Intro. I started at Caltech planning to study bioengineering and before classes had started, joined a wet-bench research project trying to design HIV/flu-neutralizing antibodies. I was quickly struck by how complex the systems and data were, and moreover, how difficult it was to develop robust intuition for them. This sentiment extends to many other scientific disciplines, and I found myself craving a principled way to approach seemingly opaque problems. Machine learning seemed to offer this, so within a few months, I'd pivoted to computer science and joined a research group where I began work to use Bayesian optimization and deep learning for adaptive experiment design. These paradigms proved effective for approaching real-world rational design problems, including designing COVID antibodies, where traditional scientific methods struggle and have each profoundly shifted the way I think about scientific problems since. What excites me most is the opportunity to reframe how we approach scientific questions through the lens of machine learning and computer science, and hence vice versa. This careful meshing of disciplines is key to developing both newly effective scientific solutions as well as interesting machine learning methods, and I look forward to continuing work at this intersection as a graduate student.

Intellectual Merit. My first machine learning (ML) research experience was in Yisong Yue's group (Caltech), where I worked closely with biology domain experts to reframe how they performed protein engineering. I sought out this project as an extension of my experience doing experiment design in biology labs: here, the challenge was designing a model that could effectively represent high-dimensional protein sequence data to learn the complex relationship between an antibody sequence and its ability to bind a coronavirus particle. I spent several weeks reviewing online content on ML, linear algebra, Bayesian optimization (BO), and neural networks because I hadn't any formal education on them yet. Once I'd gotten up to speed, I reviewed the literature and decided with my professor that integrating deep learning with BO would be an interesting and novel approach. I spent my freshman summer designing and coding a BO framework from scratch to support the use of a so-called deep kernel: instead of a state-of-the-art Gaussian process (GP) model, my code first passed input protein representations through a deep neural network (DNN) and then to a GP regressor, which I trained together by backpropagating gradients sequentially through both models. This way, my BO framework had the advantage of arbitrary learning capacity because DNNs are universal function approximators and can often extract useful embeddings and relationships from noisy, lossy input data. It kept the strength of standard GP models too though, which give posterior uncertainty measures, as uncertainty quantification is crucial to balancing exploration and exploitation in adaptive experiment design settings. I then applied my code to various synthetic optimization objectives and presented its promising performance to my PI, who asked me to join a collaboration with LLNL and UChicago seeking to develop an ML solution to protein engineering.

Prior to my involvement, our LLNL collaborators would run costly simulation campaigns on thousands of mutant antibodies and send the best batch (often fewer than 50) for more expensive lab testing. They were aware of BO, but available GP kernels are known to struggle with complex functions, high-dimensional data, and scaling to the amount of data LLNL had. These are all problems DNNs excel at, so I began running experiments using deep kernel BO on their protein datasets and comparing with baselines from the literature. I found that the DNN often overfit the noisy training data, so I implemented a form of Bayesian regularization called Monte Carlo Dropout to better account for this uncertainty. I also performed surveys over network architectures and acquisition functions to better understand how adding a DNN affected the BO procedure. These led me to realize that using Thompson sampling tended to give better results, probably because like Monte Carlo Dropout, its Bayesian sampling component helps counter the overconfidence observed across many deep learning models. For several months during my sophomore year, I presented my progress to our collaborators biweekly and deliberated to better tailor my method to their domain. The deep kernel model improved upon common GP models and was able to discover near-optimal antibody designs in as few as 300 trials (versus many thousands). This led to the submission of a first-author paper at the end of my sophomore year, which I'm reworking for submission in January. Based on the domain intuition that most mutant proteins are bad, I helped formulate an extension of deep kernel BO which maintains a region of interest on the search space to filter out poor candidates that are unlikely to provide info about the optimum. During this time, I began collaborating closely with a second-year graduate student from UChicago, so I spent time cleaning up and commenting my code to make sure he could understand and build on top of it. We devised a way to utilize untested candidate sequences by pre-training an autoencoder on all possible sequences. Thus, the DNN learns relationships across the whole search domain prior to seeing training data, so is less likely to overfit to the few labeled training points it receives. This improved performance and was only possible due to the deep kernel model (normal GPs aren't conducive to representation learning). This work was published as a co-author workshop paper at ICML 2022 and has just been submitted as a full conference paper to AISTATS 2023. We've begun to run live antibody optimization campaigns on LLNL's servers using our method, and I've been corresponding with a Cornell graduate student who is optimizing parameters of a particle accelerator with my code. I'm continuing to extend the deep kernel BO framework to: proposing diverse batches of candidates with Bayesian sampling methods, creating a graph neural network kernel for molecule graphs with positional data, multi-fidelity BO for when access to multiple experiment types is available, and transfer learning to use data from previous campaigns when a new, but similar virus arises. The deep kernel BO framework allows myriad new ways to think about many of these problems and I find this transformative potential for protein engineering and other domains very promising.

I've since branched out to explore other ways of combining ML with science. During my junior year, I worked in Anima Anandkumar's group (Caltech) to train diffusion models more efficiently by reframing very high-dimensional data (e.g., 200k-D) with the fundamental hierarchical property of signals. Under this paradigm, images are Fourier transformed and lower frequencies are viewed as "global" info and higher frequencies as high-resolution signal. We then train a model for the target resolution by sequentially conditioning on lower-frequency models, which should be cheaper overall. I also identified that default convolutional neural networks used for diffusion models were overfit to image data and hence needlessly expensive on frequency data, which allowed us to cut diffusion sampling time from ~14 hours to ~2 hours with simpler architectures. I also joined Katie Bouman's group (Caltech) on an explainability project, where we're exaggerating parts of an image that a classifier bases its predictions on. This is inspired by medicine, where a doctor needs to understand the basis for a classifier's diagnosis before making any decisions. I analyzed various GAN models in the literature to inform construction of our model and introduced a chest x-ray dataset as a medium-difficulty test case for our experiments.

This past summer, I worked with Ryan Adams' group (Princeton) to reframe mechanical structure optimization based on advances in deep learning and automatic differentiation. Domain knowledge says that 1) points must have binary mass values and 2) fabrication requires smooth topologies. I wrote code that parametrizes a topology as the level-set of a DNN and derived gradients to optimize the surface. In contrast to the standard parametrization as a pixel grid, our approach meets both domain criteria naturally: the level-set ensures binary mass values and allows us to reason about smoothness in a functional manner. I am implementing this with a graduate student and will formulate priors for smoothness and symmetries such that results are consistent with scientist expectations; we are aiming to submit a paper next year and one of my experiments is being included in a co-author paper we are preparing for ICML 2023.

During this time, my curiosity quickly outpaced the content provided by Caltech's undergraduate CS program. I took several applied math courses, participated in special topics classes in neural network theory, control, and representation learning in which I read papers and completed independent research projects, and joined graduate student reading groups on variational inference and uncertainty quantification, where I engaged in weekly discussions and even presented recaps/led discussions for a few papers. I started a slide deck of key points from each paper (and those I was reading for my research), which reached close to 50 slides across my junior year. I've continued to augment my biology intuition with classes, enjoyed learning mechanics this past summer, and plan to take advanced science classes in graduate school and collaborating with more researchers in natural science and engineering fields.

