PROJECT REPORT

MACHINE LEARNING FOR REGULATORY GENOMICS
SOSE 2023

18.04.2023 - 18.07.2023

Group members:

Mahima Arunkumar (mahima.arunkumar@tum.de)
Berfin Erdoğan (berfin.erdogan@tum.de)
Ufuk Demir (ufuk.demir@tum.de)
Joshua Günther (joshua.guenther@tum.de)

Supervisor:

Johannes Hingerl

Contents

1 Introduction and Motivation

Birds exhibit a wide range of sizes, shapes, colors, and behaviors. This diversity has captured the attention of researchers for centuries. By expanding our understanding of avian biology, we hope to get a deeper insight into bird evolution, adaptation, ecology or even conservation. Specifically, unraveling the genetic code that defines this plethora of avian coloration is of particular interest for this project. We leveraged the deep learning model CLIP (Contrastive Language-Image Pretraining) from OpenAI in order to analyze the intricate relationship between bird images and their respective genetic information. The CLIP image embeddings were utilized as input to train and evaluate multiple classifiers with the aim to predict the presence or absence of certain coloration genes of interest. The performance of these classifiers was then thoroughly inspected and benchmarked.
In the following sections, we will delve into the details of our approach, discuss the data used, present the results of our explorative analysis, and highlight the implications of our findings.

2 Methods

2.1 Data overview

In the project "Unveiling Bird Morphology Evolution through Deep-Learning Based Image Embedding and Gene Association Testing", we explored the correlation between bird phenotypes, represented by a dataset of approximately 23,000 bird images, an annotation dataframe containing the family, subspecies and scientific name of the presented birds, and the corresponding genomes of the birds.
The bird images stem from Birds of the World (“Birds of the world”, 2020) which is propriatary. All images were organized into directories based on the respective bird families, encompassing a total of 249 different families. Additionally, we had access to genomic data for 265 bird species.

2.2 Preprocessing and Data exploration

After mapping the genomes to the respective images, we were left with 948 annotated genome-image combinations. Given the limited availability of genome information at the species level, a decision was made to choose a representative subspecies for each species. This approach ensures the avoidance of duplicate genomes, which could potentially introduce biases and distortions in our results.

The genomes were present both as exon and protein sequences in FASTA format with the headers of each sequence being their GenBank id. Both the exon and protein files were identical in regards to their GenBank ids.

To extract morphological bird features from our images, we utilized image embeddings. Image embeddings are a technique for representing images in a lower dimensionality space. They are often used to measure similarity between images or to group images in a way based on the content of the image. One approach to creating image embeddings is by using a special type of neural network called a Convolutional Neural Network (CNN). By using the output of a specific layer in the CNN, we can generate a unique representation of an image that captures its important features in a simpler form.

Specifically, the model used for extracting the embeddings was OpenAI's CLIP (Radford et al., 2021).

The CLIP model combines the understanding of visual content and language by associating images with their textual descriptions. It creates a joint embedding space where images and text are represented as vectors. CLIP has demonstrated remarkable performance on standard datasets and has been applied in diverse domains without the need for fine-tuning. It is not necessary to provide text annotations during inference for CLIP, meaning for our project, we obtained image embeddings as depicted in Table 2, by directly inputting images into the model.

		All 512 CLIP embedding features				
All 258 species	Nothoprocta ornata	0.4072875	0.36193112	-0.31036815	...	0.13646944
	Smithornis capensis	0.3131959	0.20072602	-0.43035		0.18086748
	Formicarius rufipectus	0.37125307	-0.050516717	-0.52256465		0.25555477
	Sylvia atricapilla	0.21032035	0.24915919	-0.6387599		0.5333118
	Lanius ludovicianus	0.6003466	0.18232825	0.022367803		0.04154638
	Amazona guildingii	0.33798376	0.14806776	-0.47427034		-0.19996244
	Probosciger aterrimus	0.36870858	-0.056442954	-0.33242458		-0.04306676
	Eolophus roseicapilla	0.40817946	0.23949404	-0.13612387		0.18286693
	Chunga burmeisteri	0.3471905	0.34163687	-0.21079393		0.046399638
	Herpetotheres cachin...	0.49282703	0.13153149	-0.42115825		0.22466694

Table 1: Table with all species with genomic data in the rows and their corresponding CLIP image embeddings in the columns.

