BMC Bioinformatics

RESEARCH

Open Access

ESAE-SDA: ensemble sparse autoencoder framework for epigenomics-informed snoRNA-disease associations prediction

Xinqing Jiang[1], Xiaojun Chen[2], Lifeng Xu[1], Feng Zhang[3*], Jiawei Chen[4] and Wenqian Zhang[5*]

*Correspondence:
Feng Zhang
fengzhang@wmu.edu.cn
Wenqian Zhang
wenqian.zhang@hotmail.com
[1]Jinhua Graduate Joint Training Base, Zhejiang Chinese Medical University, Quzhou People's Hospital, Quzhou 324000, China
[2]Department of Endocrine and Metabolic Disease, The Quzhou Affiliated Hospital of Wenzhou Medical University, Quzhou 324000, China
[3]Department of Respiratory and Critical Care, The Quzhou Affiliated Hospital of Wenzhou Medical University, Quzhou 324000, China
[4]Ophthalmology Department, The Third Affiliated Hospital of Wenzhou Medical University, Wenzhou 325000, China
[5]Department of Thoracic Surgery, Beijing Chaoyang Hospital, Capital Medical University, 8th Gongtinanlu Rd, Chaoyang District, Beijing 100020, China

Abstract

Small nucleolar RNAs (snoRNAs), a class of non-coding RNAs broadly distributed in eukaryotes, are emerging as pivotal regulators in the field of epigenomics. In addition to guiding 2′-O-methylation and pseudouridylation modifications at specific rRNA sites to maintain ribosomal stability and support protein synthesis, snoRNAs have been increasingly implicated in epigenetic regulation, influencing gene expression, chromatin architecture, and RNA modification patterns. Accurate identification of potential snoRNA-disease associations (SDAs) is therefore essential for understanding epigenomic dysregulation in complex diseases and facilitating early intervention and drug repurposing. Although artificial intelligence (AI) methods have advanced SDA prediction, they are still hindered by issues such as sample imbalance and high false-negative rates. To address these challenges, we propose ESAE-SDA, a novel model integrating sparse autoencoders with an ensemble learning framework. ESAE-SDA first constructs a comprehensive snoRNA–disease representation using multi-source similarity metrics. It then applies k-means clustering to select high-confidence negative samples and employs a deep sparse autoencoder with sparsity constraints to learn compact, discriminative embeddings. Finally, multiple GNN-based learners are independently trained on dynamically resampled data, and ensemble inference is performed via weighted fusion, substantially enhancing robustness and generalization. Experiments on a public SDA dataset demonstrate that ESAE-SDA consistently outperforms state-of-the-art methods. Notably, a case study on ophthalmic diseases highlights the model's ability to uncover epigenetically relevant snoRNAs with potential regulatory and therapeutic significance, underscoring its value in epigenomics-driven disease research and target discovery.

Keywords Epigenomics, snoRNA-disease associations (SDAs), Ensemble learning framework, Sparse autoencoder, Dynamically sampling, Artificial intelligence (AI)

Introduction

Small nucleolar RNAs (snoRNAs) are a class of non-coding RNAs approximately 60–400 nucleotides in length, predominantly localized in the nucleoli of eukaryotic cells, where they form ribonucleoprotein (snoRNP) complexes with specific proteins to exert their

functions [1]. Beyond their canonical roles in ribosomal RNA modification, snoRNAs have also been implicated in epigenomic regulation, contributing to chromatin remodeling, transcriptional control, and RNA modification landscapes in a disease-specific context. These RNAs are structurally stable and typically classified into two canonical families: C/D box snoRNAs, containing conserved RUGAUGA and CUGA motifs, and H/ACA box snoRNAs, characterized by a double hairpin structure with a terminal ACA motif [2]. While most snoRNAs originate from intronic regions of host genes, a subset arises from long non-coding RNAs (lncRNAs) [2]. Based on sequence and function, snoRNAs are categorized into three types: C/D box snoRNAs guide 2'-O-methylation of rRNA (e.g., SNORD28), H/ACA box snoRNAs catalyse pseudouridylation (e.g., SNORA73), and MRP RNAs mediate 5.8S rRNA cleavage and mitochondrial DNA replication, the latter being unique among snoRNA types [3]. Beyond their classical roles, snoRNAs contribute to a range of regulatory processes, including splicing (e.g., SNORD27 targeting E2F7), mRNA processing (e.g., SNORD50A on Fip1), and protein translocation (e.g., SNORA73 linking mRNAs to 7SL RNA) [4]. They have also been implicated in disease mechanisms. For instance, SNORD126 activates PI3K/AKT signalling in hepatocellular carcinoma, and SNORD46 modulates IL-15 signalling, affecting lipid metabolism and immune responses [3, 5].

snoRNAs exhibit aberrant expression across a wide range of pathological conditions and are implicated in the development of cancer, metabolic, and neurodegenerative diseases, highlighting their potential as diagnostic and prognostic biomarkers. For example, SNORA21 promotes colorectal cancer metastasis, SNORD126 activates the PI3K/AKT pathway to drive hepatocellular carcinoma progression, and knockdown of SNHG1 delays Alzheimer's disease progression. Additionally, SNORA58/68 and exosomal snoRNAs have been explored for non-invasive detection of oesophageal and lung cancers, respectively [3, 5, 6]. Recent technological advances, such as KARR-seq and CRISPR-based screening, have enabled high-resolution characterisation of snoRNA function, while the SNORic database has aggregated large-scale expression profiles to support functional exploration [7]. Although snoRNAs were initially studied for their role in guiding rRNA modifications (e.g., SNORD28 and SNORA73) [3], subsequent findings have linked them to diverse regulatory processes, including post-transcriptional control, RNA splicing, and epigenetic modulation through key pathways such as p53, STAT3, and STING [4]. Notably, SNORD46 has been shown to regulate lipolysis and immune function, underscoring its relevance beyond cancer. Despite improvements in experimental tools, the systematic identification of SDAs remains constrained by high cost and low throughput. Efficient computational approaches are therefore urgently needed. However, existing SDA prediction methods remain scarce and typically depend on extensive manual feature engineering, which hampers generalisation and scalability. Recent advances in machine learning, particularly GNNs, autoencoders, and Transformer-based models, offer promising alternatives for modelling biological networks.

Deep learning has emerged as a transformative approach for large-scale biological data analysis due to its capacity for automated feature learning and nonlinear modelling [8–10]. For example, the PIONEER framework integrates graph convolutional networks (GCNs) and recurrent neural networks (RNNs) to improve the resolution of protein complex interfaces and has successfully identified 586 disease-associated mutations significantly linked to patient survival [11]. In the context of RNA–disease

association, HoRDA employs higher-order graph attention mechanisms to reduce reliance on handcrafted features and uncovers a potential link between SNORD46 and the IL-15 pathway in breast cancer [12]. BigRNA leverages a Transformer-based architecture to predict tissue-specific RNA splicing variants with enhanced disease-resolution capability [13]. In drug–target interaction (DTI) prediction, deep learning has shifted from black-box models to interpretable frameworks. CoaDTI combines GraphSAGE with a Transformer encoder and introduces a bidirectional co-attention mechanism to accurately identify key residues and ligand-binding sites of the SARS-CoV-2 main protease [14]. MHTAN-DTI incorporates metapath semantic information in heterogeneous graphs to improve the recognition of hard-to-predict targets such as GPCRs [15]. DrugSchizoNet constructs a deep learning framework for psychiatric disorders and achieves 98.7% accuracy in predicting novel dopamine receptor ligands, offering new opportunities for drug repositioning [16]. Furthermore, integrating deep learning with network-based approaches has accelerated the understanding of complex disease mechanisms in multi-omics data. For instance, combining WGCNA with machine learning has enabled the identification of key biomarkers in type II diabetes (e.g., APOA4 and IGFBP6) with an AUC of up to 0.963 [17]. The UNAGI model further applies a generative network to reconstruct the dynamic evolution of insulin resistance at the single-cell level [11].

