniRef30 (2202) and PDB70 (220313)) and attempting to create the MSAs ourselves (red), and **4)** querying the databases after the MMSEQS team rebuilt them for our use via colabfold (UniRef30 (2202) and PDB70 (220313)) (blue).

		MMSEQS 2022 Fall	Local 2022 Rebuild	MMSEQS 2022 Rebuild	MMSEQS 2023 Winter	Single Sequence
HA	WT	M	✓	-	✓	-
	15F11	✓	✓	-	M	-
	2E2	M	✓	-	-	-
Myc	WT	M	✓	-	-	-
	15F11	✓	-	-	-	-
	2E2	✓	-	✓	✓	-
mBG17	WT	✓	✓	✓	✓	-
	15F11	✓	✓	✓	✓	-
	2E2	✓	✓	✓	✓	-

Figure 2.19 Comparison of how well each MSA generation scheme accurately identified the experimentally derived epitope within the top 5 epitope sequences. A green checkmark shows that it was found by both the consensus model and the top single model, a yellow "M" means the simple max method correctly identified the experimental epitope in the top 5 epitopes, and the red dash means both methods failed. The consensus model did not identify the epitope correctly when the simple max method failed to. The colored background behind the titles is the same color as **Figure 2.18** to help guide the eye.

Chapter 3 – Automation of Protein Crystallization Scaleup via Opentrons 2 Liquid Handling

3.1 Individual Contributions

One of the unusual research themes for the Snow research group is the engineering of biomolecular crystals for purposes beyond structural biology. Automated crystallization trials represent a way to increase the engineering rigor and throughput for the sensitive nucleation and growth processes underlying crystal growth. As such, crystallization automation had been on Dr. Chris Snow's wish list for a while, and I saw it as an opportunity to pursue something I'm passionate about (the increased access of science to all – in this case, by lowering the cost barrier associated with automation of protein crystallization) with my strengths (the merge of both computational and wet lab techniques). My major contributions to this project [2]were (1) developing the Python code necessary for the Opentrons 2 (OT2) to run, (2) extensive troubleshooting and bug fixing, and (3) OT2 crystallization trials.

3.2 Summary

This study presents an innovative approach to protein crystallization scaleup, utilizing the general purpose Opentrons-2 liquid handling robot. The research demonstrates the robot's capability to automate the creation of 24-well sitting drop protein crystal plates. Using Python

[2] The authors on this paper are Jacob DeRoo, Alec Jones, Tim Ahr, Sam Stroup, Grace Thompson, and Christopher Snow. This paper has been submitted to SLAS Technology to their special issue of Robotics in Laboratory Automation, and is still under peer review, submitted January 15th, 2024.

scripts for precise control, the study explores the robot's application in mixing and setting up crystallization plates with a model protein (hen egg white lysozyme) and a non-model system isoprenoid binding protein from *Campylobacter jejuni*. Successful crystallization indicates the approach can reduce manual labor in protein crystallization scale-up experiments, and may ultimately reduce variability, offering an economical and versatile tool for laboratories. The study showcases the potential of automated systems in enhancing the protein crystallization workflow, making it more accessible and consistent. All developed liquid handling routines and relevant data files, in addition to demonstration videos are available at https://github.com/jbderoo/Opentrons2-Protein-Crystallization

3.3 Introduction

Protein crystallization is an important protein engineering operation task, where the objective is to induce protein monomers to self-assemble into a precise nanostructure with coherent long-range order. These assemblies are crystals, and are in some cases highly porous scaffold materials[73–75]. The more precisely ordered the crystal, the higher resolution X-ray diffraction data we can collect about the protein's structure. Crystal structures often provide key information about protein-protein interactions or critical details to understand protein functions. Observing these features is often essential for rational drug design. Thus, accelerating protein crystallization trials remains highly beneficial to the structural biology community, complementing the new structures enabled by cryo-EM. Solving the structure of proteins is vital to validate computational protein structure prediction[76–78], drug design[79].

Protein crystallization is a relatively time-consuming experiment with no guarantee of success. Frequently, large experimental search grids are needed to find crystallization conditions

when attempting to crystallize a completely novel protein. A nonexhaustive list of parameters frequently varied during crystallization screening for a protein include salt concentration, salt type, pH, buffer type, buffer concentration, protein concentration, and the identity and concentration of precipitant. While rational assumptions can be made about two proteins' similarities (sequence, charge, and thus crystallization conditions) to reduce the vast search space, finding the ideal conditions to induce proteins to crystalize is still an iterative, trial and error process. One of the most commonly reported approaches to crystal screening is sitting drop vapor diffusion, wherein a volume of concentrated protein is mixed with a volume containing a precipitant. The combined volume is then sealed inside a well containing a separate volume of the undiluted buffer. Gradually, water vapor equilibrates between the sitting drop containing protein and the undiluted buffer, with a net loss of the water from the sitting drop. This results in a very gradual increase in protein and precipitant concentration, ultimately triggering crystal nucleation and metastable growth. Several liquid handling machines exist to help accelerate this brute force process such as the Crystal Gryphon from Art Robbins Instruments and the Mosquito ® from SPT Labtech. However, these machines are not perfect; as standalone instruments they carry maintenance costs and may not be sufficiently economical for all laboratories.

The Gryphon protein crystallization machine is recognized for its high-throughput capabilities, allowing for the screening of 96 conditions simultaneously[80–83]. It supports various crystallization plate sizes and offers a range of screening blocks, enhancing accessibility for high throughput screening. The Gryphon's design minimizes evaporation and human error, providing fast and consistent liquid drop distribution. This efficiency is particularly beneficial when setting up multiple plates. However, there are notable drawbacks. Per Art Robbins' catalog, the Gryphon's design was was finalized in 2016. Learning a scratch-like language is required to

appropriately build protocols and loop with the Gryphon, and stringent cleaning is required. Specifically, due to the reuse of needles, ensuring thorough cleanliness is difficult, yet critical to prevent the buildup of salts and PEGs. Another concern for machines that reuse needles is the accuracy of volume dispensation. Reagent buildup on the needle can affect accuracy and reproducibility, especially when working with microliter volumes.

Additionally, liquid handling robots can struggle with highly viscous protein solutions, as viscous protein drops tend to accumulate on the pipettes, hindering successful transfer into the wells ([81]). Although there are settings to mitigate this, the problem occasionally persists. Ultimately machines that are dedicated to cannot easily be repurposed for other tasks. While the Gryphon streamlines the process of setting up crystallization plates, a major part of the time savings comes from using pre-made reservoir solution screens. Alternative liquid handling instruments (e.g. Rigaku Alchemist) are needed to automate the creation of such 96-well plates. We set out to determine if a more general-purpose liquid handling system could scaleup crystallization trials (i.e. at the 24 well scale), including both the reservoir mixing and sitting drop deposition setup steps.

The Opentrons company has 3 models of liquid handling robots and many different tools and attachments to fill the diverse needs of scientists[84–87]. Appealingly, the Opentrons robots can be programmed with and execute Python code, providing fine control over all liquid handling steps over time. This flexibility could be very useful for crystallography, accommodating the diverse schemes used for protein crystallization (e.g. sitting drop, hanging drop, etc.). Here, we focused on sitting drop experiments with Hampton Research's CrysChem 24-well plate. The geometry of a sitting drop protein crystallization plate is moderately complex, with multiple liquid containing volumes per well location (**Figure 3.1**). These plates are larger than standard

SBS format and are optimized for the growth of protein crystals of sufficient size for routine XRD structure determination. The CrysChem plates have the inherent advantage of being suitable for *in situ* mixing of a reservoir solution and the crystal growth sitting drop in a single piece of labware, as well as easier visualization and manipulation of synthesized crystals. For comparison, other crystallization screening labware such as the SWISSCI 96-well plate is standard SBS format, but is more suited for high throughput discovery of crystallization conditions (as opposed to crystal scaleup) due to its smaller reservoir size. Fortunately, the Opentrons pipettes can be moved more precisely than a simple "Go to A1", allowing pipetting to and from both the crystal growth podium and the reservoir "moat" for each sitting drop position. Notably, a team at Opentrons demonstrated the OT2's ability to create reservoir mixtures via a deep 96 well plate and then fill and plate an MRC Maxi 48-well from SWISSCI. This demonstration is freely available from Opentrons at https://vimeo.com/654672188 .

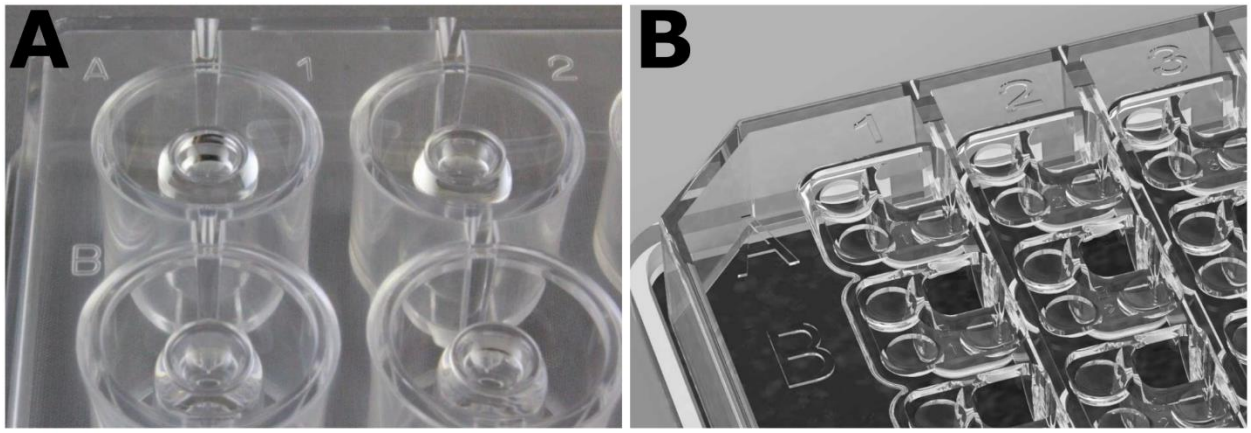


Figure 3.1 Well geometries for two common protein crystallization plates. **A)** A Hampton Research Cryschem 24-well plate for sitting drop crystallization. The outer ring "moat" is where that reservoir is prepared for that well. The center raised pedestal is where the reservoir is mixed with the purified protein solution to create sitting drops. **B)** A SWISSCI 96-well crystallization plate for sitting drop. These volumes are much smaller and are typically used for broad crystallization condition screening. Its geometry includes 3 small wells with one larger central reservoir.

3.4 Materials and Methods

3.4.1 Opentrons-2 liquid handling robot

Throughout these experiments we used the same Opentrons-2 liquid handling robot (OT2). The machine was configured with a p10 1st generation pipette arm attached to the left arm, and a p300 2nd generation pipette arm attached to the right arm. More information about the liquid handling robot can be found here: https://shop.opentrons.com/ot-2-robot/ . 10 µL tips (GEB PT0010-9B-NS) and 200 µL tips (Opentrons 999-00081) were used for each arm, respectively.

3.4.2 Custom labware for the OT2

Typical SBS-format laboratory 96 well plates measure approximately 84.5 x 127.8 mm in x and y dimensions. The deck of the OT2 is preallocated with slots that can lock plates of this size into place so that they can't move during the experiment. For our tests, we used the HR3-159 Hampton Research CrysChem 24-well plates that are larger than a standard SBS-format plate (150 x 106 mm). To overcome this barrier, an adapter was designed by the Opentrons engineering team to clip into two of the slots on the OT2's deck. The adapter narrows at the top to seat the CrysChem plates and keep them stationary for the duration of the experiment. The adapter was 3D printed on a Bambu Lab X1 Carbon printer using PLA. The entire print job took 1 hour and 24 minutes to complete. **Figure 3.2** shows the engineering sketch **(A)**, a 3D rendering **(B)**, and the final product **(C)** of this sitting drop plate OT2 adapter. This .stl file can be found at the GitHub for this project (https://github.com/jbderoo/Opentrons2-Protein-Crystallization).

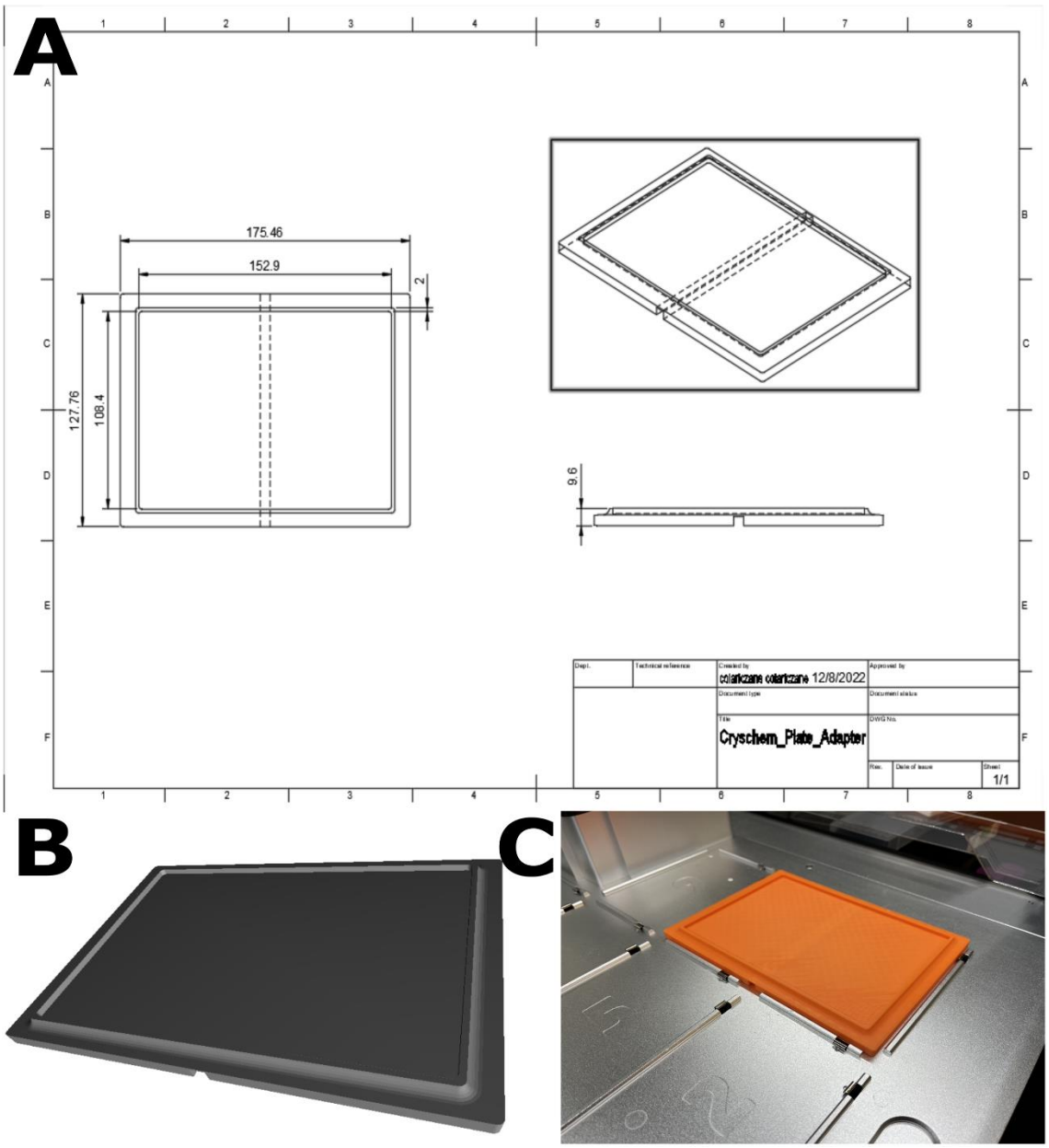


Figure 3.2 Various representations of the adapter used to seat the larger than standard SBS-format CrysChem plates. **A)** The engineering schematic for the CrysChem plate as provided by Hampton Research. **B)** A 3D rendering of the adapter designed by the Opentrons engineering team. **C)** The final product printed and clipped into the Opentrons deck.

In addition to a physical adapter, the OT2 needs a data file that tells the robot where the pipette needs to go in 3D space when location labels are used; for example, when told to go to

“A1”, go to 30.5 mm in the x direction, 12.5 mm in the y direction, and 6.1 mm in the z direction. This .json file is also available on the GitHub, and must be uploaded as a custom labware definition to the OT2. Members of the Opentrons support team were incredibly helpful in the troubleshooting and development of both files. The custom labware definition was further refined by Jacob DeRoo.

3.4.3 Python scripts

Opentrons liquid handling robots are programmable via Python, allowing both fine control of the robot and accessible customization to new projects via easy adaptation of starter scripts. All scripts written for this project were written in Python 3.9 with the Opentrons Python module. No other non-standard modules were used. For more information about the package and installation, see https://docs.opentrons.com/v2/ and https://anaconda.org/conda-forge/opentrons respectively.

Three example scripts (discussed individually in subsequent sections) are each broken down into a few major sections, and are available on the GitHub. (1) The first section of each script describes the geometry of the source tubes to prevent the OT2 from completely submerging the pipette into the source liquid. (2) The second section of each script describes the volume of every source liquid that must be transferred into every well on the CrysChem plate. (3+) The remaining sections describe how to transfer liquid to the intended destinations in sitting drop plate.

Script 1, mixing demonstration with food dye: To mimic a conventional sitting drop 24-well experiment, we programmed the OT2 to mix a variety of candidate buffers (blue and red

food dye in water) in the reservoir sections of the plate. An increasing amount of blue dye was added every column (0-150 µL from a pure blue stock solution, stepping by 30) and an increasing amount of red dye was added every row (0-150 µL from a pure red stock solution, stepping by 50). The remainder of each reservoir was filled with pure water, raising the total reservoir volume to 400 µL. Upon completion, 5 µL of yellow dye stock solution was aspirated, then immediately after 5 µL of the reservoir was aspirated in the same pipette tip. This 10 µL volume was then dispensed into the corresponding pedestal.

