BMC Bioinformatics

RESEARCH Open Access

GNNs and ensemble models enhance the prediction of new sRNA-mRNA interactions in unseen conditions

Shani Cohen[1], Lior Rokach[1] and Isana Veksler-Lublinsky[1*]

*Correspondence:
vaksler@post.bgu.ac.il

[1] Department of Software & Information Systems Engineering, Faculty of Engineering, Ben-Gurion University of the Negev, 8410501 Beer-Sheva, Israel

Abstract

Bacterial small RNAs (sRNAs) are pivotal in post-transcriptional regulation, affecting functions like virulence, metabolism, and gene expression by binding specific mRNA targets. Identifying these targets is crucial to understanding sRNA regulation across species. Despite advancements in high-throughput (HT) experimental methods, they remain technically challenging and are limited to detecting sRNA-target interactions under specific environmental conditions. Therefore, computational approaches, especially machine learning (ML), are essential for identifying strong candidates for biological validation. In this paper, we hypothesize that ML models trained on large-scale interaction data from specific conditions can accurately predict new interactions in unseen conditions within the same bacterial strain. To test this, we developed models from two families: (1) graph neural networks (GNNs), including *GraphRNA* and *kGraphRNA*, that learn transformed representations of interacting sRNA-mRNA pairs via graph relationships, and (2) decision forests, *sInterRF* (Random Forest) and *sInterXGB* (XGBoost), which use various interaction features for prediction. We also proposed Summation Ensemble Models (SEM) that combine scores from multiple models. Across three seen-to-unseen conditions evaluations, our models —particularly *kGraphRNA*— significantly improved the area under the ROC curve (AUC) and Precision-Recall curve (PR-AUC) compared to *sRNARFTarget*, *CopraRNA*, and *RNAup*. The SEM model combining *GraphRNA* and *CopraRNA* outperformed *CopraRNA* alone on a low-throughput (LT) interactions test set (HT-to-LT evaluation). Beyond enhanced performance, our models enable target prediction for species-specific sRNAs, a capability lacking in some existing tools. Furthermore, GNN models remove the dependency on external tools like *RNAplex* or *RNAup* to compute hybridization duplex or energy features, enhancing scalability and runtime efficiency. While this study focuses on *E. coli K12 MG1655* interactions, our methods are fully adaptable to predict interactions in other bacterial strains, given sufficient data for training. Our comprehensive feature importance analysis revealed the complexity of sRNA-mRNA interactions across environmental conditions, underscoring the significance of RNA sequence composition and duplex structure characteristics, like base pairing and energy factors; findings that align with biological evidence from previous studies. As HT experiments expand sRNA-target interaction data across conditions in various bacteria, our ML methods

with features analysis offer promising advances in sRNA-target prediction and deeper insights into sRNA regulatory mechanisms across diverse species.

Keywords: sRNA-target prediction, Machine learning, Graph neural networks, Bacterial gene regulation, kGraphRNA, GraphRNA, sInterRF, sInterXGB

Introduction

Bacterial small RNAs (sRNAs), which are relatively short non-coding RNA molecules (~50–500 nt), act as post-transcriptional regulators either by binding specific proteins to alter their activity or by base-paring with target mRNAs [1], impacting various bacterial functions, such as virulence, environmental sensing, metabolism, and gene expression [2]. *Cis*-encoded and *trans*-encoded are the two major classes of base-pairing sRNA. C*is*-encoded sRNAs, a.k.a. antisense RNA, are transcribed from the complementary strand of the target mRNA, whereas *trans*-encoded sRNAs share only a partial sequence complementarity with their targets and thus regulate multiple genes. Similar to eukaryotic microRNAs, *trans*-encoded sRNAs modulate the translation, processing, and/or stability of their target mRNAs through short interactions [3].

In most cases, the sRNA represses the mRNA translation as it binds it at or in the surroundings of the ribosome binding site (RBS) [1]. In other cases, the sRNA binds outside the RBS, inducing changes in its secondary structure and accessibility [4] or altering the stability of the mRNA [5], thereby exerting either negative or positive gene control. In many bacterial species, the base-paring between sRNAs and their targets is mediated by the RNA-binding chaperone Hfq [6], though other proteins such as ProQ [7], CsrA [8], and RNase E [9] were also found to be involved in sRNAs regulation processes. The pairing region between sRNAs and target mRNAs typically includes a conserved minimal seed region of 6–8 successive base pairs, with a preference for base pairing with the 5' end of the mRNA rather than the 3' end [6].

Although numerous sRNAs have already been discovered, the characterization of their regulatory mechanism across bacterial species remains limited and heavily relies on the identification of their mRNA targets [10]. Early experimental methods, such as reporter assays, mutagenesis, knockouts, sRNA deletions, and footprinting, provided strong evidence for sRNA-target interactions. Interactions identified through these low-throughput techniques were manually curated in, e.g., sRNATarBase 3.0 [11]. In recent years, with the advent of high-throughput (HT) sequencing technologies, several experimental methods have been developed for identifying sRNA-target interactions that are facilitated by RNA binding protein (RBP) [12] such as RIL-seq [13] and CLASH [14, 15] (detailed in the supplementary materials 1.1).

Despite the great advancement in HT experimental methods, applying such protocols is technically challenging in practice [16]. Consequently, experimental data of sRNA-target interactions is currently available only for a few model organisms. Another limitation of the experimental protocols stems from the low efficiency (1%-5%) of the cross-linking [13], in addition to the low efficiency of the following RNA-RNA ligation step, resulting in a set of chimeric reads representing direct interactions that usually constitute less than 2% of the total sequencing reads [17]. Most importantly, these experimental methods are typically conducted under specific conditions, such as growth stages, stress responses, or medium composition. However, sRNA-mRNA interactions are known to

vary significantly due to these or other environmental factors, primarily because of the dynamic alterations in the expression levels of both sRNAs and mRNAs under different conditions [13]. This variability makes it challenging to capture the full spectrum of interactions with a single experimental setup. Therefore, the prediction of new interactions remains highly valuable, as it can provide insights into interactions that are otherwise undetectable by experiments limited to specific contexts. We hypothesize that by learning from large-scale data of interactions captured in studied conditions, we can make reliable predictions about interactions occurring in other, yet unexplored, environmental conditions.

Computational methods for sRNA-target prediction play a crucial role in complementing experimental approaches by providing *bona fide* candidates for biological validation. Unlike experimental methods, computational tools can integrate knowledge from interactions captured under varied conditions, allowing predictions that extend beyond specific growth conditions and cover a broader range of bacterial species. Although several computational tools have already been developed, this task remains challenging due to the significant heterogeneity in the size and structure of sRNAs [18], the presence of partial and non-contiguous base-pairing regions with frequent internal secondary structures, and the relatively short sRNA-mRNA hybrids. This complexity makes it difficult to differentiate *bona fide* sRNA-mRNA interactions from random hybrids formed between any transcripts within a cell [19, 20].

Current computational methods for sRNA-mRNA interaction prediction can be broadly classified into two categories: non-learning methods and machine-learning (ML) methods. Non-learning methods, including those based on local alignment, minimum free energy (MFE) calculations, and evolutionary conservation, utilize established rules to predict sRNA-target interactions without relying on ML algorithms. Tools based on local alignment or MFE use dynamic programming or specific MFE approaches for interaction prediction, e.g., *RNAduplex* [21], *RNAup* [22], *RNAplex* [23], *IntaRNA* [24], *TargetRNA* [25], *IntaRNAhelix* [26] and more. These tools may also be used as the underlying mechanism of other, more complex, prediction tools, such as *CopraRNA* [18] utilizing *IntaRNA*, *RNApredator* [27] utilizing *RNAplex*, and *SPOT* [10] incorporating four distinct tools for interaction prediction. Still, tools like *RNAup* and *IntaRNA* may not always return an output due to energy constraints, limiting its predictive power in certain cases. Comparative tools such as *TargetRNA2* [28] and *CopraRNA* [18], designed specifically for bacterial sRNA-target interactions, exploit homology and evolutionary conservation for prediction. Since *CopraRNA* is restricted to conserved sRNAs, it is unsuitable for species-specific sRNAs such as those originating from horizontally acquired virulence regions [12].

ML methods excel at capturing complex relationships in data, making them particularly effective for a wide range of biological prediction tasks [29, 30]. Their ability to model non-linear patterns and leverage diverse features allows for more accurate predictions compared to non-learning approaches. In the context of sRNA-target prediction, ML-based tools include *sRNATarget* [31, 32], *sTarPicker* [33], and the more recent *sRNARFTarget* [34] and *TargetRNA3* [35]. These tools differ in their ML algorithms, the features used to characterize interactions, the negative data generation methods, and the datasets selected for training and testing (details for

the latter two tools are provided in the supplementary materials 1.2). Some of these tools utilize features derived from non-learning tools, such as minimum free energy (MFE) and/or evolutionary conservation measures.

Notably, the definition of the specific train and test sets is essential, not only to avoid data leakage but also to allow researchers to test specific hypotheses and properly evaluate these tools' performance. For example, *sRNARFTarget* was trained on all available bacterial interactions regardless of experimental conditions and evaluated on its ability to predict novel interactions both in *E. coli* (which had interactions included in the training set) and in bacterial species excluded from training. *TargetRNA3* was evaluated on its ability to predict interactions in new genomes not present in the training data without accounting for environmental conditions.

In this paper, we present the first evaluation of the ability of ML models to learn from interactions observed under specific environmental conditions and accurately predict new interactions in different unseen conditions (seen-to-unseen conditions evaluations). To this end, we compiled a dataset of high-throughput (HT) sRNA-mRNA interactions in *E.coli* under various conditions. We divided the dataset into three train-test splits, ensuring that the test set of each unseen condition (e.g., growth stage, medium composition, or stress response) comprised new interactions not presented in the train set. In addition, we assessed the ability to learn from all HT interactions to predict positive and negative interactions from low-throughput (LT) experiments, which are generally considered to have stronger experimental support (HT-to-LT evaluation).

We developed multiple machine-learning models from two distinct model families to accurately predict sRNA-mRNA interactions. The first model family comprises graph neural networks (GNNs) that learn transformed representations of the interacting sRNA-mRNA pairs based on graph relationships and subsequently use these representations to predict new interactions. The two GNN-based models are *GraphRNA and kGraphRNA*, where *kGraphRNA* employs sRNA and mRNA composition features to initialize the graph nodes. The second model family consists of two types of decision forests, Random Forest (*sInterRF*) and eXtreme Gradient Boosting (*sInterXGB*), that learn from a rich set of interaction features computed over the interacting sRNA-mRNA pairs, including local-interaction-based features and 3-mer frequency differences. In addition, we suggested several Summation Ensemble Models (SEM) combining the prediction scores of multiple individual models.

We have assessed the performance of our models in seen-to-unseen conditions and HT-to-LT evaluations, comparing to the existing tools *RNAup*, *sRNARFTarget*, and *CopraRNA*. We showed that GNN-based models significantly improved the predictive performance of the existing tools in terms of both the area under the ROC curve (AUC) and the area under the Precision-Recall curve (PR-AUC), with *kGraphRNA* outperforming in the seen-to-unseen evaluations and the SEM model of *GraphRNA* and *CopraRNA* significantly improving upon *CopraRNA* alone, the best competing model, in the HT-to-LT evaluation. We also provided a comprehensive analysis of feature contributions to model predictions, shedding light on the complex nature of sRNA-mRNA interactions in different environmental conditions.