GNNs have emerged as a powerful tool in bioinformatics due to their capacity to efficiently model graph-structured data [18–21]. They have been widely applied to tasks such as molecular interaction prediction [22–26], disease mechanism elucidation [27], and drug-target identification [28]. For instance, ZHMolGraph integrates GNNs with large language models (LLMs) to alleviate annotation imbalance and improve prediction of novel molecules [29]. GNN-ASAPool applies hierarchical clustering to extract plasma protein communities, surpassing traditional markers in Alzheimer's disease diagnosis [29]. PIONEER combines GNNs with RNNs to accurately identify cancer-related protein interaction interfaces and 586 survival-associated mutations, outperforming AlphaFold-Multimer [30]. MINDG introduces higher-order graph attention with multiview enhancement, boosting AUC by over 8% across multiple DTI datasets [31]. MHTAN-DTI enhances GPCR target recognition via heterogeneous graphs and a metapath-aware Transformer module [32]. In dynamic transcriptomic analysis, UNAGI leverages GCNs to reconstruct time-resolved single-cell trajectories and predict drug responses in pulmonary fibrosis [30]. Additionally, GNNs integrated with WGCNA have identified key metabolic hubs such as APOA4 and IGFBP6 in type II diabetes, improving co-diagnosis capabilities [33]. These studies collectively underscore the central role of GNNs in advancing interpretable modelling of complex biological networks. Recent studies have applied GNNs to snoRNA–disease prediction, such as the SAGESDA model based on GraphSAGE, which achieved an AUC of ~ 0.92 on public datasets [34]. For other non-coding RNAs, the K-MGCMLD framework integrating multi-graph contrastive learning and GCNs attained high performance across multiple association types and was validated in lung cancer and Alzheimer's disease [35].

While deep learning methods have achieved promising results in SDA prediction, several limitations remain. Random sampling of negative instances during training often introduces label noise and spurious associations, thereby reducing model performance. Moreover, the common assumption that all unobserved pairs are negative can result in elevated false-negative rates. To address these challenges, we propose a novel framework

that integrates deep sparse autoencoders with ensemble GNNs for accurate SDA prediction. Specifically, we first construct four complementary similarity measures based on diverse biological features to generate comprehensive representations of snoRNA–disease pairs. Known associations are treated as positive samples, while unvalidated pairs are initially labeled as negative. To minimize the impact of false negatives, a k-means clustering strategy is employed to identify high-confidence negative samples with a distribution comparable to the positives. These filtered samples are then input into a deep sparse autoencoder to learn low-dimensional, discriminative representations. In the prediction stage, we adopt an ensemble learning strategy wherein multiple base GNN learners are trained using dynamically sampled negatives. Their outputs are integrated to enhance robustness and reduce the false-negative rate. Systematic evaluations on benchmark datasets demonstrate the superior performance and robustness of our framework, and case studies highlight its utility in disease diagnosis and target discovery, particularly in ophthalmic disorders. Overall, our main contributions are as follows:

(1) A k-means-based strategy is introduced to select representative and low-noise negative samples from unvalidated pairs, effectively mitigating sampling bias and label imbalance.
(2) A deep sparse autoencoder is employed to compress and de-redundancy the fused features, retaining discriminative signals while improving generalizability and reducing overfitting risk in high-dimensional spaces.
(3) An ensemble learning strategy is designed by training multiple base GNN learners with dynamically sampled negative sets and aggregating their predictions to improve accuracy and reduce false-negative rates.
(4) Extensive experiments on public datasets confirm the effectiveness of our framework, and case studies demonstrate its potential for application in disease diagnosis and target discovery, particularly for ophthalmic conditions.

Results

To evaluate the effectiveness of ESAE-SDA, we conducted comprehensive experiments on the RNADisease dataset, including baseline comparisons, ablation studies, and case analyses. The parametric and ablation experiments were designed to assess the model's robustness and to quantify the contribution of individual components. Additionally, two case studies were performed to validate the practical utility of ESAE-SDA in real-world disease scenarios. To ensure a balance between evaluation reliability and computational efficiency, we adopted a tenfold cross-validation strategy. This approach enables effective utilization of limited samples, stabilizes training variance, and controls computational cost. To comprehensively assess model performance on the SDA prediction task, we employed six evaluation metrics: area under the ROC curve (AUC), area under the precision-recall curve (AUPR), recall, precision, accuracy, and F1-score. These indicators collectively provide a multi-dimensional evaluation of the model's ability to distinguish between positive and negative associations.

Performance comparison

To comprehensively evaluate the effectiveness of the proposed ESAE-SDA model, we conducted comparative experiments against six state-of-the-art GNN-based methods on two benchmark datasets: RNADisease [36] and ncRPheno [37]. The selected baselines

include NIMCGCN [12], AMHMDA [38], NSAMDA [39], iPiDA-GCN [40], vGAMF [41], and IGCNSDA [42]. NIMCGCN, originally developed for miRNA-disease association (MDA) prediction, utilizes a dual-channel GCN to extract topological features from miRNA and disease similarity networks and performs inference via matrix completion. AMHMDA incorporates dual GCNs, a hypergraph structure, and a hierarchical attention mechanism to capture higher-order semantic relations and improve node representation learning. NSAMDA integrates a nearest-neighbor strategy with a graph attention network (GAT) to learn representations over heterogeneous networks and predicts associations through inner product decoding. vGAMF adopts a variational graph autoencoder (VGAE) to extract latent features and applies matrix factorization for final prediction. iPiDA-GCN, initially proposed for piRNA-disease association (PDA) prediction, leverages GCNs to capture sequence-level and structural relationships within the interaction network. IGCNSDA is a recent SDA-specific method that applies graph collaborative filtering to infer associations, and currently represents one of the most targeted baselines in this domain. It is noteworthy that although some models (e.g., NIMCGCN, iPiDA-GCN) were originally designed for MDA or PDA prediction, they were adapted in this study to suit the SDA prediction setting by modifying their input features accordingly. For fair and consistent evaluation, all models were trained and tested under the same data split strategy, with identical training, validation, and test sets. This ensures a reliable and comparable assessment of model performance across datasets and evaluation metrics.

The comparative results on the RNADisease dataset demonstrate a clear trend: SDA prediction performance improves progressively with enhancements in model architecture. Early models such as AMHMDA and NIMCGCN performed suboptimally, achieving AUCs of only 65.72% and 72.44% (see Fig. 1(A)), and AUPRs of 68.03% and 66.55% (see Fig. 1(B)), respectively. These outcomes reflect the limitations of conventional GNNs in capturing the sparse and heterogeneous association patterns between

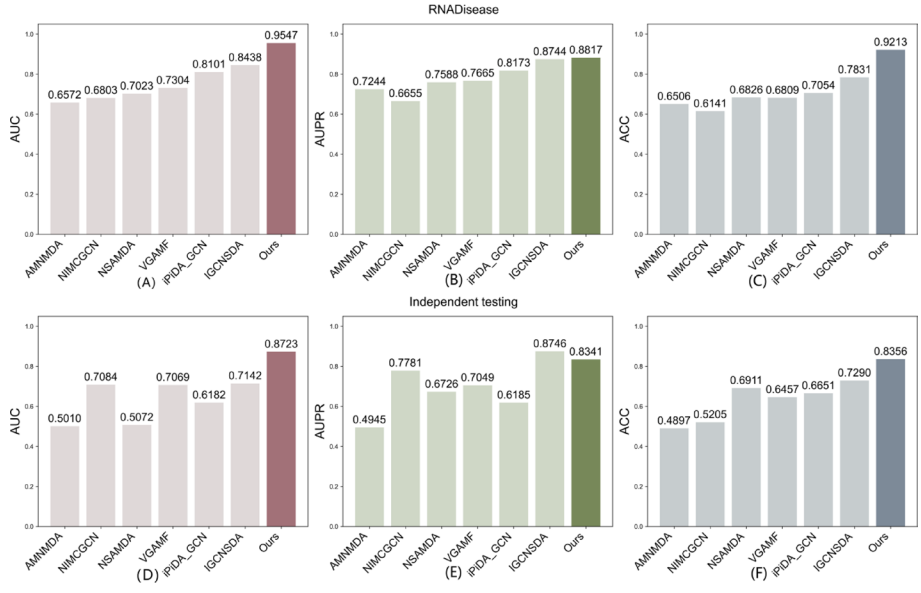

Fig. 1 Comparison of all models in terms of **A** AUC, **B** AUPR, and **C** ACC on the RNADisease dataset, and **D** AUC, **E** AUPR, and **F** ACC on the external test set. And The results of all comparison methods are obtained from this source [42]

snoRNAs and diseases. Subsequent models, NSAMDA and vGAMF, introduced attention mechanisms and variational encoders, which led to modest gains, AUCs improved to 70.23% and 73.04%, respectively—but their accuracy remained around 68%, indicating insufficient robustness in distinguishing positive from negative samples. With the integration of heterogeneous graph structures and enhanced high-order representation learning, iPiDA-GCN and IGCNSDA achieved more substantial improvements. In particular, IGCNSDA reached an AUC of 84.38% (see Fig. 1(A)), AUPR of 87.44% (see Fig. 1(B)), and ACC of 78.31% (see Fig. 1(C)), marking it as the strongest baseline in SDA-specific modeling. In contrast, the proposed ESAE-SDA model delivers a marked performance breakthrough, achieving an AUC of 95.47% and accuracy of 92.13%, substantially outperforming all compared methods. This superior performance is attributed to the deep sparse autoencoder's ability to effectively compress and de-redundancy high-dimensional, multi-source biological features, as well as the optimized high-confidence negative sampling strategy and ensemble learning framework that jointly mitigate the false-negative rate. The results validate the model's generalizability, resistance to overfitting, and its effectiveness in handling label imbalance.