Broader Impacts. Sharing concepts that excite me is one of my favorite activities, which is why I began teaching as soon as I could: starting my sophomore year, I served as a TA for the entire CS intro course sequence. As someone who had come from a non-computational science background but came around thanks to the insight of the CS TAs my freshman year, I wanted to pay it forward by showing students the broad usefulness of CS regardless of scientific discipline. To have the most impact, I stepped up and held my office hours at the busiest times, often staying 2 hours beyond the end to make sure every student got helped. I helped make important decisions regarding course organization and took lead on developing some of the programming assignments. As a result, I was selected as Head TA for our CS2 my junior year where I continued the above while mentoring and supporting 18 other TAs. In this position, I was trusted with more agency to improve the course, and worked with another TA to conceptualize and implement bridge group, a DEI initiative to allow students from disadvantaged backgrounds to self-select into a special peer group with extra TA support. I helped the other TA develop teaching material for the group each week and oversaw the program to ensure it ran smoothly. The students in bridge group did really well and showed increased confidence in their programming skills; we are planning to continue this in perpetuity and published our work at RESPECT 2022, a CS education conference focused on DEI. I also served as TA this summer for a pre-Caltech intro CS course offered to first year students from disadvantaged backgrounds, where I held daily office hours, mentored students on capstone projects, and helped develop engaging CS + STEM assignments, two of which we are preparing for submission to SIGCSE 2023. In addition, I began TAing Caltech's machine learning track during my junior year, which I've enjoyed a lot. I'm looking forward to TAing more advanced ML courses as a graduate student and teaching people about the more cutting-edge ML concepts that excite me. NSF GRFP will help me be well-positioned to continue passing along my CS/ML knowledge to the next generation of scientists.

As a result of my DEI efforts as a TA, I was invited to serve on Caltech's CS department's DEI committee this year, where I'll help push multiple aspects of the department to be more inclusive and equitable. At our first meeting of the year, I brought attention to how the difficulty associated with finding research in our CS department as an undergraduate affects less privileged students disproportionately and suggested an open market matching system where graduate students and postdocs can post projects for anyone to apply to; faculty, graduate students, and undergraduates on the committee liked the idea so it's currently being developed. I also served on Caltech's Academics & Research Committee, where I advocated for student interests and helped bridge the communication gap between faculty and undergraduates. When we received concerns that a new instructor had set expectations too high for the introductory probability class that all CS sophomores take, I helped explain the disconnect to him respectfully such that he reorganized his curriculum for the remainder of the term. I plan to use my communication and collaboration skills not only for research purposes, but also to make the CS community a more accessible place for everyone interested.

Outside of publishing/presenting research and teaching classes, there's a lot of less formal communication that can play a huge role in education. For instance, much of my machine learning education came from online resources, since a lot of the content I wanted to learn was not taught in the ML courses offered at Caltech. I've also noticed that there's not as much crosstalk between CS and other fields as I'd like there to be. I understand that it's not enough to just publish within ML conferences during graduate school and am interested in disseminating knowledge outside of ML/CS silos by blog posting and sharing content via Twitter or other social media platforms where there are strong, diverse research communities. I plan to utilize my experience communicating with broader audiences; each summer, I've presented my research to students and faculty from other disciplines as well as non-technical backgrounds. I've also learned an enormous amount from the fantastic graduate student and faculty mentors I've had thus far, and always do my best to pass this along as well. Both through and outside of my teaching roles, I've found myself mentoring tens of students with respect to CS, ML, navigating undergraduate research, internship advice, and so on. I served as an Ambassador for Caltech's summer research program twice, where I advised students on the research process itself as well as effective writing and presentation. In particular, I can't wait to take on undergraduate researchers as a graduate student myself and provide immersive mentorship to those who will come after me.

Future. I'm eager to work full-time on projects at the intersection of ML and science in graduate school and take a more active role in the scientific community and its posterity. I suspect that my future beyond graduate school lies in some collaborative combination of academia and industry—I want to stay close to both the difficult scientific problems society faces today and the interesting computational and mathematical thinking that is helping us reframe how we approach these questions. With the support of the GRFP, I plan to unite these components effectively to chip away at the many challenges of our time.

Intellectual Merit – Motivation and Background. One of the central challenges in mechanical engineering is that of designing a structure or material with certain characteristics. For centuries, new mechanical properties have been captured by altering the composition of the material itself. Such methods require significant physical fabrication and experimentation (and thus time and money) because they are difficult to control precisely. The ability to specify the desired properties of a structure given constraints and then quickly find a satisfying topology is considered the "holy grail" of mechanical structure design. Particularly, the behavior of a structure under variable amounts of force or deformation is captured by its strain energy curve. For a certain structure size, material, etc. we would like to be able to choose an arbitrary target strain energy curve and find a corresponding structure topology. The strain energy curve allows us to describe *multi-stability*, a highly sought-after property for many tasks [1].

Recently, two key advances have drastically increased capabilities: 1) computational power and methods now allow us to accurately simulate mechanical behavior *and* explicitly frame design problems as topology optimization problems that are computable on reasonable time scales [2], and 2) cellular metamaterials (as in Figure 1), a rich class of structures capable of implementing complex nonlinear physical "functions" to yield desired output behaviors given input actuations [3]. Importantly, the advent of metamaterials makes it feasible to significantly alter mechanical behavior of a structure via its topology instead of needing to change its material composition. Metamaterials draw their flexibility from the large space of substructure interactions between cells, which makes optimizing macroscopic behavior very challenging [4]. Although there has been significant work looking to utilize modern computational methods like deep learning for mechanical structure optimization (e.g., [5], [6], [7], [8]), these methods largely consist of *applying* machine learning (ML) methods in the way of naïve prediction tasks. It remains to be seen what kinds of complex structures and mechanical functions can be discovered by rebuilding topology optimization from the ground up with modern ML and computational techniques.

Approach. In this project, we aim at thoughtful integration of machine learning with mechanical structure optimization, particularly of cellular metamaterials, to allow for complete specification of material properties. We would like to discover complex multi-stable designs matching arbitrary strain energy curves. To do so, we reframe several aspects of the topology optimization pipeline.

As input to the optimization algorithm, a user will provide a desired strain energy curve. At each of *n* stable configurations (i.e., local minima of strain energy curve), the user can provide a desired shape of the structure (e.g., want a structure that is circular at rest but creates a polygon of certain side lengths when 5 Newtons are applied along the y-axis). This draws upon recently submitted work from my graduate student collaborator in Ryan Adams' group at Princeton called a *neuromechanical autoencoder* (NMA): under this paradigm, a neural network encoder learns to output actuations that yield the input shape together with a neural network decoder learning to output corresponding cell shapes given a base metamaterial ([9]; see Figure 1). NMA then uses of a fully-differentiable variational material integrator (from Adams group) to simulate deformation of the structure based on the actuations.

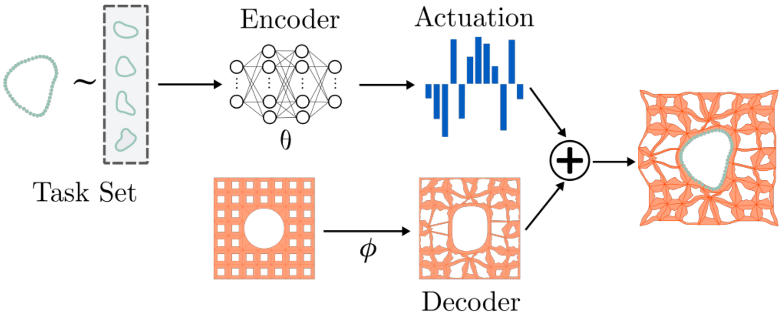

Figure 1: Neuromechanical autoencoder used for metamaterial design, from [9].

