Operon Prediction Without Embeddings: A SHAP-Guided Logistic Regression Baseline

Chandana Magapu
hmagapu@umass.edu

Abstract

This project investigates whether a simple, interpretable machine learning model can approach the performance of large foundation models on the Operonic Pair Prediction task from the Diverse Genomic Embedding Benchmark (DGEB). I chose a supervised classifier (elastic net-regularised logistic regression model) trained on biologically motivated features selected through domain-informed curation, and validated by SHAP values. The model achieved promising results, outperforming the `cos_sim_ap` baseline of `esm3_sm_open_v1`. The code for this project is available at https://github.com/cmagapu/690U-Project.

1. Introduction

The Operonic Pair Prediction task in DGEB frames operon detection as a binary classification problem: for each pair of adjacent protein-coding genes in E. coli K-12 MG1655, the label is 1 if both genes belong to the same transcription unit (operon) according to BioCyc annotations, and 0 otherwise. After dereplication and a 10% identity-based train/test split, the dataset contains 4,315 training/validation examples and 310 test examples. DGEB benchmarks performance by embedding each gene individually using foundation models, computing pairwise similarity scores using cosine, dot product, Euclidean, and Manhattan metrics, and reporting average precision (AP) for each. The primary metric, top_ap, is defined as the best AP across these four similarity measures [7].

In contrast, my approach avoids embeddings entirely, and instead predicts operon membership directly from six biologically meaningful [5] features: intergenic distance, strand concordance, orientation pattern, promoter and terminator motifs, GC-content difference, and shared functional category (COG). I trained an elastic net-regularised logistic regression model on these features, tuning both the regularisation strength and the l_1/l_2 penalty balance via cross-validation [6]. To guide model behaviour and enhance biological alignment, I computed SHAP values [4] after an initial fit, and use these to adjust the regularisation penalty applied to each feature: highly informative features are regularised less strongly, while low-contribution features are penalised more. This allows the model to emphasize biologically meaningful signals without pruning the feature set outright.

Logistic regression and shallow neural networks have both been widely used in genomics tasks [3], and logistic regression in particular has shown success in previous operon prediction efforts using similar genomic signals [5]. The central question this project addresses is: Can a compact, interpretable model trained on biologically grounded features rival the AP of DGEB's highest ranked foundation model for E. coli (`esm3_sm_open_v1`), while remaining transparent and lightweight?

2. Method

This project uses the DGEB-provided *E. coli* Operonic Pair dataset, consisting of 4,315 training/validation examples and 310 test examples after 70% dereplication and a 10% identity-based train/test split. Each example represents a pair of adjacent protein-coding genes in the *E. coli* K-12 MG1655 genome, labeled 1 if both genes belong to the same transcription unit (operon) according to BioCyc, and 0 otherwise [7].

For each gene pair, I will extract the following six biologically motivated features:

- **Intergenic distance**, computed as the number of base pairs between the end of gene A and the start of gene B (can be negative for overlapping genes).

- **Strand concordance** (binary: same or opposite) and a four-way **orientation pattern** (→→, ←←, →←, ←→), derived from gene strand annotations.
- **GC-content difference**, computed from the genomic sequences of the two genes.
- **Functional category match**, based on shared Cluster of Orthologous Groups (COG) assignments from `eggNOG-mapper v2` [2].
- **Promoter and terminator motifs** detected in the intergenic region using position-weight matrices from bacterial transcription factor databases, scanned using `MEME-Suite 5.5` [1].

I trained an elastic net-regularised logistic regression model using `scikit-learn` [6]. With hyperparameters inverse regularisation strength $C \in \{0.01, 0.1, 1, 10, 100\}$ and the L1/L2 penalty mixing ratio $\alpha \in \{0.1, 0.5, 0.9\}$, tuned using five-fold cross-validation on the combined training and validation data.