Script 2, Hen Egg White Lysozyme (HEWL) crystallization: HEWL crystals were synthesized using several different stock solutions: (1) Hampton Research's premixed "15-Minute Lysozyme Crystallization Reagent" (HR2-805), (2) Hampton Research's purified lysozyme protein (HR7-110) at 20 mg/mL in a 20 mM Sodium Acetate buffer at a pH of 4.6, and (3) a 200 mM stock solution of 200 mM sodium acetate of varying pH; either 4.6, 4.7, or 4.8. A 24 well plate was set up where each crystallization trial used 3 stock solutions. To prepare the reservoirs, 300-350 µL of solution (1) was added to every column (increasing by 10 µL every column). 50 µL of solution (3) was added to every well; the buffer at pH 4.6 was added to row A, the buffer at pH 4.7 was added to row B, and the buffer at pH 4.8 was added to rows C and D. To bring every reservoir up to 400 µL, DI water was added as necessary (**Figure 3.3**). After all the reservoirs were mixed, 2 µL of solution (2) was aspirated into the pipette. Immediately after, 2 µL of the reservoir was aspirated. The 4 µL sitting drop volume was then dispensed into the elevated crystal formation pedestal. We then manually sealed the plate with Duck HP260 clear packaging tape after all 24 sitting drops had been prepared.

		Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	
Row A	pH 4.6	50 µL buffer 300 µL HR2-805 50 µL water						
Row B	pH 4.7							
Row C	pH 4.8							
Row D	pH 4.8						50 µL buffer 350 µL HR2-805 0 µL water	
		0.75	0.775	0.8	0.825	0.85	0.875	**precip (fraction of HR2-805)**
		300	310	320	330	340	350	**HR2-805 (µL)**
		50	40	30	20	10	0	**water (µL)**

Figure 3.3 A matrix representation of the CrysChem 24-well plate sitting drop used to synthesize HEWL crystals.

Script 3, CJ crystallization: The CJ protein is an engineered variant of CJ0420, a putative periplasmic protein from *Campylobacter jejuni*. CJ crystals were grown according to previously optimized conditions[88]. The crystal structure known (PDB entry: 5W17). A 24-well plate was set up using three stock solutions: (1) 1M Bis-Tris at pH 6.0 or 6.5, (2) 4M ammonium sulfate, and (3) CJ protein at 12 mg/mL in 50 mM HEPES, 500 mM ammonium sulfate, 10% glycerol at pH 7.4. Stock solution (3) was aliquoted in bulk into 100 µL batches in 1.5 mL low adhesion microcentrifugation tubes. 40 µL of solution (1) was added to every well; rows A and C were pH 6.0, and rows B and D were 6.5. Between 335 µL and 360 µL of solution (2) (stepping 5 µL per column) was added to every well (**Figure 3.4**). The remaining volume added was water, to bring the total each reservoir volume to 400 µL. After all the reservoirs were mixed, 2 µL of stock (3) was aspirated into the pipette. Immediately after, 2 µL of the reservoir was aspirated. The 4 µL sitting drop volume was then dispensed into the elevated crystal formation pedestal. We then

manually sealed the plate with Duck HP260 clear packaging tape after all 24 sitting drops had been prepared.

		Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	
Row A	pH 6.0	40 µL buffer 335 µL AmmSulf 25 µL water						
Row B	pH 6.5							
Row C	pH 6.0							
Row D	pH 6.5						40 µL buffer 360 µL AmmSulf 0 µL water	
		3.35	3.4	3.45	3.5	3.55	3.6	**Ammonium Sulfate (M)**
		335	340	345	350	355	360	**Ammonium Sulfate (µL)**
		25	20	15	10	5	0	**water (µL)**

Figure 3.4 A matrix representation of the CrysChem 24-well plate sitting drop used to synthesize CJ crystals.

3.5 Results and Discussion

Script 1: As an initial proof of concept, we used food dye (red, blue, yellow) and colorless water to visualize the process of setting up a crystal plate. This strategy is highly recommended per Opentrons when developing liquid handling scripts to trouble shoot any errors that may arise and to ensure reagents are going to the intended locations. A gradient of red food dye was increased as we moved down the plate (more red dye in row D, less red dye in row A). A gradient of blue dye was increased as we moved across the plate (more blue dye in column 6, less blue dye in column 1). The process was repeated by manual pipetting, and no differences were discernable (**Figure 3.5**). The wells were correctly constructed with red, blue, and clear water. Importantly the sitting drops were prepared correctly by combining 5 µL yellow dye

(standing in for the biomolecules) with 5 µL from reservoir. From this process, we learned that an additional step was needed for the OT2 to mimic manual pipetting. Specifically, at the 400 µL volume, the reservoir solutions can have enough surface tension to avoid uniformly dispersing throughout the moat volume when all reservoir components were dispensed at the top of the reservoir (i.e. at the 12:00 position on a clock). This effect was stochastic, affecting approximately 20 of the reservoirs. In so far as this effect represents a potential source of variability for vapor-liquid equilibrium, we proceeded to develop a flattening technique. To this end, another step was added that withdrew 200 µL from the 12:00 position of the reservoir, then moved to the 3:00 position and dispensed 22 µL, then moved to the 4:00 position and dispensed 22 µL, and was repeated until the distributed dispensing steps no longer held any volume. In most cases this extra step successfully overcame the surface tension causing the reservoir liquid to be uniformly distributed around the central sitting drop pedestal. The time required for the OT2 to construct this plate was 38 minutes.

For clarity, we have also included a short YouTube video of our OT2 preparing the major diagonal of the color plate for visualization, available at our GitHub.

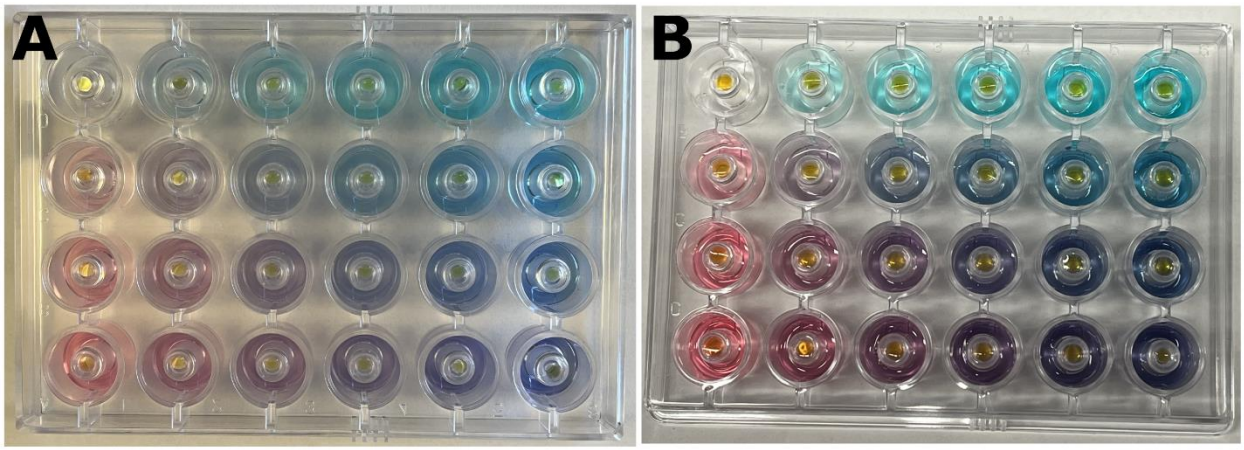


Figure 3.5 Proof of concept sitting drop plates with food coloring. The OT2 prepared a sitting drop plate with dyes to visualize correct mixing. A blue gradient increases from left to right, and

a red gradient increases from top to bottom. Each reservoir mixture was then mixed with a yellow dye stock to mimic combining each reservoir solution with purified protein in a sitting drop. After preparation, plates were taped manually. **A)** is the dye plate prepared by the OT2 whereas **B)** is a plate prepared via manual pipetting.

Script 2: With a proof of concept established, we moved onto hen egg white lysozyme (HEWL), a protein that readily crystallizes under suitable conditions[89–91]. Crystals had formed in a few wells within 15-30 minutes, but after 24 hours were readily visible in numerous wells (**Figure 3.6**). Similar pipetting techniques to the food dye were employed in the HEWL script, with two notable differences. First, the picking up and redispensing of 200 µL at 3:00-9:00 was not needed due to a difference in viscosity of the crystallization media. Second, a "tip shaking" step was added to ensure the purified protein:well reservoir mixture was completely removed from the pipette tip. In our group, a 2 µL sitting drop volume is most common when growing protein crystals. We tested the OT2 script at this volume but found the OT2 with the Gen1 p10 pipette could not consistently eject the 2 µL into the sitting drop pedestal. Frequently, the protein:reservoir mixture was too "sticky" to leave the tip and arrive at the bottom of the podium, and instead would curl up the side of the pipette tip, even with the pipette tip shake step. Doubling the sitting drop volume from 2 µL to 4 µL rectified this issue. The time required for the OT2 to construct this plate was 32 minutes. It took approximately 12 minutes to create the sitting drops.

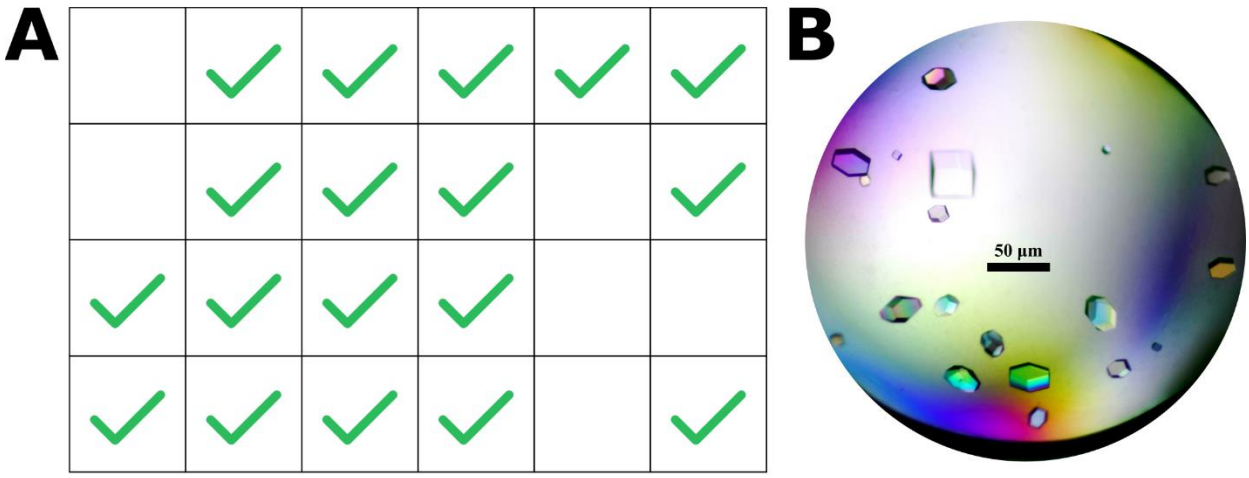


Figure 3.6 Crystallization results from the OT2 HEWL 24-well plate sitting drop. A) A cartoon schematic of the CrysChem plate, where green check marks highlight successfully grown HEWL crystals. 18/24 wells successfully yielded HEWL crystals. B) One of the wells (D4) that successfully yielded HEWL crystals.

Script 3: With a solid grasp of the OT2's capabilities and a solid script foundation, we next sought to demonstrate crystallization of a non-model system we have extensively crystallized in the past[88]. CJ crystals have unusually large solvent channels, enabling intra-crystal macromolecule transport, leading to diverse applications such as textile conjugation ([92]), enzyme immobilization[73], and mosquito tracking[93]. A 24-well plate was prepared, with two buffers at pH 6.0 and 6.5 and with an ammonium sulfate salt concentration ranging from 3.35M – 3.60M. CJ crystals typically grow in 24 – 72 hours, but can take as long as several weeks to form. Within 24 hours, CJ microcrystals had formed with the expected, signature hexagonal habit in three wells (**Figure 7**). The reservoir flattening technique was employed. Tip shaking was not needed in this case. A total sitting drop volume of 4 µL was used, similar to the HEWL plate. The time required for the OT2 to construct this plate was 40 minutes. It took approximately 12 minutes to create the sitting drops.

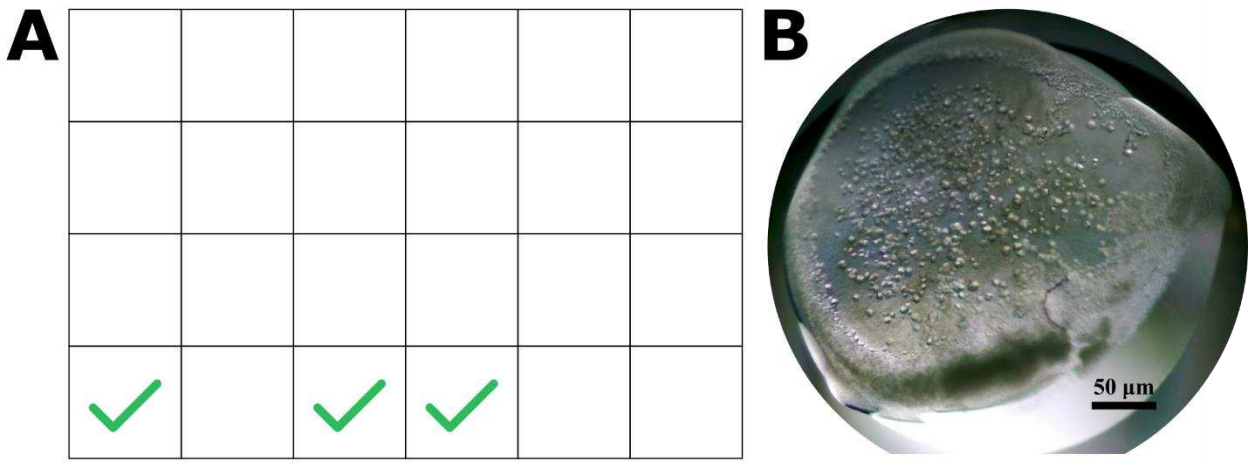


Figure 3.7 Crystallization results from the OT2 making a CJ 24-well plate sitting drop. A) A cartoon schematic of the CrysChem plate, where green check marks highlight successfully grown CJ crystals. 3/24 wells yielded CJ crystals. B) One of the wells (D3) that yielded CJ microcrystals.

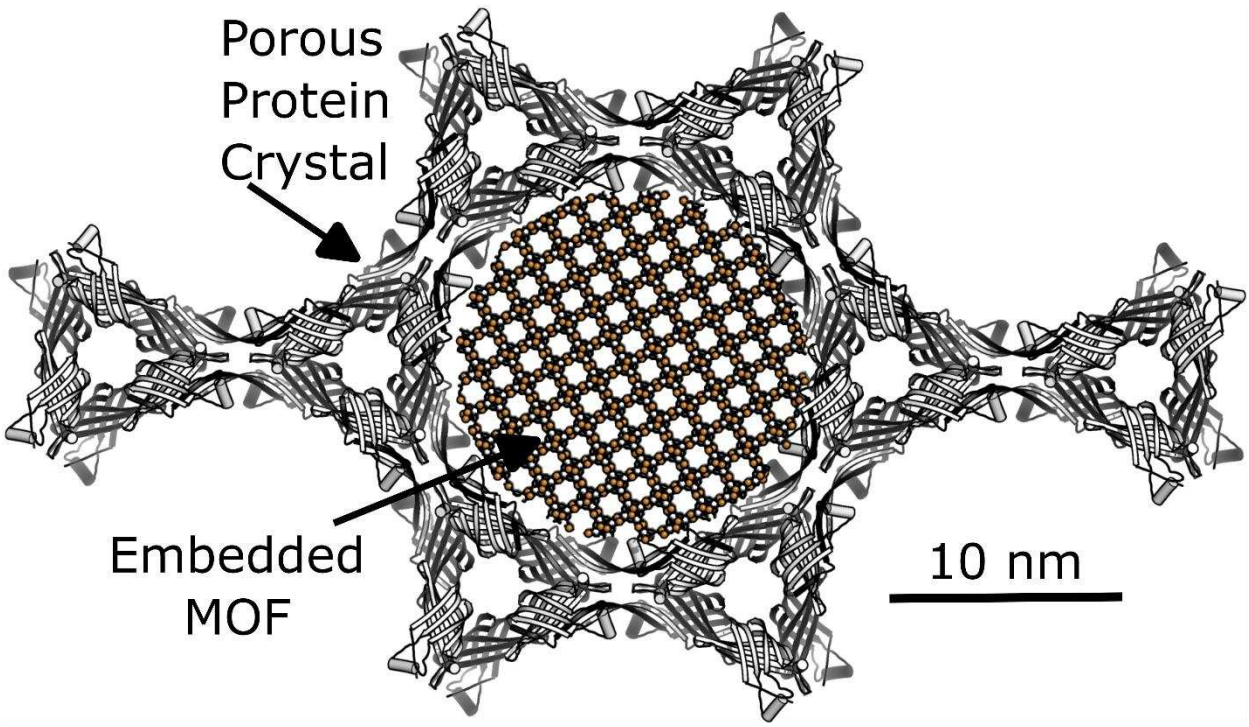
3.6 Conclusion

In this study, we presented an economical and easy to use liquid handling protocol that enables fine control over pipetting via Python scripts to construct mock and authentic 24-well sitting drop protein crystal plates with colored water, hen egg white lysozyme (HEWL), and CJ protein. While 35 – 40 minutes is not a fast setup time compared to a trained scientist or technician, it does eliminate the need for a trained technician to create these plates. Perhaps more importantly, we suspect robotic plate setup will greatly reduce plate variability from person to person. A lab scientist is only needed for OT2 setup, plate sealing with tape, and OT2 tear down. A single OT2 costs $13,500 at the time of this writing, significantly cheaper than a Gryphon machine priced at $55,000-$75,000. With the adapter, custom module, and starter scripts presented in this paper, the protein structure community at large could affordably test scale-up sitting drop plates at the 4 µL sitting drop scale. Notably, the OT2 could also prove useful for

other screen preparation steps in service of biomolecular crystallography. Additionally, the new Opentrons Flex instrument reportedly has precision down to 1 µL +/- 0.13 µL, which may overcome the 4 µL sitting drop issue encountered here. In principle, a humidity controlling device (e.g. the MiTeGen Watershed ™) could be used to humidify the interior of the OT2, limiting variability due to sitting drop evaporation. Alternatively, the script could easily be modified to pause after setting up a row of sitting drops, allowing for the attending scientist to tape over the completed wells.

