applied microbiology

MDPI

Article

Predicting Microbiome Growth Dynamics under Environmental Perturbations

George Sun [1] and Yi-Hui Zhou [2,*]

[1] Bioinformatics Research Center, North Carolina State University, Raleigh, NC 27695, USA; 3litus@gmail.com
[2] Departments of Biological Sciences and Statistics, North Carolina State University, Raleigh, NC 27695, USA
* Correspondence: yihui_zhou@ncsu.edu

Abstract: MicroGrowthPredictor is a model that leverages Long Short-Term Memory (LSTM) networks to predict dynamic changes in microbiome growth in response to varying environmental perturbations. In this article, we present the innovative capabilities of MicroGrowthPredictor, which include the integration of LSTM modeling with a novel confidence interval estimation technique. The LSTM network captures the complex temporal dynamics of microbiome systems, while the novel confidence intervals provide a robust measure of prediction uncertainty. We include two examples—one illustrating the human gut microbiota composition and diversity due to recurrent antibiotic treatment and the other demonstrating the application of MicroGrowthPredictor on an artificial gut dataset. The results demonstrate the enhanced accuracy and reliability of the LSTM-based predictions facilitated by MicroGrowthPredictor. The inclusion of specific metrics, such as the mean square error, validates the model's predictive performance. Our model holds immense potential for applications in environmental sciences, healthcare, and biotechnology, fostering advancements in microbiome research and analysis. Moreover, it is noteworthy that MicroGrowthPredictor is applicable to real data with small sample sizes and temporal observations under environmental perturbations, thus ensuring its practical utility across various domains.

Keywords: microbiome dynamics; prediction uncertainty; environmental applications

Citation: Sun, G.; Zhou, Y.-H. Predicting Microbiome Growth Dynamics under Environmental Perturbations. *Appl. Microbiol.* **2024**, *4*, 948–958. https://doi.org/10.3390/applmicrobiol4020064

Academic Editor: Bong-Soo Kim

Received: 7 May 2024
Revised: 4 June 2024
Accepted: 7 June 2024
Published: 10 June 2024

1. Introduction

The human microbiome, an intricate ecosystem of trillions of microorganisms residing in and on the human body, plays a crucial role in maintaining physiological homeostasis, metabolic functions, and immune responses [1]. Disruptions in the microbiome have been linked to a plethora of conditions, ranging from gastrointestinal disorders to more systemic diseases such as diabetes, obesity, and even neurological disorders [2]. This symbiotic host–microbe interaction underscores the necessity to understand the dynamic nature of the human microbiome [3], particularly how it changes over time and in response to various environmental stimuli [4,5].

Under normal conditions, the gut microbiome is composed of a diverse community of bacteria, with Firmicutes and Bacteroidetes being the predominant phyla. Environmental perturbations, such as changes in diet, antibiotic usage, and exposure to pollutants, can significantly alter the composition and function of the microbiome, leading to potential health implications. For instance, antibiotic treatment can dramatically reduce microbial diversity, often resulting in an overgrowth of resistant bacteria and a decrease in beneficial microbes, which can disrupt metabolic processes and immune functions [6]. Understanding these population dynamics is crucial for developing strategies to mitigate the adverse effects of such perturbations on human health.

High-throughput sequencing technologies, particularly 16S rRNA sequencing, have ushered in a new era in microbiome studies, allowing for detailed assessments of microbial diversity and relative abundance across different human populations and conditions [7]. However, the vast data generated by these technologies present both opportunities and challenges.

One of the primary challenges is deciphering the temporal patterns and predicting future states of the microbiome, essential for preventive and therapeutic healthcare applications.

Predictive modeling in biometrics has historically employed various statistical methods, but these traditional approaches often fall short in handling the high-dimensionality and nonlinearity of microbiome data. The advent of machine learning, and more specifically deep learning, offers promising new avenues for such complex data [8]. Recurrent Neural Networks (RNNs) [9] and their advanced variant, Long Short-Term Memory (LSTM) networks [10], excel in analyzing and predicting temporal sequences, providing an excellent framework for modeling microbiome dynamics.

In this study, we introduce the MicroGrowthPredictor model which aims to harness the power of LSTM networks to predict changes in the human microbiome in response to environmental perturbations, a critical step towards personalized medicine and targeted therapeutic interventions.

2. Materials and Methods

2.1. Long Short-Term Memory (LSTM) Model

The Long Short-Term Memory (LSTM) network, a specialized form of the recurrent neural network (RNN) architecture, is explicitly engineered to address the challenges of learning from sequential data, notably long-term dependencies. Traditional RNNs, while theoretically capable of handling such dependencies, often fall short in practice due to the vanishing gradient problem, wherein information is lost over each time step during training. LSTM networks are designed to overcome this limitation, thereby making them particularly suitable for applications across diverse fields such as time series analysis, natural language processing, and, pertinent to our work, microbiome data analysis.

