RNA Knowledge Graph Analysis via Embedding Methods

FRANCESCO TORGANO, EMANUELE CAVALLERI, JESSICA GLIOZZO,
FEDERICO STACCHIETTI, EMANUELE SAITTO, MARCO MESITI,
ELENA CASIRAGHI, GIORGIO VALENTINI
AnacletoLab, Dipartimento di Informatica, Università degli Studi di Milano
Via Celoria 18, Milano
ITALY

Abstract: Recent advances in RNA technologies opened the avenue to the design of novel vaccines as witnessed by the success of the COVID-19 vaccine and also by new ongoing vaccines for cancer. New drugs based on non-coding RNA can also be developed at lower costs considering the relatively simple structure of these molecules with respect to classical recombinant protein technologies. We recently developed RNA-KG, a biomedical Knowledge Graph focused on RNA, collecting information from more than 50 public databases and bio-medical ontologies to support the study of RNA and the design of novel RNA-based drugs. In this work we show that, by applying inductive machine learning methods on top of embedded node and edges obtained by applying classical Graph Representation Learning methods, we can accurately predict the entities and the relationships between entities included in RNA-KG. Our results open the way to the analysis and the discovery of novel relationships between RNAs and other bio-molecules and medical concepts represented in RNA-KG.

Key-Words: Artificial Intelligence methods for graph analysis, Graph Representation Learning, Knowledge Graphs, RNA.

Received: January 26, 2024. Revised: August 3, 2024. Accepted: September 4, 2024. Published: October 3, 2024.

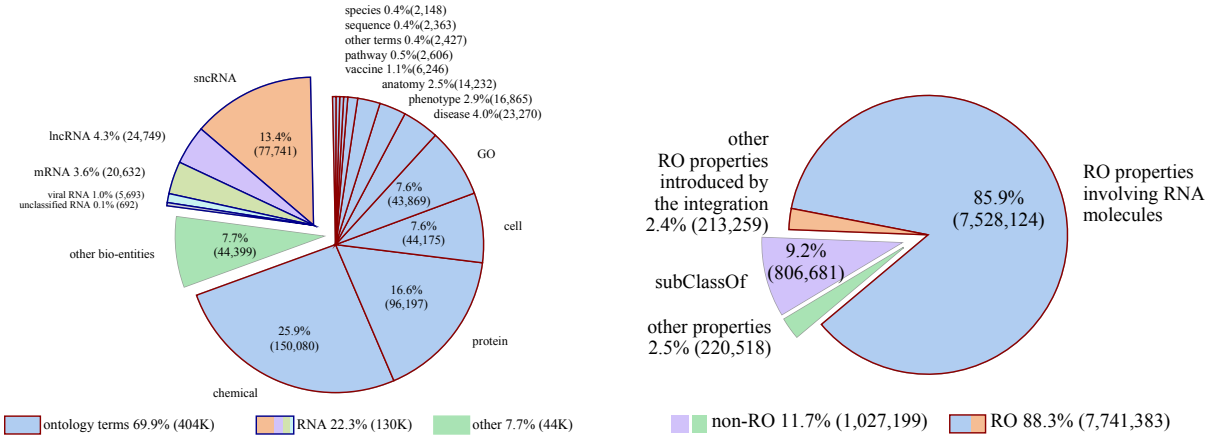
1 Introduction

RNA-based technologies introduced novel therapeutics for the treatment and prevention of human diseases, [1]. Indeed RNA molecules play a fundamental role in cell biology, performing a wide range of functions either directly by regulating gene expression, exhibiting enzymatic activity, modifying and regulating other RNAs and other bio-molecules, or indirectly by being translated into proteins. Different types of RNA are involved in regulatory processes: small non-coding RNAs (sncRNAs) are associated with RNA interference pathways, including short interfering RNAs (siRNAs), microRNAs (miRNAs), short hairpin RNAs (shRNAs), antisense oligonucleotides (ASOs), piwi-interacting RNAs (piRNAs), tRNA-derived fragments (tRFs), and tRNA-derived small RNAs (tsRNAs). sncRNAs modulate mRNA expression by inhibiting translation or facilitating the degradation of the target transcript via complementary base pairing. Long non-coding RNAs (lncRNAs) hold crucial importance in the onset and advancement of diseases, [2], and are involved in competitive endogenous RNA (ceRNA) regulation, transcriptional and epigenetic regulation, [3].

More in general several studies revealed the functional characteristics of a large variety of RNA molecules, [4], [5], thus opening the door for the design of mRNA-based vaccines for the COVID-19 pandemic, [6], for the treatment of melanom,a [7], and for the development of new drugs that can target both proteins and mRNA, as well as other non-coding RNA, [8].

Recently we proposed RNA-KG, the first ontology-based knowledge graph (KG) for representing coding and non-coding RNA molecules and their interactions with other bio-molecules as well as with pathways, abnormal phenotypes, and diseases to support the study and the discovery of the biological role of the "RNA-world", [9]. RNA-KG represents relationships between bio-molecules and bio-medical concepts through Resource Description Framework (RDF) triples extracted from more than 50 public data sources and also integrates related bio-medical concepts coded through biomedical ontologies including the Human Phenotype Ontology, [10], the Monarch Merged Disease Ontology, [11], Chemical Entities of Biological Interests, [12], and other fundamental biomedical ontologies, [9].

