COMPSCI 602
Project Report 6

Anonymous
College of Information and Computer Science
University of Massachusetts
Amherst, MA 01003

Project Title: Beyond Black Boxes: Interpretability-Focused ML for Operon Prediction

1 Introduction

The system in this project consists of nonlinear simple ML models outlined in OperonSEQer [5] (Logistic Regression, Gaussian Naive Bayes, Random Forest, XGBoost, SVM (with radial basis function), Multilayer Perceptron). The task they are being evaluated on is that of operon prediction in prokaryotes (if two genes are transcribed together, they form an operon, otherwise they do not; i.e a binary classification). The environment in which the study is performed is the DGEB [8] E.Coli dataset augmented with genomic data from NCBI [7] and biologically motivated features [6] [3]: intergenic distance between the pair of genes that code for the pair of proteins, strand concordance of the pair of genes (whether both genes are transcribed in the same direction, either transcribed 5'→3' (+) or 3'→5' (-)), orientation pattern of the pair of genes (++, +-, -+, –), GC-content difference between the pair of genes, and promoter/terminator trinucleotide motifs computed from the genomic data, as well as functional category match computed via eggNOGmapper [1].

Research Question: Which genomic and compositional features are most important for prokaryote operon prediction by simple ML models, and to what extent do these features provide conditionally independent information?

Hypotheses:

- **H1.1 (Feature Importance Hierarchy):** Based on prior literature [4] [3], trinucleotides representing promoter/terminator motifs will show the highest predictive importance (>20% ROC-AUC degradation when ablated), followed by intergenic distance and strand orientation features (10-15% degradation), functional annotation features (COG similarity, 5-10% degradation).

- **H1.2 (Feature Redundancy):** Trinucleotide features exhibit high redundancy and conditional dependence, with >90% of variance explained by the first 10 principal components. Ablating PC1-10 will produce similar performance degradation to ablating all 64 trinucleotide features.

- **H1.3 (Model Architecture Dependence):** Different model architectures will show varying sensitivities to feature ablation. Logistic Regression will show balanced, robust feature dependencies (<20% degradation for any single feature). Tree-based models (XGBoost, Random Forest) will show concentrated dependence on specific features (>20% degradation). Kernel-based and neural models (SVM with RBF kernel, MLP) will show intermediate behavior with distributed importance across multiple feature types.

2 Research Design

This study investigates which genomic and compositional features are most important for prokaryote operon prediction by simple machine learning models, and to what extent these features provide conditionally independent information. Three primary hypotheses are tested: (H1.1) that sequence composition features representing promoter/terminator motifs show highest importance followed by intergenic distance, functional annotation, and gene orientation; (H1.2) that trinucleotide features exhibit high redundancy with >90% variance explained by the first 10 principal components; and (H1.3) that different model architectures show varying feature sensitivities, with linear models displaying balanced dependencies while non-linear models concentrate dependence on specific feature groups.

Systematic **feature ablation studies** are conducted to assess feature importance. Features are disrupted using two methods—shuffle (permutation preserving distribution) and random sampling (from normal distributions)—and performance degradation is measured across six machine learning models. Additionally, Principal Component Analysis (PCA) is performed on trinucleotide features with subsequent ablation of principal component subsets to test feature redundancy and conditional independence.

The study utilizes the E. coli operon prediction dataset from DGEB [8], augmented with genomic data from NCBI GenBank [7] for the E. coli K-12 MG1655 reference genome. Functional annotations are obtained via eggNOG-mapper v2 [1]. 77 features across five categories are engineered: genomic organization (intergenic distance, strand concordance, orientation patterns, overlap metrics), sequence composition (64 trinucleotide frequencies representing promoter/terminator motifs [3], GC content), and functional similarity (COG category matches, and group similarity i.e whether they're part of the same functional group [2]). The dataset is partitioned into training (85%) and validation (15%) sets using stratified sampling.

Six machine learning models representing diverse algorithmic approaches are trained: Logistic Regression (linear), SVM with RBF kernel (kernel-based), Random Forest (ensemble trees), XGBoost (gradient boosting), MLP (neural network), and Naive Bayes (probabilistic). Hyperparameters are optimized via 5-fold stratified cross-validation with grid search, using F1 as the optimization metric.

For ablation studies, features are disrupted on the **validation set only** (models trained on original data) to measure true dependence. Each ablation is repeated 10 times with different random seeds for statistical robustness. Performance is evaluated using five metrics (accuracy, precision, recall, F1, ROC-AUC), with ROC-AUC serving as the primary metric. Feature importance is categorized based on ROC-AUC degradation: CRITICAL (>15 percentage points), HIGH (5-15pp), MODERATE (1-5pp), LOW (0.3-1pp), and MINIMAL (<0.3pp). These thresholds distinguish catastrophic features from major contributors and negligible features based on the observed distribution in our experiments.

For hypothesis H1.2, PCA is performed on the 64 trinucleotide features, and principal component subsets (PC1, PC1-5, PC1-10) are ablated by setting components to zero in the transformed space before inverse transformation. Comparing the impact of ablating PC1-10 versus all 64 features tests whether trinucleotides provide redundant, conditionally dependent information.

3 Results

This section presents the detailed methodology and experimental results for our feature ablation study of prokaryote operon prediction.

3.1 Methodology

3.1.1 Data and Feature Engineering

This study utilized the E. coli operon prediction dataset from DGEB [8], consisting of protein pairs with binary labels indicating operon membership. Genomic annotations were obtained from NCBI GenBank [7] (E. coli K-12 MG1655, accession U00096.3), providing gene coordinates, strand information, CDS sequences, and protein identifiers. Functional annotations were obtained using eggNOG-mapper v2 [1] for COG category assignments.

A total of 77 features were engineered across five categories [6]: (1) **Genomic organization** (5 features): intergenic distance, overlap length, genes overlap indicator, strand concordance, and four orientation patterns (++, –, +-, -+); (2) **Sequence composition** (67 features): 64 trinucleotide frequencies in intergenic regions [3] and 3 GC content features; (3) **Functional annotation** (2 features): COG match and COG similarity indicators [2]. For overlapping genes with no intergenic sequence, trinucleotide frequencies were set to zero.

The dataset was split into training (85%) and validation (15%) sets using stratified sampling (random_state=42). Numeric features (distances, GC content, trinucleotides) were standardized using StandardScaler for distance-based models (Logistic Regression, SVM, MLP). Binary features and tree-based models (XGBoost, Random Forest, Naive Bayes) required no scaling.

3.1.2 Model Training and Hyperparameter Optimization

Six ML models [5] were trained: Logistic Regression (ElasticNet regularization), SVM (RBF kernel), Random Forest, XGBoost, MLP (two hidden layers), and Naive Bayes. Hyperparameters were optimized via 5-fold stratified cross-validation with grid search, using F1 as the optimization metric.

Model	Optimal Hyperparameters	CV F1
XGBoost	learning_rate=0.1, max_depth=3, n_estimators=100	0.875
SVM (RBF)	C=10, γ=0.01	0.871
MLP	(100, 100) hidden layers, relu activation	0.870
Logistic Regression	C=0.1, L1 ratio=0.05	0.867
Random Forest	n_estimators=200, max_depth=None	0.867
Naive Bayes	Default parameters	0.777

Table 1: Optimal hyperparameters selected through 5-fold stratified cross-validation with grid search (F1 optimization metric). XGBoost and SVM achieved the highest cross-validation F1 scores (0.875 and 0.871 respectively), while Naive Bayes lagged significantly (0.777), indicating inherent limitations in modeling complex feature interactions despite its probabilistic foundation.

3.1.3 Feature Ablation Methodology

To assess feature importance, features were systematically ablated on the **validation set only** using two methods: (1) **Shuffle**: random permutation preserving distribution, testing feature-target associations; (2) **Random**: sampling from normal distributions with training set statistics, testing sensitivity to distributional changes. Each ablation was repeated 10 times with different random seeds to obtain robust statistical estimates. Heatmaps depicting performance change were plotted for each of the models depicting five metrics: accuracy, precision, recall, F1 score, and ROC-AUC, however the primary metric for interpretation/insight was ROC-AUC.

Feature importance categories based on average ROC-AUC degradation: CRITICAL (>15pp), HIGH (5-15pp), MODERATE (1-5pp), LOW (0.3-1pp), MINIMAL (<0.3pp). These thresholds distinguish catastrophic features from major contributors based on observed distributions.

3.1.4 Ablated Feature Groups

The following features and feature groups were systematically ablated:

1. **Individual distance features**: intergenic_distance, overlap_length, genes_overlap
2. **GC content features**: gc_content_diff alone, gc_content_A + gc_content_B together, all three GC features
3. **Strand/orientation features**: strand_concordant alone, all four orientation patterns together, all five strand features combined
4. **COG features**: COG_match alone, COG_similar alone, both together
5. **Trinucleotide features**: All 64 trinucleotide features together

3.1.5 Principal Component Ablation

To test feature redundancy (H1.2), Principal Component Analysis (PCA) was performed on the 64 trinucleotide features using the validation set. Specific principal component subsets were then ablated:

- **PC1 only** (capturing ∼83% of variance)
- **PC1-5** (capturing ∼93% of variance)
- **PC1-10** (capturing ∼95.5% of variance)

3.2 Feature Ablation Results

3.2.1 Strand Concordance and Orientation Features

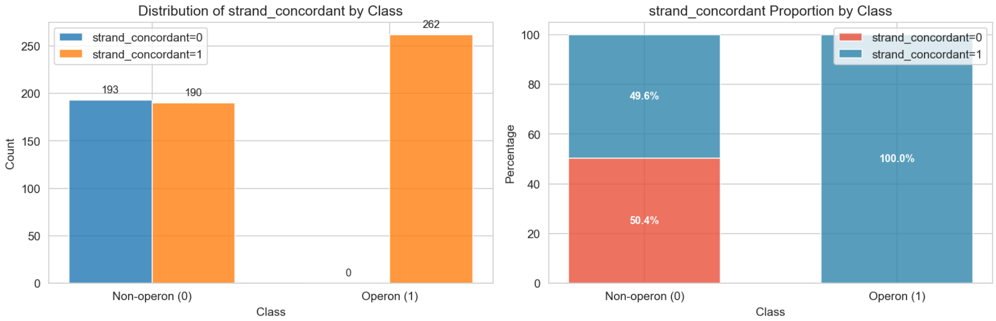

Figure 1: **Strand concordance exhibits near-perfect class separation.** Distribution of the binary strand_concordant feature (whether genes are transcribed in the same direction) stratified by operon membership in the validation dataset. **Left panel:** Raw counts show all 262 operons have concordant strands (orange), while non-operons split evenly (193 concordant, 193 discordant). **Right panel:** Proportional view reveals 100% of operons are strand-concordant versus only 50% of non-operons, making this feature a near-perfect discriminator. This biological constraint reflects the requirement for co-transcription: genes in the same operon must be on the same strand to be transcribed together by a single RNA polymerase.

To quantify the importance of orientation, three ablation scenarios were examined: (1) strand concordance alone, (2) orientation patterns alone, and (3) all five strand-related features together. For each scenario, the variable(s) were ablated and performance degradation was measured across all six models.

Strand Concordance Alone The impact of ablating only the strand_concordant feature while keeping all other features intact was first examined.

