Background and Motivation: Rising greenhouse gas emissions are accelerating climate change, posing risks to humans and the environment [1]. The transportation sector accounts for almost one fourth of the total greenhouse gas emissions from the global energy sector [2]. To decarbonize transportation, the International Energy Agency has projected that liquid biofuel production has to increase 2.6-fold by 2030 [3]. Ethanol produced from cellulose (cellulosic ethanol) exhibits potential to meet the rising demand for liquid biofuels, as cellulose is an abundant renewable resource found in plants, and ethanol can be efficiently converted into biofuels to displace fossil fuels [4]. However, achieving high ethanol titers (concentrations in gram/liter) is a **problem** limiting the commercialization of cellulosic biofuels. Microbes that can naturally break down cellulosic biomass (organic matter from plants) are a promising candidate to engineer for cellulosic ethanol production [5]. One example is *Clostridium thermocellum*, a microbe that grows at high temperatures (a thermophile). **State-of-the-art** approaches aimed at improving ethanol titer in *C. thermocellum* have included: 1) performing experiments that enable the microbe to acquire mutations that improve its ability to make ethanol, 2) expressing enzymes (catalytic proteins) from other microbes in *C. thermocellum* to improve inefficient chemical reactions in metabolism, and 3) deleting branching metabolic pathways to divert more carbon to ethanol [5]. These approaches have fallen short of developing a strain of *C. thermocellum* that can produce 40 g/L ethanol, the benchmark titer required for commercialization [5]. This prior research indicates that we do not fully understand what is limiting high ethanol titer in this microbe. This raises the **question**: What factors are limiting high ethanol titer in *C. thermocellum*?

Research Idea: A **novel approach** to advance our understanding of the factors that limit high ethanol titer in *C. thermocellum* is to identify a thermophilic pyruvate transporter (TPT). This is a protein embedded in the cell membrane that enables pyruvate transport into the microbe under high temperatures. Pyruvate is a compound of interest, because it is a key intermediate in the metabolic pathway from cellobiose (a product of cellulose) to ethanol in *C. thermocellum*. Expressing a TPT in *C. thermocellum* would enable the study of pyruvate transport into the cell and its impact on ethanol production. This would help to identify if the factors limiting high ethanol titer are located upstream or downstream of pyruvate in metabolism. To date, I am not aware of any TPTs that exist. To identify a TPT, I plan to use R to construct a model based on Bayesian inference to computationally generate a ranked list of 1000 candidate transport proteins based on predictions of TPT activity, experimentally test the top 30 candidates for activity in *C. thermocellum*, and use the results to refine the model structure. I **hypothesize** that candidate proteins predicted by the model to have high pyruvate transport activity (prediction ≥ 0.80), will enable pyruvate uptake in *C. thermocellum*.

Research Approach: My method for identifying a TPT is divided into two Aims (Figure 1). I plan to use a personal desktop computer to construct the computational model in R and the resources and equipment available in the research lab at my institution to experimentally test the top candidates for activity in *C. thermocellum*. **Aim 1: Generate an initial ranked list of candidate proteins and experimentally test the top candidates for activity in *C. thermocellum***. The **first step** of Aim 1 is to define the model: $y = f(X,\theta) + \varepsilon$, where y is the yield (prediction), X is a vector of five different biological features important for identifying a TPT, θ is the parameters, and ε is the model-observation mismatch (residual). This step also includes defining the data generating function (encapsulated in ε), my biological intuition about TPTs as priors on the parameters (using beta distributions), and the candidate function as: $\text{yield} = X_1\theta_1 + X_2\theta_2 + X_3\theta_3 + X_4\theta_4 + X_5\theta_5$ normalized between 0-1. The **second step** is to collect the feature data for each candidate protein from biological databases. The **third step** is to calculate the yield (prediction) for each candidate. The **fourth step** is to generate a ranked list of candidates in descending order based on yield. The **fifth step** is to experimentally test the top 30 candidates from the ranked list for activity in *C. thermocellum*. This involves expressing the gene encoding each candidate in *C. thermocellum* and performing an assay to assess the uptake of different concentrations of pyruvate into the cell compared to the wild-type strain as a control. **Aim 2: Refine the model structure and ranking based on the experimental results**. After obtaining the initial experimental results, the **sixth step** is to define and calculate the likelihood function (assuming a Gaussian distribution). The **seventh step** is to calibrate the model and refine the model structure and

choices. This involves updating the priors on the parameters and running a Markov Chain Monte Carlo method to produce a posterior distribution of parameters. The **eighth step** is to calculate the posterior predictions and re-rank the candidates. The **ninth step** is to select a new set of candidates to experimentally test. This method can iterate through steps 5-9 to further refine the model based on experimental results. However, the effect of neglecting structural uncertainty is unknown in this method because only one model is used. Extending Aim 2 to use multiple models and find the model with the lowest cross-validation prediction error would be one approach to account for structural uncertainty within the proposed method.

