Developing a Deep Learning Model to Predict Systemic Lupus Erythematosus Methylation Gene Expression Levels

Ahana Mangla

Received March 15, 2024
Accepted May 29, 2024
Electronic access June 15, 2024

Systemic Lupus Erythematosus (SLE) is an autoimmune disorder characterized by inflammation in areas of the body such as the joints, skin, brain, and lungs. Unfortunately, its progression is hard to track on a genetic level. This requires healthcare workers to take a symptom-based approach, making treatment difficult. To aid in genetic understanding of the condition and reveal patterns in SLE's effect on the immune system, we developed a deep learning model that can accurately predict the diseased states of methylated healthy cells for SLE. We utilized preexisting data from a database (ADEx). Datasets were selected to ensure balanced representation between SLE patients and healthy individuals. Data formatting followed rigorous criteria, focusing on methylated formats to maintain uniformity, and subsequent model construction employed a feedforward neural network with ReLU activation and dense output layers. The model was trained using an adaptive learning rate (determined using Mean Absolute Error). Overall, the analysis of the model's performance indicated a fairly high degree of precision. Nearly 60% of predictions fell within 10% of actual values, and almost 90% fell within an error of 20%. However, a small portion of the dataset (1.9%-2.5%) displayed significant errors. These outliers suggest potential unpredictability in very specific gene expressions, possibly due to unique immune responses in SLE. The model was very accurate for T-cells but tended to struggle with Monocytes. Overall, the findings underscore the model's efficacy in capturing the nuanced dynamics of SLE progression and its potential as a marker for autoimmune disease development.

Introduction

Systemic Lupus Erythematosus (SLE), one of the most predominant types of lupus, is an autoimmune disorder characterized by inflammation of multiple organs such as joints, skin, brain, lungs, kidneys, and blood vessels (CDC). Despite lacking a definitive cure, SLE management involves treatments and lifestyle adjustments to alleviate symptoms (CDC). In biomedical research, the application of machine learning techniques has emerged as a pivotal approach to gaining deeper insights into biomedical conditions in general[1]. By analyzing extensive datasets derived from clinical observations and molecular profiling, machine learning facilitates the identification of disease patterns, prediction of disease progression, and customization of treatment strategies tailored to individual patients.

In the context of SLE, machine learning algorithms have the potential to enable the discovery of correlations and the development of predictive models for early detection and personalized intervention. Through interdisciplinary collaboration among computational scientists, clinicians, and biologists, deep learning can drive innovation in bioinformatics, promising more effective management and improved outcomes for individuals affected by this chronic autoimmune condition.

This research aims to create a deep learning model for predicting values of diseased methylated CpG (cytosine-phosphate-guanine) sites in a patient's genome using healthy baseline data. These sites are markers on various different genes within the T-cells, B-cells, and monocytes of the human body[2]. Their expression levels in methylated data seem to change significantly when comparing healthy versus diseased data. Methylation is a molecular process in which a methyl group is transferred into the DNA of the specified sample in order to influence its output[3,4]. It does so by altering DNA's interactions with both chromatin proteins and specific transcription factors (proteins responsible for the rate of communication between DNA and RNA)[5]. Methylated datasets specifically have been especially relevant in the field of genetic studies of lupus because the process introduces autoreactivity[6]. By utilizing the average values of healthy CpG sites in the training data, the model aims to offer insights into the expected trends of diseased CpG sites. Specifically, it predicts the values for sites that have been shown to change to some degree when SLE is introduced to the system.

The hope for this study is to uncover the underlying mechanisms that drive SLE progression, contributing to the refinement of predictive models and a deeper understanding of autoimmune disorders. This can shed light on the underlying genetic and cellular mechanisms of this complex autoimmune disease in the context of understanding the progression of disease.

Ultimately, this research seeks to bridge the gap between genomic data and clinical applications, with the potential to

improve early detection and personalized treatment based on an individual's genomic profile and to also open the possibility of expanding datasets in order to further research. This study aims to become one of a growing set of developments in the field of omics and machine learning, and ultimately serves as a foundation for better, targeted care for patients with autoimmune conditions.