LSTM networks introduce a more sophisticated cell structure than traditional RNNs [11]. Each LSTM cell contains mechanisms called gates that regulate the flow of information into and out of the cell. There are three types of gates within an LSTM cell (Figure 1A):

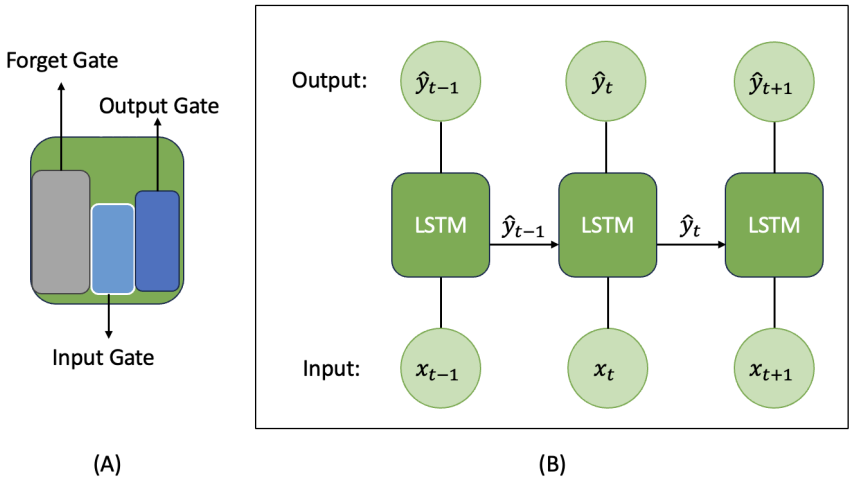

Figure 1. Long Short-Term Memory (LSTM) architecture: (**A**) A zoom-in on an LSTM cell, showing its three gates: the input gate, forget gate, and output gate. (**B**) The flow of the input and output data in an LSTM network from time step $t-1$ to time step t.

- Input gate: Modulates the amount of new information to be added to the cell state.
- Forget gate: Determines the extent of information to be discarded from the cell state. The forget gate helps to eliminate irrelevant or outdated microbial information, thus maintaining only the most pertinent data for accurate modeling.
- Output gate: Controls the quantity of information to be outputted from the cell. For microbiome data, the output gate helps decide which processed microbial information should influence the network's predictions or analyses at each time step.

These gates work together to update the cell's state and allow the LSTM to both remember and forget information over long sequences (Figure 1B), which is crucial for learning long-term dependencies. Figure 1B illustrates the transition of data through the LSTM network from one time step to the next. It shows the input data and output data as they flow from time step $t-1$ to time step t. At each time step, the input data, along with the cell state from the previous time step, is processed by the LSTM cell. This processing results in an updated cell state and an output, which are then passed on to the next time step. This sequential mechanism allows the LSTM network to effectively handle temporal dependencies, ensuring that information is carried forward and utilized across different time steps for improved prediction and analysis in time-series tasks.

In the realm of microbiome analysis, understanding temporal dynamics and sequential patterns is essential, given the nature of microbial communities' evolution and interaction over time. Here, we adopt specific notation to elucidate the mechanics of the LSTM model. Consider a training dataset $\mathcal{D} = \{(x_t, y_t)\}_{t=1}^{T}$, where x_t denotes the vector of relative abundances [12] of all microbial taxa at the t-th time step and y_t signifies the corresponding desired output. The LSTM takes these input sequences and processes them through its intricate cell structure, capturing valuable temporal dependencies present in the data that are critical for accurate predictions and analyses in microbiome studies.

2.2. Model Structure for Microbiome Growth Prediction

The LSTM model employed in this study is designed for both simplicity and power. The input layer is tailored to process the relative abundance levels of taxa, accommodating an extensive array of microbial taxa denoted as x_t. Comprising n_{taxa} nodes, each representing the relative abundance of a particular taxon, this layer corresponds to the total count of unique taxa identified in the microbiome dataset.

Moving into the architecture, our model consists of two hidden layers positioned between the input and output stages. The primary hidden layer incorporates an LSTM with n_h hidden states, functioning within a single layer. This configuration is crucial, allowing the model to capture and interpret temporal dynamics inherent in the input sequence, courtesy of the LSTM's characteristic memory cells.

To address overfitting and enhance the model's robustness, a dropout strategy is implemented following the LSTM layer. This strategy, governed by a pre-specified dropout probability p, involves the arbitrary deactivation of nodes, strengthening the model's generalization capacity. Nodes unaffected by dropout are then passed to the subsequent layer—a fully connected stratum containing n_{fc} nodes.

The secondary hidden layer employs the Rectified Linear Unit (ReLU) activation function on data points derived from the fully connected layer. This imparts essential nonlinearity, preparing the model to discern complex patterns within the dataset. Predictions are formulated based on the output of this layer.

In summary, our MicroGrowthPredictor model for predicting microbiome dynamics integrates purpose-built layers, each designed to interpret the nuanced temporal dynamics in microbiome data. The architecture begins with an input layer hosting n_{taxa} taxa-representative nodes, transitioning into a single-layer LSTM with n_h hidden states.

