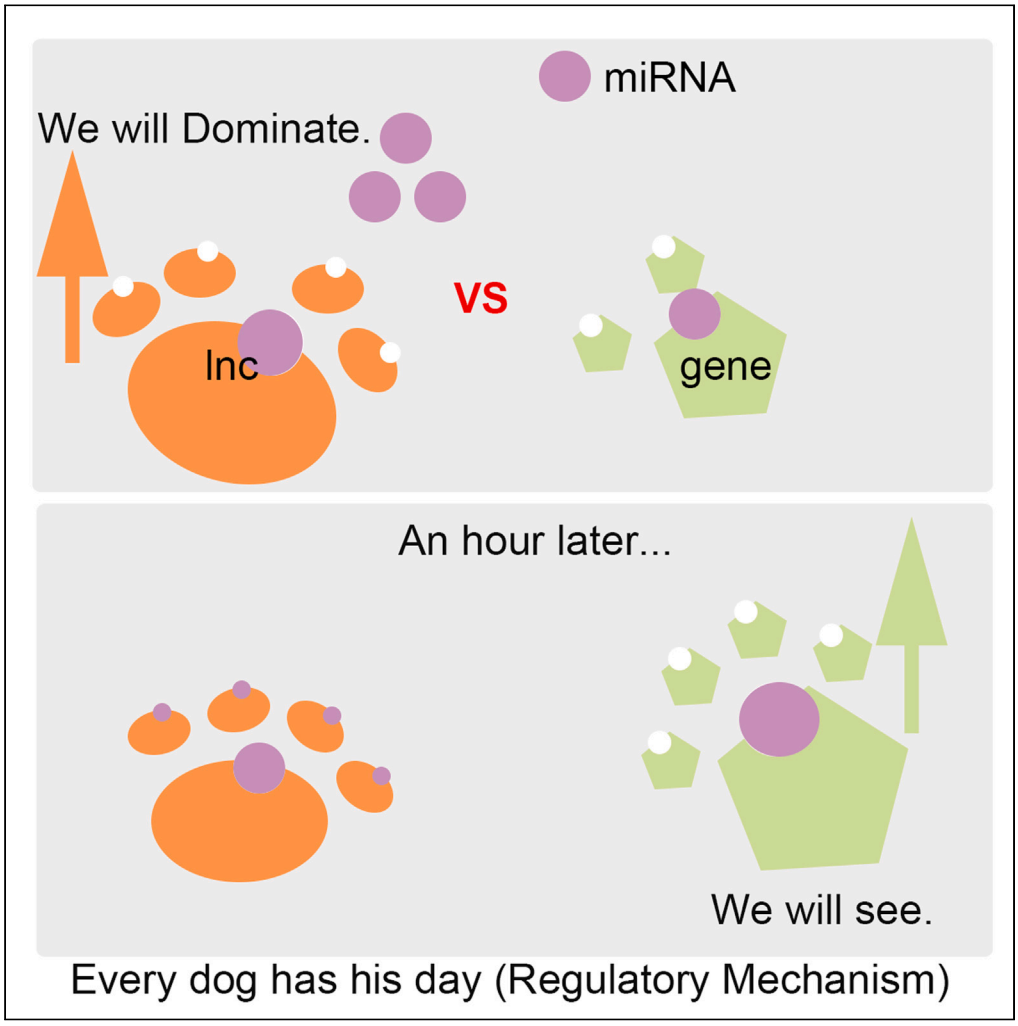
iScience

CellPress
OPEN ACCESS

Article

LncRNA-Top: Controlled deep learning approaches for lncRNA gene regulatory relationship annotations across different platforms



Weidun Xie, Xingjian Chen, Zetian Zheng, ..., Qiuzhen Lin, Yanni Sun, Ka-Chun Wong

kc.w@cityu.edu.hk

Highlights

Predict lncRNA-gene relations by considering the regulatory mechanism

Represent lncRNA/gene by features across different platforms

Multi-dimensional analysis to reveal the predictive ability of the proposed model

LncRNA-Top enhanced robustness and potential top-predicted relations

Xie et al., iScience *26*, 108197
November 17, 2023
https://doi.org/10.1016/j.isci.2023.108197

iScience

CellPress
OPEN ACCESS

Article

LncRNA-Top: Controlled deep learning approaches for lncRNA gene regulatory relationship annotations across different platforms

Weidun Xie,[1] Xingjian Chen,[1] Zetian Zheng,[1] Fuzhou Wang,[1] Xiaowei Zhu,[2] Qiuzhen Lin,[3] Yanni Sun,[4] and Ka-Chun Wong[1,5,6,7,*]

SUMMARY

By soaking microRNAs (miRNAs), long non-coding RNAs (lncRNAs) have the potential to regulate gene expression. Few methods have been created based on this mechanism to anticipate the lncRNA-gene relationship prediction. Hence, we present lncRNA-Top to forecast potential lncRNA-gene regulation relationships. Specifically, we constructed controlled deep-learning methods using 12417 lncRNAs and 16127 genes. We have provided retrospective and innovative views among negative sampling, random seeds, cross-validation, metrics, and independent datasets. The AUC, AUPR, and our defined precision@k were leveraged to evaluate performance. In-depth case studies demonstrate that 47 out of 100 projected top unknown pairings were recorded in publications, supporting the predictive power. Our additional software can annotate the scores with target candidates. The lncRNA-Top will be a helpful tool to uncover prospective lncRNA targets and better comprehend the regulatory processes of lncRNAs.

INTRODUCTION

The long non-coding RNAs (lncRNAs) are not translated into proteins and can span over 200 nucleotides.[1,2] More than 16,000 lncRNAs were reported in the *Homo sapiens*, as revealed in the human GENCODE project.[3] Although lncRNAs are less characterized than protein-coding genes, their annotations may impact downstream research.[4] By controlling target gene expression levels, lncRNAs have various activities, influencing motility, invasion, differentiation, apoptosis, and proliferation.[5] Researchers have linked lncRNA dysregulation and mutations with a greater propensity for complex disorders such as cancer.[6,7]

Finding new lncRNA relationships with targets as genes may be crucial for developing tailored therapies or discovering disease mechanisms. However, as lncRNAs may affect genes[8,9] in various ways,[10] finding new relationships can be difficult. *In vivo* validation tests are time-consuming and labor-intensive; thus, the *in silico* screen has great potential. *In silico* experiment relies on prior knowledge such as databases. Fortunately, researchers have curated databases concerning lncRNA, such as LNCipedia[11] for lncRNA sequences, lncRNAfunc[12] for lncRNA functions, LncRNADisease[13] for lncRNA illnesses, and LncRNA2Target[14] and LncTarD[15] for lncRNA-target pairings. Those datasets provide prerequisites for *in silico* experiments.

Previously, researchers have conducted *in silico* experiments on lncRNA. For instance, to predict latent lncRNA-disease interactions, Xuan et al.[16] combined heterogeneous networks with graph convolutional networks (GCN) and convolutional neural networks (CNN). DeepMNE[17] merged the disease similarities derived from disease semantic information, functions, Gene Ontology terms, and lncRNA similarities originating from sequences and expression files, using known lncRNA-disease relations to predict new lncRNA-disease associations.

Other methods directly predict the relationships between lncRNA and other molecules. For instance, Huang[18] designed a GCN autoencoder to predict lncRNA- microRNA (miRNA) relations from molecular networks. Fukunaga[19] collected tissue-specific expression profiles and subcellular localization information to predict potential lncRNA-mRNA relationships. A recent study that employs deep learning to reveal potential lncRNA-gene pairs is DeepLGP,[20] which collects expression and genomic location features and constructs positive pairs based on LncRNA2Target[14] and negatives from random lncRNA-gene pairs. Additionally, GCN is applied to extract features, while CNN is employed to make predictions farther down the line. IRDL[21] integrated sequencing, gene expression, and chromatin accessibility to identify divergent lncRNAs with target genes. The LPI-deepGBDT[22] takes gradient-boosting decision trees and features from Pyfeat and BioProt for potential lncRNA-protein relationship prediction. DNNMC[23] proposes a lncRNA-protein-coding gene (PCG) computational method combining deep

[1]Department of Computer Science, City University of Hong Kong, Kowloon Tong, Hong Kong SAR
[2]Department of Neuroscience, Jockey Club College of Veterinary Medicine and Life Sciences, City University of Hong Kong, Kowloon Tong, Hong Kong SAR
[3]College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China
[4]Department of Electrical Engineering, City University of Hong Kong, Kowloon Tong, Hong Kong SAR
[5]Shenzhen Research Institute, City University of Hong Kong, Shenzhen, China
[6]Hong Kong Institute for Data Science, City University of Hong Kong, Kowloon Tong, Hong Kong SAR
[7]Lead contact
*Correspondence: kc.w@cityu.edu.hk
https://doi.org/10.1016/j.isci.2023.108197

