Microbiome states have been correlated with health and disease across biology. Specific relative abundances of microbes can influence phenotypes such as obesity, autoimmunity and cognition in humans[1,2,3] as well as pathogen resistance, drought resilience and bountiful yield in agriculture[4]. With the acceleration of microbial community research, the field has produced accurate models for predicting the trajectory of given consortia[5,6]. The question of driving the system toward a favorable diversity for the aforementioned phenotypes, however, remains open. While strategies for microbiome modulation exist (e.g. probiotics and antibiotics) these tools are blunt at best and can harm[7]. The tools themselves are not flawed, rather, the complexity of microbial communities demands sophisticated alteration programs; effects cascade in indirect and cyclic paths, yielding nonlinear behavior sensitive to randomness. This complexity bars microbiome control, obstructing research from application in medicine and industry.

There is headway on this topic: in the past year, researchers have drawn from control literature to propose a method for identifying a minimal set of species to drive the community with a Model Predictive Controller (MPC) toward a desired microbiome state[8]. However, *I hypothesize that this control framework may be limited in practice because of the dynamic and stochastic nature of microbial interactions and microbiomes respectively*, which hinder the accuracy of the generalized Lotka-Volterra model (gLV), the best-known model of microbiomes. My senior thesis demonstrates the first issue that microbial interactions, which are invoked as constants in the models, are functions of their molecular environment, and, thus, the overall community dynamics are as well. Additionally, researchers have shown multiple cases of the stochastic nature of microbiomes[9]. Finally, in communities with many members, parameter inference is a nonconvex optimization problem for which solvers may produce different interactions that predict the same dynamics with similar scores[6]. Therefore, the application of the published control architecture might fail in the transition from microbiome research in test tubes, where parameterization occurs in defined and fixed molecular environments, to live settings like the intestine or rhizosphere. **To overcome this problem, I propose two extensions of the gLV-MPC framework[8] that utilize Reinforcement Learning (RL) to improve the performance of driving a microbiome to a given state.** MPC's that utilize RL, the branch of artificial intelligence (AI) concerned with optimization of policies to achieve desired states for given actions, lead in myriad applications including self-driving vehicles, autonomous portfolio management, and robotic manipulation by making controllers robust to dynamics absent from the models[10]. If RL proved to make microbiome control robust in noisy, biological settings, a bridge would exist between existing lab microbiome research to medical, agricultural and industrial applications.

Underlying all control framework is the gLV model that follows as such[8],

$$\dot{\mathrm{x}}(\mathrm{t}) = Diag(\mathrm{x}(\mathrm{t}))[\mathrm{r} + \mathrm{A}\mathrm{x}(\mathrm{t})] + Diag(\mathrm{x}(\mathrm{t}))\mathrm{B}\mathrm{u}(\mathrm{t})$$

where $\mathrm{x}(\mathrm{t})$ is a vector of species populations at time t, r is a vector of their innate growth rates, A is the matrix of interaction coefficients between species of x, u is the vector of control inputs (transplants, probiotics, bactericides) and B is the sensitivity matrix of species' to certain inputs. I propose that the dynamics fluctuate with the non-constant growth rates, r, interaction network, A, and probiotic/antibiotic sensitivities, B, and, thus, our controller should be flexible to their non-constant nature.

As devised and demonstrated *in silico*[8], a linear MPC with quadratic cost function J succeeds in overcoming the nonlinearities of the gLV to drive the system to desired equilibria,

$$\mathrm{J}(\widehat{\mathrm{X}}_k, \mathrm{U}_\mathrm{k}) = \sum_{\mathrm{i}=\mathrm{k}}^{\infty} [\hat{\mathrm{x}}(\mathrm{t}_\mathrm{i}) - \mathrm{x}_\mathrm{d}]^\top \mathrm{Q}[\hat{\mathrm{x}}(\mathrm{t}_\mathrm{i}) - \mathrm{x}_\mathrm{d}] + \mathrm{u}(\mathrm{t}_\mathrm{i})^\top \mathrm{S}\mathrm{u}(\mathrm{t}_\mathrm{i})$$

where $\widehat{\mathrm{X}}_k$ is a series of predicted states from current step k that will occur by taking the control input sequence U_k, Q is a positive semidefinite matrix penalizing deviations of model-predicted states $\hat{\mathrm{x}}(\mathrm{t}_\mathrm{i})$ from desired trajectory x_d, and S is a positive semidefinite matrix penalizing the magnitude of control inputs. The MPC law is to take the input u^* that is the first in the series of inputs U_k that minimizes J subject to the linearized dynamics of $\dot{\mathrm{x}}(\mathrm{t})$. Note that Q and S are design parameters that decide which species the controller should prioritize by giving varying weights to species deviations and species inputs respectively.