I will implement two main modifications to the NMA framework. Instead of a simple shape-matching objective (current), I will optimize *closeness to the target strain energy curve*. To do this, I will modify our differentiable simulator code to compute strain energy at intervals along the deformation, yielding a simulated energy curve for a given topology and set of actuations. I will then test various function similarity metrics such as L1, L2 distances between simulated and target strain energy curves. However, we also would like to be able to specify what shapes the metamaterial takes on at each of its stable minima. I propose adding a weighted version of the shape matching objective used in [9] for *each* stable configuration. When the shape is not fixed at a certain stable configuration, its weights can be set to 0. A user might also like to specify the actuation magnitudes that yield each shape, in which case an input could be used instead of the encoder prediction. These actuations will likely be incompatible with a fixed target strain energy curve, for which I propose a relative target strain energy curve with adjustable learned additive and multiplicative factors on the target function. The loss calculation would minimize distance w.r.t. these two factors such that the resultant strain energy curve is similar in behavior to the target but may be scaled/translated w.r.t. the deformation applied, such that target actuations can be accommodated. This would also be useful in cases where the user specifies only the desired shape of the target strain energy curve. These changes mainly deal with the top half of Figure 1.

I propose to modify the bottom half of Figure 1 by using implicit surface topology optimization instead of the current shape optimization over cells of the metamaterial. I spent this past summer developing a new, fully-differentiable topology optimization framework in Ryan Adams' group that parametrizes the design as an implicit surface instead of discrete pixels or voxels as in [2]. I've calculated derivatives using the Implicit Function Theorem such that gradients are propagated from the objective all the way to the implicit surface parameters, here a neural network [10]. This allows a more flexible basis for the topology and frees us of a base metamaterial structure and decoder as seen in Figure 1, which [9] relies on because it can only perform shape optimization over the cells. This means the topology can be completely rearranged from initial grid structure such that we optimize over a larger space containing more complex designs. I am still developing this component and testing its empirical capabilities, but results are promising so far. This functional parametrization allows for explicit smoothness and symmetry priors on the resultant topology (e.g., via covariance parameters of the Gaussian process in a deep kernel [11]); the use of such priors still needs to be further explored.

I will then combine all aforementioned components into an end-to-end topology optimization pipeline capable of specifying target strain energies, target structure shapes plus their actuations, and expectations on design smoothness or symmetry. Longer term goals include extending to 3D (currently only 2D), which will likely require code optimization. All code is written in JAX and I plan to improve its runtime by using just-in-time compilation and vector-Jacobian product methods. I am currently working on using differentiable fiber sampling instead of quadrature to compute expensive integrals faster.

Broader Impacts. Being able to specify aspects of a structure's behavior more fully would revolutionize mechanical engineering by allowing the design process to be largely abstracted out. When parts with complex specifications can be quickly designed by an optimization algorithm, engineering workflows can be that much faster. Our approach promises to make possible new kinds of metamaterials and mechanical properties by more completely utilizing the flexibility of the metamaterial class through co-design of the topology and the actuations required to create desired configurations. This approach is very modular so will be easily adaptable and extensible to new domain requirements. I plan to collaborate with mechanical engineers in academic and industry settings to ensure our tool is accessible to these target audiences.

[1] Jin et al. Guided transition waves in multistable mechanical metamaterials. *PNAS*, 2020. [2] Andreassen et al. Efficient topology optimization… *Struct Multidisc Optim,* 2011. [3] Surjadi et al. Mechanical Metamaterials… *Advanced Engineering Materials,* 2019. [4] Beatson et al. Learning Composable Energy Surrogates… *arXiv:2005.06549v2,* 2020. [5] Kollmann et al. Deep learning for topology optimization of 2D metamaterials. *Materials & Design,* 2020. [6] Qiu et al. A deep learning approach for efficient topology… *Materials & Design 212,* 2021. [7] Xue et al. A data-driven computational scheme for the nonlinear mechanical… *Soft Matter,* 2020. [8] Yu et al. Deep learning for topology optimization design. *arXiv:1801.05463,* 2018. [9] Anonymous. Neuromechanical Autoencoders… *Under review for ICLR*, 2022. https://openreview.net/forum?id=QubsmJT_A0. [10] Kolter et al. Deep Implicit Layers. *NeurIPS*, 2020. [11] Wilson et al. Deep Kernel Learning. *PMLR*, 2016.

Personal, Background, and Future Goals Statement

Introduction: As a student at both Carnegie Mellon and Caltech, I have begun to view astrophysics as a melting pot of science and engineering, with unanswered questions on neutron stars, black holes, and spacetime, and opportunities to teach the next generation about the joys of STEM research. A PhD focused on radio astronomy was not always the goal on the horizon, but has now become the focus of my career. Throughout my journey to this point, I have in fact shifted my focus among various fields from mechanical engineering to robotics to electrical engineering to physics. Through this, I have built up a repertoire of skills applicable to interdisciplinary fields like radio astronomy and systems engineering, while also finding motivation in lifelong learning and sharing what I have learned with a wider community. This has solidified my desire to pursue academic research following my graduate work, so that I can continue to explore interdisciplinary learning and application, facilitate engagement in science and engineering, and dive into the unsolved problems in radio astronomy and pulsar science.

Experience and Intellectual Merit: While I graduated from Carnegie Mellon University (CMU) with a degree in Electrical & Computer Engineering (ECE) and Physics, this was not before exploring other fields that would later inform my intersectional research. I began as a mechanical engineering student, with intent to double major in robotics, hoping to explore novel sensing methods for unfamiliar environments. With the CMU Field Robotics Center (FRC), my work on the Lunar CubeRover included simulating and carrying out 'drop tests' on a prototype 2 kg lunar rover, while on the RadPiper project I tested a collimating radiation sensor. I also worked with another graduate student to develop a mathematical model to estimate the distribution of surrounding radiative material. On these projects I found I was most interested in the electronic details of the radiation detector and processing methods for sensory input; so, in an effort to learn more, I shifted my primary major to Electrical & Computer Engineering following my first year. To complement this, I focused on the Orbital Edge Computing project, whose goal was to use weather satellite images in a machine learning pipeline to identify natural disasters. I used a Software Defined Radio (SDR) module to read images transmitted from weather satellites, as well as on basic image transformation algorithms in C++. In addition to understanding the interdisciplinary nature of research, I also renewed my interest in space based projects, supplemented by my coursework in modern physics. This led me to change my second major to physics to learn more about engineering applications in astronomy. I am grateful today for the technical engineering groundwork laid by these early projects; I still utilize the skills I learned in radiation detection, radio and analog signal processing, modeling, and simulation in my radio astronomy work. Furthermore, in a short time, I not only experienced fast-paced research environments and gained machine shop and programming practice, but also identified a niche in STEM that matched my interests.

Since focusing on astrophysics instrumentation, I have incorporated methods from mechanical engineering, electrical engineering and physics to make unique contributions to projects in the field. During a summer Research Experience for Undergraduates (REU) at the University of Alabama (UAH), I contributed to the Terrestrial RaY Analysis and Detection (TRYAD) cubesat focused on the detection and characterization of flashes of gamma rays associated with lightning. I tested the detection system with radioactive sources, machined screws for the prototype, and re-designed the Data Acquisition Board (DAQ) Printed Circuit Board (PCB) layout, summarizing the project in the end-of-program poster presentation.. On my return to school, I continued with instrumentation under Professor Jeffrey Peterson on the High-Z monopole antenna project. The project goal was to develop novel radio antennas that could conduct continuum surveys to search for the Hydrogen 21-cm emission line associated with the first stars' formation. On this, I developed simulations using Altair Feko and Python to model antenna patterns and the expected all-sky temperature map output; this was a great opportunity to dive deep into a subfield of astronomy that merged engineering with theory. Indeed, I developed a working understanding of radio surveys, astronomical sources of radio signals, and the expected data products of continuum surveys that created a foundation for my work in radio interferometry. I chose to attend Caltech because it would allow me to continue to explore astronomy from an interdisciplinary point of view, utilizing past experiences while learning and applying concepts from adjacent disciplines.