Independent test

Furthermore, to rigorously evaluate the model's generalization, we introduced an independent test set. Specifically, we extracted novel SDAs from the ncRPheno database, removed overlaps with the RNADisease database, and constructed an independent dataset comprising 13 snoRNAs, 82 diseases, and 439 unique SDAs. In our experimental design, the ESAE-SDA model was trained on the complete RNADisease dataset and tested on the ncRPheno dataset as a fully independent test set. This strategy effectively prevents potential information leakage between training and test data, thereby ensuring a fair and rigorous evaluation of the model's generalization.

Results reveal that traditional models such as AMHMDA and NSAMDA suffered severe performance degradation, with AUCs dropping to 50.10% and 50.72% (see Fig. 1(D)), approaching the level of random guessing and highlighting clear overfitting issues. Although vGAMF and iPiDA-GCN showed some improvement, reaching AUCs of 70.69% and 61.82%, their predictive accuracy declined sharply compared to the training phase, suggesting poor adaptability to distributional shifts. IGCNSDA displayed relatively stable performance, with an AUPR of 87.46% (see Fig. 1(E)) and ACC of 72.90% (see Fig. 1(F)), indicating better balance handling between positive and negative samples. However, the proposed ESAE-SDA model continued to outperform all competitors, achieving AUC: 87.23%, AUPR: 83.41%, and Accuracy: 83.56%, with performance highly consistent with training-phase metrics and minimal fluctuation across trials. These results highlight the robustness and transferability of ESAE-SDA in real-world SDA prediction scenarios, including unknown snoRNA–disease pairs and cross-dataset validation, and underscore its potential in disease mechanism analysis and biomarker discovery.

Ablation experiments

To evaluate the contribution of different similarity types in the multi-source similarity fusion module, we conducted a series of ablation experiments. The similarity features were grouped into four categories: functional similarity, Gaussian interaction

Table 1 Ablation study evaluating the impact of similarity feature combinations on SDA prediction

Functional	Gaussian	Cosine	Sigmoid	AUC	AUPR	Recall	Pre	ACC	F1
√				93.14	85.35	88.43	88.21	89.33	82.54
√	√	√	×	95.06	87.99	83.80	88.85	91.07	85.96
√	√	×	√	94.91	88.08	85.06	89.28	91.77	85.89
√	×	√	√	94.17	87.02	80.53	91.00	90.93	85.44
×	√	√	√	94.84	87.71	84.53	89.19	94.84	85.56
√	√	√	√	95.47	88.17	83.34	89.36	92.13	85.12

Table 2 Ten-fold cross-validation performance of the model under different k values

k	AUC	AUPR	Recall	Pre	Acc	F1-Score
10	90.74 ± 0.95	86.76 ± 1.29	80.66 ± 2.21	83.36 ± 1.68	85.82 ± 0.204	80.11 ± 1.63
15	93.70 ± 1.13	87.05 ± 2.54	82.46 ± 2.45	87.37 ± 1.83	90.63 ± 1.45	84.69 ± 2.13
23	95.47 ± 0.66	88.17 ± 1.55	83.34 ± 1.31	89.36 ± 1.52	92.13 ± 0.72	86.12 ± 1.29
30	94.52 ± 0.55	87.82 ± 1.96	83.06 ± 1.69	89.14 ± 2.03	91.72 ± 1.29	85.84 ± 1.78

profile similarity, cosine similarity, and Sigmoid kernel similarity, each computed for both snoRNA and disease nodes to maintain feature consistency across modalities. All ablation settings followed the same training configuration as the full model to ensure fair comparison, and the results are summarized in Table 1.

Among all individual feature types, functional similarity exhibited the strongest discriminative capacity, achieving AUC of 93.14%, AUPR of 85.35%, and F1-score of 82.54%. However, its relatively lower recall and F1-score indicate limitations in recognizing complex or borderline cases. When Gaussian and cosine similarities were added (Functional + Gaussian + Cosine), the AUC increased to 95.06% and F1-score improved to 85.96%, highlighting that incorporating structural and distributional cues can enhance model robustness. In contrast, excluding functional features (Gaussian + Cosine + Sigmoid) led to a noticeable drop in F1-score (85.56%) despite relatively high AUC (94.84%) and AUPR (87.71%), underscoring the irreplaceable role of functional information in identifying critical associations. The full feature combination (All Features) achieved the best overall performance (AUC: 95.47%, AUPR: 88.17%, F1: 85.12%), validating the effectiveness of heterogeneous similarity integration and the synergistic gains from multimodal feature fusion. Interestingly, the combination Functional + Gaussian + Sigmoid yielded a high AUPR (88.08%) but slightly lower accuracy and F1-score, suggesting that different similarity feature subsets contribute variably across performance dimensions. These results indicate that feature design has non-uniform effects on discriminability, sensitivity, and robustness, and that a carefully engineered fusion strategy is essential for optimal SDA prediction.

In summary, the ablation results demonstrate that multi-source similarity fusion significantly improves the model's ability to capture complex snoRNA-disease relationships, enhances stability under heterogeneous data distributions, and provides empirical support for feature engineering in graph-based biomedical prediction tasks.

Parameter experiments

We conducted a sensitivity analysis on the number of k-means clusters to assess the robustness of the model under varying cluster sizes. Experiments were performed on the RNADisease dataset with k = 10, 15, 23, and 30, and the results are summarized in Table 2. Overall, the model demonstrated stable performance across different k values,

indicating insensitivity to cluster number and strong robustness. Specifically, the model achieved AUCs of 90.74% and 93.70% for k = 10 and k = 15, respectively, which were notably lower than those obtained with larger k values. This is because fewer clusters lead to coarser partitioning of unverified pairs, increasing intra-cluster heterogeneity and causing potential positives to be mixed with negatives, thereby introducing noise and reducing recall. At k = 23, the model achieved optimal performance across all metrics (AUC = 95.47%, AUPR = 88.17%, F1-score = 86.12%). This setting balanced granularity and sample size: intra-cluster samples were more homogeneous, and inter-cluster differences more distinct. This improved the reliability of negative samples and enhanced separability between positives and negatives, leading to optimal performance across accuracy, precision, and recall. With k = 30, model performance remained high (AUC = 94.52%, AUPR = 87.82%) but was slightly lower than at k = 23, as small cluster sizes increased sampling uncertainty. In summary, the sensitivity analysis showed that the method maintained stable performance across different clustering settings, with k = 23 providing the optimal balance between granularity and sample size, and thus selected as the default setting in this study.

