Inductive models for structured output prediction of lncRNA-disease associations

Felipe Kenji Nakano*† Livia Bertoni‡, Ricardo Cerri§, Celine Vens‡§,
* KU Leuven, Campus KULAK, Dept. of Public Health and Primary Care, Kortrijk, Belgium
† Itec, imec research group at KU Leuven, Kortrijk, Belgium
‡ Departamento de Genética e Evolução, Universidade Federal de São Carlos, São Carlos, SP, Brazil
§ Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, SP, Brazil
Email: felipekenji.nakano@kuleuven.be, liviabertoni@estudante.ufscar.br, cerri@icmc.usp.br and celine.vens@kuleuven.be

***Abstract*—Long non-coding RNAs have gained significant attention due to their crucial roles in the pathogenesis of complex human diseases, such as neurological diseases, cardiovascular diseases, AIDS, diabetes, and various types of cancer. In the machine learning literature, lncRNA-disease association (LDA) has been widely investigated as a binary classification problem, where each lncRNA-disease pair is seen as an independent instance. This approach presents drawbacks as it does not exploit the correlation among the diseases, aggravates the already imbalanced dataset, and substantially increases the execution time. Furthermore, the literature focuses on the transductive setting where new disease associations are predicted in lncRNAs already seen by the model, which naturally restricts its application to already seen lncRNAs. As a solution, we propose to address LDA prediction as a structured output prediction problem, namely (hierarchical) multi-label classification, where all LDAs are predicted at once for a given lncRNA. We compared several LDA methods and their structured output variants with recent (hierarchical) multi-label classification methods in an inductive setting, e.g., disease associations are predicted in unseen lncRNAs. Our experiments reveal that approaching LDA prediction with structured output prediction leads to superior or competitive results while drastically reducing the running time.**

I. Introduction

Long non-coding RNAs (lncRNAs) are RNA sequences longer than 200 nucleotides that are not translated into functional proteins [1]. lncRNAs were first identified in eukaryotes and have structural features similar to those of messenger RNAs, such as the CAP structure and polyA tail [2].

Increasing evidence shows that lncRNAs play an essential role in many biological processes, such as cell cycle regulation [3], epigenetic regulation [4], cell differentiation control [5] and post-transcriptional regulation [6]. ,

Furthermore, LncRNAs have gained significant attention due to their crucial roles in the pathogenesis of complex human diseases. Mutations and dysregulations of these RNAs are associated with the development and progression of multiple conditions, such as neurological diseases [7], cardiovascular diseases [8], AIDS [9], and various types of cancer [10]. Despite that, most lncRNAs still remain experimentally untouched. Consequently, their disease associations are unknown [11].

Thus, accurate identification of lncRNA-disease associations (LDAs) holds significant medical potential, enhancing the comprehension of the molecular mechanisms underlying human diseases [12]. These studies may contribute to the discovery of lncRNAs as potential biomarkers for the diagnosis, treatment, and prognosis of diseases, as well as potential therapeutic targets for drug development [13].

In this direction, LDA has been widely addressed using machine learning [14], since they provide a viable, rapid, and robust approach, enabling large-scale and simultaneous data analysis. Initial methods focused on employing relatively simple models, such as regularized least squares [15], whereas more recent works focus on tree ensembles [16] and deep learning-based methods [17], [18].

Regardless of the underlying model, to the best of our knowledge, all methods approach LDA as a binary classification problem. That is, features are extracted for both lncRNAs and diseases, and their Cartesian product is used to generate new instances. For example, given an lncRNA l_1 and a set of diseases $D = \{d_1, d_2, ..., d_d\}$, the features of l_1 are concatenated with d_1, generating a new instance. The same process is repeated using l_1 and d_2 until all lncRNAs and diseases are combined. Naturally, the label of each generated instance consists of the corresponding association given in the LDA matrix, e.g., 1 if their association is experimentally confirmed or 0 otherwise.

Despite being the default approach, the binary classification strategy faces some drawbacks, such as class imbalance, as the LDA matrix is already sparse itself; disregard for disease correlations, as each disease is treated as a separate outcome; and high computational complexity, as the binary transformation generates a substantial number of instances.

Furthermore, to the best of our knowledge, the literature currently focuses on the transductive learning scenario where new disease associations are predicted in lncRNAs known to the model [14]. More specifically, existing studies use the values of the LDA matrix to generate features, making them non-applicable to newly discovered lncRNAs, which can pose a major limitation since new lncRNAs are constantly being discovered [11].

As a solution, we propose approaching LDA as an inductive learning problem where features are extracted considering only the lncRNA sequences themselves, allowing the prediction of LDA in unseen lncRNAs.

Precisely, we approach LDA using structured output pre-

979-8-3503-5663-2/24/$31.00 ©2024 IEEE

diction where multiple outputs must be predicted. Given a lncRNA with unknown associations, each disease is seen as an output, and the model predicts all LDAs concomitantly, thus exploiting the label correlations and addressing it as a multi-label classification problem [19], [20]. Furthermore, we also propose to expand the output space using Medical Subject Headings (MeSH) by incorporating the associations between general and more specific diseases [21]. In this case, we transform LDA prediction into a hierarchical multi-label classification [22], a predictive task where the correlation among the outputs is pre-established by a hierarchy.

We evaluated our proposed approaches by comparing recent LDA methods [23], [24] with their (hierarchical) multi-label adaptations. We also added methods originally proposed for structured output prediction in our comparison [25], [26]. Our experiments revealed that approaching LDA as a structured output prediction problem is not only superior to the binary approach currently used, but it is also more efficient in terms of computational complexity.

The remainder of this paper is structured as follows: Section II provides a review of related work. Section III presents our proposed approach, namely using structured output prediction for LDA. Further, Section IV presents the datasets, compared methods, and evaluation setup. Section V contains the results and discussion. Lastly, Section VI covers the conclusions and future research directions.

II. Related Work

In this section, we review recent LDA prediction and (hierarchical) multi-label methods.

A. lncRNA-disease association prediction

As discussed in [14], the literature presents works ranging from methods based on rather simpler models, such as regularized least squares [15], support vector machines [27], naive bayes [28], to more complex methods that use deep neural networks [17], [18], [29] and tree ensembles [16], [23], [24], [30]. It is also noticeable that the choice of features varies significantly, where most of the works rely on the LDA matrix to extract features for both lncRNAs and diseases.