My current work with Professor Vikram Ravi is related to the Deep Synoptic Array-110 (DSA-110) radio array, and the follow up DSA-2000. On the DSA-110, I am developing a polarization analysis pipeline for Fast Radio Bursts (FRBs) detected during the array's commissioning phase. Familiarizing myself with the details of interferometry and polarization, I have not only learned about the software infrastructure needed for astronomy, but also contributed to it, producing Python code to estimate polarization and rotation measure for FRBs. Through this process, I have also immersed myself in the scientific analysis and outcomes, and intend to soon publish one of the largest sample of full-polarization data for FRBs to date. The NSF GRFP would support me in building on this endeavor with a survey of magnetars, thought to be related to FRBs, using the DSA-110. For the DSA-2000, I have conducted preliminary analysis into the computation requirements and predicted pulsar detections that imply it could significantly increase the known pulsar population. Drawing on my background in radio simulation, I created Monte Carlo simulations and pulsar searching models for the DSA-2000 to estimate a more practical signal-to-noise threshold. Applying this in conjunction with the *PSRPop.py* pulsar simulation software allowed me to predict detection numbers for different search strategies. This is a great opportunity to make a tangible impact on the design of the DSA-2000 in the early stages, and I also intend to publish a paper discussing these results. Radio astronomy's interdisciplinary nature provides an avenue for me to branch out into a more scientific role while still contributing to more technical hardware and software work. It is this aspect that attracted my attention, and I look forward to continuing my PhD journey in the field.

Following graduation I plan to continue exploring instrumentation for radio astronomy, but also explore adjacent fields like planetary spectroscopy and even space-based system design through both industry and academia. Specifically, I plan to explore industry as a research scientist to first become more well-versed in hands-on science applications and engineering. Much of the engineering involved in radio instrumentation has been applied to other fields, such as radar sensing and communications engineering; furthermore the science underlying astronomical sources, including plasma physics and quantum mechanics, are applied in fields like in space propulsion, integrated circuit design, and quantum electronics. Gaining experience in these topics from outside academia will offer a flexibility in thinking invaluable to reaching my ultimate goal of being a professor. As an academic, I hope to conduct further research to develop instrumentation for radio astronomy and adjacent fields, contribute to emission theories behind radio sources, and share my knowledge with the next generation of science and engineering students.

Broader Impacts: Complementary to my academic pursuits in my undergraduate career were my efforts to engage with the broader Pittsburgh community, both through science and engineering, and by helping improve the quality of life of those in need. As a freshman, I joined the National Society of Black Engineers (NSBE) as the Community Service Chair, with the goal of facilitating involvement of CMU's Black scientists and engineers in giving back to the community. In this effort, I organized a successful donations drive, the Donation Sensation, in which we collected food, clothing, toys, and books to distribute to various charities in the area, including the Pittsburgh Childrens' Hospital, the Brashear Association, and Jubilee Soup Kitchen. Our work garnered the attention of SPIRIT, CMU's Black Student Union, and I joined as Community Service Chair during my sophomore year. In this role I organized another drive that collected over 500 assorted school supplies for the Education Partnership charity that helped underprivileged children attend school prepared. Additionally, I encouraged student engagement in the community by inviting representatives from Pittsburgh charity groups to SPIRIT meetings to advertise volunteer opportunities. One of the most effective ways for students to give back is through community tutoring programs; with both SPIRIT and NSBE I facilitated our groups' participation in the 100 Black Men and CHOICE tutoring programs, which provide homework help and academic mentorship to Black middle and high school students. I personally tutored and mentored students in middle and high school math and science, and also provided help with college essay writing. These opportunities successfully encouraged retention of young students and built meaningful relationships for them to learn about higher education from undergraduate and graduate students.

In addition to broad community service programs, I also focused on science and engineering outreach during my undergraduate career to facilitate engagement with STEM fields from young students, especially those from traditionally underrepresented groups. As a Black student in these fields I have found it essential to offer support and mentorship to the younger generation to help them navigate the complex cultural landscape of academia, and feel motivated to pursue science as a career. During my second year I was also the Pre-College Initiatives (PCI) Chair of NSBE, through which I mainly facilitated the NSBE Jr. STEM engagement and mentorship program for Black middle school students. In this role I developed monthly labs that would teach students about a new area in STEM, such as circuits, programming, and electricity and magnetism. Further, I helped organize NSBE's A Walk for Education

(AWFE) which invited nearly 100 elementary, middle, and high schoolers to CMU for engineering workshops, college readiness presentations, and tours of the campus. We also invited admissions officers from CMU and other colleges to help high schoolers with graduate applications, and facilitate interest in higher education.

During that year, I expanded my STEM engagement efforts by working with the ECE department's educational outreach group, ECE Outreach, as a Lab Development Chair; our group organized Mobile Labs in which ECE students would go to Oakland Catholic High School to teach simple programming and circuitry using Arduinos to students, as well as the SPARK Saturday program that invited middle and high school students to CMU's campus for similar labs. We also collaborated with other campus groups like the Society of Women Engineers (SWE) for weekend lab programs for young women interested in STEM. I played a significant role in designing our Arduino and circuitry labs for the students in each case. While the COVID-19 pandemic interrupted our in-person programs, we continued to create labs for high school students to engage with, particularly an Intro to Python programming Zoom course for which I developed problem sets and helped lead interactive lectures. As an advocate of lifelong learning, I did not limit my STEM outreach to younger students, but as a Junior also helped undergraduates as an EXCEL instructor, for which I led supplemental instruction sessions for a small group of students taking Physics II: Electricity and Magnetism. Additionally I worked as a TEAM staff member at CMU's TechSpark Makerspace to help students with machining, rapid prototyping like 3D printing and laser cutting, and with general engineering projects. Both of these allowed me to share my knowledge with fellow classmates, as well as learn more about their projects and academic experience to better inform my teaching and outreach efforts.

Here at Caltech I intend to continue with outreach efforts through campus groups. I recently became a tutor with the Caltech Y RISE Program, which offers personalized tutoring and academic help for middle and high school students. As with my previous tutoring experiences, I intend to not only be supportive academically, but also as a mentor and advocate. My position in higher education allows me to offer unique insight on what to expect from college, possible career paths, and research in STEM to help students along their educational journey. I also hope to help expand this program to reach more underrepresented groups in the greater Los Angeles community. In addition, I will participate in the Caltech Science for March event, in which Caltech groups organize demos and presentations to share science and engineering topics with the Pasadena community. This would be a great avenue to introduce radio astronomy to the world in a more digestible manner. While it is often difficult to demonstrate radio astronomy due to the infrastructure required, simple demonstrations with dipole antennas could be engaging and show the underlying concepts that drive the Event Horizon Telescope (EHT), the radio camera responsible for the first images of black holes. As part of Professor Vikram Ravi's group, I will also contribute to undergraduate classes focused on building similar demos, and can utilize my experiences and resources from that to inform the Caltech Science for March presentation.

Finally, in order to share my research with a larger community, as well as my cultural experiences in STEM fields, I intend to write for Caltech Letters. Caltech Letters publishes online articles on science and culture for the wider community with non-scientific backgrounds; after gaining experience writing with the group I will join as an editor to help create an avenue for others to share their work and experiences as well. A few possible topics for articles that could be published with Caltech Letters include an overview of pulsar astronomy and what makes neutron stars so interesting, a description of radio astronomy and how observing radio sources differs from optical astronomy, and a discussion of my journey as a Black student in college and how it has affected my experience. With the support of the NSF GRFP and throughout my career, I will continue to give back to the community, facilitate participation in STEM from underrepresented groups, and actively share my knowledge as an advocate for lifelong learning.