Methods

Data

Data of *Escherichia coli K12 MG1655* (NC_000913) sRNA-mRNA interactions were retrieved from the *sInterBase* comprehensive database [36], including interacting and non-interacting sRNA-mRNA pairs and their sequences, as detailed in Table 1**.** Since an interacting sRNA-mRNA pair may appear multiple times in a high-throughput (HT) dataset, each time with different interacting fragments, we provide information about the unique number of interactions, i.e., unique sRNA-mRNA pairs, found within each dataset. Non-interacting sRNA-mRNA pairs are pairs for which no sRNA-dependent regulation of the mRNA was observed when tested under the specific experimental settings described in the original source. For simplicity, we refer to interacting pairs as positive interactions (P), and non-interacting pairs as negative interactions (N).

Sources 1 and 2 in Table 1 provide positive and negative interactions derived from low-throughput (LT) experiments, which are generally considered to have stronger experimental support [11, 18], and are collectively referred to as the LT dataset. In contrast, sources 3 to 5 provide positive interactions from three HT experiments [7, 13, 15], collected in various environmental conditions. Thus, the HT dataset included interactions from sources 3 to 5 and excluded all LT interactions (Table 1), ensuring no data was leaked between the HT and LT datasets.

The HT dataset was divided into three train-test configurations for the seen-to-unseen condition evaluations (detailed in Recovering new interactions in unseen conditions). Additionally, HT and LT datasets served for training and testing, respectively, in the HT-to-LT evaluation. Negative samples were generated using random swapping of the sRNAs and mRNA in the positive pairs, similar to [34], while also ensuring that no negative pair was already labeled as positive in the dataset. The final datasets in the seen-to-unseen conditions and HT-to-LT evaluation are described in

Table 1 Summary of *E.coli* sRNA-mRNA interaction data retrieved from *sInterBase*

Serial no	Source	High-throughput method	No. of unique sRNA-mRNA interactions	No. of unique interacting sRNA	No. of unique interacting mRNA
1	Wright et al., 2013	-	88 P: 83, N: 5	18	68
2	sRNATarBase3.0	-	383 P: 224, N: 159	43	249
	Total—LT dataset		**391** **P: 227, N: 164**	**43**	**253**
3	Melamed et al., 2016	RIL-seq	1203 P: 1203, N: 0	23	888
4	Melamed et al., 2020	RIL-seq	2441 P: 2441, N: 0	44	1519
5	Iosub et al., 2020	CLASH	1437 P: 1437, N: 0	40	929
	Total—HT dataset		**3856** **P: 3856, N: 0**	**51**	**1988**

P the number of positive samples, ***N*** the number of negative samples. The low-throughput (LT) dataset comprised interactions from sources 1 and 2, while the high-throughput (HT) dataset comprised interactions from sources 3, 4, and 5, excluding interactions from the LT dataset. The total numbers are presented separately for the HT and LT datasets

Recovering new interactions in unseen conditions and Recovering interactions from low-throughput experiments (HT-to-LT), respectively.

Feature extraction

We extracted 3 types of features to be used by different model families. Two subsets of features, namely the local-interaction-based and 3-mer-diff, represent sRNA-mRNA interactions, while features of the third subset, named 3-mer, represent the interacting sRNA and mRNA molecules by their sequences.

Local-interaction-based features

The motivation for the subset of features described below stems from previous work [37] and is based on the assumption that physical interactions between different sRNAs and mRNAs in a given bacteria adhere to complex yet shared rules. Therefore, these rules can be learned by a machine-learning model. In this framework, the sequences of the sRNA and mRNA molecules are used to predict a local interaction duplex by utilizing the *RNAup* software [22] (see supplementary Figure S1). *RNAup* identifies the optimal local interaction site based on MFE and accessibility calculations. This tool has demonstrated superior performance in terms of True Positive Rate (TPR) and Positive Predictive Value (PPV) when compared to several other RNA-RNA interaction prediction tools [38]. To generate the *RNAup* interaction duplex, the sRNA input sequence was defined as the entire sRNA molecule, and the mRNA input sequence was defined as 200 nt upstream and + 100 nt downstream with respect to the start codon, as previously done by Wright et al. [18].

From the *RNAup* output, we extracted 23 local-interaction-based features, which can be grouped into three categories (see supplementary Table S1 for a detailed description per feature): (a) Energy: energy features computed by *RNAup*, including hybridization, unfolding the sRNA, unfolding the mRNA, and total. (b) Duplex structure: features describing the base-pairing patterns of the *RNAup* duplex formed between the sRNA and the mRNA, including alignment length (nt), number of all base pairs, mismatches, bulges in the sRNA, bulges in the mRNA; number of base pairs, max consecutive base pairs, mismatches, bulges in the sRNA, and bulges in the mRNA out of the alignment length (prop.); number of specific base pairs (GC, AU, GU) out of all base pairs (prop.). (c) Target context: we extended the mRNA interacting area by 20 nt per side, and extracted context features from the flanking region (i.e., the extensions), including single nt counts (A, G, C, U) and paired nt counts (A + U, G + C) out of all the nts in the flanking region (prop.).

It is important to note that *RNAup* prediction may fail if the predicted interaction structure violates internal consistency checks within the software. In that case, no duplex or features can be computed for the particular interaction. As a result, interaction inputs that returned an error were discarded. We have also excluded interactions for which the predicted duplex comprised fewer than 5 base pairs to omit unlikely interactions [38, 39]. The final datasets per evaluation are described in the Results (Sects. "Recovering new interactions in unseen conditions" and "Recovering interactions from low-throughput experiments (HT-to-LT)").

k-mer frequency features of sRNA and mRNA (3-mer)

Inspired by *sRNARFTarget* [34] representations of the sRNA and mRNA molecules, we computed k-mer frequency features (with k = 3) for both sRNA and mRNA sequences separately, resulting in 64 features representing each RNA molecule, and 128 features in total.

k-mer frequency difference features (3-mer-diff)

To represent an sRNA-mRNA interaction, we calculated the 3-mer frequency difference between the RNA sequences, based on the hypothesis that k-mer frequency differences can capture the base pairing potential between mRNAs and sRNAs, similarly to *sRNARFTarget* [34]. Using the frequencies from k-mer frequency features of sRNA and mRNA (3-mer), we computed the frequency difference for each k-mer i as $fi, mRNA - fi, sRNA$, where fi, S is the frequency of 3-mer i in sequence S. This resulted in 64 features overall.

Feature selection using mRMR

The Minimum Redundancy Maximum Relevance (mRMR) feature selection method aims to identify features that are highly relevant to the target variable and minimally redundant among each other. mRMR balances two criteria: (1) maximizing relevance by selecting features having high Mutual Information (MI) with the target variable, and (2) minimizing redundancy by selecting features having low MI with one another. Due to its computational efficiency and effectiveness, this approach was widely applied to high-dimensional datasets in bioinformatics [40–43] and other domains [44, 45], where selecting a smaller subset of informative features was critical for model performance.

Given the relatively large number of 87 interaction features compared to the size of our train sets, we applied mRMR on the subset of 64 3-mer-diff features. This approach helps prevent the over-representation of a specific feature type and ensures that our models can leverage a more heterogeneous set of features, representing different perspectives of the interaction, i.e., the local duplex and sequence differences. We deployed an automatic mRMR algorithm based on the mean of the smoothed Mutual Information (MI) changes over the last five selected features. It stops when the mean of MI changes converges near zero ($< 1e^{-3}$), meaning that further features do not provide significant new information to target prediction. We adjusted the Python mRMR implementation provided by the MIFS[1] module and integrated it into our feature selection procedure.

Machine-learning models

Graph neural networks (GNNs)

Our approach of utilizing graph neural networks to predict interactions was inspired by the idea of "collaborative filtering" (or interactions' similarity) with regard to sRNA-mRNA interactions [46]. It operates on the assumption that if sRNAs s1 and s2 both interact (or not) with mRNA target m1, then it is more likely that s1 will exhibit similar behavior to s2 (i.e., interact or not interact) with other mRNA targets, compared to a

[1] https://github.com/danielhomola/mifs

randomly selected sRNA. This concept also served as a basic assumption in previous research on interaction prediction conducted in other settings (e.g., miRNA-mRNA interactions), employing a similarity-based algorithm [47].

A graph is a data structure that models a set of relations (i.e., edges or links) between a set of entities (i.e., nodes or vertices); formally, $G = (V, E)$, where V is the set of nodes and E is the set of edges. A bipartite graph is a graph composed of two disjoint sets of nodes, V_1 and V_2, such that every edge in the graph connects a node in V_1 with a node in V_2, formally $B = (V_1, V_2; E)$ [48]. Graph neural networks (GNNs) are powerful deep-learning-based methods for learning rich context-informed representations of nodes in graphs by propagating and aggregating semantic and structural information from a node and its local neighborhood (a.k.a., the message-passing mechanism). The transformed representations (a.k.a., embeddings) can serve various downstream tasks, such as node classification or link prediction, e.g., the prediction of new interactions between nodes [49]. These methods have demonstrated great success in many biology and healthcare-related tasks [50–53]. In spatial-based convolutional GNN, the model consists of multiple graph convolutional layers. Each layer applies the convolution operator to aggregate and transform information from the node and its neighbors [54]. In some cases, depending on the architecture selected for the task, the convolutional layer is followed by a non-linear transformation (a.k.a., activation function) that is applied to the output vectors [54].

In this study, we constructed a bipartite graph consisting of two types of nodes: 'sRNA' and 'mRNA', and two types of edges: ('sRNA', 'regulates', 'mRNA') and ('mRNA', 'is regulated by', 'sRNA'), so that the presence of one edge forces the presence of the other edge between a given pair of nodes. We developed two GNN models for predicting new sRNA-mRNA interactions, *GraphRNA* and *kGraphRNA*, that differ in the initial embeddings of graph nodes and the size of their hidden layers. As illustrated in Fig. 1, each GNN model comprised an *encoder* part that learns transformed embeddings of the sRNA and mRNA nodes from graph-structured data and a *decoder* part that uses the transformed embeddings to make predictions. Due to the heterogeneity in the attributes of nodes from different types (i.e., sRNA vs. mRNA), our models learn individual message-passing functions for each node type.