Literature Review

Systemic autoimmune diseases pose a significant challenge due to their unpredictable nature and reliance on symptom-based diagnosis. However, recent advancements in omics research offer a powerful lens to illuminate the complexities of these conditions, including but not limited to the condition Systemic Lupus Erythematosus (SLE).

Omics studies encompass analyses of biological molecules such as genes, proteins, and metabolites. In systemic autoimmune diseases, these studies[7] have emerged as powerful tools. One notable study, focused on SLE and other systemic autoimmune diseases, utilized a multi-omics approach[7]. Researchers employed transcriptome, epigenome, genome, cytokine, and metabolome analyses to gain a deeper understanding of these conditions. This comprehensive strategy not only improved diagnosis and reclassification of lupus by uncovering molecular subtypes, but also opened doors for personalized treatment based on an individual's unique biological makeup.

Beyond diagnosis and treatment, omics investigations are shedding light on the regulatory networks within the body, pinpointing potential biomarkers, and even suggesting therapeutic targets for treatment of these lifelong conditions. This paves the way for the development of more effective and targeted interventions that can improve a patient's quality of life.

By integrating data from various omics platforms, scientists can create a more holistic picture of the molecular landscape. This approach offers a deeper understanding of how these diseases progress and vary from person to person.

A groundbreaking study in inflammatory myopathies (muscle diseases) employed machine learning to predict patient response to different immunoglobulin treatments[8]. This innovative approach not only showcased the power of machine learning in prediction, but also highlighted the potential of artificial intelligence in the field. By analyzing clinical factors and disease activity alongside omics data, these computational approaches hold the promise of personalized treatment plans, ultimately improving clinical outcomes.

In conclusion, omics studies, coupled with machine learning techniques, are revolutionizing our understanding and management of systemic autoimmune diseases. They provide a transition from symptom-based diagnosis towards a comprehensive view of the molecular underpinnings of these conditions. The field can revolutionize disease management and improve the lives of countless individuals affected by autoimmune diseases.

Methodology

Data Collection and Cleaning

In this study, we harnessed preexisting data from an online data analyzer and database called ADEx[9]. This dataset served as the foundation for this research, enabling us to train and test the deep learning model effectively. The data itself was obtained through a widespread collection and analysis of patient blood. T-cells, B-cells, and Monocytes were all isolated from freshly collected peripheral blood and were separated in Dynabeads[6]. Dynabeads are extremely small superparamagnetic particles that are commonly used to study cell-level biological properties[10]. For the purposes of this data collection, immunoprecipitation was applied, a process in which the beads help extract cells of specific types using antigens[10]. The three T-cell varieties of naive, regulatory, and memory were all collected using a distribution of various antibodies. All 6 cell types had their DNA isolated from the blood for analysis using QIAGEN DNAeasy kits - a product designed for DNA extraction[6,11]. It was discovered that CpG's (cytosine-phosphate-guanine) relevance to the expression of Systemic Lupus Erythematosus (SLE) remained stable across their two-year study of the patients, with a t-test false discovery rate of <0.5. The final selection of datasets was based on the consideration of the widest amount of patient data and a balanced distribution between individuals with the SLE and healthy individuals - i.e. maintaining enough of each such that contrasting trends between both could be identified and that intrinsic differences would remain relatively inconsequential regarding the final goals of the model.

In the data collection phase of this study, we adhered to a meticulous process to ensure the reliability and validity of the datasets used for analysis. A deliberate criterion for dataset selection was the expression format; datasets were chosen in a methylated format to maintain uniformity and relevance to the research objectives regarding SLE expression in methylated cases. Furthermore, the decision to utilize datasets from the same study was made to mitigate potential extenuating factors that could introduce bias into the analysis (i.e., batch effects). This choice was informed by the need for a robust and unbiased representation of CpG sites across a range of patients. In the subsequent data collection procedures, we constructed the foundation of the deep learning model. The next step was to prepare the input data for the model, enabling it to generate the desired outputs accurately. The metadata and raw data files obtained from the designated source were processed into the desired input through the programming language Java. It is relatively easy to read information into an array within the program and manipulate it for the purposes of this research. Here, Java

programming language was used but it is possible to use other programming methods. Some of the data functionality was later moved to Python as the model creation was occurring within the same codebase. This was to further streamline the technical and logistical flow of the code.

