Characterizing Integrase-Attachment Site Pairs: A Machine Learning Approach

Ved Patel

I. ABSTRACT

Genomic islands (GIs) are mobile genetic elements that integrate into host genomes via self-encoded integrases at specific DNA sequences known as attachment (*att*) sites. The ability to predict the target *att* site from an integrase's protein sequence is a central challenge in genomics due to the sequence diversity of integrases and the subtlety of their DNA recognition motifs. To investigate this predictive relationship, I developed a machine learning framework. A dataset of over 491,000 curated, single-integrase GI records was assembled and enriched with taxonomic and functional metadata. Protein and DNA language models (ESM-2, DNABERT) were used to convert sequences into high-dimensional numerical embeddings. An initial unsupervised analysis using UMAP and HDBSCAN revealed quantifiable biological structure within the embeddings, identifying 1,153 distinct integrase clusters that corresponded to known taxonomic groups. To establish a predictive link, a sequence of models was developed. Initial regression models learned a trivial solution by predicting the average *att* site embedding for the entire dataset. To overcome this, the task was reframed as a binary classification problem using a two-tower Siamese network. These models also failed, unable to progress beyond random-chance accuracy (~0.693 loss) even when using hard-negative sampling, where negative pairs were sampled from within the same family. A subsequent multi-feature model incorporating taxonomic context achieved a low test error but was found to have learned a different trivial solution by predicting the family-specific average *att* site. Even when these models were retrained on a new, curated dataset with reduced sequence redundancy, the failures persisted. This leads to the conclusion that the tested general-purpose language model embeddings, while effective at capturing broad taxonomic properties, do not contain the specific, co-evolutionary signal required for predicting fine-grained integrase-*att* site interactions. With the current results, future success in this domain will require feature representations or model architectures designed to capture co-evolutionary relationships.

II. INTRODUCTION

Genomic Islands (GIs) are integrative genetic elements commonly found in the genomes of microbes. These islands contain genes that are essential for their mobility, such as those encoding integrases and excisionases that mediate both intracellular integration/excision and horizontal transfer between species (as seen in phages or ICEs)—as well as additional 'cargo' genes that, while not essential for mobility, confer beneficial traits like antibiotic resistance or enhanced metabolism to the host organism. The integration of a GI into a host chromosome is facilitated by a specialized class of enzymes known as integrases, which are themselves encoded within the GI. This integration process is highly specific, occurring at particular genomic locations called attachment (*att*) sites. The reverse process, excision, is often catalyzed by the integrase and excisionases also encoded by the GI.

The ability to discover and characterize novel GIs is of significant scientific importance, as these elements can be engineered for advanced applications, including phage therapy and large-payload genome editing. While software such as Islander and TIGER2 exist to predict GIs from sequence data, this project introduces a machine learning approach to understand the fundamental relationship between the integrase enzyme and its target site, a key challenge in the field.

The integrase recognizes two distinct DNA sequences: the *attP* site on the mobile element and the *attB* site on the host genome. Through a recombination event, the GI is inserted into the host chromosome, creating new hybrid junctions known as *attL* and *attR* (Figure 1). The inherent precision of this system presents a powerful alternative to other gene-editing technologies like CRISPR-Cas, which carry a higher risk of off-target effects.