In *GraphRNA*, we initialized embedding vectors of size 64 for both sRNA and mRNA nodes using random uniform sampling. The *encoder* comprised two convolutional layers, each with 64 channels and the 'mean' aggregation function, and each layer followed by the Rectified Linear Unit (ReLU) activation function. The *decoder* comprised a dot product function applied to the transformed embeddings of the sRNA and mRNA, followed by a sigmoid activation function that yields the interaction probability. In *kGraphRNA*, we used the 64 features of 3-mer frequency (3-mer) as the initial representations of both the sRNA and mRNA nodes. The *encoder* comprised two convolutional layers, each with 32 channels and the 'mean' aggregation function, and each layer followed by the Rectified Linear Unit (ReLU) activation function. The *decoder* was the same as in *GraphRNA*. Both models applied the bipartite convolutional layer of SAGEConv, a.k.a., the GraphSAGE algorithm [55], to pass messages from source nodes to target nodes in the *encoder* part. GraphSAGE is an inductive approach for neighbor sampling that results in pruned computational graphs for generating node embedding. It significantly improves the

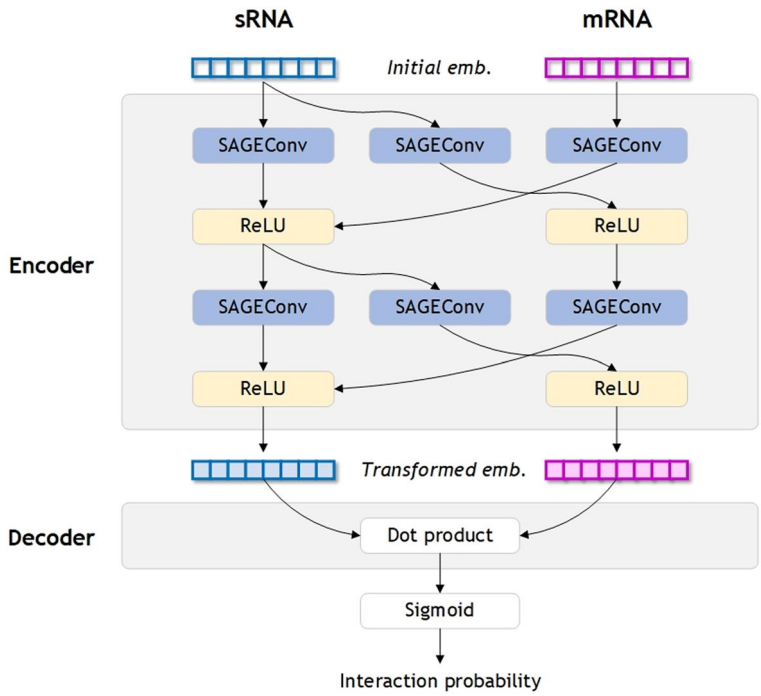


Fig. 1 An illustration of the architecture of ***GraphRNA*** and ***kGraphRNA*** models for predicting sRNA-mRNA interactions. In *GraphRNA* we initialized embedding vectors of size 64 for both the sRNA and the mRNA nodes using random uniform sampling, and each hidden layer had 64 channels. In *kGraphRNA*, we used the 64 features of 3-mer frequency (3-mer) as the initial embeddings for both the sRNA and the mRNA nodes, and each hidden layer had 32 channels. Both models were trained with binary cross-entropy loss

computational efficiency and robustness of the GNN model and can be used for generating transformed embeddings of previously unseen nodes. The *GraphRNA* and *kGraphRNA* models were implemented in Python using the *PyTorch Geometric* library [56] and were trained with binary cross-entropy loss and a learning rate of 0.001. *GraphRNA* was trained with 25 epochs, and *kGraphRNA* was trained with 120 epochs. The loss curves along the training epochs in the HT-to-LT evaluation are presented in supplementary Figure S2.

Decision forests

We examined two decision forest models for binary classification, Random Forest (*sInterRF*) and XGBoost (*sInterXGB*), both learned from a rich set of interaction features computed over the interacting sRNA-mRNA pairs, including local-interaction-based features and 3-mer frequency differences. XGBoost is a regularized variant of stochastic gradient boosting, a.k.a. Gradient Boosting Machines (GBM). In GBM, new 'weak' models are added sequentially to the ensemble to correct the errors of prior models. The XGBoost ensemble is comprised of Classification and Regression Trees (CART) that are combined to make the final prediction. This algorithm has demonstrated superior performance in many machine-learning competitions and research studies [57, 58], including bioinformatics research [59–61]. Random Forest is an ensemble of decision trees, such that each tree is constructed based on a subset of m samples selected at random from the dataset, with replacement (bootstrap sample). A subset of k features is

randomly selected and employed as the candidate splitters of each tree node, and each tree is grown to the largest extent possible without pruning. As an instance is being classified, each tree "votes" for one of the classes, and the final decision is made based on the majority of "votes". Utilizing multiple trees and random selections in RF prevents overfitting and reduces bias [62].

The XGBoost, if well-calibrated, can achieve superb results but has a higher risk of overfitting, whereas the Random Forest is more robust with less tendency to overfit. Both classifiers demonstrated great success when applied to tabular data, especially for small-medium datasets (< 100 K samples) [58]. The *sInterXGB* model was initialized with 500 boosting rounds and trained with an early stopping after 3 rounds based on log loss using a random validation set of 10% of the train data. The *sInterRF* model was initialized with 500 tree estimators, a maximum tree depth of 9, and a maximum number of features to be considered at each split set to log2.

Summation ensemble models (SEM)

We also examined three Summation Ensemble Models (SEM) that combine the probabilities given by several individual models: (1) *GraphRNA* and *CopraRNA* (2) *GraphRNA*, *CopraRNA*, and *sInterRF* (3) *kGraphRNA* and *sInterRF*. These combinations were selected to balance complementary strengths and avoid redundancy across models. Specifically, *CopraRNA* was included as it predicts sRNA-mRNA interactions based on evolutionary conservation across bacterial species, a perspective not addressed by our feature-based models. *sInterRF* was selected over *sInterXGB* due to its superior performance and shared feature space with *sInterXGB*, making the latter redundant in ensemble settings. Similarly, other baseline models such as *sRNARFTarget* and *RNAup* were excluded, as their core functionalities (e.g., sequence-based k-mer features and RNA duplex energetics) were already incorporated into our models.

In each SEM, a uniform probability summation was applied across the individual model scores per interaction to generate the final prediction score. *CopraRNA* p-values were converted to 'probability-like' scores before integration (see supplementary material 3.1.1 for details).

Features contributions

We used SHAP (SHapley Additive exPlanations) [63] to measure the relative contribution of different interaction features to model predictions. SHAP values are theoretically optimal feature attribution values that measure the contribution of each feature to the prediction of a specific sample, hence providing local explanations per sample. Aggregating these values over all dataset samples yields a global feature importance measure (I), known as the mean SHAP value magnitude [64]. For simplicity in comparing feature contributions to the model's predictions across different datasets, we derived a normalized contribution measure (NC) for each feature based on the global feature importance measure (I). Formally, considering S as the set of all model features, the normalized contribution measure (NC) of feature x is defined as: $NC_x = \frac{I_x}{\sum_{x \in S} I_x}$. Similarly to the global importance measure (I), the normalized contribution measure (NC) helped us identify the most important features of the model when predicting interactions.

Performance metrics

We evaluated the models' performance using both *ranking-based* and *threshold-based* metrics. *Ranking-based* metrics assess the performance across the entire (or a limited) range of thresholds, providing a more comprehensive view of how well the model distinguishes between interacting (label = 1) and non-interacting (label = 0) sRNA-mRNA pairs. We report three *ranking-based* metrics: the AUC, a.k.a. the Area Under the Receiver Operating Characteristic (ROC) Curve, the pAUC, a.k.a the partial Area Under the ROC Curve, and the PR-AUC, a.k.a. the Area Under the Precision-Recall Curve. The AUC metric is the probability that a binary classifier will rank a random positive sample higher than a random negative sample. It ranges between 0 and 1 and its baseline value equals 0.5. To test whether two ROC curves with similar AUC are significantly different, we calculated the p-value of *Delong's test* [65] using the pROC implementation [66]. We also measured the partial Area Under the ROC Curve (pAUC) for a False Positive Rate ≤ 0.15, a desirable threshold often applied in the practical utility of sRNA-target prediction tools. The PR-AUC metric is the average precision computed across all recall values, or the probability that if a positive sample is selected from the dataset, then a sample with a higher rank will also be positive. It ranges between 0 and 1, and its baseline value equals the proportion of positive samples to the total number of samples in the evaluated dataset.

Threshold-based metrics rely on setting a specific decision threshold to classify predictions as positive or negative. We defined a threshold using the Youden Index criterion [67], which maximizes the J statistic defined as $J = sensitivity + specificity - 1$. This criterion applies to prediction scores of different ranges (e.g., *RNAup* scores that extend beyond the 0 to 1 range) and has been widely used in medical research [68, 69]. We report six threshold-based metrics: *Accuracy, Precision, Recall* (Sensitivity), *Specificity, F1, and Matthew's Correlation Coefficient (MCC).*

Results

Recovering new interactions in unseen conditions

It is known that sRNA-mRNA interactions respond to environmental changes, primarily due to the dynamic alterations in the expression levels of both sRNAs and mRNAs under different conditions. Therefore, different high-throughput (HT) experiments, even when applied to the same bacterial strain, can capture distinct interactions depending on the specific growth conditions. Our first hypothesis was that a machine-learning model trained on interactions collected under one set of growth conditions can accurately predict *new* interactions observed under a new, unseen condition.

We examined this hypothesis through three experimental evaluations, each focusing on a different growth condition applied to the *E.coli K12* bacterial strain: growth stage, medium composition, and stress response (see Table 2). In the first evaluation, models were trained on interactions collected during the log (exponential) growth stage to predict new interactions specific to the stationary stage. In the second evaluation, models trained on interactions observed in the bacteria grown on the rich Luria–Bertani (LB) medium were used to predict new interactions in bacteria grown on the minimal m63 medium. In the third evaluation, models trained on log-growth stage interactions in

Table 2 Summary of train-test splits of positive sRNA-mRNA pairs at each evaluation of seen-to-unseen conditions

No	Seen-to-unseen Conditions	Dataset	High-throughput method	Growth stage	Medium	Stress response	No. of positive sRNA-mRNA unique pairs
1	Log-to-stationary	Train	RIL-seq, CLASH	**Log**	LB	N	3097 [2416 + 681]
		Test	RIL-seq, CLASH	**Stationary**	LB	N	478
2	LB-to-m63	Train	RIL-seq, CLASH	Log, Stationary	**LB**	N	3575 [3179 + 396]
		Test	RIL-seq	Log	**m63**	N	123
3	Normal-to-stress	Train	RIL-seq, CLASH	Log	LB	**N**	3097 [2488 + 609]
		Test	RIL-seq	Log	LB	**Y**	182

Each train set included all pairs of interacting sRNA-mRNA captured under the observed condition, comprising interactions recovered only under the seen condition plus those recovered under both the seen and the unseen conditions, as specified in the left and right elements in the parentheses (last column). Each test set consists of new pairs of interacting sRNA-mRNA captured only under the respective unseen condition. Of note, CLASH interactions from both the log and transition growth stages originating from [15] are referred to as Log. These positive interactions were supplemented with negative interactions. The final number of interactions post-filtering is indicated in the relevant figures for each evaluation below

normal conditions were tested on their ability to predict log-growth stage interactions under iron limitation stress response.

In each seen-to-unseen evaluation, we supplemented the train and test sets with negative samples, resulting in an equal number of positive and negative samples within each set. Post-filtering of interactions due to *RNAup* error or < 5 base pairs, we had the following final datasets: log-to-stationary, train: I = 5625 (P: 2808, N: 2817) and test: I = 883 (P: 443, N: 440); LB-to-m63, train: I = 6539 (P: 3251, N: 3288) and test: I = 218 (P: 114, N: 104); normal-to-stress, train: I = 5670 (P: 2808, N: 2862) and test: I = 326 (P: 163, N: 163).

For each evaluation, we report the performance metrics of all models in a comprehensive table. In addition, we present ROC and Precision-Recall curves for our top two models (in terms of AUC and PR-AUC) in comparison with existing tools: *RNAup*, *sRNARFTarget*, and *CopraRNA*. A detailed description of how we ran each of the existing tools is provided in the supplementary materials (3.1).