Site Name	x1	x2	x3	x4	x5	...	x1620
Averaged Value	y1	y2	y3	y4	y5	...	y1620

Table 1 A representation of the final input data that was fed into the deep learning model

These data files were arranged into an array, with patients represented as rows and CpG levels as columns. Note that this array contains not just one cell type, but expands across T-cells, B-cells, Monocytes, Regulatory T-cells, Memory T-cells, and Naive T-cells. Regarding the expression of SLE, the biological differences between these categories of cells should not have a significantly different effect on the actual levels of methylation that they correspond with. This means that, rather than conducting per-cell analysis in order to identify trends and predict diseased states, the model can instead be used for a bulk analysis that is aggregated across all cell types. This allows for higher generalizability of the model's use and demonstrates a wider progression within methylated cells. Building upon the findings from the study PLoS Genetics[6], we identified the CpGs of interest related to SLE and integrated them into the program, ensuring that only those CpGs were present in the input data.

The final input data the model is trained on, as depicted by Table 1, is a collection of averages of portions of these patients. For each of the 1680 CpG-cell pairs, all its respective patient data was condensed into one averaged value. These averages were taken in order to mitigate the effects of outliers on the training of the model. Since the model is already being trained upon different cell types and biologically different genetic sites, precautions needed to be taken regarding patient-to-patient data. By averaging the data for training, it is also possible to use larger subsets of the dataset, as the model can be tested on two entirely different averages, which is how we processed the input data. Essentially, for every CpG site, all the patients are taken and averaged into one value to be fed into the training of the model. The same is done for all six cell types for both healthy and SLE data. The final input is a set of two arrays, each with dimensions 1 by 1620. One is for healthy data and the other is that of diseased samples.

It is true, however, that the decision to take the average of all the patients may raise concerns as it is cloaking some of the variations that might be present in the data. One thing that is important to note is that this model is not analyzing just one specific CpG site: it is analyzing a collection of them and attempting to draw a pattern from its results. The main objective is not just to predict one value for a specific patient; it aims to show that the changes caused by SLE can be generalized across multiple different genetic values. This highlights the changes that occur in the body upon the development of SLE and the fact that changes are not sporadic in nature. They fall into a certain pattern. The ultimate decision to average was also a technical one. Had we been able to find a dataset with many more patients, we may have chosen not to average. In the case of a larger selection of patients, variations in the data would help the model as they would make its processes less formulaic. Unfortunately, there are only 50 patients for both the healthy and the SLE group. This means that the variations, while important, of extreme values could affect the model in a way that makes it less accurate for the majority of the data. The effect of an outlier would inhibit the model, rather than help it. At some points within the data, we did come across null values. These usually took the form of the lack of 1 or 2. CpGs per patient, but the type of CpG that was not available varied per patient. Ultimately, for each averaged value in the dataset, only the non-null values were considered. For example, if a certain CpG was missing for 10 out of 55 patients, then the average for that CpG is calculated using the remaining 45 patients who did have a value for that CpG. Had the entire patient's information been deleted, there would be very few patients remaining, making the model's results less generalizable and overly specific. This would have led the results to be significantly less useful than they turn out with our chosen method.

Model Creation

Initial Learning Rate	**Decay**	**Multiplier**
0.001	0.9	0.12

Table 2 This table depicts the parameters of the adaptive learning rate that was used on the training of the model

In this study, the primary goal was to identify whether a deep learning model is trained on a set of healthy patient CpG sites in their methylated form and on a set of the methylated CpG sites of a patient infected with SLE, the model will be able to predict diseased states with low levels of loss because there is not only a genetic connection between specific CpG sites and the development of SLE, but also the progression of the disease can be generalized and applied to a broad population.