While not explicitly detailed, we presume that the LSTM layer retains the conventional composition of LSTM cells, including input, forget, and output gates for effective information transfer. This structure is instrumental in enabling the model to learn and preserve long-term dependencies inherent in sequential data.

After the LSTM layer, a dropout technique with a designated probability p is applied to serve as a regularization mechanism, mitigating overfitting risks. Subsequently, a fully connected layer with n_{fc} nodes is introduced, culminating in a dense layer adept at capturing nonlinear interdependencies in the data. The final phase of the model incorporates a ReLU activation function, introducing nonlinearity and enhancing the model's complexity for detailed data interpretation. This stage is crucial in shaping the ultimate output, ensuring precise and fluid predictions amid the dynamically shifting landscape of microbiome data.

2.3. Training of the MicroGrowthPredictor Model

When working with time series data and using LSTM to predict the impact of environmental perturbations, cross-validation must account for the temporal dependencies inherent in the data. To ensure robust and accurate predictions, we employed a time series cross-validation method using a rolling-window approach. The dataset was divided into K consecutive folds without shuffling. For each fold k, the model was trained on the first k folds and tested on the $k+1$ fold, repeating until each fold served as the test set. This method ensures temporal dependencies are respected and avoids data leakage.

Evaluation metrics, such as the mean square error (MSE), were collected for each fold, and the average performance across all folds was computed to assess the model's robustness.

2.4. Prediction Interval

While traditional approaches to establishing confidence or prediction intervals in deep learning models face considerable challenges due to these models' nonlinearity and complex architecture, recent advancements have begun to pave the way for more robust solutions. One such advancement is the work of [13], in which the Monte Carlo (MC) dropout framework was leveraged to introduce a method that, while effective, leaves room for further refinement and application in new domains, such as microbiome data analysis.

Our research builds upon this foundational work, adopting the principle of stochastic dropouts after each hidden layer in the neural network architecture. However, we extend this concept by tailoring the dropout process and the subsequent interpretation of the model's outputs specifically to the characteristics and complexities of microbiome data. This adaptation not only allows for the theoretical interpretation of the model's output as a random sample from the posterior predictive distribution but also acknowledges the unique data behavior in microbiome studies.

The process of constructing an empirical distribution of predicted values by treating each prediction during the dropout as a sample from the underlying data distribution represents a nuanced approach in our study. It diverges from classical techniques by providing a window into the model's predictive capabilities and uncertainties specifically fine-tuned to the microbiome context, thereby bolstering the robustness of decision-making based on these predictions.

In our approach, we denote the test data sought to be predicted with the superscript $*$. The prediction interval's foundation lies in the conditional probability $p(y^*|x^*, \mathcal{D})$. This probability can be expressed as the integral of the product of $p(y^*|x^*, \theta)$ and $p(\theta|\mathcal{D})$ over the parameter vector θ, denoted as follows:

$$p(y^*|x^*, \mathcal{D}) = \int_\theta p(y^*|x^*, \theta) p(\theta|\mathcal{D}) d\theta.$$

θ represents the parameter vector of the deep learning model, and $p(\theta|\mathcal{D})$ corresponds to the posterior distribution. However, deriving an analytical form for $p(y^*|x^*, \theta)$ is generally infeasible. To overcome this challenge, an approximation technique utilizing a variational distribution denoted as $q(\theta)$ is proposed in ref. [14]. Consequently, the following approximation is obtained:

$$p(y^*|x^*, \mathcal{D}) \approx \int_\theta p(y^*|x^*, \theta) q(\theta) d\theta \approx \frac{1}{K} \sum_{k=1}^{K} p(y^*|x^*, \hat{\theta}_k), \quad (1)$$

where $\hat{\theta}_k \sim q(\theta)$. This final approximation, achieved through the sampling of $\{\hat{\theta}_k\}_{k=1,\ldots,K}$ from the variational distribution $q(\theta)$, employs the technique of Monte Carlo integration. Moreover, this approximation is equivalent to implementing the Monte Carlo dropout algorithm introduced in [13]. In essence, for a given testing data point (x^*, y^*), the predictive output y^* is evaluated multiple times at x^* with random dropout of nodes, and the resulting empirical distribution serves as an estimate of $p(y^*|x^*, \mathcal{D})$. Prediction intervals

capture the variability originating from two primary sources: model uncertainty (η_1) and inherent noise (η_2).