Log-to-stationary growth stage evaluation

kGraphRNA outperformed all other models across all metrics, except for specificity, where *RNAup* achieved the highest result (Fig. 2). Moreover, *kGraphRNA* demonstrated a significant improvement over *sRNARFTarget*, the top-performing competitive tool (as indicated by Delong's test comparing ROC curves, $p = 3.24e^{-7}$). Similarly, the *SEM_kGraphRNA_sInterRF* model also showed significant improvement over *sRNARFTarget* ($p = 9.73e^{-6}$). All models performed better than the random baseline (AUC = 0.5 and PR-AUC = 0.5).

LB-to-m63 medium evaluation

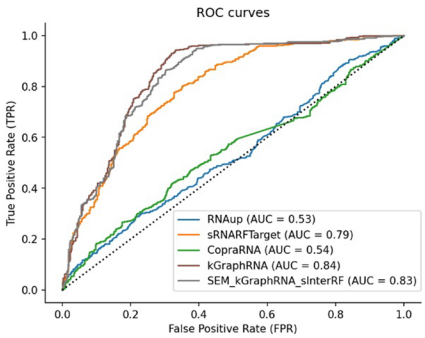
The *SEM_kGraphRNA_sInterRF* model achieved the best performance in most metrics when predicting new interactions in the m63 medium, including AUC, pAUC, PR-AUC, accuracy, F1, and MCC (Fig. 3), significantly outperforming *RNAup* ($p = 2.39e^{-16}$) and *CopraRNA* ($p = 5.52e^{-8}$). Similarly, the *kGraphRNA* model achieved the best AUC score (= 0.91), surpassing *sRNARFTarget*, *RNAup*, and *CopraRNA*, in all performance metrics.

Log-to-stationary

A.

Model	AUC	pAUC	PR-AUC	Accuracy	Precision	Recall	Specificity	F1	MCC
RNAup	0.53	0.53	0.55	0.53	0.57	0.28	**0.79**	0.38	0.08
sRNARFTarget	0.79	0.62	0.76	0.72	0.70	0.77	0.67	0.74	0.45
CopraRNA	0.54	0.53	0.55	0.55	0.57	0.42	0.68	0.48	0.10
sInterXGB	0.72	0.58	0.70	0.70	0.65	0.87	0.53	0.75	0.43
sInterRF	0.80	0.62	0.77	0.75	0.69	**0.91**	0.59	0.79	0.53
GraphRNA	0.71	0.60	0.71	0.66	0.67	0.65	0.67	0.66	0.32
kGraphRNA	**0.84**	**0.63**	**0.80**	**0.81**	**0.75**	**0.93**	0.68	**0.83**	**0.63**
SEM_GraphRNA_CopraRNA	0.61	0.57	0.64	0.59	0.64	0.42	**0.75**	0.51	0.19
SEM_GraphRNA_CopraRNA_sInterRF	0.73	0.61	0.72	0.68	0.70	0.63	0.73	0.66	0.35
SEM_kGraphRNA_sInterRF	**0.83**	**0.63**	**0.79**	**0.78**	**0.74**	0.86	0.70	**0.80**	**0.57**

B.



C.

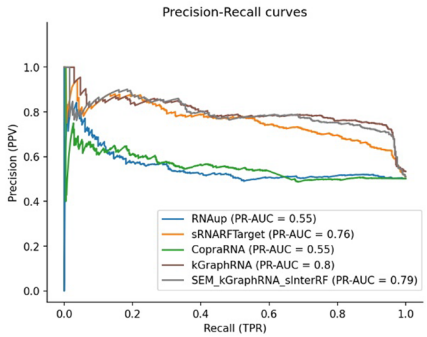


Fig. 2 **A** Performance metrics of all models trained on 5625 log growth stage interactions (P: 2808, N: 2817) and tested on 883 stationary growth stage interactions (P:443, N: 440). **B.** ROC curves and **C.** Precision-Recall curves of our top performing models (in terms of AUC and PR-AUC) compared to existing models: *RNAup*, *sRNARFTarget*, and *CopraRNA*. In panel **A**, the two top scores per metric are marked with green and gray, respectively. The pAUC is computed for $FPR \leq 0.15$. In panel **B**, the dotted diagonal line represents the baseline of a random classifier

GraphRNA achieved the highest precision and specificity, while *sInterXGB* led in recall. All models exceeded the random baselines (AUC = 0.5 and PR-AUC = 0.53).

Normal-to-stress response evaluation

The *sInterRF* model achieved the highest AUC (Fig. 4), showing a significant improvement over *sRNARFTarget* ($p = 0.004$) and *CopraRNA* ($p = 0.03$). The *SEM_GraphRNA_CopraRNA_sInterRF* model achieved the highest pAUC, PR-AUC, precision, and specificity, and significantly outperformed *CopraRNA* in AUC ($p = 0.0001$). The *SEM_kGraphRNA_sInterRF* model achieved the best F1, MCC, and recall scores, with performance comparable to *sRNARFTarget* and *CopraRNA*, while significantly outperforming *RNAup* ($p = 0.02$). All models exceeded the random baselines (AUC = 0.5 and PR-AUC = 0.5). Although the results in this task are acceptable, they are lower compared to the previous two tasks, and the differences between the models are less significant.

Recovering interactions from low-throughput experiments (HT-to-LT)

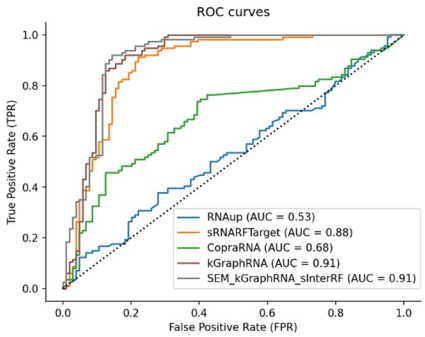
The second hypothesis we tested was whether an ML model trained on interactions generated through high-throughput (HT) techniques could accurately predict interactions (P) and non-interactions (N) that were verified through low-throughput (LT)

LB-to-m63

A.

Model	AUC	pAUC	PR-AUC	Accuracy	Precision	Recall	Specificity	F1	MCC
RNAup	0.53	0.52	0.56	0.54	0.59	0.37	0.72	0.45	0.10
sRNARFTarget	0.88	0.66	0.81	0.84	0.82	0.90	0.78	0.86	0.69
CopraRNA	0.68	0.59	0.68	0.67	0.67	0.73	0.61	0.70	0.34
sInterXGB	0.76	0.56	0.71	0.73	0.67	**0.96**	0.49	0.79	0.51
sInterRF	0.80	0.60	0.77	0.76	0.70	**0.95**	0.55	0.80	0.55
GraphRNA	0.89	0.69	0.83	**0.88**	**0.88**	0.88	**0.88**	**0.88**	**0.75**
kGraphRNA	**0.91**	**0.71**	**0.84**	0.87	0.85	0.91	0.82	**0.88**	0.73
SEM_GraphRNA_CopraRNA	0.76	0.68	0.79	0.72	0.71	0.77	0.65	0.74	0.43
SEM_GraphRNA_CopraRNA_sInterRF	0.82	**0.71**	**0.84**	0.77	0.79	0.75	0.79	0.77	0.53
SEM_kGraphRNA_sInterRF	**0.91**	**0.72**	**0.88**	**0.89**	**0.87**	0.91	**0.86**	**0.89**	**0.77**

B.



C.

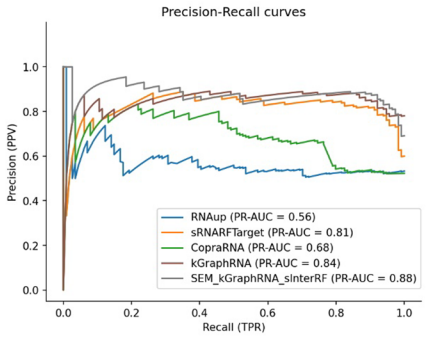


Fig. 3 **A.** Performance metrics of all models trained on 6539 LB medium interactions (P: 3251, N: 3288) and tested on 218 m63 medium interactions (P:114, N: 104). **B.** ROC curves and **C.** Precision-Recall curves of our top performing models (in terms of AUC and PR-AUC) compared to existing models: *RNAup*, *sRNARFTarget*, and *CopraRNA*. In panel **A**, the two top scores per metric are marked with green and gray, respectively. The pAUC is computed for $FPR \leq 0.15$. In panel **B**, the dotted diagonal line represents the baseline of a random classifier

experiments, which are generally considered to have stronger experimental support. Demonstrating such capability is particularly powerful given the vast amount of data continuously generated through HT technologies across diverse bacterial species [14, 70–72]. We supplemented the HT dataset of 3856 positive samples, with an equal number of negative samples. Post-filtering of interactions due to *RNAup* error or < 5 base pairs, the final HT train set consisted of 6585 interactions (P: 3356, N: 3229).

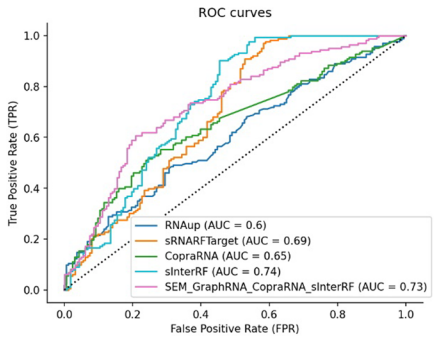
We report the performance metrics of all models on a test set of 391 (P: 227, N: 164) LT interactions (see Fig. 5). Generally, the task of predicting LT interactions from HT interactions proved more challenging for all models compared to predicting new HT interactions under unseen conditions. The *SEM_GraphRNA_CopraRNA* model achieved the best results in this task in most metrics, including AUC, pAUC, PR-AUC, accuracy, F1, and MCC, significantly outperforming *CopraRNA* ($p = 0.004$) and *sRNARFTarget* ($p = 3.64e^{-5}$). While *SEM_GraphRNA_CopraRNA* also performed better than *GraphRNA and RNAup*, the differences between the ROC curves of the models were not statistically significant (p=0.373 and $p = 0.11$, respectively). As individual models, *GraphRNA* and *CopraRNA* achieved comparable AUC, pAUC, and PR-AUC results, with a p-value of 0.596 between their ROC curves. Notably, *GraphRNA* achieved the highest precision (=0.85) and specificity (=0.91), which

Normal-to-stress

A.

Model	AUC	pAUC	PR-AUC	Accuracy	Precision	Recall	Specificity	F1	MCC
RNAup	0.60	0.56	0.62	0.58	0.61	0.47	0.69	0.53	0.17
sRNARFTarget	0.69	0.54	0.62	0.69	0.62	**0.96**	0.42	0.76	0.45
CopraRNA	0.65	**0.57**	0.65	0.63	**0.66**	0.55	**0.72**	0.60	0.27
sInterXGB	0.64	0.51	0.52	0.53	0.53	0.53	0.53	0.53	0.06
sInterRF	**0.74**	0.54	0.65	**0.72**	**0.66**	0.90	0.55	**0.76**	**0.47**
GraphRNA	0.63	0.54	0.60	0.62	0.58	0.86	0.38	0.69	0.27
kGraphRNA	0.65	0.55	0.62	0.66	0.60	0.93	0.39	0.73	0.38
SEM_GraphRNA_CopraRNA	0.66	0.56	**0.66**	0.65	**0.66**	0.63	0.68	0.64	0.31
SEM_GraphRNA_CopraRNA_sInterRF	**0.73**	**0.57**	**0.71**	0.70	**0.74**	0.60	**0.79**	0.66	0.40
SEM_kGraphRNA_sInterRF	0.69	0.55	0.63	**0.71**	0.63	**0.99**	0.43	**0.77**	**0.50**

B.