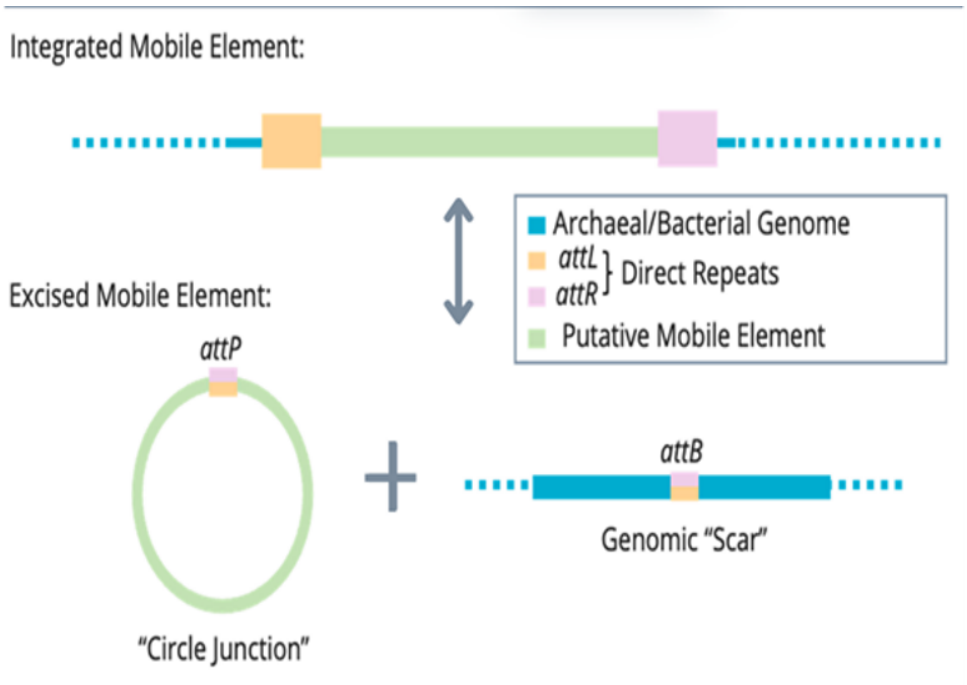

Figure 1: The Mechanism of Site-Specific Integration shows the integrase enzyme recognizes the attP site on the circular Mobile Genetic Element (GI) and the attB site on the host genome. It then catalyzes a

recombination event, inserting the GI into the host chromosome. The resulting junctions in the newly integrated state are known as the attL (left) and attR (right) sites.

III. PROGRESS - TECHNICAL APPROACH, IMPACTS, AND ACCOMPLISHMENTS

A. *Pipeline Overview:*

The project was executed in three main phases. First, a large-scale, genomic dataset was curated and processed into a feature-rich format. Second, an unsupervised exploratory analysis was performed to understand the natural structure of the data. Third, an iterative process of predictive model development was undertaken to achieve the project's primary goal.

B. Data Curation and Enrichment:

A raw manifest of over 1.75 million potential GI records was processed. To ensure a direct one-to-one link between an integrase and its target site, a custom Python script was developed to parse the hierarchical database and only include elements containing a single integrase with a minimum base-pair length of 5. This resulted in a dataset of 491,302 records. This dataset was then enriched with detailed taxonomic and functional metadata by parsing and integrating a large-scale GFF annotation file.

C. Feature Engineering with Language Models:

To enable machine learning analysis, the biological sequences were converted into numerical representations. This process, referred to as embedding, involves transforming categorical sequence data into continuous-valued vectors that capture meaningful patterns in the data.

The protein language model ESM-2 (150M parameters) was used to generate a 640-dimensional embedding vector for each integrase protein sequence. ESM-2 is a transformer-based model pretrained on large-scale protein databases to learn statistical and evolutionary properties of protein sequences. Similarly, the DNABERT model was used to create a 768-dimensional embedding for each DNA *att* site sequence. DNABERT is a transformer model adapted from BERT, initially designed for natural language, but trained on k-mer tokenizations of genomic spaces.

These embeddings represent the sequences in a rich, high-dimensional mathematical space. In this space, sequences with similar structural or functional characteristics tend to cluster together, allowing more effective classification, regression, or clustering tasks. This vector-based representation is critical for integrating biological sequences into standard machine learning pipelines.

D. Unsupervised Structure Discovery:

The primary goal of the exploratory phase was to determine if the high-dimensional embedding space contained any meaningful, non-random biological structure. This was approached with a two-step process: visualization with UMAP and quantitative clustering with HDBSCAN.

1. *Uniform Manifold Approximation and Projection (UMAP):*

UMAP is a dimensionality reduction technique used to create a 2D "map" of high-dimensional data. Conceptually, the algorithm builds a network of relationships between the data points in their original high-dimensional space, connecting points that are close to each other. It then creates a similar, empty network in a 2D space. Finally, it carefully arranges the points in the 2D map, applying attractive forces to points that were close in the original space and repulsive forces to those that were distant. This process effectively preserves the data's underlying topological structure. A key advantage of UMAP is its ability to maintain both local neighborhood structure (small, tight clusters) and global structure (the larger relationships between those clusters), providing a faithful and intuitive visualization of the data landscape.