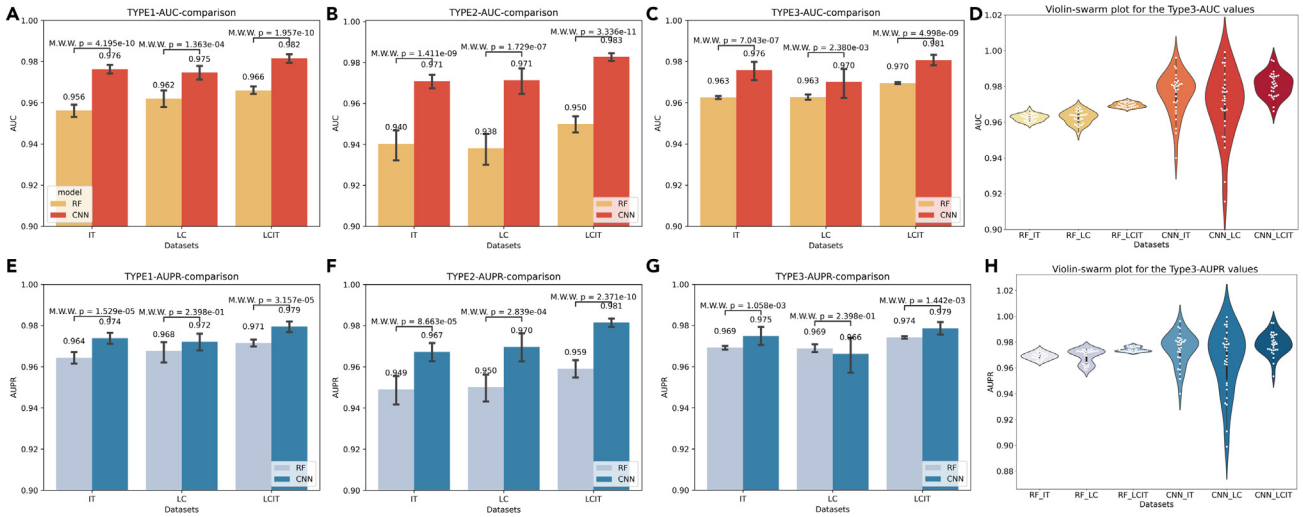


Figure 1. Verification using three different approaches on each dataset

(A–C) The barplot of AUC values inside IT, LC, and LCIT datasets with adjusted p value was annotated using the rank-sum test with Bonferroni correction using type-1, type-2, and type-3 methods.

(D) The violin-swarm plot of AUC for type-3 10-fold cross-validation.

(E–G) The barplot of AUPR values inside IT, LC, and LCIT datasets.

(H) The violin-swarm plot of AUPR for type-3 10-fold cross-validation, each white point indicates one experiment with different sampling random seeds and 10-fold cut random seeds. Mann-Whitney-Wilcoxon tests were conducted, and p-values were calculated to compare the performance of each feature extraction or prediction model. The figure's error bar indicates a specific metric's distribution range.

neural networks and inductive matrix completion using multi-omics data and known association. GAE-LGA[24] leverages multi-omics features of lncRNA and PCG as the source for the similarity calculation and applies the graph-autoencoder for potential relation interference. Most of those methods considered expressional profiles, multi-omics data, or location of gene and lncRNA as features.

Nonetheless, few considered the functional mechanism, such as lncRNA playing a role as ce-RNA to regulate the gene.[25,26] For example, lncRNA-PNUTS, an alternatively spliced lncRNA from PNUTS premature mRNA, contains seven binding sites to miR-205, enhancing ZEB1 and ZEB2 mRNAs by reducing miR-205-binding ability.[27] The miRNA can regulate gene post-transcription by binding to the 3′UTR of genes.[28] lncRNA can also provide binding sites for miRNA. Thus, building predictive models directly from lncRNA and gene 3′UTR sequences seems practical. The next thing is to select predictive models.

In other bioinformatics fields, deep learning has had considerable success, such as protein structure prediction,[29] inference of miRNA-gene links,[30] compound-protein interactions,[31] cancer survival analysis,[32] and cell fate.[33] However, deep learning is likely to show considerable variance, especially when using tiny datasets with high dimensions, such as omics characteristics from biological features,[34] which deteriorates the model's generalization ability. The model's capacity for generalization demonstrates its flexibility in responding to new data. However, many previous deep learning models, such as our previous work, only considered the area under the curves (AUC) values inside one dataset. Furthermore, when choosing the negative samples, how the random seeds may influence the results is merely explored. That lack of generalization verification demands independent datasets and more experiments on different random seeds.

Ensemble approaches are an excellent solution to increase model generalization. Combining the output from many deep-learning models can improve the model's generalizability.[35] Mixed deep-learning and machine-learning models may further improve generalization capabilities.[36] Ensemble techniques have been employed in many link prediction tasks of bioinformatics applications, including Cas9 off-target, miRNA-target, microbiome-drug, and microbiome-disease. For instance, Zhang et al.[37] used ensemble AdaBoost to predict CRISPR-Cas9 off-target activities by combining characteristics from synergizing methods, conservations, and chromatin annotations. EnANNDeep[38] is an ensemble method aggregating AkNN, DNN, and gcForst results for scoring lncRNA-protein pairs. The features of which are extracted from protein sequences and structures. To predict miRNA-target correlations, SRG-vote[39] integrated several long short-term memory models trained with features from different aspects. Long et al.[40] proposed an ensemble graph network with three models and several input sources to identify potential relationships between medications and microorganisms. Chen et al.[41] made soft voting of four CNN models for human disease-microbiome prediction. These studies indicate that integrating models may have better performance. However, those methods only considered ensembles inside one dataset with different models, not considering ensembles with different negative sampling random seeds and validating them on the independent datasets.

Inspired by this, we designed lncRNA-Top, the controlled deep-learning approach (random forest (RF)-CNN ensemble) to predict lncRNA-gene relationships leveraging sequence-based characteristics inspired by the lncRNA regulatory mechanism. In the meantime, we have retrospected and revisited some spots and details that are easily ignored in deep-learning tasks, including negative sampling with varied

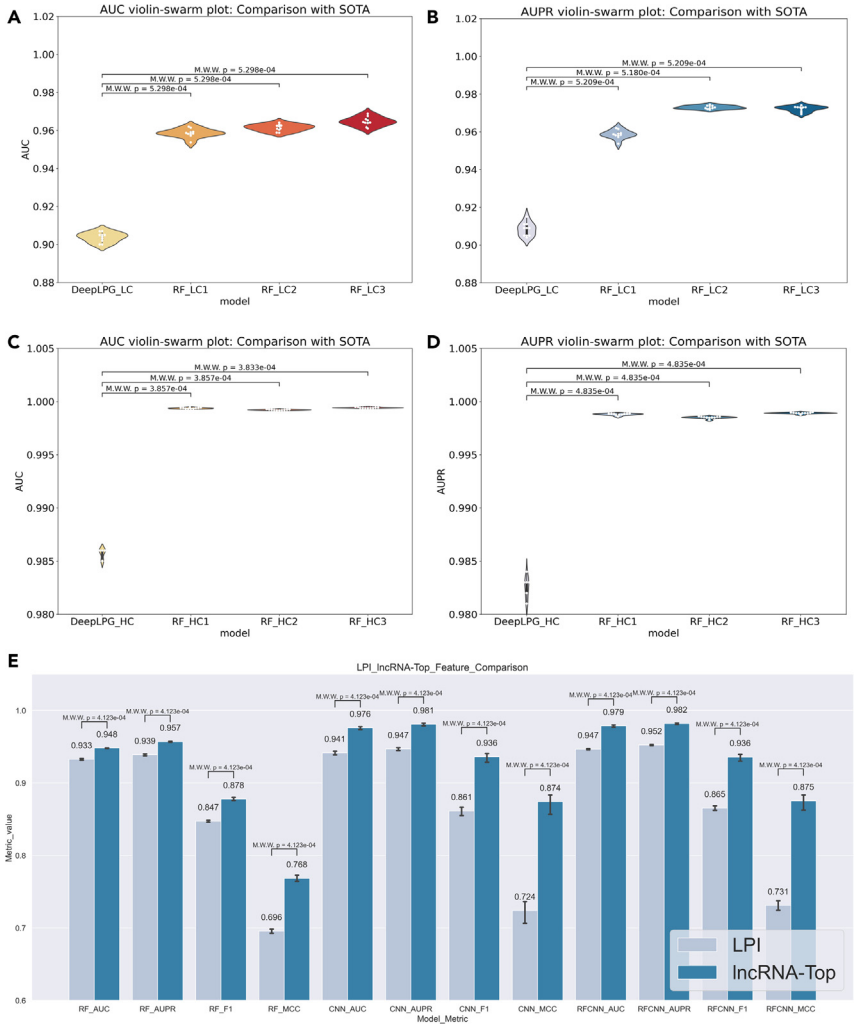


Figure 2. Comparison with the SOTA method

(A and B) The violin-swarm plot of AUC and AUPR of our methods, when compared with the SOTA method DeepLPG, in the low-throughput constructed (LC dataset) dataset, 1, 2, 3 here mean the negative sampling random seeds. Adjusted p value with Bonferroni correction is annotated.