To visualize the embedding space for all the 23.000 bird images, we reduced their dimensionality from 512 to 2 using Uniform Manifold Approximation and Projection (UMAP) (McInnes, Healy, & Melville, 2020). Using an iterative process that utilizes fuzzy membership assign-

ments via a neighbourhood graph, UMAP reduces the dimensionality of the data while preserving as much of their original euclidean distance as possible (Fig. 4).
CLIP embeddings can also be used to detect image similarity by comparing the embeddings of different images. A common way to measure similarity between vectors is by calculating the cosine similarity matrix (Fig. 5). Cosine similarity measures the angle between two vectors, and it is defined as the dot product of the vectors divided by the product of their magnitudes. Note that normalization is an important step when computing the cosine similarity because it makes the vectors unit length. This ensures that the cosine similarity only measures the angular distance between the vectors and not their magnitudes.
To investigate the impact of specific genes, we explored their ability to segregate the embedding space, which represents avian traits such as coloration. However, due to the limited availability of annotated genomes for only 258 species, a significant reduction in dataset size was necessitated.
Since the GenBank id is unique even for homologous proteins between species, we decided to map it to the definition of the protein. For instance, the identifier "NXL69168.1" was mapped to the corresponding protein definition, such as "FCGBP protein." Subsequently, the presence or absence of each protein in a given species was denoted in a tabular format.

	Nothoprocta ornata	Smithornis capensis	Formicarius rufipectus	Sylvia atricapilla	Lanius ludovicianus	Amazona guildingii	Probosciger aterrimus	Eolophus roseicapilla	Chunga burmeisteri	Herpetotheres cachinnans
104K protein	False	False	False	False	False	False	False	False	False	False
110KD protein	False	False	False	False	False	False	False	False	False	False
1433B protein	True	False	True	True	False	True	True	True	True	True
1433E protein	True	True	True	True	True	True	True	True	False	False
1433F protein	True	True	True	True	True	True	True	True	True	True
...	...	...	...	...	...	...	...	...	...	...
ZY11B protein	True	True	False	True	True	True	True	True	True	True
ZYX protein	False	True	True	False	True	True	True	True	True	False
ZZEF1 protein	True	True	False	True	True	True	True	False	True	True
ZZZ3 protein	True	True	False	True	False	True	False	False	True	True
protein	False	False	False	False	False	False	False	False	False	False

17060 rows × 258 columns

Table 2: Table with gene presence/absence as True/False values with proteins in the rows and the corresponding species in the columns.

By referencing the transcriptomics of colour patterning and coloration shifts in crows (JW et al., 2015), we discerned a collection of 40 candidate genes that are presumed to exert an influence on the coloration of hooded crows. As can be seen in Figure 1, out of the 40 genes 19 were present in our genome files.

The dataset we have contains an uneven distribution of genes of interest among the avian species. Some genes are found in a large number of species, while others are absent in all the species. This imbalance can affect our analysis and the conclusions we draw from it.