Graduate Research Plan Statement

Background: Magnetars are a species of neutron star that exhibit larger surface magnetic fields (~10^{14} G), longer pulse periods (2-12 s), and higher spin-down rates (~10^{-11}s/s) than radio pulsars[1,2]. Only 30 magnetars have been detected, and only 6 have observed transient radio emission[3]. Simulations predict a galactic population of ~10^{3-5} Ultra Long Period Magnetars (ULPM) with periods larger than 12 s; however, few have been detected, implying gaps in our understanding of magnetars' birth properties, evolution, and relation to radio pulsars[22]. Two candidates are PSR J0901-4046 (PSRJ0901) and GLEAM-X J162759.5-523504.3 (GLEAM-X) which have periods 76 s and 18.18 s respectively; while these have periods much longer than radio pulsars (< 10s), they do not exhibit the flat radio spectra, low luminosity, or high variability seen in transient magnetar emission[4,5]. Beniamini, et al. recently proposed three more candidates, identifying a distinct population forming around volumetric birth rates of ~10^5 per year per Gpc^3. High energies (~10^{40} erg) and large beaming fractions (~1%) disfavor white dwarf models, further separating ULPMs[22]. High time resolution surveys which generally target radio pulsars and unexplained, extragalactic Fast Radio Bursts (FRBs), and are insensitive to long periods (0.1 - 1s) and low flux (< 1mJy) that may relate to ULPMs[6,7]. The emerging relation of magnetars to FRBs could dramatically increase the observed sample; some theories suggest both magnetar quasi-periodic oscillations (QPOs) and trains of FRBs originate from crustal oscillations coupled to plasma core Alfvén modes[8,22].

When modeling the magnetar formation process, the small sample, broad parameter space, and many evolutionary models often render degenerate solutions. The main models are the fossil field hypothesis, which predicts magnetars evolve from the high magnetic field tail of initial radio pulsar distributions, and unique field evolution theories which invoke a dynamo process in the superfluid core that create a distinct field structure[2]. Forward modeling experiments estimate birth distributions by evolving magnetars to present day, then iteratively updating initial parameters until the estimated detections match the observed population. Experiments by Jawor, et al., Dirson, et al., and Beniamini, et al. not only find degeneracy in the mean birth period (1-2000ms) of magnetars, but also predict either exponential or super-exponential field decays, and steeper, super-exponential 100 year - 0.01 Myr timescales for ULPMs, in contrast to radio pulsars' ~0.46 Myr exponential decay[9,10,11,22]. These neglect complex field geometries expected in magnetars, and magnetohydrodynamic (MHD) simulations further posit magnetars have typical initial periods (10-100 ms), but spin-down quickly during initial cooling[12,13].

Despite that ~100 times more radio pulsars have been detected than magnetars, birth rate estimates imply the latter compose ~10% of core collapse supernova (CCSN) remnants. This makes it essential to understand the evolution and age predictions of magnetars; both rapid spin-down and decay imply the characteristic age is overestimated, and consequently that the birth rate is higher. Intermittent spin-down is often observed after magnetar giant flares or glitches, possibly leading to exponential period increase and shorter ages[22]. Contrarily, 'propeller' mass ejection theories predict fallback accretion could cause intermittent spin-up, extending magnetar ages[17]. In support of this, the predicted neutron star birth rate exceeds that of CCSN, ~1.9 per century[18]. While initial forward modeling used the ROSAT satellite's selection effects to predict a 0.15-0.3 per century birth rate, Jawor's recent simulations, using alternative decay models and a larger magnetar sample, but neglecting observational effects, imply birth rates of 0.4-0.5 per century[14,15,16]. A wide range of birth rates (~0.06-1.9 per century) are predicted for ULPMs from population synthesis, complicating our understanding of their evolution[22].

Research Plan and Methods: This study aims to improve our understanding of magnetar evolution by conducting a magnetar-targeted radio survey and a forward modeling experiment to estimate the population and birth rate. The two components are summarized below:

1. The Deep Synoptic Array-110 (DSA-110) radio array will be used to implement an image plane search for long-period magnetar emission
2. A magnetar forward modeling experiment will be conducted using three possible birth scenarios:
   A. Magnetar birth properties (period, magnetic field) are drawn from unique distributions
   B. Magnetar birth properties are drawn from the same distributions as radio pulsars
   C. Magnetar birth properties are drawn from the same pulse period distributions, but independent magnetic field distributions, and undergo a rapid spin-down process early in their life-cycle

State-of-the art surveys such as CHIME and MeerKAT do not search long timescales (>0.1 ms); the recently operational DSA-110 array has comparable sensitivity on short timescales, and will search for transients up to 3.25 s widths. The custom imaging pipeline will build on the current FRB detection with techniques developed for the Very Large Array (VLA) RealFAST survey[20]. 130 ms cadenced visibility data is nearest-neighbors gridded to the U,V plane, then de-dispersed, fast Fourier transformed, and cleaned to image. After eliminating known sources, pixels with signal-to-noise (S/N) greater than ~6-10 are then searched with boxcar filters. Arc-second localization will inform follow-up observation with the VLA. For Task (2), *PSRPop.py*, a software designed for synthesis of radio pulsars, will be modified to include alternate decay models (sub-, super-, and exponential), alternate field geometries (dipole, dipole + magnetosphere, and dipole + toroidal), and X-ray and gamma ray detection as in Gill and Heyl's experiments[21].

Intellectual Merit and Science Goals: The proposed study contributes to current research by conducting the first ULPM-targeted radio survey to build the known population. The increased 3.25 s maximum pulse width would increase the S/N of long pulses; for example, GLEAM-X and PSRJ0901 would have a 2.2x S/N increase in comparison to MeerTRAP. Probing larger pulse widths could also yield exotic objects like FRB 20191221A, a unique 3 s duration repeater exhibiting QPOs[23]. The synthesis experiments will address the relation between radio and magnetar birth distributions by modeling both fossil field and dynamo process birth scenarios. Further, by modeling rapid initial spin down and fallback accretion spin-up, the magnetar birth rate and mean age can be better constrained. Moreover, modeling X-ray and gamma ray telescope selection effects could improve Jawor, et al.'s detection model. The primary outcomes of the project will be an increased population of long-period radio transients, constraints on the birth pulse period and magnetic field distributions, a revised magnetar spin-down history, and updated magnetar birth rate and mean age estimates.

Feasibility and Timeline: The novel magnetar survey will build upon the existing DSA-110 pipeline. The main risk will be implementing this in real-time using CUDA for GPU multiprocessing, which will be conducted with the DSA-110 software team. For population synthesis, *PSRPop.py* is well-equipped to run Monte Carlo simulations; the main risk is in modifying the code, whose Python 2 implementation makes this feasible. I will apply my experience working with Professor Vikram Ravi on polarization analysis of ~20 DSA-110 FRBs and simulating pulsar searches for the DSA-2000 readily to this project. The remainder of the 2022-23 year will involve development of methods and evolution models used for population synthesis. Forward modeling experiments will be conducted during the 2023-24 year, as well as implementation of the CUDA search pipeline for the image plane survey leading up to PhD candidacy. The magnetar survey will proceed from June 2024-June 2025, and synthesis experiments will be revisited including any new candidates. During this time follow-up observations with the VLA will be conducted. Publication of results is expected in late 2025 or early 2026 in conjunction with the PhD thesis.