Random forest and feature selection based LDA prediction model (RFLDA) proposes to first perform feature selection using the feature importance of the random forest itself [23]. Further, it builds a random forest using its optimal features. The authors used Wang's distance [31] for the diseases features and functional similarity for the lncRNAs, which is calculated using their disease associations [32].

As its name suggests, the incremental principal component analysis random forest (IPCARF) preprocesses its input features using incremental PCA [24]. As features, the authors used Gaussian interaction profile kernel similarity for both lncRNAs and diseases [33]. Wang's distance was also used to obtain more disease features [31].

The work of Wu et al. [30] employs graph convolutional networks to extract features based on functional and Gaussian similarities, which are further used as input to randomized trees [34]. Similarly to the previous works, Cheng's functional similarity was used as lncRNAs features, whereas Wang's distance [31] was used as disease features. Both input spaces were complemented using the Gaussian kernel similarity [33]. The main difference lies in the fact that the authors employed KNN to first identify the most similar lncRNAs, before feeding them to the graph convolutional neural networks (GCN).

Peng et al. [17] incorporated SVD features generated using the LDA matrix. That is, the U and Vt matrices produced by the SVD decomposition are used as lncRNAs and disease features, respectively. These features are combined with embeddings obtained from a GCN constructed using functional and Gaussian kernel similarities [32], [33]. Lastly, a heterogeneous Newton boosting machine is used as the binary classifier. In a complementary work, the same authors proposed to employ non-matrix factorization instead of the SVD and a dual deep neural network as its binary classifier [17]. Also employing SVD, Zhou et al. [16] proposed LDA-SABC, a method that combines AdaBoostCNN and lightGBM using only the SVD features as input. The final prediction is obtained by the weighted average of both models.

Using only deep learning, the work of Li et al. [18] combined both functional and Gaussian similarities as input to a similarity structural network and a node adaptive feature smoothing to extract features, which are further used as input to a transformer-based classifier. Lastly, Gao et al. [29] used the same features as input to a matrix factorization method combined with GCNs, using CatBoost as the predictive model.

B. Structured output prediction

Structured output prediction is an umbrella term used to refer to prediction problems where multiple outputs must be predicted [35]–[37]. Since LDA consists of multiple binary outputs, in this work we investigate further multi-label and hierarchical multi-label classification.

1) Multi-label classification: As defined in [20], [36], multi-label classification is a structured output prediction problem whose objective consists of building a model capable of predicting multiple binary outputs. For instance, an lncRNA may be associated with Alzheimer's disease, Parkinson's disease, and cancer.

As surveyed in [20], a wide range of methods have been proposed, such as KNN [38] and shallow neural networks [39], to more complex deep neural networks [35]. Despite their relative simplicity, tree ensembles in general are mostly associated with superior results, in the context of structured data where features are presented in a tabular form [20].

Recent methods proposed deep forest for multi-label classification. Similar to deep learning, deep forest methods consist of employing multiple sequential layers of tree ensembles where each layer augments the feature vector [26]. The original deep forest, gcForest [26], proposed to use prediction probabilities from the previous layers as extra features.

As an improvement, in earlier work, we proposed to use tree-embeddings, which are features generated based on the structure of the decision trees [40]. More specifically, each

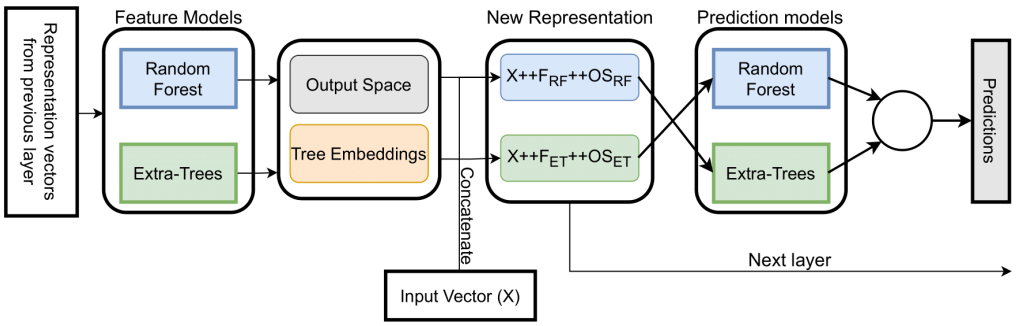

Fig. 1. An example of a layer in deep forests [40]. The box labeled 'Output Space (OS)' represents the prediction probabilities features, whereas the orange box contains the tree-embeddings. The 'Next Layer' arrow connects the new representations of this layer to the input box of the subsequent layer, which shares the same structure as the one illustrated.

decision tree is represented as a binary vector that receives the value 1 if it is traversed, or 0 otherwise. The resulting binary vectors of each tree are concatenated and used as input to a PCA, thus generating new features. By combining tree-embeddings with prediction probabilities, the authors proposed two variants: Cascade Forest Embedders (CAFE), which includes only the tree-embeddings, and Cascade Forest Embedders with Output Space features (CAFE-OS), the combination of tree-embeddings and prediction probabilities. These methods are depicted in Fig. 1.

2) Hierarchical multi-label classification: Similarly to multi-label, hierarchical multi-label classification also investigates problems with multiple outputs. Its main difference lies in the presence of a hierarchy of outputs that determines underlying correlations among the labels. For instance, the output "Bladder Cancer" is a subclass of "Cancer". Thus, "Cancer" must be predicted whenever "Bladder cancer" is predicted; this is referred to as the hierarchy constraint [22].

The first HMC method was introduced by Vens et al. [22] where the authors proposed decision trees that incorporate the hierarchy in their heuristic. Since then, the focus has shifted to neural networks, where loss functions that take the hierarchy into consideration are used. Adjacency Wrapping Matrix (AWX) proposes to replicate the hierarchy as the architecture of the network and propagate gradients from the leaves to the roots [41]. Alternatively, Cerri et al. proposed to split the hierarchy into levels and build a neural network for each of them [42]. Wehrmann et al. [43] extended both of these networks by also incorporating a third component related to the deepest known output of the instances. Lastly, Giunchiglia and Lukasiewicz [44] proposed coherent neural networks that strictly enforce the hierarchy constraint into the loss function.