C.

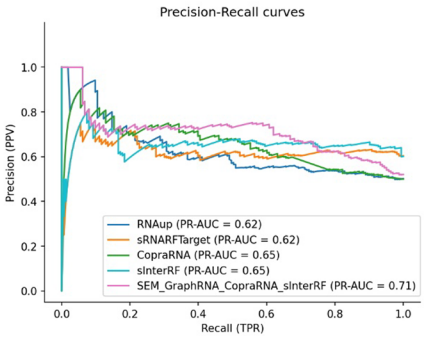


Fig. 4 **A.** Performance metrics of all models trained on 5670 normal (no stress) interactions (P: 2808, N: 2862) and tested on 326 iron limitation stress response interactions (P:163, N: 163). **B.** ROC curves and **C.** Precision-Recall curves of our top performing models (in terms of AUC and PR-AUC) compared to existing models: *RNAup*, *sRNARFTarget*, and *CopraRNA*. In panel **A**, the two top scores per metric are marked with green and gray, respectively. The pAUC is computed for $FPR \leq 0.15$. In panel **B**, the dotted diagonal line represents the baseline of a random classifier

greatly exceeded the scores of existing models, though it had a lower recall. *RNAup,* on the other hand, achieved the highest recall. All models exceeded the random baselines (AUC = 0.5 and PR-AUC = 0.58).

Features contribution analysis

We assessed the relative contribution of different interaction features to the model's prediction using SHAP (SHapley Additive exPlanations) [63]. First, we describe the k-mer frequency difference features (3-mer-diff) selected by the mRMR filter method for the train set of each unseen condition individually, as well as for the train set consisting of all HT interactions (HT-to-LT evaluation). Then, we analyze the contribution of the most important features learned by the *sInterRF* model on each train set.

mRMR selected features

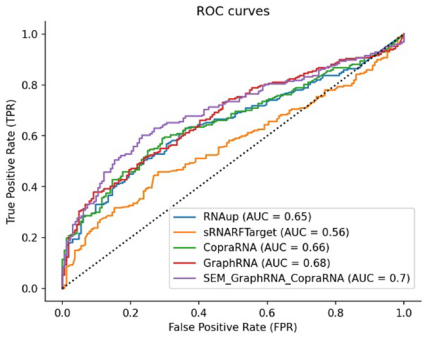
Before training the decision forests models (*sInterRF* and *sInterXGB*), we applied the mRMR method to each train set to auto-select the most relevant features out of the 64 3-mer-diff features (see Feature selection using mRMR). A total of 15 features were selected for the log-to-stationary evaluation in the following order: [TTT, TTC, CTG, CCT, TCT, AAT, GCT, GTT, CAG, TGG, GTG, TTA, TGC, TAT, ATT]. Similarly,

HT-to-LT

A.

Model	AUC	pAUC	PR-AUC	Accuracy	Precision	Recall	Specificity	F1	MCC
RNAup	0.65	0.60	0.75	0.63	0.71	**0.62**	0.65	**0.66**	0.26
sRNARFTarget	0.56	0.56	0.67	0.57	0.69	0.45	0.72	0.55	0.18
CopraRNA	0.66	0.61	0.76	**0.64**	0.73	0.59	0.71	0.65	0.29
sInterXGB	0.58	0.54	0.67	0.55	0.70	0.38	0.78	0.49	0.17
sInterRF	0.57	0.53	0.65	0.53	0.68	0.36	0.76	0.47	0.13
GraphRNA	**0.68**	**0.62**	**0.77**	0.60	**0.85**	0.37	**0.91**	0.52	**0.32**
kGraphRNA	0.56	0.55	0.67	0.57	0.70	0.44	0.74	0.54	0.19
SEM_GraphRNA_CopraRNA	**0.70**	**0.63**	**0.79**	**0.67**	**0.78**	**0.60**	0.77	**0.68**	**0.36**
SEM_GraphRNA_CopraRNA_sInterRF	0.68	0.60	0.76	0.63	0.75	0.55	0.74	0.63	0.29
SEM_kGraphRNA_sInterRF	0.57	0.56	0.68	0.53	0.76	0.29	**0.87**	0.42	0.19

B.



C.

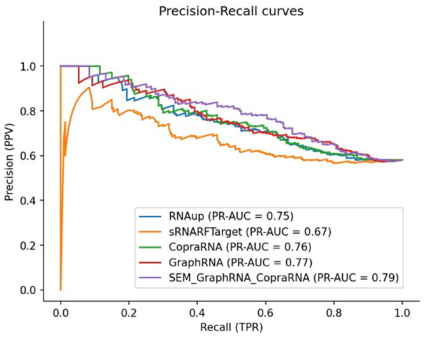


Fig. 5 **A.** Performance metrics of all models trained the final HT set of 6585 interactions (P: 3356, N: 3229) and tested on 391 LT interactions (P: 227, N: 164). **B.** ROC curves and **C.** Precision-Recall curves of our top performing models (in terms of AUC and PR-AUC) compared to existing models: *RNAup*, *sRNARFTarget*, and *CopraRNA*. In panel **A**, the two top scores per metric are marked with green and gray, respectively. The pAUC is computed for $FPR \leq 0.15$. In panel **B**, the dotted diagonal line represents the baseline of a random classifier

4 features were selected for the LB-to-m63 evaluation [TTT, TTA, CTG, ATT], 17 features were selected for the normal-to-stress evaluation [TTT, TTA, CCT, GTT, CCA, CAG, CTG, CTT, TTC, GTG, TGG, GCT, TAT, GGT, TGC, TCT, ATT], and 5 features selected for the HT-to-LT evaluation [TTT, TTA, TAT, ATT, CTG]. The selected 3-mer-diff features were added to the 23 local-interaction-based features constructing a final heterogenous feature set used by the decision forests models per evaluation.

We note that the train sets of the LB-to-m63 and HT-to-LT evaluations consist of interactions from both the log and stationary growth stages. For these sets, only a few 3-mer-diff features were auto-selected by the mRMR filter method, with 4 features selected for LB-to-m63 and 5 for HT-to-LT (4 of which were shared between the two). However, when the train sets consisted solely of log growth stage interactions, additional 3-mer-diff features were found as relevant by the mRMR method, with 15 features selected for the log-to-stationary and 17 for the normal-to-stress (14 of which were shared between the two).

SHAP

We analyzed the top 20 most important features learned by the *sInterRF* model from the individual train sets of the seen-to-unseen conditions evaluations, and the HT train set

used in the HT-to-LT evaluation. The SHAP values for each interaction sample (local) alongside the normalized contribution (*NC*) measure (global) are presented for the top features of each dataset (Fig. 6). A detailed description of all interaction features is provided in the supplementary materials (Table S1).

Across datasets, we observed that most of the 3-mers-diff features had greater contributions (*NC*) to model predictions than that of the local-interaction-based features (Fig. 6). The SHAP values, portraying local explanations, provided a detailed view of the impact of each feature on model predictions. It showed that the TTT feature, i.e., the trinucleotide frequency difference measured by subtracting the sRNA frequency from the mRNA frequency, was the most contributing feature for predicting interactions in all

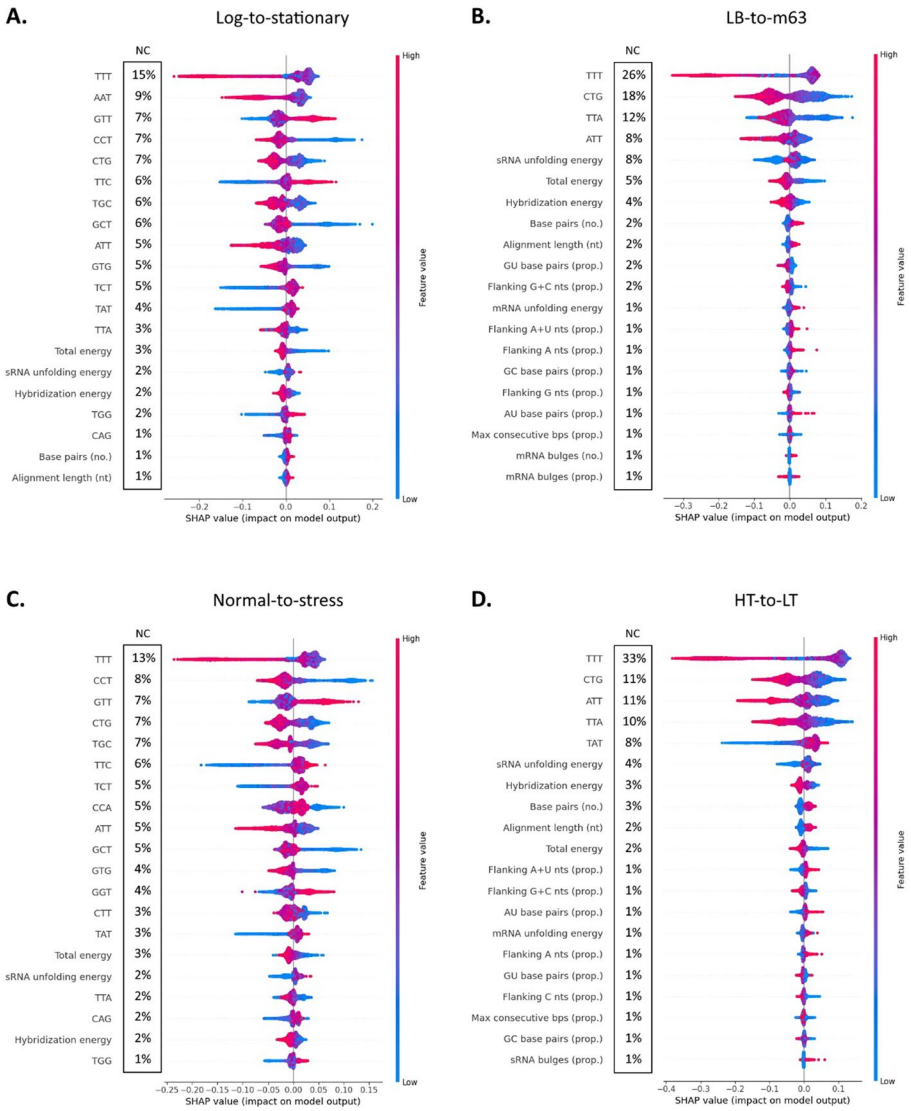


Fig. 6 SHAP values and normalized contributions (NC) of the top 20 most important features learned by the *sInterRF* model on the train set in each evaluation (feature descriptions provided in supplementary Table S1). Each dot within a subplot corresponds to a single sRNA-mRNA interaction sample; where its horizontal position reflects the impact that feature has on the model's prediction for that sample. The color of each dot represents the feature value, with red indicating high values and blue indicating low values

datasets, with a higher (red) feature value decreasing the predicted interaction probability. However, we observed variations in the SHAP value distributions of the TTT feature across datasets. Notably, some interactions with lower feature values (blue dots) were scattered along the negative x-axis, exhibiting negative SHAP values in the LB-to-m63 and HT-to-LT train sets (Fig. 6B, D). The CTG feature was ranked in the top 5 most contributing features in all datasets, with a higher (red) feature value lowering the interaction probability, and vice versa. We observed that CTG, ATT, and TTA demonstrated greater contributions in the LB-to-m63 ($NC = 18\%, 8\%, 12\%$, Fig. 6B) and HT-to-LT ($NC = 11\%, 11\%, 10\%$, Fig. 6D) train sets compared to their contributions in the log-to-stationary ($NC = 7\%, 5\%, 3\%$, Fig. 6A) and normal-to-stress ($NC = 7\%, 5\%, 2\%$, Fig. 6C) train sets. However, the trends in the SHAP values of these features were similar across all datasets, with higher (red) feature values decreasing the likelihood of interaction. The log-to-stationary and normal-to-stress datasets showed similar contributions (*NC*) and trends in the SHAP values of the 3-mers features (Fig. 6A, C), as higher (red) values of GTT and TTC increased the interaction probability. In contrast, higher (red) values of CCT and TGC decreased the interaction probability and vice versa.