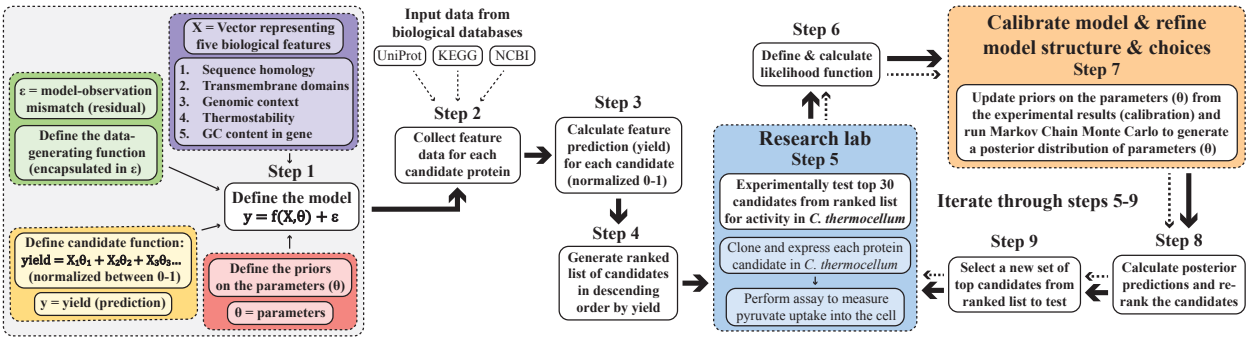

Figure 1: Data flow diagram of the method outlined in Aims 1 and 2 for identifying a functional TPT.

Initial Results: A code script has been written in R to simulate steps 1-4 of Aim 1 that defines the model, calculates the yield for 1000 hypothetical candidate transport proteins, and produces a ranked list of the top 30 candidates in descending order based on yield. Figure 2 details a table of the top, middle, and worst two candidate proteins based on calculated yield.

Candidate Protein	Protein ID	Sequence Homology	Trans-membrane Domains	Genomic Context	Thermo-stability	GC Content	Yield (Prediction)	Rank (Score)
Top 1	275	0.89	0.86	0.95	0.93	0.70	1.00	1
Top 2	373	0.76	0.93	0.93	0.97	0.43	0.92	2
Middle 1	413	0.24	0.43	0.23	0.93	0.66	0.52	500
Middle 2	496	0.18	0.07	0.58	0.96	0.69	0.52	501
Lowest 1	658	0.10	0.07	0.15	0.10	0.48	0.06	999
Lowest 2	371	0.09	0.07	0.11	0.00	0.45	0.00	1000

Figure 2: Ranked list of the top, middle, and lowest two candidates based on yield output by the model in R.

Expected Results: It is expected that 1-3 of the top 30 candidates from the initial ranked list will be able to transport pyruvate into *C. thermocellum* to varying extents, while most candidates will exhibit no activity. It is expected that sequence homology and thermostability will be the most predictive of the five biological features for determining pyruvate transport activity. It is also expected that prediction accuracy will improve in each iteration of the method from the experimental observations. It is also expected that 3-5 TPTs that are functional in *C. thermocellum* will be identified after 2-3 iterations of steps 5-9 of the method.

Intellectual Merit: This project may identify a few TPTs that are functional in *C. thermocellum*. This may inform what specific biological features enable pyruvate transport at high temperatures when compared to other pyruvate transporters. This work may provide insights into where the carbon from pyruvate is routed through metabolism in *C. thermocellum*, which would advance our understanding of its non-canonical metabolism [5]. In addition, expressing a TPT in *C. thermocellum* may inform if the factors limiting high ethanol titer are located upstream or downstream of pyruvate. This may inform future research aimed at troubleshooting these metabolic issues, which may improve ethanol titers in *C. thermocellum*.

Broader Impacts: This work may provide a computational framework that can be adopted to identify unknown transport proteins in microbes for diverse applications across the field of synthetic biology (i.e., bioremediation, biodegradation, etc.). Second, this work may provide a key step forward towards developing a strain of *C. thermocellum* that can produce cellulosic ethanol at an industrial scale. Finally, this work may advance our efforts aimed at establishing commercial biorefineries for cellulosic ethanol production to displace fossil fuels from the transportation sector and achieve net zero emissions by 2050.

References: 1. Keller, K. et al. *Annu. Rev. Earth Planet. Sci.* **49**, 95-116. https://doi.org/10.1146/annurev-earth-080320-055847. (2021). **2.** IEA. https://www.iea.org/data-and-statistics/charts/global-energy-related-co2-emissions-by-sector. (2020). **3.** IEA. https://www.iea.org/data-and-statistics/charts/bioenergy-use-by-sector-globally-in-the-net-zero-scenario-2010-2030. (2024). **4.** Lynd, L.R. et al. *Energy Environ. Sci.* **15**, 938-990. https://doi.org/10.1039/D1EE02540F. (2022). **5.** Mazzoli, R. & Olson, D.G. *Adv Appl Microbiol*. **113**, 111-161. https://doi.org/10.1016/bs.aambs.2020.07.004. (2020).