III. Structured Output Prediction for LDA

First, we present the binary classification approach in more detail. Further, we introduce our proposed approaches: multi-label and hierarchical multi-label classification for LDA.

A. Binary classification approach

Given a LDA matrix in the shape $|L| \times |D|$ where $|L|$ corresponds to the number of lncRNAs, $|D|$ to the number of diseases, where the value 1 represents that such association was validated and 0 otherwise, the binary approach adopts the transformation procedure presented in Fig. 2.

First, features are extracted for both lncRNAs (XL) and diseases (XD). Following, new instances are created using the Cartesian product of the lncRNAs and the diseases, where the features of each lncRNA are separately concatenated with XD. For instance, the features related to l_1 are concatenated with XD, creating L_1D_1. The same procedure is repeated to generate $\{L_1D_2, L_1D_3, ..., L_nD_n\}$. Lastly, their corresponding binary output is given by the LDA itself.

Thus, an LDA matrix of dimensions $|L| \times |D|$ results in a dataset with $|L| \times |D|$ instances where all LDAs are seen as a single output.

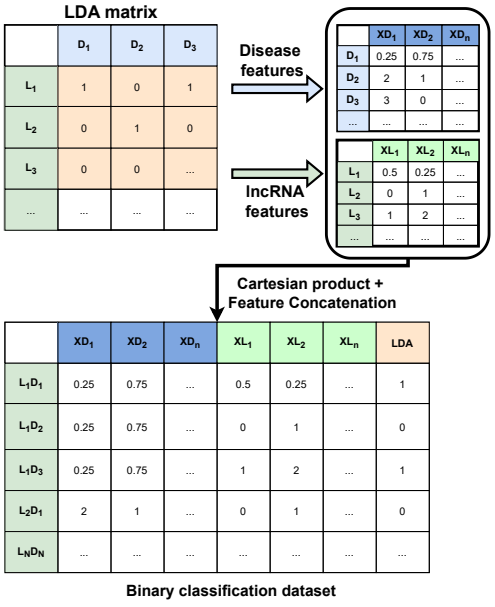

Fig. 2. Transforming LDA prediction in a binary classification problem.

B. LDA as a multi-label classification

We propose to treat each disease as an individual label in a multi-label classification task. That is, only lncRNA features are extracted (XL), which are further concatenated

with their corresponding disease associations from the LDA matrix, resulting in a multi-label dataset. This procedure is depicted in Fig. 3.

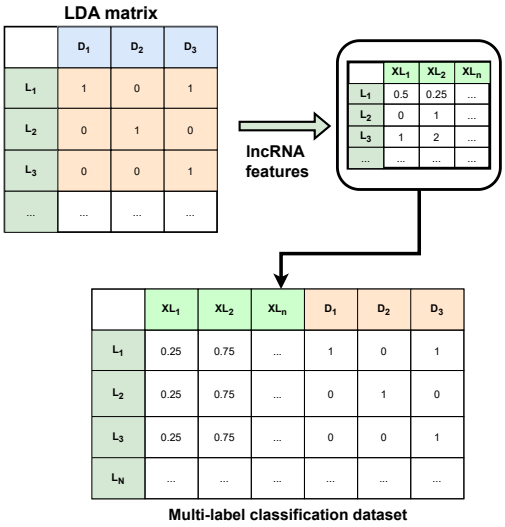

Fig. 3. Approaching LDA prediction as a multi-label classification problem.

In this case, the correlations among the diseases are exploited by a multi-label classification method, since they are presented as separate outcomes. For instance, if a lncRNA is associated with cancer and myocardial infarction, the binary classification considers them to be the same output. As opposed to that, multi-label methods can exploit such correlation during their construction [20].

C. LDA as a hierarchical multi-label classification

We propose the procedure shown in Fig. 4. Namely, we expand the LDA matrix by mapping the diseases to a predefined hierarchy of diseases, e.g., the MeSH hierarchy. More specifically, for every disease in the LDA matrix, we locate its node in the hierarchy and add all parent diseases in its path to the root. As exemplified in Fig. 4, L_1 is associated with D_1 and D_3, whose superclasses are D_5 and D_6, respectively. Thus, D_5 and D_6 are added as positive associations to L_1. The same applies to L_2, resulting in its disease associations being expanded from D_2 to $\{D_2, D_4, D_5\}$.

Although specific methods have been proposed that explicitly model the hierarchical relations, as mentioned in Section II, here we take a simple and more generalizable approach that passes the resulting dataset as input to a multi-label classification method. After obtaining its predictions, it is necessary to convert them back to the original disease space, e.g., $\{D_1, D_2, D_3, D_4, D_5, D_6\}$ to $\{D_1, D_2, D_3\}$. In this case, we simply opted to remove the prediction probabilities related to the added diseases.

IV. Materials and Methods

In this section, we present our materials and methods, which are publicly available[1].

[1]https://github.com/fkenjinakano/lncRNA_CIBCB2025

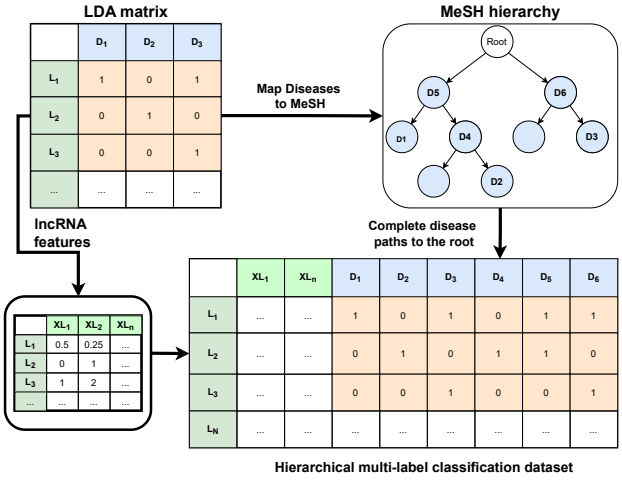

Fig. 4. Approaching LDA prediction as a hierarchical multi-label classification problem.