Research reported in this publication was supported by NIAID of the National Institutes of Health under award number 1R01AI168459-O1A1.

Chapter 4 – Metal Organic Frameworks (MOFs) with a Porous Protein Crystal Superstructure



Graphical Abstract/Splash Art

4.1 Individual Contributions

The original motivation behind this paper [3]was done completely on a whim, as a side project. At the time I had been working on the exact mirror opposite to this project, wherein I was trying to encapsulate MOF around a peptide. I decided to take some CuBTC and grow it in

[3] The authors on this paper are Jacob DeRoo, Rojina Shrestha, Christopher Snow, and Melissa Reynolds. This paper has been submitted to will be submitted within the next few weeks to a journal that specializes in crystals, either "RSC crystal engineering communications" or "MDPI crystals".

solution with a CJ crystal present – low and behold after 24 hours, the CJ crystals was bright blue. Excited about my results I went (literally) running to Dr. Snow's office to showcase what had happened, and ask every scientist's favorite question: now what?

4.2 Introduction

The last chapter included the growth of crystals from a putative isoprenoid binding protein from C. jejuni. Since 2016[73,88,94], a modified "CJ" protein has been often crystallized in the Snow group to explore applications enabled by the unusually large solvent pores in the resulting crystals. After chemical crosslinking these crystals can tolerate a wide range of solvent conditions and can uptake diverse guests. In this chapter, I took the engineering of host-guest CJ crystals to a new height by creating a highly novel hybrid MOF-protein material: the MOF@CJ crystal. This project represents my attempts to combine two different labs doing similar sciences with different materials; one a protein crystal, and one a hybrid organic-inorganic material.

Chemical warfare agents, more broadly chemical threats (CTs), are chemicals that wreak havoc on the human body and can be fatal at a high enough dose or exposure time. In particular, G-class and V-class nerve agents are especially dangerous due to their lack of color and high volatility, meaning that victims are not aware of exposure until symptoms start developing[95–97]. Nerve agents interfere with the normal functioning of the nervous system, and some symptoms include nausea, vomiting, diarrhea, abdominal pain, cramping, bronchoconstriction, shortness of breath, unconsciousness, convulsions, and coma induction. Some well-known examples of these nerve agents are Soman (GD), Sarin (GB), Tabun (GA), and VX (**Figure 4.1**). Exposure to these chemicals can result in skin blistering, eye and respiratory tract irritation, and asphyxiation[98].

The therapeutic window ranges from minutes to hours depending on exposure time and species. Interestingly, from an inactivation strategy, all these chemical compounds have an organophosphate bond, and degradation can be achieved via hydrolysis or oxidation. Of these, hydrolysis is preferred because the reaction produces safer byproducts[99].

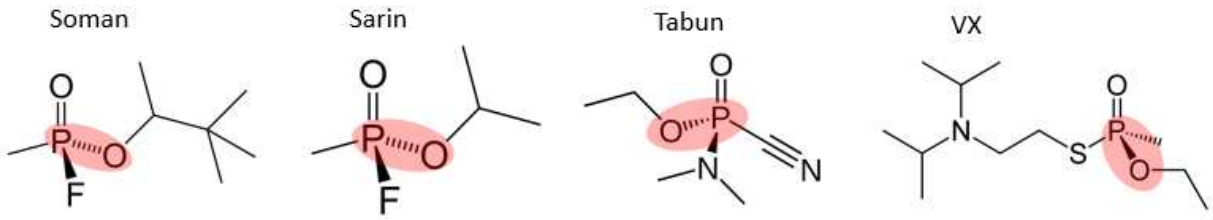


Figure 4.1 Molecular structures of various G-class and V-class nerve agents, and the organophosphate bond (red) that can be hydrolyzed to form less harmful byproducts.

Within the body of research into materials that can eliminate these threats, we will highlight three families of materials that can hydrolyze the nerve agents: organophosphate bond cleaving enzymes such as phosphotriesterase[100], sorbent materials like metal oxides[101] or zeolites[102], and metal organic frameworks[16] (MOFs). In the event of an emergency, rapid response with enzymes can be quite challenging with shelf life and biocompatibility. Most sorbent materials and some MOFs suffer from a similar problem: poor structure stability and flexibility, low sorption capacity, few active sites, and deactivation of the active sites[18]. Even in highly porous MOFs, the majority of catalysis may occur at the surface of the crystal[103]. There is a clear need for an engineered class of material to offer long term stability, easy administration (proactive or reactive), high catalytic capability, and readily accessible surface area.

We therefore sought to create a novel hybrid MOF material with the potential for improved operational parameters relative to pure MOF. Having recently explored applications for highly porous protein crystals, we sought to determine if it was possible to create a hybrid material where MOF domains, (i.e. nanocrystals or nanorods) could be deposited within the 13-nm diameter nanopores delimited by "CJ crystals"[88]. CJ crystals are composed entirely of protein building blocks, specifically "CJ," an engineered protein variant of a putative isoprenoid binding protein (GenBank cj0420) from *Campylobacter jejuni*. After crystal growth, the porous lattice can be stabilized via chemical crosslinking, commonly with dialdehydes (particularly glyoxal or glutaraldehyde) or carbodiimides. In principle, combining MOF catalytic proficiencies with the environmental stability and biocompatibility of another material could open the door to large-scale deployment via textile weaving or bioconjugation[92], or even emergency *in vivo* administration[85,93,104]. MOF encapsulation within protein crystals may provide a route to reduce immune response or other biocompatibility issue, but could perhaps still preserve or amplify the catalytic activity of the guest MOF domains. While we hope the application of these materials to nerve agent hydrolysis will be realized in the future, the focus of the present contribution is instead to demonstrate the feasibility of the synthesis of a new class of material, specifically MOFs embedded into highly porous protein crystals (PPCs).

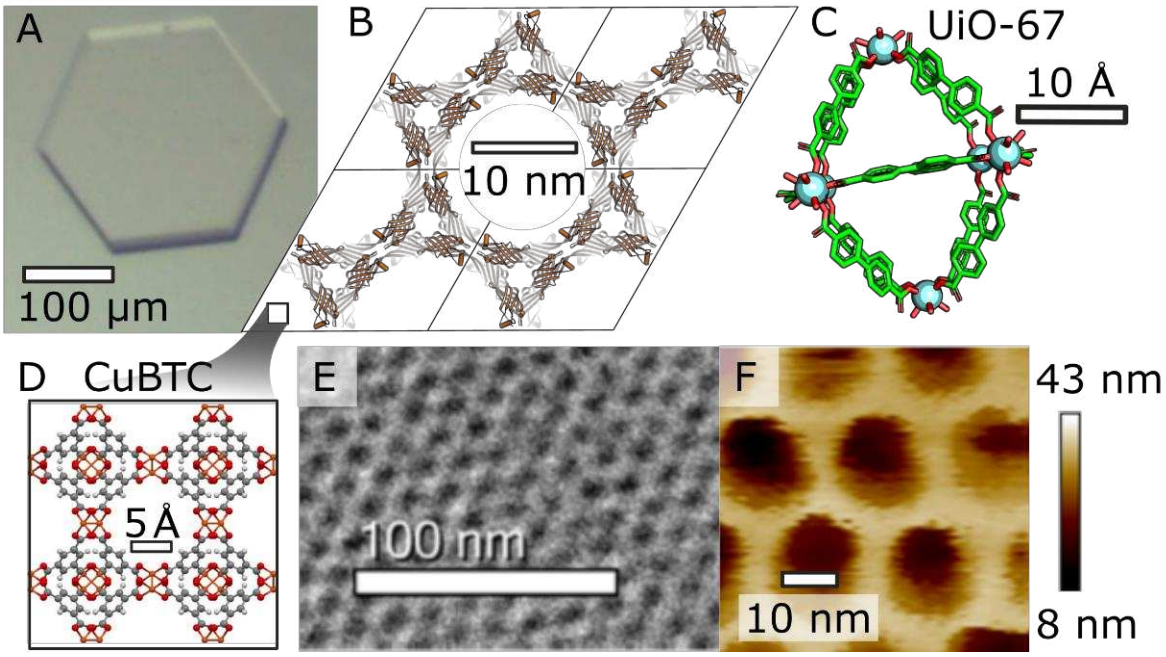


Figure 4.2 Structures of the materials of interest. **A)** CJ protein crystals tend to adopt a hexagonal prism habit. **B)** The CJ crystal nanostructure is in the P622 space group and features a 13 nm diameter pore. **C**) A single unit cell of the metal organic framework UiO-67 with space group Fm-3m. The primary pore of UiO-67 runs through the unit cell, perpendicular to the page. **D)** 4 unit cells of CuBTC, arranged in a 2x2 pattern, with space group Fm-3m. The pore of CuBTC runs through the union of 4 CuBTC unit cells. **E)** TEM of a stained microtome slice of a CJ crystal interior, highlighting the periodicity of the pores. **F)** AFM of the CJ protein crystal's surface, highlighting the porosity of the material[74].

Here, we introduce a hybrid material MOF@PPC; a metal-organic framework grown in and around a porous protein crystal. While most protein crystals have significant solvent content and retain solvent channels sufficient for the internal transport of small molecules, here the term "porous" refers to crystals with pores large enough to host many MOF unit cells (e.g. > 10 nm diameter[105]). Porous protein crystals (**Figure 4.2A**) are a highly ordered arrangement of protein monomers that self-assemble into porous, low-density materials. One such case of a PPC is the CJ protein crystal. This protein readily forms crystals of varying, controllable size (commonly between ~300 nm and 0.6 mm in diameter, depending on the precipitant and protein

concentration), and possesses 13 nm diameter pores perpendicular to their hexagonal faces (**Figure 4.2 B,E,F**). After formation CJ crystals can be crosslinked, becoming extremely tough and capable of withstanding diverse environments and stresses, such as in pure water, high salt conditions, the presence of cells[104], and ingestion by mosquitoes[85]. CJ crystals have also been explored for guest molecule installation and delivery[85]. CJ monomers first form a domain-swapped dimer and then proceed to assemble into a P622 lattice reminiscent of a honeycomb. The protein components delimit a tightly spaced hexagonal array of nanopores, one 13 nm pore repeating every 18 nm[74] (**Figure 4.2 E,F**). We hypothesize that nanoscale MOF domains deposited inside CJ crystals may feature a large accessible surface area, and thus improved transport to MOF active sites and improved catalytic performance over freestanding monolithic MOF microcrystals[17,106]. *A priori*, it was unclear if the solution conditions needed for MOF growth (typically harsh organic solvents and high temperatures) would be compatible with host protein crystals. Because of the remarkable stability of crosslinked CJ crystals, we proceeded to test their capability to withstand the harsh organic solvents required to grow both CuBTC and UiO-67. We hypothesized that the CJ crystal's high porosity and large solvent channels would permit the nucleation and growth of MOF nanocrystals that span multiple unit cells.

The MOF CuBTC (aka HKUST-1, MOF-199) is a well characterized metal organic framework for which there are robust synthetic procedures known[19,20]. CuBTC is composed of copper centers and benzene-1,3,5-tricarboxylic acid linker molecules (aka trimesic acid) and readily grows on many surfaces[22], making it a top candidate for proof of concept studies. CuBTC has been applied in carbon dioxide gas storage[23], catalytic breakdown of 2-CEES/HD[107], and more. CuBTC is a MOF whose unique building unit is composed of two copper atoms, held together by four BTC molecules in a pinwheel shape. The CuBTC unit cell has a side length of

approximately 2.6 nm, therefore it would be theoretically possible to fit 5 copies of the unit cell in the pore of the CJ crystal linearly across the diameter (**Figure 4.2 D**). The catalytic capabilities of CuBTC to degrade G-class nerve agents are less favorable, especially when compared to UiO-67[16]. UiO-67 is a newer MOF than CuBTC, and has found applications in gas storage[108], chemical warfare agent detoxification[25], bioimaging[109,110], and more. Its structure is composed of ZrO clusters and biphenyl-4,4'-dicarboxylic acid linkers.

To the best of our knowledge, this is the first report of the creation of a MOF@PPC semi-biological material. While the catalytic deactivation of CTs is a long-term application, this paper reports only the preparation, characterization and stability of the CuBTC@CJ and UiO-67@CJ.

4.3 Results

4.3.1 CuBTC grows mostly in the pores of the MOF

After exposure to CuBTC growth conditions, host CJ crystals underwent a color transformation from a pale yellow to brilliant blue (**Figure 4.3 A,B**). As assessed in physical manipulation of such crystals under a stereozoom microscope, the blue color appeared to be uniform throughout the crystal interior. We also observed green crystals after CuBTC@CJ formation (**Figure 4.4).** To ensure that the CuBTC MOF had indeed grown in and around the CuBTC@CJ crystals, scXRD and SEM data were collected. SEM images of the surface show sparse, individual CuBTC crystals deposited into surface imperfections of the protein crystal (**Figure 4.5 A,C**). scXRD data show full and intact Debeye-Scherrer rings (**Figure 4.6**) that can be converted to a pXRD "fingerprint" (**Figure 4.7 A,C**). This implies that the X-ray beam (~8 μm diameter for ALS beamline 4.2.2), while passing through the host CJ crystal, was also passing through MOF domains in all orientations. Notably, the XRD intensity did vary with

goniometer angle (angle between crystal face and x-ray source), consistent with the variable pathlength of the host crystal (**Figure 4.8**). This scXRD data is consistent with a significant portion of the CuBTC having grown within the pores of the CJ crystal. In contrast, a large single CuBTC crystal would generate a single set of Bragg reflections. For comparison, we pressed a pure CuBTC powder into a 10 µm crystal loop (**Figure 4.9 A**), where individual CuBTC crystals clung to the loop. We obtained a diffraction pattern that is intermediate between the single lattice Bragg peaks and the powder diffraction Debeye-Scherrer rings (**Figure 4.9 B**). Notably, the presence of CuBTC nanocrystals within the interior of the CJ crystal does not preclude the formation of CuBTC nano- and microcrystals adhered to the CJ crystal surface. Indeed, SEM imaging clearly reveals surface associated CuBTC microcrystals (**Figure 4.5 A**).

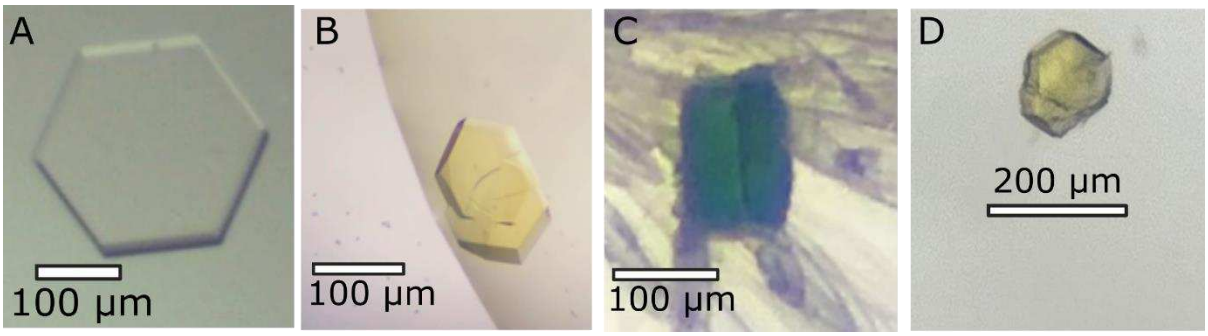


Figure 4.3 The various colors of CJ crystals after being loaded with different MOFs. **A)** An empty CJ crystal. **B)** An empty CJ crystal that has recently been crosslinked with glyoxal. **C)** A crosslinked CJ crystal with CuBTC grown in and around the crystal. **D)** A crosslinked CJ crystal with UiO-67 grown in and around the crystal.

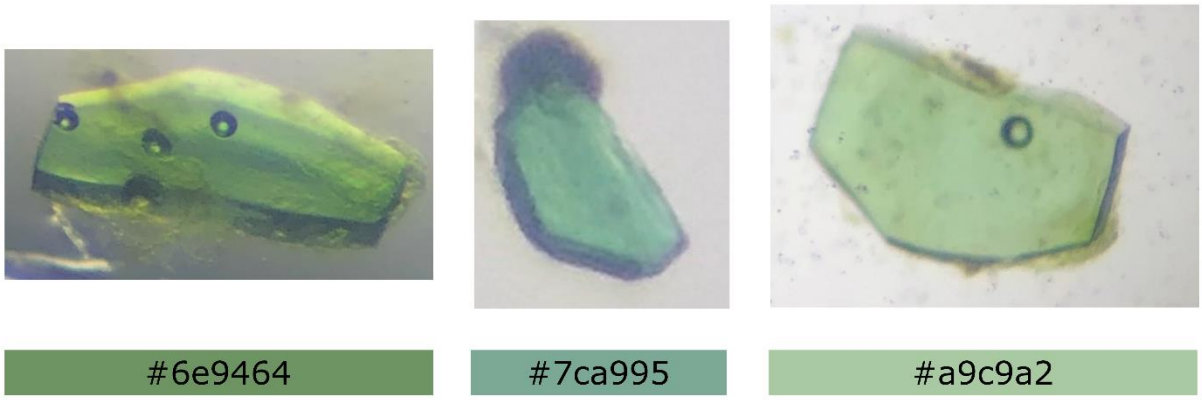


Figure 4.4 Some CuBTC@CJ crystal photos taken to highlight the color change after growth. These crystals were not taken with a scale bar attached, and subsequently are only used to further illustrate the color change that is observed after interpore and surface CuBTC growth. We can specify that the crystals comfortably fit in the 100 – 150 µm Hampton Research Mounted CryoLoops.