The following steps outline the process: For each individual data point x^* in the testing set, calculate the corresponding output $\tilde{y}^*$ by randomly dropping out each node with a given dropout probability p. Repeat this process B times to obtain a large number of predicted values $\tilde{y}^*$, each of which varies due to the random dropout of nodes. Next, compute the model uncertainty η_1 by calculating the average squared difference between each predicted value $\tilde{y}_i^*$ and the mean of all predicted values $\bar{\tilde{y}}^*$. This is done using the formula $\eta_1 = \frac{1}{B}\sum_{i=1}^{B}(\tilde{y}_i^* - \bar{\tilde{y}}^*)^2$. To quantify the inherent noise in the predictions, calculate the average squared difference between each predicted value $\tilde{y}_j^*$ and its corresponding true value y_j^* using the test dataset of length V. This gives us the inherent noise η_2, computed as $\eta_2 = \frac{1}{V}\sum_{j=1}^{V}(\tilde{y}_j^* - y_j^*)^2$. Combining the model uncertainty and inherent noise, compute the overall uncertainty η as the square root of the sum of η_1 and η_2, i.e., $\eta = \sqrt{\eta_1 + \eta_2}$. Finally, determine the upper and lower bounds of the prediction interval by adding and subtracting $z_{\alpha/2}$ times η from the mean predicted value $\bar{\tilde{y}}^*$. Here, $z_{\alpha/2}$ represents the z-score associated with the desired confidence level $(1-\alpha)100\%$. A formal algorithm is listed in Algorithm 1.

Algorithm 1: LSTM Neural Network and Prediction Interval.

Require: x, y, x^*, y^*, p, t, n_h, n_{fc}
Ensure : θ, U, L
repeat
  $z_1 \leftarrow x$ from LSTM layer with t and n_h;
  $z_2 \leftarrow z_1$ through random dropout with p;
  $z_3 \leftarrow z_2$ from the fully connected layer with n_{fc} nodes;
  Apply `ReLU` to z_3;
  $\hat{y} \leftarrow z_3$ from the output layer;
  Evaluate $\hat{y}$ with y;
  Update θ for model m_θ;
until *the last epoch*;
for $i = 1$ *to* B **do**
  $\tilde{y}_i^* \leftarrow m_\theta(x^*)$ with random dropout;
end
Compute $\bar{\tilde{y}}^*$ and η;
$U, L \leftarrow \bar{\tilde{y}}^* \pm z_{\alpha/2} \times \eta$;

2.5. Parameter Tuning

To optimize the performance of our MicroGrowthPredictor model, we employ a two-step tuning process.

In the first step, we preselect the numbers of hidden units in the LSTM layer (n_h) and fully connected layer (n_{fc}) based on preliminary experiments. We then explore different combinations of the dropout probability (p) and the sequence length (T), which represents the number of previous data points used as features for prediction. The model's performance is evaluated by calculating the mean square error (MSE) on a separate testing dataset, and we select the combination of p and T that minimizes this error.

Once the optimal dropout probability and sequence length are determined, we proceed to the second step, where we fine-tune the numbers of nodes in both the LSTM and fully connected layers. For each architectural combination, we train the model multiple times with different initializations to account for variations introduced by random dropout and initial weight settings. We calculate the MSE for each training run and select the architecture that results in the lowest error on the testing dataset.

This rigorous tuning process ensures that our MicroGrowthPredictor model is optimally configured for the specific dataset under consideration, thereby enhancing its predictive performance.

3. Results

In this study, we employ the MicroGrowthPredictor model and associated tuning procedure to two distinct datasets: the antibiotic ciprofloxacin (Cp) dataset from [15] and the artificial gut dataset detailed in [16]. Both datasets offer insights into the microbiome's temporal dynamics under varied environmental perturbations.

3.1. Ciprofloxacin Dataset

Reference [15] underscores the significant alterations imposed on the human gut microbiota composition and diversity due to recurrent antibiotic treatments. This research involved an in-depth surveillance of bacterial communities in the distal gut across three subjects (D, E, and F). Stool samples were periodically gathered over ten months, summing to 52–56 samples per individual. Within this timeframe, each subject was administered two separate 5-day regimens of the antibiotic ciprofloxacin (Cp), spaced 6 months apart. Intense sampling—daily over two 19-day spans coinciding with each Cp course—provided a detailed perspective of the microbiome during antibiotic exposure. Outside these windows, samples were acquired either weekly or monthly, capturing the microbial composition in the absence of treatment.

For illustrative purposes, we focus on subject *D*. Our optimization process involves generating a contour plot of the mean square error (MSE) against varying values of the dropout probability p and the number of time steps. Figure 2 visualizes this relationship, guiding our selection of an optimal combination for refining the MicroGrowthPredictor model. The contour plot of the mean square error is plotted with dropout probability p on the x-axis and the number of time steps on the y-axis. In the contour plot, the darker the shading is, the smaller the error is. We use an optimization function to identify the best combination of dropout probability and sequence length.