A. Datasets

1) lncRNA-disease association data: The LncRNADisease database, established in [45], was the first manually curated and experimentally verified LDA repository. In this study, we used the v2017 data from the LncRNADisease database. Furthermore, we kept only the diseases that were associated with more than five lncRNAs. Table I provides a summary of this dataset.

2) Medical Subject Headings data: The MeSH (Medical Subject Headings) provides a standardized framework for indexing and categorizing biomedical literature [21]. It organizes descriptors into 16 categories, each with hierarchical subcategories. We focused on category C (diseases), mapping the original labels to the MeSH taxonomy, following the procedure specified in Section III. Table I presents a summary of the hierarchical dataset.

TABLE I
SUMMARY OF THE DATASETS EMPLOYED.

	LncRNADisease	LncRNADisease + MeSH
# LncRNAs	355	355
# LncRNA feature	827	827
# Diseases	44	369
# Disease features	44	369
# Positive interactions	1132 / 10,175 (5%)	9,539 / 130,995 (7%)

B. Features

As detailed in Section II, most of the works employ a combination of functional and Gaussian kernel similarities to represent the lncRNAs, which rely on the LDA matrix. That is, previously known LDA associations are included in the test set, which are not applicable in the inductive setting.

Hence, we extracted eight RNA feature vectors using the package Pse-in-One [46]: K-mers, Mismatch, dinucleotide-based auto covariance (DAC), dinucleotide-based cross covariance (DAC), dinucleotide-based auto-cross covariance

(DACC), general parallel correlation pseudo dinucleotide composition (PC-PseDNC-General), and general series correlation pseudo dinucleotide composition (SCPseDNC-General). In total, it sums to 827 features, as shown in Table I.

K-mers consist of the number of occurrences of a nucleotide subsequence of size K, e.g., if K = 2, K-mers quantify the occurrences of {AA, AC, AG, AU,..., UU}. DAC, DACC, PC-PseDNC-general and SCPseDNC-General are features related to the (pseudo)-dinucleotide composition and physicochemical properties of the RNAs.

As for the diseases, we employed the widely used Wang's distance [31]. Naturally, the number of disease features corresponds to the number of diseases in the dataset (Table I). Mind that disease features are only used by binary methods.

C. Evaluation Measures

As suggested in the literature [14], we employed precision, recall, f-score, and accuracy, which are described in Equations 1, 2, 3, and 4. Furthermore, we also used (hierarchical) multi-label evaluation measures [19], [20], [25]: hamming loss, label ranking error, micro AUROC, and micro AUPRC.

The hamming loss and label ranking error are provided in Equations 5 and 6 whereas the micro AUROC is calculated using the area under the curve where recall (Equation 2) is in the x-axis and the false positive rate (Equation 10) is in the y-axis. Similarly, the micro AUPRC is measured using the area under the curve where the micro precision (Equation 8) and micro recall (Equation 7) are represented in the x and y-axis, respectively. Table II presents the symbols used to describe the evaluation measures.

TABLE II
SYMBOLS USED TO DESCRIBE THE EVALUATION MEASURES.

Y	LDA matrix
X	LncRNA test set
Z	Predicted LDA
TP	True positive
FP	False positive
FN	False negative
N	Number of lncRNAs ($\|X\|$)
M	Number of diseases in LDA ($\|Y\|$)
θ	Ranking function
$\hat{Y}_i$	Complement of Y for an instance i
$\oplus$	Boolean XOR
T	Threshold values

$$Pr = \frac{TP}{TP + FP} \tag{1}$$

$$Re = \frac{TP}{TP + FN} \tag{2}$$

$$F1 - score = \frac{2 * Pr * Re}{Pr + Re} \tag{3}$$

$$Accuracy = \frac{TP + FN}{TP + FP + TN + FN} \tag{4}$$

The hamming loss (Equation 5) quantifies the average number of incorrectly predicted labels.

$$HammingLoss = \sum_{i}^{N} Z_i \oplus Y_i \tag{5}$$

$$RankingError = \frac{1}{N} \frac{1}{|Y_i||\hat{Y}_i|} |(y_1, y_2) : \theta_i(y_1) > \theta_i(y_2), (y_1, y_2) \in Y_i \times \hat{Y}_i| \tag{6}$$

In addition, the average precision score will be used. Given the imbalance in the label set, we are using its micro variant as described in Equations 7-9.

$$MicroRe = \sum_{j}^{M} \frac{TP_j}{TP_j + FN_j} \tag{7}$$

$$MicroPr = \sum_{j}^{M} \frac{TP_j}{TP_j + FP_j} \tag{8}$$

$$PrecisionScore = \sum_{t}^{T} \frac{(MicRe_t - MicroRe_{t-1})}{MiPr_t} \tag{9}$$

$$FPR = \frac{FP}{FP + TP} \tag{10}$$

Due to high class imbalance, a default threshold of 50% is not recommended. Hence, the Youden index was used to find the optimal thresholds [47].

D. Validation setup and statistical test

We employed 10-fold cross-validation. A fold consists of selecting 90% of the dataset as training data and the remaining 10% as the test set. This procedure is repeated 10 times, using different training and test sets. Results are reported considering the average and the standard deviation over the 10 folds.

To attest statistical significance, we used the pairwise Wilcoxon signed-rank test with correction for multiple comparisons to compare all the methods and their corresponding structured output prediction variants.

E. Comparison methods

- RFLDA: A random forest-based method that selects the most relevant features according to its feature importance [23];
- IPCARF: A random forest-based method that employs an incremental PCA to construct new features [24];
- LDA-SABC: A binary method that employs an Adaboost-CNN and a lightGBM [16];
- GcForest: the original deep forest method proposed in [26], which uses prediction probabilities as extra features;
- CAFE: A deep forest method that uses tree-embeddings as extra features [25];
- CAFE-OS: A deep forest method that uses tree-embeddings and prediction probabilities as extra features [25];

- RF + ET: A method that employs a random forest and randomized trees. It is equivalent to a single layer of the deep forest methods.