The initial model configuration was based on a basic feed-forward neural network. The coding was done with the use of Jupyter Notebook. Using the libraries Numpy, Pandas, and Keras, the model was coded entirely in Python. Matplotlib was used to create the result and analysis graphs. Leveraging the information pertaining to healthy versus diseased CpG site levels, the model was trained to make predictions based on their average values. Its layout is depicted in Figure 1. The design

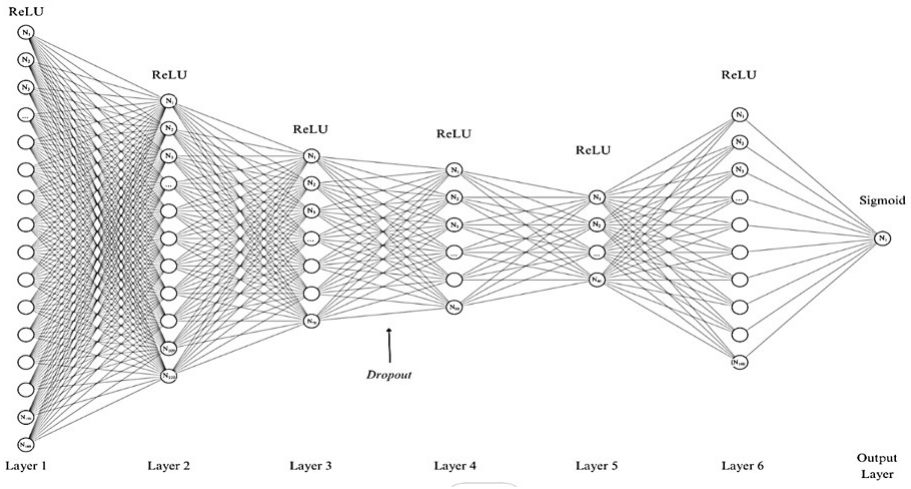

Fig. 1 Model creation

is that of a standard feedforward network, in which a variety of data is inputted, and one value is output. The difference in size between all the layers is important to avoid creating too many links from one group to the next, a process which could overfit the data. In the case of this model, it is also important to bring the information taken from all 1620 CpGs down to one final value, which would be the prediction for the diseased state. For the network to function accurately, the layers must gradually shrink in size to reach their final value. Specific sizes were chosen primarily through a process of trial and error, but the general shell of the diagram was kept similar. Every layer except the third is a Dense layer, in which information is input and a different set of values is output. The third is a Dropout that works upon the previous Dense layer, in which 10% of the values are randomly dropped to introduce randomness, a choice that helps make the model more robust because it prevents the phenomenon of overfitting. Had some sort of arbitrary change to the data not been made, the model may have become too good at predicting based solely on the training dataset, making it perform worse on unseen data, and therefore, ungeneralizable. When observing training values, a dropout of 10% performed the best, with other levels of dropout resulting in an increase of high error (likely due to loss of data or overfitting). The ReLU activation function is used for every layer other than the last. The last is put through the sigmoid activation to bring it to a value between 0 and 1. The figure was created using (NN SVG). Regarding the activation functions chosen, the rectified linear unit (ReLU) was used to enhance the network's capacity to capture non-linear relationships within the data. The final layer employed the sigmoid activation function, facilitating the output of probability distributions over the possible categories, allowing for accurate predictions and interpretations of the CpG site values associated with the SLE condition. The decision to use sigmoid may seem unorthodox, due to it being more prevalent in cases of binary classification. However, when looking at the input and training data for this model, all have been normalized to a float between 0 and 1. Since the goal of this model is to use these values to predict another normalized value between the 0 and 1, the sigmoid seemed like the best option to determine the final output. It would undeniably bring the value from the alterations it went through in the Dense layers to a reasonable output in the context of this model. The model itself was trained on 16 GBs of RAM, on an M1 processor chip GPU. Given the relatively small size of the dataset, it was possible to train locally.

