





# Computational Understanding of Pairwise Interactions in ncRNA Data

Marco Nicolini<sup>1, </sup>, Federico Stacchiotti<sup>1, </sup>, Elena Casiraghi<sup>1,2,3,4, </sup>, and Giorgio Valentini<sup>\*,1,2, </sup>

<sup>1</sup> AnacletoLab - Dipartimento Informatica, Università degli Studi di Milano, Milan, Italy.

<sup>2</sup> ELLIS - European Lab for Learning and Intelligent Systems, Milan Unit.

<sup>3</sup> Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, CA, United States.

<sup>4</sup> Department of Computer Science, Aalto University, Espoo, Finland.

\*corresponding author email: giorgio.valentini@unimi.it

**Keywords:** ncRNA-ncRNA interaction, Machine Learning, non-coding RNA, Large Language Models.

**Abstract.** Non-coding RNAs (ncRNAs) are central to regulating diverse cellular processes, yet their complex interaction networks remain poorly characterized due to experimental and computational challenges. We present *ncRNA-CUPID* (Computational Understanding of Pairwise Interactions in ncRNA Data), a novel deep learning framework that predicts pairwise ncRNA interactions solely from primary sequence information. By leveraging embeddings from a pre-trained ncRNA language model and a dedicated feed-forward neural network classifier, *ncRNA-CUPID*, differently from previous methods, can learn and predict virtually any type of ncRNA interactions and represents the first attempt to predict ncRNA interactions directly from RNA sequences using a transformer-based model.

## 1 Introduction

Understanding RNA-RNA interactions is critical for deciphering the regulatory circuits that orchestrate gene expression, RNA processing, and signal transduction. Non-coding RNAs (ncRNAs), despite lacking protein-coding potential, play pivotal roles in these processes. However, experimental mapping of ncRNA interactions remains challenging due to the limitations of existing experimental and computational techniques [1].

Methods such as Minimum Free Energy (MFE) calculations and accessibility-based models have been used to predict RNA-RNA interactions [2, 3], yet these approaches rely on predefined parameters and simplified energy models. Moreover, experimental techniques such as RNA Antisense Purification (RAP-RNA) offer validation but remain limited by their high cost and labor intensity [4].

Recent advances in deep learning have enabled direct modeling of complex biological sequences. Methods such as Convolutional Neural Networks (CNNs) and deep forests [5, 6] have been applied to RNA-protein interaction prediction, while graph-based approaches embed heterogeneous networks of ncRNAs and diseases using multigraph contrastive learning [7] or graph representation learning techniques [8]. While effective, these methods often rely on predefined feature extraction, graph structures, or supervised training, limiting their adaptability to novel ncRNA sequences.

In contrast, Large Language Models (LLMs) can directly learn from large corpora of proteins or RNA data [9, 10, 11, 12], capturing intricate interaction motifs beyond predefined energy models or graph-based constraints. GenerRNA [10], for instance, learns long-range dependencies via masked language modeling, processing full-length ncRNA sequences without truncation.

In this work we introduce *ncRNA-CUPID*, a deep learning framework that predicts ncRNA interactions using only sequence information. *ncRNA-CUPID* extracts embeddings from a pre-trained ncRNA language model and feeds a Feed-Forward Neural Network (FFNN) to automatically learn intricate sequence interaction features. This design circumvents the need for explicit thermodynamic parameterization and manually engineered features, offering a scalable and efficient alternative for uncovering novel regulatory interactions [13].

## 2 Methods

**Dataset** Our dataset comprises a subset of multispecies ncRNA interaction pairs from RNA-KG [14]. Due to the constraints imposed by the ncRNA Language Model (LM) in our pipeline (GenerRNA [10]), we filtered the dataset to retain only sequences that fit within the model’s token limit (approximately 4096 nucleotides). After applying these filters, the dataset contains: 99841 interaction pairs (down from an initial 130310 pairs), and 10644 unique sequences (selected from 19624 potential sequences) belonging to different RNA molecule types, including long non-coding RNA (lncRNA), circular RNA (circRNA), microRNA (miRNA), small nuclear RNA (snRNA), small nucleolar RNA (snoRNA), Small Cajal body-specific RNAs (scaRNAs), small cytoplasmic RNAs (scrRNA) and other types of ncRNAs.