Statistical significance analysis

To rigorously assess the performance advantages of the proposed model in SDA prediction, we constructed an AUC-based evaluation dataset comprising seven representative models: AMHMDA, NIMCGCN, NSAMDA, vGAMF, iPiDA-GCN, IGCNSDA, and the proposed method. All models were evaluated under consistent tenfold cross-validation settings, and AUC was used as the primary performance metric. To evaluate the overall performance variance across models, we conducted a one-way analysis of variance (one-way ANOVA) [43], which revealed statistically significant differences between models ($p < 0.05$). To further explore pairwise model differences, we applied Tukey's Honest Significant Difference (HSD) test. The results in Fig. 2 indicated that the proposed method significantly outperformed all baseline models, with adjusted p-values < 1e − 7 across all pairwise comparisons. In addition, we visualized the pairwise adjusted p-values using a

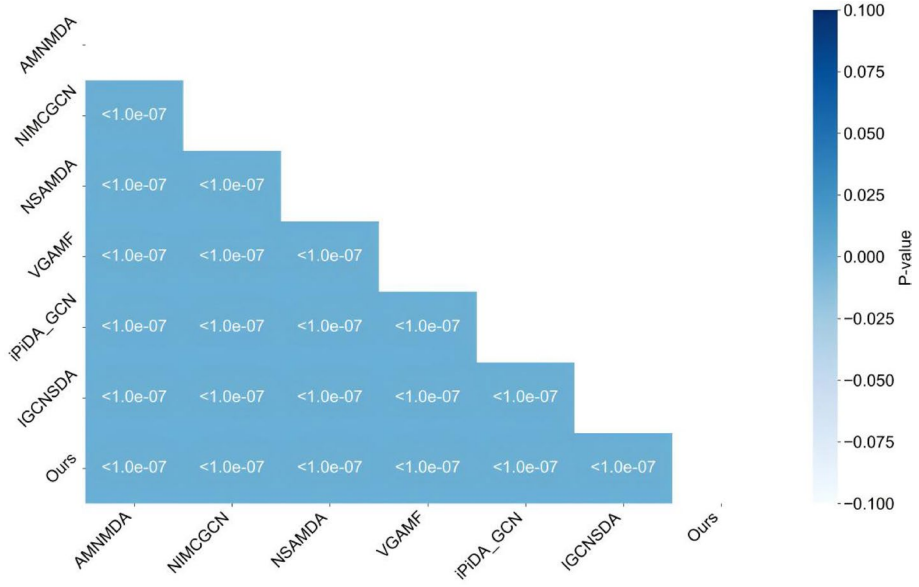

Fig. 2 Statistical significance analysis of multiple methods based on AUC

heatmap, which clearly illustrates the significance distribution across model comparisons. This visualization intuitively supports the conclusion that the proposed method consistently delivers statistically superior performance. Together, these statistical analyses confirm that the improvements achieved by the proposed model in terms of prediction accuracy are not only empirically effective but also statistically significant. The results underscore the model's robustness in feature extraction, representation learning, and classification strategy design, highlighting its competitive advantage in complex SDA modeling tasks.

Case analysis

To further assess the real-world applicability of the proposed ESAE-SDA model, we conducted two case studies focused on Retinoblastoma and SNORA3B, evaluating its ability to predict novel SDAs.

Retinoblastoma is the most common primary intraocular malignancy in children under five years of age, accounting for ~ 3% of all childhood tumors. It originates from immature neuroectodermal retinal cells and is primarily driven by mutations in the RB1 oncogene located on chromosome 13. Approximately 40% of cases are hereditary, typically bilateral or multifocal, while the remaining 60% are non-hereditary, usually unilateral and sporadic. A rare subset (< 2%) is associated with MYCN gene amplification, leading to more aggressive disease even in the absence of RB1 mutations [44]. Treatment strategies are guided by IIRC staging (grades A-E). While eye-preserving therapies (e.g., topical chemotherapy, intravenous agents, or intra-arterial chemotherapy) are preferred in most bilateral or intermediate-stage cases, enucleation remains necessary in advanced or treatment-resistant tumors. For high-risk metastatic patients, high-dose chemotherapy, radiotherapy, and stem cell transplantation are commonly employed.

SNORA3B is a member of the H/ACA box class of small nucleolar RNAs, functioning primarily to guide site-specific pseudouridylation of rRNAs and snRNAs through snoRNP complexes [45]. These modifications are critical for rRNA structural integrity, ribosome biogenesis, and spliceosome assembly. Although no direct evidence linking SNORA3B to diffuse large B-cell lymphoma (DLBCL) currently exists in public databases (e.g., PubMed, NCBI Gene, snoDB), our model identified a high-confidence predictive association between the two. This suggests a potential role for SNORA3B in B-cell malignancy, possibly through regulation of ribosome production, protein synthesis load, apoptosis evasion, or immune escape.

Previous studies have reported aberrant expression of SNORA3B in solid tumors such as pancreatic and hepatocellular carcinoma, implicating its involvement in tumor proliferation, metastasis, and clinical prognosis. These findings further support the hypothesis that SNORA3B may participate in malignant transformation pathways relevant to DLBCL. To validate the model's predictive ability, we performed association inference for lncRNAs related to Retinoblastoma and diseases potentially linked to SNORA3B, using the LncRNADisease v3.0, RNADisease, and ncRPheno databases. To simulate novel discovery scenarios, known associations involving Retinoblastoma or SNORA3B were intentionally excluded from the training phase. During inference, the model computed similarity-based association scores and ranked potential lncRNA-disease and snoRNA-disease pairs.

Table 3 Predicted lncRNAs associated Retinoblastoma and their validation in LncRNADisease v3.0

ID	Symbol	Category	Species	Causality
LDA0006254	XIST	LncRNA	Homo sapiens	Yes
LDA0007164	NEAT1	LncRNA	Homo sapiens	Yes
LDA0005484	hsa_circ_0000527	CircRNA	Homo sapiens	Yes
LDA0001478	MALAT1	LncRNA	Homo sapiens	Yes
LDA0006227	ZFPM2-AS1	LncRNA	Homo sapiens	Yes
LDA0000809	FEZF1-AS1	LncRNA	Homo sapiens	Yes
LDA0007919	MEG3	LncRNA	Homo sapiens	Yes
LDA0005886	UCA1	LncRNA	Homo sapiens	Yes
LDA0005471	hsa_circ_0075804	CircRNA	Homo sapiens	Yes
LDA0007400	NKILA	LncRNA	Homo sapiens	Yes
LDA0000826	FAM238C	LncRNA	Homo sapiens	Yes
LDA0000514	KCNQ1OT1	LncRNA	Homo sapiens	Yes
LDA0001349	LINC-PINT	LncRNA	Homo sapiens	Yes
LDA0022781	CASC8	LncRNA	Homo sapiens	Unknown

Table 4 Predicted diseases associated SNORA3B and validation in RNADisease and ncRPheno

Disease	Causality
Pancreatic ductal adenocarcinoma	Yes
Multiple sclerosis	Yes
Cancer	Yes
Hepatocellular carcinoma	Yes
Diffuse large B-cell lymphoma	Yes

As shown in Table 3, the model retrieved 15 top-ranked lncRNAs potentially associated with Retinoblastoma, among which 14 pairs were confirmed in the LncRNADisease v3.0 database and one remained unreported, demonstrating strong predictive accuracy. For SNORA3B, the model predicted the top 5 disease associations, and all 5 were validated in the RNADisease or ncRPheno databases (see Table 4), indicating the model's ability to generalize to unseen data. These results demonstrate that the proposed ESAE-SDA model not only accurately recovers known associations but also effectively predicts previously unreported snoRNA-disease links, providing valuable clues for future experimental studies. The identified associations may serve as candidate biomarkers or therapeutic targets, especially for early diagnosis and precision treatment in diseases such as Retinoblastoma and DLBCL.

DLBCL is the most common aggressive non-Hodgkin's lymphoma in adults and exhibits marked biological heterogeneity [46]. Traditional diagnosis relies on pathology and immunohistochemistry combined with COO classification (GCB and ABC subtypes) and molecular features (e.g., LME classification, double/triple-hit lymphoma, double-expression lymphoma) to assess prognosis. However, 30–40% of patients relapse or develop refractory disease after first-line R-CHOP treatment, highlighting the limitations of current classification systems and therapeutic strategies [47]. In recent years, small nucleolar RNAs (snoRNAs) have attracted increasing attention for their roles in gene regulation, epigenetic modification, and tumor progression. Dysregulated snoRNA expression is closely associated with DLBCL development, molecular subtypes, and treatment response, suggesting their potential as prognostic biomarkers and therapeutic targets with important implications for precise classification and personalized therapy.

