1. Results & Analysis

Midway through the Thesis, technical issues with the Tecan came to light that prevented the analysis of the data using the first calibration: As previously suspected, and later confirmed, the Tecan measurements were uncharacteristic and severely noisy - it is not clear for how long this had been going on and hence how much error the measurements contain. As a result, the single strain calibration data and likely some of the Tecan measured data given for this Thesis is as well affected. After the issues with the Tecan were resolved, new calibration data was obtained (SynCom (SC) calibration data), which serves as the primary model creation and testing data set. After the creation of the models, a more fundamental and intriguing problem was encountered that did not permit the analysis in their original form. After considerable effort in finding a resolution to this problem, one is believed to have been found and is described below. Analysis of the models and their behavior led to considerable improvement overall and provides a future strategy for monitoring the phycosphere modeling system under development. Having a solution, the data was subsequently analyzed; however, other issues were encountered and limited the amount of information that could be extracted from the data.

4.1 Modeling III

Although the visualization (**FigureX**) of the SS data does initially not suggest that the Tecan’s measurement data are flawed, upon comparison with the SC data (**FigureX Show Later**) there are obvious differences, especially considering they are prepared similarly and thus are expected to behave as such. Although the SS calibration models are not used for the analyses, they nonetheless served to establish much of the code and methods for handling and investigating the data using the SC calibration models. The early calibration data also revealed the un-intuitiveness of using absorbance data as a measure of concentration for comparing the contents of the SynCom and ultimately guided the selection of one regressor, which cut down the time spent on model selection. The initial coding base and methods, along with the selected regressor, are applied to the SC calibration and the appropriate models are made for analysis. The resulting models, however, quickly showed that their application would not be trivial and a solution would need to be engineered before an analysis could be made, let alone pursue monitoring the system in the way the lab has envisioned.

|  |
| --- |
|  |
| SSmodel |

4.1.1 Single-Strain Calibration Data Contributions

The first models constructed were used for the analysis of the Screen 4 (S4) data measurements. At the time it was unknown which wavelength would serve as a proxy concentration, henceall wavelengths of the measured spectrum were modeled (Aλ ~ A680 + A750). Browsing through any of those model’s predictions against the known concentration (as in **figureSSmodel),** it becomes obvious that only one regressor wavelength would suffice since only subtle differences in the range of the predictions themselves are present in each of these models. To avoid choosing a wavelength arbitrarily, scientific literature was consulted for information indicating the optimal wavelength to serve as a proxy concentration measurement. The literature provided information as to which wavelengths to use as regressands in single component systems, often citing the usage of the sharpest and maximally absorbing wavelength for a given compound **[chlamy papers]**. For Chlamy, this meant using A680 **[chemistry tome, interfrence by],** which the spectrum data in (**FigureSpectrum)** confirms as there is a spike at 680nm, as is expected given Chlamy’s chlorophyll content. Although 680nm is not the maximally absorbing peak, it is the sharpest peak as the two other peaks are overlapping and broad. Continued probing for a representative absorbance wavelength to use for subsequent analyses reveals that neighboring wavelengths produce similar values as noted in the continuous nature of the absorbance spectrum **FigureSpectrum.** It was reasoned that the wavelength with the least correlation to both A680 and A750 might be the least arbitrary choice and Panda’s built-in Pearson’s correlation implementation is used to obtain the spectrum correlation values **[cite pearsons]**. A special function is written to find the sums of minimums with respect to each of S4’s constrained wavelengths, which reveal the minima at A560 asseen in **FigureSum**. Note that in (**FigureAbsorbption**) this is the lowest absorbing wavelength (560nm) between the maximum absorbing peaks and is the furthest wavelength from the two main regressands (680 and 750nm) for the screening models. Given these observations, A560 is chosen to serve as the absorbance wavelength at which to represent concentration for all other work conducted.

|  |
| --- |
|  |
| **FigureSum** |

4.1.2 SynCom Calibration & Its General Failings

After the new calibration data (SC) set was obtained, the necessary models were created to be applied to the data. These models are derived from System 2 below and solved for the concentration of Chlamy and Strains as illustrated with System 1 in Methodology. These variables are denoted with the 560 subscript to symbolize that the predicted concentrations are absorbance measures at that particular wavelength. Each data set (Screens or PBR) is measured at explicit wavelengths and also dictates which wavelengths to extract extinction coefficients from and subsequently model **(TableX)**. The R2 value does not fall below 0.995 for any of the coefficient determination regressions, however, this metric is only good when the individual regressions are used to regress back to either measured component given an absorbance. Nevertheless, they do inform on the extinction coefficients themselves being precise.