We also experimented with different optimizers and layers. We decided to use the Adam optimizer in this model construction. This optimizer has proven efficient for the purposes of this model construction as it has proven to be empirically better than other optimization methods [12]. An adaptive learning rate is employed as depicted in Table 2, in which the current learning rate is manipulated through a function that decreases its size when the Mean Absolute Error (MAE) is low and increases it when it is high. It started with an ILR of 0.001. When performing with an error rate of lower than the previous loss, the learning rate decayed to 0.9 of what it originally was and multiplied by 1.12 when performing unreliably. This loss was determined using Mean Absolute Error (MAE).This calculation serves as the cost function of the model and is illustrated in Equation 1. The variable xi represents the actual value while the x represents the predicted value. The variable n stands for the number of outputted values. Given all of these factors, the model was trained on 200 epochs (iterations). Initial experiments brought us to around 100 before the MAE began to stagnate. However, after adding the adaptive learning rate, the loss only stagnates around 200 epochs, after which it does not decrease further.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|x_i - x|$$

Results

To assess the accuracy of our model, we primarily used the Mean Absolute Percentage Error (MAPE) as explained in Equation 2. At represents the actual value while Ft represents the predicted value. The variable n represents the number of predicted values.

$$\text{MAPE} = \frac{1}{N}\sum_{T=1}^{N}\left|\frac{A_T - F_T}{A_T}\right|$$

MAPE is optimal as a primary metric since it is a digestible representation of the outcome of this model. It is not squared, so it does not overly inflate the results of outliers, limiting the possibilities of skewed data due to biological differences. Its calculation as a percentage serves as a clear indication of the model's accuracy. This calculation is done for each CpG (cytosine-phosphate-guanine) site value in the test data, providing an idea of the model's accuracy across multiple CpGs and cell types.

The first method of validation was the standard one used with deep learning models: a percentage of the original batch of training data that was set aside and then tested against the newly trained model, from which the results are displayed in Figure 2. Each of these values was calculated in the same way as the training data, by taking the average of all available patients for that CpG.

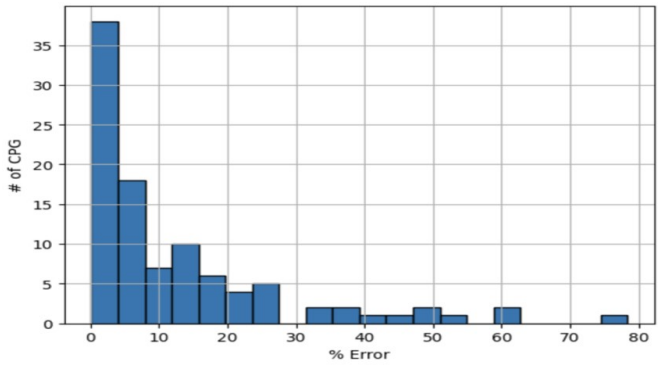

Fig. 2 Results

In the analysis of percentage loss within the context of this developed deep learning model, its predictive accuracy was tested in the above graph. The findings reveal that almost 60% of the predicted values align within a narrow margin of 10% in relation to the actual values for the given CpG sites. This observation is indicative of a high degree of precision in the model's predictions, portraying its efficacy in capturing the nuanced dynamics of diseased cell states from healthy cell states.

The validation dataset itself for Figure 3 was derived from the averaged values of the other section of the CpG values across different sets of patients. Note that because this is completely different type of validation, the error and general performance will appear slightly different. Figure 3 is NOT a per cell breakdown of Figure 2, it is a different dataset. Rather than using a fraction of the overall data, a new set of data was generated specifically for validation. Since each value in the training data was the average of all the patients provided, these new validation sets were generated using a small, random subset of the patients. For example, had the 245th value in the training data been the average of all 55 patients, the validation would be the average of a random 15 of them. This is a different method of validation than depicted in Figure 2. The CpG values were then tested against groupings of specific cell types, thereby refining the analysis to a more granular level. These subsequent graphs revealed an even more robust predictive capability.