We included RFLDA and IPCARF, since they can be easily adapted to (hierarchical) multi-label classification problems. In structured output prediction, outputs are normally represented by a binary vector where each position corresponds to an output [22], [48]. For instance, given a set of diseases $D = \{D_1, D_2, D_3\}$, and a lncRNA, which is associated to D_1,D_3, its binary vector would consist of $\{1, 0, 1\}$. This conversion is applied to the associations of all lncRNAs present in the dataset, and the heuristic is averaged over each output.

Furthermore, they are often used as comparison methods, as seen in [16], [17]. LDA-SABC is included as a stronger competitor, since it was recently proposed [16]. However, it does not support structured output prediction. Thus, we only compared its binary version.

As for multi-label classification methods, we included the deep forest methods gcForest [26], CAFE [40], and CAFE-OS [40], as they are often associated with state-of-the-art results.

Regarding hierarchical multi-label classification, we simply extended the deep forest methods. Despite the popularity of deep learning methods, tree-ensemble method still yield competitive or superior results in this context [49].

We employed 150 trees for all tree ensemble-based methods. For gcForest and CAFE, the depth was set to three layers. The number of components in PCA used by CAFE was set to 85%. For LDA-SABC, the number of estimators in AdaBoostCNN was set to 150, the number of epochs and batch size to 10, both learning rates in lightGBM and AdaBoostCNN to 0.1, and α to 0.4, as recommended by the authors [16].

The remaining features were optimized using 5-fold inner cross-validation, following the hyperparameters suggested by the authors. That is, the number of optimal features in RFLDA was selected using steps of 50 features, i.e. $\{50, 100, 150, ... 827\}$ [23]. The number of components in IPCARF was optimized considering the values: $\{2, 4, 8, 16, 32, 64, 128\}$ [24].

V. Results and Discussion

In this section, we present our results. Initially, we report the predictive performance of the compared methods, followed by their running time.

A. Predictive performance

As can be seen in Table III, our proposed methodology, approaching LDA as a structured output problem, leads to superior results in most cases. More specifically, using the multi-label approach very frequently improves on the binary approach, whereas the hierarchical multi-label approach is better than the multi-label counterpart, in several cases.

IPCARF (HMC) leads to better results in all measures, except for hamming loss, precision, and accuracy. Precisely, IPCARF (HMC) achieved 0.24, 0.83, and 0.44 in ranking error, micro AUROC, and microAUPRC, respectively. Regarding recall and f-score, IPCARF (HMC) yielded 78% and 48%. A very similar performance was observed for IPCARF (MLC). This performance improvement was further highlighted by the Wilcoxon test, which detected statistical difference between their binary and structured output prediction is most cases.

As opposed to that, the binary methods from the LDA prediction literature performed systematically worse than their (hierarchical) multi-label variants. In fact, LDA-SABC, a recently proposed method, presented rather underwhelming performance, except for accuracy, where it was consistently outperformed by considerably simpler methods, such as IPCARF and RFLDA.

Furthermore, the deep forest methods, gcForest, CAFE, and CAFE-OS, did not manage to overcome RFLDA and IPCARF, apart from CAFE-OS (B), which achieved the lowest hamming loss (0.06), the highest precision (42%) and the highest accuracy (20%), with statistical significance over its structured output prediction. This is rather surprising since they were originally proposed to address structured output prediction problems. Moreover, the best-performing method, IPCARF (HMC), can be seen as a simpler version of CAFE, since IPCARF employs incremental PCA to construct new features. This finding highlights the capacities of methods specifically validated in LDA, despite their rather simple nature.

The main difference between the deep forest-based methods and IPCARF lies in the fact that CAFE uses multiple layers of tree ensembles to process the features, which is seemingly less beneficial in LDA. A possible explanation lies in the sparseness seen in LDA datasets (Table I), where these methods have originally been validated in less sparse datasets [25], [49].

Further, LDA datasets present only the experimentally confirmed associations, meaning that negative associations can be seen as unconfirmed rather than non-interactive, e.g., 0s in the LDA matrix could be changed to 1s in future versions of the dataset. This possibly has a more pronounced negative effect on more complex models, resulting in overfitting, whereas methods IPCARF and RFLDA are able to generalize better.

These results attest to our hypothesis that incorporating label dependencies by using structured output prediction can indeed improve performance. Furthermore, it also highlights the fact that the current literature may struggle in inductive scenarios as considerably superior results are reported when the associations of the test instances are partially known [16], [17], [23], [24], [28], [30].

B. Running time

Table IV contains the average running time of our experiments. As can be seen, multi-label classification methods are the fastest, followed by hierarchical multi-label and binary classification methods. Experiments were performed on a server equipped with an Intel Xeon Gold 6326 processor with 252GB of RAM.

Since the binary transformation approach increases the number of instances (Section III), binary methods present substantially longer running times. Specifically, CAFE-OS, the slowest binary method, may take approximately 48 and

TABLE III

AVERAGE AND STANDARD DEVIATION PREDICTIVE PERFORMANCE OF ALL METHODS CONSIDERING 10-FOLD CROSS-VALIDATION. HERE, B STANDS FOR BINARY, MLC FOR MULTI-LABEL CLASSIFICATION, AND HMC FOR HIERARCHICAL MULTI-LABEL CLASSIFICATION. THE SYMBOL ASTERISK (∗) REPRESENTS STATISTICAL DIFFERENCE BETWEEN A STRUCTURED OUTPUT PREDICTION METHOD AND ITS BINARY VARIANT.