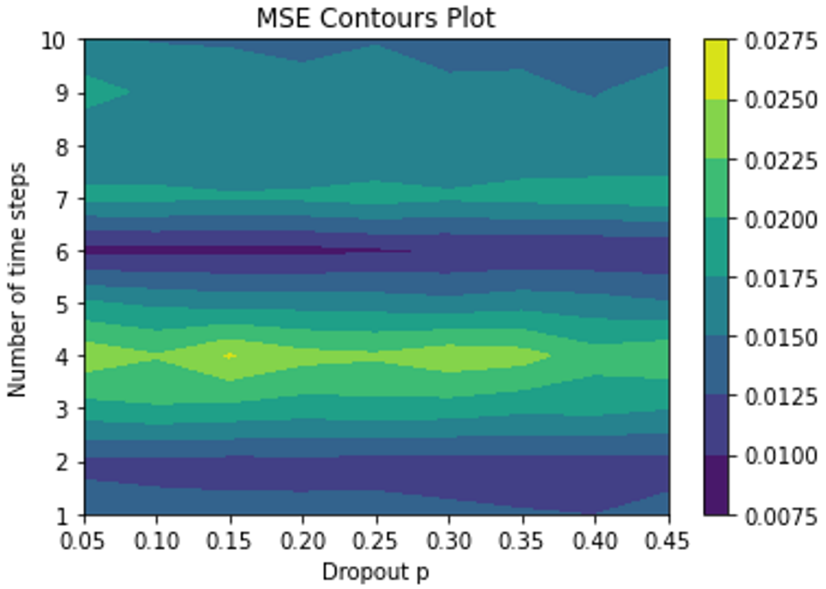

Figure 2. Contour plot of the mean square error over p and t for Subject D EU766613: The darker the contour plot is, the smaller the error is. We can identify the best combination of dropout probability and sequence length.

Subsequently, our focus shifts to ascertaining the optimal node count for both the LSTM and fully connected layers, as depicted in Figure 3. The x-axis represents the number of hidden states in the single LSTM layer, and the y-axis represents the number of nodes in the fully connected layer. Different combinations result in changes in the mean square error value. The contour plot provides a direct representation of the smallest MSE, indicated by the darkest area on the figure.

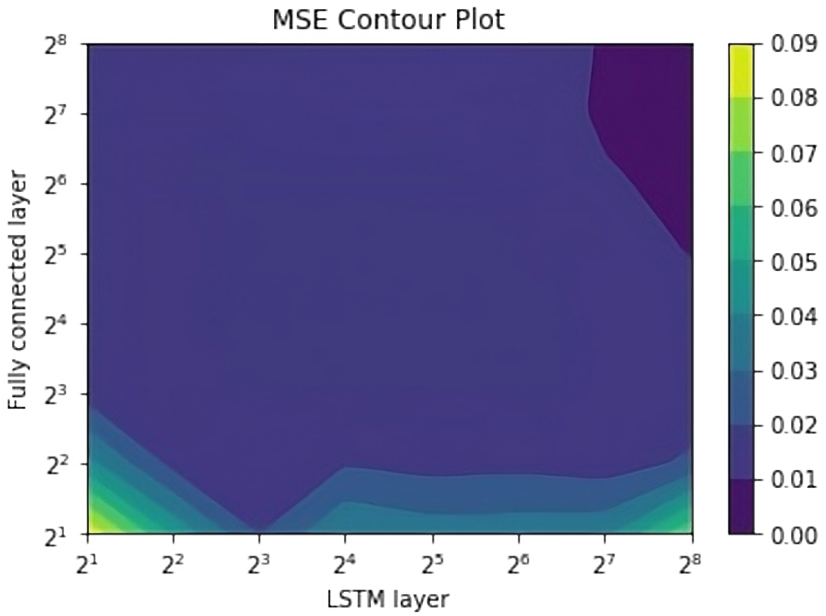

Figure 3. Contour plot of the mean square error over n_{fc} and n_h for Subject D EU766613: The x-axis represents the number of hidden states in the single LSTM layer, and the y-axis represents the number of nodes in the fully connected layer. With different combinations, the mean square value changes. The contour plot basically gives us a direct impression of the smallest MSE, which is represented by the darkest area on the figure.

Through this systematic exploration, our objective remains consistent: pinpointing a configuration that minimizes testing dataset error, thereby enhancing the efficacy of the MicroGrowthPredictor.

It is essential to note that in our training dataset, we included two-thirds of the observed data, aiming to provide a robust foundation for the model. Notably, there were two data points corresponding to antibiotic administration for each patient. One of these points was included in the training set, while the other was reserved for the prediction set. Based on our observations, the reaction to the first antibiotic exhibited a delayed response compared to the second one. This observation explains why our predicted data demonstrate a delayed pattern in Figure 4.

The temporal insight provided by the visualization of the trajectories of relative microbiome abundance was crucial for understanding the dynamics of microbiome changes and their potential implications for host health. To elucidate further, Figure 4 presents an analysis and prediction of the relative abundance of Bacteroid EU766613 for subject D, utilizing the aforementioned optimal parameters. The antibiotic administration intervals are denoted by a blue dotted vertical line, while the red dotted demarcation segregates the training and testing periods. In our study on repeated antibiotic treatments, we prioritize the inclusion of extensive data on antibiotic interventions to bolster the predictive power of our model. This data-driven approach enhances the accuracy of subsequent treatment predictions, offering a critical tool in combating antibiotic resistance through informed, strategic application of therapies.