When tested against the averaged values of the other 30% of the data as per each individual cell type, a majority of these values were within 20% of their expected values. The only two test charts showing more variation and more tests across the x-axis are those of the Monocytes and the B-cells. This can likely be attributed to the limitations imposed by the primary datasets being composed of different subclasses of T-cells. With a majority of the training being done on these biological processes, the model becomes slightly less reliable for significantly different types of cell processes. T-cells are lymphocytes by fighting against threats within the body itself (most famously, against cancerous cells). It makes sense that their activity is highly predictable, as their processes correspond directly with the nature of the autoimmune condition. B-cells are also lymphocytes, but they work against outside threats. This means they are slightly different cells. However, the model is still a reasonably accurate marker of the changes that occur when a patient becomes infected with SLE. This is an indication that these changes in genetic value are relatively predictable and are all strongly related to the development of this autoimmune condition and to each other. Monocytes, on the other hand, are leukocytes that mainly work to boost immune response. This biological difference in immune role likely played a part in the higher error rates for the monocytes. This is also why they did worse than B-cells, despite holding a similar fraction of the dataset.

Interestingly, when tested on data that was the averaged values of smaller amounts of patients, the amount of error/highly inaccurate values actually shrunk. This is contrary to what one might assume, since it would make more sense for outliers to affect a batch in which less patients were used for each value. For example, when tested against the averaged values of a more averaged subset of the data, 11.1% of them were at a signifi-