Method	Hamming Loss	Ranking Error	micro AUROC	micro AUPRC	Precision	Recall	F-score	Accuracy
CAFE (B)	0.08 (0.03)	0.57 (0.06)	0.65 (0.08)	0.21 (0.14)	0.34 (0.12)	0.52 (0.12)	0.40 (0.11)	0.11 (0.06)
CAFE (MLC)	0.09 (0.02)	0.27 (0.06)*	0.81 (0.06)*	0.42 (0.14)*	0.33 (0.10)	0.77 (0.10)*	0.45 (0.10)*	0.12 (0.09)
CAFE (HMC)	0.09 (0.03)	0.27 (0.05)*	0.81 (0.06)*	0.40 (0.15)*	0.34 (0.10)	0.77 (0.09)*	0.47 (0.10)*	0.15 (0.09)
CAFE-OS (B)	**0.06 (0.02)**	0.51 (0.06)	0.69 (0.09)	0.29 (0.15)	**0.42 (0.13)**	0.54 (0.14)	0.46 (0.12)	**0.20 (0.07)**
CAFE-OS (MLC)	0.09 (0.03)*	0.27 (0.06)*	0.81 (0.07)*	0.37 (0.14)*	0.33 (0.10)*	0.77 (0.10)*	0.45 (0.10)*	0.07 (0.08)*
CAFE-OS (HMC)	0.11 (0.03)*	0.31 (0.06)*	0.79 (0.07)*	0.38 (0.16)*	0.28 (0.09)*	0.74 (0.11)*	0.40 (0.10)*	0.07 (0.08)*
gcForest (B)	0.12 (0.02)	0.37 (0.06)	0.77 (0.08)	0.28 (0.12)	0.26 (0.07)	0.73 (0.12)	0.38 (0.08)	0.15 (0.05)
gcForest (MLC)	0.11 (0.03)	0.30 (0.06)*	0.79 (0.07)*	0.37 (0.15)*	0.29 (0.09)	0.75 (0.12)	0.41 (0.10)	0.06 (0.10)*
gcForest (HMC)	0.12 (0.03)	0.35 (0.06)	0.73 (0.08)*	0.31 (0.16)	0.26 (0.09)	0.72 (0.12)	0.38 (0.11)	0.12 (0.05)*
IPCARF (B)	0.11 (0.04)	0.38 (0.12)	0.72 (0.13)	0.27 (0.16)	0.28 (0.10)	0.69 (0.20)	0.38 (0.13)	0.18 (0.09)
IPCARF (MLC)	0.08 (0.02)*	**0.24 (0.05)***	0.82 (0.06)	**0.44 (0.12)***	0.36 (0.09)*	0.77 (0.11)	**0.48 (0.09)***	0.14 (0.07)
IPCARF (HMC)	0.08 (0.02)*	**0.24 (0.07)***	**0.83 (0.07)**	**0.44 (0.13)***	0.36 (0.08)*	**0.78 (0.13)**	**0.48 (0.08)***	0.18 (0.08)
RFLDA (B)	0.11 (0.02)	0.25 (0.04)	0.80 (0.07)	0.33 (0.09)	0.29 (0.06)	**0.78 (0.11)**	0.42 (0.07)	0.16 (0.06)
RFLDA (MLC)	0.08 (0.01)*	0.24 (0.05)	0.81 (0.07)	0.41 (0.11)*	0.34 (0.06)*	0.75 (0.12)	0.46 (0.07)	0.17 (0.07)
RFLDA (HMC)	0.09 (0.02)*	0.25 (0.05)	0.81 (0.06)	0.42 (0.11)*	0.33 (0.07)*	0.75 (0.10)	0.46 (0.08)*	0.12 (0.06)
LDA-SABC (B)	0.08 (0.03)	0.27 (0.08)	0.69 (0.11)	0.28 (0.09)	0.33 (0.06)	0.59 (0.17)	0.41 (0.08)	0.19 (0.09)
RF + ET (B)	0.12 (0.02)	0.37 (0.05)	0.76 (0.08)	0.31 (0.12)	0.26 (0.07)	0.73 (0.11)	0.38 (0.08)	0.12 (0.05)
RF + ET (MLC)	0.09 (0.02)	0.26 (0.07)*	0.81 (0.06)	0.36 (0.13)*	0.33 (0.08)*	0.77 (0.10)*	0.46 (0.09)*	0.13 (0.06)
RF + ET (HMC)	0.10 (0.02)	0.27 (0.06)*	0.80 (0.06)	0.35 (0.13)*	0.32 (0.09)*	0.77 (0.10)*	0.44 (0.10)*	0.10 (0.05)*

TABLE IV

AVERAGE AND STANDARD DEVIATION OF THE RUNNING TIME IN HOURS CONSIDERING 10-FOLDS CROSS-VALIDATION. HERE, B STANDS FOR BINARY, MLC FOR MULTI-LABEL CLASSIFICATION AND HMC FOR HIERARCHICAL MULTI-LABEL CLASSIFICATION.

Method	Mean(std) [hours per fold]
CAFE (B)	6.322 (0.253)
CAFE-OS (B)	6.491 (0.394)
RF+ET (B)	0.064 (0.001)
IPCARF (B)	0.297 (0.042)
LDA-SABC (B)	3.414 (0.792)
RFLDA (B)	0.600 (0.019)
CAFE (MLC)	0.136 (0.018)
CAFE-OS (MLC)	0.134 (0.016)
RF+ET (MLC)	0.001 (0.001)
IPCARF (MLC)	0.035 (0.009)
RFLDA (MLC)	0.129 (0.003)
CAFE (HMC)	0.221 (0.012)
CAFE-OS (HMC)	0.223 (0.014)
RF+ET (HMC)	0.001 (0.001)
IPCARF (HMC)	0.055 (0.013)
RFLDA (HMC)	0.476 (0.010)

29 times longer to run in comparison to its multi-label and hierarchical multi-label versions, respectively.

When compared among themselves, multi-label methods are naturally more efficient, since the output space is smaller, as shown in Table I. However, this difference is still rather negligible, given that HMC provides either competitive or slightly superior performance.

VI. CONCLUSIONS AND FUTURE WORK

In this work, we proposed to address LDA prediction as a (hierarchical) multi-label classification problem. We focused on the inductive learning setting, i.e., situations where predictions are made for new lncRNAs that were not available at training time. Our results showcased that using structured output prediction improves on the state-of-the-art both in terms of predictive performance and computational complexity. More specifically, the performance of LDA methods consistently improved when they were adapted to structured output prediction. Thus, future research on LDA should be based on the vast literature of structured output prediction [2], [19], rather than transforming it into a binary problem.

As future work, we would like to incorporate disease features into the structured output approach, since we currently rely only on the lncRNAs features. Furthermore, we would like to explore the unconfirmed associations further. As aforementioned, they do not necessarily correspond to negative associations, but rather to unconfirmed ones. Thus, using weakly-supervised methods could potentially increase performance [50]. Lastly, we would like to validate our methods with more datasets and incorporate associations among miRNAs as well.