The visualization underscores the MicroGrowthPredictor model's capability in understanding microbiome dynamics and formulating predictions rooted in these identified patterns. This is achieved with the model trained over 200 epochs using a learning rate of 0.001. Furthermore, the mean square error loss for the training data is 0.00081, and for the testing data, it is 0.01021.

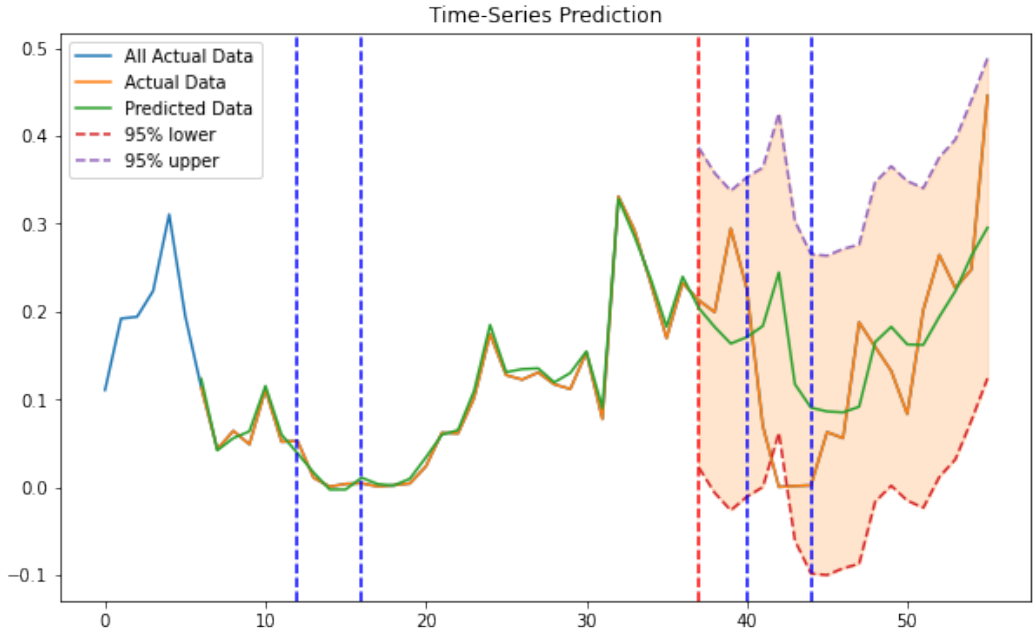

Figure 4. Trajectories of relative abundance of Bacteroid EU766613 for subject D. The chosen parameters are $p = 0.05$, $t = 6$, $n_h = 256$, and $n_f = 256$. The blue vertical bands represent the two antibiotic treatment periods, and the red dotted line splits the data into training and testing.

3.2. Artificial Gut Dataset

The dataset provided by [16] comprises time-resolved readings of the gut microbiota sourced from an artificial human gut. These data, captured both daily and hourly, originate from an artificial gut constructed using continuous-flow anaerobic bioreactor systems, ensuring an accurate representation of human gut microbiota dynamics. Over a month, four ex vivo vessels, each initialized with an identical human fecal inoculum, were cultured. To ensure experimental fidelity, key parameters like pH, temperature, media input rate, and oxygen concentration were stringently maintained. On Day 23, microbial dynamics received a deliberate stimulus via the introduction of a *Bacteroides ovatus* bolus, a strain isolated from the stool donor. However, unforeseen disruptions to the feed supply in two vessels between days 11 and 13 introduced unplanned microbial variations. Notably, we observed significant changes in the population of Rikenellaceae, a family of bacteria known for its role in the human gut microbiome. Rikenellaceae are involved in the breakdown of complex carbohydrates and play a crucial part in maintaining gut health and metabolic functions. The changes in this population are particularly interesting because they can provide insights into how disruptions in diet and microbial introductions influence gut microbiota stability and function.

In this example, the first vessel serves as our training set, while the second vessel functions as our testing set. Our MicroGrowthPredictor tool, configured with an optimal dropout probability (p) of 0.25, utilized the preceding five time points to identify four parameters and achieve optimal predictions. The fully connected layer was equipped with 256 nodes, and the LSTM layer comprised 128 nodes. The model underwent training for 800 epochs. The mean square error for the training data is 0.00057, and for the testing data, it is 0.01456. Without using our predictive model, a generalized additive model (GAM) has an MSE of 0.0048 for the training data, which is approximately 8.42 times higher. The performance on the testing data is significantly worse, so it is not included for comparison.

The trajectories of relative microbiome abundance visualized in Figure 5 provide critical insights into the dynamics of micribome changes over time. The blue line in Figure 5 represents all the actual data, while the orange line is highlighted simultaneously with the predicted line (green). In our deep learning algorithm, we utilized the previous five time points to predict the next one. Notable variations, especially for Rikenellaceae,

were observed due to the disruption of the first two vessels between days 11 and 13. These visualizations reveal significant shifts in microbial populations, underscoring the model's accuracy in capturing temporal changes. The observed patterns align with our statistical analyses, confirming substantial changes in microbiome composition during perturbations. This alignment strengthens our understanding of microbiome dynamics and their responses to experimental conditions.