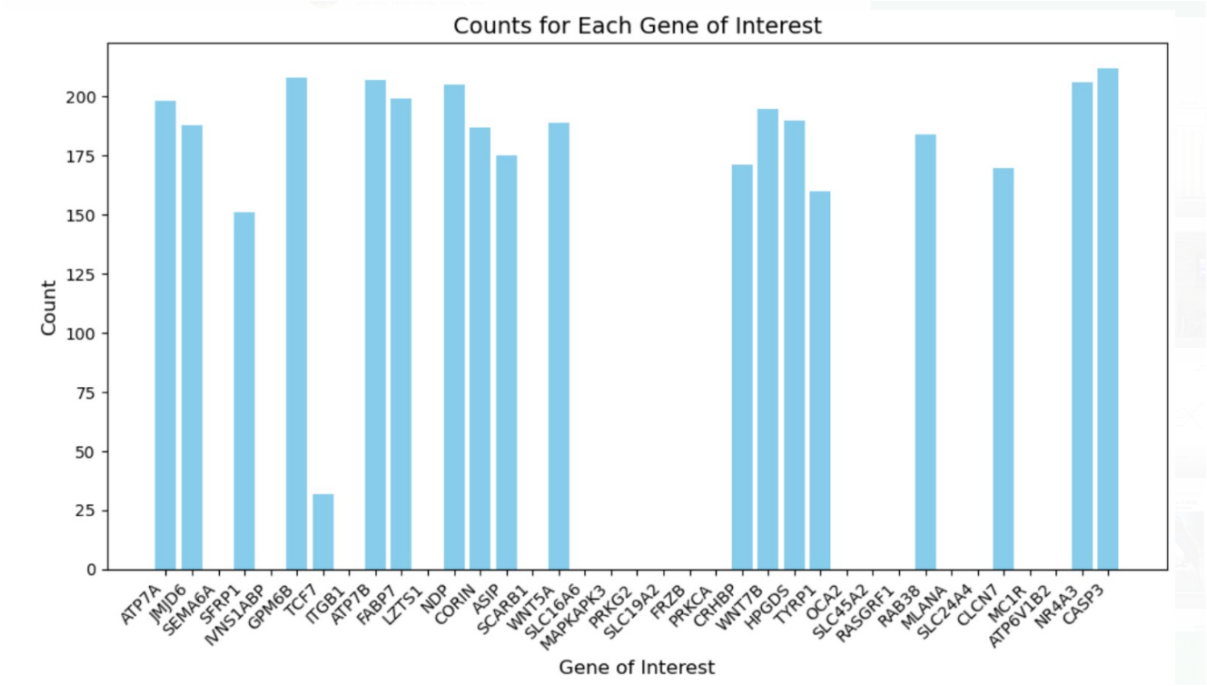

Figure 1: Counts of the 40 identified hooded crows colouration genes in our avian genome dataset.

Given the notion that coloration is most likely encoded within embedding space, the plausibility of training a classifier on the embeddings, to predict the presence or absence of these candidate genes, emerges.

2.3 Modeling

The main objective of this project was to train various machine learning and deep learning models to predict gene presence or absence from our CLIP image embeddings.
The input data, upon which we performed a 64-20-16 train-validation-test split in all of our models, looks as follows:

		All 512 CLIP embedding features				Y
All 258 species	Nothoprocta ornata	0.4072875	0.36193112	-0.31036815		0.13646944	1
	Smithornis capensis	0.3131959	0.20072602	-0.43035		0.18086748	0
	Formicarius rufipectus	0.37125307	-0.050516717	-0.52256465		0.25555477	1
	Sylvia atricapilla	0.21032035	0.24915919	-0.6387599		0.5333118	0
	Lanius ludovicianus	0.6003466	0.18232825	0.022367803		0.04154638	1
	Amazona guildingii	0.33798376	0.14806776	-0.47427034	...	-0.19996244	1
	Probosciger aterrimus	0.36870858	-0.056442954	-0.33242458		-0.04306676	0
	Eolophus roseicapilla	0.40817946	0.23949404	-0.13612387		0.18286693	0
	Chunga burmeisteri	0.3471905	0.34163687	-0.21079393		0.046399638	1
	Herpetotheres cachin...	0.49282703	0.13153149	-0.42115825		0.22466694	...

Figure 2: The input data for the classifiers consists of rows representing the bird species and columns corresponding to the 512-dimensional CLIP embeddings. The target variable, denoted as Y represents the presence (1) or absence (0) of a specific gene of interest.

In the Results section of this report, we evaluated each model using several different metrics and determined which model performs best for predicting candidate genes related to bird coloration. Considering the imbalanced nature of the data, we carefully assessed the models's performance and tried to select the one that achieves a balance between accurately predicting both positive and negative samples.

2.3.1 Classical Machine Learning Models

We selected several classical machine learning models to predict the presence or absence of candidate genes associated with bird coloration using image embeddings from CLIP. The chosen models were logistic regression, decision tree, random forest, Support Vector Machine (SVM), and Multi-Layer Perceptron (MLP).

Logistic regression was chosen as a baseline model due to its simplicity and interpretability. It is a linear model that estimates the probability of a binary outcome. While logistic regression may not capture complex relationships as well as other models, it can provide insights into the importance of different features in the classification task.

The decision tree model was selected for its ability to handle non-linear relationships. It creates a hierarchical structure of decisions based on features and can capture complex patterns in the data. However, decision trees have a tendency to overfit, meaning they may perform well on the training data but struggle to generalize to unseen data.

Random Forest, an ensemble model, was chosen to overcome the limitations of individual decision trees. It combines multiple trees to make predictions, reducing overfitting through aggregation. Random forest can handle high-dimensional data and is generally robust and accurate.