ACKNOWLEDGMENT

This study was financed by the São Paulo Research Foundation (FAPESP) grants 2022/02981-8, 2022/14762-9, and 2024/05438-9, and Research Fund Flanders (FWO) mandate 1235924N.

REFERENCES

[1] T. R. Mercer, M. E. Dinger, S. M. Sunkin, M. F. Mehler, and J. S. Mattick, "Specific expression of long noncoding rnas in the mouse brain," *Proceedings of the National Academy of Sciences*, vol. 105, no. 2, pp. 716–721, 2008.

[2] L. Xu, S. Jiao, D. Zhang, S. Wu, H. Zhang, and B. Gao, "Identification of long noncoding rnas with machine learning methods: a review," *Briefings in Functional Genomics*, vol. 20, no. 3, pp. 174–180, 2021.

[3] X. Liu, D. Li, W. Zhang, M. Guo, and Q. Zhan, "Long non-coding rna gadd7 interacts with tdp-43 and regulates cdk6 mrna decay," *The EMBO journal*, vol. 31, no. 23, pp. 4415–4427, 2012.

[4] K. Hirota, T. Miyoshi, K. Kugou, C. S. Hoffman, T. Shibata, and K. Ohta, "Stepwise chromatin remodelling by a cascade of transcription initiation of non-coding rnas," *Nature*, vol. 456, no. 7218, pp. 130–134, 2008.

[5] W. Hu, J. R. Alvarez-Dominguez, and H. F. Lodish, "Regulation of mammalian cell differentiation by long non-coding rnas," *EMBO reports*, vol. 13, no. 11, pp. 971–983, 2012.

[6] C. Carrieri, L. Cimatti, M. Biagioli, A. Beugnet, S. Zucchelli, S. Fedele, E. Pesce, I. Ferrer, L. Collavin, C. Santoro *et al.*, "Long non-coding antisense rna controls uchl1 translation through an embedded sineb2 repeat," *Nature*, vol. 491, no. 7424, pp. 454–457, 2012.

[7] R. Johnson, "Long non-coding rnas in huntington's disease neurodegeneration," *Neurobiology of disease*, vol. 46, no. 2, pp. 245–254, 2012.

[8] A. Congrains, K. Kamide, R. Oguro, O. Yasuda, K. Miyata, E. Yamamoto, T. Kawai, H. Kusunoki, H. Yamamoto, Y. Takeya *et al.*, "Genetic variants at the 9p21 locus contribute to atherosclerosis through modulation of anril and cdkn2a/b," *Atherosclerosis*, vol. 220, no. 2, pp. 449–455, 2012.

[9] Q. Zhang, C.-Y. Chen, V. S. Yedavalli, and K.-T. Jeang, "Neat1 long noncoding rna and paraspeckle bodies modulate hiv-1 posttranscriptional expression," *MBio*, vol. 4, no. 1, pp. 10–1128, 2013.

[10] R. A. Gupta, N. Shah, K. C. Wang, J. Kim, H. M. Horlings, D. J. Wong, M.-C. Tsai, T. Hung, P. Argani, J. L. Rinn *et al.*, "Long non-coding rna hotair reprograms chromatin state to promote cancer metastasis," *nature*, vol. 464, no. 7291, pp. 1071–1076, 2010.

[11] J. Mattick and P. Amaral, *RNA, the epicenter of genetic information*. Taylor & Francis, 2023.

[12] C. Li, Y.-Q. Ni, H. Xu, Q.-Y. Xiang, Y. Zhao, J.-K. Zhan, J.-Y. He, S. Li, and Y.-S. Liu, "Roles and mechanisms of exosomal non-coding rnas in human health and diseases," *Signal transduction and targeted therapy*, vol. 6, no. 1, p. 383, 2021.

[13] S. Chandra Gupta and Y. Nandan Tripathi, "Potential of long non-coding rnas in cancer patients: from biomarkers to therapeutic targets," *International journal of cancer*, vol. 140, no. 9, pp. 1955–1967, 2017.

[14] N. Sheng, L. Huang, Y. Lu, H. Wang, L. Yang, L. Gao, X. Xie, Y. Fu, and Y. Wang, "Data resources and computational methods for lncrna-disease association prediction," *Computers in Biology and Medicine*, vol. 153, p. 106527, 2023.

[15] X. Chen and G.-Y. Yan, "Novel human lncrna–disease association inference based on lncrna expression profiles," *Bioinformatics*, vol. 29, no. 20, pp. 2617–2624, 2013.

[16] L. Zhou, X. Peng, L. Zeng, and L. Peng, "Finding potential lncrna–disease associations using a boosting-based ensemble learning model," *Frontiers in Genetics*, vol. 15, p. 1356205, 2024.

[17] L. Peng, L. Huang, Q. Su, G. Tian, M. Chen, and G. Han, "Lda-vghb: identifying potential lncrna–disease associations with singular value decomposition, variational graph auto-encoder and heterogeneous newton boosting machine," *Briefings in Bioinformatics*, vol. 25, no. 1, p. bbad466, 2024.

[18] G. Li, P. Bai, C. Liang, and J. Luo, "Node-adaptive graph transformer with structural encoding for accurate and robust lncrna-disease association prediction," *BMC genomics*, vol. 25, no. 1, p. 73, 2024.

[19] G. Tsoumakas, I. Katakis, and I. Vlahavas, *Mining Multi-label Data*. Springer US, 2010, pp. 667–685.

[20] J. Bogatinovski, L. Todorovski, S. Džeroski, and D. Kocev, "Comprehensive comparative study of multi-label classification methods," *Expert Systems with Applications*, vol. 203, p. 117215, 2022.

[21] C. E. Lipscomb, "Medical subject headings (mesh)," *Bulletin of the Medical Library Association*, vol. 88, no. 3, p. 265, 2000.

[22] C. Vens, J. Struyf, L. Schietgat, S. Džeroski, and H. Blockeel, "Decision trees for hierarchical multi-label classification," *Machine learning*, vol. 73, pp. 185–214, 2008.