RNA-KG exploits PheKnowLator, [13], a software system for the construction of semantically rich, large-scale biomedical KGs that are Semantic Web compliant and amenable to automatic OWL reasoning. The current version of RNA-KG includes about 600K nodes and 9M of edges and can be exported in different data formats. RNA-KG has been designed not only to represent information related to RNA in a



(a) Distribution of nodes. (b) Distribution of edges.

Fig. 1: Pie-chart of: (a) node distribution according to node types. (b) edge distribution according to edge types

relational graph format but also to provide a KG ready to be analyzed through graph-based AI methods for inferring new knowledge about the "RNA world" and supporting the discovery of new RNA-based drugs.

In this paper we show that Graph Representation Learning (GRL) methods, [14], [15], can be applied to the analysis of RNA-KG to both visually represent and to classify the different types of nodes and edges that characterize the heterogeneous graph. We model the entity and relation predictions in RNA-KG as a multi-class classification problem, using graph embedding methods to transform nodes and edges into their vector representation and applying inductive machine learning methods to classify the nodes and edges of the KG.

2 The RNA Knowledge Graph (RNA-KG)

RNA-KG is the first KG that aggregates biological knowledge about RNAs from over 50 public databases, integrating functional relationships between genes, proteins, chemicals, and ontologically grounded biomedical concepts. The current release of RNA-KG [16] has a single component with approximately 600K nodes and 9M edges, and it can be queried via a SPARQL endpoint from the laboratory website [17]. The nodes are typically mapped to reference biomedical vocabularies and ontologies, such as NCBI Gene Entrez identifiers for unique identification of genes and various types of non-coding RNAs (ncRNAs), the Human Phenotype Ontology for phenotypes, the Monarch merged disease ontology for diseases, and the Gene Ontology, [18], for annotating genes. Furthermore, all possible interactions are represented using the Relation Ontology (RO, [19]), which ensures consistent semantics for the different relationships extracted from the sources.

Fig 1a (adapted from [9]) illustrates the distribution of nodes within RNA-KG. Nodes are divided into those representing ontology terms and bio-entities without a direct mapping. The bio-entities category is further split into RNA nodes (which includes sncRNA, mRNA, lncRNA, viral RNA, and unclassified RNA nodes), and non-RNA nodes (termed `other bio-entities`), including for example gene and variant (SNP) nodes. Fig 1b displays the distribution of edges in RNA-KG. Edges are sorted into three groups: (i) edges representing RO terms that denote interactions among RNA molecules from various sources, (ii) edges representing the `subClassOf` relationships, and (iii) edges representing other types of relationships not covered by RO. The `subClassOf` relationship arises from the integration of bio-ontologies into RNA-KG, along with the absence of a dedicated ontology for RNA molecules. When RNA molecules cannot be precisely mapped to a reference ontology, they are classified as `subClassOf` an appropriate category within the Sequence Ontology, [20], such as `SO_0000276` for miRNA molecules.

3 RNA-KG embedding

We constructed embedded representation of nodes and edges of RNA-KG to assess whether their vector representations can be used to visualize the resulting graph in an Euclidean space and to predict the node and edge types of the graph. In particular, we applied node2vec, [21], and LINE, [22], embedding methods.

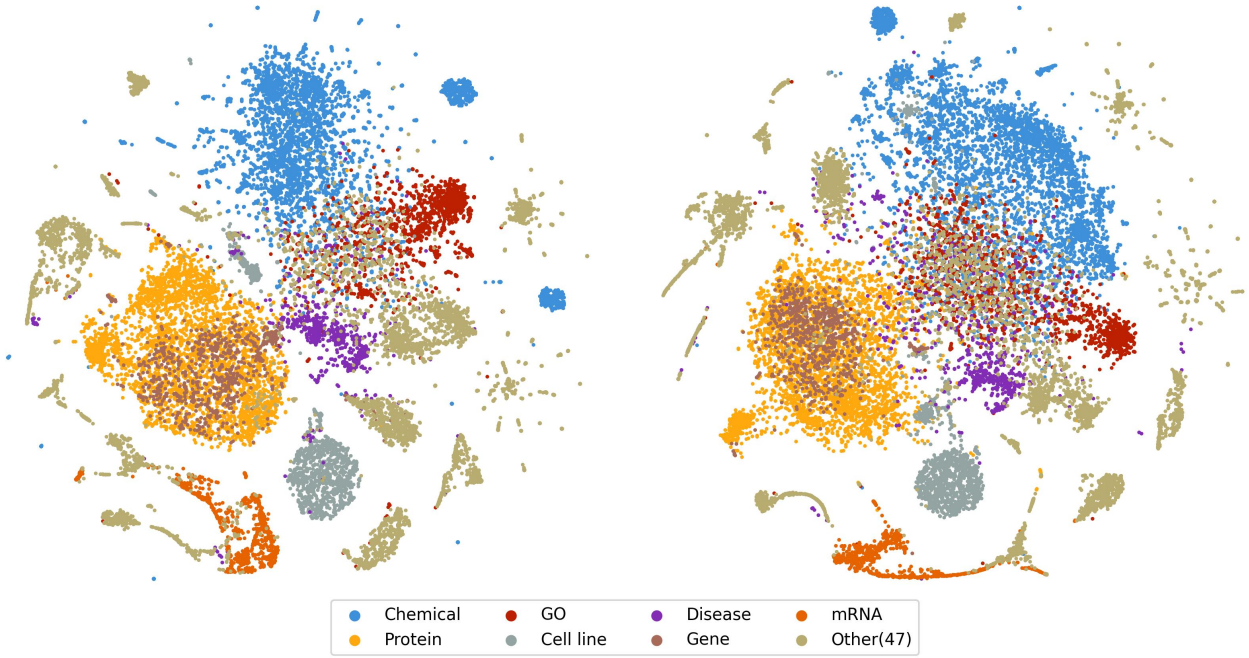


Fig. 2: t-SNE projections of RNA-KG embedding of the main node types generated by LINE. (Left) First order LINE; (right) Second order LINE