Broader Impacts: This project provides a great opportunity to inform a wider audience about radio astronomy and pulsar science. One way I will do this is by organizing a radio astronomy demo for Caltech's Science for March event, in which campus researchers and organizations organize demos and talks to share their research with non-scientists. As part of Professor Vikram Ravi's group, I will be contributing to a hands-on undergraduate radio astronomy class in which students create small interferometer demos using dipole antennas. I will then use resources from these demos in the Science for March event to present the basics of radio astronomy. Additionally, DSA-110 data could be converted to audio signals to help younger students visualize radio observations. Furthermore, Caltech Letters, which publishes online science and cultural articles for non-scientific audiences, provides an avenue to share results from this project with a broader audience. I intend to contribute as a writer to discuss magnetars and pulsars, as well as topics surrounding culture in STEM fields. Both initiatives are avenues to increase STEM engagement and public interest in pulsar science and radio astronomy as a whole.

References: [1]Condon, et al., 2016, Princeton University Press; [2]Kaspi, et al., 2017, ARAA, 55,1; [3]Olausen, et al. 2014, AJSS, 212, 1; [4]Caleb, Manisha, et al. 2022, Nature, 6; [5]Hurley-Walker, N., et al. 2022, Nature, 601, 7894; [6]Rajwade, K.M. , et al., 2022, MNRAS 514, 2; [7]CHIME/FRB, et al., 2018, ApJ, 863, 1; [8]Wadiasingh, et al., 2020, 903, 2; [9]Jawor, et al., 2022, MNRAS, 509,1; [10]Faucher-Giguere, et al., 2006, AIP, 983; [11]Dirson, et al., 2022, arXiv e-prints, arXiv:2206.13837; [12]Prasanna, Tejas, et al., 2022, MNRAS; [13]Mereghetti, Sandro, et al., 2015, SSSI, 54; [14]Ferrario, et al., 2008, MNRAS, 389, 1; [15]Gill, et al., 2007, MNRAS, 381, 1; [16]Kouveliotou, et al., 1998, Nature, 393, 6682; [17]Gibson, et al., 2017, MNRAS, 470, 4; [18]Keane, et al., 2008, MNRAS, 391, 4; [19]Ravi, et al., 2019, Nature, 572, 7769; [20]Law, C.J., et al., 2018, AJSS 236, 1; [21]Bates, et al. 2014, MNRAS 493, 3; [22]Beniamini, et al., 2022, arXiv e-prints, arXiv:2210.09323; [23]CHIME/FRB, et al., 2022, Nature, 607, 256-259

Personal, Background, and Future Goals Statement

Background: When I was in middle school, my grandfather was misdiagnosed with tuberculosis. Although "state-of-the-art" technology existed, the suburban Chinese clinic had limited access to these expensive blood cultures. For over a week, my grandfather's viral pneumonia was mistreated with tuberculosis antibiotics, and he developed a persistent lung infection and permanent nerve damage. This event profoundly shaped my views on the importance of creating effective and accessible diagnostics.

Motivated by this experience, I decided to attend Caltech to pursue bioengineering, a field that stands out for its ability to **apply engineering on an observable personal and societal level**. My undergraduate experience has helped me realize that diagnostics are widely applicable and have the capacity to revolutionize environmental surveillance and healthcare.

Intellectual Merit: Since spring term of freshman year, I have been an active member of the Caltech Ismagilov Lab. I was initially interested in their work on rapid antibiotic susceptibility detection but soon realized that diagnostic tools are useful beyond clinical settings. My first project was to quantify and sequence microbes at low cell counts, a collaboration with the Jet Propulsion Laboratory to detect microbes on their spacecrafts for planetary protection purposes. The project was postponed due to the onset of the pandemic, and I shifted to virtually analyzing microbial "disruptor taxa" in small intestine biopsies of inflammatory bowel disease patients. To perform sequencing data analysis, I learned to code in Python and use bioinformatic tools such as QIIME2. Over the summer, I discovered how to perform Poisson regression, sensitivity analyses such as limit of detection calculations, and principal component analyses. These core bioinformatic skills were applicable to later projects in the lab and will be useful for developing diagnostic tools in graduate school. I also presented my findings at Caltech's Summer Undergraduate Research Fellowship (SURF) Seminar Day, where any member of the public could join, and it was exciting to learn how to best **share my research with both scientific and general audiences**.

When the pandemic continued into the summer, the lab launched a community-based COVID-19 project to quantify longitudinal viral loads from the incidence of infection and determine the optimal testing strategy (sample type and test sensitivity) to reduce household transmission. After learning about the lab's initiative, I joined the team, excited to **tackle a global crisis**. I performed a comprehensive epidemiological literature review on previous household transmission studies and helped design the survey instrument used to gather participant data. I also suggested ways to maintain participant engagement and make participation more accessible. During the study enrollment period, I digitally logged participants' daily symptoms and learned new programming languages such as PostgreSQL to help maintain the database quality. Through these experiences, I was able to suggest improvements to study structure and questionnaire wording when the study relaunched in late 2021 for Delta and Omicron variants. These involvements led to **two co-author papers, in *Journal of Clinical Microbiology* and *Microbiology Spectrum***, in which we compared efficacies of nasal and saliva tests for early viral detection and identified advantages of morning sample collection. The experience helped me better appreciate the **value of collaboration and the power of community engagement**.

As part of my 2021 SURF project, I **led an extensive epidemiological analysis** to identify modulating and risk factors of COVID-19 household transmission. The work involved combining the measured laboratory viral loads with survey data, and I worked closely with our local health department, biostatisticians at UCLA, and graduate students in public health. By **forging these meaningful interdisciplinary connections**, I built a model of household transmission that controlled for socioeconomic and demographic differences to evaluate the efficacy of different infection control practices. I also taught myself STATA to calculate adjusted odds ratios and secondary attack rates. I shared my findings on non-pharmaceutical interventions through a **poster at the American Society of Tropical Medicine and Hygiene Annual Conference** and as an oral presentation for the following SURF Seminar Day, where I was a **semifinalist in the Doris S. Perpall Speaking Competition**. Most importantly, when my graduate student mentor applied findings from this epidemiological analysis to his 7-person, multigenerational, high-risk household in Colombia, he was able to successfully prevent household transmission of COVID-19 in a context where testing and antiviral medications were limited. This reminded me that my **research results can lead to first-hand and real-world impacts**. Moreover, our viral load data suggested nasal tests delay detection of infected individuals, so I created a novel causal model linking diagnostic test type with household transmission. Currently, I am preparing a **first-author manuscript that demonstrates how the type of COVID-19 test has a large impact on transmissibility**. This finding stresses the need for better-designed testing strategies to combat increasingly transmissible viral variants. Overall, my work on COVID-19 has helped me realize how high-sensitivity assays are significantly more effective, but these expensive technologies are inaccessible to vulnerable communities. I aim to ameliorate this issue through my current and future research.

After returning to campus, excited to pursue hands-on research and diversify my skillset, I began working with the Ismagilov Lab team that was developing a fungal quantitative sequencing ("quant-seq") pipeline that combines highly sensitive digital droplet-PCR readouts with relative-abundance measurements from taxon-level sequencing. I initially focused on clinical applications, but as I read case studies regarding impacts of soil fungi, I realized the **importance of applying diagnostics to soil communities as a marker of ecosystem sustainability and environmental health**. I participated in this project from planning to execution, helping draft and secure a **grant from Caltech's Resnick Sustainability Institute** and reaching out to potential collaborators. In contrast to established bacterial standards, fungal barcodes were lacking in coverage and specificity, so I evaluated new primers and applied restriction enzymes to remove off-target amplification. Although the procedure still requires optimization, we have tested our workflow by measuring fungal absolute abundance in contrived soil and human microbiome samples, and I shared the initial findings at the 2022 SURF Seminar.

Over the course of three academic years and summers, I have tackled three diverse real-world problems with different graduate mentors. The Ismagilov lab's **interdisciplinary nature, collaborative culture, and focus on translational research** has further cemented my interest in diagnostics and equipped me with valuable tools to address global problems. At the end of my junior year, due to my initiatives with the COVID-19 study and fungal quant-seq, I was **nominated by Dr. Ismagilov and subsequently awarded the Caltech George and Bernice E. Green Memorial Prize**, which honors up to two undergraduates from any class for original research and evidence of creative scholarship.