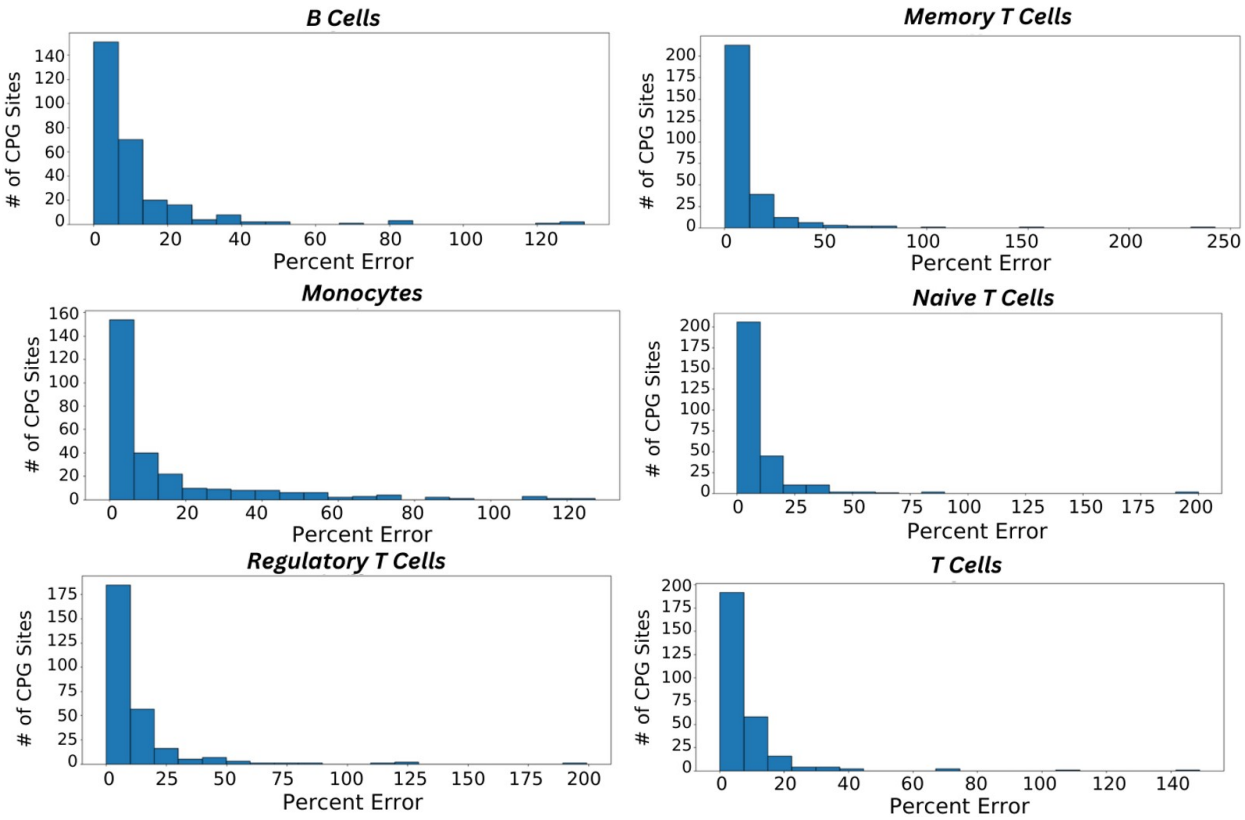

Fig. 3 A panel of the per-cell breakdown of the second method of validation

cant level of error. However, when tested against two smaller averages of the dataset, only 7.8% and 7.6% were significantly off. It is possible that, when taking a larger subset of the patient data to average, the average has a higher chance of including biological outliers and therefore is more difficult to predict.

Of course, it is important to analyze the values with high error. Across all three subsets of testing, 1.9% to 2.5% of the entire dataset is above an error rate of 50%. This translates to 32-43 values out of the total 1680. Observation of the genes linked to the CpGs that faced rates above 50% yielded a few commonalities. The gene BST-2 repeatedly had issues, especially within the B-cells. This gene is responsible for the growth and the development of B-cells (BST2)[13]. This would explain why, while the gene itself may have warped with the development of SLE (making it relevant to the model), its actual behavior did not align with the CpGs from other cells. Overall, the genes that did not do well were more related to other processes of their respective cells, rather than directly being involved in the immune response. This means that SLE may have changed them in more unpredictable ways. In the future, it is possible that these genes and CpGs could be removed. Within all the T-cells, there was one gene that was above a 50% error threshold, regardless of whether the data was tested on an average of 60%, 30%, or 20% of the data. It had an extremely high error rate, erroring 97 - 161% off from its actual value. This was cg24154161, located at position 32820421 on the TAP1 gene ("TAP1 Gene: MedlinePlus Genetics")[14]. TAP1 is responsible for making proteins that are responsible for a significant role specifically in the immune system. It works with another gene, TAP2, to transport proteins away from foreign invaders into the endoplasmic reticulum. Its role makes it significant to the change of the immune response due to SLE. It is possible that, as such a direct component of the immune response, it gets highly and more unpredictably changed when SLE is introduced to the system. Upon evaluating the general trend that many of the high error genes seem to follow, a few of them seem to be related to the function of TAP1 - They bring proteins to the surface in an immune response. Perhaps there is something unpredictable about this process that makes the values hard to gauge. It is also possible that these kinds of genes could be predicted given a dataset of related CpGs, and that they just follow a different trend than the rest of the CpGs do.

The trends outlined in the two validation metrics shown above highlight precision in the prediction of diseased methylated cell

states when examining the condition of SLE. They do reveal, however, that there are some processes where, although change may occur with the onset of the condition, this change is not genetically predictable and should not be considered in further studies. As this research study is a foundational one, the correlation provided by a majority of these CpG sites is the most important use for this model. Rather than being used directly on patients in clinical trials as it currently is, its contents and work can be optimized in further research.

Discussion

The model's performance provides not only a useful tool in the expansion and study of the genetic development of Systemic Lupus Erythematosus (SLE), but could also prove that the changes at these specific sites can be construed to a similar pattern.

One of the significant advantages of this approach is its potential to significantly reduce the cost and time associated with patient genetic testing, particularly when conducted at a large scale. Traditional genetic testing can be prohibitively expensive and time-inefficient, making it challenging to study and diagnose conditions like SLE comprehensively ("Systemic Lupus Erythematosus - SLE — Choose the Right Test.")[15]. By creating a system in which a patient might view a reasonably accurate progression of their disease and understand where the limitations of healthy CpG (cytosine-phosphate-guanine) levels lie, major amounts of time and funding can be saved in the diagnosis of the general population. Such a model can also significantly contribute to the understanding of the molecular landscape associated with SLE at a single-cell level. This deep learning model offers a cost-efficient alternative by providing a concrete visualization of how a patient's CpG sites might change as the disease progresses. By accurately categorizing methylated cells associated with the disease, researchers can prioritize specific cell types or molecular pathways for further investigation. This targeted approach could help the discovery of potential drug targets. It could also help in the development of precision medicine strategies tailored to individual patients.