This study investigated snoRNAs associated with DLBCL using the LncRNADisease v3.0, RNADisease, and ncRPheo databases. To evaluate the model's ability to predict novel associations, DLBCL-related samples were intentionally excluded from the training phase. During inference, the model scored and ranked potential snoRNA–DLBCL associations based on snoRNA and disease features, and identified the top 10 candidates with the highest scores. As shown in Table 5, all snoRNA–disease associations from LncRNADisease v3.0 have been reported in the literature, demonstrating the model's strong capability to identify potential snoRNA–disease relationships. In summary, the proposed model successfully predicted key potential associations related to DLBCL, offering new insights into disease mechanisms and identifying candidate targets for targeted intervention strategies. Future work will experimentally validate these predictions and explore their potential for early diagnosis and treatment.

Conclusion

In this study, we propose a deep learning framework, ESAE-SDA, that integrates multi-source information to accurately predict potential associations between snoRNAs and human diseases. The model is systematically optimized across multiple stages—from feature construction to discriminative learning—to address key challenges in SDA prediction, including sample imbalance, high-order node dominance, feature redundancy, and a high false-negative rate. Specifically, ESAE-SDA constructs a multi-source snoRNA–disease representation by integrating four types of similarity metrics: Gaussian kernel similarity, cosine similarity, Sigmoid kernel similarity, and functional semantic similarity, thereby overcoming the limitations of relying on a single feature source. To reduce label noise, high-confidence negative samples are selected via k-means clustering, avoiding the common pitfall of treating all unknown associations as negatives. To enhance feature representation, a deep sparse autoencoder is employed to learn compact and discriminative embeddings, improving the model's ability to capture complex biological relationships. At the prediction stage, ESAE-SDA incorporates a dynamic ensemble learning module. This strategy assigns distinct negative subsets to each base learner to introduce sampling diversity, and aggregates predictions through weighted fusion to improve overall accuracy and sensitivity to rare interactions. In tenfold cross-validation on the RNA-Disease dataset, ESAE-SDA significantly outperformed state-of-the-art baselines (e.g., IGCNSDA with AUC of 84.38%), achieving AUC: 95.47%, AUPR: 88.17%, and Accuracy: 92.13%. On the external ncRPheno dataset, the model maintained strong generalizability (AUC: 87.23%, AUPR: 83.4%), demonstrating robustness across datasets. Furthermore, case studies confirmed the biological relevance of the predictions: after excluding known associations from training, the model successfully identified high-confidence snoRNA–disease associations such as the link between SNORA3B and Retinoblastoma, highlighting its explanatory power and application value.

Table 5 Predicted snoRNAs associated DLBCL and validation in LncRNADisease v3.0

snoRNA	Causality	snoRNA	Causality
SNORA37	Yes	SNORA64	Yes
SNORA13	Yes	SNORA70C	Yes
SNORA67	Yes	SNORA27	Yes
SNORA41	Yes	SNORA71A	Yes
SNORA5A	Yes	SNORA60	Yes

In summary, the proposed ESAE-SDA framework achieves efficient and robust SDA prediction through synergistic mechanisms of multi-source feature fusion, negative sample optimization, sparse representation learning, and ensemble inference. It offers a generalizable paradigm for modeling complex biological networks and holds great promise for extension to other non-coding RNA classes such as lncRNAs and miRNAs. Future work will focus on experimental validation, cross-species prediction, and the clinical translation of model outputs for biomarker discovery and therapeutic target identification.

Materials and methods

Data preparation

This study employed RNADisease v4.0 and ncRPheno databases to evaluate the performance and generalizability of the proposed model. The RNADisease v4.0 database [36] was used as the primary dataset. It contains both experimentally validated and computationally predicted SDAs. After removing duplicate entries and samples with missing values, the final dataset included 471 snoRNAs, 84 diseases, and 1,095 SDAs. This dataset was randomly split into a training set and test set in a 4:1 ratio for cross-validation. The second dataset was obtained from the ncRPheno database [37], which was used as an external test set to assess the model's performance under distribution shift and isolation settings. After data cleaning, the ncRPheno dataset comprised 13 snoRNAs, 82 diseases, and 439 SDA pairs.

Model architecture

The overall framework of the proposed ESAE-SDA model is illustrated in Fig. 3. To construct the input feature space, two curated datasets—RNADisease v4.0 and ncRPheno—were integrated to generate a bimodal similarity matrix between snoRNAs and diseases (see Fig. 3A). For each disease, four types of similarity metrics were calculated: functional semantic similarity (F_DS), Gaussian interaction profile kernel similarity (GIP_DS), cosine similarity (COS_DS), and Sigmoid kernel similarity (SIG_DS). Similarly, snoRNA similarities were defined by functional similarity (F_SS), GIP similarity (GIP_SS), cosine similarity (COS_SS), and Sigmoid similarity (SIG_SS). These heterogeneous similarity matrices were fused into a unified multi-source representation, which served as the initial feature vector for each snoRNA–disease pair. To capture compact and informative representations, a deep sparse autoencoder was employed (see Fig. 3B). It consists of an input layer that receives the fused similarity vector, a hidden layer for nonlinear transformation and compression, and an output layer for reconstruction. The encoder's output is used as the low-dimensional embedding of each association. The training set comprises verified SDAs as positive samples and high-confidence negative samples screened from unverified pairs via k-means clustering, aiming to avoid the bias of treating all unknown associations as negatives and to maintain class balance. To further enhance the model's sensitivity to rare associations, a dynamic ensemble learning strategy was introduced (see Fig. 3C). Multiple base learners were trained in parallel using different subsets of resampled negative samples to encourage diversity. The optimal positive-to-negative sample ratio (1:1.4) was determined via grid search, and final predictions were obtained by weighted fusion of the base learners' outputs. This end-to-end framework integrates

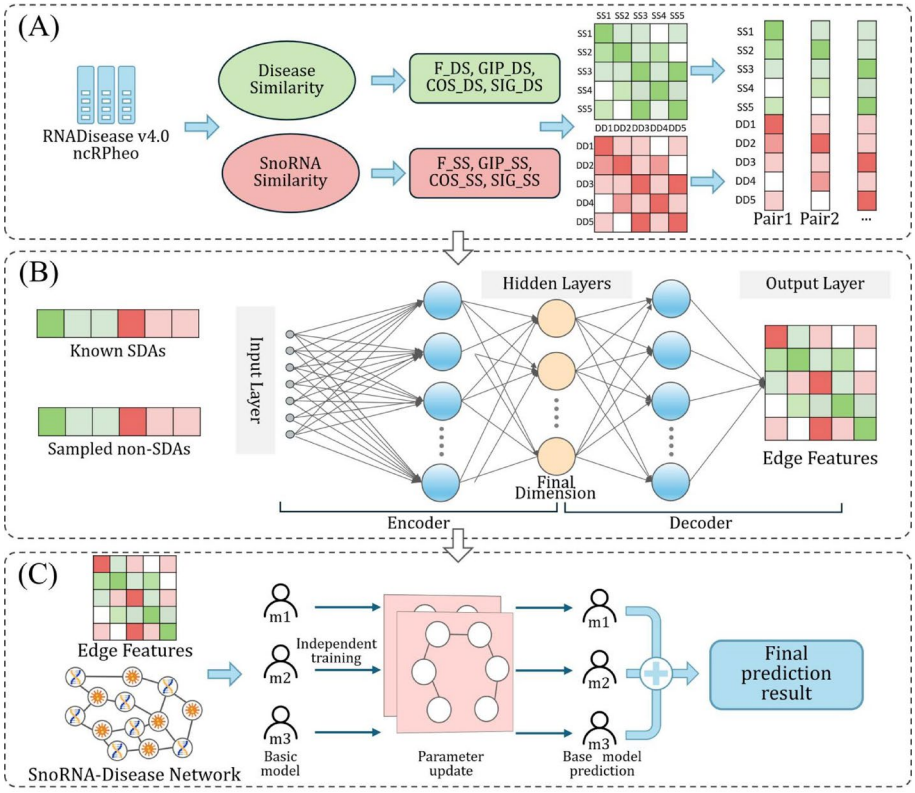

Fig. 3 The ESAE-SDA framework for snoRNA–disease association (SDA) prediction comprises three main components: **A** construction of snoRNA–disease pair representations, **B** representation learning using a sparse autoencoder to encode both verified SDAs and sampled negative pairs, and **C** a dynamic ensemble learning strategy for robust model training and prediction

multi-source similarity fusion, sparse representation learning, and dynamic ensemble inference, enabling robust and generalizable SDA prediction.