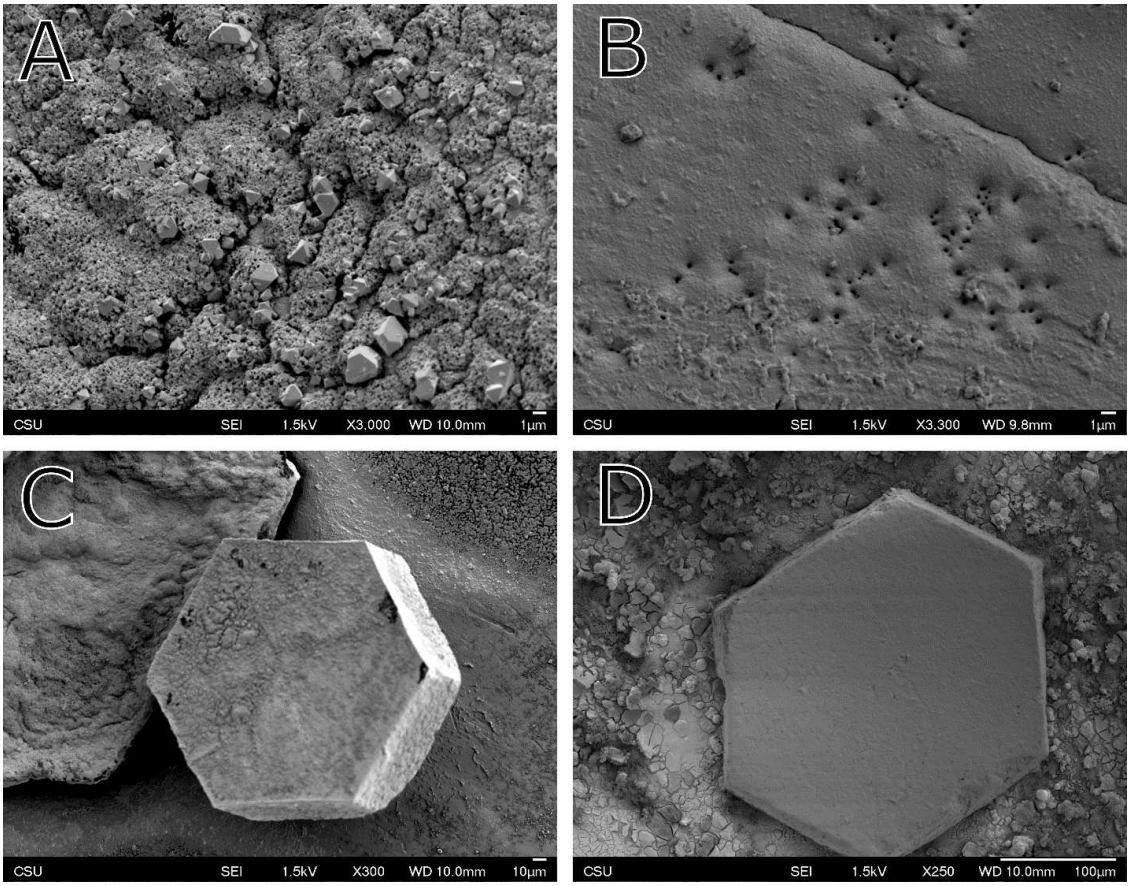


Figure 4.5 SEM images of CuBTC@CJ and UiO-67@CJ. **A)** SEM of the surface of a CuBTC@CJ crystal at 3,000x magnification. Individual micro CuBTC crystals are observable with their expected octahedral habit. **B)** SEM of the surface of a UiO-67@CJ crystal at 3,000x magnification. Instead of the expected octahedral morphology of UiO-67, a "blanket" is observed. **C)** The surface of a CuBTC@CJ crystal at 300x magnification. **D)** The surface of a UiO-67@CJ crystal at 300x magnification.

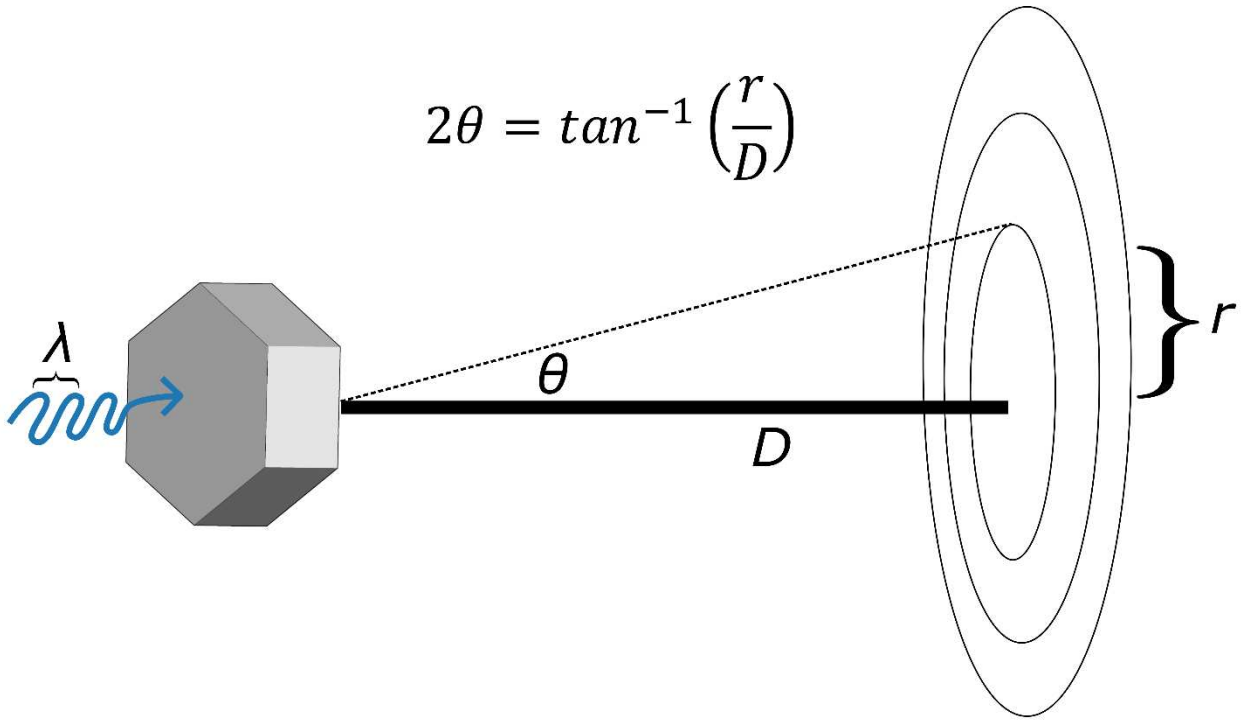


Figure 4.6 A physical representation of Debye-Scherrer ring generation in a scXRD experiment. An X-ray beam strikes the face of a MOF@CJ crystal and is refracted at a certain angle, depending on the orientation of the MOF nanocrystal lattice it strikes. These individual refractions combine to form Debye-Scherrer rings on the detector (x, y, intensity). To create a pXRD plot, we begin at the center of the image, then average all of the pixel intensities at every radial step outwards to create a plot of average intensity vs radius. Radius can then be converted into 2θ to compare the computed pXRD to the expected one. For example scripts, see https://github.com/jbderoo/MOF_in_CJ

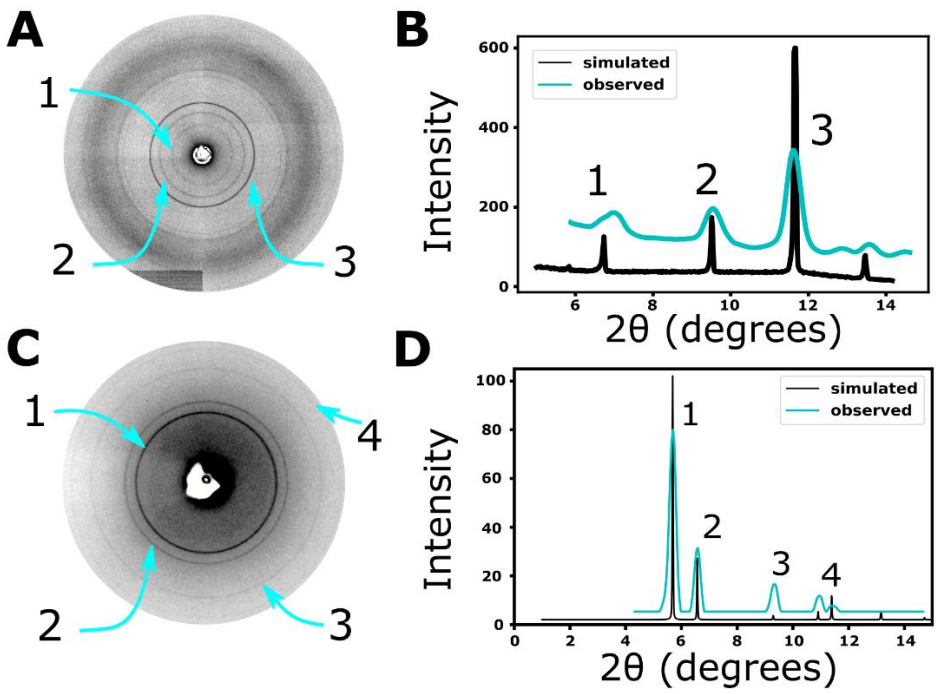


Figure 4.7 Single crystal x-ray diffraction (collected on the ALS 4.2.2 beamline) patterns and Debye-Scherrer rings of **A)** CuBTC@CJ and **C)** UiO-67@CJ, respectively. Extracted powder x-ray diffraction patterns from the scXRD, overlayed with the simulated PXRD patterns of **B)** CuBTC and **D)** UiO-67, respectively.

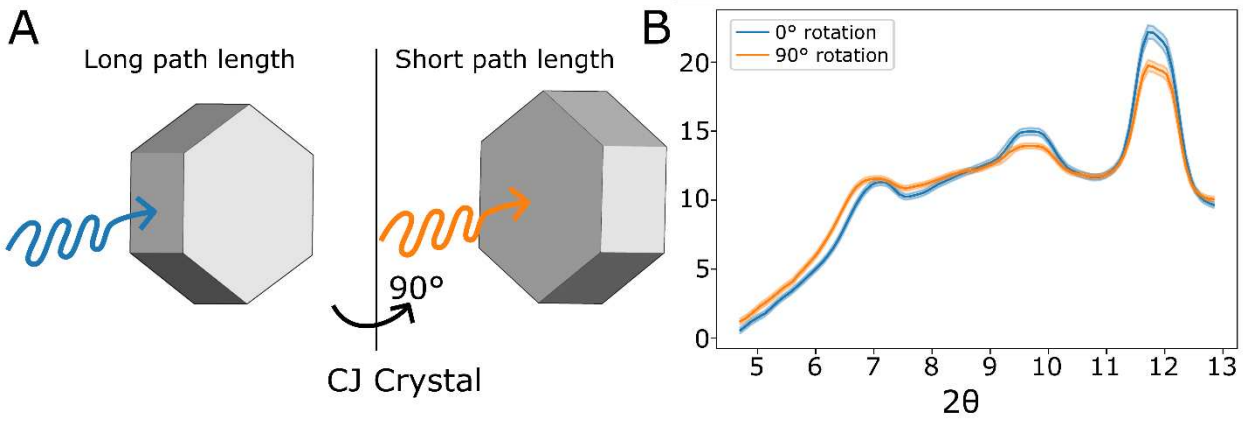


Figure 4.8 A) A CuBTC@CJ crystal was shot at two different orientations. When the X-ray beam was perpendicular to the major nanopore axis (blue) the path length through the crystal interior was longer, when the X-ray beam was parallel to the major nanopore axis (orange), the

path length was shorter. **B)** The pXRD plots (+/- 1 standard deviation in intensity along the azimuthal ring, highlighted) of both orientations reveal a clear anisotropy. While there are other possible explanations, the higher diffraction intensity for longer path length is consistent with deposition of CuBTC nanocrystals throughout the CJ crystal interior.

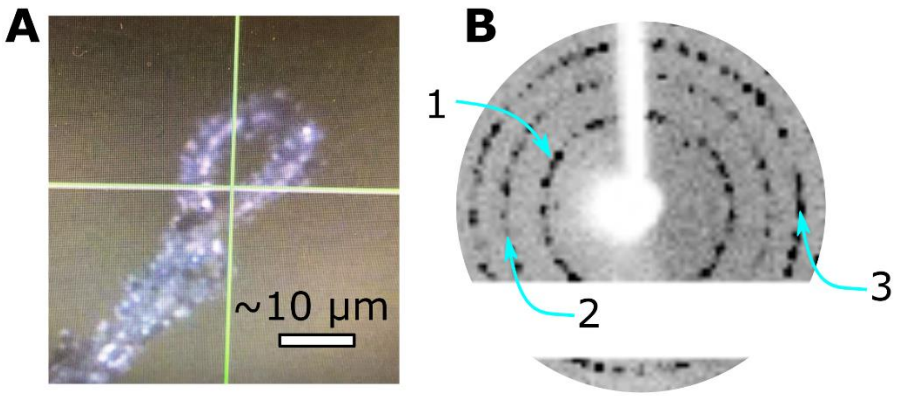
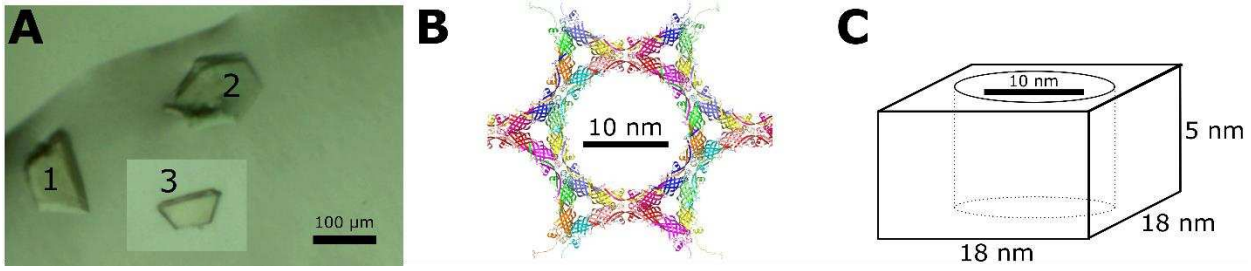


Figure 4.9 A) A picture of a ~10 µm Hampton Research crystallography loop with CuBTC powder pressed into it. **B)** scXRD (taken on the homelab Rigaku) of CuBTC crystals pressed into a fiber loop, then shot on the single crystal diffractometer. This experiment illustrates how the beam passing through individual crystals creates individual Braggs spots. Here, shooting through a moderate number of microcrystals results in distinguishable Bragg spots that are found at the same angle that is populated by Debye-Scherrer rings in a pXRD sample.

To further demonstrate that the MOF was growing throughout the protein crystals' interior, the CuBTC was leached out of the CJ crystal by soaking in 10 µL water for 24 hours. The supernatant was analyzed for copper concentration via inductively coupled plasma atomic emission spectroscopy (ICP-AES) by Huffman Hazen Labs (Golden, CO). The copper content in the supernatant would have been sufficient to fill 74% of the crystal's interior with CuBTC (**Figure 4.10**). After 24 hours, the crystals had lost much of their visible color and had not changed in size significantly before and after CuBTC growth (**Figure 4.11**).