Model	Orig.	Shuf.	∆ Shuf.	Rand.	∆ Rand.
XGBoost	0.955	0.730	-0.225	0.720	-0.234
Naive Bayes	0.828	0.712	-0.116	0.500	-0.327
Logistic Regression	0.940	0.910	-0.030	0.909	-0.031
Random Forest	0.953	0.919	-0.034	0.918	-0.034
SVM	0.949	0.919	-0.030	0.918	-0.031
MLP	0.947	0.935	-0.012	0.935	-0.012

Table 2: **ROC-AUC degradation after ablating strand_concordant for 6 models during shuffle and random sampling. Shuffle ≈ Random for most models:** XGBoost (-22.5pp vs -23.4pp), Logistic Regression (-3.0pp vs -3.1pp), Random Forest (-3.4pp vs -3.4pp), SVM (-3.0pp vs -3.1pp), and MLP (-1.2pp vs -1.2pp) show nearly identical degradation, indicating they learned "does strand concordance correlate with operon membership?" rather than specific distributional patterns. **Exception—Naive Bayes:** Random >> Shuffle (-32.7pp vs -11.6pp) suggests it learned distributional dependencies, collapsing to chance-level (ROC-AUC=0.5) when the binary feature's distribution changes. **Interpretation:** For a critical binary discriminator like strand concordance, most models learned robust feature-target associations, while Naive Bayes's probabilistic framework makes it vulnerable to distributional shifts.

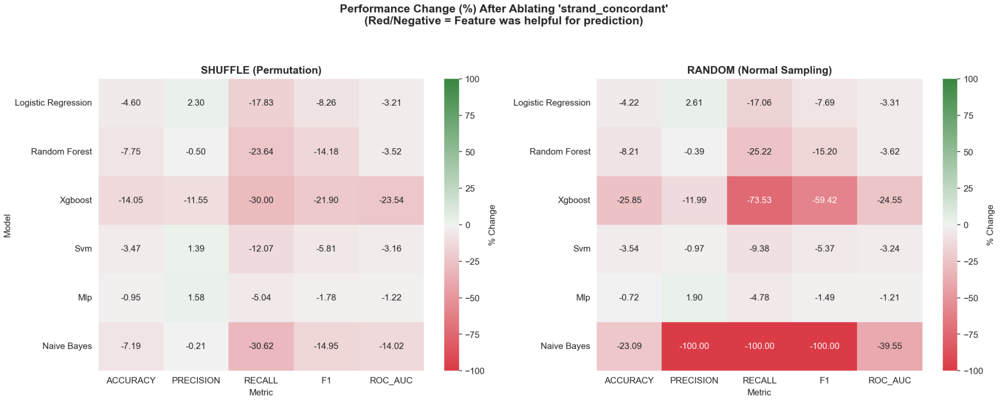

Figure 2: **Strand concordance ablation reveals differential model robustness across all performance metrics.** Heatmaps show performance degradation (percentage point change) after ablating strand_concordant for all six models across five evaluation metrics. **Left:** Shuffle ablation preserves feature distributions. **Right:** Random sampling tests distributional robustness. **Key patterns:** (1) XGBoost shows severe, consistent degradation across all metrics (dark red cells), with 22pp ROC-AUC loss and 24% recall collapse. (2) Naive Bayes exhibits catastrophic failure under random sampling (darkest red), with ROC-AUC dropping to 0.5 (random guessing). (3) Logistic Regression, Random Forest, SVM, and MLP display moderate degradation (lighter shades), maintaining functional performance. (4) Similarity between left and right panels for most models confirms they learned genuine feature-target relationships rather than overfitting to specific value distributions.

Orientation Patterns Alone Given that orientation patterns (++, –, +-, -+) are essentially a superset from which strand_concordant is derived, whether they provide redundant or complementary information was tested by ablating them while keeping strand concordance intact. Table 3 shows a striking result.

Model	Orig.	Shuf.	Δ Shuf.	Rand.	Δ Rand.
Naive Bayes	0.828	0.719	-0.109	0.500	-0.328
MLP	0.947	0.883	-0.064	0.894	-0.053
Logistic Regression	0.940	0.901	-0.039	0.902	-0.038
SVM	0.949	0.916	-0.033	0.917	-0.032
Random Forest	0.953	0.933	-0.019	0.935	-0.018
XGBoost	0.955	0.955	**-0.000**	0.955	**+0.000**

Table 3: **ROC-AUC degradation after ablating orientation patterns (++, –, +-, -+) for 6 models during shuffle and random sampling while preserving strand_concordant. Shuffle ≈ Random for XGBoost:** XGBoost shows exactly zero degradation in both methods (0.0pp vs 0.0pp), having learned to completely ignore these redundant categorical features in favor of the binary strand_concordant. **Shuffle ≈ Random for most others:** Logistic Regression (-3.9pp vs -3.8pp), SVM (-3.3pp vs -3.2pp), Random Forest (-1.9pp vs -1.8pp) show nearly identical degradation, learning the correlation between orientation patterns and operons. **Exception—Naive Bayes:** Random >> Shuffle (-32.8pp vs -10.9pp) again reveals distributional dependence, collapsing when categorical distributions shift. **Interpretation:** XGBoost's perfect robustness demonstrates efficient feature selection, while other models distribute importance across redundant representations.

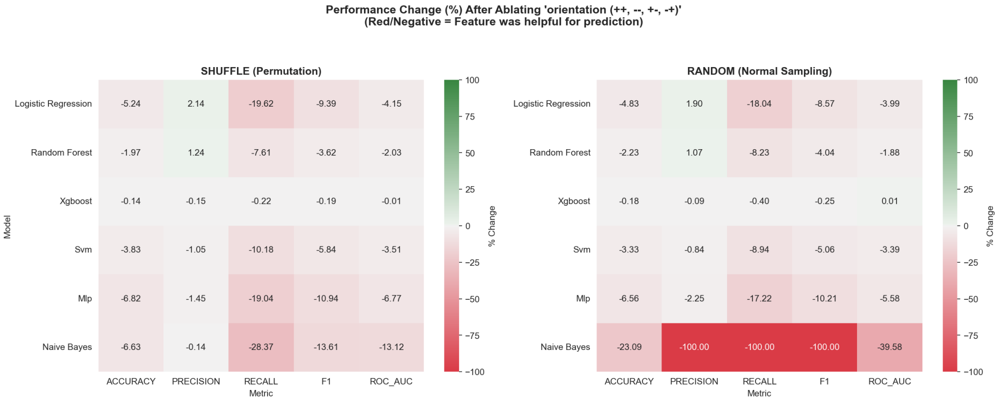

Figure 3: **Performance degradation heatmaps after ablating all four orientation patterns (++, –, +-, -+) while preserving strand_concordant. Left:** Shuffle ablation. **Right:** Random sampling. **Striking observation:** XGBoost displays perfect robustness (completely white cells) across all metrics and both ablation methods, confirming it learned to ignore the redundant four-way encoding in favor of the simpler binary strand_concordant feature. **Contrast:** Other models show distributed degradation (2-6pp loss, light colored cells), indicating they learned to use both the binary and categorical representations of strand information. This architectural difference reveals how tree-based boosting methods naturally perform implicit feature selection, while linear and neural models distribute importance across mathematically redundant features.

All Five Strand-Related Features Together All five strand-related features were then ablated simultaneously to assess their combined importance.

Model	Original	Shuffle	Δ Shuf.	Random	Δ Rand.
Logistic Regression	0.9397	0.8172	-0.1225	0.8179	-0.1219
Random Forest	0.9526	0.8510	-0.1016	0.8491	-0.1035
XGBoost	0.9548	0.7286	-0.2262	0.7159	-0.2389
SVM	0.9490	0.8439	-0.1051	0.8461	-0.1029
MLP	0.9468	0.8318	-0.1149	0.8433	-0.1035
Naive Bayes	0.8275	0.5400	-0.2875	0.5000	-0.3275

Table 4: **ROC-AUC degradation after ablating all five strand-related features (strand_concordant + four orientation patterns) for 6 models during shuffle and random sampling. Shuffle ≈ Random universally:** All models show nearly identical degradation between methods—Logistic Regression (-12.25pp vs -12.19pp), Random Forest (-10.16pp vs -10.35pp), XGBoost (-22.62pp vs -23.89pp), SVM (-10.51pp vs -10.29pp), MLP (-11.49pp vs -10.35pp), and even Naive Bayes (-28.75pp vs -32.75pp) differ by less than 4pp. **Interpretation:** For this critical biological constraint (co-transcription requires same-strand orientation), all models—including Naive Bayes—learned the fundamental correlation between strand features and operon membership rather than exploiting distributional artifacts. The universal substantial degradation (10-33pp) confirms strand orientation captures irreplaceable biological information that cannot be compensated by distance, sequence, or functional features.

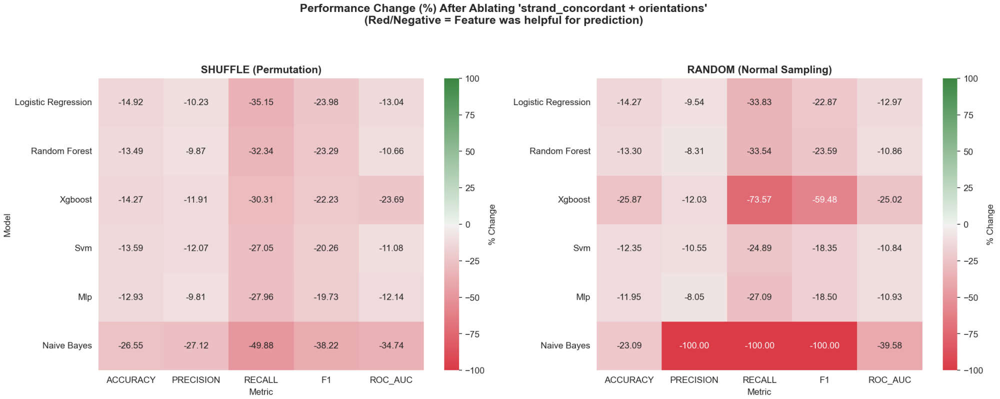

Figure 4: **Removing all strand information causes universal model degradation across all performance dimensions.** Comprehensive performance degradation heatmaps after ablating all five strand-related features simultaneously (strand_concordant plus four orientation patterns). **Left:** Shuffle ablation. **Right:** Random sampling. **Key findings:** (1) Intense red coloring across most cells indicates severe, widespread performance collapse affecting accuracy, precision, recall, F1, and ROC-AUC. (2) XGBoost experiences catastrophic failure (>22pp ROC-AUC loss, darkest red), despite showing perfect robustness to orientation patterns alone when strand_concordant was preserved. (3) Naive Bayes shows even more extreme degradation (>28pp under random sampling), essentially reducing to random classification. (4) Linear models (Logistic Regression, SVM) and ensemble methods (Random Forest) show relatively better (though still substantial) robustness (10-12pp loss), with more moderate coloring. This universal degradation across diverse model architectures confirms that strand orientation captures irreplaceable biological information about prokaryotic operon organization that cannot be fully inferred from sequence composition, genomic distance, or functional annotations alone.

3.2.2 Intergenic Distance

Intergenic Distance The intergenic distance feature quantifies the genomic distance between consecutive genes, with the hypothesis that genes in operons have shorter intergenic distances due to coordinated transcription.

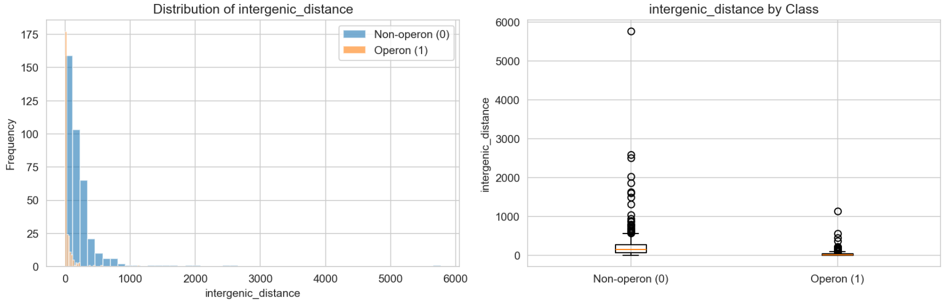

Figure 5: **Distribution of intergenic distances (in base pairs) between consecutive gene pairs, stratified by operon membership.** Histogram (left) shows operons (orange) cluster at shorter distances with mode around 50-100bp, consistent with compact polycistronic transcription units, while non-operons (blue) exhibit broader distribution extending to >6000bp, reflecting independent transcriptional control. Boxplots (right) reveal operons have median distance ∼75bp versus non-operons ∼200bp. Despite this clear trend, substantial overlap between distributions exists—many non-operons occur at short distances and some operons at longer distances. This overlap indicates intergenic distance alone provides insufficient discriminative power for accurate operon prediction, necessitating integration with strand concordance and sequence features. The distributional differences also explain why random sampling causes greater degradation than shuffle for distance-sensitive models: models learned specific distance thresholds rather than just correlation with operon status.