Among the local-interaction-based features, the 5 most contributing features were the sRNA unfolding energy, the total energy, the hybridization energy, the number of base pairs, and the alignment length in slightly different orderings across datasets. Greater contributions of these features were observed in the LB-to-m63 ($NC = 8\%, 5\%, 4\%, 2\%, 2\%$, Fig. 6B) and HT-to-LT ($NC = 4\%, 2\%, 3\%, 2\%, 2\%$, Fig. 6D) train sets compared to their contributions in the log-to-stationary and normal-to-stress train sets ($NC = 2\%, 3\%, 2\%, \leq 1\%, \leq 1\%$, Fig. 6A, C). The energy features had shown similar trends in the SHAP values across the datasets. Particularly, greater absolute values of total (ΔG) and hybridization (ΔG_h) energies (blue, as $\Delta G < 0$ and $\Delta G_h < 0$) increased the interaction probability, and vice versa. The SHAP values, reflecting the impact of sRNA unfolding energy on model predictions, revealed different distributions across datasets. While lower (blue) sRNA unfolding energy decreased the likelihood of interaction in the log-to-stationary and normal-to-stress train sets (Fig. 6A, C), interactions with lower feature values (blue dots) were scattered along the x-axis, exhibiting both positive and negative SHAP values in the LB-to-m63 and HT-to-LT train sets (Fig. 6B, D). Features based on the duplex structure, have shown similar trends in SHAP values across datasets. We note a rise in interaction probability for higher (red) values of the number of bps, the alignment length (nt), and the proportion of A and U in the mRNA flanking region (Fig. 6A, B, D), as well as for lower values (blue) of the proportion of G and C in the mRNA flanking region (Fig. 6B, D).

Discussion

Computational sRNA-target prediction and HT interaction datasets

Reliable computational methods for sRNA target prediction can significantly enhance the discovery of regulatory relationships in bacteria by providing trustworthy candidates for biological validation. Yet, the development of such methods is quite challenging due to the high variability in length and structure of the sRNA molecules, as well as the uncertainty in the location and length of the sRNA interacting seed and the mRNA target site. Due to the complex nature of sRNA-target interactions in bacteria, advanced

computational methods that consider different aspects and various features of interaction can potentially improve the predictive performance of the current tools. In particular, ML methods that excel at capturing complex data relationships and modeling non-linear patterns have already demonstrated remarkable success in addressing various computational challenges in biology and medicine, outperforming non-learning methods [29, 30].

ML models represent a relatively new approach in the field of sRNA-mRNA interaction prediction and have the potential to further improve as more data is collected. Although high-throughput (HT) experimental techniques, such as CLASH and RIL-Seq, can recover numerous sRNA-target interactions simultaneously, they are inherently limited by inefficiencies in certain steps (e.g., crosslinking and ligation), capturing only a subset of interactions within the sample. Moreover, since these methods are performed under specific environmental conditions, they are restricted to detecting interactions between sRNAs and targets expressed under those particular conditions. In *E. coli*, for example, multiple datasets have been collected for the *K12 MG1655* strain under varying growth conditions, related to growth stage, medium, and stress response. This variability introduces new research questions regarding the capabilities of ML methods in different prediction tasks, such as learning from experimentally observed interactions under certain conditions to predict new interactions in unseen conditions in specific bacteria.

Seen-to-unseen conditions evaluations

We assessed the performance of our newly developed models in predicting new interactions observed in different unseen environmental conditions within the same *E.coli* strain. By comparing several learning and non-learning tools, we have generally demonstrated that ML models can succeed in this task, with our models reaching an AUC of 0.74–0.91. However, the degree of success varied by the unseen condition. Particularly, we observed lower performance of all models in predicting new interactions under stress response (normal-to-stress) versus interactions in the stationary growth condition (log-to-stationary), although similar train sets were used in both evaluations. Our *kGraphRNA* and *SEM_kGraphRNA_sInterRF* models achieved the top two AUC and PR-AUC scores in the log-to-stationary and LB-to-m63 evaluations, significantly outperforming all existing tools in the log-to-stationary evaluation. In the LB-to-m63 evaluation, these two models significantly suppressed both *CopraRNA* and *RNAup*, yet their relative improvement in the ROC curve compared to *sRNARFTarget* was not significant. Given that the test set of the LB-to-m63 evaluation was relatively small, we hypothesize that utilizing a larger test set specific to m63 interactions would yield more statistically significant results. Additionally, *SEM_kGraphRNA_sInterRF* achieved results comparable to those of existing tools in the normal-to-stress evaluation.

HT-to-LT evaluation

In this study, we also evaluated the models' ability to predict interactions (P) and non-interactions (N) identified in low-throughput (LT) experiments, while learning from HT interactions in the same *E. coli* strain. LT methods, which preceded the emergence of HT technologies, infer sRNA-target interactions based on functional evidence, making them

generally considered to have stronger experimental support. These methods include approaches such as overexpression or depletion of specific sRNAs, followed by assessing their effects on the expression levels of potential targets. This assessment determines whether a positive interaction (i.e., upregulation or downregulation of the target) or a negative interaction (i.e., no significant effect) occurs under specific laboratory conditions. However, it is important to note that among the set of positive interactions, some may represent indirect targets of the examined sRNAs. As for HT interactions, they may represent transient connections and not necessarily indicate functional consequences. Moreover, HT methods detect only positive interactions, necessitating that models trained on these datasets be complemented by negative interactions without experimental support, such as sRNA-mRNA pairs generated by randomly swapping sRNAs in the positive pairs. Another noteworthy distinction between HT and LT experimental methods is that HT approaches are often designed to capture sRNA-target interactions mediated by specific proteins, such as Hfq in our case. In contrast, LT methods may detect interactions that are not necessarily Hfq-mediated.

The results of the HT-to-LT evaluation clearly show that all models struggled with the task of predicting LT interactions, likely due to the methodological differences outlined above. Despite this, our *SEM_GraphRNA_CopraRNA* model achieved satisfactory performance (AUC = 0.7, PR-AUC = 0.79), significantly improving upon *CopraRNA* alone, the best competing model. Notably, previously reported methods did not differentiate between HT and LT datasets in their evaluations, and some studies ignored negative LT interactions.

sRNA-mRNA interaction features

The set of interaction features used by the decision forests incorporated local-interaction-based features extracted from *RNAup* duplexes and 3-mer-diff features. Given the relatively large number of 3-mer-diff features compared to the train set size, we applied the mRMR feature selection algorithm on the train sets to identify diverse and informative features of that type. This process significantly reduced the number of 3-mer-diff features, ranging from 4 to 17, depending on the train set. These selected features, combined with the 23 local-interaction (duplex-based) features, composed the final feature sets for SHAP analysis. The analysis of the top 20 most important features learned by *sInterRF*, revealed that the contribution of most 3-mers-diff features to model predictions consistently surpassed that of local-interaction-based features. Notably, the TTT feature emerged as the most influential across all datasets, with the highest normalized contribution (NC) values. Negative impact on model output (i.e., decrease in interaction probability) was generally associated with high (red) TTT values. However, in some samples, high TTT values had a positive impact on model prediction, as indicated by the distribution of red dots along the positive x-axis. We also observed a heterogeneous distribution of low (blue) TTT values across the x-axis, especially in the LB-to-m63 and HT-to-LT evaluations. This lack of separation between high and low feature values indicates an ambiguous influence of the TTT feature on model predictions, making it challenging to draw definitive conclusions about its effect on interaction probability.

Among the local-interaction-based features, the five most contributing features were related to energy and hybridization duplex strength, i.e., the sRNA unfolding energy,

the total energy, the hybridization energy, the number of base pairs, and the alignment length. Specifically, we observed that lower values (larger magnitude) of total and hybridization energies and higher values of the number of base pairs and alignment length increased the likelihood of interaction. These findings are consistent with previous findings on RNA-RNA interactions. Thermodynamics, in particular, free energy, has been established as a key characteristic of RNA interactions [73]. In accordance, previous studies have highlighted the significance of energy-related features in predicting general RNA-RNA interactions [26] and specifically sRNA-target interactions [27]. The impact of sRNA and mRNA unfolding energy observed in Hfq-mediated HT interactions aligns with previous experimental studies suggesting that Hfq's primary function in interaction mediation is to unfold the target site, making it accessible for sRNA binding [74, 75]. This role of Hfq in sRNA unfolding has also been demonstrated for a few specific sRNAs [76, 77].

sInterRF incorporates local interaction features alongside a selected subset of 3-mer-diff features adapted from *sRNARFTarget*. We compared the contributions of the 3-mer-diff features to *sInterRF* predictions against the contributions of these features to *sRNARFTarget* predictions reported in [34]. This comparison highlighted different influential features. This difference is likely due to the composition of the training sets — our model was trained solely on *E. coli* interactions, while *sRNARFTarget* used data from multiple bacterial species.

In the *kGraphRNA* model, we used 3-mer frequency features of the sRNA and mRNA to initialize the graph nodes. Interestingly, this addition of 3-mer features proved beneficial (as seen in the comparison between *kGraphRNA* and *GraphRNA*) across all seen-to-unseen conditions evaluations [0.84 vs 0.71; 0.91 vs 0.89; 0.65 vs 0.63], but not in the HT-to-LT evaluation [0.56 vs 0.68].

The most recent ML-based tool, *TargetRNA3,* used 9 most informative features from an initial set of 111 features for model training. SHAP analysis of these 9 features, revealed that *RNAplex: energy considering accessibility* was the most influential feature, with low energy values increasing the probability of interaction, consistent with the impacts of *RNAup*-based energy features measured in our model. An additional *TargetRNA3* feature related to duplex strength (specifically whether there is a *seed of length 7 bps*) influenced the model output similarly to duplex strength features (*the number or proportion of base pairs*) in our model, with higher feature values increasing the probability of interaction. There was no overlap between other features of *TargetRNA3* and *sInterRF*.

Challenges in comparing ML models' performance

It is crucial to emphasize the challenges involved in comparing ML models' performance. To facilitate a fair comparison, models must be retrained and tested on the same datasets, ensuring a controlled evaluation and minimizing biases that could arise from inconsistent datasets or evaluation strategies. Moreover, the ability to retrain models on diverse train sets is essential for exploring new hypotheses and adapting to a wide range of research objectives.