(C and D) The violin-swarm plot of AUC and AUPR of our methods, when compared with the SOTA method DeepLPG, in the high-throughput constructed dataset (HC dataset).

(E) The barplot indicated the performance of features from LPI and ours. The x axis is the models and metrics. We introduced AUC, AUPR, F1, and MCC in this experiment. Mann-Whitney-Wilcoxon tests were conducted, and p-values were calculated to compare the performance of each feature extraction or prediction model. The figure's error bar indicates a specific metric's distribution range.

random seeds, different types of cross-validation, results confidence by the statistical calculation, varied independent datasets, ensemble verification on independent datasets, and differentiated metrics. Except for AUC, we also introduced other metrics, such as area under precision-recall (AUPR) curve and our defined precision@K. If one model can get a higher AUPR, it can control the false positive rate (FPR),[42] which also means that the top-predicted pairs are more likely to be the "real-positive" pair. These metrics help us polish our ensemble policy when predicting the actual scenario.

The contributions of lncRNA-Top are listed as follows.

(1) We designed a model to predict lncRNA-gene regulation relations from the lncRNA sequence and gene UTR 3' sequence by considering the mechanism of how they interact. The features with simple random forest can achieve higher AUC/AUPR than the SOTA (state-of-the-art) deep-learning methods.

(2) We have introduced multi-dimensional analysis to reveal the predictive ability of the proposed model (including statistical analysis, transfer verification, and precision@k).

(3) We utilized hybrid ensemble controlled deep-learning approaches, significantly improving the deep-learning models' robustness and AUC/AUPR for predicting potential relations between lncRNAs and genes.

Table 1. Comparison with the SOTA method

model	No. of gene	No. of lncRNA	Feature dimension	Feature sources
DeepLGP[20]	6782	427	16	Manually selected
LPI[22]	–	–	410/351	Pyfeat/BioProt+PCA
GAE-LGA[24]	256–716	208–268	312	Multi-omics
Ours	16127	12417	4096	iLearn+kmer+KPCA

LPI is designed to predict lncRNA and protein relationships and provides feature generators from sequences. GAE-LGA is a graph-based method proposed for lncRNA and PCG (protein-coding gene) relation prediction. In the framework of GAE-LGA, features were leveraged for similarity calculation. We compared our method in their framework by replacing its original multi-omics features with LPI generated and our features.

RESULTS

Verification using three different approaches on each dataset

Three types of verification results for each sub-model, such as RF and CNN, on different datasets are shown in Figure 1. Rank-sum tests were leveraged to calculate the adjusted p value with Bonferroni correction. Except for the type-1 and type-3 AUPR in the low-throughput constructed (LC) dataset, all datasets show that CNN performs better than RF with adjusted p values less than 0.05. From the results, some interesting conclusions can be drawn, such as: Integrating the IT and LC datasets may improve the model's performance and lower error bars; leveraging more training data (90% of data instead of 80%) results in higher performance in type-3 than in type-1 and type-2.

Figures 1D and 1H demonstrate the overall performance of merging the results of type-3 verification of AUC and AUPR. We can see from the images that for the 10-fold cross-validation, RF's results are less variable than CNN's, demonstrating that deep learning exhibits significant variance, which means that the deep-learning models (CNN) are more easily affected by the random seeds. The results further indicate that, when conducting deep learning experiments inside one dataset, it is necessary to repeat the experiments from the negative sampling.

Comparison with the SOTA methods

Although our method is specially designed for the lncRNA-gene regulation relations, given that we share the same training datasets with the SOTA method DeepLGP,[20] we could still compare our AUC/AUPR with theirs. We also leveraged the 10-fold cross-validation. Details of the results are shown in Figures 2A–2D.

From the results, we could see that the RF with Poly-kernel-PCA-transformed features could get significantly higher results (adjusted p value with Bonferroni correction $\leq$ 0.05) in both the LC and high-throughput constructed (HC) datasets. DeepLGP leverages GCN and CNN to discover potential relationships between lncRNA and genes, a deep-learning-based model. We only leveraged our non-deep-learning algorithm and got higher AUC/AUPR, demonstrating that our mechanism involved sequence-based methods that are working.

One rationale is that we utilized more advanced features than the SOTA approaches; and another is that we created more gene and lncRNA features. We further explored the feature generated from different lncRNA-related predictive models. A detailed method comparison can be found in Table 1. We compared those features in two steps. The first step is leveraging LPI-generated features in our predictive framework; the results are shown in Figure 2E. The second step is to compare the performance of different features under the framework GAE-LGA. Those features include original multi-omics features, LPI's features, and our features. The results of different features under GAE-LGA can be found in Table 2.

From the figure, we could find that the features generated by the lncRNA-Top can achieve better performance for all sub-models and metrics. Table 2 demonstrates that the original multi-omics features are better than LPI's generated features but inferior to ours

Table 2. Results of GAE-LGA with different features as input

Method	D1_ AUC	D1_ AUPR	D1_F1-score	D1_ MCC	D2_ AUC	D2_ AUPR	D2_F1-score	D2_ MCC	D3_ AUC	D3_ AUPR	D3_F1-score	D3_ MCC
GAE-LGA_ (Adjust)	**0.9522**	**0.5479**	0.7691	0.6071	0.923	0.5984	0.8265	0.6907	0.8149	0.4169	0.7761	0.6007
GAE-LGA_ (LPI_Features)	0.9482	0.5407	0.7692	0.6072	0.918	0.5833	0.7901	0.633	0.8404	0.5312	0.8172	0.6524
GAE-LGA_ (Our_Features)	0.9489	0.5453	**0.7726**	**0.6124**	**0.9242**	**0.6176**	**0.834**	**0.7019**	**0.8568**	**0.5683**	**0.8359**	**0.6809**
Increment (%)	−0.346	−0.476	0.4551	0.8730	0.1300	3.2085	0.9074	1.6215	5.142	36.315	7.7051	13.351

We first experimented in the adjusted network (after removing the lncRNA/gene we don't have). We acquired a benchmark value of GAE-LGA, named GAE-LGA (Adjusted), as the adjusted network is smaller than the original one. The metrics calculated are slightly inferior to the original paper's description. Then, we replaced the original multi-omics features with LPI features and our features and reconducted experiments again. The bold value corresponds to the best performance method for each metric. The results show that our features can outperform most of the metrics.

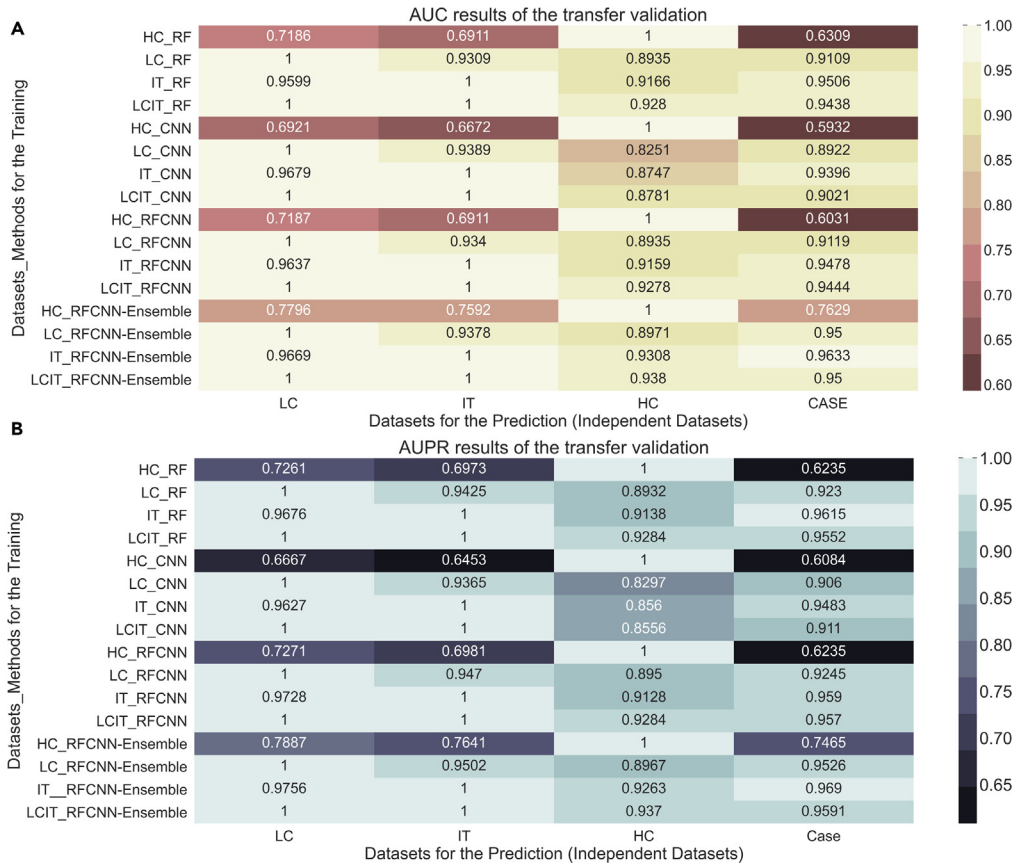


Figure 3. Transfer verification for independent datasets (Heatmap)

(A and B) Heatmap of average AUC/AUPR of different methods and training-testing sets. Each cube zipped one machine-learning method, one training dataset, and an independent testing dataset in the heatmap. The AUC/AUPR values are marked with light to dark colors, denoting the value from large to small. Every four rows indicate models or ensemble methods, each column indicates each independent dataset for testing, and each row indicates model-trained datasets, including HC, LC, IT, and LCIT.