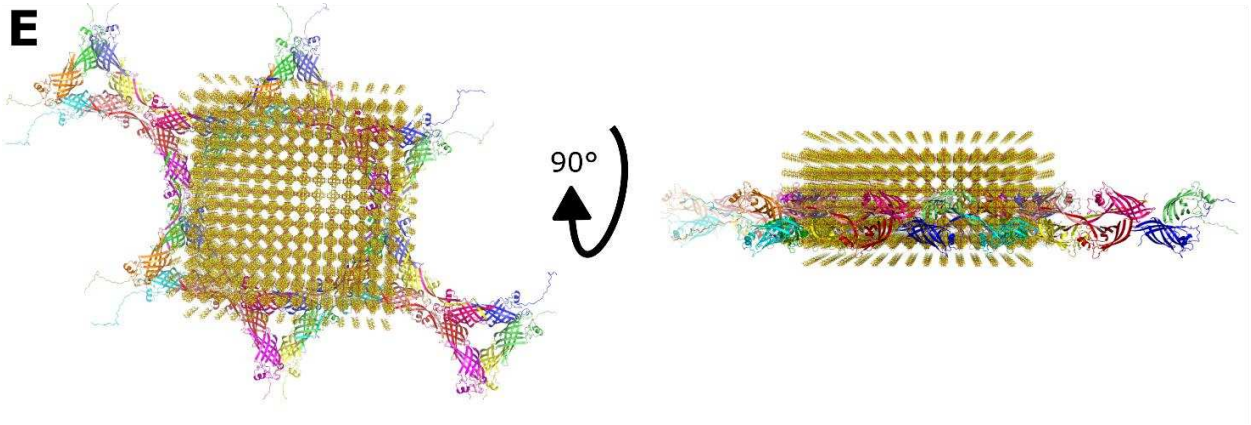
D

Section 1: Total Crystal Volume		
Crystal	Shape	Measurement Order
1	Trapezoidal prism	base, base, width, height
2	Hexagonal prism	diameter, height
3	Trapezoidal prism	base, base, width, height
Measurement 1 (µm)		Measurement 2 (µm)
138		74
66		N/A
100		55
Measurement 3 (µm)		Measurement 4 (µm)
76		17
N/A		17
42		10
Volume (µm^3)		Total volume (µm^3)
136952		361894
192392		
32550		

Section 2: CJ Nanostructure	
Diameter of pore (nm)	13
Depth of pore (nm)	5
volume of pore (nm^3)	663.7
Total volume (nm^3, Section 1)	3.62×10^{14}
length of CJ unit cell (nm)	18
width of CJ unit cell (nm)	18
volume of CJ unit cell (nm^3)	1403
Number of unit cells (N pores)	2.58×10^{11}

Section 3: Copper Quantity via ICP-AES	
Cu (µg/L) in water leached from CJ	0.86
Cu (µg/L) in water (neg control)	0.03
Cu (µg/L) in water leached due to CJ	0.83
sample water (mL) sent for analysis	50
sample water (L) sent for analysis	0.05
Cu (µg) present in sample	0.0415
Cu (g) present in sample	4.15×10^{-8}
Molecular weight (g/mol) Cu	63.546
Cu (mol) present in sample	6.53×10^{-10}
Cu (atoms) present in sample	3.93×10^{14}
Cu (atoms) in a single pore (experimental)	1524

Section 4: CuBTC packing fraction in CJ	
Cu (atoms) in a single pore (theoretical)	2059
Cu (atoms) in a single pore (experimental)	1524
Packing fraction CuBTC in CJ (%)	74



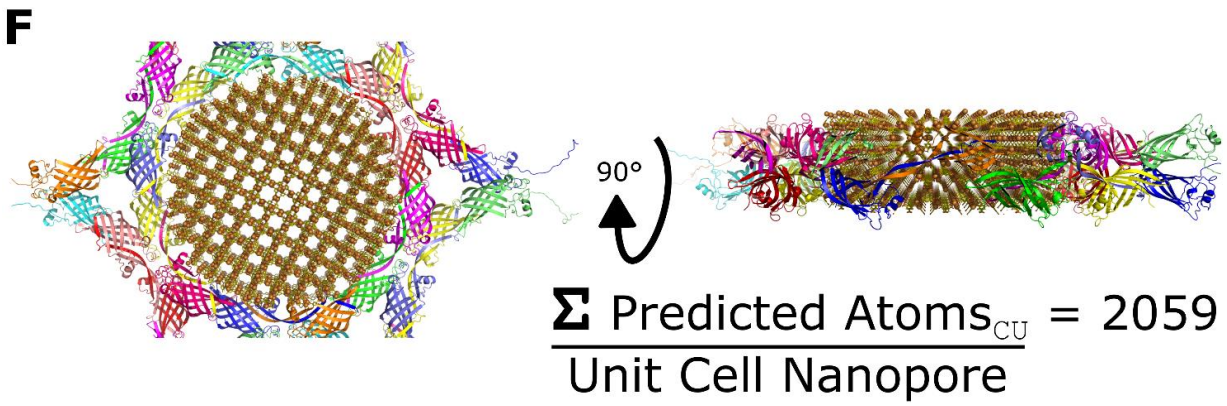
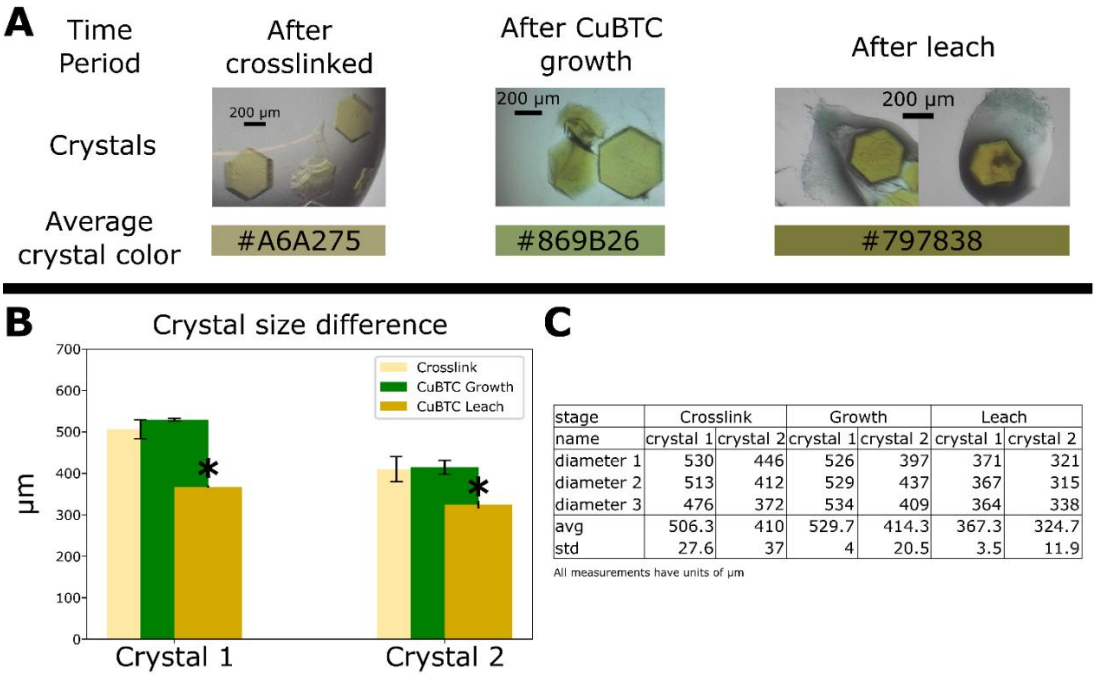


Figure 4.10 Analysis of the amount of CuBTC that was adsorbed into the CJ protein crystals. **A)** A composite image highlighting the crystals that were used to CuBTC grown inside, then subsequently leached out. Note: crystal 3 had drifted very far from crystals 1 & 2, and the image was composited for ease of visualization. Crystals images were taken prior to CuBTC growth and leaching. **B)** The crystal structure of the CJ protein, PDB code 5W17. **C)** The geometry of the CJ crystal nanostructure, highlighting the 13 nm diameter pore and the 18 nm unit cell. The total volume of the CJ unit cell is approximately 1403 nm^3, but not all of this volume is available for a guest MOF – only the interior cylindrical pore is accessible, with a volume of 663.7 nm^3. **D)** A four-section, step-by-step explanation of the calculation of the 74% packing fraction. First, measurements of the crystals in **A)** were taken via ImageJ, enabling a calculation of protein volume (**D, Section 1**). The total pore volume and unit cell volume was calculated, then finally a total number of unit cells (**D, Section 2**). The copper concentration was determined experimentally by Huffman Hazen in Golden, CO via ICP-AES (**D, Section 3**). The theoretical amount of CuBTC that could fit inside the CJ pore was determined by creating a large structured grid of CuBTC and overlaying this on the CJ structure (**E**). The MOF was then limited to exist only just beyond the major axial nanopore (13.5 nm diameter) the pore wall, and then a final clash detection step was used to create a perfect fit between the MOF and the solvent accessible surface of the CJ nanopore (**F**). This is analogous to taking a ring mold (CJ nanostructure) and using it to cut perfect, evenly sized cookies from a sheet of cookie dough (large CuBTC grid). The total number of copper atoms (per unit-cell nanopore) were finally estimated as 2060 atoms per CJ pore. With a theoretical maximum number of Cu atoms and an experimental number of Cu atoms determined via leaching, a packing fraction of 74% is calculated (**D, Section 3**). Scripts for calculating the theoretical maximum number of Cu atoms in the pore can be found in our github at https://github.com/jbderoo/MOF_in_CJ .



stage	Crosslink		Growth		Leach	
name	crystal 1	crystal 2	crystal 1	crystal 2	crystal 1	crystal 2
diameter 1	530	446	526	397	371	321
diameter 2	513	412	529	437	367	315
diameter 3	476	372	534	409	364	338
avg	506.3	410	529.7	414.3	367.3	324.7
std	27.6	37	4	20.5	3.5	11.9

All measurements have units of µm

Figure 4.11 Analysis of the change in size and color of CJ crystals before and after CuBTC had been grown inside them. **A)** 2 ideal crystals are shown after crosslinking but before CuBTC growth (left), after CuBTC growth but before CuBTC leaching (middle) and after CuBTC leaching into water (right). CuBTC was allowed to grow for 24 hours and leach out of the CJ crystals for 24 hours. Additionally, the average color of the crystals is provided both as a color and its corresponding hex color code. Interestingly, partially formed CuBTC or CuBTC components were visible outside the CJ crystal after leaching. **B)** Crystals were analyzed for size change with ImageJ, and each of the 3 treatments are plotted for both crystals' diameters. There was a statistically significant shrinkage in the crystals after leaching, but this is likely due to the drops being allowed to nearly dry out rather than due to MOF dissolution. Shrinking during desiccation also explains the phenomenon of the crystal becoming a darker yellow; the crystal is shrinking, but the number of blue-light-adsorbing glyoxal end products remains the same, thus increasing the spatial density of the yellow color. **C)** The raw data (in µm) used to generate the plots in **B)**.

We performed azimuthal integration (see https://github.com/jbderoo/MOF_in_CJ for extensive explanation and demonstration) on the scXRD data to generate pXRD data (**Figure 4.7 A, B**) and compared it to the expected pXRD pattern of CuBTC (computed with the structure of CuBTC and VESTA). While CJ crystals maintain macroscopic stability and overall shape in

organic solvents like DMSO, the precise nanostructure is lost. No high-resolution Bragg reflections were detected in the scXRD from the protein crystal. In contrast, the MOF pXRD "fingerprint", particularly peaks (2,0,0), (2,2,0), and (2,2,2), were extremely similar with peak angular displacement difference of less than 1%. This indicates that the CuBTC metal organic framework is present. From the computed pXRD curve, the Debye-Scherrer equation (Equation 1) can be used to approximate the size of the polycrystalline material present (D), from the dimensionless shape factor (K, assumed to be 0.9), X-ray wavelength (λ = 0.107 nm), the full width half max value ($\beta = 0.273$), and the Bragg angle ($\theta = 11.65$).

Doing so yields an average particle size of 5.7 Å. This is not a physically reasonable value since the unit cell edge length of CuBTC is 26 Å[111]. However, this analysis is consistent with the average CuBTC@CJ crystal diameter being extremely small, presumably due to confinement within the CJ crystal nanopores.

$$D = \frac{K\lambda}{\beta \cos\theta} \qquad \text{Equation 1}$$

While most CuBTC growth conditions rely on high temperature, here we are growing CuBTC inside and outside the host CJ crystals at room temperature (298 K). We hypothesize that solvent-exposed histidines (including disordered histags) within the CJ nanopores could bind Cu atoms, thus favoring CuBTC nucleation and growth within the pores. In this model for

nucleation, several BTC molecules will come and bind to a copper-histidine complex, followed by more coppers onto the recently bound BTC molecules, and the process repeats. A possibly over-simple model for MOF growth would be to assume uniform deposition of MOF nanocrystals throughout the body of host CJ crystals. In this case, varying the crystal orientation via the goniometer would change the number of MOF unit cells intercepted by the X-ray beam proportional to the path length of the host CJ crystal, thereby resulting in angle dependence of the total pXRD diffraction intensity (**Figure 4.6**).

4.3.2 The porous protein crystal can host different MOFs

UiO-67 was grown in/outside of the CJ crystal, demonstrating the ability to grow a variety of MOFs in conjunction with the CJ crystal. Crosslinked CJ crystals were rested in a UiO-67 precursor solution inspired by UiO-67 synthesis procedures[24,112] for 24 hours. A subtle color change of pale yellow to opaque yellow/white (**Figure 4.3 C**) was observed. SEM images of the UiO-67@CJ crystals did not reveal individual crystals of UiO-67 latched onto the surface of the CJ crystals, but instead showed a smooth "blanket" on the surface of the SEM (**Figure 4.5 B,D**). Critically, these crystals produced continuous Debye-Scherrer rings (**Figure 4.7 C**). After azimuthal integration, the observed pXRD pattern was again a close match to the expected pattern (**Figure 4.7 D**), specifically comparing peaks with Miller indices (1,1,1), (2,0,0), (2,2,0), and (3,1,1). The observed diffraction pattern and subtle color change were consistent with the growth of UiO-67 in and around the CJ protein crystal.

4.3.3 The metal organic frameworks could be used as a capping or protecting agent for guests installed into the protein crystal

Given the limited resolution of the SEM images, it was not possible to directly observe the presence or absence of open nanopores from the original crystal. Therefore, to provide supporting evidence for the supposition that UiO-67 growth sealed the CJ nanopore array, we turned to confocal microscopy. Confocal microscopy allows us to image a Z-stack with focal planes inside large CJ crystals. We can incubate such crystals in a volume of fluorescent protein after UiO-67 growth, and observe the change in fluorescence of the UiO-67@CJ crystals.

After observing the change in fluorescence of some UiO-67@CJ protein crystals, we found that UiO-67@CJ crystals nearly completely blocked the entry of super folder GFP (sfGFP) from entering the crystal interior (**Figure 4.12 A**). Initial values of fluorescence are taken as the zero point. After 60 seconds an additional 1 μL drop of super folder GFP (sfGFP) was added to create a strong driving force for sfGFP uptake. The resulting sfGFP transport caused a gradual increase in background fluorescence. The crystal that underwent UiO-67 growth remained significantly darker than the sister crystal that was incubated in water for an equal amount of time (**Figure 4.12 B,C**).

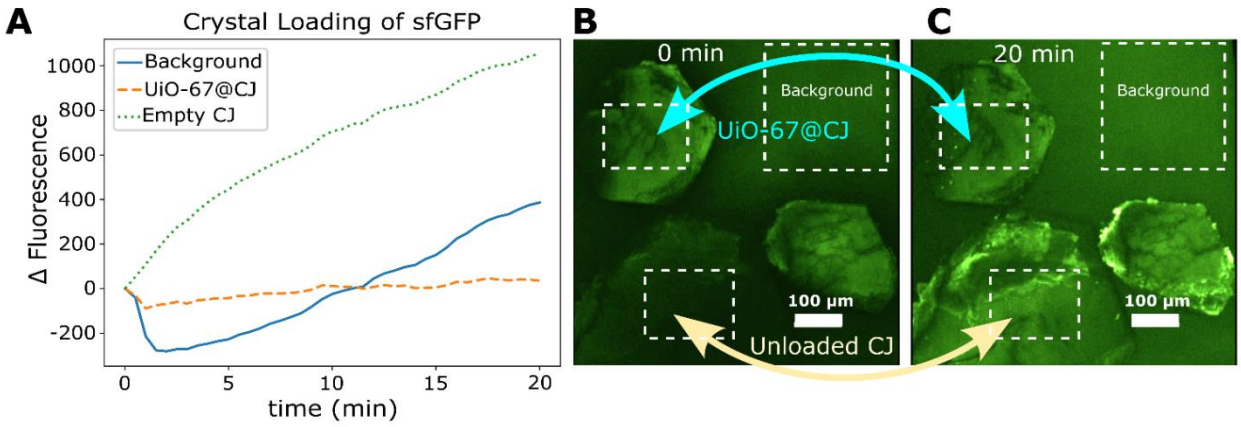


Figure 4.12 Change in fluorescence over time for 2 crystals and the background. 2 crystals are monitored for their change in fluorescence (accumulation of sfGFP, ROI highlighted with white dotted boxes) over time. **B)** A confocal image of the crystals (both UiO-67 loaded and not) as soon as the sfGFP is added. ROI's are shown in white dotted boxes for each region, corresponding to the labels in figure **A)**. **C)** The same crystals 20 minutes later, allowing sufficient time for sfGFP to diffuse into the pores of the CJ crystals.

4.4 Conclusions

Here, we present a new combination of materials with a protein crystal acting as a scaffold for internal and surface growth of metal organic frameworks. Both protein crystals and MOFs have their own expansive history and use cases, and this paper marks the first reported combination of these porous materials. The MOF@PPC combination has the potential to support highly organized, spatially optimized chemical reactions. Candidate application for the semi-biological materials include hyper stabilization, dual catalysis, *in vivo* gas storage, separation, delivery, and protective guest delivery. Future work includes exploring unique doping combinations, finding greener and more protein friendly MOF deposition conditions, and material synthesis scaleup. Comparing the relative catalytic activity of solution-suspended versus protein-encapsulated MOFs will be particularly interesting. Hypothetically, the high surface area and lattice defects expected for UiO-67 grown within the protein crystal could enhance catalytic

degradation of organophosphorus nerve agents - a critical avenue requiring further exploration. This heterogeneous catalyst represents a steppingstone to a new frontier for potential methods of chemical inactivation of harmful chemical nerve agents.

4.5 Materials and Methods

4.5.1 CuBTC

$Cu(NO_3)_2$ x 2.5 H_2O (copper (II) nitrate hemipentahydrate) and BTC (Benzene-1,3,5 tricarboxylic acid) were purchased from Sigma Aldrich and used without further purification. 2.45 g $Cu(NO_3)_2$ and 1.16 g BTC were sonicated into pure 10 mL DMSO (dimethyl sulfoxide)[19] to form a stable precursor solution of CuBTC. Vortexing and sonication were used until dissolution was complete.

4.5.2 UiO-67

$ZrCl_4$ (Zirconium tetrachloride) and BPDC (biphenyl-4,4'-dicarboxylate) were purchased from Sigma Aldrich and used without any further purification. A 1 mL sample of DMF (n,n-dimethyl formamide) was prepared with the $ZrCl_4$ and the BPDC linker at 1:1 stoichiometry at 17 mM. 5 µL water was added to the solution. The solution was vortexed and sonicated until complete dissolution was reached[24,112].