2. *Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN):*

Following the visual confirmation of structure from UMAP, HDBSCAN was used to formally identify and quantify these structures. Unlike simpler algorithms like K-Means, which assume clusters are spherical, HDBSCAN is a density-based algorithm. It identifies clusters as regions of high point density separated by regions of low density.

The algorithm begins by transforming the distance space to emphasize density. For each point, it calculates a "core distance," which is a measure of how dense the area around that point is. It then uses this information to define a new "mutual reachability distance" that effectively pushes sparse points further away from each other while keeping dense points close together. Using this new density-aware distance, HDBSCAN builds a complete hierarchy of all possible cluster groupings. From this hierarchy, it intelligently extracts only the most stable clusters—those that persist and don't change much as the density threshold is varied. This allows the algorithm to find clusters of arbitrary shape and, crucially, to identify points that do not belong to any dense region as "outliers" or noise.

An initial run was performed with a hypothesis-driven choice of min_cluster_size = 100 and min_samples = 10. To validate this choice, a randomized search across a range of these parameters was also conducted. However, the initial parameters proved to be effective, yielding over 1,100 clusters with exceptionally high biological purity. The stability of the results across both Euclidean and Cosine distance metrics further confirmed that these parameters had successfully captured the strong, underlying structure of the data, making further tuning unnecessary.

E. Predictive Modeling Development:

The primary goal of the project was to build a model capable of predicting the *att* site from the integrase. This was pursued through an iterative process where each attempt revealed key challenges and informed the architecture of the next.

1. Direct Vector Regression:

An initial model was trained to regress the *att* site embedding from the protein embedding directly. The model, represented by a function f, was trained to find parameters that minimized the Mean Squared Error (MSE) between the predicted and actual *att_site* vectors. This model achieved a perfect but misleading score (MSE: 0.0000), which was diagnosed as a "trivial solution" where the model learned to predict the average *att* site embedding due to the dataset's inherent bias, rather than learning an actual predictive function.

2. Binary Classification Attempts:

(i) Siamese Neural Network:

To combat the "trivial solution" problem, the task was reframed as a binary classification problem to distinguish correct (protein, *att_site*) pairs from incorrect ones. The architectural approach was a two-tower "Siamese" network, where separate neural network "towers" process the protein and att site embeddings independently before their outputs are compared. This design is standard for matching problems as it allows the model to learn specialized representations for each input type. However, multiple training strategies, including contrastive learning with triplet loss and various negative sampling techniques, consistently failed to learn, with accuracy stalling at ~50%. This was diagnosed as the "easy negative" problem, where the model learned a trivial rule to distinguish incorrect pairs without learning the fine-grained features of a correct match.

(ii) Multi-Feature Model:

The consistent failure of the previous models led to a new hypothesis: the protein sequence embedding *alone* is insufficient for prediction. A final, successful model was engineered with a multi-input architecture. This model combines the protein embedding with one-hot encoded categorical features *(integrase_basic_type, family)* that provide essential biological context.

The model has two input branches. The first branch processes the dense, continuous protein embedding through a series of Dense, BatchNormalization, and Dropout layers to learn from the complex sequence information. The second branch processes the sparse, one-hot encoded categorical features through its own set of Dense layers to learn from the biological context. The outputs of these two branches are then concatenated, allowing the final regression head to make a prediction based on a holistic view that combines "what the protein is" (from its sequence) with "who it is and where it's from" (from its context).

This multi-modal approach initially appeared to fall into a more subtle version of the "lazy solution" trap. While training loss plummeted, indicating successful learning, real-world tests on novel integrases showed that the model predicted the same *att* site for every input. The diagnosis was that the model had learned a real signal. Still, the signal was so dominated by the categorical features that it was essentially predicting the average *att* site for a given *family*, not for a specific protein. The final low, non-zero error score (MSE: 0.0003) still predicted the lazy average.