For *sRNARFTarget,* we were able to adapt their code for model initialization, feature extraction, and retraining on new train sets according to our evaluations. In contrast,

while *TargetRNA3* provides a pre-trained model alongside code and methods for inference, we could not use this model for comparison as we understood it was trained on *E.coli* datasets overlapping with our test sets. Additionally, adapting *TargetRNA3* for retraining on new datasets is not straightforward due to the complex features, which require additional data for calculation. As a result, a direct comparison with *TargetRNA3* was not feasible in the scope of this paper.

Based on our research and the realization that model retraining is essential for meaningful comparisons to other future tools, we provided the full source code, including model initialization, feature extraction, feature selection, training, and testing procedures for all models. In addition, we trained our models on the entire labeled data of *Escherichia coli K12 MG1655* and provided the prediction scores at https://doi.org/10.5281/zenodo.10134390 for any pair of sRNA and mRNA of the bacterial strain. By making the complete source code publicly available, we hope to encourage researchers to retrain and evaluate our models on new datasets for further studies and applications.

Model advantages, applications, and future work

In addition to enhancing the performance of sRNA-mRNA interaction prediction compared to other tools, our proposed models enable the prediction of interactions involving species-specific sRNAs, capabilities that tools like *sRNARFTarget* and *TargetRNA3* possess but *CopraRNA* lacks. Furthermore, GNN-based models offer an additional advantage by eliminating the dependency on external tools, such as *RNAplex* or *RNAup*, to compute the hybridization duplex or energy features, making them more scalable and run-time efficient. We anticipate that HT data will significantly expand in scale, incorporating additional environmental conditions in *E. coli* and other bacteria. This growth will allow for the refinement of our ML models and enable their application to an increasing number of bacterial species with available data. While this study rigorously evaluated our methods using interaction data specifically curated for the *E. coli K12 MG1655* strain, our methods are fully adaptable and can be applied to predict new interactions within other bacterial strains, provided that sufficient interaction data of the bacteria is available for training.

While other tools, such as *sRNARFTarget* and *TargetRNA3*, used datasets from multiple species spanning a broad phylogenetic range for training, it remains unclear whether this approach is advantageous, as we do not yet know if sRNA-mRNA interactions follow the same principles across all bacteria. Future research should investigate different strategies—such as how restricting the training data to a single strain, species, or closely related species affects the models' performance when predicting new interactions within species and cross-species. Generally, most of our methods can, in principle, be applied to predict interactions in new strains that were not part of the training set. The decision forest-based models, *sInterRF* and *sInterXGB*, can be trained on *E. coli* or any other species and then applied directly to sRNA-mRNA interactions from other bacterial strains, though this task was not evaluated in the current study.

kGraphRNA, which initializes sRNA and mRNA nodes using 3-mer frequency features, can be applied to entirely new sRNA and mRNA pairs from other strains or species. However, this capability was not evaluated in the current study as well. In contrast, the standard graph-based model, *GraphRNA*, is less suited for this purpose, as its

random initialization of sRNA and mRNA representations limits its ability to generalize to unseen RNAs. Future work will explore methods for transferring interaction knowledge across species to enhance the prediction of sRNA-target interactions in bacteria with limited or no existing interaction data, similar to studies performed on miRNA-target interactions [78, 79].

Conclusion

Advanced ML methods applied to HT interaction data can significantly enhance the predictive performance of the current non-learning computational tools for sRNA-target prediction, specifically for interactions in new unseen environmental conditions, while also shedding light on important characteristics that impact the likelihood of interaction, as demonstrated for the bacterial strain of *Escherichia coli K12 MG1655*. Furthermore, the increasing number and improved quality of sRNA-target interaction datasets, collected for multiple bacteria under various environmental conditions, are expected to further enhance the performance and utility of our proposed ML-based methods over time.

GNN models, whether used independently or as components of SEM models, consistently improved the predictive performance over existing methods across all evaluations, as demonstrated by ranking metrics and threshold-based metrics. Our methods are adaptable to any bacterial species or strain with available sRNA and mRNA sequences and a sample of sRNA-mRNA interactions for training. Future work will aim to develop state-of-the-art prediction methods for bacteria with limited interaction data and further enhance GNN-based models by incorporating additional types of nodes, relations, and interaction features.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-025-06153-w.

Additional file 1.

Acknowledgements
We thank Sahar Melamed for his guidance on bacterial environmental conditions.

Author contributions
S.C.: Conceptualization, Methodology, Software,Validation, Data Curation, Writing—Original Draft L.R.: Methodology, Supervision, Writing—Review & Editing I.V.L.: Conceptualization, Methodology, Data Curation, Supervision, Writing—Original Draft, Funding acquisition.

Funding
This work was supported by the Israel Science Foundation [520/20]. The funding agency had no role in study design, data collection, analysis and interpretation, or manuscript preparation.

Availability of data and materials
The prediction scores of our models can be found at https://doi.org/10.5281/zenodo.10134390 for any pair of sRNA and mRNA of the bacterial strain of Escherichia coli K12 MG1655 (NC_000913). The complete source code, including documented examples for API usage and datasets, is available at https://github.com/IsanaVekslerLublinsky/sInterModels.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 11 January 2024 Accepted: 30 April 2025
Published online: 21 May 2025

References

1. Storz G, Vogel J, Wassarman KM. Regulation by small rnas in bacteria: expanding frontiers. Mol Cell. 2011;43(6):880–91. https://doi.org/10.1016/j.molcel.2011.08.022.
2. Bloch S, Węgrzyn A, Węgrzyn G, Nejman-Faleńczyk B. Small and smaller—sRNAs and MicroRNAs in the regulation of toxin gene expression in prokaryotic cells: a mini-review. Toxins (Basel). 2017;9(6):181. https://doi.org/10.3390/toxins9060181.
3. Wagner EGH, Romby P. Small RNAs in Bacteria and Archaea. Amsterdam: Elsevier Ltd; 2015.
4. Jagodnik J, Brosse A, Le Lam TN, Chiaruttini C, Guillier M. Mechanistic study of base-pairing small regulatory RNAs in bacteria. Methods. 2017;117(2017):67–76. https://doi.org/10.1016/j.ymeth.2016.09.012.
5. Jagodnik J, Chiaruttini C, Guillier M. Stem-loop structures within mRNA coding sequences activate translation initiation and mediate control by small regulatory RNAs. Mol Cell. 2017;68(1):158-170.e3. https://doi.org/10.1016/j.molcel.2017.08.015.
6. Gottesman S, Storz G. Bacterial small RNA regulators: versatile roles and rapidly evolving variations. Cold Spring Harb Perspect Biol. 2011;3(12):a003798–a003798. https://doi.org/10.1101/cshperspect.a003798.
7. Melamed S, Adams PP, Zhang A, Zhang H, Storz G. RNA-RNA interactomes of ProQ and HfQ reveal overlapping and competing roles. Mol Cell. 2020;77(2):411-425.e7. https://doi.org/10.1016/j.molcel.2019.10.022.
8. Hör J, Matera G, Vogel J, Gottesman S, Storz G. Trans-acting small RNAs and their effects on gene expression in escherichia coli and salmonella enterica. EcoSal Plus. 2020. https://doi.org/10.1128/ecosalplus.ESP-0030-2019.
9. Park S, et al. Dynamic interactions between the RNA chaperone Hfq, small regulatory RNAs, and mRNAs in live bacterial cells. Elife. 2021;10:1–25. https://doi.org/10.7554/eLife.64207.
10. King AM, Vanderpool CK, Degnan PH. sRNA target prediction organizing tool (SPOT) integrates computational and experimental data to facilitate functional characterization of bacterial small RNAs. mSphere. 2019;4(1):1–19. https://doi.org/10.1128/mSphere.00561-18.
11. Wang J, et al. sRNATarBase 3.0: an updated database for sRNA-target interactions in bacteria. Nucleic Acids Res. 2016;44(D1):D248–53. https://doi.org/10.1093/nar/gkv1127.
12. Hör J, Gorski SA, Vogel J. Bacterial RNA biology on a genome scale. Mol Cell. 2018;70(5):785–99. https://doi.org/10.1016/j.molcel.2017.12.023.
13. Melamed S, et al. Global mapping of small RNA-target interactions in bacteria. Mol Cell. 2016;63(5):884–97. https://doi.org/10.1016/j.molcel.2016.07.026.
14. Waters SA, et al. Small RNA interactome of pathogenic E. coli revealed through crosslinking of RNase E. EMBO J. 2017;36(3):374–87. https://doi.org/10.15252/embj.201694639.
15. Iosub IA, et al. Hfq CLASH uncovers sRNA-target interaction networks linked to nutrient availability adaptation. Elife. 2020;9:1–33. https://doi.org/10.7554/eLife.54655.
16. Saliba A-E, Santos SC, Vogel J. New RNA-seq approaches for the study of bacterial pathogens. Curr Opin Microbiol. 2017;35:78–87. https://doi.org/10.1016/j.mib.2017.01.001.
17. Helwak A, Tollervey D. Mapping the miRNA interactome by cross-linking ligation and sequencing of hybrids (CLASH). Nat Protoc. 2014;9(3):711–28. https://doi.org/10.1038/nprot.2014.043.
18. Wright PR, et al. Comparative genomics boosts target prediction for bacterial small RNAs. Proc Natl Acad Sci. 2013;110(37):E3487–96. https://doi.org/10.1073/pnas.1303248110.
19. Han K, Tjaden B, Lory S. GRIL-seq provides a method for identifying direct targets of bacterial small regulatory RNA by in vivo proximity ligation. Nat Microbiol. 2016;2(3):16239. https://doi.org/10.1038/nmicrobiol.2016.239.
20. Pain A, Ott A, Amine H, Rochat T, Bouloc P, Gautheret D. An assessment of bacterial small RNA target prediction programs. RNA Biol. 2015;12(5):509–13. https://doi.org/10.1080/15476286.2015.1020269.
21. Lorenz R, et al. ViennaRNA Package .20. Algorithms Mol Biol. 2011. https://doi.org/10.1186/1748-7188-6-26.
22. Mückstein U, Tafer H, Hackermüller J, Bernhart SH, Stadler PF, Hofacker IL. Thermodynamics of RNA–RNA binding. Bioinformatics. 2006;22(10):1177–82. https://doi.org/10.1093/bioinformatics/btl024.
23. Tafer H, Hofacker IL. RNAplex: a fast tool for RNA–RNA interaction search. Bioinformatics. 2008;24(22):2657–63. https://doi.org/10.1093/bioinformatics/btn193.
24. Busch A, Richter AS, Backofen R. IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. Bioinformatics. 2008;24(24):2849–56. https://doi.org/10.1093/bioinformatics/btn544.
25. Tjaden B. TargetRNA: a tool for predicting targets of small RNA action in bacteria. Nucleic Acids Res. 2008;36:109–13. https://doi.org/10.1093/nar/gkn264.
26. Gelhausen R, Will S, Hofacker IL, Backofen R, Raden M. IntaRNAhelix - composing RNA–RNA interactions from stable inter-molecular helices boosts bacterial sRNA target prediction. J Bioinform Comput Biol. 2019;17(05):1940009. https://doi.org/10.1142/S0219720019400092.
27. Eggenhofer F, Tafer H, Stadler PF, Hofacker IL. RNApredator: fast accessibility-based prediction of sRNA targets. Nucleic Acids Res. 2011. https://doi.org/10.1093/nar/gkr467.
28. Kery MB, Feldman M, Livny J, Tjaden B. TargetRNA2: identifying targets of small regulatory RNAs in bacteria. Nucleic Acids Res. 2014;42(W1):W124–9. https://doi.org/10.1093/nar/gku317.
29. Wang H, et al. Scientific discovery in the age of artificial intelligence. Nature. 2023;620(7972):47–60. https://doi.org/10.1038/s41586-023-06221-2.