3.1 Node2vec and LINE embedding

Node2vec uses random walks (RWs) to obtain sequences of nodes to "linearize" the graph and then applies a shallow neural network to obtain a vectorial representation of the nodes and edges, using an approach similar to word2vec to embed text, [23]. More precisely, the node2vec second order RW is defined by a transition probability of the form:

$$\pi_{rvx} = \alpha_{p,q}(x, v, r) \cdot w_{v,x}. \tag{1}$$

where π_{rvx} is the probability of moving from node v to node x coming from node r. In eq. 1 the term $w_{v,x}$ denotes the weight of the edge $(v, x) \in E$, and $\alpha_{p,q}$ are node2vec parameters defined as:

$$\alpha_{p,q} = \begin{cases} \frac{1}{p} & \text{if } d_{r,x} = 0 \\ 1 & \text{if } d_{r,x} = 1 \\ \frac{1}{q} & \text{if } d_{r,x} = 2 \end{cases}$$

where $d_{r,x}$ denotes the graph distance between the nodes r and x, whereby $d \in \{0, 1, 2\}$. By tuning the parameters p and q we can both leverage the homophily (through a Depth-First Sampling (DFS)-like visit) and the structural (through a Breadth-First Sampling (BFS)-like visit) characteristics of the input graph, thus obtaining embeddings that can capture the topological features of the graph.

LINE (Large-scale Information Network Embedding) provides embeddings that scale nicely with big graphs, [22]. The proposed model optimizes an objective function that preserves the first and second order proximity of nodes in the embedded spaces. First-order proximity is defined as the local pairwise proximity between two vertices, indicated by the weight of the edge connecting them. Second-order proximity is instead defined as the similarity between the neighborhoods of the two vertices. The first-order LINE model optimizes an objective function that considers the following function:

$$O_1 = \sum_{i,j \in E} KL(\hat{p}_1(v_i, v_j), p_1(v_i, v_j))$$

where KL is the Kullback-Leibler divergence between the joint probability $p_1(v_i, v_j) = \frac{1}{1+\exp(-u_i^T \cdot u_j)}$ and the empirical probability $\hat{p}_1(v_i, v_j) = \frac{w_{ij}}{W}$, with v_i and $v_j \in V$ and u_i, u_j are their corresponding embedding in a vector space, and w_{ij} is the weight of the edge (v_i, v_j), with $W = \sum_{(i,j) \in E} w_{ij}$.

Similarly second-order LINE minimizes the KL divergence between the second-order proximity empirical distribution and the second-order proximity distribution in the embedded space, [22].

3.2 Embedded representations of RNA-KG

We generated 100-dimensional representations of the nodes through the LINE algorithm. Fig. 2 shows the t-SNE, [24], two-dimensional projections of the nodes

obtained by LINE, while Fig. 3 shows the same embeddings obtained with node2vec. In both figures the main seven most numerous node types are shown. We can observe that all the embeddings methods can reasonably separate the different node types. However node2vec can neatly separate the node types, especially with the second-order Depth-First-like (DFS: $p = 5, q = 0.2$) and Breadth-First-like (BFS: $p = 0.2, q = 5$) RWs; in particular while node2vec is able to separate genes (brown points) from proteins (orange points), this is not the case with LINE embeddings (Fig. 2 and Fig. 3).

A reasonable separation has been also achieved with respect to edge types, even if for both LINE and node2vec the separation of the different edge types is not so clearly defined as for node types (Fig. 4).

4 Classification of embedded nodes and edges

The embedded representations of nodes and edges are used to train different learning machines for the prediction of the node and the edge types of RNA-KG.

4.1 Experimental set-up

We applied the following models to classify the nodes and edges of RNA-KG: Decision tree classifier, [25], random forest ensembles, [26], a linear perceptron classifier, a support vector machine classifier with Gaussian kernel and a multi-layer perceptron (MLP) classifier with one hidden layer.

We applied multiple hold-out (70% training and 30% test) on 2-dimensional t-SNE projections, [24], of the 100-dimensional embedding of randomly sampled nodes (20K from about 600K) and edges (10K from about 9M) of RNA-KG. We also evaluated different multi-class classification tasks, considering for node type prediction the classification of the 7 most represented classes of nodes, and also the multi-class classification of the 20 and 54 most represented node types. For edge type classification we considered respectively the $7, 15, 40$ and 74 most represented edge types.

We performed limited model selection with decision trees (tuning of the maximum depth of the tree) and random forest (tuning of the number of base learners) and no model selection at all for the Perceptron, MLP (one hidden layer with 100 neurons, ReLU activation function, ADAM for weight optimization and maximum number of iterations set to 500), and a gaussian SVM with regularization parameter $C = 1$ and maximum number of iterations set to 200. All the models have been implemented using the scikit-learn Python library and the embeddings have been computed using the GRAPE library, [27].

4.2 Classification of node and edge types of the RNA-KG

Fig. 5 summarizes RNA-KG node and edge type predictions across the different classification tasks and the different models using the 2-dimensional t-SNE projections of the BFS-like node2vec embeddings.