Contrary to the notion that more data lead to better predictions, our experiment involving additional training vessels (including 1, 3, and 4) for predicting the second vessel yielded a mean square error for testing of 0.0265, almost double the original testing error. Interestingly, the correlation between the predicted value and the actual value for testing vessel 2 is 0.70, which is 18% higher than the case when we include vessels 1, 3, and 4. This suggests that a careful balance in the selection of training data is crucial for achieving accurate predictions.

In the realm of scientific studies, there is often a prevailing belief that incorporating more datasets or information for training leads to increased accuracy. However, a critical consideration arises when the environment in which the model is trained significantly differs from the environment in which it will be applied for testing. This disjunction in environmental conditions can introduce unforeseen disruptions and challenges.

In our experiment, the initial assumption that more training data (including vessels 1, 3, and 4) would inherently improve predictions was challenged by the observed results. The disruptions to the feed supply in the first two vessels between days 11 and 13 created variations in the microbial dynamics that were not adequately captured by the additional training data. The unforeseen disruptions underscore the importance of aligning the training data with the conditions and disturbances expected in the testing environment.

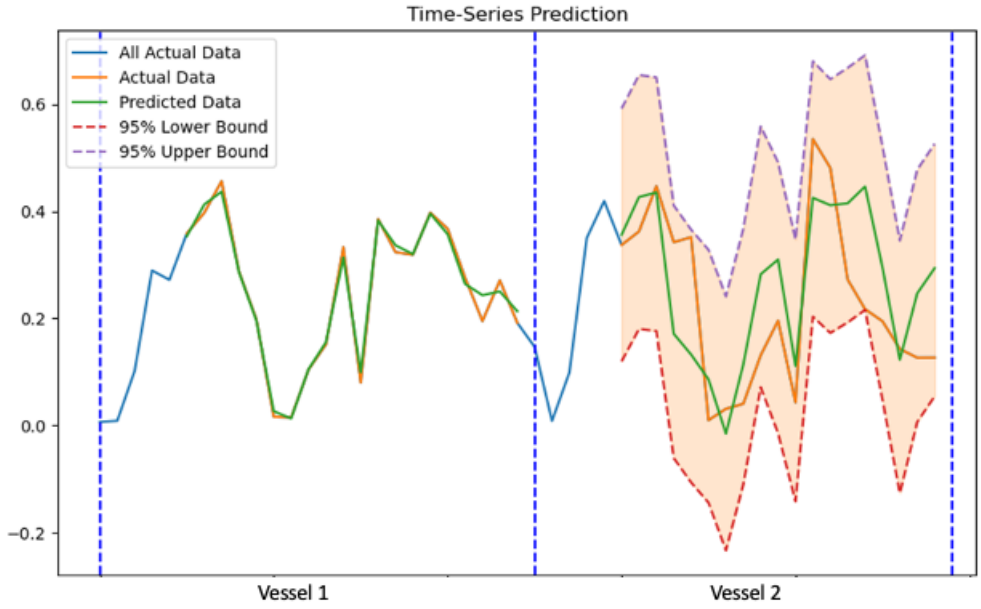

Figure 5. Trajectories of the relative abundance of Rikenellaceae in Vessels 1 and 2. The entire trajectory of Vessel 2 is predicted by the MicroGrowthPredictor model trained on data from Vessel 1. Confidence intervals are provided for the testing data of Vessel 2. In this experiment, the optimal dropout probability p of 0.25 was used. The model utilized the preceding five time points to identify four parameters and achieve optimal predictions. The fully connected layer was equipped with 256 nodes, and the LSTM layer comprised 128 nodes. The model underwent training for 800 epochs.

While it is tempting to assume that a larger sample size will inherently lead to better predictions, the key lies in the relevance of the training data to the testing conditions. In cases where the testing data involve different environmental interruptions or perturbations,

blindly including diverse datasets may lead to suboptimal predictions. The delicate balance between the quantity and relevance of training data becomes crucial in ensuring the model's adaptability to real-world scenarios.

4. Discussion

The MicroGrowthPredictor model leverages knowledge from observations that repeated antibiotic treatment disrupts the gut microbial community, affecting the diversity and abundance of specific bacterial groups. By analyzing the data, the model accurately predicts how the microbiome will change over time in response to antibiotic perturbation. This provides a deeper understanding of antibiotics' impact on the gut microbiota and potential human health implications.

Additionally, the model's versatility is demonstrated through its application to an artificial gut dataset. Insights drawn from this controlled environment show MicroGrowthPredictor's adaptability to diverse microbiome systems. The artificial gut dataset validates the model's predictive capabilities under specific conditions, highlighting its proficiency in capturing intricate temporal dynamics. This positions the model as valuable for understanding antibiotics' effects and broader applications in environmental sciences, healthcare, and biotechnology.