To enhance biological alignment and interpretability, I computed SHAP values [4] after the initial model fit to rank feature contributions. Rather than pruning features outright, I first computed the mean absolute SHAP value m_j for each feature $j = 1, \ldots, p$, and then rescaled via

$$\mathrm{imp\_norm}_j \;=\; \frac{m_j \;-\; \min_{k=1,\ldots,p} m_k}{\max_{k=1,\ldots,p} m_k \;-\; \min_{k=1,\ldots,p} m_k}$$

where $\min_{k=1,\ldots,p} m_k$ and $\max_{k=1,\ldots,p} m_k$ denote the minimum and maximum mean SHAP values over all p features. Given the global regularisation constant C_0 and lower-bound parameter $\mathrm{lb} = 0.5$, I defined per-feature regularisation strengths

$$C_j \;=\; C_0\,\bigl(\mathrm{lb} + (1 - \mathrm{lb})\,\mathrm{imp\_norm}_j\bigr).$$

To emulate these C_j in a solver that accepts only a single C, each feature column was rescaled as

$$X_j \;\longrightarrow\; \frac{X_j}{\sqrt{C_j}},$$

and an Elastic-Net model was refit with the best C and mixing ratio from the Grid Search. This ensures that features with higher SHAP importance receive a weaker penalty, while lower-importance features are more strongly regularised.

Because my classifiers produce calibrated probabilities, I will evaluate model performance using average precision (AP), computed directly from predicted probabilities. For fair comparison with DGEB, I benchmark my model against the AP derived from cosine similarity (`cos_sim_ap`) in the `esm3_sm_open_v1` foundation model [7]. I will also report F1-score, precision, recall, and accuracy at a fixed threshold of 0.5, to provide interpretable binary classification metrics.

Planned experiments include:

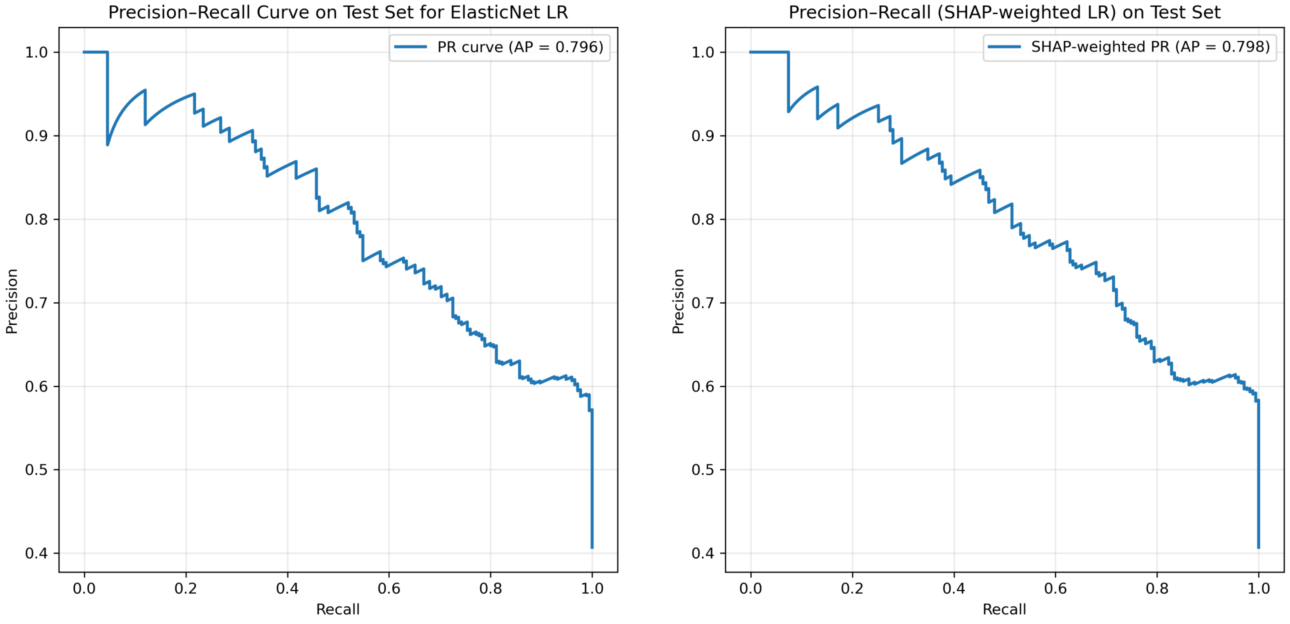
- Comparison of elastic net logistic regression with and without SHAP-informed regularisation.
- Direct benchmarking of my best model against `esm3_sm_open_v1`'s `cos_sim_ap`.