Node type classification, decision trees, random forests and SVMs achieved a balanced accuracy across the 7 most represented classes larger than 90% and also with 20 classes we obtained a balanced accuracy close or larger than 90%. With 54 node types, a reasonable accuracy of about 50% is obtained (consider that a random balanced accuracy with 54 classes would be about 2%). The linear perceptron obtained worse results, since classes are surely non linearly separable (see Fig. 3).

Reasonable, but significantly worse results are obtained for edge type classification (Fig. 5, bottom). With random forests (the best performing method) we obtained a balanced accuracy of about 75% with 7 classes but performances decrease with other models or, as expected, when the number of classes is higher.

Fig. 6 shows the effect of model parameters in decision trees (Fig. 6 a) and b) and random forests (Fig. 6 c).

Summarizing, results show that edge and especially node types of RNA-KG are predictable using also simple prediction models (e.g., decision trees) trained on top of the node and edge node2vec embeddings.

4.3 Prediction of the overall nodes and edges of RNA-KG

Previous results were obtained on a random sample of 20K nodes of RNA-KG. Here we present the results obtained on the analysis of larger numbers of nodes till to the overall about 600K nodes of RNA-KG.

Table 1 and Table 2 report the results of the decision trees and random forest trained on different number of classes and including the original sample of 20K nodes but also a larger random sample of 100K till to the overall 600K nodes of RNA-KG. Also in this case we performed a multiple hold-out (repeated 5 times) by splitting the available data with 70% of training and 30% test set.

5 Discussion

We predicted node and edge types of RNA-KG, using relatively simple embedding and classification methods. Embeddings of nodes and edges visualized through t-SNE projections show that the different types of nodes and edges can be separated in the euclidean space, even if edge types show a less clear separation (Fig. 2, Fig.3, Fig. 4).

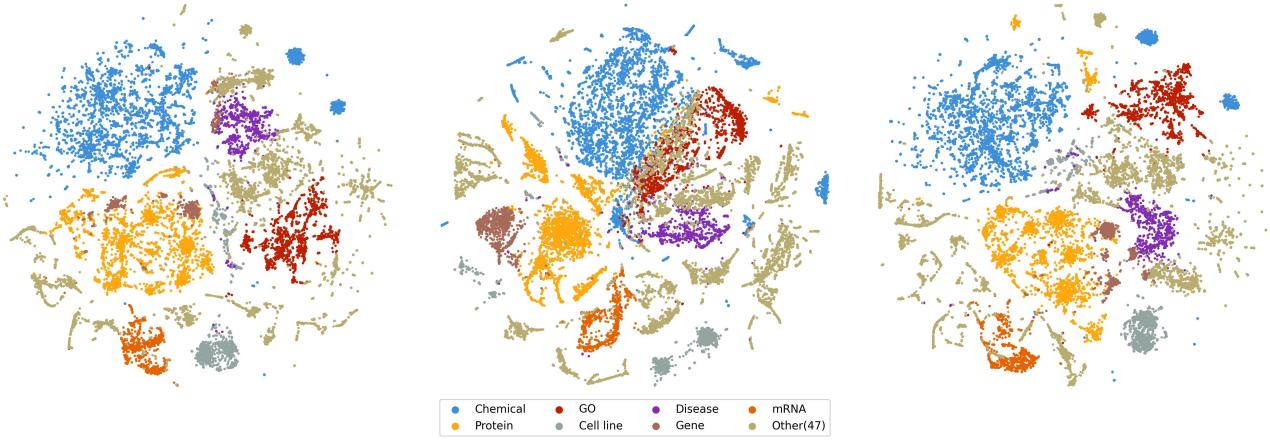


Fig. 3: t-SNE projections of RNA-KG embedding of the main node types using node2vec according to three different graph visiting strategies: a) DFS-like RW b) BFS-like RW c) First order RW

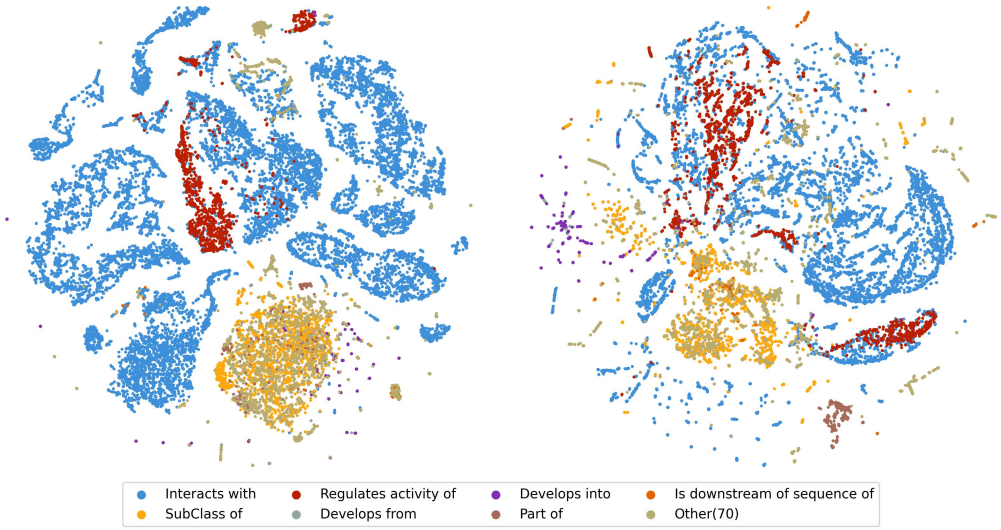


Fig. 4: Embedding of the main edge types of RNA-KG using second-order LINE (left) and DeepWalk (right)