A SVM is a powerful model that can handle both linear and non-linear classification tasks. It finds an optimal hyperplane to separate classes with the largest margin. By using the 'balanced' class weight parameter, SVM accounts for class imbalances in the dataset. SVM can perform well on complex datasets but may require careful tuning of hyperparameters.
Lastly, the MLP is a type of artificial neural network with multiple hidden layers. It can learn complex patterns in the data and adapt to non-linear relationships. It has a high capacity for modeling intricate interactions. However, with limited training data. Also, MLP models may be prone to overfitting and may require careful regularization techniques.
Given the presence of imbalanced data in our dataset, where the number of samples for each class may differ significantly, metrics such as accuracy alone may not provide an accurate assessment of the models' performance. Therefore, we considered additional metrics such as balanced accuracy, precision, recall, F1-score, AUC-ROC and MCC.
Specifically, balanced accuracy takes into account class imbalances by averaging the accuracy of each class. Precision measures the proportion of correctly predicted positive samples, while recall calculates the proportion of actual positive samples that have been correctly identified. The F1-score then combines precision and recall into a single metric. AUC-ROC measures the model's ability to discriminate between positive and negative samples.
In contrast, the Matthews correlation coefficient (MCC) is a measure that yields a high score only when the prediction achieves good results across all four categories of the confusion matrix. This score is proportional to both the number of positive elements and the number of negative elements in the dataset and will serve as the main determiner of model quality (Chicco, 2020).

Figure 3: Example for the different metrics used to compare model performance on the genes of interest.

2.3.2 Deep Learning Models

We wanted to leverage the computational power of deep learning models to predict the presence of candidate genes associated with bird coloration. In the experiments, we employed several models: CLIP (Radford et al., 2021), Gene2Bird, ResNet (He et al., 2016) (ResNet18 and ResNet50), and ResNeSt (Zhang et al., 2020).
Gene2Bird is a model we developed from scratch, designed to compare against the models mentioned above. We decided to make a less complex model to avoid overfitting considering the size and complexity of our data. Furthermore, we used a model with the following values for the number of filters, kernel size, stride, and padding:

- Conv1: 8 Filters of Size 5, Stride 2, Padding 4
- Conv2: 64 Filters of Size 5, Stride 2, Padding 4
- Conv3: 128 Filters of Size 5, Stride 2, Padding 4
- MaxPool: Filter Size 2, Stride 2, No Padding
- FC1: input feature 2176, output feature 1024
- FC2: input feature 1024, output feature 128
- FC3: input feature 128, output feature 1

ResNet18 consists of 18 layers, including residual blocks. It strikes a good balance between model complexity and computational efficiency.
Compared to ResNet18, the more complex ResNet50 is equipped with 50 layers. These provide a better representation of learning and improve performance. Additional layers in ResNet50 and parameters make it a deeper and more powerful model capable of capturing finer details and learning more complex features from images.
ResNeSt combines Squeeze-and-Excitation (SE) blocks to increase the representativeness of ResNet, leading to improved accuracy and performance. In comparison to ResNet50, ResNeSt have similar depth but further strengthens its predictive capabilities by leveraging the benefits of SE blocks. These blocks enable adaptive recalibration of feature responses and better utilization of channel-wise information in the network.

3 Results

3.1 Data exploration

When we visualize the image embeddings and color them based on bird family, as can be seen in Figure 4, we see a certain degree of clustering.