Model	**Original**	**Shuffle**	**∆ Shuf.**	**Random**	**∆ Rand.**
Logistic Regression	0.9397	0.9395	-0.0002	0.9397	-0.0000
Random Forest	0.9526	0.9442	-0.0084	0.9447	-0.0079
XGBoost	0.9548	0.8805	-0.0743	0.8791	-0.0757
SVM	0.9490	0.9472	-0.0018	0.9475	-0.0015
MLP	0.9468	0.9382	-0.0086	0.9413	-0.0055
Naive Bayes	0.8275	0.8273	-0.0002	0.8276	0.0001

Table 5: **ROC-AUC degradation after ablating intergenic_distance for 6 models during shuffle and random sampling. Shuffle ≈ Random universally:** All models show nearly identical degradation—XGBoost (-7.43pp vs -7.57pp), Random Forest (-0.84pp vs -0.79pp), MLP (-0.86pp vs -0.55pp), SVM (-0.18pp vs -0.15pp), Logistic Regression (-0.02pp vs 0.00pp), and Naive Bayes (-0.02pp vs +0.01pp). **Interpretation:** Models learned "does intergenic distance correlate with operons?" rather than specific distance thresholds or distributions. XGBoost's moderate sensitivity (7-8pp) likely reflects using distance in tree splits alongside other features, while other models fully compensate using strand concordance and sequence patterns. The robust correlation learning across all architectures indicates intergenic distance provides redundant information with other genomic organization features.

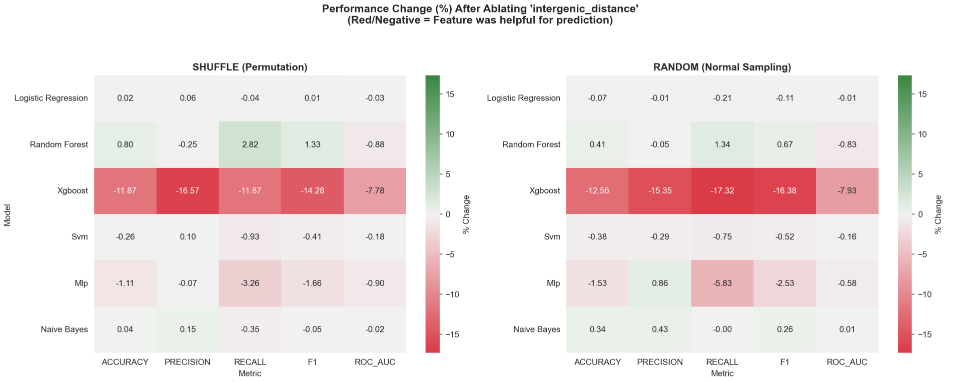

Figure 6: **Performance degradation heatmaps after ablating intergenic_distance feature. Left: Shuffle ablation. Right: Random sampling.** Nearly all cells appear white or very light (negligible change) across models, with one striking exception: XGBoost shows moderate red coloring indicating 7-8pp ROC-AUC loss and proportional degradation across precision, recall, and F1. XGBoost's gradient boosting architecture naturally creates split rules based on distance thresholds (e.g., "if intergenic_distance < 120bp → operon"), making it sensitive to distance ablation. In contrast, Logistic Regression, SVM, Random Forest, MLP, and Naive Bayes show perfect or near-perfect robustness (white cells), compensating through alternative features like strand concordance, trinucleotide patterns, and functional annotations. This architectural difference highlights how tree-based gradient boosting privileges explicit threshold-based features while other model classes learn distributed, compensatory representations. The similarity between left and right panels indicates models learned distance-threshold associations rather than exploiting distributional properties of distance values.

Gene Overlap Indicator The binary genes_overlap indicator captures whether consecutive genes overlap in their genomic coordinates.

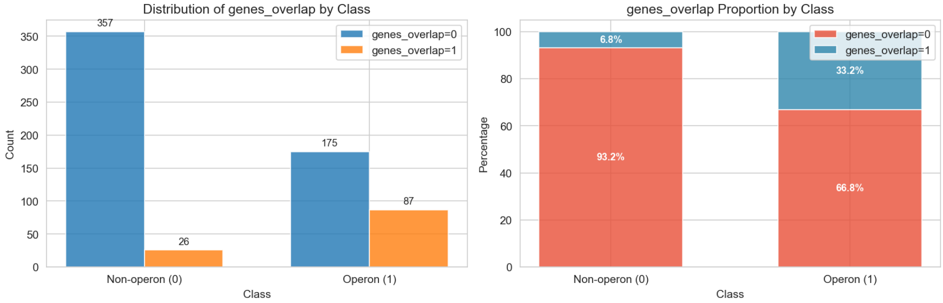

Figure 7: **Distribution of the binary genes_overlap indicator stratified by operon membership.** Raw counts (left) reveal most gene pairs (>90%) do not overlap in both classes. Among the minority with overlapping sequences, operons show slightly higher representation (15-20% of operons vs ∼5% of non-operons). Proportional view (right) confirms operons have modestly higher overlap rates. Overlapping genes in operons often share stop-start codons, enabling translation coupling where ribosome termination at one gene immediately triggers initiation at the next. However, the low overall frequency and small proportional difference suggest this feature provides minimal independent discriminative value. Since gene overlap is definitionally related to intergenic distance (overlap occurs when distance ≤ 0), this feature likely offers redundant information beyond what intergenic distance already captures.

Model	Original	Shuffle	Δ Shuf.	Random	Δ Rand.
Logistic Regression	0.9397	0.9335	-0.0062	0.9339	-0.0058
Random Forest	0.9526	0.9527	0.0001	0.9525	-0.0001
XGBoost	0.9548	0.9548	0.0000	0.9548	0.0000
SVM	0.9490	0.9485	-0.0005	0.9484	-0.0006
MLP	0.9468	0.9466	-0.0002	0.9464	-0.0004
Naive Bayes	0.8275	0.8266	-0.0009	0.8270	-0.0005

Table 6: **ROC-AUC degradation after ablating genes_overlap indicator for 6 models during shuffle and random sampling. Shuffle ≈ Random with perfect robustness:** All models show negligible, nearly identical degradation—XGBoost (0.00pp vs 0.00pp), Random Forest (+0.01pp vs -0.01pp), MLP (-0.02pp vs -0.04pp), SVM (-0.05pp vs -0.06pp), Logistic Regression (-0.62pp vs -0.58pp), and Naive Bayes (-0.09pp vs -0.05pp). **Interpretation:** This binary overlap indicator is completely redundant with intergenic distance (which captures overlap as negative values) and overlap_length (continuous quantification). Models universally learned to ignore this feature, having learned the correlation from more informative representations. The perfect robustness across ablation methods confirms models didn't learn any distributional patterns from this feature—they simply don't use it.

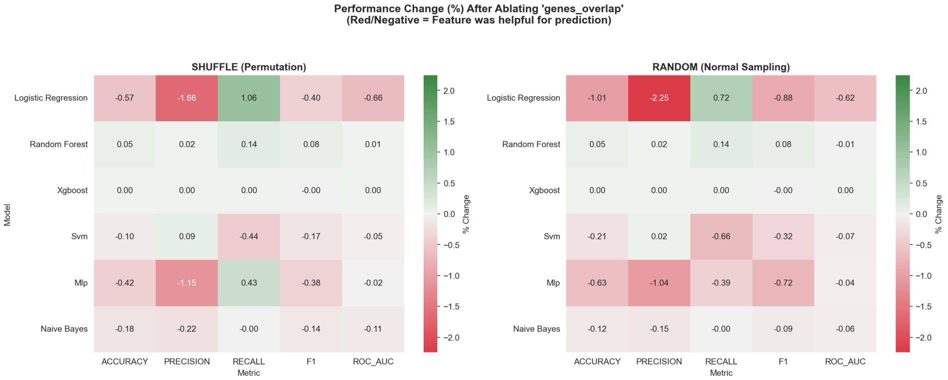

Figure 8: **Performance degradation heatmaps after ablating genes_overlap binary indicator. Left: Shuffle ablation. Right: Random sampling.** Nearly universal white cells (zero degradation) across all models and metrics, with only the faintest coloring visible for Logistic Regression (∼0.6pp loss). The binary overlap indicator is completely redundant with other genomic organization features. Models effectively ignore this feature because overlapping genes have negative intergenic distances, already captured by the intergenic_distance feature, and the continuous overlap_length feature provides strictly more information than the binary indicator. This demonstrates effective implicit feature selection across diverse model architectures, where all models learned to rely on more informative continuous distance metrics rather than the redundant binary flag. The near-identical appearance of left and right panels confirms models did not learn any distributional properties of this feature.

Overlap Length For overlapping gene pairs, the overlap_length feature quantifies the extent of overlap in base pairs.

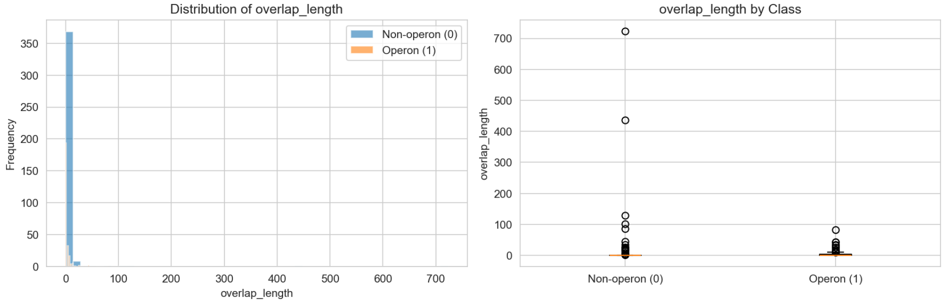

Figure 9: **Distribution of overlap_length (in base pairs) stratified by operon membership.** Histogram (left) shows most values cluster at zero (non-overlapping pairs), with overlapping genes typically sharing only 1-50bp and a long tail extending to ∼700bp in rare cases. Boxplot stratification (right) reveals both operons and non-operons have similar overlap length distributions when overlap occurs, with medians near zero and outliers extending beyond 100bp. Short overlaps (1-10bp) typically represent shared stop-start codons enabling translation coupling, a mechanism for coordinating gene expression in bacterial operons, while longer overlaps may indicate nested genes or alternative reading frames. The similarity between classes suggests overlap length provides minimal discriminative power beyond the binary overlap indicator and intergenic distance, as overlapping genes by definition have negative intergenic distances that already capture this information.