(for 10 out of 12 metrics). D1 (Dataset1), D2, and D3 are introduced datasets by GAE-LGA.[24] D1 contains 208 lncRNAs, D2 contains 238 lncRNAs, and D3 contains 263 lncRNAs. Notably, the performance increased when more lncRNA was introduced to the dataset.

Transfer verification for independent datasets

The results (sorted by the models) of the transfer verification among independent datasets are shown in Figure 3. We compared the mean value of AUC/AUPR of RF, CNN, RF-CNN, and our controlled ensemble methods, RF-CNN ensemble. As can be seen from the picture of Figures 3A and 3B, the RF-CNN-ensemble (last four rows) can get lighter color in both AUC (yellow) and AUPR (blue), indicating that aggregating the scores predicted from different random-seeds generated training sets can increase the metric of AUC/AUPR.

Ensemble composition

As ensemble policy from different training random seeds can increase the overall performance, the next thing is to figure out the optimal composition of the ensemble. The LCIT-HC is more likely to be the actual scenario, where the pairs to be predicted are enormous, and the training dataset is limited. Thus, we further generated ROC and PR curves of the CNN and RF ensemble on the LCIT training set. We also introduced the sub-model-only results as a comparison. The results are shown in Figure 4.

Prediction ability comparison

To further assess the RF ensemble, CNN ensemble, and RF-CNN ensemble's capacity for prediction, we trained models on the LCIT dataset. All pairs of lncRNA and genes were taken into consideration. The precision@K is leveraged for the comparison. The results are listed in Table 3. Here, we select k between 10 and 100. The results demonstrate that the RF-CNN ensemble could get much higher precision@K in all scenarios, which denotes better predictive performance.

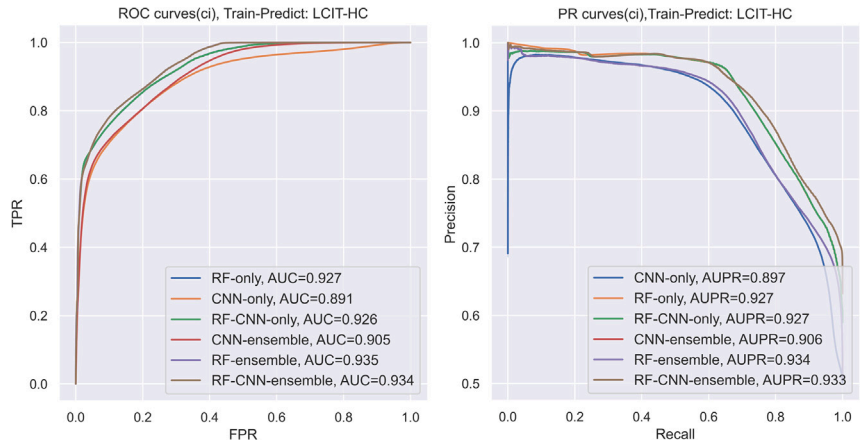


Figure 4. Ensemble composition

The figure described the ROC and PR curves when using LCIT to predict the HC dataset. This project also leverages the small dataset to predict the large dataset, which is the closest to the actual scenario. In the first picture, RF-only is barely overlapped with RF-CNN-only, and RF-ensemble is barely overlapped with the RF-CNN ensemble. The ROC and PR curves are drawn based on the concatenated label and scores of the independent dataset (for example, if we leverage the LCIT_rs1 to train the model and generate a score for the HC_rs2, rs3 sample, we would concatenate scores and labels to generate RF-only or CNN-only ROC curves). The independent sample score is the sum of different random seeds trained models for ensemble methods. As can be seen from the picture, the performance of ensemble methods like RF-ensemble and CNN-ensemble are increased compared with their sub-model. Also, the RF ensemble can get the highest AUC and AUPR in this task. The RF-CNN-ensemble ranked second in this task.

DISCUSSION

Parametrical analysis

In the framework of lncRNA-Top, we have plenty of parameters that might contribute to the model robustness. To further explore those factors, we conducted a parametrical analysis. The variables include kernels and dimensions of Kernel PCA when transforming the features, different sub-machine-learning models (Figure S1), negative sampling rate (Figures S2–S4), and random seeds (Figure S5; Table S1) for cross-validation and ensemble.

RF-CNN ensemble zoomed in comparison

As shown in Figure 5, the RF-ensemble model seems to be the best option as it can achieve higher true positive rate or grow fast when the FPR rises in all predicting HC tasks. However, when we further screen the predictive results from the RF ensemble with all pairs of 12417 lncRNA and 16127 genes, their top results frequently tend to be indistinguishable. (RF-ensemble predicted pairs shared identical scores, resulting in the same rank.) Thus, by adding the CNN scores, the controlled deep-learning algorithms (RF-CNN ensemble) may produce more separable top outcomes with slight deterioration to AUC/AUPR but a significant increase to the top-predicted results (precision@K). Another possible reason why the RF-CNN ensemble is better in real predictive situations is that the RF-CNN ensemble is the aggregate results of six models. In comparison, RF ensembles and CNN ensembles only contain three models.

Top predictive results

To further illustrate the predictive ability of lncRNA-Top, we searched the literature for the highly predicted lncRNA-gene relationships. The top ten predicted results (those that have appeared in the training set are removed) are shown in Table 4.

Table 3. The prediction ability of the different ensemble strategies

Precision@K	RF ensemble (training included)	CNN ensemble (training included)	RF-CNN ensemble (training included)
10	0.1	0	0.5
20	0.15	0.05	0.65
30	0.17	0.07	0.57
50	0.1	0.06	0.48
70	0.07	0.07	0.47
90	0.06	0.08	0.46
100	0.05	0.11	0.47

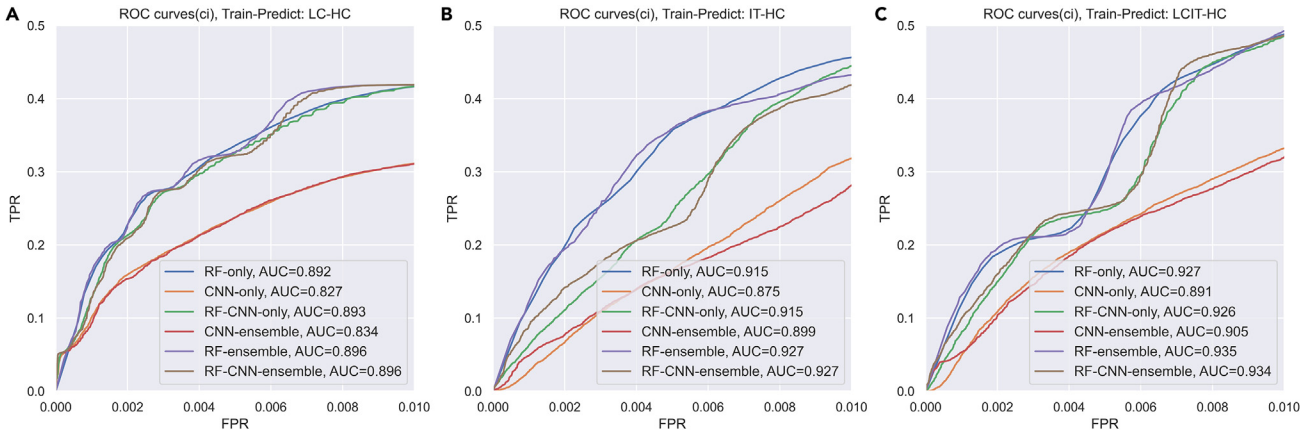


Figure 5. RF-CNN ensemble zoomed in comparison

(A and B) Zoomed in ROC curves of (A) LC predicts HC, (B) IT predicts HC, and (C) LCIT predicts HC. Here, the FPR is cut off by 0.01. As can be seen from the picture, when the false positive rate (FPR) is small and begins to grow, the true positive rate (TPR) rises quickly. Considering we have 12417 lncRNAs and 16127 genes, the total pair is 200,000,619. Thus, even if a small part of the TPR increases, it can significantly increase the predictive ability.

The regulatory mechanism that inspired us to construct this predictive model is that the lncRNA can regulate miRNA (microRNA) while miRNA can regulate genes. Thus, lncRNA can be a competitor to some genes that share the same target miRNAs. Our previous publications[30,39] demonstrated that modelling miRNA-gene relationships between miRNA and 3′UTR gene sequences worked. Inspired by this, in this model, we also constructed our predictive models using 3′UTR gene sequences.