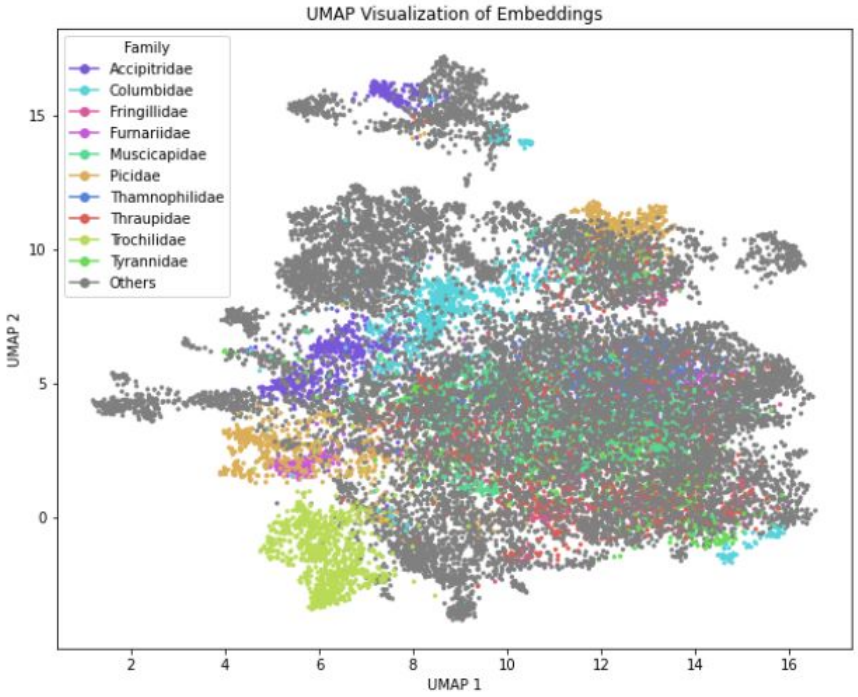

Figure 4: A UMAP representation of the 512 dimensional embedding space. The top 10 most abundant bird families were colored to observe potential clustering. All other bird families were grouped together as "Others".

A clustering of bird families in UMAP space is expected, since more closely related birds tend to look similar and therefore should be closer in embedding space.
For all bird images the cosine similarity matrix in Figure 5 shows an overall high level of similarity.

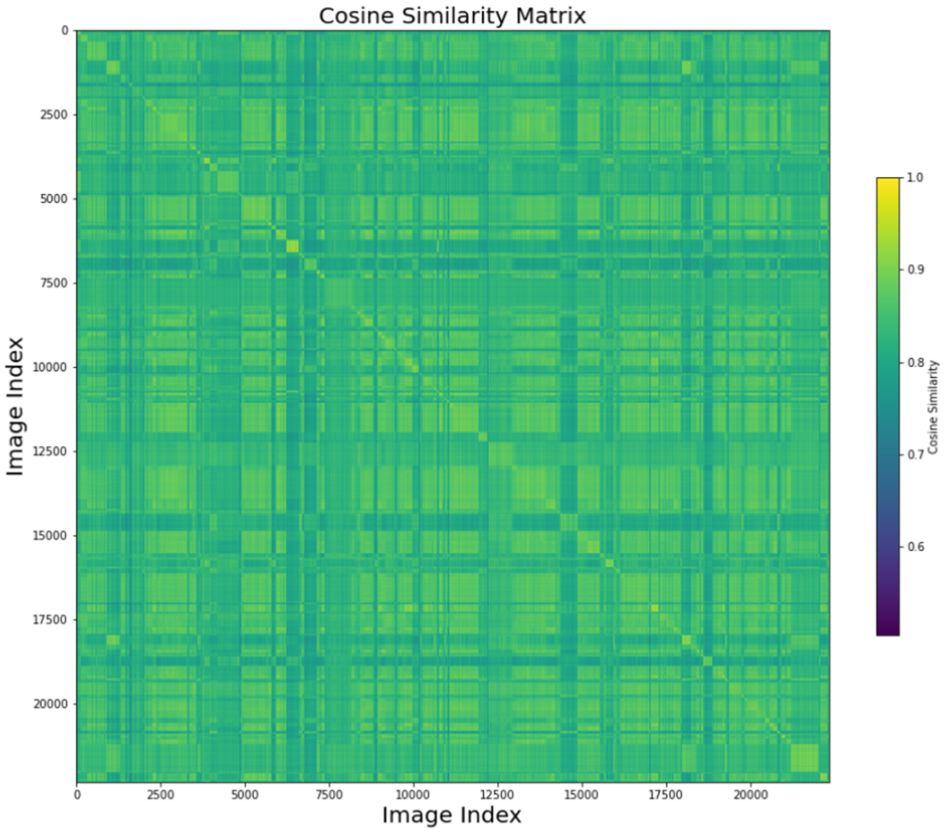

Figure 5: Cosine similarity matrix for all bird images.

We see that the similarity ranges from 0.5 to 1, whereas the cosine similarity can take on values between -1 and +1. Smaller angles between vectors produce larger cosine values, indicating greater cosine similarity. This is expected to a certain degree as all birds are related. Especially, within bird families, different bird species are expected to look very similar.
Given a species of interest we were then able to display the most similar images based on the cosine similarity. For example, in Figure 6 we show the top 5 most similar birds to the Gray's lark (also known as Ammomanopsis grayi). The Gray's lark is a species belonging to the family Alaudidae. It is commonly found in south-western Africa in its natural habitat of hot deserts ("Animalia Bio", n.d.).

