Candidate Overview: I am currently an undergraduate student at Northeastern University studying Mathematics and Physics. I am applying to MIT's EECS doctoral program so that I can pursue a PhD building architectures with relaxed equivariance constraints for atomic- and materials-property prediction. I am fortunate to have had broad exposure through my undergraduate research, including work on approximation error bounds of equivariant functions, property prediction models for atomic data, computational imaging, and estimators for correlation functions relevant to Cosmology.

My research has resulted in several first author papers. One paper is in submission to Transactions on Machine Learning Research (TMLR) and two others are published in some of the best astrophysics venues, the Astronomical Journal and the Open Journal of Astrophysics. I also have a co-first author paper that was awarded an Oral Spotlight as part of the top 4/300+ submissions at the NeurIPS Machine Learning and the Physical Sciences workshop, which I wrote with one other undergraduate student and no advisory supervision. Additionally, I have over 14 total publications, including co-authorship, highlighted by a paper currently in submission to Nature Astronomy. I am the first student in my university's Physics department to be nominated for the computing research association outstanding undergraduate research award. Similarly, I am the first Northeastern student to have received national recognition for research, service, and academic performance through the leadership scholarship presented by the Society of Physics Students National Chapter.

Research Interests: Based on my research interests and preparation for graduate research, I am most interested in working with **The Atomic Architects Group led by Professor Tess Smidt. I would also be excited to work with Professor Justin Solomon as a primary or co- advisor.** I would like to spend my PhD researching the design of approximate equivariant neural networks and distributional symmetry breaking on molecular and materials datasets. This research thread is both timely and impactful: recent work has shown that relaxed and approximate symmetries can outperform both fully equivariant neural networks and fully unconstrained networks, which in turn has provided industry practitioners an improved suite of tools for identifying molecules and materials with desirable properties. In parallel, distributional symmetry breaking can be used to understand when equivariance constraints are the most useful and provide a theoretical basis for the design of networks with approximate symmetry.

On the design of approximate equivariant architectures, I am specifically interested in constructing generative models that use these relaxed symmetries. The end goal for this area of machine learning is to be able to generate molecules and materials conditioned on desirable properties such as drug likeness or stability. So far, approximate symmetries have mostly been used in common regression tasks, but the potential for these networks to inform the design of materials with inexact symmetry remains an untapped and exciting area for further research. On distributional symmetry breaking, I am interested in developing methods that not only measure *if* a distribution breaks group invariance, but also on *which parts* of the distribution. Characterizing distributional symmetry breaking in this way highlights where model relaxations are the most useful.

In the long term, my goal is to become a professor and lead my own research lab. I see my PhD as a stepping stone that will allow me to mature and refine my research interests within geometric deep learning. For the reasons mentioned above, I think that the EECS doctoral program MIT is uniquely suited to provide this for me.

Research Background: The primary thread of my undergraduate research has studied the generalization and approximation error of equivariant functions. In particular, I was concerned with how well equivariant functions could minimize calibration error, loosely defined as the discrepancy between a

model's accuracy and purported confidence. The main contribution of this work was a theorem that placed calibration error bounds under various assumptions of symmetry breaking between the model and the data, expressed through the language of correct, incorrect, and extrinsic equivariance. I also showed experimentally how symmetry mismatch can cause a detriment to model calibration in both regression and classification settings. The work provides practitioners a useful guide for using equivariance when model calibration is crucial. This led to a first author publication currently in submission to Transactions on Machine Learning Research (TMLR). This work also produced workshop papers at NeurReps and EquiSystems, which were held at the conferences NeurIPS and RSS. My contributions were especially significant considering that I proposed the main idea, what to prove and how, how to substantiate the results experimentally, and wrote up the final paper myself.

Eager to demonstrate the applicability of my theory, I joined the Center for Astrophysics Harvard and Smithsonian as a visiting scientist with the AstroAI group, where I was advised by Cecilia Garraffo and worked closely with Bill Freeman and Sara Seager. During this time, I showed that my previous work was applicable to the biomarkers retrieval problem. The biomarkers retrieval problem aims to identify signs of life in exoplanet atmospheres in the form of Carbon based macromolecules. Specifically, chemical spectra and uncertainty information are used with the Hamiltonian Monte Carlo (HMC) algorithm to estimate relative chemical abundances. I contributed to this problem by developing an E(3)-invariant machine learning model that predicted these chemical properties *and* embedded uncertainty estimates into the prediction using a technique called evidential regression. This allows us to look for Carbon based macromolecules whose spectra may not be measured on Earth. I assessed the reliability of the uncertainties in terms of the calibration error bounds derived in my previous work. Once I deemed that the uncertainties were well calibrated, I used them to build the covariance information necessary to perform HMC. A paper is currently in preparation**.**

A parallel thread of my undergraduate research was focused on developing algorithms for large scale astronomical surveys in order to perform weak gravitational lensing analysis. Specifically, I developed a method for modeling optical blurring artifacts intrinsic to the James Webb Space Telescope Near Infrared Camera (JWST NIRCam) for use by the COSMOS-Web astronomical survey collaboration. This led to a first author publication in the Astronomical Journal and an open source library for the community. I am proud of this work not only because I conceived of the idea itself but also because I have already seen an immediate impact on the field: The suite of benchmarks I developed for my paper were used in subsequent analysis; this includes the discovery of >100 rare strong lens systems, the highest resolution map of dark matter ever measured, and the public data release of the COSMOS-Web survey. These results were spread across 5 papers in submission to prestigious astrophysics journals. In the main, my work has enabled me to stand out in a large collaboration of over 200 scientists and supported my colleagues in their own research directions. In developing algorithms and libraries tools for other researchers to use, I became intimately familiar with best practices in software development, and in particular I became a big fan of the Julia programming language. So much so, that I wrote a co-first author perspectives paper with one other undergraduate student on the state and future of the language. This paper got an oral spotlight at the ML4PS workshop at NeurIPS. More importantly, the paper gave me the platform to engage with many leading developers on the language and conversations on how to increase its adoption.

Beyond imaging algorithms and libraries, I also developed algorithms and estimators for a suite of correlation functions relevant to cosmology. Initially motivated by my project characterizing JWST NIRCam, I became frustrated by data scarcity making traditional evaluation statistics inconsistent.

Ultimately, I realized the data scarcity issues were exacerbated by clustering approximations used to make the calculation of the statistics computationally feasible. To quantify this, I proposed an algorithm that calculated the uncertainty due to clustering on the fly when computing 2-point and 3-point correlations. This led to a first author publication in a prestigious astronomy journal, the Open Journal of Astrophysics.