Interestingly, although no miRNA information (as we directly modeled on lncRNA sequence and gene sequence) was introduced to our method, they were mentioned in most publications. We list the miRNA in the table and find that five out of the top 10 lncRNA-gene relations predicted by our framework are related to miRNA. Here are some examples: the miRNA miR-370-3p can target gene mitogen-activated protein kinase (MAPK1), and therefore, lncRNA taurine-upregulated gene 1 (lncRNA TUG1) could reduce the level of functional miR-370-3p and facilitate MAPK1 expression.[43] The elevated level of miR-197 in cells treated with lipopolysaccharide (LPS) was inhibited by transfection with TUG1, which can sponge miR-197 to enhance the level of *p*-MAPK/MAPK, thereby inducing autophagy, indicating the upregulating of TUG1 might inhibit the increase of miR-197 and enhance MAPK.[44] SNHG1 and MAPK1 were significantly upregulated, while miR-125b-5p was downregulated in the MPTP-induced PD mouse and MPP+-induced PD cell models.[45] SNHG1 competitively binds to the miR-221/222 cluster and indirectly regulates the expression of cyclin-dependent kinase inhibitor 1B (CDKN1B/p27).[46] NEAT1 forms double-stranded RNA with miR-222-3p, thus limiting miR-222-3p's binding with CDKN1B, which indicates that the upregulate of NEAT1 will reduce the miR-222-3p. Thus, CDKN1B would be upregulated.[47] lncRNA NEAT1 was upregulated while miR-27a-3p was downregulated in SH-SY5Y cells, and CASP-3 protein and its lytic cell protein were upregulated.[48] Neat1_2 (a transcript variant of NEAT1) competitively binds to miR-129-5p and prevents miR-129-5p from decreasing FADD, CASP-8, and CASP-3 levels, ultimately facilitating TEC apoptosis. Neat1 binding to miR-129-5p, then the level of CASP-3 is increased.[49] We further marked the direction of regulation that the papers indicated, and these cases

Table 4. Top unknown lncRNA-target relations predicted by lncRNA-Top

Rank	gene_name	lnc_name	Score	Evidence/Year	PMID (Regulating miRNA)
1	SNAI1	HOXA-AS2	0.9959		
2	MAPK1 ↑	TUG1 ↑	0.9952	[43]/2019 [44]/2019	31819520 (miR-370-3p)↓ 31115515 (miR-197) ↓
3	MAPK1 ↑	SNHG1 ↑	0.9952	[45]/2021	33911864 (miR-125b-5p)↓
4	SNAI1	TUG1	0.9946		
5	CDKN1B ↓	SNHG1 ↓	0.9943	[46]/2019	31499060 (miR-221/222) ↑
6	GSK3B	HOXA-AS2	0.9943		
7	CDKN1B ↑	NEAT1 ↑	0.994	[47]/2021	33585566 (miR-222-3P) ↓
8	CDKN1B	TUG1	0.994		
9	HAS2	HOXA-AS2	0.9934		
10	CASP3 ↑	NEAT1 ↑	0.9933	[48]/2021 [49]/2022	34540002 (miR-27a-3p) ↓ 35619557 (miR-129-5p) ↓

The ↑ means the molecule is upregulated, ↓ means the molecule is downregulated according to the publications.

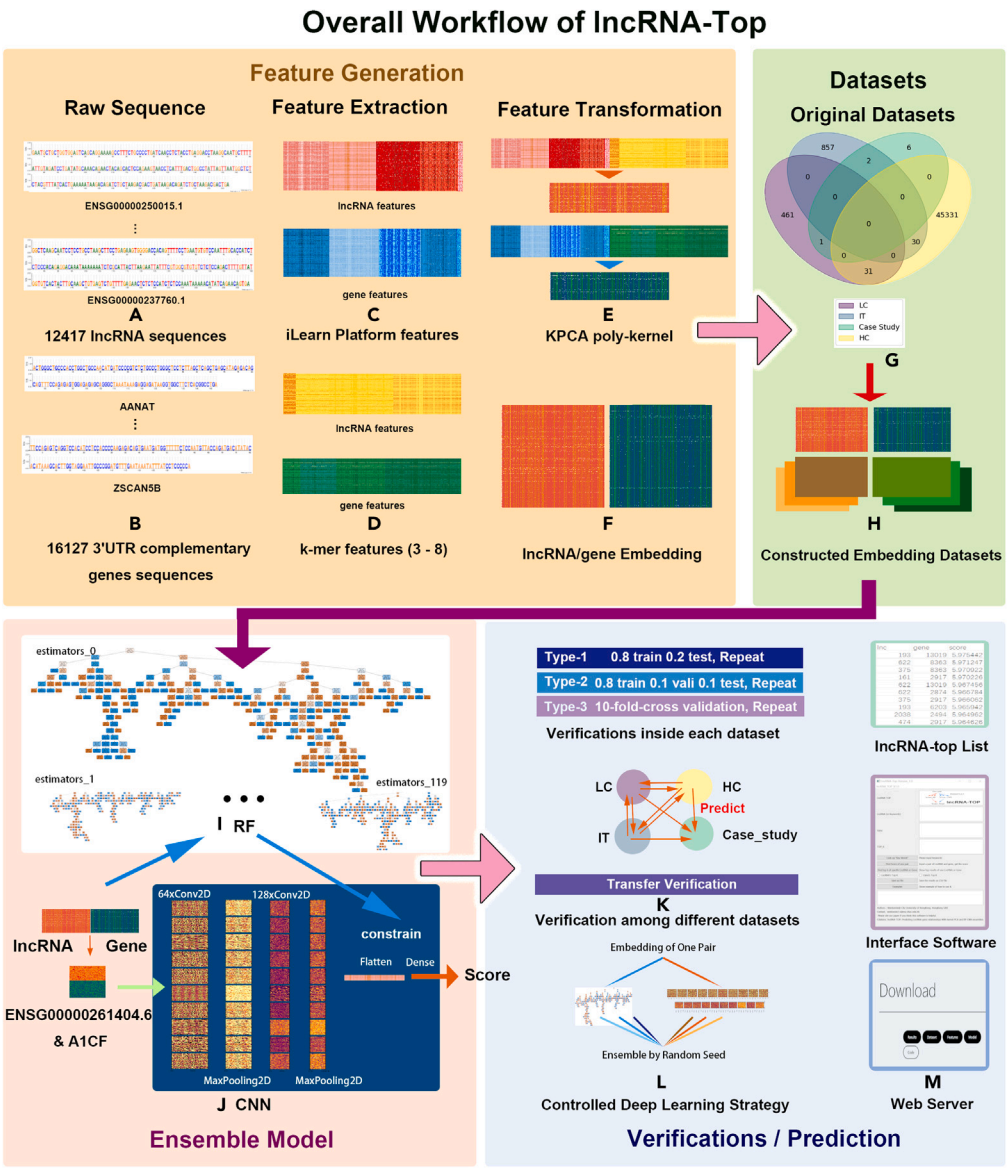


Figure 6. The overall workflow of lncRNA-Top

The workflow can be divided into four parts. The yellow parts denote the feature generation, and the green part reveals the dataset for downstream works. The reddish part indicates the ensemble model we take for the prediction, and the bluish part illustrates the verification we have done to validate our models.

(A and B) lncRNA/gene 3'UTR sequences examples.

(C) Generation of iLearn features from the iLearn platform.

(D) Generation of k-mer (k = 3 to k = 8) features.

(E) LncRNA and gene feature transformation by the KPCA with poly kernels.

(G and H) Using the LC, IT, HC, and Case_study datasets to construct four embedding datasets using the previously generated features (with different random seeds).

(I) RF model trained by datasets.

(J) CNN models trained by datasets.

(K) Validation methods (three types) inside each dataset and the transfer verification among each dataset.

(L) Controlled deep-learning strategy, using different random seeds to construct a dataset and train sub-models. We predict the final results with RF and CNN models (controlled deep learning), differentiating the final scores and improving predictive performance.

(M) Software and web server demo.

Table 5. Datasets and their usage in our method

Datasets	Usage/Purpose
Original datasets	Provide positive pairs (known pairs, such as[14,15])
Constructed datasets	Provide positive/negative pairs with random seeds.
Constructed embedding datasets	Generate from constructed datasets and provide training/testing set for downstream machine-learning algorithms.
Independent datasets	To verify the performance of the trained model
LNCipedia v5[11]	Provide lncRNA sequences
'biomaRT' R package[52]	Provide gene 3'UTR sequences (13815/16127 genes were protein-coding genes)
LncRNA2Target v2.0[14]	Provide positive samples, an independent dataset for[15]
lncTarD[15]	Provide positive samples, an independent dataset for[14]
Case study datasets[10]	Extracted from,[10] the Independent dataset for[14,15]

indicate that the lncRNA and gene are regulated in the same direction. If the lncRNA is upregulated and the corresponding gene is upregulated, the target miRNA is sponged and downregulated. The top-predicted lncRNA-gene pairs further indicate that building a predictive model based on the lncRNA sequences and gene 3'UTR sequences directly can reveal authentic gene transcriptional regulations by lncRNA through mechanisms such as competitive endogenous RNA. We also marked the year of publication. Those recent publications are not included in any dataset demonstrating lncRNA-Top's predictive performance.