4.5.3 Protein expression and purification

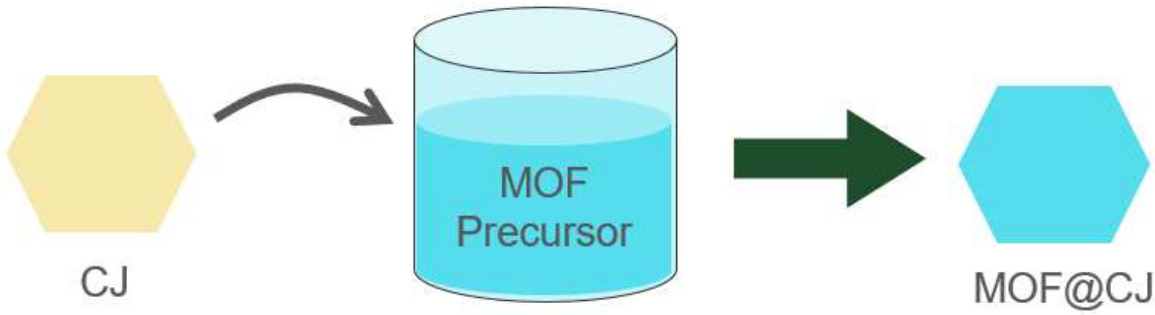
As detailed by Huber et al., a gene encoding "CJ" an optimized variant of GenBank ID CJ0420 cloned into the pSB3 expression vector[88].This was expressed in BL21 *E. coli.* The target protein was purified via metal affinity chromatography column[73].

4.5.4 CJ porous protein crystal fabrication and crosslinking

The aforementioned protein was crystalized via sitting drop vapor diffusion by mixing 1 µL of 14 mg/mL purified protein with 1 µL of crystallization buffer (3.2 M ammonium sulfate, pH 6.5, 0.1 M Bis-tris buffer). Crystals formed over the course of 3-7 days. After sizeable crystals had formed (100-150 µm diameter) they were washed three times in 4.2 M TMAO (trimethylamine oxide) at pH = 7.5 for 10 minutes per wash. The crystals were then crosslinked for 2 hours in the TMAO solution with 1% glyoxal and 50 mM DMAB. These conditions were previously optimized to ensure complete crosslinking and crystal integrity[88].

4.5.5 MOF@CJ combination

The crosslinked protein crystals were submerged in the respective MOF precursor solutions for 24 hours. They were then transferred to a clean solution of their respective organic solvents and given a gentle swirl to knock off any loosely adhered surface MOF that had grown on the surface of the crystal. These crystals are the final result, and are named CuBTC@CJ and UiO-67@CJ crystals, respectively.



4.5.6 MOF@CJ material analysis

Putative CuBTC@CJ single crystals were shot on a single crystal x-ray diffractometer (scXRD). Data were collected at either the Advanced Light Source (ALS) Beamline 4.2.2 at Berkeley National Laboratory, or on a local Rigaku HomeLab. When a polycrystalline material is shot on an scXRD, the resulting diffraction patterns contain Debye-Scherrer rings instead of individual Braggs spots. scXRD data can be converted to powder x-ray diffraction data (pXRD) data by azimuthal integration (averaging the radial intensity). Examining the meta data of the scXRD image provides a conversion of pixels to radial distance, as well as the sample to detector distance. Using Equation 1, meta data from the image, and the pixel intensity vs radial distance data, we compute pXRD data from scXRD for direct comparison (**Figure 4.6**). Example processing scripts are available at our github. The crystal lattice of the CJ crystal is disrupted by the organic solvents, and so its Braggs spots are lost. Equation 2 (Bragg's Law) relates the diffraction angle (2θ) to the radial distance (r) and the distance between the sample and the detector (D).

$$2\theta = tan^{-1}\left(\frac{r}{D}\right) \qquad \text{Equation 2}$$

4.5.7 SEM images of MOF@CJ crystals

Crystals were made following the MOF@CJ combination protocol previously described. Afterwards, the crystals were gently cleaned in their respective growth solvent (DMF/DMSO). SEM imaging was performed using a JEOL JSM-6500F microscope. An accelerating voltage of

15.0 kV was used to image the MOF@CJ crystals. All samples were prepared by looping them from cleansing media onto an aluminum SEM stage into 2 μL of DI water. The samples were allowed to dry overnight before they were placed under vacuum and coated with 20 nm of gold prior to imaging.

Chapter 5 – Concluding Remarks and a Summary of Additional Research Projects

This concludes the description of a substantial body of work I have completed under the guidance of Dr. Melissa Reynolds and Dr. Christopher D. Snow. For the sake of brevity, I have focused above on a limited subset of my CSU research. The remainder of this section is dedicated to conclusions and future works, brief blurbs about outstanding side projects I have contributed to (as first author or otherwise), and other projects.

The contents of this dissertation have made significant contributions towards the field of interdisciplinary protein engineering, both experimentally and computationally. Throughout the chapters, the work discussed has made strides towards identifying how an antibody and antigen bind together computationally via AlphaFold2, decreasing the cost-barrier of high throughput protein crystallization via an inexpensive, general-use liquid handling robot, and presentation of a hybrid material that is part porous protein crystal, and part metal organic framework. Chapter 2 is an article that will be published in eLife. We recently received reviewer comments and I am currently drafting the initial responses and revisions. The experimental verification presented in Chapter 2 was conceived of and executed by Dr. Brian Geiss and James Terry (I conceived, wrote, and executed the computational work, which is the majority of the manuscript). Chapter 3 is an article intended for publication in SLAS Technology's special issue of Robotics in Laboratory Automation, and is still under peer review, submitted January 15th, 2024. The article was mostly written by Jacob DeRoo. All code and experiments were written by Jacob DeRoo. Chapter 4 is undergoing final editing and will be submitted within the next few weeks to "RSC crystal engineering communications" or "MDPI crystals". The Chapter 4 experiments were

conducted by Jacob DeRoo withassistance from Rojina Shrestha. The code and the paper was written by Jacob DeRoo. All code is available at https://github.com/jbderoo/ in their respective project's repositories. None of this would have been possible without the continual guidance provided by Melissa Reynolds and Christopher Snow. Members of the TagTeam group, consisting of the Snow, Geiss, and Stasevich labs were instrumental in the success of Chapter 2 – particularly Dr. Brian Geiss, Dr. Ning Zhao, and (very nearly Dr.) James Terry.

5.1 – Chapter 2 PAbFold Future Outlook

In Chapter 2, a pipeline was made in Python centralized around AlphaFold to evaluate where an antibody might bind to an antigen. This was done by dividing the antigen sequence up into many subsequences, then predicting a structure of both the antibody and the antigen-subsequence. After this has been done for every antigen-subsequence, we can evaluate how well all of the subsequences bind with the antibody using the evaluation metric pLDDT. Investigating a plot of residue number vs pLDDT and rank sorting all the antigen-subsequences lets us identify which of them bound to the antibody well, thus elucidating which of the antigen-subsequences may be the experimental epitope. This was then computationally tested and experimentally verified on mBG17:SARS-CoV-2 Nucleocapsid spike protein, a structure that AlphaFold didn't have access to in its learning set. We verified that what was computationally predicted is actually observed experimentally. Future directions for this project include the exploration of more antibody:antigen pairs, a deep dive into how the multiple sequence alignments are generated and why they influence AlphaFold's predictions so heavily, investigating how well AlphaFold2 can

discern correct epitope sequences when placed in direct contest with other epitopes, and a reimplementation of the pipeline with the advent of AlphaFold 3.

5.2 – Chapter 3 Protein Crystallization Automation Outlook

In Chapter 3, the development of several Python protocols for the Opentrons for protein crystallization are presented, in addition to both key 3D printing files (.stl) and corresponding data files for the Opentrons' interpretation of those custom print attachments. The automation of protein crystallization would reduce human error in plate preparation, and would allow for continuous use – a robot can't get too tired to work! While high throughput protein crystallization machines already exist (such as the Gryphon and Mosquito), these machines are dedicated to crystallization trials and are costly, in the neighborhood of $75,000 for the Gryphon machine. These machines also require substantial upkeep and thorough cleaning procedures. The Opentrons is cheap ($15,000 for the Opentrons2) and cleaning procedures aren't required as routinely as with the permanent syringes on the Gryphon. We successfully grew Hen Egg White Lysozyme crystals, an ideal and model protein crystal, as well CJ, as a practically useful protein crystal the Snow lab uses in bulk quantities. Future directions for this project include an exploration of the Opentrons' operating parameters, to induce larger CJ crystal formation. These parameters include the timing and quantities for liquid pipetting, the consumable crystallization plate type, speed of plate preparation, and when the plate is sealed with tape for vapor-liquid equilibration. Additionally, the addition of AI vision via a camera that can see into the growth wells, and keep a record of how well each condition produces crystals of a given size, would be a

way to “close the loop”, and more fully automate the search for optimal protein crystallization conditions.

5.3 – Chapter 4 MOF@PPC Outlook

In Chapter 4 two metal organic frameworks, CuBTC and UiO-67, were grown in and around large CJ protein crystals. Python scripts were made to do the data analysis of any subsequent MOF@Porous protein crystal growth, with the intention to produce a pXRD pattern that one could then compare to the expected pXRD pattern for verification of the atomic structure. SEM images of the crystals were taken, and sparse, individual CuBTC crystals were seen on the surface of the protein crystal – however, a “blanket” of MOF was observed on the UiO-67@CJ surface. To elucidate whether the CJ crystal had MOF grown only on the surface or indeed throughout the pores, we collected data at two different orientations of the protein crystal – thus changing the path length, and the number of individual crystals detected – and observed a significant difference between the two. Future directions for this project include an exploration of other MOFs that could be grown inside the CJ crystal, and a refinement of the UiO-67@CJ growth conditions. Additionally, the growth of a MOF under less harsh conditions to try and preserve the diffraction of the CJ crystal to simultaneously see both CJ and MOF would allow us to solve the structure, and see exactly how a MOF is growing in conjunction with the CJ pore wall and surface. A very interesting application for this would be the elimination of chemical warfare threats, particularly G-class nerve agents. UiO-67 catalyzes the organophosphate bonds present in these chemicals quickly. While I was able to replicate UiO-67 powder deactivating dimethyl n-propylamine (a pesticide that is a simulant for these nerve agents), I was not able to

produce enough UiO-67@CJ crystals to reliable degrade DMNP faster than it decomposes naturally in water.

5.4 – Additional Projects

The remainder of this chapter are extra projects that I have contributed to, listed in no particular order. Often these projects are documented elsewhere, and as such only receive short project descriptions here.

5.4.1 NREL L-sugar

In this unpublished work, I collaborated with Dr. James Henriksen and Dr. Rich Conant during the summer of 2022 to develop a pipeline for quantifying the compatibility of various L-sugars with a given enzyme. This automation was necessary as Dr. Henriksen generated extensive lists of proteins from several databases, ranging from approximately 1,200 to 500,000 proteins. The pipeline, currently hosted in a private GitHub repository (https://github.com/conant-csu/LSugarDockML), is designed to be deployed on large compute servers (Google Cloud, Microsoft Azure, AWS) due to the computational intensity of generating metrics for each protein pairing.

I completed the pipeline development, which folds enzyme sequences from InterPro and Chebi/Rheia, docks over 30 L-sugars and NADH into the enzymes using AutoDock Vina, and refines the docked molecules with Rosetta. The pipeline's initial application to a proof-of-concept set of L-sugars and enzyme sequences yielded preliminary results that were used to secure a $3 million grant from Microsoft. Following this, a machine learning model will be developed to classify the data based on the pipeline results and experimental parameters, aiming to identify the most promising enzyme-sugar combinations for catalytic efficiency.

This project, jointly conducted between CSU and NREL, aims to identify enzymes capable of producing L-sugars, which Dr. Conant's lab seeks to use in plants to deposit L-sugars into the soil. This process has been shown to recruit microbes that regenerate soil health, enhance soil longevity, scrub CO2 from the air, and contribute to a positive environmental cycle.

5.4.2 scFv Optimization

Single chain antibody fragment optimization work was done with the collaborative group called TagTeam, where the significant contributors are the Snow lab, Geiss lab, and Stasevich lab. A key protein structure is a single chain variable fragment (scFv) as is discussed and presented in Chapter 2. TagTeam uses scFvs for several applications, but the most common application is tagging intracellular proteins of interest with a fluorophore, so that we can monitor a specific intracellular process. Commonly, when an antibody is reduced to an scFv, functional issues arise – aggregation, solubility, and binding affinity are increased or decreased. One approach to circumnavigating these issues is by creating scFv chimeras: taking the framework of one scFv, and the CDR loops of another antibody, and grafting them together to create a new scFv. While this process (and the process of creating an scFv) aren't necessarily labor-intensive tasks, it can take around 5-7 minutes to complete and is prone to human error. Here I present two simple Python scripts: the first script intakes both V_H and V_L antibody sequences and returns an scFv sequence, and the second script intakes two scFv sequences and returns a single scFv sequence that has the framework from one of the input sequences and the CDR loops from the other. These simple scripts have seen extensive use by both the Geiss and Stasevich labs to create

and test new scFvs and chimeric scFvs (**Figure 5.1**). As is exemplified in Chapter 2, this is done with the 15F11 and 2E2 antibody frameworks.

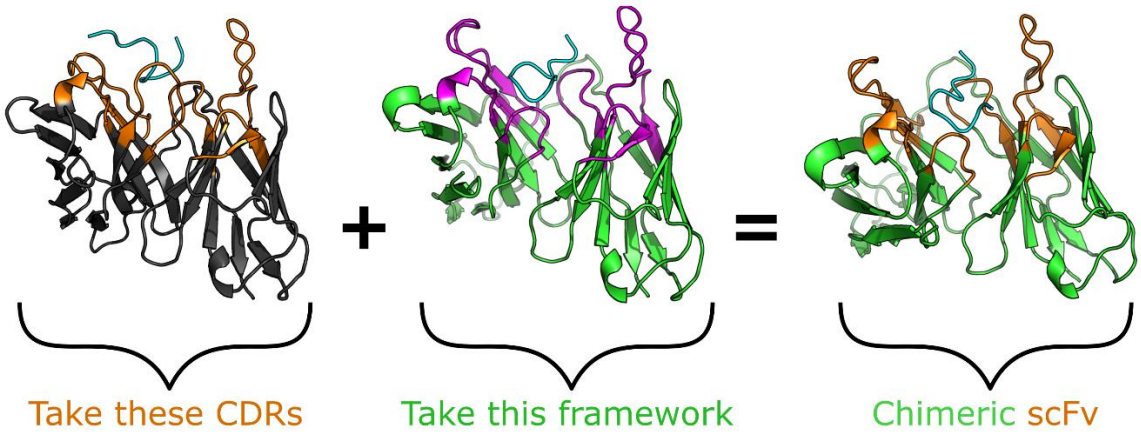


Figure 5.1 Chimeric scFv creation example, where the scFv frameworks and loops are identified, then recombined.

There have also been significant efforts to rescue scFvs that have performed poorly, and try to discern why some scFv sequences perform well when others perform poorly when they have the same binding target. To this end, I helped to launch multiple projects.

For scFv optimization and rescue, we adopt two strategies: (1) CDR loop redesign and (2) scFv framework redesign. To redesign the CDR loops, every residue in the CDR loops was mutated to every amino acid, and the binding change was measured in two ways: ddG via Rosetta[113,114], or epitope pLDDT[115] change via AF2. After identifying the single best mutant, this was deemed the new "parent" scFv, and the process was repeated an additional 3 times, ending on a quad mutant. While Rosetta and AF2 disagreed on the best mutants, they did largely agree on which mutants were beneficial and deleterious. Finally, we probed top ranking AF2 mutants' binding affinities via pulling experiments followed by umbrella sampling with GROMACS. Redesigned CDR loops with increased binding affinity predicted by all 3 methodologies were

ordered and tested experimentally. Of the 4 ordered mutants, 1 bound 10% tighter than WT, 2 bound equally as tight, and 1 bound less tightly. Binding affinities were tested by Dr. Ning Zhao via FRAP, a method to quantify intracellular binding affinities between two targets if one target is relatively fixed in space and the other is bound to a fluorescent protein[116].

Secondly, a pipeline in Python was written to propose mutations that might overcome scFv solubility, stability, and expression challenges. In this pipeline, an antibody sequence or a V_H and V_L chain are provided and an scFv is created. The CDR loops are determined automatically via Kabat numbering or structurally by identifying what's in the paratope. Everything else in the scFv framework is then allowed to mutate via ProteinMPNN[76]. The new sequences are folded in AF2[1], and the best sequences are determined via metrics provided by AF2, namely pLDDT and pTM. This pipeline has approximately a 40% success rate in preliminary experimental testing, where a success is a sequence that performs equally as well or better with respect to solubility or ease of expression and still binds to its target.

5.4.3 Sasya

SasyaBio is a biotech company in Minneapolis, Minnesota, and there is a collaborative project between the Snow lab and Sasya. Some of the key information here has been intentionally kept vague, as I signed an NDA in 2022 pertaining to some aspsects of this project. In this project, we redesigned an enzyme to act on a homologous substrate. Catalytically active residues were identified from the literature, and additional residues proximal to the substrate were allowed to vary. We also incorporated insights from the initial library design round proposed by Dr. Christopher Snow. Utilizing Rosetta, we generated 10,000 enzyme variants with

the substrate docked and an additional 10,000 variants with the product docked. From these structures, we synthesized a library that integrated both sets of results. The initial library generation, led by Dr. Snow, identified two new "parent" mutants that exhibited strong experimental performance. These mutants served as the starting points for a second round of mutation discovery, wherein new regions of the enzyme are being explored for potential mutations. Additionally, I have written a thorough data analysis script for the folks at Sasya to increase their limited throughput of enzyme mutant analysis testing.