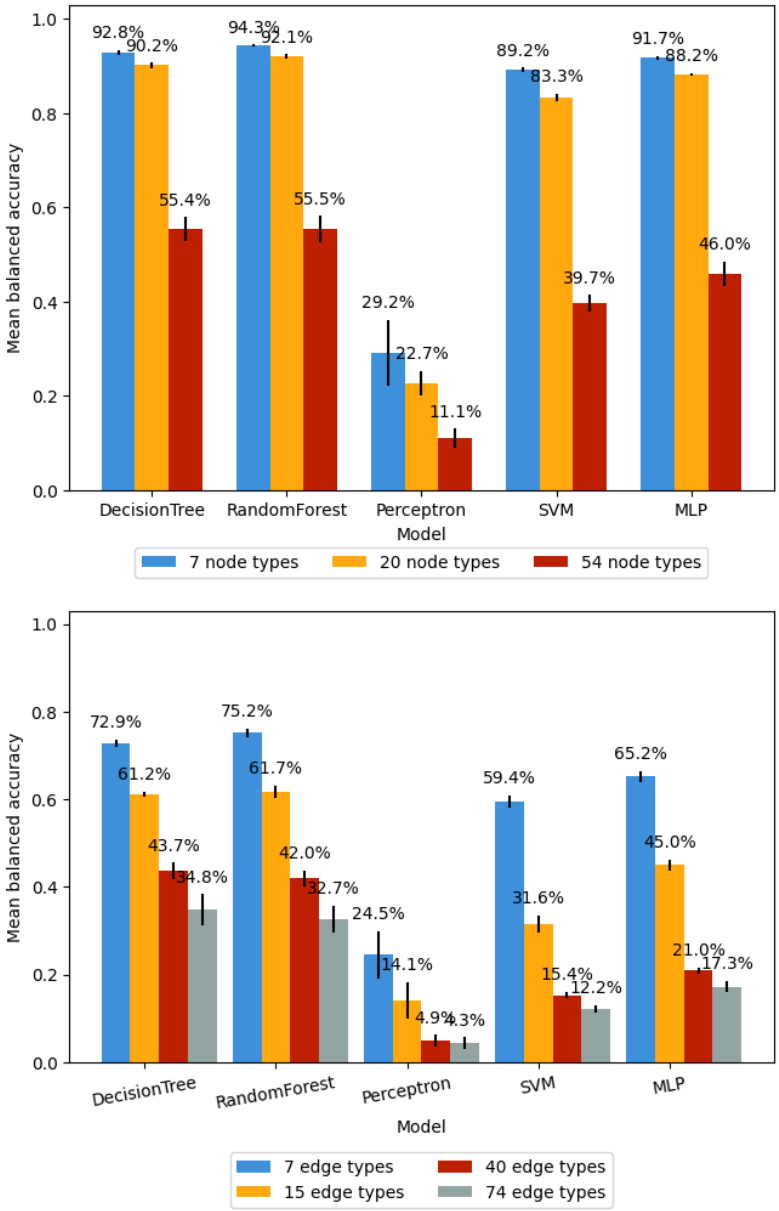


Fig. 5: RNA-KG node and edge type classification. (top) Comparison of balanced accuracy results between models on the most represented 7, 20, 54 node types; (bottom) Comparison of balanced accuracy results between models on the most represented 7, 14, 40, 71 edge types. Vertical lines on top of the bars represent the standard deviation