**Data Augmentation** To address the issues due to the limited cardinality of the available training data, especially for specific types of ncRNA interactions (e.g., snoRNA-lncRNA or miRNA-circRNA), we employed a data augmentation strategy that effectively increases the dataset size by a factor of 4. For each original training instance represented as a pair of interacting ncRNA  $(s_i, s_j)$  we generate three additional augmented instances: 1) Molecule Order Reversal: swap the order of the molecules:  $(s_j, s_i)$ ; 2) Sequence Flipping: reverse the nucleotide order in both molecules (denoted by the superscript  $F$ ):  $(s_i^F, s_j^F)$ ; 3) Combined Augmentation: reverse both the molecule order and the nucleotide sequences  $(s_j^F, s_i^F)$ . This augmentation introduces invariance to both the order and orientation of sequences, thereby enabling the model to better capture the underlying biological patterns and improving its robustness against input variability.

**Generation of negative examples.** Since only positive non-coding RNA-RNA interactions are explicitly provided, we generated a set of negative examples  $\mathcal{N}$  that is  $n$  times larger than the positive examples ( $n = 20$  in our experiments) to address the imbalance problem in actual RNA interaction data. We corrupted a randomly sampled tuple  $(s_i, s_j)$  of each interacting pair by substituting its second element,  $s_j$ , with another randomly chosen ncRNA sequence,  $s_t$ , that has the same type of  $s_j$ . After checking that neither  $(s_i, s_t)$  or  $(s_t, s_i)$  are existing in the positive interaction set we add the tuple to the set of negative tuples; we generate a number of negatives per interaction pair that matches the relative frequency distribution of the positive interaction pair itself.

**Model Architecture** Our model follows a two-stage pipeline, as illustrated in Figure 1. It first extracts ncRNA sequence embeddings using a pre-trained ncRNA Language Model (GenerRNA [10]) and then processes these embeddings through a Feed-Forward Neural Network (FFNN) to predict interaction probabilities.

The GenerRNA architecture mimics the GPT-2-medium model [15], and is composed of 24 stacked transformer-decoder layers, each incorporating a self-attention mechanism - that models pairwise interactions among all positions in its input sequence. GenerRNA uses a context window of 1024 tokens, corresponding to input sequences with a length of approximately 4096 nucleotides coded through byte pair encoding.

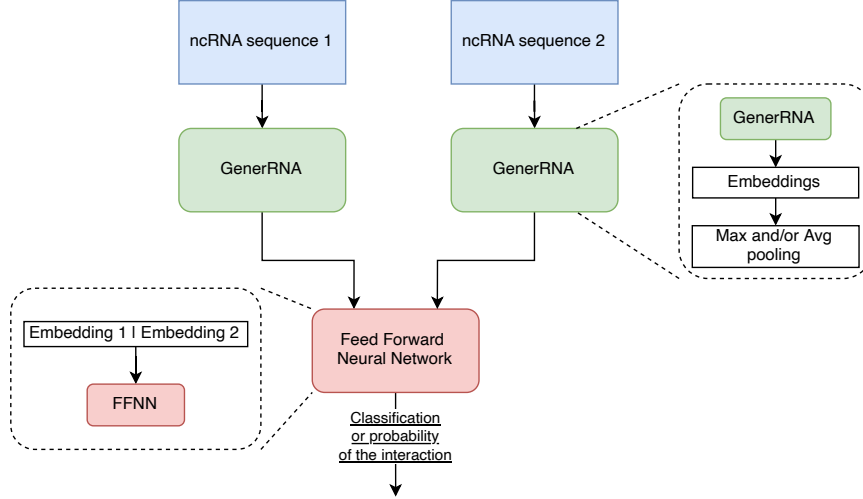


Figure 1: High-level *ncRNA-CUPID* architecture schema.

Each transformer block is fed with an input of size  $H \times I$ , where  $I = 1024$  is the length of the input token and  $H = 1024$  is the dimension of their embedded representations, and outputs a latent representation matrix  $\mathbf{X} \in \mathbb{R}^{H \times I}$ , whose rows  $\mathbf{x}_i \in \mathbb{R}^{1024}$  are latent representations of the  $i^{th}$  token.