On behalf of the experimentalists here at CSU, I've also written a script to perform similar analyses on the plate reader data for the machine at CSU as to the one I wrote for the plate reader at Sasya. In an effort to help optimize how quickly the enzymatic reaction occurs, we also attempted to optimize how the enzyme produces the product from the reactant. The rate limiting step in this process is the C-terminus tail's movement around the active site. In an effort to expedite this process, we probed the idea of creating circular permutants of the enzyme. Circular permutation is the process of stitching the N terminus and the C terminus together, and then creating a new N-terminus and C-terminus elsewhere in the sequence. Sister sequences with new "cut sites" (i.e. new termini placements) were generated, then protein structures were predicted with AF2 for these new sequences. In an attempt to predict whether or not these sequences would be useable, from all of these structures I quantified 4 characteristics (**Figure 5.2**): (1) how far away was the cut site from the active site and center of mass of the protein, (2) global RMSD of the structures and RMSD of a key substructure within the enzyme, (3) contact order of the protein with the new cut site, and (4) the average pLDDT of the key substructure previously mentioned. The reasoning behind (1) is that placing the N-terminus and C-terminus regions on the exterior of the protein are the least likely positions to destabilize the protein and

interfere with the active site. The reasoning behind (2) is that we want our new proposed protein to have a similar shape to the WT structure. The reasoning behind (3) is that contact order shows how well a protein is touching itself, and that a protein is more likely to fold when all the contacts are neighboring residues and not far from one another in sequence space. There is a statistically significant relationship between contact order and how well a protein folds in solution[117]. The reasoning behind (4) is that structures with a high pLDDT predicted by AF2 are likely to have the shape predicted by AF2. From these candidates, we looked at sequences that would have a cut site far from the active site and center of mass, low RMSDs, high contact order, and high pLDDT. These sequences (and subsequently the remainder of the project) was handed off to Austin Knight, a PhD student in the Snow lab, for testing and verification.

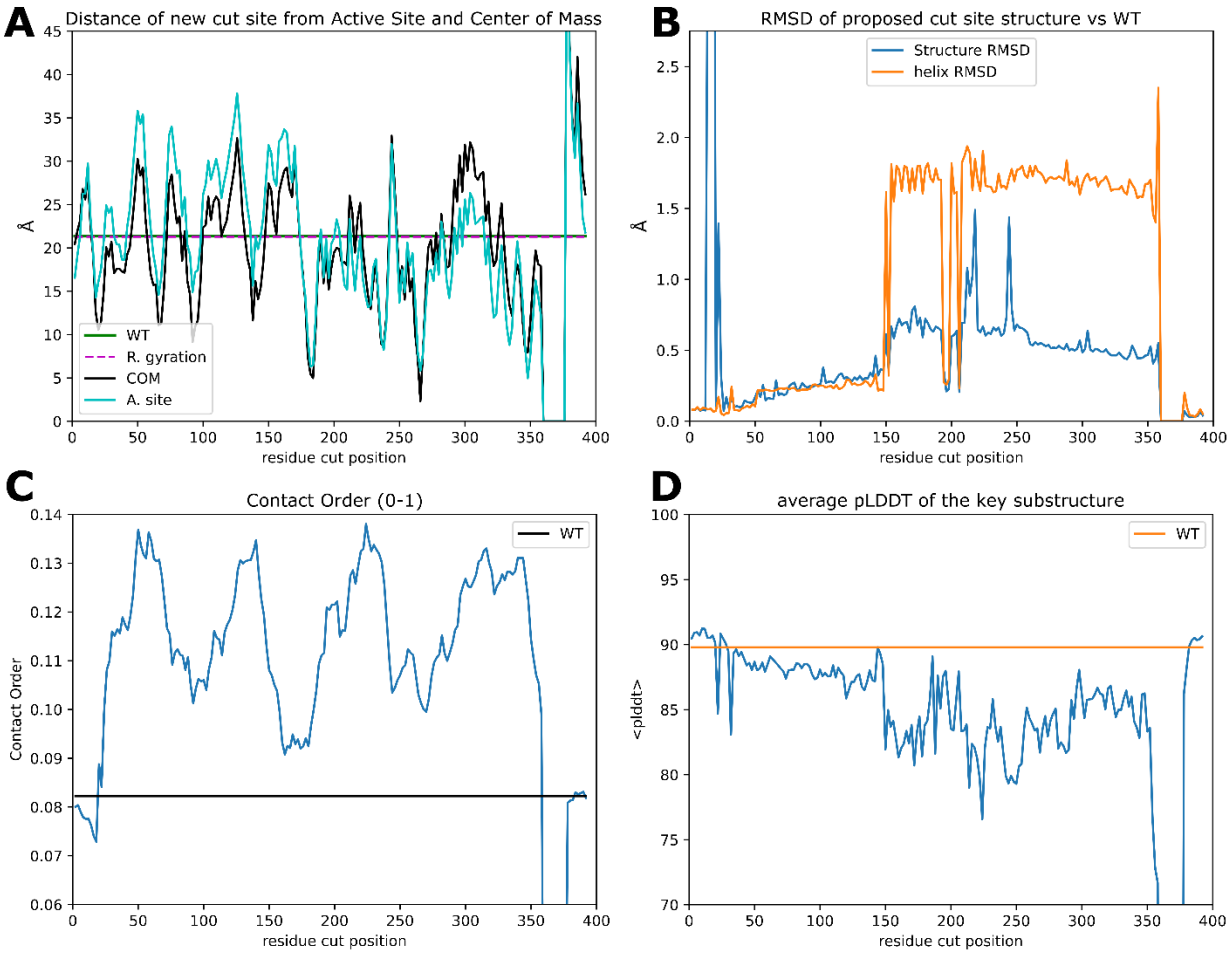


Figure 5.2 The four characteristics predicted for every proposed cut site in the enzyme. **A)** highlights the distance the new cute site is from both the active site and the center of mass. Wild type values are highlighted in green and dashed purple, and are very similar to one another (21 Å). **B)** showcases the difference in global RMSD (blue) and relevant substructure RMSD (orange) from the new predicted structure vs the WT structure. **C)** Contact order for each new proposed structure. While technically CO exists between 0-1, very few protein structures have a higher CO than 0.3. Ubiquitin, a small and ideal protein, has a contact order of about 0.23. **D)** Average pLDDT of the proposed structures.

5.4.4 Engineered Enzyme Small Molecule Dependence

In this unpublished work, Michael Scroggins, M.Sc., and I, under the guidance of Dr. Christopher Snow, aimed to create a dependency of the catalytic enzyme Salicylic Acid Carboxyl

Methyltransferase (SAMT) on the small molecule theophylline. SAMT produces salicylic acid, the compound responsible for the wintergreen smell. Initially, Michael and Dr. Snow identified potential pockets in SAMT that could spatially accommodate theophylline by mutating all side chains to hydrogens (glycine), leaving only the backbone intact. With these pockets identified, I generated 1,000 specific mutants and 1,000 fuzzy mutants for each pocket using Rosetta, simultaneously docking the ligand into the pocket and allowing the side chains to mutate either randomly (specific mutants) or to any amino acid of equal or smaller size than the wild type.

Subsequently, I relaxed the top mutant structures of each pocket 100 times with theophylline present and absent, aligning them all to the wild type (WT) SAMT to calculate the RMSD. This approach allowed us to indirectly measure the active site's dependence on theophylline: a high RMSD indicated significant movement of the active site, suggesting that salicylic acid production might be hindered. The results were then reviewed by the team to identify promising pockets and initiate the ordering of libraries from Twist Bioscience for expression and testing.

These protein variants were further categorized based on their respective pockets and subjected to relaxation protocols in Rosetta, both with and without the presence of theophylline. Following relaxation, we analyzed the residues involved in the binding and catalysis of salicylic acid to determine which pocket collapses had the most significant impact on key residue positioning. The top three pockets, demonstrating the greatest effect on key residue movement, were selected for experimental synthesis and testing. Currently, these sequences are awaiting experimental validation.

5.4.5 Mentoring Junior Protein Design Students

I led a small team of undergraduate and high school students in the development of a novel, smaller GFP mimic. Key residues responsible for chromophore positioning and maturation were identified from literature. Using RFDiffusion, a new protein backbone was designed around these key residues, and ProteinMPNN was employed to fix a sequence to this backbone. The sequences were then folded using AlphaFold2 (AF2). This process was repeated 1,000 times, and the folded structures were aligned with the native GFP to identify novel structures that conserved the orientation and spatial relationships of key residues. Structures with low RMSD were retained, further mutated using ProteinMPNN, and refolded with AF2. The top results from the second round were visually inspected, with some selected for expression and testing. Additionally, the same group of students was guided through redesigning the back end of a single-chain variable fragment (scFv) to host a small fluorescent molecule. This was achieved through a divide and conquer approach, where one student used Rosetta and another used LigandMPNN[118], an advanced version of ProteinMPNN capable of recreating protein sequences with consideration of a small molecule's presence.

5.4.6 Evaluation of the Adsorption-Accessible Surface Area of MIL-53(Al) using Cannabinoids in a Closed System

In this publication[119], I collaborated with Dr. Jamie Cuchiaro and Dr. Jon Thai under the supervision of Dr. Melissa Reynolds. My contributions included estimating the approximate surface area of several cannabinoids, specifically CBD, CBN, and THC, as well as the interior surface area of the metal-organic framework (MOF) MIL-53(Al) using the MSMS program.

Furthermore, I calculated the radius of gyration for each of these small molecules. Equipped with these parameters, I assisted Dr. Cuchiaro in refining her adsorption models for the cannabinoids within the MOF, enhancing our understanding of the interaction dynamics between these compounds and the framework. This work provided critical insights into the adsorption properties and potential applications of MIL-53(Al) for cannabinoid separation.

5.4.7 Cu-Based Metal–Organic Framework Nanosheets Synthesized via a Three-Layer Bottom-Up Method for the Catalytic Conversion of S - Nitrosoglutathione to Nitric Oxide

In this publication[120], I collaborated with Dr. Jon Thai, Dr. Rob Tuttle, and Dr. Jamie Cuchiaro under the mentorship of Dr. Melissa Reynolds. My contributions to this study involved assisting Dr. Jon Thai in synthesizing MOF nanosheets using the three-layer method. Following synthesis, I conducted powder x-ray diffraction (pXRD) analysis on these nanosheets to determine their crystalline structure. Additionally, I sent selected samples to Huffman Hazen Laboratories in Golden, CO, for elemental analysis using inductively coupled plasma atomic emission spectroscopy (ICP-AES). Utilizing the elemental composition data provided by Huffman Hazen, I calculated the atomic mass of the nanosheets. These analyses were crucial for understanding the structural and compositional properties of the MOF nanosheets, thereby supporting further research and applications of these materials. Future work on this project, which I had intended on doing but was limited by the data collection process and time availability, was to use aspect ratio data (obtained on the AFM via Jon Thai and his undergraduate students) and synthesis data to create a model that would let us find synthesis

conditions that would maximize the aspect ratio. A majority of catalysis is done on the surface of MOFs[103], thus allowing us to optimize the catalytic efficiency per mass of MOF.

5.4.8 Riboswitch Identification in the Human UTR

In this preprint pending reviews, Dr. Will Raymond and I developed 19 positive unlabeled classifiers using leave-one-out cross-validation on features extracted from 67,683 known RNA riboswitch sequences. Notably, there are currently no identified riboswitches within the human 5' untranslated region (UTR). Utilizing this ensemble of 19 models, we classified 48,031 human 5' UTR regions and identified 456 sequences that were highly likely to contain riboswitches based on our model predictions.

We further mapped these 456 candidate sequences to their most structurally similar riboswitches from our training dataset, ranking them according to their similarity scores and the output probabilities from the ensemble models. Our ranked results are made available to the broader scientific community through our GitHub repository: https://willraymond.github.io/human_riboswitch_hits_gallery/about/. This work represents a significant step toward identifying potential riboswitches in the human genome, providing valuable insights and resources for further research in RNA biology and gene regulation.

5.4.9 Envelope Protein Phylogenetic Analysis

We utilized ancestral sequence reconstruction (ASR) and AlphaFold2 to reconstruct and analyze the envelope protein of extinct ancestral tick-borne flaviviruses using modern-day

progeny. By employing phylogenetic analysis for the reconstruction of ancestral sequences, we were able to identify and model evolutionary changes within the phylogenetic tree and pinpoint structures that underwent the greatest changes between subsequent nodes. Specifically, we evaluated structural differences within Domain 3 of the envelope protein that might not be reflected solely through multiple sequence alignments, aiming to identify residues potentially linked to adaptation to host immune responses. Ultimately, our analysis demonstrated the viability of using AlphaFold2 for modeling and documenting protein structure evolution.

5.4.10 Course Instructor

In the Spring of 2022, Dr. Christie Peebles, the designated instructor for CBE 320 Reactor Design, took a sabbatical. Consequently, the course was left without an instructor, prompting Dr. David Dandy to request that I assume the teaching responsibilities. I have a strong passion for teaching, and reactor design/chemical kinetics is perhaps my favorite area within chemical engineering. The subject's blend of mastery, mathematical rigor, and coding simulations is particularly stimulating to me. I accepted the challenge of teaching the course and found the experience highly rewarding, with the exception of having to fail two students and grade 75 exams.

Bibliography

1. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

2. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science (80-.).* **379**, 1123–1130 (2023).

3. Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science (80-.).* **373**, 871–876 (2021).

4. Ahdritz, G. *et al.* OpenFold: retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization. *Nat. Methods* (2024) doi:10.1038/s41592-024-02272-z.

5. Mo, Y., Ovchinnikov, S., Mirdita, M. & Steinegger, M. LocalColabFold. *LocalColabFold* https://github.com/YoshitakaMo/localcolabfold (2021).

6. Durbin, S. D. & Feher, G. PROTEIN CRYSTALLIZATION. *Annu. Rev. Phys. Chem.* **47**, 171–204 (1996).

7. McPherson, A. & Gavira, J. A. Introduction to protein crystallization. *Acta Crystallogr. Sect. F, Struct. Biol. Commun.* **70**, 2–20 (2014).

8. Giegé, R. A historical perspective on protein crystallization from 1840 to the present day. *FEBS J.* **280**, 6456–6497 (2013).

9. Subramaniam, S. The cryo-EM revolution: fueling the next phase. *IUCrJ* vol. 6 1–2 (2019).

10. Saibil, H. R. Cryo-EM in molecular and cellular biology. *Mol. Cell* **82**, 274–284 (2022).

11. Renaud, J.-P. *et al.* Cryo-EM in drug discovery: achievements, limitations and prospects. *Nat. Rev. Drug Discov.* **17**, 471–492 (2018).

12. Lyumkis, D. Challenges and opportunities in cryo-EM single-particle analysis. *J. Biol. Chem.* **294**, 5181–5197 (2019).

13. Carroni, M. & Saibil, H. R. Cryo electron microscopy to determine the structure of macromolecular complexes. *Methods* **95**, 78–85 (2016).

14. Hu, Y. *et al.* NMR-Based Methods for Protein Analysis. *Anal. Chem.* **93**, 1866–1879 (2021).

15. Purslow, J. A., Khatiwada, B., Bayro, M. J. & Venditti, V. NMR Methods for Structural Characterization of Protein-Protein Complexes. *Front. Mol. Biosci.* **7**, (2020).

16. Vellingiri, K., Philip, L. & Kim, K.-H. Metal-organic frameworks as media for the catalytic degradation of chemical warfare agents. (2017) doi:10.1016/j.ccr.2017.10.010.

17. Ballester, P. *et al.* Confinement Effects in Catalysis Using Well-Defined Materials and Cages. **6**, 623 (2018).

18. Mondloch, J. E. *et al.* Destruction of chemical warfare agents using metal–organic frameworks. *Nat. Mater. 2014 145* **14**, 512–516 (2015).

19. Doan, H. V *et al.* Controlled Formation of Hierarchical Metal−Organic Frameworks Using CO 2-Expanded Solvent Systems. **33**, 50 (2017).

20. Kim, J., Cho, H. Y. & Ahn, W. S. Synthesis and Adsorption/Catalytic Properties of the Metal Organic Framework CuBTC. *Catal. Surv. from Asia* **16**, 106–119 (2012).

21. Lin, K. S., Adhikari, A. K., Ku, C. N., Chiang, C. L. & Kuo, H. Synthesis and characterization of porous HKUST-1 metal organic frameworks for hydrogen storage. in *International Journal of Hydrogen Energy* vol. 37 13865–13871 (Pergamon, 2012).

22. Neufeld, M. J., Harding, J. L. & Reynolds, M. M. Immobilization of Metal–Organic Framework Copper(II) Benzene-1,3,5-tricarboxylate (CuBTC) onto Cotton Fabric as a Nitric Oxide Release Catalyst. *ACS Appl. Mater. Interfaces* **7**, 26742–26750 (2015).

23. Teo, H. W. B., Chakraborty, A. & Kayal, S. Evaluation of CH4 and CO2 adsorption on HKUST-1 and MIL-101(Cr) MOFs employing Monte Carlo simulation and comparison with experimental data. *Appl. Therm. Eng.* **110**, 891–900 (2017).

24. Katz, M. J. *et al.* A facile synthesis of UiO-66, UiO-67 and their derivatives. *Chem. Commun.* **49**, 9449–9451 (2013).

25. Zhao, J. *et al.* Ultra-Fast Degradation of Chemical Warfare Agents Using MOF–Nanofiber Kebabs. *Angew. Chemie Int. Ed.* **55**, 13224–13228 (2016).

26. Larsen, J. E. P., Lund, O. & Nielsen, M. Improved method for predicting linear B-cell epitopes. *Immunome Res.* **2**, 2 (2006).

27. Ponomarenko, J. *et al.* ElliPro: A new structure-based tool for the prediction of antibody epitopes. *BMC Bioinformatics* **9**, 1–8 (2008).

28. Saha, S. & Raghava, G. P. S. Prediction of continuous B-cell epitopes in an antigen using recurrent neural network. *Proteins* **65**, 40–8 (2006).