Furthermore, the predictive model serves as a valuable tool for amassing a prediction of diseased-state CpGs from the data of healthy patients. It can be hard to obtain diseased patient data, so having a model that can accurately predict what these states look like in the data of healthy people can greatly increase our current databases. This can even be extended to future studies of the condition on mouse models, allowing scientists to run healthy mice data through a mass simulation.

Unfortunately, however, the model's abilities may be limited to T and B-cells. Due to the dataset being primarily biased towards T-cells, its predictions for the monocytes were more inaccurate. This means that, while the model does show a level of generalization and patterning for all these types of cells, it is still more precise regarding T-cells specifically. Since the monocytes being inaccurate is mainly due to their biological dissimilarities to the T and B-cells, a separate model might have to be trained for them. It is also possible that monocytes simply are not as relevant to this general trend and should be removed from consideration entirely, however there may be more of a benefit to analyzing them separately. Regarding future applications and expansions, its own training dataset and robustness could be expanded. It could be altered to include more omic datasets of various cell types, and a similar model could also be developed for RNA methylation. In general, the model's structure could likely be replicated to apply to multiple autoimmune conditions. Now that this paper has devised a formula for the ideal type of input file, the ADEx database could be used in order to generate models of a very similar format for conditions such as Multiple Sclerosis and Diabetes.

Conclusion

This research successfully developed a deep learning model for predicting values of diseased CpG (cytosine-phosphate-guanine) sites in a patient's genome, utilizing average values from healthy CpG sites in the training data. The focus on Systemic Lupus Erythematosus (SLE) provided valuable insights into the genetic and cellular mechanisms underlying this autoimmune disease. By bridging the gap between genomic data and clinical applications, this study contributes to refining predictive models and enhancing our understanding of autoimmune disorders. The potential impact lies in improved early detection and personalized treatment strategies based on an individual's genomic profile, marking a significant step towards advancing precision medicine.

Acknowledgements

Thank you for the guidance of Anish Roy, my mentor in the development of this research paper.

References

1 K. P. Kording, *The Roles of Machine Learning in Biomedical Science*, Www.ncbi.nlm.nih.gov, National Academies Press (US.

2 S. Acharjee, *Chapter Three - Mechanisms of DNA Methylation and Histone Modifications*, `http://www.sciencedirect.com/science/article/abs/pii/S1877117323000017`.

3 A. Razin and H. Cedar, *Microbiological Reviews*, **55**, 451–458,.

4 R. K. Perez, *Science*, **376**, year.

5 D. Latchman, *Int. J. Exp. Path*, **74**, 417–422,.

6 D. M. Absher, *PLoS Genetics*, **9**, 1003678,.

7 T. M., C. C. and A.-R. M. E, *Rheumatol. Oxf*, **56**, 78– 87.

8 M. Danieli, T. A., P. A., L. E., M. G. and A. A.

9 J. Martorell-Marugán, R. López-Dominguez and A. García-Moreno, *BMC Bioinformatics*, **22**, 343,.

10 A. Neurauter, M. Bonyhadi, E. Lien, L. Nøkleby, E. Ruud, S. Camacho and T. Aarvak, *Advances in Biochemical Engineering/Biotechnology*, **106. Springer**, 41–73.

11 Q. DNEasy, *DNEasy Blood Tissue Kits"*, https://www.qiagen.com/us/products/discovery-and-translational-research/dna-rna-purification/dna-purification/genomic-dna/dneasy-blood-and-tissue-kit.

12 D. Kingma and J. L. Ba, *ADAM: A METHOD for STOCHASTIC OPTIMIZATION*.

13 BST2, *bone marrow stromal cell antigen 2"*, https://www.ncbi.nlm.nih.gov/gene/684.

14 *Medlineplus.gov*.

15 *Arupconsult.com*.