To broaden my perspectives and to learn protein-based techniques, I also joined the Caltech Mazmanian Lab at the end of junior year. I initiated conversation with Mazmanian lab members for a potential collaboration applying fungal quant-seq to their mice and became interested in their work studying the microbiome's influence on Parkinson's disease. I made meaningful progress on projects in both labs through effective time-management and prioritization of goals and experiments. Over the summer, I visualized and compared markers along the p53 pathway and levels of mitochondrial oxidative stress between germ-free and control Parkinson's mice models. To do so, I developed mice handling skills and optimized an immunofluorescence procedure that reduced resource and time consumption.

My experiences at Caltech have taught me to 1) identify pressing societal issues and 2) to plan and execute creative research projects to address them. For my PhD, I aim to research and make high-sensitivity diagnostics for in-field and point-of-care use at Harvard's Wyss Institute, which uniquely leverages cross-disciplinary interactions to bridge the gap between research and practical applications.

Broader Impacts: Growing up in Southern California, I participated in Caltech-run Science Olympiad (Scioly) events throughout high school. The program provided me a platform to explore diverse STEM topics such as herpetology and minerology. To **give back to the community that piqued my interest in science**, I joined the SoCal Scioly planning team, a chapter that organizes annual tournaments for over 6,500 local students across 10 counties. Ever since freshman year, I have written and reviewed tests, corresponded with hundreds of volunteers, and planned awards ceremonies for regional events. Last year, **Caltech hosted the national tournament** for the first time; in addition to my normal duties, I designed virtual games to increase student engagement. Taking on **leadership roles**, I served as Treasurer and coordinated a materials subgroup that ordered and managed tournament supplies. However, since much of the work had been behind the scenes, I also actively searched for opportunities to **directly engage with students**. I was selected as the elementary division coordinator and worked closely with a team of dedicated middle and high schoolers to prepare elementary-level events. I soon realized that the students had many questions about what life as a scientist was like and were worried if they were prepared for it. I **dispelled**

common stereotypes about careers in science and **encouraged them to pursue futures in STEM** by sharing snippets of fun personal experiences as an undergraduate researcher and how, just like them, I started my journey as a Scioly competitor. Seeing student responses first-hand spurred me to **push Scioly to become more accessible to underrepresented communities**. I proposed rule changes to include more household items, presented at workshops with new coaches, and secured funding to promote local outreach. I look forward to continuing my involvement with Scioly during graduate school.

Moreover, I discovered a passion for tutoring and mentoring. I have been a RISE tutor for the Caltech Y since freshman year. The afterschool program connects Caltech students to local public high school students struggling with math and science. Many students are from underrepresented and low-income communities with limited resources at school. I learned to work with students from diverse backgrounds and tailor each of my tutoring sessions to best accommodate their learning style. I focused on filling knowledge gaps, teaching valuable self-study skills, and providing flexibility not available in large classrooms. I came to realize that many of these children liked STEM, but they previously did not have the proper support to pursue it. I remain in contact with many of my past tutees as they pursue college degrees and move beyond the program, and I am excited to see their future adventures. I will also participate in similar community-based tutoring events in the future, where I can interact with younger kids to **bridge socioeconomic gaps and better democratize science**. I have also been a teaching assistant for an introduction to complex analysis course (ACM95A), one of the largest non-core classes at Caltech. Last year, 172 undergraduates and postgraduates from various engineering majors were enrolled, and I worked to address problems that students from different backgrounds might have. As the class does not have recitations—a feature I found challenging when I took the course—I converted my office hours into a partial recitation. In addition to answering specific homework questions, I reviewed important formulas, conditions behind applying theorems and lemmas, and example problems that weren't covered in lecture. When I was grading, I provided detailed feedback and reached out to individuals regarding specific errors. Receiving positive student feedback further affirmed my decision to teach the course again the following year and to **serve as a teaching assistant during graduate school**.

To expand my sights outside of my field and explore interdisciplinary relations, I have been co-leading a project on media biases with Harvard anthropology professors. The research aims to show that mainstream U.S. media has biases in covering archaeological findings from different regions, and I have been leading the data analysis portion. Using my knowledge of data analysis from class and research, I showed that there is statistically significant lower news coverage of research from East Asia versus those from Western Europe. I gave an in-person oral presentation of this finding at that Annual Society for American Archaeology conference last spring, and an editor at *Science Advances* was interested in this research. The work has since expanded, and I am collaborating **with professors and researchers from 12 different universities to publish a joint paper titled "Changing the Landscape of Archaeological Publishing."** This experience has pushed me to reflect on representation of my cultural background in mainstream U.S. media outlets, **become more familiar with the scientific publishing industry**, and identify changes needed to improve reachability of scientific journals outside of archaeology.

Further broadening my global perspectives, I am currently **studying abroad at University College London (UCL)** for the first term of my senior year. To complement the applied engineering aspect of my studies at Caltech with traditional biology coursework, I am taking classes on bioethics, physiology, and medical anthropology. This program has provided me the opportunity to engage with a wide range of communities within the large international student population, diversify my perceptions of STEM, and reflect on cultural differences. Through sharing my experiences and engagements in societies at UCL, I am **encouraging other Caltech students to pursue study abroad opportunities**.

I plan to remain in academia and hope to have my own research group developing high-quality environmental and clinical diagnostics that are accessible in limited resource settings. **Mentorship** has largely shaped my undergraduate experience and helped me grow personally and academically, and I would like to provide future students with the same support. The NSF Graduate Research Fellowship is instrumental to my future career goals and would provide me necessary resources to continue my involvement with mentorship and community outreach programs during my PhD.

Graduate Research Plan Statement

The Food and Agriculture Organization of the United Nations estimates that crop pathogens reduce total agricultural yield and quality by 10-16% and cost the global economy over $220 billion every year[1]. Due to climate change and globalization, new plant diseases are emerging at accelerated paces and pose additional threats to food security[2]. There is an urgent need to establish an affordable surveillance system for early and rapid detection of novel plant pathogens. Traditionally, high-analytical sensitivity plant diagnostics are performed only in laboratory settings, have long turn-around times, and are cost prohibitive at scale[3]. Affordable sensors have been developed for common viruses[4], but these methods are only capable of identifying specific pathogens and provide limited help in detecting uncharacterized agricultural infections. Novel rapid and portable DNA and RNA sequencing devices (e.g., MinION) are promising options for affordable in-field plant diagnostics and surveillance[2]. Although shotgun sequencing provides insight into environmental biodiversity by detecting genetic sequences from all kingdoms, it frequently misses organisms with low relative abundances of DNA, some of which may disrupt soil health or emerge as plant pathogens. Deep sequencing may identify these microbes, but it is costly for ecologically complex communities such as soil. Rather, *a selective sample preparation method that allows for robust detection of low abundance sequences across different organismal kingdoms is needed to provide critical, novel information about the biotic community inhabiting an environmental sample*.

Intellectual Merit: Compared with DNA, RNA is more abundant in cells and has a shorter average half-life, making it a better marker for detecting both *low-abundance* and *living* organisms[5]. However, RNA sequencing lacks standardization across sampling practices, making it difficult for use in the field, where scientific equipment is limited. Furthermore, it is crucial to minimize time to detection to prevent outbreaks. **I propose to develop and validate a rapid RNA sample preparation procedure that can be performed in a field setting to detect even very low abundance sequences across different viral, bacterial, and fungal kingdoms in specimens with complex, mixed organismal composition**. To do this, I will develop in-field sample-processing methods for environmental samples that will allow for parallel sequencing of different kingdoms, such that low abundance sequences from each kingdom can be detected by RNA sequencing. Data obtained using this method will provide novel insights into the structure and diversity of soil microecosystems to allow for subsequent interventions that support agriculture and biosphere health.