Multi-source similarity feature fusion

To improve the predictive accuracy of potential SDAs, this study comprehensively integrates four similarity modeling strategies to characterize the complex relationships between snoRNAs and diseases. First, a disease similarity matrix is constructed based on semantic ontology information, capturing associations from a biological knowledge perspective. Second, functional annotation data are used to compute snoRNA functional similarity, reflecting shared biological roles and potential mechanistic overlaps. In addition to semantic and functional similarities, three interaction profile-based similarity measures are introduced: cosine similarity, Gaussian interaction profile (GIP) kernel similarity, and Sigmoid kernel similarity. These metrics are derived from known snoRNA–disease interaction profiles and provide complementary insights into latent structural patterns. By fusing these heterogeneous sources, a comprehensive high-dimensional representation of snoRNA–disease relationships is constructed, enabling the model to leverage both structural and functional semantics. This multi-source integration is crucial for deep learning-based SDA prediction, as it facilitates the learning of more biologically meaningful and discriminative features.

Specifically, for snoRNA similarity modeling, four metrics—functional similarity (F_SS), Gaussian interaction profile similarity (GIP_SS), cosine similarity (COS_SS), and

Sigmoid kernel similarity (SIG_SS)—are combined to form the integrated similarity matrix SS. The similarity between any two snoRNAs, s_i and s_j, is computed as follows:

$$\mathrm{Sim}\left(\mathrm{s_i},\mathrm{s_j}\right) = \mathrm{w_1}\cdot \mathrm{F\_SS}\left(\mathrm{s_i},\mathrm{s_j}\right) + \mathrm{w_2}\cdot \mathrm{GIP\_SS}\left(\mathrm{s_i},\mathrm{s_j}\right) + \mathrm{w_3}\cdot \mathrm{COS\_SS}\left(\mathrm{s_i},\mathrm{s_j}\right) + \mathrm{w_4}\cdot \mathrm{SIG\_SS}\left(\mathrm{s_i},\mathrm{s_j}\right), \quad (1)$$

where $\boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{w}_3, \boldsymbol{w}_4$ denote the respective fusion weights (uniform or learned). A similar fusion scheme is applied for disease similarity. This integration strategy ensures a more robust and generalizable similarity representation, thereby enhancing the downstream performance of the predictive model.

Similarly, the integrated similarity between two diseases, $\mathrm{d_i}$ and $\mathrm{d_j}$, denoted as $\mathrm{Sim}\left(\mathrm{d_i},\mathrm{d_j}\right)$, is computed by fusing four types of similarity measures:

$$\begin{aligned}\mathrm{Sim}\left(\mathrm{d_i},\mathrm{d_j}\right) = {} & \mathrm{w_1}\cdot \mathrm{F\_DS}\left(\mathrm{d_i},\mathrm{d_j}\right) + \mathrm{w_2}\cdot \mathrm{GIP\_DS}\left(\mathrm{d_i},\mathrm{d_j}\right) \\ & + \mathrm{w_3}\cdot \mathrm{COS\_DS}\left(\mathrm{d_i},\mathrm{d_j}\right) + \mathrm{w_4}\cdot \mathrm{SIG\_DS}\left(\mathrm{d_i},\mathrm{d_j}\right),\end{aligned} \quad (2)$$

Negative sampling

To improve the model's ability to distinguish true SDAs from non-associations, we introduce a k-means-based negative sampling strategy, inspired by the work [48, 49]. Specifically, all unvalidated snoRNA–disease pairs are partitioned into k sub-clusters using the k-means algorithm. From each cluster, a subset of representative samples is randomly selected to serve as negative training instances. This strategy enhances both the representativeness and distributional balance of negative samples in the feature space, thereby reducing the risk of biased learning. To preserve all known positive associations, the positive sample set remains unchanged throughout training. To address class imbalance, the number of negative samples is controlled to approximately match that of positive samples, ensuring balanced supervision and stable gradient updates during model optimization. Based on empirical studies and prior work, the number of clusters k is set to 23, which yields better performance in similar biological association tasks. Overall, this negative sampling strategy significantly improves the quality of the training dataset and provides more informative and discriminative signals for subsequent model learning.

Sparse autoencoder

The sparse autoencoder is an unsupervised neural network architecture that introduces sparsity constraints—typically via Kullback–Leibler (KL) divergence in the loss function—to efficiently learn latent representations. In this study, inspired by previous work [50], we employ a sparse autoencoder to extract informative features for snoRNA–disease pairs. For each snoRNA–disease pair, a high-dimensional joint feature vector is constructed by concatenating the row vector of the corresponding disease from the disease similarity matrix (DD) with that of the snoRNA from the snoRNA similarity matrix ($\boldsymbol{SS}$). Given the high dimensionality and potential redundancy of this concatenation, we further adopt a deep sparse self-coding network to compress these joint features into compact and discriminative low-dimensional embeddings, facilitating efficient and robust downstream prediction.

The sparse autoencoder design follows the bottleneck principle, using a low-dimensional embedding layer to enforce compact and discriminative feature representations. Unlike variational autoencoders (VAEs), which model latent distributions at the bottleneck for generative learning, our approach applies a sparsity constraint (KL divergence,

target $p = 0.05$) at the bottleneck to regulate activation patterns and enhance feature discriminability. Furthermore, the model adopts a symmetric, fully connected architecture with four encoder and four decoder layers. This design makes the model more suitable for snoRNA–disease association prediction than for generative applications. During the encoding phase, the model uses a series of fully connected layers with gradually decreasing neuron counts and Sigmoid activation functions to capture non-linear dependencies and extract intrinsic structures. To enforce sparsity, only a subset of neurons is activated in each layer, promoting feature selectivity and eliminating noise.

The decoder mirrors the encoder in reverse, reconstructing the original input from the compressed representations. Training is guided by reconstruction error minimization, ensuring the preservation of meaningful semantic information. Overall, the sparse autoencoder effectively transforms high-dimensional, heterogeneous snoRNA–disease features into low-dimensional, information-dense representations. This not only reduces computational complexity but also improves feature expressiveness and robustness, thereby enhancing the predictive performance in SDA tasks.

Ensemble learning

Traditional association prediction models often regard all unverified snoRNA–disease pairs as negative samples, which can result in a high false-negative rate and degrade model performance. Although ensemble learning has been applied to alleviate this issue, most existing frameworks suffer from significant overlap in training data across base learners, leading to limited model diversity and restricted performance gains. To address these challenges, we propose a novel ensemble learning framework that incorporates dynamic negative sampling to enhance both robustness and generalizability. By dynamically generating diverse and high-confidence negative samples, this strategy reduces false-negative interference and constructs a differentiated negative sample space, thereby increasing the complementarity among base learners.

The ensemble framework comprises five GNN models—GraphNet1, GraphNet2, GraphNet3, GraphNet4, and GraphNet5—designed to comprehensively capture snoRNA–disease associations. GraphNet1 is a graph transformer network that employs self-attention to overcome the local limitations of traditional GNNs, enabling the capture of long-range dependencies and superior performance on complex graph tasks [51]. GraphNet2 and GraphNet3 are VGNAE models that use L2 normalization to address the limited ability of GAE/VGAE to model isolated nodes [52], thereby improving SDA prediction performance; they share the same architecture but differ in model weights. GraphNet4 and GraphNet5 extend the VGNAE architecture with a global attention mechanism, allowing nodes to dynamically attend to information from all other nodes, thereby enhancing long-range dependency modeling and feature representation. Each base learner was first trained on the complete training set, after which the others were optimized sequentially using four base learners. During inference, a weighted strategy was applied to fuse predictions. Grid search within the range 1:1 to 1:3.5 was used to determine the optimal weight, with 1:1.4 set as the default. This configuration effectively balances sensitivity and specificity, improving the recognition of rare yet meaningful associations.

Overall, this ensemble framework significantly reduces overfitting risk, improves prediction accuracy, and enhances model sensitivity in detecting rare interactions—key requirements for reliable SDA prediction.