To obtain a fixed-length embedding for the entire sequence, we tested two types of pooling over the sequence (i.e., across the  $H$  tokens), as well as their concatenation:

- **Average (AVG) Pooling:** obtained as the mean of the embeddings of all the tokens:  $\mathbf{e}_{\text{avg}} = \frac{1}{H} \sum_{i=1}^H \mathbf{x}_i$ .
- **Max Pooling:** Compute the element-wise maximum over all token embeddings:  $\mathbf{e}_{\text{max}} = \max_{i=1}^H \mathbf{x}_i$ .
- **Concatenation of [AVG, Max]:** Combine both pooled representations into a single embedding vector:  $\mathbf{e} = [\mathbf{e}_{\text{avg}}; \mathbf{e}_{\text{max}}] \in \mathbb{R}^{2048}$ .

To predict the interaction we concatenated the embedding pairs and used them as input to a FFNN.

Due to the imbalance in our dataset we designed a training strategy to prevent the model from learning predominantly from the negatives. To address this, we constructed mini-batches that contain a controlled mix of positive and negative examples. Each mini-batch  $B$  of size  $m$  is formed by randomly selecting  $m_p$  positive examples (using a uniform distribution with replacement) and  $m_n$  negative examples (using a uniform distribution without replacement). The ratio of negatives within each mini-batch is defined by  $r = \frac{m_n}{m_n + m_p}$ , with  $m_n + m_p = m$ . Here,  $r$  can vary between 0 and 1. A value of  $r = 0.5$  implies that 50% of the mini-batch consists of negatives. In our experiments, we set  $m = 512$  and  $r = 0.7$ . The choice of sampling positives with replacement is driven by their limited number, ensuring sufficient representation even in large batches, whereas sampling negatives without replacement allows for a broader coverage of these more abundant examples.

## 2.1 Experimental Evaluation

**Data preparation and model selection.** In all our experiments the negative examples were sampled according to the relative frequency of the interacting pair types with negative:positive ratio equal to 20:1. The dataset was partitioned into stratified training and test sets (train:test ratio = 90:10). The training set was further split into a stratified set for training (80% of interaction

pairs) and the remaining 20% for validation. The validation set was used for early stopping and for model selection via maximization of the Matthews correlation coefficient [16].

*Baseline for comparison.* Besides the random classifier—whose expected performance are AUROC = 0.5 (Area Under the Receiver Operating Characteristic Curve) and AUCPR = 0.047 (Area Under the Precision-Recall Curve)—we employed the IntaRNA method [2] as a baseline for comparison. IntaRNA estimates the interaction energy between RNA molecules. While interaction energy can be thresholded to obtain binary predictions, we opted to limit the comparison to threshold-independent metrics AUROC and AUCPR.

### 3 Results and discussion

Table 1: Comparison of AUROC and AUCPR across different pooling strategies and data augmentation techniques. Random baseline refer to the expected performance of the random classifiers. *ncRNA-CUPID* models are sorted in increasing order of both AUROC and AUCPR

Experiment	AUROC	AUCPR
Random baseline	0.5	0.047
Baseline-concat	0.658	0.078
Data-aug-Max	0.810	0.147
Data-aug-concat	0.862	0.222
Data-aug-AVG	<b>0.919</b>	<b>0.364</b>
IntaRNA	0.544	0.055

Table 1 summarizes the overall performance achieved by adopting different pooling strategies and data augmentation techniques with our proposed model:

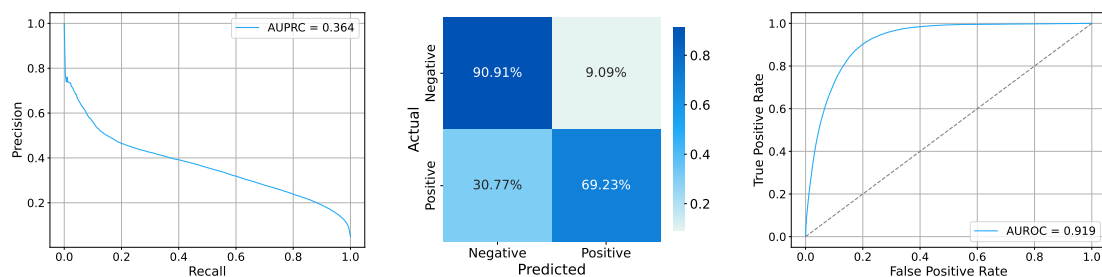
- (1) no data-augmentation and computation of the molecule embeddings by concatenating the AVG and Max pooling embedded representations (Baseline-concat setting in Table 1)
- (2) data-augmentation using three different embedding strategies: (a) Max pooling (Data-aug-Max), (b) Concatenation of AVG and Max pooling (Data-aug-concat) (c) Average pooling (Data-aug-AVG).

The best model is the one using data-augmentation and the average pooling for molecule embedding. This result was somehow expected. Data augmentation not only reduces the problems due to a limited sample set, but importantly improves model generalization with respect to the order of input molecules in the interacting pair, as well as to the order of the nucleotide sequences.

IntaRNA seems to fail on this task, but its poor results could be due to the fact that IntaRNA has been designed to detect interactions between small ncRNA and mRNA in bacteria, while our dataset includes a larger set of ncRNA interactions, involving also eukaryotic ncRNA.

Detailed *ncRNA-CUPID* results for each of the different types of ncRNA interactions are reported in Fig. 2.

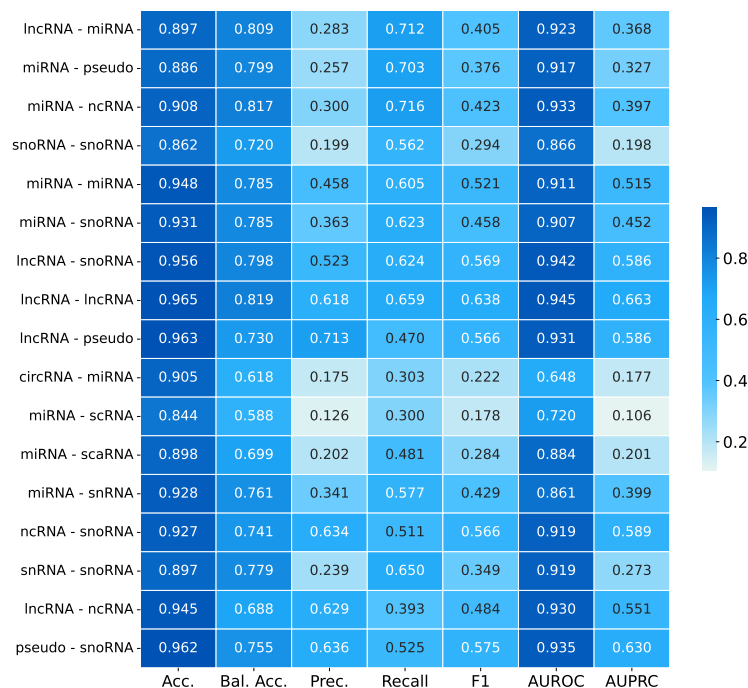
In conclusion, results show that *ncRNA-CUPID* can successfully predict ncRNA interactions, achieving for different types of ncRNA interactions AUROC larger than 0.9 (Fig. 2d). This approach opens the way for the large scale in silico prediction of ncRNA interactions using only their sequences. To our knowledge, this is the first transformer-based method to predict ncRNA interactions directly for sequence and the first computational method able to predict any type of ncRNA interactions. Future directions include evaluating other RNA-specific representations, such as those provided by RNABERT [17]; and broadening the performance comparison including deep learning methods for RNA interaction prediction.



(a) Test PRC curve

(b) Test confusion matrix

(c) Test ROC curve



(d) Test RNA Type Metrics Heatmap

Figure 2: *ncRNA-CUPID* results for Augmented train and test Classification Experiments with types-dependent sampling, and average pooling. (a) Precision Recall Curve (PRC) on the test set including all the type of ncRNA interactions; (b) Confusion matrix on the test set; (c) Receiver Operating Characteristic curve (ROC) on the test set including all the type of ncRNA interactions; (d) *ncRNA-CUPID* results on the test set across different types on ncRNA interactions (rows) for different types of metrics (columns).