|  |  |
| --- | --- |
|  |  |
|  |  |

|  |  |
| --- | --- |
|  |  |
|  |  |

|  |
| --- |
|  |
| **FigureSpectrum** |

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  | | |

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
| b | | |

The created models are first applied to the wells of the calibration data which are not used for the regressions (A-G:1-11). As is explained in Methodology 3.4.4, those rows and columns have a specific and known concentration of Chlamy and Strain as measured by and correspond the values used for the SLR regressions. The Beer-Lambert law states that the models should return the known and respective values at row H and column 12 as predictions and that their sums should be equal to the empirical measurements of their respective wells. When the sums of the predicted values are plotted against the empirical measurements, they should line up along the line as seen in **FigureX=Y.** For Model 1, the sums of the component predictions have an R2 value of 0.9972;on average, each predicted sum is only slightly overestimated by 0.023 absorbance units. However, this is an illusion, and although the linear combination that encompasses the Beer-Lambert law models is working as it should (per the predicted sums below), the actual component predictions are quite variable for the bacterial component and hint towards Chlamy being responsible for this observation. As seen in **FigureComponentPred** the gradient of Chlamy has the general effect of forcing the models to underpredict the Strain component’s content with a strength variable and reliant on the amount of bacteria at that row.

|  |
| --- |
|  |
| **FigureX=Y** |

The predicted values in **FigureXcomponentPred** are expected to line up along the red line for each respective component. Instead of doing so, the bacterial fractions predictions for are distributed along the y-axis with an average standard deviation of 0.033 absorbance units. The predictions for fare much better and have an average standard deviation of 0.006 absorbance units. Given the relatively small magnitude of absorbance measurements in general, the standard deviations for the bacterial components do not warrant their use for the analysis of the given data, as any prediction would have an associated error range of about 100x106 cells per cell-count regressions. Although the application of the models is not meant to provide quantitative measure of the contents of the phycosphere models themselves, it would be best if the models provide a sound estimate from which to form and test hypotheses and quickly iterate over experimental designs without the need to constantly measure the cellular content of the modeled phycospheres. Towards this extent, a general solution is devised and expanded upon in section 4.3.2.

4.1.3 Analysis of General Model Behavior

**FigureComponentPreds** shows that the interaction of light with these two organisms is not entirely linear for the bacterial fractions, as the Beer-Lambert law-based models should have produced. This observation is true for Models 1-3 and it is not completely surprising since the usual application of these types of models is towards determining concentrations of compounds or macromolecules, **[cite applications]** which are less complex and magnitudes smaller than cells. As is the case in any system studied through spectroscopy, light will not only interact with the analyte of interest; instead, every compound and macromolecule comprising the system up to and including the cell’s internal contents interact with the incident wavelength of light. The level of interaction or the affinity with which an analyte absorbs light is determined by their respective extinction coefficient **[optics, light interaction with cells, absorbance].** The goodness of Chlamy’s concentration predictions are largely based on/due to?? the effect of Chlamy’s cholorophyll content as it is responsible for absorbing a large fraction of light at 680nm, represented by the larger slope in **FigureInitReg**. The composition of Chlamy and Strains are similar to the point that, even at the carefully chosen concentration wavelength (), three of the extinction coefficients themselves are almost equivalent. As the models apply the coefficients to the given input, the similar light absorbtion of those wavelengths and Chlamy’s larger size have the net effect of attributing absorption to Chlamy, causing the range of bacterial predictions seen in **FigureComponentPreds**.

4.1.4 Specific Failure of Models 4 & 5

During the generation of the models required for the analysis of the given data, the subsequent visualization of Model 4 on the test data revealed an unexpected consequence of the similarity of coefficients. As is seen in **FigureWackyA600Model,** this model’s predictions are nonsense. A potential reason for this could either be the data at this wavelength being compromised or an error in the math causing the observed chaos. Previous inspection of the data had not revealed any issues with it and therefore the mathematics, and in particular, the relationship of the coefficients in the denominator had to be contributing to the ballooning of the predicted values.

|  |  |
| --- | --- |
|  | Model 4 [B] |

|  |
| --- |
|  |
| **figureInitreg** kB\_A680: 0.7928459513331864 kC\_A680: 1.6622985479003543 kB\_A750:  0.6960260403690639 kC\_A750: 0.7758826809892051 |

Note the arrangement of the coefficients in the denominator of Model 4: Each half is composed of opposing member and wavelength values. The difference of these products and the extinction coefficient in the numerator are multiplied by the empirical measurements of the modeled wavelengths (600nm & 750nm for Model 4). The calculation of the denominator, K**,** yields 0.0079 for this model: The resulting proximity to zero is the reason why the values are relatively large compared to the prediction in **FigureComponentPred.** This observation played a critical role in determining an optimized model (section?) and data collection strategy, and is further described in **the Discussion.** Given the incoherent results of this model, it cannot be applied to the S1 and S3 data; consequently, S1 and S3 data are no longer considered for further study. Moreover, S2 measurements were only made at F680 and A750 – seeing that at least two absorbance measurements are required to generate these types of models, it is not possible to extract information from S2 using the method under study hence S2 is also excluded from further study.