29. Ambrosetti, F., Jiménez-García, B., Roel-Touris, J. & Bonvin, A. M. J. J. Modeling Antibody-Antigen Complexes by Information-Driven Docking. *Structure* **28**, 119-129.e2

(2020).

30. Dominguez, C., Boelens, R. & Bonvin, A. M. J. J. HADDOCK: A protein-protein docking approach based on biochemical or biophysical information. *J. Am. Chem. Soc.* **125**, 1731–1737 (2003).

31. Chen, R., Li, L. & Weng, Z. ZDOCK: an initial-stage protein-docking algorithm. *Proteins* **52**, 80–7 (2003).

32. Cai, H. *et al.* Pretrainable Geometric Graph Neural Network for Antibody Affinity Maturation. (2021).

33. He, H. *et al.* De novo generation of antibody CDRH3 with a pre-trained generative large language model. (2023).

34. Jin, W. *et al.* DSMBind: SE(3) denoising score matching for unsupervised binding energy prediction and nanobody design. *bioRxiv* 1–24 (2023).

35. Jaszczyszyn, I. *et al.* Structural modeling of antibody variable regions using deep learning—progress and perspectives on drug discovery. *Front. Mol. Biosci.* **10**, 1–8 (2023).

36. Evans, R. *et al.* Protein complex prediction with AlphaFold-Multimer. *bioRxiv* 2021.10.04.463034 (2022) doi:10.1101/2021.10.04.463034.

37. Ofek, G. *et al.* Structure and Mechanistic Analysis of the Anti-Human Immunodeficiency Virus Type 1 Antibody 2F5 in Complex with Its gp41 Epitope. *J. Virol.* **78**, 10724–10737 (2004).

38. Ekiert, D. C. *et al.* Antibody recognition of a highly conserved influenza virus epitope :

implications for universal prevention and therapy. *Science (80-.).* **324**, 246–251 (2009).

39. Stanfield, R. L., Gorny, M. K., Williams, C., Zolla-Pazner, S. & Wilson, I. A. Structural Rationale for the Broad Neutralization of HIV-1 by Human Monoclonal Antibody 447-52D. *Structure* **12**, 193–204 (2004).

40. Zhou, T. *et al.* Structural definition of a conserved neutralization epitope on HIV-1 gp120. *Nature* **445**, 732–737 (2007).

41. Ko, J. & Lee, J. Can AlphaFold2 predict protein-peptide complex structures accurately? *bioRxiv* 2021.07.27.453972 (2021) doi:10.1101/2021.07.27.453972.

42. Tsaban, T. *et al.* Harnessing protein folding neural networks for peptide–protein docking. *Nat. Commun.* **13**, 1–12 (2022).

43. Ghani, U. *et al.* Improved Docking of Protein Models by a Combination of Alphafold2 and ClusPro. *bioRxiv* 2021.09.07.459290 (2022) doi:10.1101/2021.09.07.459290.

44. Johnson, G. & Wu, T. Te. Kabat Database and its applications: 30 years after the first variability plot. *Nucleic Acids Res.* **28**, 214–218 (2000).

45. Warr, G. W., Clem, L. W. & Söderhäll, K. The international imMunoGeneTics database IMGT. *Dev. Comp. Immunol.* **27**, 1 (2003).

46. Zhao, N. *et al.* A genetically encoded probe for imaging nascent and mature HA-tagged proteins in vivo. *Nat. Commun.* **10**, (2019).

47. Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. *Nat. Methods* **19**, 679–682 (2022).

48. Desta, I. T. *et al.* The ClusPro AbEMap web server for the prediction of antibody

epitopes. *Nat. Protoc.* **18**, (2023).

49. Zeng, Y. *et al.* Identifying B-cell epitopes using AlphaFold2 predicted structures and pretrained language model. *Bioinformatics* **39**, (2023).

50. Desta, I. T. *et al.* Mapping of antibody epitopes based on docking and homology modeling. *Proteins Struct. Funct. Bioinforma.* **91**, 171–182 (2023).

51. Ahdritz, G. *et al.* OpenFold: Retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization. *bioRxiv* 2022.11.20.517210 (2022).

52. Lee, J. H. *et al.* EquiFold: Protein Structure Prediction with a Novel Coarse-Grained Structure Representation. *bioRxiv* 2022.10.07.511322 (2023).

53. Ruffolo, J. A. & Gray, J. J. Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies. *Biophys. J.* **121**, 155a-156a (2022).

54. Abanades, B. *et al.* ImmuneBuilder: Deep-Learning models for predicting the structures of immune proteins. *Commun. Biol.* **6**, 575 (2023).

55. Ruffolo, J. A., Sulam, J. & Gray, J. J. Antibody structure prediction using interpretable deep learning. *Patterns* **3**, 100406 (2022).

56. Mariani, V., Biasini, M., Barbato, A. & Schwede, T. IDDT: A local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics* **29**, 2722–2728 (2013).

57. Terry, J. S. *et al.* Development of a SARS-CoV-2 nucleocapsid specific monoclonal antibody. *Virology* **558**, 28–37 (2021).

58. Interactions, P. A Single-Chain Antibody / Epitope System for Functional Analysis of.

Society 12729–12738 (2002).

59. Krauß, N. *et al.* The structure of the anti-c-myc antibody 9E10 Fab fragment/epitope peptide complex reveals a novel binding mode dominated by the heavy chain hypervariable loops. *Proteins Struct. Funct. Genet.* **73**, 552–565 (2008).

60. Sato, Y. *et al.* A Genetically Encoded Probe for Live-Cell Imaging of H4K20 Monomethylation. *J. Mol. Biol.* **428**, 3885–3902 (2016).

61. Fujiwara, K. *et al.* A single-chain antibody/epitope system for functional analysis of protein-protein interactions. *Biochemistry* **41**, 12729–38 (2002).

62. Kabsch, W. A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr. Sect. A* **34**, 827–828 (1978).

63. Lawrence, J., Bernal, J. & Witzgall, C. A purely algebraic justification of the Kabsch-Umeyama algorithm. *J. Res. Natl. Inst. Stand. Technol.* **124**, 1–6 (2019).

64. Churchill, M. E. A. *et al.* Crystal structure of a peptide complex of anti-influenza peptide antibody Fab 26/9: Comparison of two different antibodies bound to the same peptide antigen. *Journal of Molecular Biology* vol. 241 534–556 (1994).

65. Pace, C. N. & Scholtz, J. M. A helix propensity scale based on experimental studies of peptides and proteins. *Biophys. J.* **75**, 422–427 (1998).

66. Polonsky, K., Pupko, T. & Freund, N. T. Evaluation of the Ability of AlphaFold to Predict the Three-Dimensional Structures of Antibodies and Epitopes. *J. Immunol.* **211**, 1578–1588 (2023).

67. Guarra, F. & Colombo, G. Computational Methods in Immunology and Vaccinology:

Design and Development of Antibodies and Immunogens. *J. Chem. Theory Comput.* **19**, 5315–5333 (2023).

68. Giulini, M. *et al.* Towards the accurate modelling of antibody-antigen complexes from sequence using machine learning and information-driven docking. *bioRxiv* 2023.11.17.567543 (2023).

69. Hummer, A. M., Abanades, B. & Deane, C. M. Advances in computational structure-based antibody design. *Curr. Opin. Struct. Biol.* **74**, 102379 (2022).

70. Shashkova, T. I. *et al.* SEMA: Antigen B-cell conformational epitope prediction using deep transfer learning. *Front. Immunol.* **13**, 1–11 (2022).

71. Lo, Y. T. *et al.* Conformational epitope matching and prediction based on protein surface spiral features. *BMC Genomics* **22**, 1–16 (2021).

72. Mirdita, M., Steinegger, M., Breitwieser, F., Söding, J. & Levy Karin, E. Fast and sensitive taxonomic assignment to metagenomic contigs. *Bioinformatics* **37**, 3029–3031 (2021).

73. Kowalski, A. E. *et al.* Porous protein crystals as scaffolds for enzyme immobilization. *Biomater. Sci.* **7**, 1898–1904 (2019).

74. Wang, D., Stuart, J. D., Jones, A. A., Snow, C. D. & Kipper, M. J. Measuring interactions of DNA with nanoporous protein crystals by atomic force microscopy. *Nanoscale* **13**, 10871–10881 (2021).

75. Ward, A. R. & Snow, C. D. Porous crystals as scaffolds for structural biology. *Curr. Opin. Struct. Biol.* **60**, 85–92 (2020).

76. Dauparas, J. *et al.* Robust deep learning-based protein sequence design using ProteinMPNN. *Science* **378**, 49–56 (2022).

77. Mcbride, J. M., Polev, K., Reinharz, V., Grzybowski, B. A. & Tlusty, T. AlphaFold2 can predict single-mutation effects on structure and phenotype. doi:10.1101/2022.04.14.488301.

78. Krishna, R. *et al.* Generalized Biomolecular Modeling and Design with RoseTTAFold All-Atom. *bioRxiv* 2023.10.09.561603 (2023) doi:10.1101/2023.10.09.561603.

79. Anderson, A. C. The Process of Structure-Based Drug Design. *Chem. Biol.* **10**, 787–797 (2003).

80. Stewart, P. S. & Mueller-Dieckmann, J. Automation in biological crystallization. *Acta Crystallogr. Sect. F, Struct. Biol. Commun.* **70**, 686–696 (2014).

81. Marin, E. *et al.* Custom Design of a Humidifier Chamber for In Meso Crystallization. *Cryst. Growth Des.* **24**, 325–330 (2024).

82. Martin, L., Vernède, X. & Nicolet, Y. Methods to Screen for Radical SAM Enzyme Crystallization Conditions BT - Fe-S Proteins: Methods and Protocols. in (ed. Dos Santos, P. C.) 333–348 (Springer US, 2021). doi:10.1007/978-1-0716-1605-5_17.

83. Li, D., Boland, C., Walsh, K. & Caffrey, M. Use of a robot for high-throughput crystallization of membrane proteins in lipidic mesophases. *J. Vis. Exp.* e4000 (2012) doi:10.3791/4000.

84. Grosjean, H. *et al.* High-throughput crystallography for rapid fragment growth from crude arrays by low-cost robotics Harold. *chemRxiv* 1–20 (2023).

85. Stuart, J. D. *et al.* Scalable Combinatorial Assembly of Synthetic DNA for Tracking Applications. *International Journal of Molecular Sciences* vol. 24 (2023).

86. Wierenga, R. P., Golas, S. M., Ho, W., Coley, C. W. & Esvelt, K. M. PyLabRobot: An open-source, hardware-agnostic interface for liquid-handling robots and accessories. *Device* **1**, 100111 (2023).

87. Bryant Jr., J. A., Kellinger, M., Longmire, C., Miller, R. & Wright, R. C. AssemblyTron: flexible automation of DNA assembly with Opentrons OT-2 lab robots. *Synth. Biol.* **8**, ysac032 (2023).

88. Huber, T. R., Hartje, L. F., McPherson, E. C., Kowalski, A. E. & Snow, C. D. Programmed Assembly of Host–Guest Protein Crystals. *Small* **13**, 1602703 (2017).

89. Iwai, W. *et al.* Crystallization and evaluation of hen egg-white lysozyme crystals for protein pH titration in the crystalline state. *J. Synchrotron Radiat.* **15**, 312–315 (2008).

90. Hampton Research. Lysozyme User Guide. HR7-110 (page 1–2) https://hamptonresearch.com/uploads/support_materials/HR7-110_UG.pdf (2019).

91. Forsythe, E. L., H. Snell, E., Malone, C. C. & Pusey, M. L. Crystallization of chicken egg white lysozyme from assorted sulfate salts. *J. Cryst. Growth* **196**, 332–343 (1999).

92. Hartje, L. F. *et al.* Textile Functionalization by Porous Protein Crystal Conjugation and Guest Molecule Loading. *Crystals* vol. 13 (2023).

93. Stuart, J. D. *et al.* Mosquito tagging using DNA-barcoded nanoporous protein microcrystals. *PNAS Nexus* **1**, pgac190 (2022).

94. Kowalski, A. E. *et al.* Gold nanoparticle capture within protein crystal scaffolds.

Nanoscale **8**, 12693–12696 (2016).

95. Gupta, R. D. *et al.* Directed evolution of hydrolases for prevention of G-type nerve agent intoxication. *Nat. Chem. Biol.* **7**, 120–125 (2011).

96. Mukherjee, S. & Gupta, R. D. Organophosphorus Nerve Agents: Types, Toxicity, and Treatments. *J. Toxicol.* **2020**, 3007984 (2020).

97. Costa, L. G. Organophosphorus Compounds at 80: Some Old and New Issues. *Toxicol. Sci.* **162**, 24–35 (2018).

98. CDC. ToxFAQs for Nerve Agents (GA, GB, GD, VX). https://wwwn.cdc.gov/TSP/ToxFAQs/ToxFAQsDetails.aspx?faqid=524&toxid=93 (2014).

99. Peterson, G. W. & Wagner, G. W. Detoxification of chemical warfare agents by CuBTC. *J. Porous Mater.* **21**, 121–126 (2014).

100. Bigley, A. N. & Raushel, F. M. The evolution of phosphotriesterase for decontamination and detoxification of organophosphorus chemical warfare agents. *Chem. Biol. Interact.* **308**, 80–88 (2019).

101. Wagner, G. W., Peterson, G. W. & Mahle, J. J. Effect of Adsorbed Water and Surface Hydroxyls on the Hydrolysis of VX, GD, and HD on Titania Materials: The Development of Self-Decontaminating Paints. *Ind. Eng. Chem. Res.* **51**, 3598–3603 (2012).

102. Meng, Q. *et al.* Adsorption of organophosphates into microporous and mesoporous NaX zeolites and subsequent chemistry. *Environ. Sci. Technol.* **45**, 3000–3005 (2011).

103. Tuttle, R. R., Finke, R. G. & Reynolds, M. M. CuIILewis Acid, Proton-Coupled Electron

Transfer Mechanism for Cu-Metal-Organic Framework-Catalyzed NO Release from S-Nitrosoglutathione. *ACS Catal.* **12**, 8055–8068 (2022).

104. Hartje, L. F. *et al.* Characterizing the Cytocompatibility of Various Cross-Linking Chemistries for the Production of Biostable Large-Pore Protein Crystal Materials. *ACS Biomater. Sci. Eng.* **4**, 826–831 (2018).

105. Jones, A. A. & Snow, C. D. Porous Protein Crystals: Synthesis and Applications. *Chem. Commun.* (2024) doi:10.1039/D4CC00183D.

106. Melvin, A. C., Tuttle, R. R., Mohnike, M. & Reynolds, M. M. MOF Polymer Composites Exhibit Faster Nitric Oxide Catalysis than MOF Crystallites. *ACS Mater. Lett.* **4**, 2434–2439 (2022).

107. Roy, A. *et al.* Degradation of sulfur mustard and 2-chloroethyl ethyl sulfide on Cu–BTC metal organic framework. *Microporous Mesoporous Mater.* **162**, 207–212 (2012).

108. Chavan, S. *et al.* H2storage in isostructural UiO-67 and UiO-66 MOFs. *Phys. Chem. Chem. Phys.* **14**, 1614–1626 (2012).

109. Chen, R. *et al.* Ruthenium(II) Complex Incorporated UiO-67 Metal-Organic Framework Nanoparticles for Enhanced Two-Photon Fluorescence Imaging and Photodynamic Cancer Therapy. *ACS Appl. Mater. Interfaces* **9**, 5699–5708 (2017).

110. Li, Y. A. *et al.* Nanoscale UiO-MOF-based luminescent sensors for highly selective detection of cysteine and glutathione and their application in bioimaging. *Chem. Commun.* **51**, 17672–17675 (2015).

111. Wang, T. *et al.* Bottom-up Formation of Carbon-Based Structures with Multilevel

Hierarchy from MOF–Guest Polyhedra. *J. Am. Chem. Soc.* **140**, 6130–6136 (2018).

112. Kaur, G. *et al.* Controlling the Synthesis of Metal-Organic Framework UiO-67 by Tuning Its Kinetic Driving Force. *Cryst. Growth Des.* **19**, 4246–4251 (2019).

113. Kaufmann, K. W., Lemmon, G. H., Deluca, S. L., Sheehan, J. H. & Meiler, J. Practically useful: what the Rosetta protein modeling suite can do for you. *Biochemistry* **49**, 2987–2998 (2010).

114. Sormani, G., Harteveld, Z., Rosset, S., Correia, B. & Laio, A. A Rosetta-based protein design protocol converging to natural sequences. *J. Chem. Phys.* **154**, (2021).

115. Mariani, V., Biasini, M., Barbato, A. & Schwede, T. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics* **29**, 2722–2728 (2013).

116. Sprague, B. L. & McNally, J. G. FRAP analysis of binding: proper and fitting. *Trends Cell Biol.* **15**, 84–91 (2005).

117. Plaxco, K. W., Simons, K. T. & Baker, D. Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.* **277**, 985–994 (1998).

118. Dauparas, J. *et al.* Atomic context-conditioned protein sequence design using LigandMPNN. *bioRxiv* 2023.12.22.573103 (2023) doi:10.1101/2023.12.22.573103.

119. Cuchiaro, J., Deroo, J., Thai, J. & Reynolds, M. M. Evaluation of the Adsorption-Accessible Surface Area of MIL-53(Al) using Cannabinoids in a Closed System. *ACS Appl. Mater. Interfaces* **14**, 12836–12844 (2022).

120. Thai, J. E., Tuttle, R. R., DeRoo, J., Cuchiaro, J. & Reynolds, M. M. Cu-Based Metal–

Organic Framework Nanosheets Synthesized via a Three-Layer Bottom-Up Method for the Catalytic Conversion of S-Nitrosoglutathione to Nitric Oxide. *ACS Appl. Nano Mater.* **5**, 486–496 (2022).