[23] D. Yao, X. Zhan, X. Zhan, C. K. Kwoh, P. Li, and J. Wang, "A random forest based computational model for predicting novel lncrna-disease associations," *BMC bioinformatics*, vol. 21, pp. 1–18, 2020.

[24] R. Zhu, Y. Wang, J.-X. Liu, and L.-Y. Dai, "Ipcarf: improving lncrna-disease association prediction using incremental principal component analysis feature selection and a random forest classifier," *BMC bioinformatics*, vol. 22, pp. 1–17, 2021.

[25] F. K. Nakano, K. Pliakos, and C. Vens, "Deep tree-ensembles for multi-output prediction," *Pattern Recognition*, vol. 121, p. 108211, 2022.

[26] Z.-H. Zhou and J. Feng, "Deep forest," *National science review*, vol. 6, no. 1, pp. 74–86, 2019.

[27] Q. Chen, D. Lai, W. Lan, X. Wu, B. Chen, J. Liu, Y.-P. P. Chen, and J. Wang, "Ildmsf: inferring associations between long non-coding rna and disease based on multi-similarity fusion," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 18, no. 3, pp. 1106–1112, 2019.

[28] J. Yu, Z. Xuan, X. Feng, Q. Zou, and L. Wang, "A novel collaborative filtering model for lncrna-disease association prediction based on the naïve bayesian classifier," *BMC bioinformatics*, vol. 20, pp. 1–13, 2019.

[29] Q. Gao, G. Han, J. Li, and L. Zeng, "Gmf-mgcn-lda: Prediction of incrna-disease association based on novel generalized matrix factorization and graph neural networks," in *Proceedings of the 15th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, 2024, pp. 1–10.

[30] Q.-W. Wu, R.-F. Cao, J.-F. Xia, J.-C. Ni, C.-H. Zheng, and Y.-S. Su, "Extra trees method for predicting lncrna-disease association based on multi-layer graph embedding aggregation," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 19, no. 6, pp. 3171–3178, 2021.

[31] J. Z. Wang, Z. Du, R. Payattakool, P. S. Yu, and C.-F. Chen, "A new method to measure the semantic similarity of go terms," *Bioinformatics*, vol. 23, no. 10, pp. 1274–1281, 2007.

[32] X. Chen, C. Clarence Yan, C. Luo, W. Ji, Y. Zhang, and Q. Dai, "Constructing lncrna functional similarity network based on lncrna-disease associations and disease semantic similarity," *Scientific reports*, vol. 5, no. 1, p. 11338, 2015.

[33] T. Van Laarhoven, S. B. Nabuurs, and E. Marchiori, "Gaussian interaction profile kernels for predicting drug–target interaction," *Bioinformatics*, vol. 27, no. 21, pp. 3036–3043, 2011.

[34] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine learning*, vol. 63, pp. 3–42, 2006.

[35] W. Liu, H. Wang, X. Shen, and I. W. Tsang, "The emerging trends of multi-label learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 11, pp. 7955–7974, 2021.

[36] G. Tsoumakas, I. Katakis, and I. Vlahavas, "Mining multi-label data," *Data mining and knowledge discovery handbook*, pp. 667–685, 2010.

[37] D. Xu, Y. Shi, I. W. Tsang, Y.-S. Ong, C. Gong, and X. Shen, "Survey on multi-output learning," *IEEE transactions on neural networks and learning systems*, vol. 31, no. 7, pp. 2409–2429, 2019.

[38] M.-L. Zhang and Z.-H. Zhou, "Ml-knn: A lazy learning approach to multi-label learning," *Pattern recognition*, vol. 40, no. 7, pp. 2038–2048, 2007.

[39] ——, "Multilabel neural networks with applications to functional genomics and text categorization," *IEEE transactions on Knowledge and Data Engineering*, vol. 18, no. 10, pp. 1338–1351, 2006.

[40] F. K. Nakano, K. Pliakos, and C. Vens, "Deep tree-ensembles for multi-output prediction," *Pattern Recognition*, vol. 121, p. 108211, 2022.

[41] L. Masera and E. Blanzieri, "Awx: An integrated approach to hierarchical-multilabel classification," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2018, pp. 322–336.

[42] R. Cerri, R. C. Barros, A. C. PLF de Carvalho, and Y. Jin, "Reduction strategies for hierarchical multi-label classification in protein function prediction," *BMC bioinformatics*, vol. 17, pp. 1–24, 2016.

[43] J. Wehrmann, R. Cerri, and R. Barros, "Hierarchical multi-label classification networks," in *International conference on machine learning*. PMLR, 2018, pp. 5075–5084.

[44] E. Giunchiglia and T. Lukasiewicz, "Coherent hierarchical multi-label classification networks," *Advances in neural information processing systems*, vol. 33, pp. 9662–9673, 2020.

[45] G. Chen, Z. Wang, D. Wang, C. Qiu, M. Liu, X. Chen, Q. Zhang, G. Yan, and Q. Cui, "Lncrnadisease: a database for long-non-coding rna-associated diseases," *Nucleic acids research*, vol. 41, no. D1, pp. D983–D986, 2012.

[46] B. Liu, H. Wu, K.-C. Chou *et al.*, "Pse-in-one 2.0: an improved package of web servers for generating various modes of pseudo components of dna, rna, and protein sequences," *Natural science*, vol. 9, no. 04, p. 67, 2017.

[47] R. Fluss, D. Faraggi, and B. Reiser, "Estimation of the youden index and its associated cutoff point," *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, vol. 47, no. 4, pp. 458–472, 2005.

[48] D. Kocev, C. Vens, J. Struyf, and S. Džeroski, "Tree ensembles for predicting structured outputs," *Pattern Recognition*, vol. 46, no. 3, pp. 817–833, 2013.

[49] F. K. Nakano, M. Lietaert, and C. Vens, "Machine learning for discovering missing or wrong protein function annotations: A comparison using updated benchmark datasets," *BMC bioinformatics*, vol. 20, pp. 1–32, 2019.

[50] P. Ilidio, R. Cerri, C. Vens, and F. K. Nakano, "Deep forests with tree-embeddings and label imputation for weak-label learning," in *2024 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2024, pp. 1–8.