|  |
| --- |
|  |
| **FigureComponentPredfracts Model 1** |

|  |
| --- |
|  |
| **Wacky600** |

4.2 Pivoting the Thesis Scope

Several factors converged to force an exhaustive search for an optimization of the Beer-Lambert models. First, the model’s predictions for the bacterial fraction had a relatively large associated standard deviation. Second, preliminary application of these poorly behaving models to PBR and S4 resulted in predictions that consistently fell under the Tecan’s absorbance measurement limits (A**lambda** < 0.1) **[cite].** Last, the provided relative abundance measurements in PBR4’s **extra data** could not be directly compared to the preliminary ratios composed of absorbance predictions from Model 3: This comparison is necessary to validate the usability of the method for monitoring the simulated phycospheres. The reason behind this ratio problem is simple: The two types of organisms under study do not absorb light equivalently per unit cell making their absorbance ratios meaningless. This problem can be addressed by regressing absorbances (A560) to cell counts, and creating ratios from the regressed counts instead. Hence the regression table in the Methodology: It should facilitate comparing cell fractions and relative abundances going forward. However, the small valued results from preliminary model application and the associated large standard deviation problems remained. These issues forced the project, originally concocted to glean information for the respective taxa in the screening and PBR measurements, to change its scope from analytical study to exercise in mathematics and software engineering in order to interpret future measurements more meaningfully.

4.3 Modeling IV: Engineering Solutions

After failing to derive meaningful insights from the application of Models 1-3, it became obvious that a more exhaustive study of the results of the SC calibration would be necessary. The continued probing of the data yielded two key insights. The first was the uncovering of the hidden dependency of the bacterial fraction predictions to the known concentrations of Chlamy, which allowed the optimization of the Models 1-3. The second insight was derived from the failure of Model 4 and its chaotic predictions that result from the difference of the products of the extinction coefficients, **K**, of the denominator. The combination of these two results are combined into a general solution and strategy going forward in the continued development of the phycosphere modeling system.

4.3.1 Special K

The observation that the ballooning of the predicted values of Model 4 (e.g. **FigureWacky600)** could be due to small values in the denominator is an intuitive result from experience with the properties of zero. After conducting the necessary calculation, this hunch was confirmed. Upon comparison to Model 1’s **K** (0.554230), it was possible to hypothesize that a better overall model might exist. The specific hypothesis was that if the models became “worst” as **K** approaches zero, there would have to be a **Ks** in the space of all **K** that these data can generate, which would behave “best”. If two **K**s are produced for every model (Az ~ Ax + Ay), equal in magnitude but opposing in sign (when accounting for the commutative property), then **Ks** should have the largest magnitude in this space: **560 ~ 680 + 750 != 750 ~ 680 + 560.**

|  |
| --- |
| kspace  A560 ~ A680 + A750  A750 ~ A680 + A540  A750 ~ A540 + A680 |
| kspace |

In order to search this space visually by constructing a plot like **FigurePredFractions** is unwise as the SC data contains 26 different measured wavelengths and three of them are required to extract the four coefficients needed to create the predictive models. This results in exactly 15600 permutations, half of which are completely different models each yielding slightly different results than a neighboring model. Thus, a brute force approach was used to extract the four coefficients required to calculate **K**. The computation of this **K-space (Figure KSpace)** was done in 30 minutes on an Ubuntu 18.04.2 machine running on an Intel Core i7 CPU. Since the hypothesis states that the largest **K** would produce the best model, it was a matter of tabulating the results with the wavelengths used as an index and finding the index of the largest value in the data. The resulting **Ks** has a value of 1.64876 and is yielded by model **A750 ~ A680 + A540** (Model S)**.** The model is tested on the testing data and shown in **FigureModelS.** In contrast, the models required for the analysis of the given data have the form **A560 ~ A7\*\* + A680**. The required models have similar wavelength to Model S but have a better separation of coefficients as seen in **FigInitegresS.** Given a larger set of measured wavelengths, a more optimal model might arise from the data, as well as different models arising for different SynCom compositions. Therefore, special care should be placed when selecting the modeling wavelengths for monitoring any future SynCom. This should also be the first step as it is unclear whether these models are dependent on SynCom members or composition at time of calibration preparation [**elaborate more].**