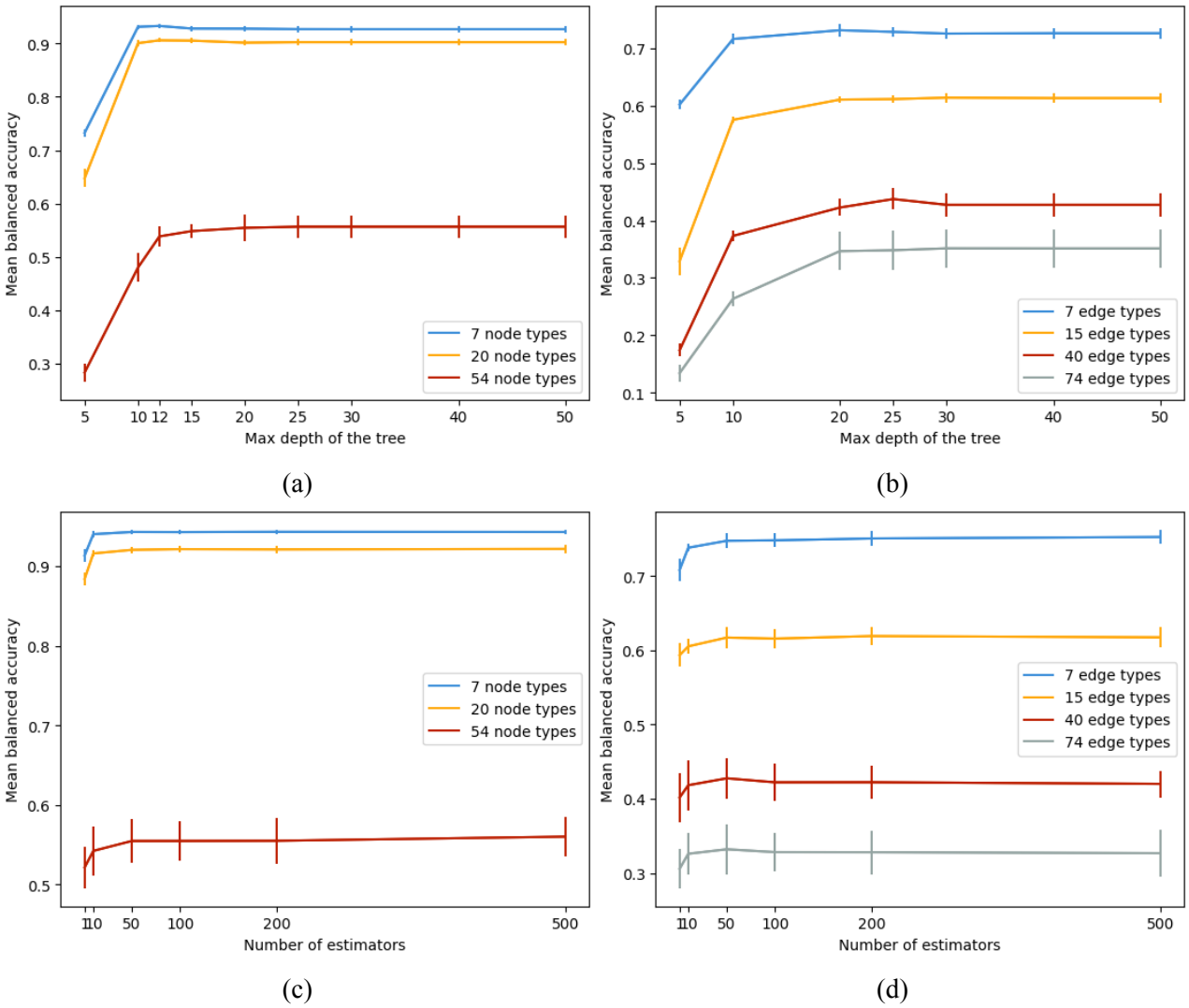


Fig. 6: Effect of the learning parameters on decision tree and random forest balanced accuracy performance. (a) Decision tree node type prediction results with respect to the maximum tree depth; (b) Decision tree edge type prediction results with respect to the maximum tree depth; (c) Random forest node type prediction results with respect to the number of base learners. (d) Random forest edge type prediction results with respect to the number of base learners

These results are also confirmed by the classification performance obtained by machine learning models trained on top of the embeddings. All the models, except the linear perceptron, achieve a reasonable balanced accuracy (Fig. 5) for all the classification tasks.

The results on edge embeddings (Fig. 5, bottom), even if significantly better than random guessing, are worse with respect to node type prediction since current edge type definitions in RNA-KG are too general to functionally characterize the distinct types of edges of RNA KG. For instance general relationships, such as "interacts with" or "regulates activity of" can involve different types of nodes, e.g. genes or proteins or miRNA and mRNA, and refers to different functional relationships all "pushed" into the same type of edge.

Classification results on the overall nodes of RNA-KG show a certain decrement in the performance of decision trees and random forest classifiers, even if with 7 classes we obtained a balanced accuracy larger than 80% with both decision trees and random forests (Table 1 and Table 2). This may be due to the reduced input dimension and also to the fact that we did not perform a thorough model selection, or to a possible larger presence of outliers.

Summarizing this preliminary analysis shows that embeddings methods coupled with off-the-shelf classification methods can obtain good results on the pre-

Table 1. Decision tree node type prediction results on RNA-KG. Balanced accuracy and empirical computational time are reported considering different numbers of classes and nodes of the graph.

#cl.	#nodes	balanced_acc	time
7	20 000	92.81% ± 0.57%	0.45
7	100 000	92.34% ± 0.27%	2.81
7	578 384	81.84% ± 0.14%	12.93
20	20 000	90.18% ± 0.58%	0.51
20	100 000	89.73% ± 0.22%	3.22
20	578 384	77.19% ± 0.06%	14.75
54	20 000	55.43% ± 2.50%	0.67
54	578 384	42.14% ± 0.41%	19.00
68	100 000	52.77% ± 1.66%	4.37
68	578 384	33.39% ± 0.39%	20.90
81	578 384	32.36% ± 0.53%	22.72

Table 2. Random forest node type prediction results on RNA-KG. Balanced accuracy and empirical computational time are reported considering different numbers of classes and nodes of the graph. Est. represents the number of estimators (base learners).

#cl.	#nodes	#est.	balanced_acc	time
7	20 000	50	94.32% ± 0.22%	6.77
7	100 000	200	94.11% ± 0.23%	136.48
7	578 384	500	83.72% ± 0.11%	194.26
20	20 000	50	92.07% ± 0.43%	7.92
20	100 000	200	91.74% ± 0.44%	133.67
20	578 384	500	80.29% ± 0.19%	248.77
54	20 000	50	55.45% ± 2.78%	10.47
54	578 384	500	42.12% ± 0.44%	369.31
68	100 000	200	52.28% ± 2.89%	169.01
68	578 384	500	33.59% ± 0.33%	412.86
81	578 384	500	32.20% ± 0.67%	452.37

dictions of node and edge types of RNA-KG. This is quite surprising since we used relatively simple prediction methods without performing an accurate model selection. We foresee that better results could be obtained through fine-tuned models (some preliminary results seem to confirm this hypothesis, data not shown).

Moreover we used only bi-dimensional t-SNE projections of the embeddings to train the classifiers, but by using full embeddings we could in principle further improve the classification performances.

We also observe that for the embeddings we used methods conceived for homogeneous graph embeddings. By applying graph embeddings for heterogeneous graphs, [28], we could further improve the results since RNA-KG is a heterogeneous graph composed of a large number of different nodes and edge types.