3. Planned results

Model	AP	F1	Precision	Recall	Accuracy
Elastic-Net LR (0.5 threshold)	0.796	0.719	0.620	0.857	0.728
LR @ tuned threshold (0.507)	0.796	0.722	0.614	0.874	0.726
SHAP-Weighted LR (0.5 threshold)	0.798	0.711	0.622	0.829	0.726
SHAP-Weighted LR @ tuned threshold (0.426)	0.798	0.722	0.607	0.891	0.721
`esm3_sm_open_v1` (cosine-sim)	0.5218	0.5864	0.4223	0.9588	0.9765

Table 1. Performance on E. coli operon-pair classification. AP: average precision; F1, precision, recall, and accuracy are computed at the indicated decision thresholds. Performance values for `esm3_sm_open_v1` are from https://huggingface.co/spaces/tattabio/DGEB.

SHAP analysis (Figure 1) highlights strand concordance and operonic overlap as the most influential features, while precision-recall curves (Figure 2) indicate a slightly better overall performance for the baseline model compared to the SHAP-weighted version.



(a) Baseline LR

(b) SHAP-weighted LR

Figure 1. Precision–Recall curves for both models.

4. Conclusion

This project demonstrated that a simple, interpretable elastic net model trained on biologically grounded features can achieve a substantially higher Average Precision on the E. coli operonic pair prediction task compared to the `cos_sim_ap` baseline of a top-performing foundation model from DGEB. SHAP analysis highlighted the importance of strand concordance and operonic overlap. While SHAP-weighted regularization did not improve overall performance very significantly, the results underscore the effectiveness of domain-relevant features for functional genomics, offering a lightweight alternative to large embedding architectures. Future work could explore non-linear models and refined feature weighting strategies.

References

[1] Timothy L Bailey, Mikael Boden, Fabian A Buske, Martin Frith, Charles E Grant, Luca Clementi, Jingyuan Ren, Wilfred W Li, and William Stafford Noble. Meme suite: tools for motif discovery and searching. *Nucleic Acids Research*, 37(suppl_2):W202–W208, 2009. 2

[2] Carlos P Cantalapiedra, Aurora Hernández-Plaza, Ivica Letunic, Peer Bork, and Jaime Huerta-Cepas. eggnog-mapper v2: Functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Molecular Biology and Evolution*, 38(12):5825–5829, 2021. 2

[3] Maxwell W Libbrecht and William Stafford Noble. Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, 16(6):321–332, 2015. 1

[4] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Adv. Neural Inform. Process. Syst.*, 30, 2017. 1, 2

[5] Gabriel Moreno-Hagelsieb and Julio Collado-Vides. A comparative genomics approach to predict operons in prokaryotes. *Bioinformatics*, 18(suppl_1):S329–S336, 2002. 1

[6] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. 1, 2

[7] Thomas West-Roberts, Anshul Kundaje, and James Zou. Dgeb: A diverse genomic embedding benchmark for functional genomics tasks. *bioRxiv*, 2024. Preprint available at `https://www.biorxiv.org/content/10.1101/2024.01.10.575096v1`. 1, 2