We tested the prediction of the case study dataset. Results can be found in Table S2.

In this study, we proposed lncRNA-Top, controlled deep-learning approaches that predict potential lncRNA-gene regulatory relations inspired by regulatory mechanisms. The overall workflow of lncRNA-Top is illustrated in Figure 6. The usage and purpose of different datasets in this manuscript are shown in Table 5. Details of the constructed datasets leveraged for machine learning model training can be found in Table 6. We explored the influence of random seeds, ratios of negative sampling, different cross-validation, and transfer verification among datasets and conducted a multi-dimensional analysis of the predicted results. The controlled deep-learning approaches hybrid ensemble datasets and results of machine-learning and deep-learning models, increasing the AUC/AUPR/Precision@k of the predictive method. The case study denotes that our suggested approaches can accurately identify those lncRNA and gene regulatory relationships with substantial evidence. The code, features, software, and website are provided.

Limitations of the study

In our framework, we only considered the sequence-based feature extractors. However, there are many features from other domains, such as multi-omics and graph-based features. Introducing those features might contribute positively to the performance. We will explore more features in the next version of lncRNA-Top.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY

Table 6. The details of the constructed datasets

Constructed datasets	lnc	Gene	Positive pairs	Negative pairs	Source
LC	118	276	493	493	LncRNA2Target[14]
HC	36	12434	45331	45331	LncRNA2Target[14]
IT	192	455	889	889	LncTarD[15]
LCIT (union set)	237	613	1382	1382	(union set)[14,15]
Case study	8	9	9	9	Review paper[10]

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2023.108197.

ACKNOWLEDGMENTS

This research was substantially sponsored by the research projects (grant no. 32170654 and 32000464) supported by the National Natural Science Foundation of China and was substantially supported by the Shenzhen Research Institute, City University of Hong Kong. The work described in this paper was substantially supported by the grant from the Research Grants Council of the Hong Kong Special Administrative Region [CityU 11203723]. This project was substantially funded by the Strategic Interdisciplinary Research Grant of City University of Hong Kong (Project No. 2021SIRG036). The work described in this paper was partially supported by the grant from City University of Hong Kong (CityU 9667265).

AUTHOR CONTRIBUTIONS

W.X. and K.-C.W. conceived the study; W.X. designed and implemented the algorithms; W.X., X.C., Z.Z., and F.W. conducted the result analysis. W.X., X.Z., Q.L., Y.S., and K.-C.W developed the software, designed the web server, and wrote the manuscript. K.-C.W. funded and supervised the study. All authors read and approved the final manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interest.

Received: April 24, 2023
Revised: August 10, 2023
Accepted: October 10, 2023
Published: October 12, 2023

REFERENCES

1. Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M.C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., et al. (2005). The transcriptional landscape of the mammalian genome. Science *309*, 1559–1563.
2. Guttman, M., Amit, I., Garber, M., French, C., Lin, M.F., Feldser, D., Huarte, M., Zuk, O., Carey, B.W., Cassady, J.P., et al. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature *458*, 223–227.
3. ENCODE Project Consortium, Good, P., Guyer, M., Kamholz, S., Liefer, L., Wetterstrand, K., Collins, F., Gingeras, T., Kampa, D., Sekinger, E., et al. (2004). The ENCODE (ENCyclopedia of DNA elements) project. Science *306*, 636–640.
4. Uszczynska-Ratajczak, B., Lagarde, J., Frankish, A., Guigó, R., and Johnson, R. (2018). Towards a complete map of the human long non-coding RNA transcriptome. Nat. Rev. Genet. *19*, 535–548.
5. Cao, H., Wahlestedt, C., and Kapranov, P. (2018). Strategies to annotate and characterize long noncoding RNAs: advantages and pitfalls. Trends Genet. *34*, 704–721.
6. Chen, X., Yan, C.C., Zhang, X., and You, Z.-H. (2017). Long non-coding RNAs and complex diseases: from experimental results to computational models. Brief. Bioinform. *18*, 558–576.
7. Wapinski, O., and Chang, H.Y. (2011). Long noncoding RNAs and human disease. Trends Cell Biol. *21*, 354–361.
8. Engreitz, J.M., Haines, J.E., Perez, E.M., Munson, G., Chen, J., Kane, M., McDonel, P.E., Guttman, M., and Lander, E.S. (2016). Local regulation of gene expression by lncRNA promoters, transcription and splicing. Nature *539*, 452–455.
9. Van Bakel, H., Nislow, C., Blencowe, B.J., and Hughes, T.R. (2010). Most "dark matter" transcripts are associated with known genes. PLoS Biol. *8*, e1000371.
10. Statello, L., Guo, C.-J., Chen, L.-L., and Huarte, M. (2021). Gene regulation by long non-coding RNAs and its biological functions. Nat. Rev. Mol. Cell Biol. *22*, 96–118.
11. Volders, P.-J., Anckaert, J., Verheggen, K., Nuytens, J., Martens, L., Mestdagh, P., and Vandesompele, J. (2019). LNCipedia 5: towards a reference set of human long non-coding RNAs. Nucleic Acids Res. *47*, D135–D139.
12. Yang, M., Lu, H., Liu, J., Wu, S., Kim, P., and Zhou, X. (2022). lncRNAfunc: a knowledgebase of lncRNA function in human cancer. Nucleic Acids Res. *50*, D1295–D1306.

