Introduction: Discovery of new chemicals and materials has long been hindered by the maximum pace at which humans can run experiments and the enormous size of the search space.[1] Recently, the materials-design loop has been accelerated through several different mechanisms, including robotic high-throughput experimentation (HTE), atomistic simulations, and machine learning (ML).[1] While each of these techniques has independently shown promise for accelerating discovery, an approach that applies all three harmoniously would revolutionize chemical research with wide-ranging implications, enabling faster and cheaper discovery of new pharmaceuticals, catalysts, and photovoltaic devices. I aim to advance this effort in my own research by coordinating the interplay between these three methods for the discovery and design of new dye molecules. Dyes are a suitable class of molecules for testing an autonomous, integrated design platform because they have several readily measurable properties that must be optimized simultaneously for use cases ranging from solar cells to medical imaging. **Specifically, I am focusing on the following objectives: (1) developing ML models to predict UV-Vis absorption and emission spectra accurately given a dye molecule and solvent pair, (2) creating a generalizable, automated active machine learning framework to improve the prediction models, and (3) utilizing this framework to design a novel near-infrared (NIR) dye for biomedical sensing and diagnostics.**

Objective 1 - Model Development for UV-Vis Spectra Predictions: Accurate prediction of UV-Vis optical properties is essential to dye design for any application. Previous work toward predicting UV-Vis spectra with ML has mostly consisted of the simpler task of predicting two scalars, the wavelengths of the peaks of maximum absorption and emission (λ_{abs} and λ_{em})[2], and has been limited by data sparsity. Since starting my work with Prof. Rafael Gómez-Bombarelli in January, I have addressed this limitation by collecting all openly accessible UV-Vis data from seven online repositories (29,811 measurements in total) and standardizing it into a consistent format. I then used a combination of a directed message-passing neural network (DMPNN)[3] and a feed-forward neural network to predict a value for λ_{abs} given an input molecule-solvent pair and an analog of λ_{abs} computed with time-dependent density functional theory (TD-DFT). **Using this method, my model has achieved a test-set mean absolute error (MAE) of 8.68 nm (a 17% reduction in error over the previous best model) on the largest dataset for which ML predictions have been published.[2]** My first step toward extending this method to predict full spectra will be to train my model to predict the peak widths and intensities for each λ_{abs} using the data I assembled. I foresee the limited quantity of available data presenting a challenge for predicting full spectra since the majority of openly accessible data contains only λ_{abs} values for each molecule-solvent pair. I will address this issue with a pretraining strategy in which I train my model with lower-fidelity data and use the resulting neural network weights as the initial weights when training my final model (as opposed to a random initialization). I will then estimate the epistemic and aleatoric uncertainty in my model's predictions using a deep ensembling approach.[4] Finally, I will replicate the previous steps for predicting emission spectra. **My accurate models for predicting absorption and emission spectra will aid experimentalists in choosing which molecules to test, even before I further automate this process in the following objective.**

Objective 2 - Active Learning: My models' abilities to make predictions with corresponding uncertainties will fulfill an important prerequisite for implementing active learning (AL), which improves models by focusing the sampling of new data on molecules with high uncertainties in their predictions. The additional components needed for active learning are (1) a set of new molecules from which to sample, and (2) a method of measurement for each sample. My experimental collaborators in Prof. Klavs Jensen's group have created (1) by extracting a list of 7 million purchasable compounds from chemical vendor websites. Further, they have created a method for (2) by building an HTE apparatus for measuring 96 UV-Vis spectra simultaneously. Since TD-DFT calculations are faster and cheaper than experiments, I propose using these to augment the strategy for (2) by reducing the number of necessary experiments. I will design a computational framework to automatically deploy calculations for molecules with a high epistemic uncertainty and retrain my models using this new data. From molecules that still have high uncertainty after the calculations, my system will use the uncertainty values along with molecular similarity to choose 96 molecules to recommend for measurement in the HTE apparatus. My models will automatically receive

data from the experiments and repeat the previous steps iteratively until their predictive performances reach asymptotes of aleatoric uncertainty. **The proposed AL framework integrates TD-DFT and HTE in an automatic fashion, which will enable significant time and cost savings and will be readily generalizable to many areas of chemistry and materials science research.**