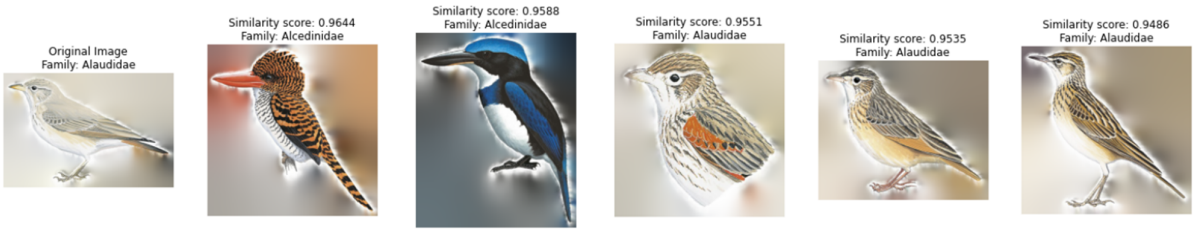

Figure 6: Original species is the Grey's lark. Shown are the top 5 most similar birds according to cosine similarity. The respective bird family for each bird is indicated as well.

We note that, 3 of the top 5 most similar images do in fact also belong to the bird family Alaudidae and look very similar to the original image. However, we can also see that cosine similarity is not a perfect measure as the top 2 most similar birds do not belong to the same bird family as the Grey's lark nor do they share similar coloration traits.

3.2 Modeling

The performance of various models was evaluated using the Matthews Correlation Coefficient (MCC) as the evaluation metric.

3.2.1 Classical Machine Learning Models

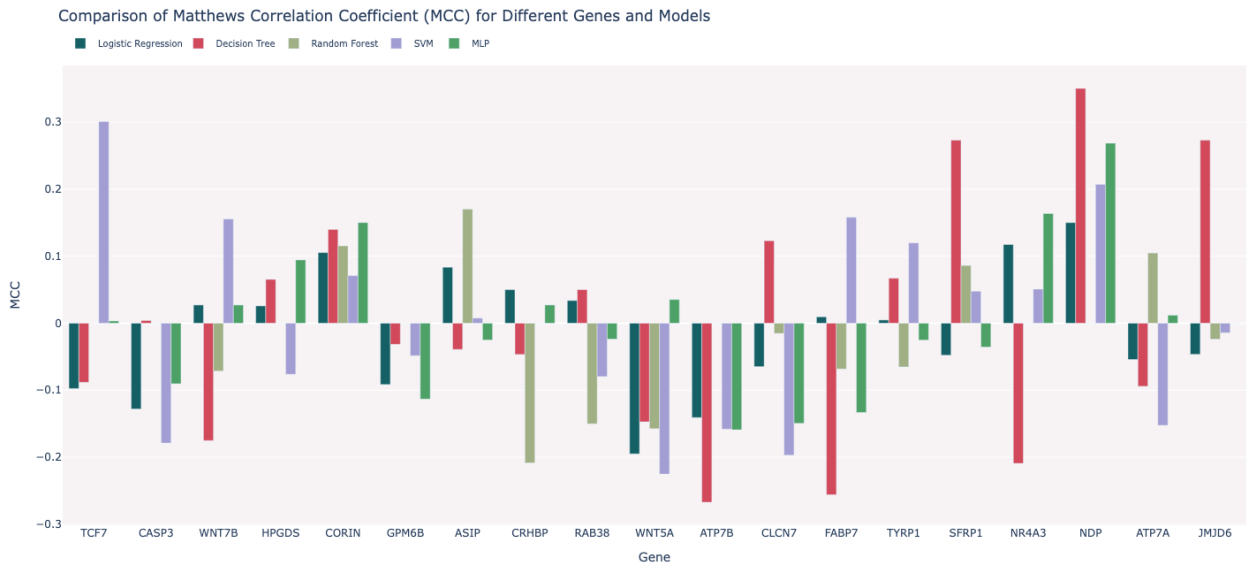

Figure 7: Comparison of MCC for Different Genes and Models

As we can see in Figure 7, the evaluation results seem to stay fairly close to zero with the highest MCC reaching around 0.3 and the lowest reaching about -0.3.
Furthermore, a basic Support Vector Machine (SVM) was employed on the complete genomic dataset with a 0.75, 0.25 train-test split to identify potential new genes outside the previously identified genes of interest that may influence bird morphology through their presence or absence. The protein with the highest MCC identified during this procedure was the AKAP5 protein with a MCC of around 0.39. When bootstrapping the MCC of AKAP5 for 10000 different test-train datasplits, we can see that the real value for the MCC is more likely to be around 0.1. Indicating close to no correlation between our real and predicted values.