|  |
| --- |
|  |
| **figureInitregS** |

Theresulting plots of the training data with respect to Model S (FigureModelS) do **not** look much different from Model 1 (**FigurePredFract**) at first glance, however, upon closer inspection, a few things become obvious. First, it would seem Model 1’s over and under predictions are an effect of the wavelengths used to model the data given that three of those coefficients are very similar (FigureBadCoeffPlot) given that they’re spead about the x=y line. For Model S, while the predictions do not line up along the x=y line, the underpredicted values seem to be a function of Chlamy’s concentration on respective bacterial load: i.e. the greater the bacterial concentrations, the greater the the model’s effect on lowering the predictions underpredicitons. Second, there is considerably less variability for low concentrations of bacteria even at high Chlamy concentrations, as noted by the predictions at B=0. There is a clear relationship in how the models attribute signal to the opposing member, which was first noticed with Model 1 and is explained further in Section 4.3.2.

|  |
| --- |
|  |
| **ModelSPlot** |

4.3.2 Optimization of Model S

Prior to uncovering Model S buried in the calibration data, a signal correction method for bacterial predictions was chanced upon when trying to understand the ratio of absorbance (B:C) for the predicted fractions of Model 3. While visualizing the data of the bacterial predictions in a table that conforms to the 96-well position (**FiguredRedTable**), a slight ‘bump’ in the data was noticed in row H columns 6-8 of the bacterial predictions. Row H has no bacteria in any of its well Chlamy gradient exists across the row. With perfect models, the value at each row would be the same and a flat line at the known concentration would be observed **(figure with flat lines)**. (Model S is used to illustrate the correction method found with Model 3). This plot, however, shows how the errors in bacterial fractions actually behave: The presence of Chlamy imparts noise to the prediction with a severity that is proportional to the concentration of bacteria. As the concentration of bacteria increases, the underpredictions become more severe with respect to Chlamy content. To observe this net effect of Chlamy only, the predictions at column 12 (where no Chlamy is present) is removed from each prediction at its respective row, effectively shifting the parabolas downward and forcing them to pass through the origin **FigureTruncated**. This behavior gave rise to an idea that allowed the manipulation of the predictions, albeit with **one/two** key concession, which results in an increased performance over the stock prediction of the modeled calibration data and is expected to work with other Tecan measurements as well as the phycosphere models themselves. this however, is yet to be determined.

|  |
| --- |
|  |
|  |

|  |
| --- |
|  |
|  |
|  |

|  |  |
| --- | --- |
|  | Standard Equation of a Parabola |
|  | Vertex Form of Parabola Equation: V = (h, k) |
|  | Relationship between the Equations above |
|  | Truncated Parabola |
|  |  |
|  |  |

In order for the manipulations to work, the concession has to be made that the observed parabolic trends seen in **FigureUntruncated** all come from a parabola having a unique breadth or **a2** parameter (i.e. the rate at which it flares open), which is not an unreasonable first approximation as seen in Un**TruncatedPlot**. Once this concession is made, the next step is to model the observed relationship between each fitted parabolas vertex (h, k) at **h** and the known bacterial concentration: **H** moves leftward as a function of B. Objectively, this relationship is true as there is an apparent movement of fitted parabolas vertex’ h value leftwards in **FigureUnctruncated**. It is not unreasonable to imagine that at any concentration of bacteria B, a theoretical parabola exists whose vertex lies along the trace and outlines the path of all possible predictions given Chlamy—lower lying parabolas have a lower rate of change than higher lying parabolas with respect to Chlamy concentration. When these theoretical parabolas are subsequently forced to pass through the origin, the Chlamy concentrations (predicted with a good degree of accuracy) determine the extent of this influence and the value is either added or removed from the original prediction depending on whether it lies below or above the x-axisas seen in truncated.

To grant the concession, the data is fitted to find the average parabolic breadth, which yields -0.4187; alternatively, the median can be used. Subsequently, it was found that the average over all the corrected Models 1-S (global **a2** average, -0.4995) gave the best performance. Next, the path that the parabolas’ vertices trace must be found and is the slope **h/B.** This parameter had to be found numerically from each of the fitted parabolas as the set of returned parameters **[a2,a1, a0]** do not yield an **h** using the analytical function to obtain **(h, k)**. Through Nympy’s **poly1d[cite]** function, which takes in a curves parameters to create a predictive linear model, a list of 100K numbers, representing a range of **h** values (-1,1), is used to obtain the **h** value at each apparent vertex (using the argmax property of Nympy’s lists). The large list of numbers is necessary approximate the true **h** as best as possible. The **h values** are plotted against the known B concentration of each row and result in the following regression **Fig\_hBRegres.** The parameters representing this line are used to compose a line **f(pB) = -m(pB) + b** which yields a **h** of (h, k) that corresponds to that particular parabola in the space of all possible parabolas which lie along that path.