Aim 1: Develop a sample-processing procedure to selectively enrich RNA from different kingdoms. A standard RNA sequencing workflow involves multiple steps from nucleic acid extraction to sequencing adaptor ligation. This aim seeks to test and identify methods that can be inserted into stages of this workflow to sequentially select for RNA sequences from different kingdoms of organisms (virus, bacteria, fungi, plant) of interest in complex, environmental samples (**Figure 1**). I will first test physical extraction efficiencies. Because viruses are orders of magnitude smaller than other cells of interest, they may be isolated via a membrane size filter. Subsequently, I will differentiate between fungi, bacteria, and plant cells via liberation of nucleic acids by selective lysis of cell wall components frequently targeted by popular antibiotics and antifungals. Further, on the molecular level, specific transcriptional targets of plants (rbcL), bacteria (16S), and fungi (ITS, TEF1α) from literature can be leveraged for species identification since they may exhibit kingdom-specific sizes and GC content; if there are sufficient differences, I will select for transcripts from a particular kingdom using molecular size selection, chemical buffer conditions, or extraction techniques that promote binding or elution of targets of interest. Bioinformatically identified unique cut sites and canonical sequences in the transcriptional targets of kingdoms being selected against can undergo targeted degradation by engineered restriction enzymes and/or CRISPR-Cas13 systems. Selection between eukaryotic and prokaryotic sequences by poly-A pull down and/or priming during RNA reverse transcription will be evaluated. Then, I will apply probe-based digestion of undesired cDNA sequences and/or pull down of desired sequences by designed synthetic oligonucleotides to increase coverage and specificity within a kingdom of interest.

Aim 2: Evaluate and optimize method as a field-compatible rapid RNA library preparation toolkit. In parallel with Aim 1, I will optimize the time and resources needed to achieve successful separation for each step. I will first identify low-cost and low-weight devices appropriate for field use. Many pieces of lab equipment (e.g., $6 centrifuge) can be readily 3D printed from open-source websites for ten times less than

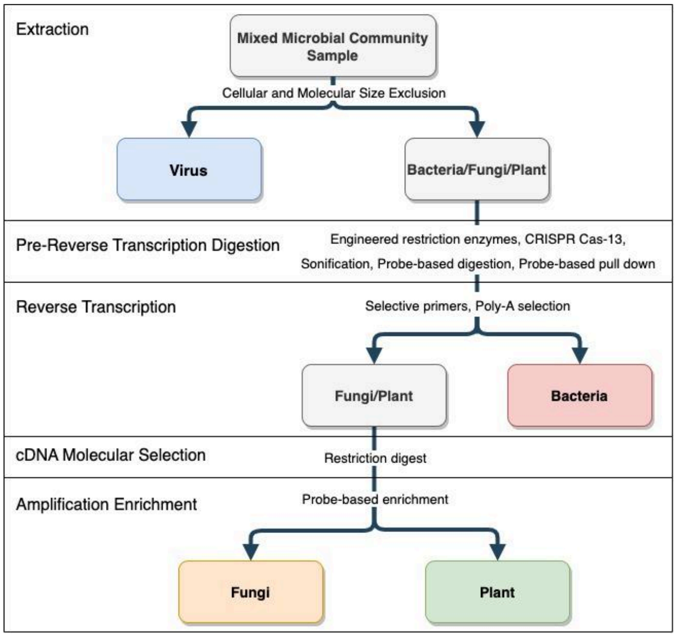
Extraction
Mixed Microbial Community Sample
Cellular and Molecular Size Exclusion
Virus
Bacteria/Fungi/Plant
Pre-Reverse Transcription Digestion
Engineered restriction enzymes, CRISPR Cas-13,
Sonification, Probe-based digestion, Probe-based pull down
Reverse Transcription
Selective primers, Poly-A selection
Fungi/Plant
Bacteria
cDNA Molecular Selection
Restriction digest
Amplification Enrichment
Probe-based enrichment
Fungi
Plant

Figure 1*: Proposed workflow separating viral, bacterial, fungal, and plant RNA.*

commercial prices[6]. I also have experience constructing 3D printed, Arduino-operated devices and will forge collaborations with LabOnTheCheap, a group dedicated to promoting the DIY science revolution, to create additional pieces of custom laboratory equipment that can be powered by a personal laptop. Furthermore, I will select for reverse transcriptase and restriction enzymes that can undergo inducible activation and isothermal activity at room temperature. For example, the enzyme PaqCI is useful for cleaving plant amplicons but is listed to function at 37°C. By running a temperature gradient, I can find if it yields comparable results as low as 30°C, which is obtainable with a DIY solar incubator. **Aim 3: Validate approach by demonstrating stratification of bacteria, fungi, and virus RNA in complex environmental samples.** Next, I will establish collaborations with institutions with well-characterized soil samples to test the sample preparation method, such as with NSF's Konza Prairie Biological Station[7]. RNA sequencing will be performed after the proposed sample preparation method, and the results will be compared to existing metagenomic data to evaluate coverage and detection of low-abundance sequences.

Prior research in the Ismagilov Lab has equipped me with skills to tackle the issue of processing complex, environmental samples for robust detection of low abundance sequences in different kingdoms. I have experience developing fungal quantitative sequencing technology, which combines relative-abundance measurements from sequencing with highly sensitive digital droplet-PCR measurements of absolute-abundance. I have explored many options to discern fungal from plant DNA via primer design features and amplification conditions. In particular, I have focused on identification and implementation of cost-efficient restriction enzymes to selectively cleave nucleic acids from different kingdoms.

Broader Impacts: Despite the existence of many national and regional plant disease surveillance systems, sample processing methods may fail to detect low abundance but high-consequence sequences, including those from novel environmental pathogens. Via innovative sample processing pipelines, previously missed data will yield substantial insights into microecosystem composition and ways to better support soil community health. Additionally, although most emerging plant diseases are detected in developed countries, this is likely due to access to laboratory-based technologies that are not available in lower resource settings[2]. Thus, affordable and rapid in-field diagnostics are needed to promote global surveillance. Further, as many human pathogens, including SARS-CoV-2, are likely the result of environmental spillover, highly sensitivity sample preparation technologies developed in this project for environmental samples may also improve surveillance and early detection of emerging pathogens with pandemic potential.

References: **1.** Chakraborty, S. & Newton, A. C. Climate change, plant diseases and food security: an overview. *Plant Pathology* 60, 2–14 (2011). **2.** Ristaino, J. B. *et al.* The persistent threat of emerging plant disease pandemics to global food security. *Proceedings of the National Academy of Sciences* 118, e2022239118 (2021). **3.** Fang, Y. & Ramasamy, R. P. Current and Prospective Methods for Plant Disease Detection. *Biosensors (Basel)* 5, 537–561 (2015). **4.** Li, Z. *et al.* Non-invasive plant disease diagnostics enabled by smartphone-based fingerprinting of leaf volatiles. *Nat Plants* 5, 856–866 (2019). **5.** Emerson, J. B. *et al.* Schrödinger's microbes: Tools for distinguishing the living from the dead in microbial ecosystems. *Microbiome* 5, 86 (2017). **6.** Lab On The Cheap! https://www.labonthecheap.com/ (2021). **7.** Terabase Metagenome Sequencing of Grassland Soil Microbiomes | Microbiology Resource Announcements. https://journals.asm.org/doi/full/10.1128/MRA.00718-20.