13. Bao, Z., Yang, Z., Huang, Z., Zhou, Y., Cui, Q., and Dong, D. (2019). LncRNADisease 2.0: an updated database of long non-coding RNA-associated diseases. Nucleic Acids Res. *47*, D1034–D1037.
14. Cheng, L., Wang, P., Tian, R., Wang, S., Guo, Q., Luo, M., Zhou, W., Liu, G., Jiang, H., and Jiang, Q. (2019). LncRNA2Target v2. 0: a comprehensive database for target genes of lncRNAs in human and mouse. Nucleic Acids Res. *47*, D140–D144.
15. Zhao, H., Shi, J., Zhang, Y., Xie, A., Yu, L., Zhang, C., Lei, J., Xu, H., Leng, Z., Li, T., et al. (2020). LncTarD: A manually-curated database of experimentally-supported functional lncRNA–target regulations in human diseases. Nucleic Acids Res. *48*, D118–D126.
16. Xuan, P., Pan, S., Zhang, T., Liu, Y., and Sun, H. (2019). Graph convolutional network and convolutional neural network based method for predicting lncRNA-disease associations. Cells *8*, 1012.
17. Ma, Y. (2022). DeepMNE: deep multi-network embedding for lncRNA-disease association prediction. IEEE J. Biomed. Health Inform. *26*, 3539–3549.
18. Huang, Y.-A., Huang, Z.-A., You, Z.-H., Zhu, Z., Huang, W.-Z., Guo, J.-X., and Yu, C.-Q. (2019). Predicting lncRNA-miRNA interaction via graph convolution auto-encoder. Front. Genet. *10*, 758.
19. Fukunaga, T., Iwakiri, J., Ono, Y., and Hamada, M. (2019). LncRRIsearch: a web server for lncRNA-RNA interaction prediction integrated with tissue-specific expression and subcellular localization data. Front. Genet. *10*, 462.
20. Zhao, T., Hu, Y., Peng, J., and Cheng, L. (2020). DeepLGP: a novel deep learning method for prioritizing lncRNA target genes. Bioinformatics *36*, 4466–4472.
21. Wang, Y., Chen, S., Li, W., Jiang, R., and Wang, Y. (2020). Associating divergent lncRNAs with target genes by integrating genome sequence, gene expression and chromatin accessibility data. NAR Genom. Bioinform. *2*, lqaa019.
22. Zhou, L., Wang, Z., Tian, X., and Peng, L. (2021). LPI-deepGBDT: a multiple-layer deep framework based on gradient boosting decision trees for lncRNA–protein interaction identification. BMC Bioinf. *22*, 1–24.
23. Gao, M., and Shang, X. (2022). Identification of lncRNA-related protein-coding genes using multi-omics data based on deep learning and matrix completion. In 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (IEEE), pp. 3307–3314.
24. Gao, M., Liu, S., Qi, Y., Guo, X., and Shang, X. (2022). GAE-LGA: integration of multi-omics data with graph autoencoders to identify lncRNA–PCG associations. Brief. Bioinform. *23*, bbac452.
25. Cesana, M., Cacchiarelli, D., Legnini, I., Santini, T., Sthandier, O., Chinappi, M., Tramontano, A., and Bozzoni, I. (2011). A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. Cell *147*, 358–369.
26. Salmena, L., Poliseno, L., Tay, Y., Kats, L., and Pandolfi, P.P. (2011). A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? Cell *146*, 353–358.
27. Grelet, S., Link, L.A., Howley, B., Obellianne, C., Palanisamy, V., Gangaraju, V.K., Diehl, J.A., and Howe, P.H. (2017). A regulated PNUTS mRNA to lncRNA splice switch mediates EMT and tumour progression. Nat. Cell Biol. *19*, 1105–1115.
28. Lewis, B.P., Burge, C.B., and Bartel, D.P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell *120*, 15–20.
29. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583–589.
30. Xie, W., Luo, J., Pan, C., and Liu, Y. (2021). SG-LSTM-FRAME: A computational frame using sequence and geometrical information via LSTM to predict miRNA–gene associations. Brief. Bioinform. *22*, 2032–2042.
31. Shen, C., Luo, J., Ouyang, W., Ding, P., and Chen, X. (2021). IDDkin: network-based influence deep diffusion model for enhancing prediction of kinase inhibitors. Bioinformatics *36*, 5481–5491.
32. Chaudhary, K., Poirion, O.B., Lu, L., and Garmire, L.X. (2018). Deep learning–based multi-omics integration robustly predicts survival in liver cancer. Clin. Cancer Res. *24*, 1248–1259.
33. Buggenthin, F., Buettner, F., Hoppe, P.S., Endele, M., Kroiss, M., Strasser, M., Schwarzfischer, M., Loeffler, D., Kokkaliaris, K.D., Hilsenbeck, O., et al. (2017). Prospective identification of hematopoietic lineage choice by deep learning. Nat. Methods *14*, 403–406.
34. Cao, Y., Geddes, T.A., Yang, J.Y.H., and Yang, P. (2020). Ensemble deep learning in bioinformatics. Nat. Mach. Intell. *2*, 500–508.
35. Ju, C., Bibaut, A., and van der Laan, M. (2018). The relative performance of ensemble methods with deep convolutional neural networks for image classification. J. Appl. Stat. *45*, 2800–2818.
36. Khagi, B., Kwon, G.-R., and Lama, R. (2019). Comparative analysis of Alzheimer's disease classification by CDR level using CNN, feature selection, and machine-learning techniques. Int. J. Imaging Syst. Technol. *29*, 297–310.
37. Zhang, S., Li, X., Lin, Q., and Wong, K.-C. (2019). Synergizing CRISPR/Cas9 off-target predictions for ensemble insights and practical applications. Bioinformatics *35*, 1108–1115.
38. Peng, L., Tan, J., Tian, X., and Zhou, L. (2022). EnANNDeep: an ensemble-based lncRNA–protein interaction prediction framework with adaptive k-nearest neighbor classifier and deep models. Interdiscip. Sci. *14*, 209–232.
39. Xie, W., Zheng, Z., Zhang, W., Huang, L., Lin, Q., and Wong, K.-C. (2022). SRG-vote: Predicting miRNA-gene relationships via embedding and LSTM ensemble. IEEE J. Biomed. Health Inform. *26*, 4335–4344.
40. Long, Y., Wu, M., Liu, Y., Kwoh, C.K., Luo, J., and Li, X. (2020). Ensembling graph attention networks for human microbe–drug association prediction. Bioinformatics *36*, i779–i786.
41. Chen, X., Zhu, Z., Zhang, W., Wang, Y., Wang, F., Yang, J., and Wong, K.-C. (2022). Human disease prediction from microbiome data by multiple feature fusion and deep learning. iScience *25*, 104081.
42. Liu, P., Luo, J., and Chen, X. (2022). miRCom: tensor completion integrating multi-view information to deduce the potential disease-related miRNA-miRNA pairs. IEEE/ACM Trans. Comput. Biol. Bioinform. *19*, 1747–1759.
43. Li, G., Zheng, P., Wang, H., Ai, Y., and Mao, X. (2019). Long non-coding RNA TUG1 modulates proliferation, migration, and invasion of acute myeloid leukemia cells via regulating miR-370-3p/MAPK1/ERK. OncoTargets Ther. *12*, 10375–10388.
44. Zhao, D., Liu, Z., and Zhang, H. (2019). The protective effect of the TUG1/miR-197/MAPK1 axis on lipopolysaccharide-induced podocyte injury. Mol. Med. Rep. *20*, 49–56.
45. Xiao, X., Tan, Z., Jia, M., Zhou, X., Wu, K., Ding, Y., and Li, W. (2021). Long noncoding RNA SNHG1 knockdown ameliorates apoptosis, oxidative stress and inflammation in models of Parkinson's disease by inhibiting the miR-125b-5p/MAPK1 axis. Neuropsychiatr. Dis. Treat. *17*, 1153–1163.
46. Qian, C., Ye, Y., Mao, H., Yao, L., Sun, X., Wang, B., Zhang, H., Xie, L., Zhang, H., Zhang, Y., et al. (2019). Downregulated lncRNA-SNHG1 enhances autophagy and prevents cell death through the miR-221/222/p27/mTOR pathway in Parkinson's disease. Exp. Cell Res. *384*, 111614.
47. Liao, L., Chen, J., Zhang, C., Guo, Y., Liu, W., Liu, W., Duan, L., Liu, Z., Hu, J., and Lu, J. (2020). Lncrna neat1 promotes high glucose-induced mesangial cell hypertrophy by targeting mir-222-3p/cdkn1b axis. Front. Mol. Biosci. *7*, 627827.
48. Dong, L.-X., Zhang, Y.-Y., Bao, H.-L., Liu, Y., Zhang, G.-W., and An, F.-M. (2021). LncRNA NEAT1 promotes Alzheimer's disease by down regulating micro-27a-3p. Am. J. Transl. Res. *13*, 8885–8896.
49. Ma, T., Li, H., Liu, H., Peng, Y., Lin, T., Deng, Z., Jia, N., Chen, Z., and Wang, P. (2022). Neat1 promotes acute kidney injury to chronic kidney disease by facilitating tubular epithelial cells apoptosis via sequestering miR-129-5p. Mol. Ther. *30*, 3313–3332.
50. Durinck, S., Spellman, P.T., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat. Protoc. *4*, 1184–1191.
51. Chen, Z., Zhao, P., Li, F., Marquez-Lago, T.T., Leier, A., Revote, J., Zhu, Y., Powell, D.R., Akutsu, T., Webb, G.I., et al. (2020). iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data. Brief. Bioinform. *21*, 1047–1057.
52. Crooks, G.E., Hon, G., Chandonia, J.-M., and Brenner, S.E. (2004). WebLogo: a sequence logo generator. Genome Res. *14*, 1188–1190.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
LNCipedia	https://lncipedia.org/	https://doi.org/10.25504/FAIRsharing.84c1a7
biomaRT	https://bioconductor.org/packages/release/bioc/html/biomaRt.html	SCR_019214
LncRNA2Target	http://123.59.132.21/lncrna2target	LC/HC
lncTarD	https://lnctard.bio-database.com/	IT
Software and algorithms		
LncRNA-Top	This study (http://lncrna.cs.cityu.edu.hk/) (https://github.com/Xshelton/LncRNA-TOP)	LncRNA-Top
iLearn Platform	https://github.com/Superzchen/iLearn	iLearn
DeepLGP	https://github.com/zty2009/LncRNA-target-gene	DeepLGP
LPI-deepGBDT	GitHub - plhhnu/LPI-deepGBDT	LPI
GAE-LGA	GitHub - meihonggao/GAE-LGA	GAE-LGA

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Dr. Ka-Chun Wong (kc.w@cityu.edu.hk).

Materials availability

This study did not generate new biological data.

Data and code availability

The authors analyze existing and publicly available data. These accession numbers for the datasets are listed in the key resources table.

All original code has been deposited at GitHub and is publicly available as of the date of publication. DOIs are listed in the key resources table.

Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

This paper analyses existing, publicly available data. The study does not use experimental models typical in life sciences.

METHOD DETAILS

Overall workflow

The workflow of lncRNA-Top is illustrated in Figure 6.

The main framework can be divided into four parts: feature generation (a-f), dataset (g-h), ensemble model (i-j), and verifications (k-m). We first extracted sequence-based features of both 3′UTR complementary genes sequence and lncRNAs from the iLearn platform[50] and the k-mer. In Figures 6A and 6B, we visualized the raw sequences by Weblogo[51] and generated the features for each sequence (c-d). Then, we concatenated all generated features and applied kernel PCA (KPCA) with the poly kernel to perform feature transformation and reduce dimension (e). We constructed embedding datasets using known lncRNA and gene relationships as positive and randomly selected pairs as negatives with random seeds, as described in Figure 6H. The deep learning (CNN) and machine learning (RF) models were then trained on constructed embedding datasets. (k-l) Different types of cross-validation, verification among varied independent datasets, random seeds-based ensemble method verification, and metrics exploration. (m) The software would list the top-predicted target and scores by inputting names of genes or lncRNA.