|  |
| --- |
|  |
| **figureRedTable** |

Next, for any predicted concentration pB, this function is used to yield an **h** of a theoretical parabola of which the only other known attribute is an averaged breadth (a2) parameter. To fit a parabola to this h, **a1 is necessary** and can be derived from the mathematical relationships above. Once **a1** is obtained, the parameters **a2** and **a1**are fitted to conform to a truncated parabola in the form **[a2, a1, 0]. Using the poly1D** function, the resulting equation is a parabola which takes in **pC** and yields the net effect of the Chlamy concentration for that particular parabola. The yielded value is then added or subtracted from the given pB. The resulting corrections are shown in **FigureCorrected.**

|  |
| --- |
|  |
| Show only the corrected version as you’ve already shown the uncorrected... |

Whereas it may at first seem that this finding is an effect of circular logic, take for example the fact that in the above uncorrected graph, any horizontal line pB would cross more than one of the predictions in the graph (e.g Bpred = .213) but each yields the exact **h** from h**(pB)** and thus the same parabola is built from the resulting parameters. However, when the truncated parabola are subsequently fed their partner predicted Chlamy concentration, which is different in each case, the amount of signal which is subsequently removed is almost exactly what has been added by the original model. Furthermore, it would be easier to create a correction model taking pC alone. However, given the data, this is impossible because the model attributes different amounts with respect to the concentration of bacteria and is believed to be captured by the trace of the parabolic vertices. Thus it appears that although it is not yet known why, there is indeed some truth behind the mathematical manipulation of data in this manner. It is also to be noted that corrections make the predictions worst at low values of bacteria, maybe/potentially prompting a limit to the range of handled pB . Moreover, it is possible that without the accurate predictions of Chlamy, it would not be possible to achieve this kind of correction and that more testing is required.

# Describe The Space C B,

# Can only be applied post predictions

4.4 Pivoting Back to Analysis

As noted in 4.3.2 Optimization of Model S, there is an observed tendency for the Strain predictions to be either overpredicted or underpredicted by the Chlamy gradient with a strength that is proportional to the true concentration of the Strains. The optimization of Models 1-S is meant to bring those model’s predictions back into focus. The correction factors resulting from the model optimizations are yet to be validated experimentally; in the following analyses, they are nevertheless applied to the predictions made by their respective models. The models’ low predictions across the given was the impetus for finding a potential solution within the calibration data as the predictions appeared to be at the limits of reliable absorbance measurements. As the calibration data pointed out, the underpredictions were likely caused by the Chlamy effect, thus making the analysis at that time an exercise in guessing.

Armed with the correction factors, the analysis of the data could be conducted; these data include SC4 and PBR2-4 and PBR4’s metadata **[mentioned in the methodology?].** Analysis of the SC4 data consists of first resolving the signal between the components and searching for the Strains that exhibit interesting interactions with Chlamy, given a community or binary context and a control consisting of Chlamy’s supernatant. The PBR data serves to create a tool which can resolve the signals of the component’s concentrations in real time as PBR data is continuously collected during the run of an experiment. Before creation of a tool, the models’ applicability to the PBR data must first be established and shown to provide similar predictions on the same sample’s measurement by either Tecan or PBR. PBR2-3 data are used to establish the transferability of the models. Establishing this is key in analyzing PBR4’s metadata which consists of daily cell count and 16S sequencing relative abundance measurements. The hope is that through the successful application of the models to the PBR, the predictions can be matched to the observed relative abundances from the 16S sequencing. This would reduce the need to perform these as well as lead to a more efficient modulation and iteration of experimental design and optimization in the continued effort to establishing a robust phycosphere modeling system.

4.4.1 Analysis of Screen 4

This Thesis was originally centered on finding Strains that exhibited noteworthy trends, such as helping Chlamy grow or finding those that grew best with Chlamy. The expectation was that resolving their signals would yield such trends and a more nuanced experimentation with the resulting Strains could be undertaken. Towards this extent the Beer-Lambert law is employed, however, as it has been established, the resulting models’ usability was not as straightforward as initially anticipated. During the preliminary analysis of SC4, the predicted values from these models were considered too low to allow for a significant analysis and required finding a solution in order to assess the data for Strains of interest. Once the calibration data divulged the correction factors, the models were re-applied to their pertaining data and the predictions were corrected. Unfortunately, out of the screening data only SC4 allows itself to be probed in this manner as mentioned in section 4.14, thus it being the only screening data analyzed.