Our results show that GRL methods can be successfully applied to the analysis of RNA-KG with accurate predictions. These findings open the door to more refined analyses of RNA-KG to detect novel edges to support the investigation of the "RNA world" and the discovery of new RNA-based drugs.

For future research we foresee that the application of graph embedding methods aware of the heterogeneity of the RNA-KG, [29], [30], can significantly improve the predictions of novel relationships between non coding RNAs and other target molecules, as well as associations between ncRNA with abnormal phenotypes and diseases. Another futuew research direction is represented by the application of graph neural networks that can solve the edge prediction problem with a direct end-to-end approach, [31], [32].

References:

[1] Sparmann, Anke and Vogel, Jörg. Rna-based medicine: from molecular mechanisms to therapy. *The EMBO Journal*, 42(21):e114760, 2023.

[2] John S. Mattick, Paulo P. Amaral, Piero Carninci, Susan Carpenter, Howard Y. Chang, Ling-Ling Chen, Runsheng Chen, Caroline Dean, Marcel E. Dinger, Katherine A. Fitzgerald, Thomas R. Gingeras, Mitchell Guttman, Tetsuro Hirose, Maite Huarte, Rory Johnson, Chandrasekhar Kanduri, Philipp Kapranov, Jeanne B. Lawrence, Jeannie T. Lee, Joshua T. Mendell, Timothy R. Mercer, Kathryn J. Moore, Shinichi Nakagawa, John L. Rinn, David L. Spector, Igor Ulitsky, Yue Wan, Jeremy E. Wilusz, and Mian Wu. Long non-coding rnas: definitions, functions, challenges and recommenda-

tions. *Nature Reviews Molecular Cell Biology*, 24(6):430–447, January 2023.

[3] Lin Liu, Zhao Li, Chang Liu, Dong Zou, Qianpeng Li, Changrui Feng, Wei Jing, Sicheng Luo, Zhang Zhang, and Lina Ma. LncRNAWiki 2.0: a knowledgebase of human long noncoding RNAs with enhanced curation model and database system. *Nucleic Acids Research*, 50(D1):D190–D195, 2022.

[4] Lucia Lorenzi, Hua-Sheng Chiu, Francisco Avila Cobos, Stephen Gross, Pieter-Jan Volders, Robrecht Cannoodt, Justine Nuytens, Katrien Vanderheyden, Jasper Anckaert, Steve Lefever, et al. The rna atlas expands the catalog of human non-coding rnas. *Nature biotechnology*, 39(11):1453–1465, 2021.

[5] Andreas Keller, Laura Gröger, Thomas Tschernig, Jeffrey Solomon, Omar Laham, Nicholas Schaum, Viktoria Wagner, Fabian Kern, Georges Pierre Schmartz, Yongping Li, et al. mirnatissueatlas2: an update to the human mirna tissue atlas. *Nucleic acids research*, 50(D1):D211–D221, 2022.

[6] Ann J. Barbier, Allen Yujie Jiang, Peng Zhang, Richard Wooster, and Daniel G. Anderson. The clinical progress of mrna vaccines and immunotherapies. *Nature Biotechnology*, 40(6):840–854, May 2022.

[7] Thiago Carvalho. Personalized anti-cancer vaccine combining mrna and immunotherapy tested in melanoma trial. *Nature Medicine*, 29(10):2379–2380, August 2023.

[8] Melanie Winkle, Sherien M. El-Daly, Muller Fabbri, and George A. Calin. Noncoding rna therapeutics — challenges and potential solutions. *Nature Reviews Drug Discovery*, 20(8):629–651, June 2021.

[9] Cavalleri, E and Cabri, A and Soto-Gomez, M and Bonfitto, S and Perlasca, P and Gliozzo, J and Callahan, T and Reese, J and Robinson, P and Casiraghi, E and Valentini, G and Mesiti, M. Rna-kg: An ontology-based knowledge graph for representing interactions involving rna molecules. *Scientific Data, Nature Publishing*, (in press), 2024.

[10] Peter N. Robinson, Sebastian Köhler, Sebastian Bauer, Dominik Seelow, Denise Horn, and Stefan Mundlos. The human phenotype ontology: A tool for annotating and analyzing human hereditary disease. *The American Journal of Human Genetics*, 83(5):610–615, November 2008.

[11] Lynn M Schriml, James B Munro, Mike Schor, Dustin Olley, Carrie McCracken, Victor Felix, J Allen Baron, Rebecca Jackson, Susan M Bello, Cynthia Bearer, Richard Lichenstein, Katharine Bisordi, Nicole Campion Dialo, Michelle Giglio, and Carol Greene. The human disease ontology 2022 update. *Nucleic Acids Research*, 50(D1):D1255–D1261, November 2021.

[12] K. Degtyarenko, P. de Matos, M. Ennis, J. Hastings, M. Zbinden, A. McNaught, R. Alcantara, M. Darsow, M. Guedj, and M. Ashburner. Chebi: a database and ontology for chemical entities of biological interest. *Nucleic Acids Research*, 36(Database):D344–D350, December 2007.