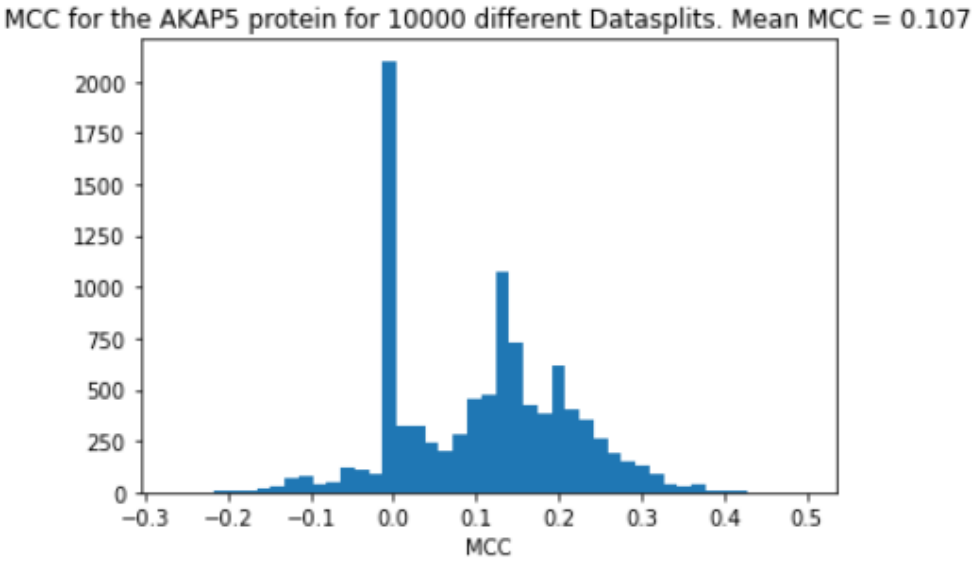

Figure 8: MCC Bootstrapping for AKAP5

3.2.2 Deep Learning Models

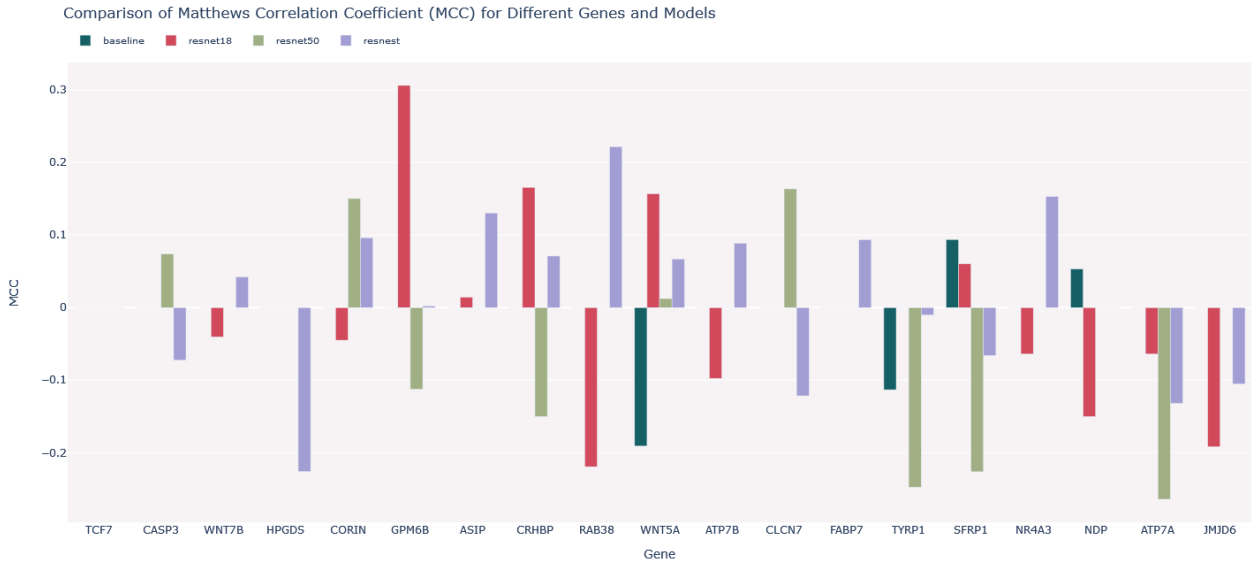

Figure 9: Comparison of MCC for Different Genes and Deep Learning Models

The results shown in Figure 9 show that MCC scores are at zero and near zero. The highest MCC value observed is approximately 0.3, while the lowest score is around -0.3. These values align with the findings from classical machine learning models as illustrated in Figure 7.