This experiment was conducted to answer two major questions in determining potential candidates for further experimentation: How does the Strain-to-Chlamy cell count inoculation ratios affect either component’s growth; and is the individually tested Strains’ growth dependent on a SynCom context or is a binary context (or Chlamy supernatant) sufficient. To test this, several microplateswere made which incorporate those conditions as explained in Methodology. It was expected that at least some of the Strains would exhibit differential growth and would show up as higher valued predictions of the measurements taken during the course of a week.

Upon comparing the corrected and uncorrected versions of the predictions for the data, no significant changes are observed between them. **FigCorrVunCoor** shows only slight, almost unperceivable, changes to the predictions. Considering that a low concentration of Chlamy is in the system, the Chlamy effect is only slightly manifesting and hence making these small corrections consistent to what is expected. Furthermore, the models have a tendency to overpredict Strain content at low known Strain and Chlamy concentrations, and therefore the correction highlighted in the purple ellipse is also consistent to the direction of expected correction. For example, Model 1 **FigureCompPred750**,used for this analysis, consistently overpredicts Strain concentrations; at low concentrations of Chlamy (), however, the predictions are minimally overpredicted, making these small corrections consistent with the calibration data.

|  |
| --- |
|  |
| FigCorr Uncorr |

Thirty bacterial strains are tested in this experiment and the figures below are good illustrations of the vast majority of the results. The predictions are for the most part under 0.050 absorbance units. Given the almost unchanged nature of the corrected results, the data itself is further reviewed and points to the issue of low predictions being a product of the empirical measurements themselves. The empirical absorbance data falls into the range of 0 - 0.192 and 0 - 0.181 absorbance units respectively, however, the vast majority of measurements are considerably lower than this and have an average of 0.030 and 0.015 absorbance units; inevitably, these measurements can only lead to the observed predictions. The instances of growth spikes in the results (e.g. ICL\_41 at 100Hrs, ratio=16), which would symbolize growth, are likely errors in the measurements as they only appear during that time for many other Strains and immediately shoot back down in the subsequent measurement. Had growth truly been observed, the spikes would have a flatter top or continue growing for more than just one measurement.

Despite predicted and corrected values remaining low, they nevertheless provide a minimal boundary for subsequent screening experiments and a reassurance that the models themselves can be viable. For example, take the side by side comparison of the ICL strain vs. SynCom and Chlamy-only context in **FigSC\_C,B\_C**. In this figure, the predicted values for Chlamy at ratio 1.6 are consistent with the cell count vs. regressions (**RegressionTable**), which show a detectable signal for the order of magnitude of Chlamy cells (105) used to conduct this experiment. In the neighboring plots, the higher ratios reflect a decrease of Chlamy relative to a constant bacterial load (1x106 cells), which is one order of magnitude below what is required (5.0x107 cells) to obtain a strong enough signal to model. These cell counts correspond to the lowest reliable values of and that can be modeled, 0.18 and 0.15 respectively, in their uncorrected form with respect to the medium’s (TP) absorption. Given these results, a good course of action would be to increase the inoculation concentrations for these types of experiments to correspond to the models’ limits as these boundaries effectively force a zero on the system from which growth can be subsequently observed and measured more accurately.

|  |
| --- |
|  |
| **FigSidebySide** |
|  |

4.4.2 Model Transferability to PBR

The verification of the transferability of **Model 3** to PBR data is conducted to ensure that the envisioned monitoring tool will accurately assess the system’s contents in real time. Gaining the ability to do so would allow testable hypotheses to be iterated over each modeled phycosphere and to modulate the system to tease out and resolve fundamental principles governing the dynamics of real phycospheres, which can be subsequently applied to plant microbiota. While it would be possible to obtain the necessary coefficients to create models from the PBR itself, transferring the coefficients from the Tecan’s calibration measurements is far simpler given the time and effort required to setup the PBR. Per the Beer-Lambert law, any calibration, whether through the PBR or Tecan, should yield the same extinction coefficients. Once corrected for the path length, the coefficients at 680 and 720nm wavelengths with respect to concentration are modeled as established in System 1. Correction of the PBR’s much longer path length requires the Tecan-derived coefficients to be multiplied by the path-length (2.7 cm) of the vessels (cite manual] as demonstrated in Model 2 PBR). However, when this was applied to the PBR measurements, the data was severely underpredicted. It is likely that the system is already calibrated much like the Tecan to correct the path length internally. Hence, the extinction coefficients are applied as obtained from the Tecan.

|  |  |
| --- | --- |
|  | Model 2 PBR |

The system is solved for each of the two components and the data is entered into the models to retrieve the predictions. In theory, the predicted values from PBR measurements at sampling time should equal the sample’s predicted values given its Tecan measurements since the aliquot taken at sampling time retains the same concentrations of the components. Once the models have been applied, the predicted values for each component should line up along the line, have a positive correlation if growth is continuous, or alternatively, have strong precision about theline if the system’s growth is constant.