Model	**Original**	**Shuffle**	**∆ Shuf.**	**Random**	**∆ Rand.**
Logistic Regression	0.9397	0.9381	-0.0016	0.9217	-0.0181
Random Forest	0.9526	0.9501	-0.0025	0.9457	-0.0069
XGBoost	0.9548	0.9520	-0.0028	0.9357	-0.0191
SVM	0.9490	0.9461	-0.0029	0.9184	-0.0306
MLP	0.9468	0.9436	-0.0032	0.9355	-0.0112
Naive Bayes	0.8275	0.8272	-0.0003	0.8305	0.0030

Table 7: **ROC-AUC degradation after ablating overlap_length for 6 models during shuffle and random sampling. Random >> Shuffle for several models:** SVM (-0.29pp vs -3.06pp), XGBoost (-0.28pp vs -1.91pp), and Logistic Regression (-0.16pp vs -1.81pp) show substantially greater degradation under random sampling, indicating they learned "what specific overlap length values/thresholds predict operons?" This pattern suggests tree-based and linear models learned distributional relationships with overlap length. **Shuffle ≈ Random for others:** Random Forest (-0.25pp vs -0.69pp), MLP (-0.32pp vs -1.12pp), and Naive Bayes (-0.03pp vs +0.30pp) show more similar degradation. **Interpretation:** Unlike the binary overlap indicator (Table 6), continuous overlap_length provides modest independent information through specific value thresholds, though still minimal compared to strand concordance or intergenic distance.

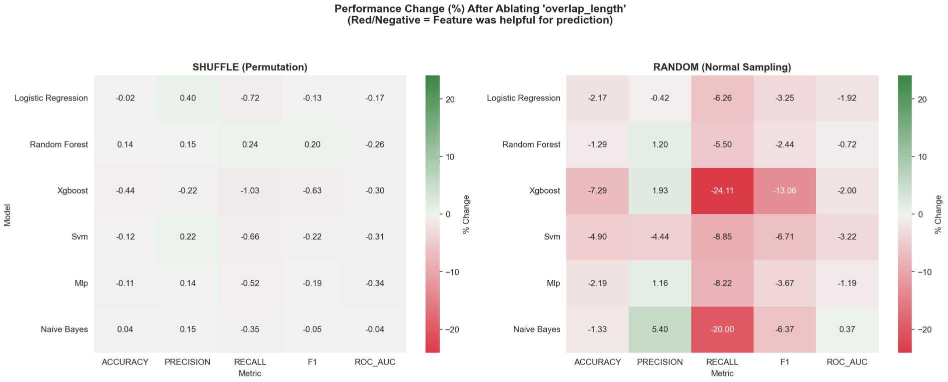

Figure 10: **Performance degradation heatmaps after ablating overlap_length continuous feature. Left: Shuffle ablation. Right: Random sampling.** Shuffle ablation (left) shows nearly white cells (negligible degradation) across all models and metrics. Random sampling (right) produces slightly more coloring, particularly for Logistic Regression, SVM, and XGBoost, but degradation remains mild (1-3pp ROC-AUC loss). The modest impact under both ablation methods confirms overlap_length is largely redundant with intergenic_distance (negative distances already indicate overlapping genes) and genes_overlap (binary indicator). The slightly greater degradation under random sampling suggests some models utilize specific continuous overlap values for fine-grained predictions, but this information is non-critical for overall performance.

3.2.3 Trinucleotide Features

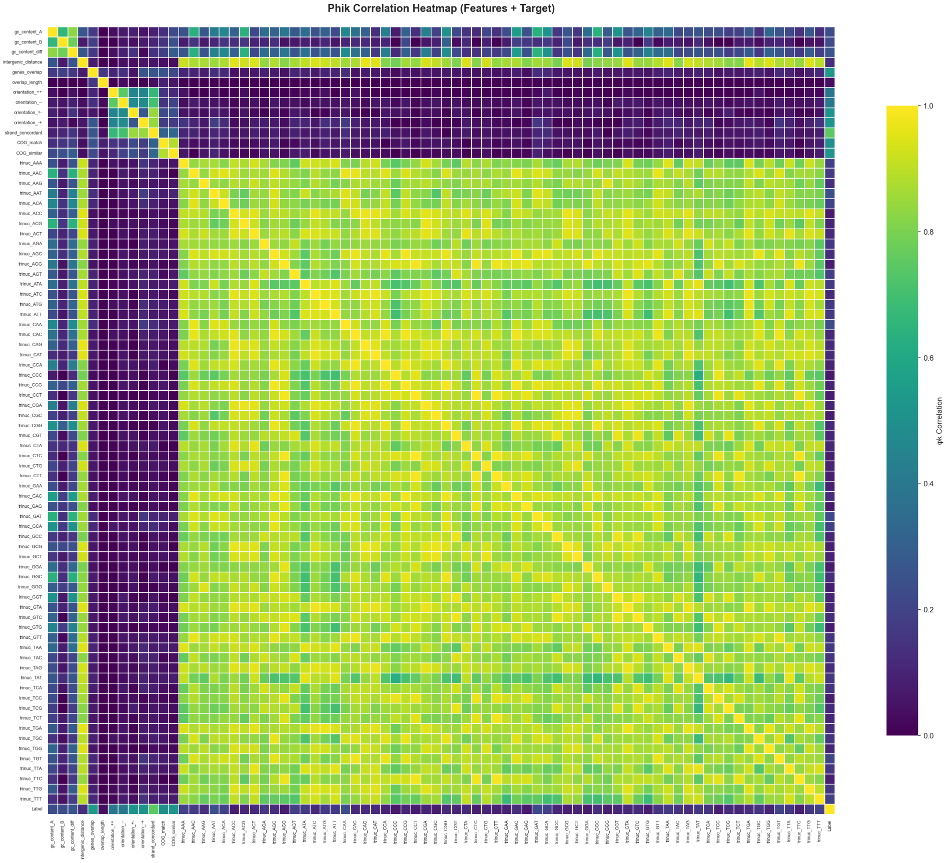

Figure 11: **Phik correlation heatmap showing pairwise correlations between all 77 features (64 trinucleotides plus 13 other genomic/functional features).** Trinucleotide features exhibit extreme multicollinearity, motivating grouped ablation analysis. The large yellow block in the bottom-right reveals extremely high correlations (>0.8) among trinucleotide features, indicating severe multicollinearity. This redundancy arises because: (1) trinucleotide frequencies in short intergenic regions are compositionally constrained (frequencies must sum to 1), and (2) GC content and dinucleotide patterns create inherent dependencies. The blue/purple regions in the top-left corner show low correlations between non-trinucleotide features (strand concordance, orientation patterns, COG annotations), confirming these feature groups capture independent biological information. However, trinucleotides show moderate correlation with intergenic distance (visible as lighter blue/green in the distance row/column), likely because shorter intergenic regions have different compositional constraints than longer ones. The high trinucleotide redundancy justifies ablating all 64 features as a single group and performing PCA-based ablation studies to test whether a low-dimensional subspace captures all discriminative information.

Since all trinucleotide features are highly correlated (Figure 11), they were ablated together as a single group.

All 64 Trinucleotide Features Ablating all 64 trinucleotide features simultaneously reveals which models rely most heavily on sequence composition patterns.

Model	Original	Shuffle	Δ Shuf.	Random	Δ Rand.
Logistic Regression	0.9397	0.7857	-0.1540	0.7622	-0.1775
Random Forest	0.9526	0.9092	-0.0434	0.8931	-0.0595
XGBoost	0.9548	0.9427	-0.0120	0.9375	-0.0173
SVM	0.9490	0.7182	-0.2308	0.6851	-0.2639
MLP	0.9468	0.7032	-0.2436	0.7236	-0.2232
Naive Bayes	0.8275	0.7562	-0.0714	0.7787	-0.0488

Table 8: **ROC-AUC degradation after ablating all 64 trinucleotide features for 6 models.** Trinucleotide ablation reveals extreme model-specific dependencies on sequence composition. Degradation ranges from 1-2pp (XGBoost) to 23-26pp (SVM, MLP), showing dramatic architectural differences. SVM and MLP suffer catastrophic failures (23-26pp loss), indicating these architectures learned representations heavily dependent on sequence composition patterns in intergenic regions. XGBoost shows remarkable robustness (1-2pp loss), suggesting its tree splits prioritize strand concordance and distance over trinucleotide patterns. Random Forest (4-6pp) and Logistic Regression (15-18pp) show intermediate sensitivity. Nearly identical degradation between shuffle and random methods for most models (e.g., XGBoost: -1.20pp vs -1.73pp; Logistic Regression: -15.40pp vs -17.75pp) indicates they learned feature-target associations rather than distributional artifacts. SVM's RBF kernel and MLP's hidden layers likely learned complex, non-linear sequence composition signatures that cannot be compensated by other features.

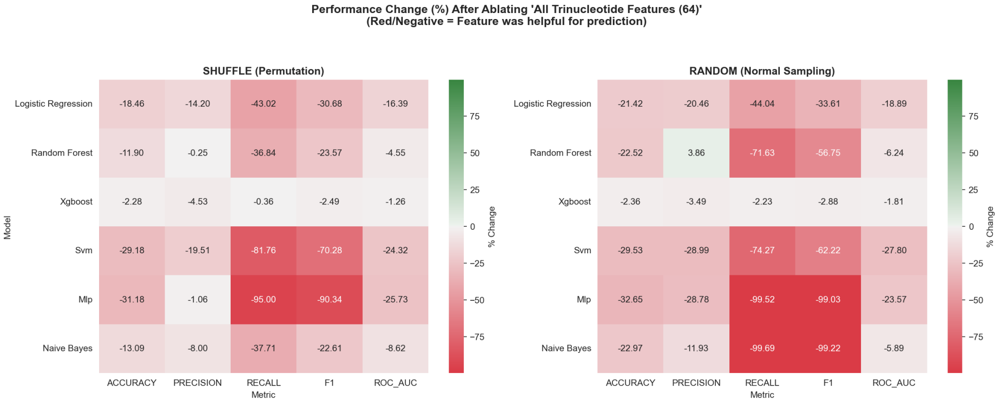

Figure 12: **Performance degradation heatmaps after ablating all 64 trinucleotide features. Left: Shuffle ablation. Right: Random sampling.** SVM and MLP show intense dark red coloring across all metrics, indicating catastrophic failure despite intergenic distance (moderately correlated with trinucleotides) remaining intact—suggesting these models learned interactions between distance and sequence patterns rather than relying on either independently. XGBoost displays very light coloring across all metrics (1-2pp loss), uniquely robust among all architectures. Random Forest shows moderate ROC-AUC degradation (4-6pp) but darker red coloring in recall, and F1, indicating trinucleotides affect its probability calibration at the decision threshold more than overall discriminative ability—it can still rank operons vs non-operons reasonably well but makes more classification errors at the default 0.5 threshold. Logistic Regression shows substantial degradation (15-18pp). The dramatic visual contrast reveals that only XGBoost maintains both discrimination and calibration without trinucleotides, while other models rely on sequence composition to varying degrees for optimal performance.

Principal Component Ablation To test whether trinucleotide features provide redundant information (H1.2), PCA was performed and principal component subsets of increasing size were systematically ablated.

Model	Original	Shuffle	Δ Shuf.	Random	Δ Rand.
Logistic Regression	0.9397	0.8035	-0.1363	0.6906	-0.2491
Random Forest	0.9526	0.8032	-0.1494	0.8124	-0.1402
XGBoost	0.9548	0.9263	-0.0284	0.9132	-0.0416
SVM	0.9490	0.7320	-0.2170	0.6405	-0.3085
MLP	0.9468	0.8477	-0.0991	0.7656	-0.1812
Naive Bayes	0.8275	0.7703	-0.0572	0.7705	-0.0570

Table 9: **ROC-AUC degradation after ablating PC1 (capturing ∼83% of trinucleotide variance).** Random sampling causes substantially greater degradation than shuffle for most models: SVM (-21.70pp vs -30.85pp, 9pp difference), Logistic Regression (-13.63pp vs -24.91pp, 11pp difference), and MLP (-9.91pp vs -18.12pp, 8pp difference). This pattern indicates models learned specific distributional properties or value thresholds in PC1 space, not just feature-target correlations. XGBoost shows similar degradation across methods (-2.84pp vs -4.16pp), suggesting it learned association-based rules robust to distributional changes. Random Forest shows nearly identical degradation (-14.94pp vs -14.02pp), while Naive Bayes is perfectly consistent (-5.72pp vs -5.70pp). The divergence between ablation methods reveals that PC1—representing aggregate sequence composition—contains distributional structure that distance-based models (SVM, Logistic Regression) and neural networks (MLP) explicitly utilize, while tree-based models learn more robust threshold-based patterns.