Objective 3 - Design of a Novel Dye for Biomedical Imaging: Once I am able to demonstrate that my ML models are sufficiently accurate over a large region of chemical space, I will adapt my AL framework to design novel molecules with optimized properties for biomedical imaging. Specifically, it is favorable for dyes to have absorption and emission peaks in the NIR-II range (1000-1700 nm) because this range has deeper tissue penetration compared to visible or shorter-wavelength NIR-I light.[5] High Stokes shift (λ_{em} - λ_{abs}) and high quantum yield are also desirable.[5] Additionally, I will leverage the ongoing work of my collaborators in Prof. Bill Green's group who are predicting solubility, toxicity, and photodegradation, as these are also important properties for this application.[5] I will employ the generative models of Jin et al.[6] to create new molecules out of substructures that are likely responsible for desired properties of interest in known molecules. Next, I will make predictions on these new molecules with my ML models. I will then modify my AL framework to explore the new chemistries proposed by the generative models; it will deploy TD-DFT calculations as necessary for molecules with uncertain predictions and ultimately recommend novel molecules with predicted properties in the target ranges to my experimental collaborators in the Jensen group. They will use automatic retrosynthesis methods[7] to synthesize the novel compounds and will then use their HTE apparatus to test which proposed molecules indeed have the desired properties. Finally, I will propose the best-performing molecules to Prof. Angela Belcher's group for further study and *in vivo* testing. **If successful, this strategy could serve as a blueprint for combining experiments, theory, and ML for multi-objective molecular design across the field of chemistry.**

Intellectual Merit: Design problems in chemistry and materials science often suffer from a combinatorial explosion of configurations to explore, which makes solution of these problems intractable by brute force, or with HTE, physics-based calculations, or ML alone. By using all three methods simultaneously and automating the interactions between them, my work will be an advancement toward a "closed-loop" system that can explore massive chemical spaces with minimal need for human intervention beyond the specification of design objectives. Conducting my work at MIT gives me the opportunity to collaborate with experts who have proven records of integrating chemistry and computer science methods, and it gives me access to computing resources to run atomistic calculations and train ML models. **An NSF fellowship would supplement my current computing resources with access to XSEDE and would ensure the necessary funding for my completion of this project.**

Broader Impacts: I will design a novel dye that could be applied to guide surgery or to detect cancer at earlier stages. My work's flexible multi-objective optimization will also be able to design new dyes for additional applications such as dye-sensitized solar cells. Furthermore, the AL framework I develop could be widely adopted to design other types of molecules and materials, such as those in batteries and catalysts. I plan to make all code and datasets I develop openly available online with detailed documentation, which will enable other researchers to replicate and build upon my work more easily. Additionally, **I will host a workshop to demonstrate my framework**, with the goal that even experimentalists with little computational experience would learn to utilize the AL component of my framework to accelerate their progress in molecular or materials design. My work ultimately aims to encourage greater collaboration between experimental, theoretical, and computational researchers by automating the connections between their work in pursuit of design challenges that would otherwise be intractable.

References: [1] *Angew. Chemie Int. Ed.,* 2019, doi:10.1002/anie.201909987. [2] ChemRxiv, 2020, doi:10.26434/chemrxiv.12111060.v1. [3] *J. Chem. Inf. Model.,* 2019, 59 (8), 3370–3388. [4] *J. Chem. Inf. Model.,* 2020, 60 (6), 2697–2717. [5] *J. Mater. Sci.*, 2020, 55 (23), 9918–9947. [6] ICLR, 2020, arXiv: 2002.03244. [7] *Science*, 2019, 365 (6453), eaax1566.