[13] Tiffany J. Callahan, Ignacio J. Tripodi, Adrianne L. Stefanski, Luca Cappelletti, Sanya B. Taneja, Jordan M. Wyrwa, Elena Casiraghi, Nicolas A. Matentzoglu, Justin Reese, Jonathan C. Silverstein, Charles Tapley Hoyt, Richard D. Boyce, Scott A. Malec, Deepak R. Unni, Marcin P. Joachimiak, Peter N. Robinson, Christopher J. Mungall, Emanuele Cavalleri, Tommaso Fontana, Giorgio Valentini, Marco Mesiti, Lucas A. Gillenwater, Brook Santangelo, Nicole A. Vasilevsky, Robert Hoehndorf, Tellen D. Bennett, Patrick B. Ryan, George Hripcsak, Michael G. Kahn, Michael Bada, William A. Baumgartner, and Lawrence E. Hunter. An open source knowledge graph ecosystem for the life sciences. *Scientific Data*, 11(1), April 2024.

[14] M.M. Li, K. Huang, and M. Zitnik. Graph representation learning in biomedicine and healthcare. *Nat. Biomed. Eng.*, 6:1353–1369, 2022.

[15] Luca Cappelletti, Lauren Rekerle, Tommaso Fontana, Peter Hansen, Elena Casiraghi, Vida Ravanmehr, Christopher J Mungall, Jeremy J Yang, Leonard Spranger, Guy Karlebach, J Harry Caufield, Leigh Carmody, Ben Coleman, Tudor I Oprea, Justin Reese, Giorgio Valentini, and Peter N Robinson. Node-degree aware edge sampling mitigates inflated classification performance in biomedical random walk-based graph representation learning. *Bioinformatics Advances*, 4(1):vbae036, 03 2024.

[16] Emanuele Cavalleri et al. RNA-KG: data and experiments code. Available at: `https://doi.org/10.5281/zenodo.10418431`. Accessed: 14 March 2024.

[17] RNA-KG website. Available at: http://RNA-KG.anacleto.di.unimi.it. Accessed: 22 December 2023.

[18] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, May 2000.

[19] Chris Mungall, Nico Matentzoglu, Jim Balhoff, David Osumi-Sutherland, Bill Duncan, pgaudet, Shawn Tan, Charles Tapley Hoyt, Clare Pilgrim, James A. Overton, Lauren, Anita Caron, Nomi Harris, Sierra Moxon, lschriml, Nicole Vasilevsky, Sabrina Toro, Damien Goutte-Gattat, Matthew Brush, Vasundra Touré, Anthony Bretaudeau, Scott Cain, Melissa Haendel, diatomsRcool, Bide Zhang, Clint Dowland, Damion Dooley, actions user, and Jen Hammock. oborel/obo-relations: 2023-08-18 release. Available at https://doi.org/10.5281/zenodo.8263469, August 2023.

[20] Karen Eilbeck, Suzanna E Lewis, Christopher J Mungall, Mark Yandell, Lincoln Stein, Richard Durbin, and Michael Ashburner. The sequence ontology: a tool for the unification of genome annotations. *Genome Biology*, 6(5), April 2005.

[21] Aditya Grover and Jure Leskovec. Node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 855–864, New York, NY, USA, 2016. Association for Computing Machinery.

[22] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. Line: Large-scale information network embedding. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, page 1067–1077, Republic and Canton of Geneva, CHE, 2015. International World Wide Web Conferences Steering Committee.

[23] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 3111–3119, Red Hook, NY, USA, 2013. Curran Associates Inc.

[24] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

[25] L. Breiman, Jerome H. Friedman, Richard A. Olshen, and C. J. Stone. Classification and regression trees. *Biometrics*, 40:874, 1984.

[26] L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[27] L. Cappelletti, T. Fontana, E. Casiraghi, V. Ravanmehr, T.J. Callahan, C. Cano, M.P. Joachimiak, C.J. Mungall, P.N. Robinson, J. Reese, and G. Valentini. Grape for fast and scalable graph processing and random walk-based embedding. *Nature Computational Science*, 3:552–568, 2023.

[28] Y. Xie, B. Yu, S. Lv, C. Zhang, G. Wang, and M. Gong. A survey on heterogeneous network representation learning. *Pattern Recognition*, 116(107936), 2021.

[29] Ayush Noori, Michelle M Li, Amelia LM Tan, and Marinka Zitnik. Metapaths: similarity search in heterogeneous knowledge graphs via meta-paths. *Bioinformatics*, 39(5):btad297, 2023.

[30] Dengju Yao, Yuexiao Deng, Xiaojuan Zhan, and Xiaorong Zhan. Predicting lncrna-disease associations using multiple metapaths in hierarchical graph attention networks. *BMC Bioinformatics*, 25(1), January 2024.

[31] I. Chami, S. Abu-El-Haija, B. Perozzi, C. Ré, and K. Murphy. Machine Learning on Graphs: A Model and Comprehensive Taxonomy. *Journal of Machine Learning Research*, 23(89):1–64, 2022.

[32] Yixuan Liang and Yuan Wan. Learning on heterogeneous graph neural networks with consistency-based augmentation. *Applied Intelligence*, 53(22):27624–27636, 2023.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

Francesco Torgano implemented the code.
Emanuele Cavalleri curated the graphs used in the experiments.

Francesco Torgano, Jessica Gliozzo, Federico Stacchietti and Emanuele Saitto performed the experiments, results visualization and evaluation.
Giorgio Valentini conceptualized the work, drafted the original paper, and acquired fundings.
Marco Mesiti, Elena Casiraghi and Giorgio Valentini validated the results and supervised the work.
All the authors revised, read and approved the final manuscript.

Funding sources

This research was supported by the National Center for Gene Therapy and Drugs based on RNA Technology, PNRR-NextGenerationEU program (G43C22001320007).

Conflicts of Interest

The authors have no conflicts of interest to declare that are relevant to the content of this article.