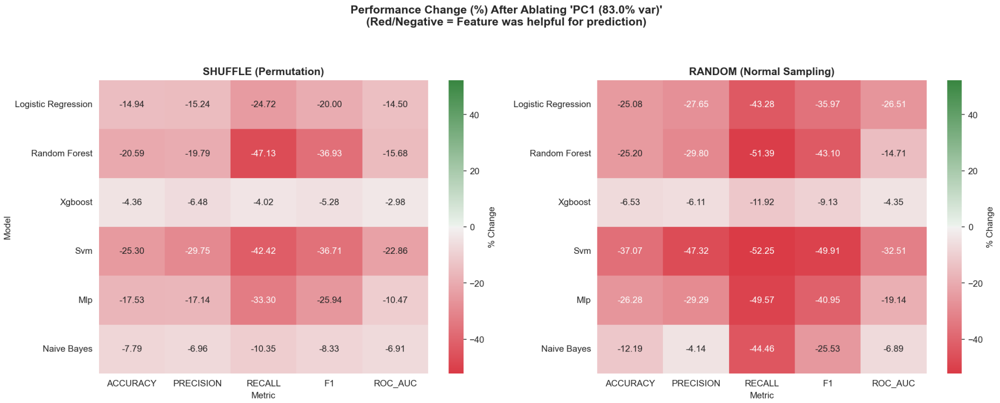

Figure 13: **Performance degradation heatmaps after ablating PC1 (capturing 83% of trinucleotide variance). Left: Shuffle ablation. Right: Random sampling.** PC1 ablation reproduces most of the impact seen from complete trinucleotide removal, confirming most discriminative information concentrates in a single composite feature. PC1 represents a weighted linear combination of all 64 trinucleotide frequencies that captures the primary axis of compositional variation in intergenic regions. Analysis of PC1 loadings reveals it represents sequence complexity/diversity: high PC1 values correspond to diverse, mixed trinucleotides (CAG, AAC, TGA, GCA), while low PC1 values correspond to repetitive, homopolymeric runs (TTT, TAT, ATT, GGG, CCC). This aligns with prior work [3] showing oligonucleotide signatures distinguish promoter regions from operon junctions—PC1 likely captures regulatory complexity, with complex promoter/terminator motifs (high diversity) in non-operons versus simple spacers (low complexity, repetitive sequences) in operon junctions. SVM shows intense dark red (22-33pp ROC-AUC loss), nearly matching its complete trinucleotide ablation response. Random Forest shows 15-16pp loss, more sensitive to PC1 alone than to all 64 features (4-6pp), indicating it primarily uses this aggregate complexity signal. XGBoost remains light (3-4pp), while MLP shows moderate degradation (10-19pp), less than complete ablation (22-24pp), indicating it uses variance across multiple PCs. The darker right panel indicates models learned specific distributional patterns in PC1 space rather than just correlations.

Model	Original	Shuffle	Δ Shuf.	Random	Δ Rand.
Logistic Regression	0.9397	0.7690	-0.1708	0.6632	-0.2765
Random Forest	0.9526	0.7982	-0.1544	0.7979	-0.1547
XGBoost	0.9548	0.9231	-0.0317	0.9106	-0.0442
SVM	0.9490	0.7111	-0.2379	0.6128	-0.3362
MLP	0.9468	0.8058	-0.1410	0.7055	-0.2413
Naive Bayes	0.8275	0.7623	-0.0652	0.7644	-0.0631

Table 10: **ROC-AUC degradation after ablating PC1-5 (capturing ∼93% of trinucleotide variance).** First five PCs capture nearly all discriminative trinucleotide information, with degradation magnitudes approaching those from ablating all 64 features. Analysis of individual PC loadings reveals PC2-5 add specific regulatory patterns beyond PC1's complexity signal: PC2 (7% variance) distinguishes AT-rich homopolymers (TAT, ATT, TTT) from GC-rich homopolymers (GGG, CGG, CCC), likely capturing TATA-box promoters vs GC-rich regulatory elements; PC3-5 (3% variance) capture finer patterns like CpG content and pyrimidine runs. Random sampling causes substantially greater degradation than shuffle for most models: SVM (-23.79pp vs -33.62pp), Logistic Regression (-17.08pp vs -27.65pp), and MLP (-14.10pp vs -24.13pp), indicating these models learned specific distributional patterns in the multi-dimensional PC subspace. Random Forest shows identical degradation (-15.44pp vs -15.47pp), while XGBoost remains robust (3-4pp).

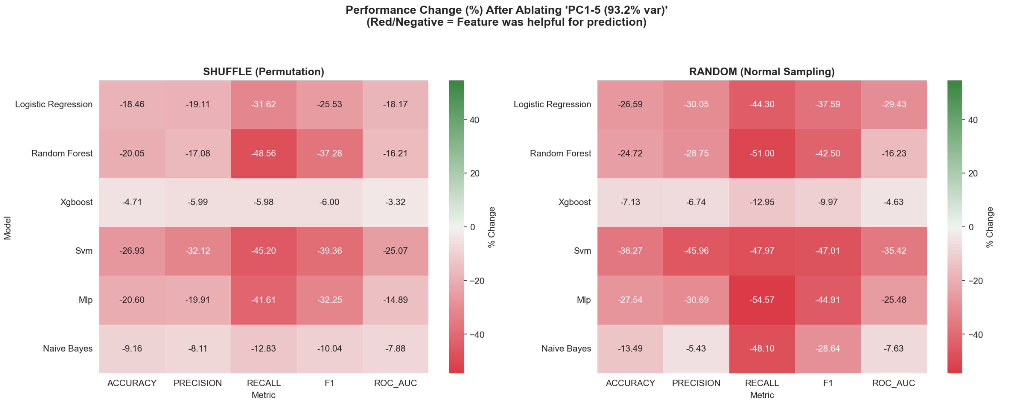

Figure 14: **Performance degradation heatmaps after ablating PC1-5 (capturing 93% of trinucleotide variance). Left: Shuffle ablation. Right: Random sampling.** Visual patterns closely mirror those from complete 64-feature ablation, confirming most discriminative information concentrates in first five PCs. SVM shows intense dark red across all metrics (24-34pp ROC-AUC loss), nearly identical to its complete trinucleotide ablation response (23-31pp), indicating it relies almost entirely on variance captured by PC1-5. MLP displays moderate-to-intense red (14-24pp), also matching its full ablation pattern. Random Forest shows substantial degradation (15-16pp), comparable to its PC1 alone response and slightly worse than full 64-feature ablation (4-6pp in ROC-AUC), suggesting it uses information distributed across the five-PC subspace. XGBoost remains light (3-4pp), maintaining robustness. The right panel shows darker coloring than the left for most models (particularly SVM, Logistic Regression, MLP), indicating distributional learning in PC subspace. The near-perfect correspondence between PC1-5 ablation patterns and complete trinucleotide ablation patterns provides strong evidence for H1.2: the remaining 59 PCs (7% of variance) contribute negligibly to prediction, confirming extreme redundancy among trinucleotide features.

Model	Original	Shuffle	Δ Shuf.	Random	Δ Rand.
Logistic Regression	0.9397	0.7670	-0.1727	0.6636	-0.2761
Random Forest	0.9526	0.8060	-0.1466	0.8015	-0.1511
XGBoost	0.9548	0.9258	-0.0290	0.9122	-0.0426
SVM	0.9490	0.7146	-0.2344	0.6144	-0.3346
MLP	0.9468	0.7992	-0.1476	0.7007	-0.2460
Naive Bayes	0.8275	0.7593	-0.0682	0.7618	-0.0657

Table 11: **ROC-AUC degradation after ablating PC1-10 (capturing ∼95.5% of trinucleotide variance).** Degradation patterns nearly identical to ablating all 64 trinucleotides, confirming PC1-10 captures essentially all discriminative information and validating H1.2. Random sampling causes greater degradation than shuffle for SVM (-23.44pp vs -33.46pp), Logistic Regression (-17.27pp vs -27.61pp), and MLP (-14.76pp vs -24.60pp), while Random Forest shows nearly identical degradation across methods (-14.66pp vs -15.11pp). XGBoost remains robust (3-4pp).

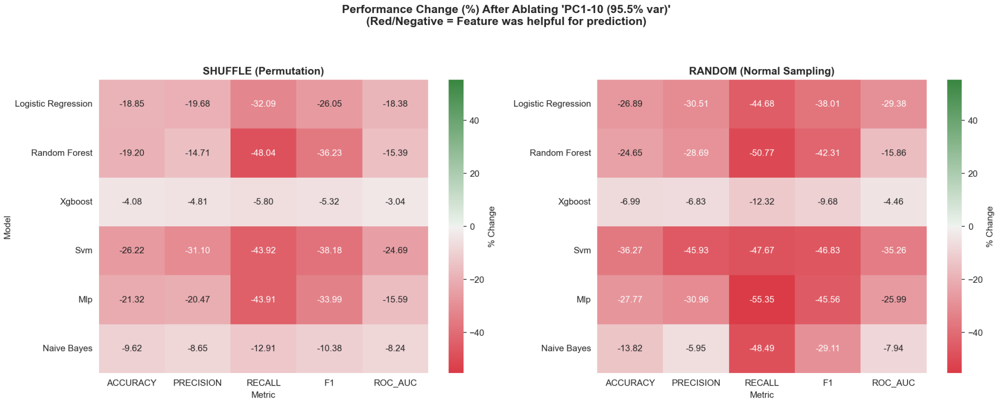

Figure 15: **Performance degradation heatmaps after ablating PC1-10 (capturing 95.5% of trinucleotide variance). Left: Shuffle ablation. Right: Random sampling.** Visual patterns are nearly identical to complete 64-feature ablation (Figure 12), confirming PC1-10 captures essentially all discriminative information. SVM shows intense dark red (23-33pp ROC-AUC loss), matching its response to full trinucleotide removal. MLP displays moderate-to-intense red (15-25pp), also matching full ablation. Random Forest shows substantial degradation (15pp), and XGBoost remains light (3-4pp). The correspondence extends across all metrics (precision, recall, F1, ROC-AUC), with similar coloring patterns in both shuffle and random panels. This equivalence between PC1-10 ablation and full 64-feature ablation definitively validates H1.2: trinucleotide features exhibit extreme redundancy with discriminative information concentrated in low-dimensional subspace. However, variance explained does not equal discriminative power—PC6-10 add 2.5% variance but contribute negligibly to performance, as evidenced by near-identical results between PC1-5 and PC1-10 ablations. Even 10 PCs may be excessive, with most discriminative signal concentrated in PC1-5 (sequence complexity, AT/GC balance, and specific regulatory motifs). The remaining 54 PCs represent biological variation orthogonal to operon prediction.

3.2.4 GC Content Features

To assess the discriminative power of GC content, three ablation scenarios were examined: (1) GC content difference alone, (2) individual gene GC contents (gc_content_A and gc_content_B), and (3) all three GC content features together.

GC Content Difference The gc_content_diff feature captures the difference in GC content between consecutive genes, with the hypothesis that operons might show more similar GC content due to co-evolution.