### Conflict of interests

The authors have no competing interests to declare that are relevant to the content of this article.

### Acknowledgments

This work was supported by National Center for Gene Therapy and Drugs Based on RNA Technology—MUR (Project no. CN.00000041) funded by NextGeneration EU program, and by FAIR (Future Artificial Intelligence Research) project, funded by the NextGenerationEU program within the PNRR-PE-AI scheme (M4C2, Investment 1.3, Line on Artificial Intelligence). We thank Emanuele Cavalleri for help us in the preparation of the data.

### Availability of data and software code

The *ncRNA-CUPID* code, and the scripts to reproduce the experiments and tutorials are available from GitHub (<https://github.com/AnacletoLAB/ncRNA-CUPID>).

### References

- [1] Lucia Lorenzi, Hua-Sheng Chiu, Avila Cobos, et al. The RNA atlas expands the catalog of human non-coding RNAs. *Nature biotechnology*, 39(11):1453–1465, 2021.
- [2] Martin Mann, Patrick R Wright, and Rolf Backofen. IntaRNA 2.0: enhanced and customizable prediction of RNA–RNA interactions. *Nucleic acids research*, 45(W1):W435–W439, 2017.
- [3] Sinan Uğur Umu and Paul P Gardner. A comprehensive benchmark of RNA–RNA interaction prediction tools for all domains of life. *Bioinformatics*, 33(7):988–996, 2017.
- [4] Jesse M Engreitz et al. RNA-RNA interactions enable specific targeting of noncoding RNAs to nascent Pre-mRNAs and chromatin sites. *Cell*, 159(1):188–199, 2014.
- [5] Babak Alipanahi, Andrew Delong, Matthew T Weirauch, and Brendan J Frey. Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. *Nature biotechnology*, 33(8):831–838, 2015.
- [6] Xiongfei Tian, Ling Shen, Zhenwu Wang, Liqian Zhou, and Lihong Peng. A novel lncRNA–protein interaction prediction method based on deep forest with cascade forest structure. *Scientific reports*, 11(1):18881, 2021.
- [7] Si-Lin Sun, Yue-Yi Jiang, Jun-Ping Yang, Yu-Han Xiu, Anas Bilal, and Hai-Xia Long. Predicting noncoding RNA and disease associations using multigraph contrastive learning. *Scientific Reports*, 15(1):230, 2025.
- [8] F. Torgano et al. RNA Knowledge-Graph analysis through homogeneous embedding methods. *Bioinformatics Advances*, 5, 2025.
- [9] G Valentini et al. The promises of large language models for protein design and modeling. *Frontiers in Bioinformatics*, 3, 2023.
- [10] Yichong Zhao, Kenta Oono, Hiroki Takizawa, and Masaaki Kotera. GenerRNA: A generative pre-trained language model for de novo RNA design. *PLoS One*, 19(10):e0310814, 2024.
- [11] Tao Shen et al. Accurate RNA 3D structure prediction using a language model-based deep learning approach. *Nature Methods*, 21:2287–2298, 2024.
- [12] M Nicolini, E Saitto, Jimenez-Franco RE, E Cavalleri, Galeano A, D Malchiodi, A Paccanaro, PN Robinson, E Casiraghi, and G Valentini. Fine-tuning of conditional Transformers improves in silico enzyme prediction and generation. *Computational and Structural Biotechnology Journal*, 27:1318–1334, 2025.
- [13] Muller Fabbri, Leonard Girnita, Gabriele Varani, and George A Calin. Decrypting noncoding RNA interactions, structures, and functional networks. *Genome research*, 29(9):1377–1388, 2019.
- [14] E. Cavalleri et al. An ontology-based knowledge graph for representing interactions involving RNA molecules. *Scientific Data*, 11(1):906, 2024.
- [15] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- [16] Brian W Matthews. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451, 1975.
- [17] Manato Akiyama and Yasubumi Sakakibara. Informative RNA base embedding for RNA structural alignment and clustering by deep representation learning. *NAR genomics and bioinformatics*, 4(1), 2022.