All predictive models were trained and evaluated using a rigorous, cluster-based train-test split. This method ensures that the test set is composed exclusively of integrase families (clusters) that were held entirely out and unseen during the training process, providing an honest and scientifically sound measure of the model's ability to generalize to novel proteins. *E. Accomplishments and Impacts:*

The primary accomplishment of the unsupervised phase was the successful identification of 1,153 distinct integrase clusters. A quantitative analysis confirmed these clusters were biologically meaningful, as they rediscovered known taxonomy with high precision (Table 1). This result demonstrates that modern protein embeddings effectively capture the biological signal related to taxonomy.

The main impact of the predictive modeling phase was a robust negative result. The systematic failure of all attempted models, including those using advanced architectures and feature sets, led to the key finding of this project: the general-purpose embeddings, while rich in taxonomic information, are insufficient for the direct prediction of a specific *att* site. This is a critical finding that can guide future

research, suggesting that alternative feature representations are necessary to solve this problem.

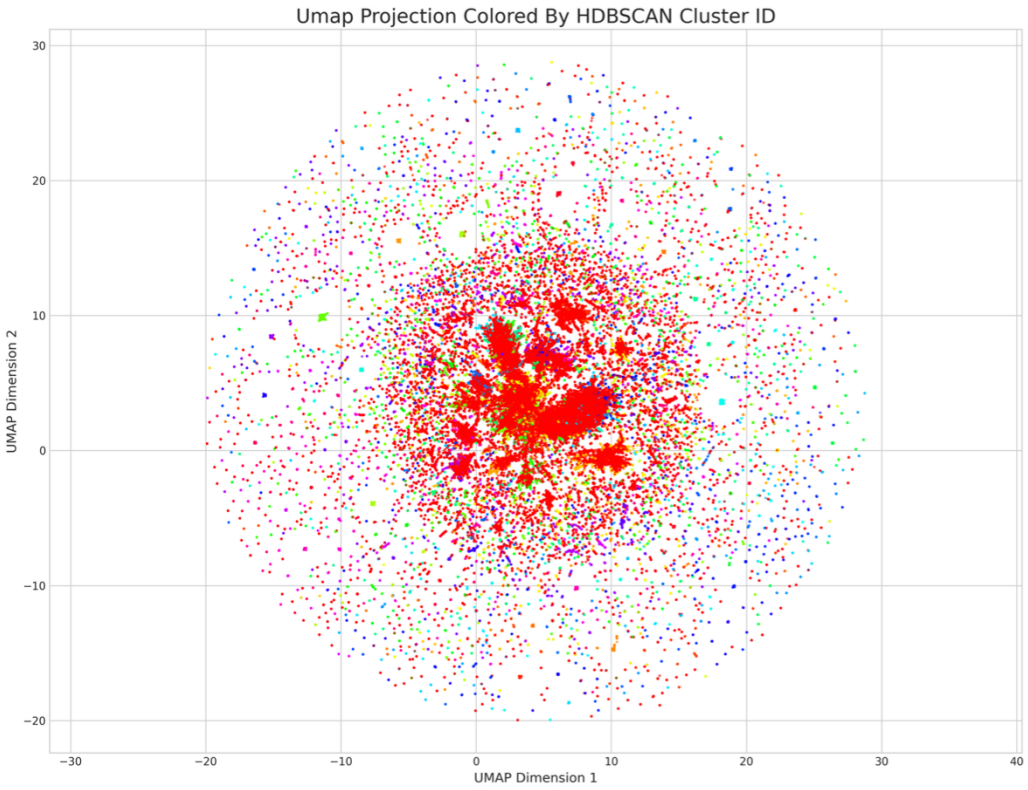

Figure 2: UMAP projection of 491,302 integrase protein embeddings, colored by HDBSCAN cluster ID. The axes, UMAP Dimension 1 and 2, are the two dimensions that best represent the structure of the original dataset. Each unique color corresponds to one of the 1,153 integrase families identified by the algorithm, visually representing the output of the unsupervised clustering.