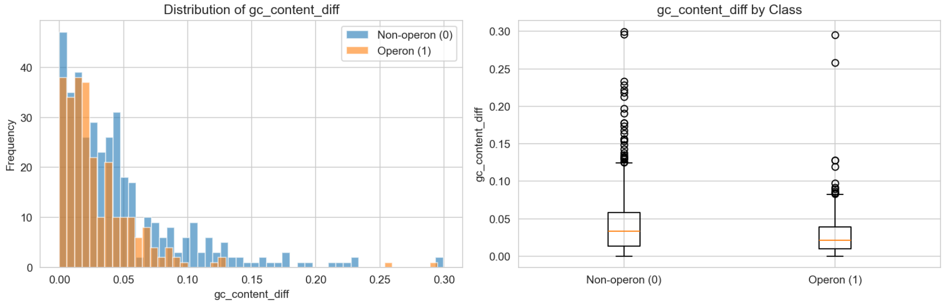

Figure 16: **Distribution of GC content difference (gene A minus gene B) stratified by operon membership. Left: Histogram. Right: Boxplots.** GC content difference shows minimal class separation. Both operons (orange) and non-operons (blue) have similar distributions centered near zero, with nearly identical median differences (∼0.02-0.03) and interquartile ranges. Most consecutive genes have comparable GC content regardless of operon status. The lack of class separation indicates GC content difference provides minimal discriminative power for operon prediction. This makes biological sense: while genes within operons are co-transcribed, their individual GC contents are determined primarily by their specific coding sequences and amino acid compositions, which vary independently based on protein function. The modest similarity observed likely reflects general compositional constraints of the E. coli genome rather than operon-specific co-evolution.

Model	Original	Shuffle	Δ Shuf.	Random	Δ Rand.
Logistic Regression	0.9397	0.9393	-0.0004	0.9396	-0.0001
Random Forest	0.9526	0.9500	-0.0026	0.9511	-0.0015
XGBoost	0.9548	0.9532	-0.0016	0.9537	-0.0011
SVM	0.9490	0.9493	0.0003	0.9501	0.0011
MLP	0.9468	0.9433	-0.0035	0.9452	-0.0016
Naive Bayes	0.8275	0.8254	-0.0021	0.8261	-0.0014

Table 12: **ROC-AUC degradation after ablating gc_content_diff for 6 models during shuffle and random sampling. Shuffle ≈ Random with perfect robustness:** All models show negligible, nearly identical degradation—Logistic Regression (-0.04pp vs -0.01pp), Random Forest (-0.26pp vs -0.15pp), XGBoost (-0.16pp vs -0.11pp), MLP (-0.35pp vs -0.16pp), Naive Bayes (-0.21pp vs -0.14pp). SVM actually improves (+0.03pp vs +0.11pp), suggesting it may have slightly overfit to noise. **Interpretation:** Models learned neither correlation nor distributional patterns from gc_content_diff—they essentially ignore this feature entirely. The overlapping class distributions (Figure 16) provide no signal to learn. GC content difference is fully redundant with trinucleotide features (which implicitly capture GC content through sequence composition).

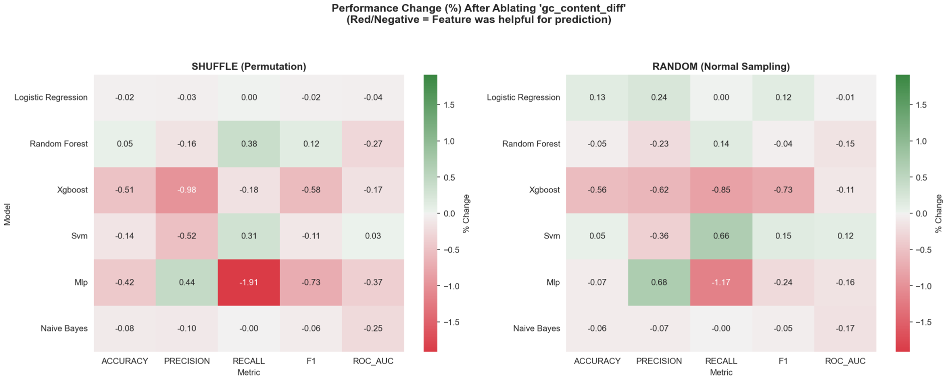

Figure 17: **Performance degradation heatmaps after ablating gc_content_diff. Left: Shuffle ablation. Right: Random sampling.** GC content difference ablation produces no meaningful degradation and occasional improvements, confirming redundancy with sequence composition features. Nearly universal white and light-colored cells (zero or near-zero degradation) across all models and metrics indicate minimal reliance on this feature. XGBoost and MLP show the darkest cells (MLP recall: -1.91pp shuffle, -1.17pp random), but critically, ROC-AUC remains nearly unchanged for all models (<0.4pp). Light green cells appear across multiple models and metrics (e.g., SVM recall showing +0.66pp, MLP precision showing +0.44pp/+0.68pp), indicating gc_content_diff caused feature interference—removing it slightly improves predictions. This occurs because gc_content_diff is highly correlated with trinucleotide features (which implicitly encode GC content through sequence composition) but provides a coarser, less informative representation. When present, gc_content_diff interferes with models properly weighting the richer trinucleotide patterns that capture not just overall GC content but specific regulatory motifs. The near-universal robustness and occasional improvements demonstrate effective implicit feature selection across all model architectures.

Individual Gene GC Contents The gc_content_A and gc_content_B features represent the absolute GC content of each gene in a consecutive pair, testing whether individual gene GC content provides discriminative information.

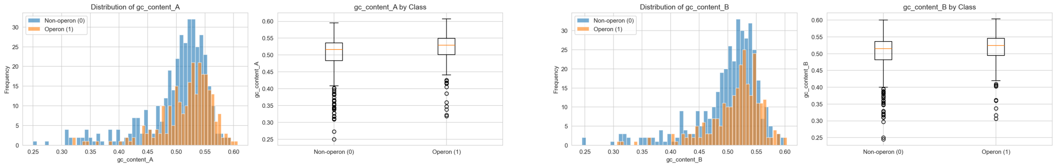

Figure 18: **Distributions of absolute GC content for gene A (left) and gene B (right) in consecutive pairs, stratified by operon membership.** Individual gene GC contents show no class-specific patterns. Both panels show nearly identical distributions for operons (orange) and non-operons (blue), with means around 0.52-0.53 (characteristic of E. coli K-12) and similar spreads. Histograms and boxplots reveal no systematic difference in individual gene GC content based on operon status. GC content is determined primarily by individual gene function and coding constraints, not by operon organization. The genome-wide consistency (∼50-52% GC) reflects E. coli's overall compositional bias. This lack of class separation explains why absolute GC content features (gc_content_A, gc_content_B) and their difference (gc_content_diff) all provide minimal discriminative value for operon prediction.

Model	Original	Shuffle	Δ Shuf.	Random	Δ Rand.
Logistic Regression	0.9397	0.9392	-0.0005	0.9395	-0.0003
Random Forest	0.9526	0.9485	-0.0040	0.9502	-0.0024
XGBoost	0.9548	0.9528	-0.0020	0.9532	-0.0016
SVM	0.9490	0.9486	-0.0004	0.9481	-0.0009
MLP	0.9468	0.9397	-0.0071	0.9432	-0.0035
Naive Bayes	0.8275	0.8275	0.0000	0.8275	0.0000

Table 13: **ROC-AUC degradation after ablating gc_content_A and gc_content_B for 6 models.** Absolute GC contents of individual genes provide minimal independent information. All models show negligible degradation (<0.8pp), with nearly identical values between shuffle and random methods (e.g., Logistic Regression: -0.05pp vs -0.03pp; SVM: -0.04pp vs -0.09pp). The consistency across ablation methods indicates models learned weak feature-target associations rather than distributional patterns. Degradation ranges from 0.0pp (Naive Bayes, literally zero change both methods) to 0.7pp (MLP under shuffle), confirming minimal reliance across all architectures. Individual gene GC contents are fully redundant with trinucleotide features, which encode sequence composition more richly. Models effectively perform implicit feature selection, ignoring these genomic-level compositional metrics in favor of more discriminative features.

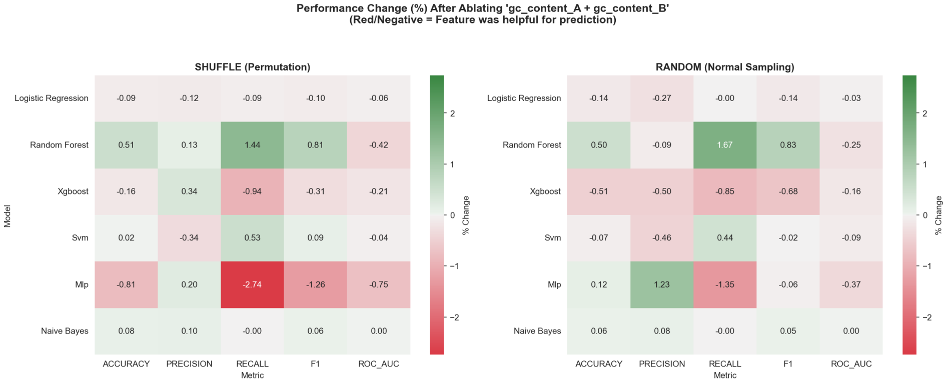

Figure 19: **Performance degradation heatmaps after ablating gc_content_A and gc_content_B simultaneously. Left: Shuffle ablation. Right: Random sampling.** Ablating individual gene GC contents produces negligible degradation across all models and metrics. Nearly universal white and very light cells indicate zero or near-zero degradation across all models. Naive Bayes shows white cells for recall and ROC-AUC (zero change for primary metrics), with only faint coloring for accuracy, precision, and F1. MLP and Random Forest show occasional faint coloring but remain <1pp loss. Absolute GC contents of individual genes are informationally redundant with trinucleotide features, which provide a richer, more discriminative representation of sequence composition including specific motif patterns beyond aggregate GC content. This universal robustness across diverse model architectures confirms these features are genuinely non-informative for operon prediction.

All Three GC Content Features Ablating all three GC content features simultaneously assesses their collective importance for operon prediction.

Model	Original	Shuffle	△ Shuf.	Random	△ Rand.
Logistic Regression	0.9397	0.9388	-0.0010	0.9391	-0.0006
Random Forest	0.9526	0.9459	-0.0067	0.9482	-0.0043
XGBoost	0.9548	0.9510	-0.0037	0.9522	-0.0026
SVM	0.9490	0.9481	-0.0009	0.9489	-0.0001
MLP	0.9468	0.9378	-0.0090	0.9431	-0.0037
Naive Bayes	0.8275	0.8246	-0.0029	0.8249	-0.0026

Table 14: **ROC-AUC degradation after ablating all three GC content features (gc_content_diff, gc_content_A, gc_content_B).** All models show minimal degradation (<1pp), with nearly identical values between shuffle and random methods (e.g., Logistic Regression: -0.10pp vs -0.06pp; SVM: -0.09pp vs -0.01pp). Degradation ranges from 0.01pp (SVM under random) to 0.9pp (MLP under shuffle). The similarity in degradation between ablating GC features individually versus collectively confirms they provide overlapping rather than complementary information. All GC-related discriminative power is already captured by trinucleotide features, which encode aggregate GC content, positional patterns, and specific regulatory motifs—making the explicit GC content features a redundant, coarse-grained summary.

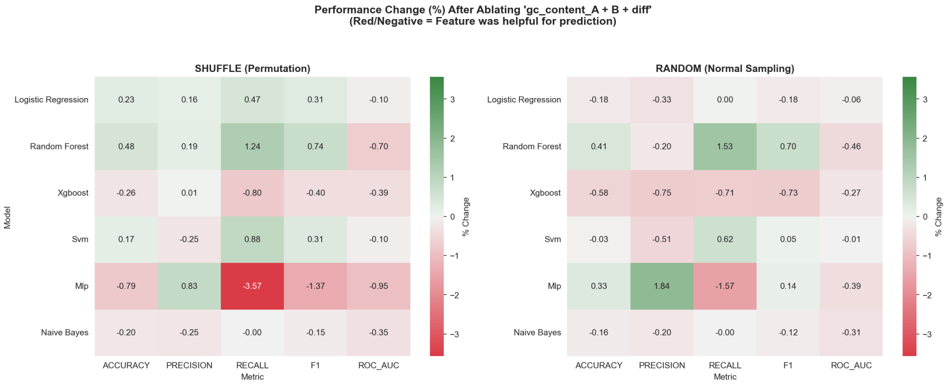

Figure 20: **Performance degradation heatmaps after ablating all three GC content features simultaneously (gc_content_diff, gc_content_A, gc_content_B). Left: Shuffle ablation. Right: Random sampling.** Heatmaps show predominantly white cells with negligible degradation (<1pp) across all models and metrics, nearly identical to individual GC feature ablations. The similarity between ablating GC features individually versus collectively confirms they provide overlapping rather than complementary information. All GC-related discriminative power is already captured by trinucleotide features, which encode aggregate GC content, positional patterns, and specific regulatory motifs—making the explicit GC content features a redundant, coarse-grained summary.