In order to ensure the data is modeled properly, the PBR data has to be devoid of the medium’s absorption effects as the need to perform the inoculation of the phycosphere vessels in a laminar airflow hood and the time consuming nature of a PBR experiment set-up, blank measurements are not taken. However, it makes sense that the measurements immediately post-inoculation are dilute enough to use them as a blank correction. This ensures that the predicted measurements are not overpredicted because of the medium’s absorption. The correcting post-inoculation blank values are 0.066003 and 0.037617 for and measurements respectively; following this correction, the predictive models are applied to the data and corrected for the Chlamy effect.

**FigureX2masTree and X3,** shows the corrected results of selected vessels for this analysis. These results are indicative of what is expected from the transferability of the models, especially those of PBR2. In the selected vessels for PBR2, the values for vessels 3-5 indicate a positive correlation symbolizing growth in those vessels and line up along the line indicating that the predictions on both systems are similar. However, the limited amount of data used to validate the transferability of the models does not make a strong argument on behalf of this method for monitoring the system.

|  |
| --- |
|  |
| **FigureXmasTree** |

|  |
| --- |
|  |
| **Figxmas3, no grow exp3** |

Not only does the limited amount of data limit the validity of this method for monitoring the phycospheres, but the data was also measured incorrectly even though the effects should be minimal. The PBR2-3 data were collected before the start of this Thesis and were measured by the Tecan at 730nm instead of the 720nm wavelength used by the PBR. Predictions according to each data’s respective constraints are as follows: Model 2 is used for the Tecan measurements and Model 3 is used for the PBR measurements. Since the predictions do not stem from the same model, the SC calibration data was used to conduct a similar analysis for comparison. Those results indicate that bacterial fraction predictions should be, on average, 0.05 absorbance units higher for PBR-measured data; this is in agreement with the observed results in the figures. Chlamy’s predictions are only slightly affected and both systems should yield similar values . In case of more accurate data, it is expected that the viability of this method will be upheld and the tool will yield reliable results especially when fitted with the model corrections.

4.4.2 PBR Data Modeling

The Tecan-generated model’s transferability is as of now inconclusive given the low yielding results collected from PBR2 & 3 as well as the incorrectly measured wavelengths. Furthermore , the latest iteration of PBR experiments does unfortunately not have Tecan measured values as it was out of commission during the data collection stage of this Thesis. Given its observed stability, planned Tecan-measured samples will be key in demonstrating the viability of the models in monitoring the PBR as well as verifying whether the observed Chlamy effect is manifested in the phycosphere models given their larger volumes. Nevertheless, in the following, PBR4 data is modeled to illustrate how the proposed monitoring would facilitate analysis.

|  |
| --- |
|  |
| **FigurePBR4RAw** |

Each PBR experiment is monitored after each of the phycosphere modeling vessels are inoculated with a specific SynCom comprised of specific strains of the ICL and Chlamy. Once the instrument is connected to the network, it begins taking measurements; through its control software, the measurements can be viewed in real-time and monitored remotely. However, monitoring only shows the absorbance measurements at 680 and 720nm for each individual vessel and thus solely informs on the overall growth of the system. Alternatively, the system also measures the input of fresh medium over time, which continuously dilutes the system to maintain a constant absorbance. The amount of fresh medium serves as proxy to estimate the total growth **[turbidostat]**. The Garrido-Oter lab, however, is interested in knowing the state of both components in the system in real time and this part of the Thesis is undertaken to validate Beer-Lambert law as a viable approach towards this extent. Given the particular instrument model currently in use, the predictive models that can be constructed are limited to one: Model 3 . At these wavelengths, the calibration data indicates that it is particularly prone to the Chlamy effect as seen in **figureexp720.A**. Fortunately, the corrections devised for this model should bring the predictions back into focus as seen in the corrected version. The corrections for this model remain preliminary, however, and are still in need of validation for both the PBR and Tecan. They arenevertheless applied to PBR4 in order to illustrate their use.

|  |
| --- |
|  |
| Uncorr v corr@720 |

**FigurePBR4Raw** shows the state of the PBR4 measurements untilthe end of the data collection stage of this Thesis. The figures show the data re-sampled at a three hour window after removal of negative measurements, which occur mainly post-inoculation and are a calibration issue of the instrument. Each vessel represents one phycosphere and is independent of each other, visualized together in **FigureRaw** for simplicity. The figure shows that the phycosphere models exhibit growth within the first few days post-inoculation, although it is unclear as to whether the observed growth is due to Chlamy or Strain and at what combinations of the two. The calibration data established that the system can in fact be modeled, albeit inaccurately due to the Chlamy effect, and can be brought into focus given the correction factors. In order to illustrate the model and correction factor, the phycosphere model at vessel # 4 is chosen as it has the highest range of empirical measurements as well as corrected and uncorrected prediction forms. The rest of the modeled vessels are to be found in Supplementary Data.