AUC optimization strategy

Due to the scarcity of experimentally validated snoRNA-disease pairs, SDA prediction models are highly susceptible to topological noise and cross-context regulatory interference. Although dynamic negative sampling in heterogeneous graphs can effectively expand the training space [53], it tends to aggravate the issue of class imbalance, particularly given the significantly larger number of snoRNAs compared to diseases. AUC quantifies the probability that a randomly selected positive interaction is ranked higher than a negative one, making it well-suited for imbalanced data scenarios. Building on recent advances in imbalanced learning [54], we design an AUC-oriented loss function to mitigate data skewness and improve ranking performance under long-tailed distributions.

Let the heterogeneous graph node space be $\mathcal{X} = \boldsymbol{V_{sno}} \cup \boldsymbol{V_{dis}}$, and the label space be $\mathcal{Y} = \{0, 1\}$. The model learns an embedding function $f_\theta : \mathcal{X} \to \mathbb{R}$, which assigns an association score to each snoRNA–disease pair. The AUC can be defined as:

$$\mathrm{AUC}\left(\mathrm{f}_\theta\right) = \mathrm{P}\left(\mathrm{f}_\theta\left(\mathrm{x}^+\right) > \mathrm{f}_\theta\left(\mathrm{x}^-\right)\right), \tag{3}$$

where x^+ $\mathrm{P}(\mathrm{y}{=}1)$ and x^- $\mathrm{P}(\mathrm{y}{=}0)$ are positive and negative samples, respectively.

To optimize AUC, we employ a stochastic saddle-point optimization framework [30], which introduces average score statistics for positive and negative samples:

$$\hat{\mu}^+ = \mathrm{E}_{\mathrm{x}^+}\left[\mathrm{f}_\theta\left(\mathrm{x}^+\right)\right], \quad \hat{\mu}^- = \mathrm{E}_{\mathrm{x}^-}\left[\mathrm{f}_\theta\left(\mathrm{x}^-\right)\right]. \tag{4}$$

Along with a trade-off hyperparameter λ. Let $\boldsymbol{p} = \mathbb{P}(\boldsymbol{y} = 1)$ be the prior probability of a positive sample. We define the saddle-point objective $\boldsymbol{F}(\boldsymbol{m}, \boldsymbol{d}, \boldsymbol{e}, \boldsymbol{\tau}, \boldsymbol{z})$, where $\boldsymbol{m} \in \mathbb{R}^{\boldsymbol{d}}$ and $\boldsymbol{d}, \boldsymbol{e}, \boldsymbol{\tau}, \boldsymbol{z}$ are hyperparameters for balancing the equilibrium loss, to optimize ranking under imbalanced distributions.

To further enhance robustness and generalizability, we formulate a hybrid loss function by combining the AUC loss with binary cross-entropy (BCE) classification loss:

$$\mathrm{L} = \gamma \cdot \mathrm{L}_{\mathrm{AUC}} + (1 - \gamma) \cdot \mathrm{L}_{\mathrm{BCE}}, \tag{5}$$

where $\gamma \in [0, 1]$ controls the balance between global ranking and local discriminative structure.

This hybrid strategy effectively improves the model's ability to recognize infrequent but strong associations—such as those between snoRNAs and rare or orphan diseases—under long-tailed data distributions, thereby enhancing both precision and robustness in real-world SDA prediction tasks.

Acknowledgements
Not applicable.

Author contributions
Xinging Jiang and Xiaojun Chen conceived and designed the experiments, and wrote the initial draft. Lifeng Xu and Jiawei Chen performed data collection. Feng Zhang and Wenqian Zhang supervised the study and revised the manuscript. All authors reviewed and approved the final version.

Funding
Not applicable.

Data availability
The code and preprocessed datasets used and analysed during the current study are available at: https://github.com/shmildsj/ESAE-SDA-main The original data were sourced from the RNADisease v4.0 and ncRPheno databases, available at http://www.rna-society.org/mndr/ and http://lilab2.sysu.edu.cn/ncrpheno, respectively. The external test dataset has been compiled and made publicly available at: https://github.com/shmildsj/ESAE-SDA-main/tree/master/ESAE-SDA/Data/indepent.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 22 July 2025 / Accepted: 1 October 2025
Published online: 31 October 2025

References

1. Esteller M. Non-coding RNAs in human disease. Nat Rev Genet. 2011;12(12):861–74.
2. Liao J, Yu L, Mei Y, Guarnera M, Shen J, Li R, et al. Small nucleolar RNA signatures as biomarkers for non-small-cell lung cancer. Mol Cancer. 2010;9:1–10.
3. Huang Z-H, Du Y-P, Wen J-T, Lu B-F, Zhao Y. SnoRNAs: functions and mechanisms in biological processes, and roles in tumor pathophysiology. Cell Death Discov. 2022;8(1):259.
4. Liu B, Wu T, Miao BA, Ji F, Liu S, Wang P, et al. snoRNA-facilitated protein secretion revealed by transcriptome-wide snoRNA target identification. Cell. 2025;188(2):465-483.e422.
5. Zhang Y, Zhao Z, Huang LA, Liu Y, Yao J, Sun C, et al. Molecular mechanisms of snoRNA-IL-15 crosstalk in adipocyte lipolysis and NK cell rejuvenation. Cell Metab. 2023;35(8):1457-1473.e1413.
6. Li Y, Chen X, Xiao S, Wang H, Li B, Zhang M, et al. Unlocking the life code: a review of SnoRNA functional diversity and disease relevance. Cell Commun Signal. 2025;23(1):266.
7. Gong J, Li Y, Liu C-J, Xiang Y, Li C, Ye Y, et al. A pan-cancer analysis of the expression and clinical relevance of small nucleolar RNAs in human cancer. Cell Rep. 2017;21(7):1968–81.
8. Wang R, Wang T, Zhuo L, Wei J, Fu X, Zou Q, et al. Diff-AMP: tailored designed antimicrobial peptide framework with all-in-one generation, identification, prediction and optimization. Brief Bioinform. 2024. https://doi.org/10.1093/bib/bbae078.
9. Wei J, Zhu Y, Zhuo L, Liu Y, Fu X, Li F. Efficient deep model ensemble framework for drug-target interaction prediction. J Phys Chem Lett. 2024;15(30):7681–93.
10. Wei J, Zhuo L, Fu X, Zeng X, Wang L, Zou Q, et al. Drugrealign: a multisource prompt framework for drug repurposing based on large language models. BMC Biol. 2024;22(1):226.
11. Xu Z, Zhang L, Lyu J, Cai M, Ji T, Bai L, et al. Integrated clinical and proteomic-based model for diagnostic and prognostic prediction in pRCC. J Hematol Oncol. 2025;18(1):1–5.
12. Li J, Zhang S, Liu T, Ning C, Zhang Z, Zhou W. Neural inductive matrix completion with graph convolutional networks for miRNA-disease association prediction. Bioinformatics. 2020;36(8):2538–46.
13. Celaj A, Gao AJ, Lau TT, Holgersen EM, Lo A, Lodaya V, Cole CB, Denroche RE, Spickett C, and Wagih O. An RNA foundation model enables discovery of disease mechanisms and candidate therapeutics. BioRxiv, 2023. 2023.2009. 2020.558508.
14. Huang L, Lin J, Liu R, Zheng Z, Meng L, Chen X, et al. CoaDTI: multi-modal co-attention based framework for drug–target interaction annotation. Brief Bioinform. 2022;23(6):bbac446.
15. Zhang R, Wang Z, Wang X, Meng Z, Cui W. Mhtan-dti: metapath-based hierarchical transformer and attention network for drug–target interaction prediction. Brief Bioinform. 2023;24(2):bbad079.
16. Sherine Glory J, Durgadevi P, Ezhumalai P. Enhancing drug discovery in schizophrenia: a deep learning approach for accurate drug-target interaction prediction–DrugSchizoNet. Comput Methods Biomech Biomed Eng. 2025;28(2):170–87.
17. Xiang L, Zhou B, Luo Y, Bi H, Lu Y, Zhou J. Hub biomarkers and their clinical relevance in glycometabolic disorders: a comprehensive bioinformatics and machine learning approach. Chin Med J. 2025;138:10–1097.
18. Cai L, He Y, Fu X, Zhuo L, Zou Q, Yao X. AEGNN-M: a 3D graph-spatial co-representation model for molecular property prediction. IEEE J Biomed Health Inf. 2024;29:1726–34.
19. Ma X, Fu X, Wang T, Zhuo L, Zou Q. Graphadt: empowering interpretable predictions of acute dermal toxicity with multi-view graph pooling and structure remapping. Bioinformatics. 2024;40(7):btae438.
20. Wang T, Li Z, Zhuo L, Chen Y, Fu X, Zou Q. MS-BACL: enhancing metabolic stability prediction through bond graph augmentation and contrastive learning. Brief Bioinform. 2024;25(3):bbae127.
21. Wang T, Du Z, Zhuo L, Fu X, Zou Q, Yao X. Multicblo: enhancing predictions of compound-induced inhibition of cardiac ion channels with advanced multimodal learning. Int J Biol Macromol. 2024;276:133825.
22. Zhou Z, Zhuo L, Fu X, Lv J, Zou Q, Qi R. Joint masking and self-supervised strategies for inferring small molecule-miRNA associations. Mol Ther. 2024. https://doi.org/10.1016/j.omtn.2023.102103.
23. Zhou Z, Zhuo L, Fu X, Zou Q. Joint deep autoencoder and subgraph augmentation for inferring microbial responses to drugs. Brief Bioinform. 2023. https://doi.org/10.1093/bib/bbad483.
24. Wang T, Zhuo L, Chen Y, Fu X, Zeng X, Zou Q. ECD-CDGI: an efficient energy-constrained diffusion model for cancer driver gene identification. PLoS Comput Biol. 2024;20(8):e1012400.