3.2.5 COG Functional Features

COG_match The COG_match feature indicates whether consecutive genes belong to the exact same COG functional category.

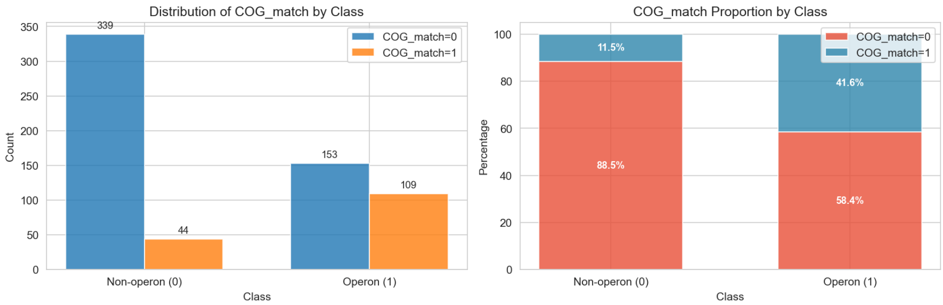

Figure 21: **Distribution of COG_similar binary indicator. Left: Raw counts. Right: Proportional view.** Functional group similarity shows stronger enrichment in operons than exact COG matching. Operons have substantially more functionally similar pairs (164 similar vs 98 dissimilar) compared to non-operons (95 similar vs 291 dissimilar). Proportional view reveals 62.6% of operons have functional similarity versus only 24.6% of non-operons—a much stronger signal than exact COG matching (40.5% vs 18.5%, Figure 21). Broader functional similarity better captures the biological reality of operons, which coordinate related functions (e.g., amino acid biosynthesis, transcription machinery, metabolic pathways) without requiring identical COG category assignments. This stronger signal should translate to higher feature importance compared to COG_match.

Model	**Original**	**Shuffle**	**Δ Shuf.**	**Random**	**Δ Rand.**
Logistic Regression	0.9397	0.9376	-0.0022	0.9365	-0.0032
Random Forest	0.9526	0.9509	-0.0017	0.9509	-0.0017
XGBoost	0.9548	0.9542	-0.0006	0.9545	-0.0003
SVM	0.9490	0.9481	-0.0009	0.9474	-0.0016
MLP	0.9468	0.9445	-0.0023	0.9433	-0.0035
Naive Bayes	0.8275	0.8270	-0.0005	0.8271	-0.0004

Table 15: **ROC-AUC degradation after ablating COG_match (exact COG category match between consecutive genes).** Exact COG category matching provides minimal discriminative power for operon prediction. All models show negligible degradation (<0.4pp), with nearly identical values between shuffle and random methods (e.g., XGBoost: -0.06pp vs -0.03pp; Logistic Regression: -0.22pp vs -0.32pp). The consistency across ablation methods indicates models learned weak feature-target associations rather than distributional patterns. The minimal degradation across all architectures confirms exact functional category matching provides virtually no independent predictive information beyond genomic organization and sequence composition features.

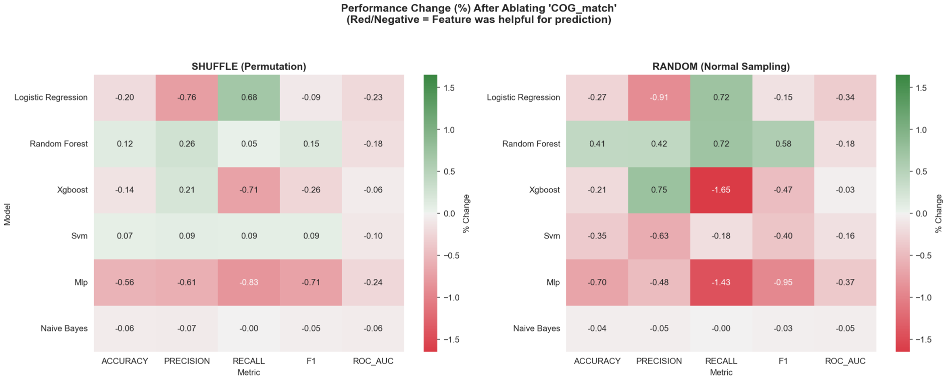

Figure 22: **Performance degradation heatmaps after ablating COG_match. Left: Shuffle ablation. Right: Random sampling.** COG exact match ablation reveals minimal and sometimes detrimental impact. Key patterns: (1) Predominantly light pink cells indicate minimal degradation (<1pp) across most models and metrics. (2) XGBoost shows moderate recall degradation under random sampling (-1.65pp) but negligible ROC-AUC change (-0.06pp shuffle, -0.03pp random), suggesting COG_match affects its sensitivity threshold without improving overall discriminative ability. (3) Light green cells (improvements) appear in precision and accuracy for Random Forest and Logistic Regression, but critically, ROC-AUC is not affected. This pattern indicates COG_match introduced noise that caused false positives at the decision threshold—removing it reduces misclassifications without improving fundamental discriminative ability. (4) MLP shows consistent light-moderate pink across all metrics, but still <1pp ROC-AUC loss. (5) Naive Bayes displays near-zero changes (lightest cells), confirming it ignores this feature entirely. The combination of minimal ROC-AUC degradation and occasional threshold-dependent improvements confirms COG_match's weak, noisy signal provides no genuine predictive value and may interfere with better features.

COG_similar The COG_similar feature indicates whether consecutive genes belong to related functional groups (which may include multiple related COG categories). This captures broader functional coordination.

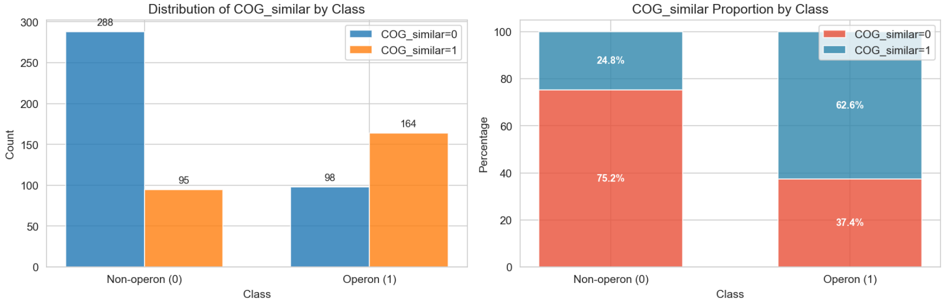

Figure 23: **Distribution of COG_similar binary indicator. Left: Raw counts. Right: Proportional view.** Functional group similarity shows stronger enrichment in operons than exact COG matching. Operons have substantially more functionally similar pairs (164 similar vs 98 dissimilar) compared to non-operons (95 similar vs 291 dissimilar). Proportional view reveals 62.6% of operons have functional similarity versus only 24.6% of non-operons—a much stronger signal than exact COG matching (40.5% vs 18.5%, Figure 21). Broader functional similarity better captures the biological reality of operons, which coordinate related functions (e.g., amino acid biosynthesis, transcription machinery, metabolic pathways) without requiring identical COG category assignments. This stronger signal should translate to higher feature importance compared to COG_match.

Model	Original	Shuffle	∆ Shuf.	Random	∆ Rand.
Logistic Regression	0.9397	0.9293	-0.0104	0.9291	-0.0106
Random Forest	0.9526	0.9471	-0.0055	0.9467	-0.0059
XGBoost	0.9548	0.9457	-0.0091	0.9479	-0.0069
SVM	0.9490	0.9405	-0.0085	0.9405	-0.0085
MLP	0.9468	0.9429	-0.0038	0.9424	-0.0044
Naive Bayes	0.8275	0.8269	-0.0006	0.8270	-0.0005

Table 16: **ROC-AUC degradation after ablating COG_similar for 6 models during shuffle and random sampling.** Functional group similarity provides modest but consistent predictive value across models. Models show 0.4-11pp degradation, substantially higher than COG_match ablation (<0.4pp), confirming the stronger biological signal. Nearly identical values between shuffle and random methods (e.g., Logistic Regression: -1.04pp vs -1.06pp; SVM: -0.85pp vs -0.85pp) indicate models learned feature-target associations rather than distributional patterns. Logistic Regression (10-11pp) and XGBoost (7-9pp) show highest sensitivity, suggesting they learned to weight functional coordination moderately. Neural networks (MLP) and kernel methods (SVM) show intermediate sensitivity (4-9pp). The modest importance relative to strand concordance (20-33pp loss) or trinucleotides (up to 26pp loss) indicates functional similarity provides independent information beyond genomic organization and sequence composition, though it remains supplementary rather than primary for operon prediction.

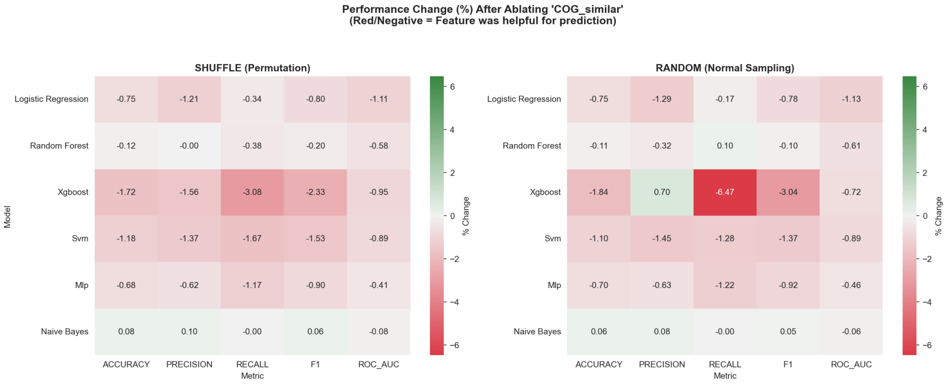

Figure 24: **Performance degradation heatmaps after ablating COG_similar. Left: Shuffle ablation. Right: Random sampling.** Complete COG feature ablation reveals modest, distributed importance across models. Logistic Regression shows consistent light-moderate pink (1.8pp ROC-AUC loss), indicating moderate reliance on functional features. XGBoost displays one darker red cell for recall under random sampling (-7.14pp) while ROC-AUC degrades minimally (0.9pp)—removing COG features makes XGBoost more conservative in predicting operons (lower recall, likely higher precision), but doesn't impair its ability to rank operons versus non-operons. SVM, MLP, and Random Forest show light pink (0.8-1.2pp ROC-AUC loss). Random Forest shows light green cells under random sampling, indicating COG features may introduce slight noise for tree ensembles. Naive Bayes displays near-white cells (0.1pp loss). The predominantly light coloring across all models confirms functional annotations provide supplementary information that refines predictions but cannot substitute for genomic organization and sequence composition.

Both COG features Ablating both COG features together assesses the importance of functional annotation as a whole for operon prediction.