Dataset collection

We considered different dataset types in this manuscript. The usage and purpose of different types of datasets are shown in Table 5. lncRNA sequences were downloaded from LNCipedia,[11] and gene sequences were downloaded from biomaRT′ R package.[52] In detail, we leveraged

the lncRNA sequences and 3′ UTR complementary sequences from genes as input data for feature extraction. The LncRNA2Target v2.0 [14] contains low-throughput and high-throughput datasets. High-throughput datasets contain more genes but fewer lncRNAs than low-throughput datasets. We constructed three independent datasets from the low-throughput data (LC), the high-throughput data (HC), the lncTarD[15] dataset (IT), and the union set of LC and IT (named LCIT). Another independent dataset is manually curated from the literature[10] and named a Case study dataset. We filtered the datasets in advance to include unique lncRNA-gene pairs to avoid AUC/AUPR inflation. The case study dataset overlaps HC, LC, IT, and LCIT differently, as depicted in Figure 6G. Details of the constructed datasets can be found in Table 6.

Negative sampling with random seeds

We preserved the positive pairs and randomly chose the negative pairs to build embedding datasets. The random seed is noted after the dataset. For example, If we use the random seed 'one 'to generate the negative samples for the LC dataset, we remark them as LC_1. For our final ensemble predictor, we trained RF and CNN on the LCIT_1,2,3, tested it on the HC_1,2,3, and verified the top 100 predicted pairs with precision@k.

Feature extraction

We leveraged the iLearn platform[50] and the k-mer for sequence-based features. The iLearn platform is an open-sourced sequence feature calculation platform (software). We can generate a series of sequence-based features by inputting DNA, RNA, or protein sequences into the platform. We first generated all the features that do not need alignment in advance with the default parameters. Those features include Pseudo k-tuple composition (PseKNC), Dinucleotide-based Auto Covariance (DAC), and Series correlation pseudo dinucleotide composition (SCPseDNC) et al. The k-mer is one of the rudimental features for sequence analysis. It contains essential information such as the statistical distribution, palindromic clips, and possible motifs. Taking "3-mer" as an example, assuming we have a sequence of "GTAC," then applying a sliding window to the sequence from right to left, such as "GTA" and "TAC." The count of those "k-mer" is the feature of one sequence. Herein, we set k from three to eight to guarantee variety but not the sparsity of the features. Features details can be found in Table S3.

Kernel principal component analysis

After the feature generation, the total feature dimensionality is exceptionally high (more than 84,000). If we build our machine-learning model directly and integrate all these characteristics, it is not computationally efficient. Thus, we applied feature transformation algorithms. The original feature space is usually non-linear separable. Thus, we applied Kernel Principal Component Analysis (Kernel PCA) with the poly kernel to extract the most variance-explaining features from the initial feature space. Mathematically, we denote the original features as x_i. Applying a high-dimensional non-linear feature transformation ø to the data x_i. We get:

$$x_i \rightarrow \text{ø}(x_i) \tag{1}$$

If we assume v is a linear combination of a high-dimension vector:

$$V = \sum_{i=1}^{n} a_i \, \text{ø}(x_i) \tag{2}$$

Where a_i are learned weights from the component of v. Thus, for a new point x_*, the coefficient is based on the data points' similarity. The kernel PCA coefficient w for v is calculated as:

$$w = \text{ø}(x_*)^{Tv} = \sum_{i=1}^{n} a_i \, \text{ø}(x_*)^T \text{ø}(x_i) = \sum_{i=1}^{n} a_i \, k(x_*, x_i) = k_*^T a \tag{3}$$

Here, k is the p-order polynomial kernel function:

$$k(x, x_i) = [(x \cdot x_i)+1]^p \tag{4}$$

In this study, we selected 4096 features for the final prediction due to their overall good performance among different datasets.

Constructed embedding datasets and metrics

The lncRNA-gene pairs from the original datasets[14,15] were positive samples. The negative samples were randomly selected with random seeds.[20] The number of negative samples is equal to the number of positive samples. We generated the constructed datasets several times to increase the robustness and calculate metrics statistics. The merge of positive and negative samples was regarded as the constructed dataset. We concatenated the lncRNA and gene sequence-based transformed features for each positive/negative lncRNA-gene pair to build embedding datasets and append them with label 1/0. Assuming $F_{i,j}$ as one row of lncRNA i and gene j, and $L_{i,n}$ and $G_{j,n}$ represent the n-th value of embedding, then the $F_{i,j}$ can be represented as:

$$F_{i,j} = \left[L_{i,0}\ ,\ L_{i,1} \ldots L_{i,n}, G_{j,0}, G_{j,1} \ldots G_{j,n}\right], \tag{Equation 5}$$

Those embedding datasets were fed into machine-learning/deep-learning models for training and fine-tuning. The machine learning method distinguishes positive and negative samples by leveraging extracted features.

The metric we leveraged to evaluate model performance is the AUC/AUPR values. We also defined another metric, precision@K, to evaluate the predictive ability. For a model that predicted k unknown pairs, we cut the top k results and searched the pair of keywords in PubMed (https://pubmed.ncbi.nlm.nih.gov/). We denoted a publication containing both keywords as a hit. The precision@K is defined as:

$$\text{precision@k} = \frac{\text{hits}}{\text{k}} \quad \text{(Equtaion 6)}$$

Random forest and convolutional neural network

For our controlled deep-learning strategy, we choose Random Forest (RF) as our basic machine-learning algorithm and the convolutional neural network (CNN) as the deep-learning model. We generated construct datasets via different random seeds and trained an equal number of RF and CNN models between those datasets. The algorithms' details are as follows: An RF implemented bagging to vote from the Decision Tree classifier and aggregate prediction results from different tree classifiers. RF can be directly applied to the transformed dataset, while for CNN, to apply the 2D-convolutional layer, we need to first cut the number of the features into the power of the largest integer. Then, we reshaped the embedding into a squared shape and applied the CNN. For the 2D convolution, assume we have the W to form a 2D filter template, then the filter response h can be calculated by

$$h = f\left(\sum_{x,y} W_{x,y} P_{x,y}\right) \quad \text{(Equation 7)}$$

Where P is an image patch, its size is the same as the size of W. *f*(x) is the activation function, which is applied to each dimension to get the output. We leveraged the 'tanh' function as the activation function for the convolutional layers:

$$\tanh(x) = \frac{\sinh(x)}{\cosh(x)} = \frac{e^{x} - e^{-x}}{e^{x}+e^{-x}} \quad \text{(Equation 8)}$$

For the output layer, the sigmoid function was applied:

$$\sigma(x) = 1 / (1 + e^{-x}) \quad \text{(Equation 9)}$$

For the loss function, we chose categorical cross-entropy.

$$\text{Loss} = -\sum_{i=1}^{\text{output\_size}} y_i \cdot \log \hat{y}_i \quad \text{(Equation 10)}$$

We leveraged the RMSProp (Root Mean Squared Propagation) for the optimizer with a learning rate of 2e-06.

Three types of verification inside each dataset

Most methods may only consider one type of validation. As we want to explore further the relations between different verification methods inside each dataset, we designed three types of verification inside each dataset (Figure 6K). We set another random seed for splitting the dataset. The first type is to cut the dataset to 80% of training and 20% of testing, repeat ten times with different random seeds. The random seed guaranteed that the split could be reproducible and unique. For the type two verification, we divided the dataset into 80%, 10%, and 10% for training, validation, and testing, respectively. We grid-searched the best parameters on the validation set and applied the best model to obtain test scores on the testing set. For the type three verification, we implemented 10-fold cross-validation. Each dataset was split into ten parts for cross-validation, each serving as a different test set. SOTA approaches also employ this technique.[20] In our experiments, each dataset underwent all three verification methods.

Transfer verification

We introduced four independent datasets here. They are LC, IT, HC, and the case_study. The transfer verification means each dataset will be tested in turn. First, we trained RF and CNN on LC, IT, LCIT, and HC, and regarding the rest of the datasets as independent datasets. We recorded the average AUC/AUPR to indicate the predictive transfer ability of each sub-model. As for each independent dataset, we varied with different random seeds. We further explored the random seeds ensemble policy, which leverages different random seeds to generate training sets and ensemble their predictive results to generate the final scores.

Ensemble predictor

We selected the RF as a robust model among discrepant datasets, while CNN can fit for the best performance inside each embedding dataset alone. Applying the ensemble will train three RF and three CNN models via different random seeds generated training sets. Finally,

each pair of lncRNA and gene ensemble would have individual scores derived from several models. They would all be added together to determine the final score. The score and rank for each lncRNA/gene can be retrieved by our software published on our web server.

QUANTIFICATION AND STATISTICAL ANALYSIS

In this manuscript, Mann-Whitney-Wilcoxon tests (M.W.W. in figures, also known as rank-sum tests) were conducted to compare the performance of each feature extraction or prediction model. For cross-validation, each fold's results (AUC/AUPR values) would be regarded as minimal operation elements for statistical tests. The mean value is also calculated to show the overall performance of each model.