Our method addresses real-world problems where limited sample sizes are a constraint due to logistical, ethical, or financial challenges. By developing and validating methods that perform well with limited data, we provide practical solutions for such situations. Unlike many black-box models, our approach offers clear insights into how environmental perturbations influence microbial populations over time, crucial for understanding biological processes and designing targeted interventions. Specifically, we discuss its potential to contribute to personalized treatment plans by predicting individual responses to dietary changes, antibiotic treatments, and probiotic interventions.

In summary, MicroGrowthPredictor emerges as a potent tool surpassing traditional modeling approaches. The model, driven by insights derived from data rather than direct integration of knowledge, incorporates LSTM networks with confidence interval estimation to contribute to a holistic comprehension of microbiome dynamics. The model's successful applications to both real-world human gut microbiota and artificial gut datasets underscore its efficacy and potential impact. We foresee MicroGrowthPredictor playing a pivotal role in advancing microbiome research, offering valuable insights, and contributing to well-informed decision-making across various fields.

Author Contributions: Conceptualization, Y.-H.Z.; methodology, G.S. and Y.-H.Z.; validation, G.S. and Y.-H.Z.; writing—original draft, G.S. and Y.-H.Z.; writing—review and editing, G.S. and Y.-H.Z.; visualization, G.S. and Y.-H.Z.; supervision, Y.-H.Z.; project administration, Y.-H.Z.; funding acquisition, Y.-H.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by U.S. Environmental Protection Agency, grant number 84045001, National Institute of Health P30ES025128 and the Engineering Research Centers Program of the National Science Foundation under NSF cooperative agreement No. EEC-2133504.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be constructed as a potential conflict of interest.

References

1. Altveş, S.; Yildiz, H.K.; Vural, H.C. Interaction of the microbiota with the human body in health and diseases. *Biosci. Microbiota Food Health* **2020**, *39*, 23–32. [CrossRef] [PubMed]
2. Smith, J.; Johnson, M. Microbiome dynamics under environmental perturbations. *J. Microbiome Res.* **2022**, *10*, 123–145.
3. Brown, E.M.; Sadarangani, M.; Finlay, B.B. The role of the immune system in governing host-microbe interactions in the intestine. *Nat. Immunol.* **2013**, *14*, 660–667. [CrossRef] [PubMed]
4. Candela, M.; Biagi, E.; Maccaferri, S.; Turroni, S.; Brigidi, P. Intestinal microbiota is a plastic factor responding to environmental changes. *Trends Microbiol.* **2012**, *20*, 385–391. [CrossRef]

5. Uhr, G.T.; Dohnalová, L.; Thaiss, C.A. The dimension of time in host-microbiome interactions. *mSystems* **2019**, *4*, e00216-18. [CrossRef] [PubMed]
6. Willing, B.P.; Russell, S.L.; Finlay, B.B. Shifting the balance: Antibiotic effects on host–microbiota mutualism. *Nat. Rev. Microbiol.* **2011**, *9*, 233–243. [CrossRef] [PubMed]
7. Brown, E.; Williams, D. Predictive modeling of microbiome growth using LSTM networks. *J. Comput. Biol.* **2021**, *45*, 321–335.
8. Ching, T.; Himmelstein, D.S.; Beaulieu-Jones, B.K.; Kalinin, A.A.; Do, B.T.; Way, G.P.; Ferrero, E.; Agapow, P.M.; Zietz, M.; Hoffman, M.M.; et al. Opportunities and obstacles for deep learning in biology and medicine. *J. R. Soc. Interface* **2018**, *15*, 20170387. [CrossRef] [PubMed]
9. Medsker, L.R.; Jain, L. Recurrent neural networks. *Des. Appl.* **2001**, *5*, 2.
10. Graves, A.; Graves, A. Long short-term memory. In *Supervised Sequence Labelling with Recurrent Neural Networks*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 37–45.
11. Yu, Y.; Si, X.; Hu, C.; Zhang, J. A review of recurrent neural networks: LSTM cells and network architectures. *Neural Comput.* **2019**, *31*, 1235–1270. [CrossRef] [PubMed]
12. Zhou, Y.H.; Gallins, P. A review and tutorial of machine learning methods for microbiome host trait prediction. *Front. Genet.* **2019**, *10*, 579. [CrossRef] [PubMed]
13. Zhu, L.; Laptev, N. Deep and confident prediction for time series at uber. In Proceedings of the 2017 IEEE International Conference on Data Mining Workshops (ICDMW), Orleans, LA, USA, 18–21 November 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 103–110.
14. Gal, Y.; Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Proceedings of the International Conference on Machine Learning, PMLR, New York, NY, USA, 20–22 June 2016; pp. 1050–1059.
15. Dethlefsen, L.; Relman, D.A. Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 4554–4561. [CrossRef] [PubMed]
16. Silverman, J.D.; Durand, H.K.; Bloom, R.J.; Mukherjee, S.; David, L.A. Dynamic linear models guide design and analysis of microbiota studies within artificial human guts. *Microbiome* **2018**, *6*, 202.