|  |
| --- |
|  |
|  |
| Sidebysdie1 |

The empirical plot **FigA** shows that the SynCom quickly achieves growth, stabilizes and begins to crash around July 8th, then recovers and remains stable for a number of days. The periods of stability are artificially enforced on the system by programming it to maintain a constant absorbance level as measured by . Hence, the observed growth beginning on July 17th is due to the instrument receiving a new and higher directive, i.e. to grow up to absorbance units. Another crash is then observed on all vessels per **FigurePBR4Raw** on July 28th and indicates either instrument malfunction or a forced dilution of all vessels. The system then quickly obtains its directed growth and a new experiment is begun on August 1st . That experiment, symbolized here by the oscillating curves, is simulating the day and night cycle over a 24 hour period and shows how it adapts to those conditions. The valleys of the Chlamy predictions (in green) line up with the peaks of bacterial predictions. Whether these peak-to-valley predictions are a mathematical artefact or a true phenomenon should be observable in the empirical cell count data.

|  |
| --- |
|  |
|  |
| **Side by side** |

The uncorrected predictions for this vessel indicate that growth is about the same for the two components during the initial two weeks post inoculation. In the context of absorbance and this particular SynCom, the “same or similar” predictions translate to roughly 250x more bacterial cells relative to Chlamy per the -to-cell-count regression. After the system receives its new directive, the Chlamy cells quickly ‘explode’ in growth. The divergence of the growths immediately after the change in directive appears similar to what is observed in the diverging-converging oscillations in the latter part of the experiment. However, in this case, both empirical measurements are going upwards and the divergence is likely due to how these components absorb light per their extinction coefficients. Note that three out of four of those coefficients have similar rates of absorption for both wavelengths, and that Chlamy’s absorption of 680nm light is much more efficient than the others hence its perceived faster rate of growth (see **FigureInitReg).** The correction factor seems to corroborate this as the purple ellipse indicates an increased correction of Strains. When the system reaches the directed absorbance near July 17th, it forces the system to maintain said absorbance and fresh medium is introduced to the system as necessary. The constant A720 measurement allows the observation of slight bacterial growth as it continues to grow more medium is introduced.Given that the bacteria are dividing much faster than the Chlamy, the dilution is affecting Chlamy concentrations more.

The Beer-Lambert law states that a measured absorbance at any **lambda is** equal to the sum of the individual absorbance of each of the components at that wavelength. Thus for any wavelength measured or predicted, this must be true. However, as seen with the calibration data, this is not true for individual components in this system as Chlamy’s concentration impacts the overall prediction of the Strains proportionally to the true concentration of bacteria. While Chlamy’s measurements tend to be correctly predicted by these models when true bacteria concentrations are **A560 < 0.4,** the same is not true for the Strains predictions. However, with the correction factors generated from the calibration data these bacterial fraction can be reestimated as is seen in **FigUncorrCorr. The** resulting corrections are mostly unchanged and while the elevated levels of Chlamy do indicate a stronger effect, a low prediction of bacteria is likely indicating that the system has a low abundance of them hence limiting the impact of the chlamy effect. The negative predictions in the uncorrected results are mostly brought back to positive ranges as no negative values should exist. Negative values occur when the bacterial fractions are below the limit of detection for that particular instrument. The slight correction of these data indicate the at these concentrations, Strains and Chlamy are relatively well predicted.

To get a more objective estimation of the systems contents it might be useful to use the theoretical cell-counts to monitor the system. The following figures below are by no means true concentrations of cells in the system since the instrument used to measure cell counts and hence used to make the regressions for them isn’t very reliable**[cite]**. Nevertheless they provide first approximation in more intuitive unit of measurement for the proposed monioring of the phycospheres. **FigureCellCount** below shows the descrepancy of the absorbance by Chlamy relative to the Strains. The axis for the Chlamy cell counts show that their quantities remain fairly constant throught the experiement while the bacterial fraction fluctuates as the system tries to maintain a constant A720 during the day night cycle.

|  |
| --- |
|  |
|  |

Lastly, it was hoped that cell count regressions would provide a realiable first approximation of comparison to the 16S abundances below. However, it does not seem that using cell count this is a good proxy this. Given the relative abundances in **FigureSprinklyWinkly** below it seems, it would appear that more genetic material can be extract for the lower quantities of Chlamy than bacteria

|  |
| --- |
|  |
|  |

|  |
| --- |
|  |
|  |