30. Li MM, et al. Contextual AI models for single-cell protein biology. Nat Methods. 2024;21(8):1546–57. https://doi.org/10.1038/s41592-024-02341-3.
31. Cao Y, et al. sRNATarget: a web server for prediction of bacterial sRNA targets. Bioinformation. 2009;3(8):364–6. https://doi.org/10.6026/97320630003364.
32. Zhao Y, et al. Construction of two mathematical models for prediction of bacterial sRNA targets. Biochem Biophys Res Commun. 2008;372(2):346–50. https://doi.org/10.1016/j.bbrc.2008.05.046.
33. Ying X, Cao Y, Wu J, Liu Q, Cha L, Li W. sTarPicker: a Method for efficient prediction of bacterial sRNA targets based on a two-step model for hybridization. PLoS ONE. 2011;6(7): e22705. https://doi.org/10.1371/journal.pone.0022705.
34. Naskulwar K, Peña-Castillo L. sRNARFTarget: a fast machine-learning-based approach for transcriptome-wide sRNA target prediction. RNA Biol. 2022;19(1):44–54. https://doi.org/10.1080/15476286.2021.2012058.
35. Tjaden B. TargetRNA3: predicting prokaryotic RNA regulatory targets with machine learning. Genome Biol. 2023;24(1):276. https://doi.org/10.1186/s13059-023-03117-2.
36. Cohen S, Maximof E, Rokach S, Tadeski M, Veksler-Lublinsky I. sInterBase: a comprehensive database of Escherichia coli sRNA–mRNA interactions. Bioinformatics. 2023;39(4):1–3. https://doi.org/10.1093/bioinformatics/btad172.
37. Backofen R, Hess WR. Computational prediction of sRNAs and their targets in bacteria. RNA Biol. 2010;7(1):33–42. https://doi.org/10.4161/rna.7.1.10655.
38. Umu SU, Gardner PP. A comprehensive benchmark of RNA–RNA interaction prediction tools for all domains of life. Bioinformatics. 2017;33(7):988–96. https://doi.org/10.1093/bioinformatics/btw728.
39. Peer A, Margalit H. Accessibility and evolutionary conservation mark bacterial small-RNA target-binding regions. J Bacteriol. 2011;193(7):1690–701. https://doi.org/10.1128/JB.01419-10.
40. C. Ding and H. Peng, Minimum redundancy feature selection from microarray gene expression data. *Computational Systems Bioinformatics. Proceedings of the 2003 IEEE Bioinformatics Conference.* 2003. https://doi.org/10.1109/CSB.2003.1227396.
41. Peng H, Long F, Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell. 2005;27(8):1226–38. https://doi.org/10.1109/TPAMI.2005.159.
42. Chandra B, Gupta M. An efficient statistical feature selection approach for classification of gene expression data. J Biomed Inform. 2011;44(4):529–35. https://doi.org/10.1016/j.jbi.2011.01.001.
43. Alshamlan H, Badr G, Alohali Y. mRMR-ABC: a hybrid gene selection algorithm for cancer classification using microarray gene expression profiling. Biomed Res Int. 2015;2015:1–15. https://doi.org/10.1155/2015/604910.
44. A. R. Subhani, W. Mumtaz, N. Kamil, N. M. Saad, N. Nandagopal, and A. S. Malik mRMR based feature selection for the classification of stress using EEG. *2017 Eleventh International Conference on Sensing Technology (ICST)*. 2017. https://doi.org/10.1109/ICSensT.2017.8304499.
45. Kaya D. The mRMR-CNN based influential support decision system approach to classify EEG signals. Measurement. 2020;156: 107602. https://doi.org/10.1016/j.measurement.2020.107602.
46. He X, Liao L, Zhang H, Nie L, Hu X, Chua TS. Neural collaborative filtering. In: Proceedings of the 26th international conference on world wide web. 2017. p. 173–82.
47. Jiang H, Wang J, Li M, Lan W, Wu F-X, Pan Y. miRTRS: a recommendation algorithm for predicting miRNA targets. IEEE/ACM Trans Comput Biol Bioinforma. 2018;17(3):1032–41. https://doi.org/10.1109/TCBB.2018.2873299.
48. Bang-Jensen J, Gutin G. Digraphs: theory, algorithms and applications. London: Springer, London; 2008.
49. Yoon M, Palowitch J, Zelle D, Hu Z, Salakhutdinov R, Perozzi B. Zero-shot transfer learning within a heterogeneous graph via knowledge transfer networks. NeurIPS. 2022 [Online]. Available: http://arxiv.org/abs/2203.02018.
50. Zhang X-M, Liang L, Liu L, Tang M-J. Graph neural networks and their current applications in bioinformatics. Front Genet. 2021;12(July):1–22. https://doi.org/10.3389/fgene.2021.690049.
51. Réau M, Renaud N, Xue LC, Bonvin AMJJ. DeepRank-GNN: a graph neural network framework to learn patterns in protein–protein interfaces. Bioinformatics. 2023. https://doi.org/10.1093/bioinformatics/btac759.
52. Gaudelet T, et al. Utilizing graph machine learning within drug discovery and development. Brief Bioinform. 2021. https://doi.org/10.1093/bib/bbab159.
53. E. Choi, M. T. Bahadori, L. Song, W. F. Stewart, and J. Sun, GRAM: Graph-based Attention Model for Healthcare Representation Learning. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2017, https://doi.org/10.1145/3097983.3098126.
54. Wu Z, Pan S, Chen F, Long G, Zhang C, Yu PS. A comprehensive survey on graph neural networks. IEEE Trans Neural Networks Learn Syst. 2021;32(1):4–24. https://doi.org/10.1109/TNNLS.2020.2978386.
55. W. Hamilton, Z. Ying, and L. Jure Inductive representation learning on large graphs. Adv Neural Inf Proc Syst (NIPS) 30. 2017
56. M. Fey and J. E. Lenssen, Fast Graph Representation Learning with PyTorch Geometric. 2019. http://arxiv.org/abs/1903.02428.
57. Santhanam R, Uzir N, Raman S, Banerjee S. Experimenting XGBoost algorithm for prediction and classification of different datasets. Int J Control Theory Appl. 2017;9(March):651–62.
58. Shwartz-Ziv R, Armon A. Tabular data: deep learning is not all you need. Inf Fusion. 2022;81:84–90. https://doi.org/10.1016/j.inffus.2021.11.011.
59. Cohen S, Rokach L, Motro Y, Moran-Gilad J, Veksler-Lublinsky I. minMLST : machine learning for optimization of bacterial strain typing. Bioinformatics. 2021;37(3):303–11. https://doi.org/10.1093/bioinformatics/btaa724.
60. Wang P, Zhang G, Yu Z-G, Huang G. A deep learning and XGBoost-based method for predicting protein-protein interaction sites. Front Genet. 2021;12(October):1–11. https://doi.org/10.3389/fgene.2021.752732.
61. Ma B, Yan G, Chai B, Hou X. XGBLC: an improved survival prediction model based on XGBoost. Bioinformatics. 2022;38(2):410–8. https://doi.org/10.1093/bioinformatics/btab675.
62. Breiman L. Random forests. Mach Learn. 2001;45:5–32. https://doi.org/10.1023/A:1010933404324.
63. Lundberg SM, Lee S-I. A Unified Approach to Interpreting Model Predictions. Adv Neural Inf Process Syst. 2017;30:4766–75.

64. Lundberg SM, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell. 2020;2(1):56–67. https://doi.org/10.1038/s42256-019-0138-9.
65. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44(3):837. https://doi.org/10.2307/2531595.
66. Robin X, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011;12(1):77. https://doi.org/10.1186/1471-2105-12-77.
67. Youden WJ. Index for rating diagnostic tests. Cancer. 1950;3(1):32–5. https://doi.org/10.1002/1097-0142(1950)3:1%3c32::AID-CNCR2820030106%3e3.0.CO;2-3.
68. López-Ratón M, Rodríguez-Álvarez MX, Suárez CC, Sampedro FG. OptimalCutpoints : an R package for selecting optimal cutpoints in diagnostic tests. J Stat Softw. 2014;61(8):1–36. https://doi.org/10.18637/jss.v061.i08.
69. Schisterman EF, Faraggi D, Reiser B, Hu J. Youden Index and the optimal threshold for markers with mass at zero. Stat Med. 2008;27(2):297–315. https://doi.org/10.1002/sim.2993.
70. Pearl Mizrahi S, et al. The impact of Hfq-mediated sRNA-mRNA interactome on the virulence of enteropathogenic Escherichia coli. Sci Adv. 2021. https://doi.org/10.1126/sciadv.abi8228.
71. Mediati DG, et al. RNase III-CLASH of multi-drug resistant Staphylococcus aureus reveals a regulatory mRNA 3′UTR required for intermediate vancomycin resistance. Nat Commun. 2022;13(1):3558. https://doi.org/10.1038/s41467-022-31177-8.
72. Huber M, et al. An RNA sponge controls quorum sensing dynamics and biofilm formation in Vibrio cholerae. Nat Commun. 2022;13(1):7585. https://doi.org/10.1038/s41467-022-35261-x.
73. Nowakowski J, Tinoco I. RNA structure and Stability. Semin Virol. 1997;8(3):153–65. https://doi.org/10.1006/smvy.1997.0118.
74. Hoekzema M, Romilly C, Holmqvist E, Wagner EGH. Hfq-dependent mRNA unfolding promotes sRNA-based inhibition of translation. EMBO J. 2019;38(7):1–14. https://doi.org/10.15252/embj.2018101199.
75. Geissmann TA, Touati D. Hfq, a new chaperoning role: Binding to messenger RNA determines access for small RNA regulator. EMBO J. 2004;23(2):396–405. https://doi.org/10.1038/sj.emboj.7600058.
76. Wu P, et al. The important conformational plasticity of DsrA sRNA for adapting multiple target regulation. Nucleic Acids Res. 2017;45(16):9625–39. https://doi.org/10.1093/nar/gkx570.
77. Cai H, Roca J, Zhao Y-F, Woodson SA. Dynamic refolding of OxyS sRNA by the Hfq RNA chaperone. J Mol Biol. 2022;434(18): 167776. https://doi.org/10.1016/j.jmb.2022.167776.
78. Ben Or G, Veksler-Lublinsky I. Comprehensive machine-learning-based analysis of microRNA–target interactions reveals variable transferability of interaction rules across species. BMC Bioinformatics. 2021;22(1):264. https://doi.org/10.1186/s12859-021-04164-x.
79. Hadad E, Rokach L, Veksler-Lublinsky I. Empowering prediction of miRNA–mRNA interactions in species with limited training data through transfer learning. Heliyon. 2024;10(7): e28000. https://doi.org/10.1016/j.heliyon.2024.e28000.

Publisher's Note