Therefore, we have two avenues by which our controller can learn to improve its performance: through the modeling space by learning the parameters that best predict the system at hand or through the action space by learning the design parameters that control which species to prioritize in control. The first attacks the model uncertainty and assumes that we might ascertain parameters that accurately model the system, while the second ignores the gLV altogether and improves the MPC via experiential learning. With respect to RL, we can translate the problem by considering J to be the value function for state $x(t)$ and the policy as the rule of taking the action u^* which minimizes the cost J [11]. With regard to the two avenues of learning, I propose two reward functions for scoring our policies, in a Policy Search based approach[11],

$$R_{r,A,B} := -\sum_{i=0}^{k} \left\| \hat{x}\left(x(t_{i-1}), u_{i-1}^{*}(r, A, B)\right) - x(t_i) \right\|^2 \qquad R_{Q,S} := -\sum_{i=0}^{k} \left\| x\left(x(t_{i-1}), u_{i-1}^{*}(Q, S)\right) - x_d \right\|^2$$

After performance, $R_{r,A,B}$ would penalize the deviations of the actual system changes from the model predicted changes with $r, A,$ and B for each previous step, and $R_{Q,S}$ would penalize the deviations of the state that the controller with Q and S caused and the desired trajectory. The slow pace of microbial systems allows for 'active' learning, improving between each input. Particularly, the slow dynamics allow for more computationally expensive optimization such as with Deep Neural Networks (DNN) that prove best with nonlinear, nonconvex rewards[10]. We can, thus, pair the proposed reward functions and MPC's with DNN optimization to learn the best methods for driving a microbial system actively, perhaps in the context of a sick patient. The schemes are also amenable to traditional training over several episodes of system exploration which might occur in mice or voluntary clinic trials preceding at-risk system usage.

The hypothesis of my work is that the RL-enhanced algorithms will succeed beyond the simple environments for which complete and accurate parameterizations have been garnered, and thus, I propose to validate my experiments in wild-type mice with disease-state microbiomes due to antibiotic-caused Clostridium difficile infections[6]. This approach is built on the published parameters of several common intestinal microorganisms in mice[5, 6] which form a partial view of a network that might exist in any wild type microbiome with some known and some unknown species. The two RL-controllers will require a period of training which would occur by driving several mice from a disease state to a defined healthy state[6]. I will investigate what duration of training is necessary and how frequently policies need to be updated to see performance converge. We can quantify success of the experiment by the performance of the RL-augmented controller compared with the standard MPC and an undisturbed control in a population of mice not used in controller training. The performance would be evaluated through the percent success of driving the microbiome in the experimental population given similar origin and diversities.

The success of an RL-based microbiome controller would yield diverse benefits, primarily by making a large body of academic microbiome research useable in medicine, agriculture, and industry. Additionally, as an interdisciplinary project, it would further crosstalk between bioengineering, AI and control theory: the same scheme here has been proposed for transcription circuits and cancer therapy[8] and the RL extension might prove powerful in those settings for similar reasons. If I were accepted to a PhD at my current university, I would apply for the Catalyst Program[12] for industry partnership and translate the proposed algorithm into a medical software for dictating treatments. Despite the complexity, the project still requires simple tasks such as bacterial coculture, sample processing, and coding for data analysis that provide opportunities for students to participate and learn the advanced topics. If I were at my current university, I would connect with the Empowering Leadership Alliance[13] and SMASH[14] minority outreach programs to offer my project as a platform to teach basic wet lab techniques and coding in summer sessions. This would also be a great way to find a mentee to assist me on the project. These endeavors are crucial because they provide access for underrepresented students to learn bioengineering, math and AI and pursue research in general, necessary for social equality and the most ethical and cunning science of the future.

References: 1. Rosenbaum et al. 2015 2. Lyte et al. 2014 3. Johnson et al. 2018 4. Lakshmanan et al. 2017 5. Venturelli et al. 2018 6. Bucci et al 2016 7. Cohen et al. 2018 8. Angulo et al. 2019 9. Justice et al. 2017 10. Song et. al 2020 11. Sutton and Barto 2014 12. Catalyst 13. ELA 14. SMASH