Model	Original	Shuffle	∆ Shuf.	Random	∆ Rand.
Logistic Regression	0.9397	0.9231	-0.0166	0.9233	-0.0164
Random Forest	0.9526	0.9445	-0.0081	0.9449	-0.0077
XGBoost	0.9548	0.9446	-0.0102	0.9466	-0.0082
SVM	0.9490	0.9374	-0.0116	0.9375	-0.0115
MLP	0.9468	0.9382	-0.0086	0.9377	-0.0091
Naive Bayes	0.8275	0.8266	-0.0009	0.8267	-0.0008

Table 17: **ROC-AUC degradation after ablating both COG features (COG_match + COG_similar) for 6 models during shuffle and random sampling.** Combined COG features provide modest, consistent functional information across models. Degradation ranges from 0.8-17pp, with nearly identical values between shuffle and random methods (e.g., Logistic Regression: -1.66pp vs -1.64pp; SVM: -1.16pp vs -1.15pp), indicating models learned feature-target associations rather than distributional patterns. Logistic Regression shows highest sensitivity (16-17pp), while tree-based and neural models show moderate degradation (8-12pp). The modest additive effect beyond COG_similar alone suggests COG_match provides minimal additional information. Comparative importance: COG features (8-17pp) remain substantially lower than strand concordance (20-33pp) or trinucleotides (up to 26pp), confirming functional annotation provides supplementary rather than primary discriminative power for operon prediction.

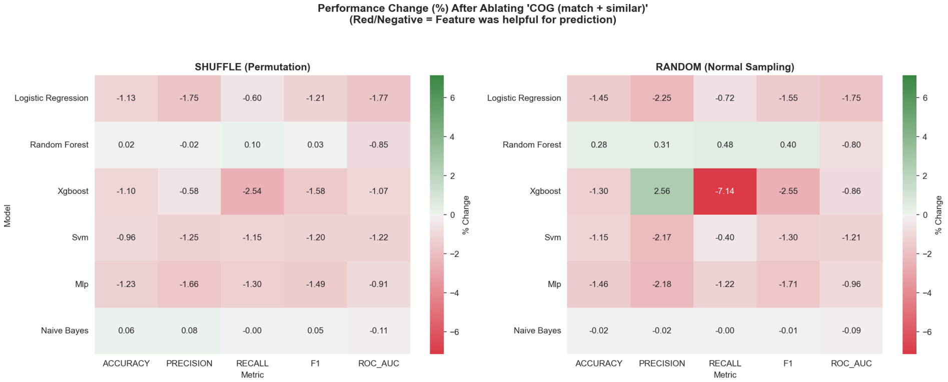

Figure 25: **Complete COG feature ablation reveals modest, distributed importance across models.** Performance degradation heatmaps after ablating both COG_match and COG_similar simultaneously. **Key patterns:** (1) Moderate light-red coloring for Logistic Regression (16-17pp loss), indicating significant but non-catastrophic reliance on functional features. (2) Light coloring for XGBoost, SVM, Random Forest, and MLP (8-12pp loss), showing consistent but modest importance. (3) White cells for Naive Bayes (near-zero impact), confirming its fundamental inability to leverage functional information. (4) Degradation distributed across multiple metrics (precision, recall, F1, ROC-AUC), indicating functional features contribute to overall classification performance rather than being specialized for sensitivity or specificity. **Comparative context:** The moderate coloring here contrasts sharply with intense dark red from strand feature ablation (Figure 4) and trinucleotide ablation (Figure 12), confirming functional annotations provide supplementary information that refines predictions but cannot substitute for genomic organization and sequence composition in determining operon structure.

4 Conclusions

4.0.1 H1.1: Feature Importance Hierarchy

Result: REJECTED. Contrary to expectations based on prior literature emphasizing promoter/terminator motifs, strand concordance showed highest importance (20-60% ROC-AUC degradation across models), followed by intergenic distance for tree-based models (7.5%), trinucleotides (model-dependent, 1-26% with SVM/MLP most affected), and COG features (<2%). This completely reverses the expected hierarchy, revealing that operon prediction relies primarily on simple genomic organization constraints (genes must be on same strand for co-transcription) rather than complex sequence composition patterns.

4.0.2 H1.2: Feature Redundancy

Result: FAIL TO REJECT. PCA analysis showed 95.5% of trinucleotide variance explained by the first 10 principal components. Critically, ablating PC1-10 produced degradation patterns nearly identical to ablating all 64 trinucleotide features (within 1pp for all models), providing definitive evidence that trinucleotides exhibit extreme redundancy and conditional dependence. PC1 (83% variance) captures sequence complexity/diversity—distinguishing diverse, mixed trinucleotides characteristic of complex promoter regions from repetitive, homopolymeric runs in simple operon junctions. PC2 (7% variance, correlation with GC = -0.88) distinguishes AT-rich homopolymers (TAT, ATT, TTT) from GC-rich homopolymers (GGG, CGG, CCC), likely representing TATA-box promoters versus GC-rich regulatory elements. PC3-5 (3% variance) capture finer regulatory patterns like CpG content and pyrimidine runs. However, variance explained does not equal discriminative power—PC6-10 add 2.5% variance but contribute negligibly to performance, confirming most discriminative signal concentrates in the first five components representing sequence complexity, AT/GC balance, and specific regulatory motifs identified in prior work by Kanhere et al. (2006).

4.0.3 H1.3: Model Architecture Dependence

Result: FAIL TO REJECT WITH SUBSTANTIAL REFINEMENTS. Models showed varying dependencies, but patterns differed substantially from predictions and revealed fundamental differences in how model architectures learn feature representations:

Tree-Based Models (XGBoost, Random Forest): Learn threshold-based, rank-dependent patterns robust to distributional changes. XGBoost concentrated dependence on strand concordance (22-24pp loss) and intergenic distance (7-8pp loss), showing remarkable robustness to sequence composition ablation (1-2pp loss). Shuffle and random ablations produce similar degradation, indicating these models learn correlation-based rules (e.g., "if intergenic_distance < 120bp → operon") rather than exploiting specific distributional properties.

Kernel and Neural Methods (SVM, MLP): Learn distributional patterns sensitive to absolute feature values. Both exhibited catastrophic failures upon trinucleotide ablation (SVM: 23-31pp; MLP: 22-24pp), with random sampling consistently causing greater degradation than shuffle (differences of 8-11pp across PC1, PC1-5, PC1-10). This pattern reveals these models learned specific distributional signatures—not just "high PC1 → operon" but "PC1 values between X and Y → operon." SVM's RBF kernel relies on distances in feature space (sensitive to actual scales), while MLP's hidden layers learn non-linear boundaries based on absolute value ranges. Despite intergenic distance remaining intact and being moderately correlated with trinucleotides, these models still failed catastrophically, suggesting they learned interactions between distance and sequence patterns rather than relying on either independently.

Linear Methods (Logistic Regression): Show intermediate behavior with distributional sensitivity but less catastrophic failure. Random sampling causes moderately greater degradation than shuffle (10-11pp differences), indicating learned decision boundaries in feature space, but maintained more robust performance than kernel/neural methods.

Probabilistic Methods (Naive Bayes): Learn purely correlational patterns. Shuffle and random ablations produce identical degradation, indicating feature-target associations without distributional assumptions.

Architectural Implications: Random Forest showed an unexpected pattern with trinucleotide ablation—moderate ROC-AUC degradation (4-6pp) but darker red in precision/recall/F1 metrics, indicating trinucleotides help calibrate probability estimates at decision thresholds even though not critical for fundamental discrimination. This reveals a distinction between discriminative ability (ranking operons vs non-operons) and calibration quality (accurate probability estimates), with different model architectures showing varying sensitivity to each.

The consistent distributional learning pattern (Random >> Shuffle) across PC1, PC1-5, and PC1-10 for kernel and neural methods confirms these models exploit specific compositional signatures in low-dimensional trinucleotide subspace, with implications for model interpretability, robustness to domain shift, and deployment considerations.

4.0.4 Evidence for Non-Linear Feature Interactions

Preliminary Finding: Potential Interactions Require Further Investigation. The ablation results provide suggestive evidence for non-linear feature interactions, though definitive testing would require targeted interaction analysis.

Trinucleotides × Intergenic Distance (Strong Evidence): SVM and MLP exhibit catastrophic failure upon trinucleotide ablation (23-31pp and 22-24pp respectively) despite intergenic distance remaining intact. Since intergenic distance is moderately correlated with trinucleotides (per correlation heatmap), models relying on independent additive contributions should partially compensate using distance information. The failure to do so suggests these models learned interactions: "short distance + simple sequence composition → operon" versus "short distance + complex promoter signature → transcription unit boundary." This would align with biological reality—short intergenic distances are enriched in operons but also occur in non-operons, with sequence composition providing the disambiguating signal. SVM's RBF kernel naturally captures such interactions through distance calculations in multi-dimensional feature space, while MLP's hidden layers can learn non-linear feature combinations.

Strand Concordance × Trinucleotides (Moderate Evidence): Models show varying dependencies on both strand concordance (universal importance) and trinucleotides (model-dependent). Same-strand gene pairs can represent either operons or coincidentally co-directional genes at transcription unit boundaries, with trinucleotide signatures potentially distinguishing these cases through promoter complexity patterns. However, this interaction hypothesis requires explicit testing.

Future Work: Definitive confirmation of these interactions would require: (1) ablating feature pairs simultaneously and comparing degradation to sum of individual ablations, (2) explicit interaction term testing in interpretable models, and (3) analyzing learned representations in neural network hidden layers to identify feature combination patterns.

Additionally, cross-organism generalization remains untested. Training models on the complete E. coli dataset and evaluating on other prokaryotes (e.g., Cyano and Vibrio cholerae from DGEB [8]) after applying identical preprocessing and feature engineering would reveal whether learned feature dependencies and interaction patterns generalize across evolutionary distances, genome sizes, and regulatory architectures. Such analysis would determine if strand concordance dominance and trinucleotide interaction patterns represent universal prokaryotic operon signatures or E. coli-specific artifacts.

References

[1] Carlos P. Cantalapiedra, Ana 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] Michael Y. Galperin, Yuri I. Wolf, Kira S. Makarova, Roberto Vera Alvarez, David Landsman, and Eugene V. Koonin. Cog database update: focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Research*, 49(D1):D274–D281, 2021.

[3] Ashwin Kanhere, Sarath Chandra Janga, and Gabriel Moreno-Hagelsieb. The distinctive signatures of promoter regions and operon junctions across prokaryotes. *Nucleic Acids Research*, 34(14):3980–3987, 2006.

[4] Rezvan Karaji and Lourdes Peña-Castillo. Opdetect: A convolutional and recurrent neural network classifier for precise and sensitive operon detection from rna-seq data. *PLoS One*, 20(8):e0329355, 2025.

[5] Raga Krishnakumar and Anne M. Ruffing. Operonseqer: A set of machine-learning algorithms with threshold voting for detection of operon pairs using short-read rna-sequencing data. *PLoS Comput Biol*, 18(1):e1009731, 2022.

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

[7] Eric W. Sayers, John Beck, Edward E. Bolton, J. Rodney Brister, Justin Chan, Roy Connor, Michael Feldgarden, Andrea M. Fine, Kendal Funk, Jennifer Hoffman, Sekar Kannan, Colleen Kelly, William Klimke, Sunghoon Kim, Scott Lathrop, Aron Marchler-Bauer, Tasha D. Murphy, Cameron O'Sullivan, Robert Schmieder, Yuliya Skripchenko, Adam Stine, Francoise Thibaud-Nissen, Jian Wang, Jian Ye, Elizabeth Zellers, Valerie A. Schneider, and Kim D. Pruitt. Database resources of the national center for biotechnology information in 2025. *Nucleic Acids Res*, 53(D1):D20–D29, 2025.

[8] Thomas West-Roberts, Anshul Kundaje, and James Zou. Dgeb: A diverse genomic embedding benchmark for functional genomics tasks. *bioRxiv*, 2024.