25. Zhou Z, Wei J, Liu M, Zhuo L, Fu X, Zou Q. AnomalGRN: deciphering single-cell gene regulation network with graph anomaly detection. BMC Biol. 2025;23(1):73.
26. Ma X, Li Z, Du Z, Xu Y, Chen Y, Zhuo L, et al. Advancing cancer driver gene detection via Schur complement graph augmentation and independent subspace feature extraction. Comput Biol Med. 2024;174:108484.
27. Wang W, Yan Q, Liao Q, Jin X, Gong Y, Zhuo L, et al. Multi-scale information fusion and decoupled representation learning for robust microbe-disease interaction prediction. J Pharm Anal. 2024;15:101134.
28. Zhou Z, Liao Q, Wei J, Zhuo L, Wu X, Fu X, et al. Revisiting drug–protein interaction prediction: a novel global–local perspective. Bioinformatics. 2024;40(5):btae271.
29. Zheng Y, Schupp JC, Adams T, Clair G, Justet A, Ahangari F, Yan X, Hansen P, Carlon M, Cortesi E. A deep generative model for deciphering cellular dynamics and in silico drug discovery in complex diseases. *Nat Biomed Eng*, 2025. 1–26. https://doi.org/10.1038/s41551-025-01423-7
30. Liu H, Jian Y, Zeng C, Zhao Y. RNA-protein interaction prediction using network-guided deep learning. Commun Biol. 2025;8(1):247.
31. Yang H, Chen Y, Zuo Y, Deng Z, Pan X, Shen H-B, et al. MINDG: a drug–target interaction prediction method based on an integrated learning algorithm. Bioinformatics. 2024;40(4):btae147.
32. Han R, Liu X, Pan T, Xu J, Wang X, Lan W, et al. Copra: bridging cross-domain pretrained sequence models with complex structures for protein-rna binding affinity prediction. Proc AAAI Conf Artif Intell. 2025. https://doi.org/10.1609/aaai.v39i1.32001.
33. Zhaori G. Looking back at the past and seeking for greater development: in commemoration of the 120th anniversary of the Chinese Medical Journal. Chin Med J. 2007;120(6):443–6.
34. Momanyi BM, Zhou Y-W, Grace-Mercure BK, Temesgen SA, Basharat A, Ning L, et al. SAGESDA: multi-GraphSAGE networks for predicting snoRNA-disease associations. Curr Res Struct Biol. 2024;7:100122.
35. Sun S-L, Jiang Y-Y, Yang J-P, Xiu Y-H, Bilal A, Long H-X. Predicting noncoding RNA and disease associations using multigraph contrastive learning. Sci Rep. 2025;15(1):230.
36. Chen J, Lin J, Hu Y, Ye M, Yao L, Wu L, et al. RNAdisease v4. 0: an updated resource of RNA-associated diseases, providing RNA-disease analysis, enrichment and prediction. Nucleic Acids Res. 2023;51(D1):D1397–404.
37. Zhang W, Yao G, Wang J, Yang M, Wang J, Zhang H, et al. NcRPheno: a comprehensive database platform for identification and validation of disease related noncoding RNAs. RNA Biol. 2020;17(7):943–55.
38. Ning Q, Zhao Y, Gao J, Chen C, Li X, Li T, et al. AMHMDA: attention aware multi-view similarity networks and hypergraph learning for miRNA–disease associations identification. Brief Bioinform. 2023;24(2):bbad094.
39. Zhao H, Li Z, You Z-H, Nie R, Zhong T. Predicting mirna-disease associations based on neighbor selection graph attention networks. IEEE ACM Trans Comput Biol Bioinform. 2022;20(2):1298–307.
40. Hou J, Wei H, Liu B. iPiDA-GCN: identification of piRNA-disease associations based on graph convolutional network. PLoS Comput Biol. 2022;18(10):e1010671.
41. Ding Y, Lei X, Liao B, Wu F-X. Predicting miRNA-disease associations based on multi-view variational graph auto-encoder with matrix factorization. IEEE J Biomed Health Inform. 2021;26(1):446–57.
42. Hu X, Zhang P, Liu D, Zhang J, Zhang Y, Dong Y, et al. IGCNSDA: unraveling disease-associated snoRNAs with an interpretable graph convolutional network. Brief Bioinform. 2024;25(3):bbae179.
43. St L, Wold S. Analysis of variance (ANOVA). Chemometr Intell Lab Syst. 1989;6(4):259–72.
44. Bao Z, Yang Z, Huang Z, Zhou Y, Cui Q, Dong D. LncRNAdisease 2.0: an updated database of long non-coding RNA-associated diseases. Nucleic Acids Res. 2019;47(D1):D1034–7.
45. Cho Y-S. Decoupled variational graph autoencoder for link prediction. Proceedings of the ACM Web Conference 2024, 2024.
46. Abid J, Al-Rawi MBA, Mahmood A, Li A, Jiang T. Identification and functional characterization of key biomarkers in diffuse large B-cell lymphoma: emphasis on STYX as a prognostic marker and therapeutic target. Hereditas. 2025;162(1):45.
47. Wang H-L, Shang C-Y, Hua W, Yin H, Li Y, Liang J-H, et al. A Novel CD58 Mutation-Related Signature Predicts Prognosis Risk in Diffuse Large B-Cell Lymphoma Patients. Cancer Res Treat. 2025. https://doi.org/10.4143/crt.2024.1143.
48. Cheng H, Wang Z, Wei Z, Ma L, Liu X. On adaptive learning framework for deep weighted sparse autoencoder: a multiobjective evolutionary algorithm. IEEE Trans Cybern. 2020;52(5):3221–31.
49. Peng L-H, Yin J, Zhou L, Liu M-X, Zhao Y. Human microbe-disease association prediction based on adaptive boosting. Front Microbiol. 2018;9:2440.
50. Wang L, Wang Y, Xuan C, Zhang B, Wu H, Gao J. Predicting potential microbe–disease associations based on multi-source features and deep learning. Brief Bioinform. 2023;24(4):bbad255.
51. Yun S, Jeong M, Kim R, Kang J, and Kim HJ. Graph transformer networks. *Adv Neural Inf Process Syst*, 2019. 32 https://proceedings.neurips.cc/paper/2019/hash/9d63484abb477c97640154d40595a3bb-Abstract.html
52. Ahn SJ, and Kim M. Variational graph normalized autoencoders. In: Proceedings of the 30th ACM international conference on information & knowledge management, 2021.
53. Yuan Z, Yan Y, Sonka M, and Yang T. Large-scale robust deep auc maximization: a new surrogate loss and empirical studies on medical image classification. In: Proceedings of the IEEE/CVF international conference on computer vision, 2021.
54. Nemirovski A, Juditsky A, Lan G, Shapiro A. Robust stochastic approximation approach to stochastic programming. SIAM J Optim. 2009;19(4):1574–609.

Publisher's Note