4 Discussion

4.1 Classical Machine Learning Models

The model performance for our different classical machine learning models converged around the random point. Any major difference in performance can most likely be attributed to the statistical noise that is created when splitting such a small dataset.
As can be seen in Figure 8, scores like the MCC can vary to a large extend just by sampling in a different way. Consequently, the MCC outcomes of approximately 0.3 attained across diverse genes, as illustrated in Figure 7, do not substantiate the efficacy of the models. This inadequacy is further exacerbated by the presence of negative MCC scores displaying comparable magnitudes, suggesting that those models are exhibiting a tendency to predict outcomes opposite to their intended predictions as good as the highest scoring models predict the

intended predictions.
These findings support that the predictive performance falls short of surpassing that of a rudimentary majority class predictor.

4.2 Deep Learning Models

The results of the deep learning models and the results of the machine learning models are generally similar. The inference that can be made about the performance of the models is similar to the one in Section 4.1. Furthermore, the deeper ResNet50 model achieved similar performance to the ResNet18 model. It may be due to the input we are using and not making enough use of the feature extraction capability of ResNet50. Although the ResNeSt model has a similar depth to the ResNet50, it achieved better results than ResNet50. It might indicate that SE blocks used in the network use better channel-based information.

4.3 Potential Causes for Poor Model Performance

The failure of the applied statistical methods may be attributed to several factors. Firstly, CLIP might not have captured all the necessary information from the images required to accurately identify avian characteristics.
Additionally, the mapping of genbank identifiers to common protein names might not have functioned as intended. Our approach relied on proper protein annotation across species, and if a homologous protein was not annotated with the same name in genbank, it would not be identified as such in our data.
Lastly, relying solely on the presence or absence of a gene may not provide sufficient information for reliably deriving morphological traits. Instead, gene content should also be taken into consideration, as even minor mutations in a gene can potentially lead to changes in traits such as colouration. An illustration of this is a single point mutation occurring in the splice donor site of the scavenger receptor B1 (SCARB1) gene, resulting in a transformation of feather color in common canaries from the wild-type yellow to white (Price-Waldman & Stoddard, 2021). Supporting gene expression data might also be useful for the task at hand.

4.4 Final Remarks

Our findings contribute to the growing body of research utilizing CLIP in various biological applications. In the future, one could embark on the journey to explore different ensemble methods for improved classification performance and also investigate the interpretability of

CLIP embeddings to gain insights into the underlying biological mechanisms. However, it is important to acknowledge the limitations of CLIP, such as its reliance on pre-trained models and potential biases present in the data. In the future, a bigger dataset would also greatly improve the performance of the different classifiers.
In summary, this project demonstrated the potential of CLIP embeddings in gene presence prediction. This may have the potential to open new avenues of research and inspire further exploration into the intricate world of birds.

References

Animalia bio [Accessed: 10.07.2023]. (n.d.). %5Curl%7Bhttps://animalia.bio/grays-lark%7D

Birds of the world [Accessed: 17.07.2023]. (2020).

Chicco, J. (2020). The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*. https://doi.org/10.1186/s12864-019-6413-7

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778. https://doi.org/10.1109/CVPR.2016.90

JW, P., N, V., MP, H., & JB, W. (2015). Transcriptomics of colour patterning and coloration shifts in crows. doi:10.1111/mec.13353

McInnes, L., Healy, J., & Melville, J. (2020). Umap: Uniform manifold approximation and projection for dimension reduction.

Price-Waldman, R., & Stoddard, M. C. (2021). Avian Coloration Genetics: Recent Advances and Emerging Questions. *Journal of Heredity*, *112*(5), 395–416. https://doi.org/10.1093/jhered/esab015

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *CoRR*, *abs/2103.00020*. https://arxiv.org/abs/2103.00020

Zhang, H., Wu, C., Zhang, Z., Zhu, Y., Zhang, Z., Lin, H., Sun, Y., He, T., Mueller, J., Manmatha, R., Li, M., & Smola, A. J. (2020). Resnest: Split-attention networks. *CoRR*, *abs/2004.08955*. https://arxiv.org/abs/2004.08955

A Appendix

A.1 Code availability

Relevant code can be found on the server under the following directory: /s/project/gene2bird/groupA/notebooks.