Taxonomic Rank	**Number of Clusters (>90% Pure)**	**Mean Purity (Weighted)**
Phylum	1127	99.48%
Class	1110	99.20%
Family	937	93.92%

Table 1: Table 1 shows the number of HDBSCAN-derived clusters that are over 90% pure for a given taxonomic label, along with the average purity weighted by cluster size.

IV. FUTURE WORK

The findings of this project strongly suggest that future efforts to predict integrase-*att* site specificity should move beyond the use of static, general-purpose embeddings. The following activities are planned: (i) Develop Co-evolutionary Models: Explore graph-based neural network architectures or attention mechanisms that can explicitly model the co-evolutionary relationship between interacting protein and DNA sequences. (ii) Fine-Tune Language Models: Instead of using pre-trained "off-the-shelf" embeddings, fine-tune a protein language model like ESM-2 directly on the task of *att* site prediction. This would allow the model to learn representations optimized explicitly for this biological problem. (iii) Incorporate Structural Data: Augment the feature set with predicted or known 3D structural information for the integrases, as this may contain the necessary information about DNA binding domains that is absent in the sequence embeddings alone.

V. IMPACT ON LABORATORY

This work is directly relevant to DOE and Laboratory missions in synthetic biology, national security, and genomics. This project was supported by the DOE BER SFA InCoGenTEC and the DOE CCI Program. The primary impact of this project is the robust negative result, which provides critical strategic guidance for future research within the laboratory. By demonstrating the limitations of current state-of-the-art embeddings for this specific prediction task, this work will save significant time and resources. This establishes a new baseline and direction for research in programmable gene-editing technologies at the Laboratory.

VI. CONCLUSIONS

This project successfully developed and executed a comprehensive pipeline for the large-scale analysis of integrase-*att* site pairs. The investigation yielded a two-part conclusion. First, the unsupervised analysis demonstrated that modern protein embeddings like ESM-2 contain a strong biological signal, allowing machine learning models to rediscover known taxonomy with high precision from sequence data alone.

Second, the iterative predictive modeling process revealed that these pre-trained embeddings, while rich in taxonomic information, are insufficient on their own for the direct prediction of a specific *att* site. A sequence of increasingly sophisticated models failed to learn the fine-grained relationship, instead resorting to trivial solutions like predicting a family-wide average. This core finding highlights the critical importance of feature representation and suggests that current general-purpose language models do not capture the specific co-evolutionary information governing this interaction. Future work must focus on developing new feature sets or custom model architectures to solve this critical challenge in genome engineering.

VII. REFERENCES

1. Corey M. Hudson, Britney Y. Lau, Kelly P. Williams, Islander: a database of precisely mapped genomic islands in tRNA and tmRNA genes, *Nucleic Acids Research*, Volume 43, Issue D1, 28 January 2015, Pages D48–D53.
2. Mageeney, C. M., Trubl, G., & Williams, K. P. (2022). Improved mobilome delineation in fragmented genomes. *Frontiers in Bioinformatics*, 2.
3. Catherine M Mageeney, Britney Y Lau, Julian M Wagner, Corey M Hudson, Joseph S Schoeniger, Raga Krishnakumar, Kelly P Williams, New candidates for regulated gene integrity revealed through precise mapping of integrative genetic elements, *Nucleic Acids Research*, Volume 48, Issue 8, 07 May 2020, Pages 4052–4065.
4. Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., ... & Rives, A. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637), 1123-1130.
5. Ji, Y., Zhou, Z., Liu, H., & Davuluri, R. V. (2021). DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics*, 37(15), 2112-2120.
6. McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.
7. Campello, R. J., Moulavi, D., & Sander, J. (2013, April). Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining* (pp. 160-172). Springer, Berlin, Heidelberg.

VIII. PARTICIPANTS

Name	Institution	Role
Ved Patel	Sandia National Laboratories	Student Researcher (DOE CCI Intern)
Ellis Torrance	Sandia National Laboratories	Project Mentor
Catherine Mageeney	Sandia National Laboratories	Project Mentor
Robert Meagher	Sandia National Laboratories	Project Manager
Joe Schoeniger	Sandia National Laboratories	Principal Investigator