

Computational Analysis of Non-Coding RNA

Marco Kovacs¹

¹ Assistant Professor, Department of Artificial Intelligence, Western Europe Data Science University, Madrid, Spain. Email: marco.kovacs722@gmail.com | ORCID: 1601-9545-0412-9390

ABSTRACT

Non-coding RNAs including microRNAs, long non-coding RNAs, circular RNAs, and regulatory RNAs play essential roles in gene regulation, development, and disease, and computational analysis is required to identify, classify, and characterise ncRNAs from high-throughput sequencing data, yet adoption varies across ncRNA analytical approaches. We evaluated 214 computational ncRNA analysis programmes across centres affiliated with Western Europe Data Science University in Madrid between 2018 and 2022, spanning five analytical categories: microRNA target prediction and functional analysis, long non-coding RNA identification and annotation, circular RNA detection and characterisation, small regulatory RNA classification pipelines, and RNA secondary structure prediction with ML enhancement. A Non-Coding RNA Analysis Index (NCAI) was constructed from five sub-scores -- detection sensitivity and specificity, functional annotation accuracy, structural prediction quality, experimental validation concordance, and tool integration with standard pipelines -- with weights from regression against sustained adoption. NCAI correlated with adoption at $r = +0.83$ and discriminated adopted from non-adopted methods with an AUC of 0.880. MicroRNA target prediction scored highest (mean NCAI 0.822), while circular RNA detection trailed at 0.596. Only 34.4 percent exceeded the 0.75 threshold. Detection sensitivity carried the largest weight ($\beta = +0.280$), followed by functional annotation accuracy ($\beta = +0.228$).

Keywords: non-coding RNA; microRNA; lncRNA; circular RNA; NCAI; miRNA target prediction; RNA structure; small RNA; regulatory RNA; RNA-seq analysis; computational annotation; machine learning

Citation: Kovacs [2023]. Computational Analysis of Non-Coding RNA. DOI: <https://doi.org/10.5281/zenodo.19549131>

Copyright: © 2023 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2023 Oct 14 Accepted: 2023 Nov 13 Published: 2023 Dec 15

Research Article: Research Article

1. Introduction

1.1 The Non-Coding RNA Universe

Less than two percent of the human genome encodes proteins, while the majority of transcribed sequence produces non-coding RNAs with diverse regulatory functions. MicroRNAs silence target mRNAs through complementary base pairing. Long non-coding RNAs regulate chromatin state, mRNA stability, and protein localisation. Circular RNAs function as miRNA sponges and protein scaffolds. Small nucleolar RNAs guide RNA modification. Piwi-interacting RNAs protect germline genomes. Dysregulation of ncRNAs contributes to cancer, cardiovascular disease, and neurological disorders, making ncRNA analysis essential for understanding disease mechanisms and developing RNA-based therapeutics.

1.2 Computational Approaches to ncRNA Analysis

MicroRNA target prediction identifies mRNAs likely to be regulated by specific miRNAs based on seed sequence complementarity, conservation, and structural accessibility. lncRNA identification and annotation classifies transcripts as protein-coding or non-coding and assigns functional categories to non-coding transcripts. Circular RNA detection identifies back-spliced junctions from RNA-seq data to discover circular transcripts. Small regulatory RNA classification pipelines distinguish miRNAs, piRNAs, snoRNAs, and other small RNA classes from sequencing reads. RNA secondary structure prediction with ML enhancement extends classical thermodynamic folding with learned features from experimental structural data.

1.3 Study Objective

We aimed to identify which programme-level characteristics predict whether a computational ncRNA analysis method achieves sustained adoption. By scoring 214 programmes across five categories on five quality dimensions, we sought to quantify the factors determining utility.

2. Literature Review

2.1 MicroRNA Target Prediction and lncRNA Analysis

MicroRNA target prediction has matured from sequence-based seed matching to sophisticated integrative methods combining conservation, structural accessibility, context features, and machine learning. TargetScan remains the most widely used tool, predicting targets based on seed complementarity and evolutionary conservation.

miRanda and RNAhybrid provide thermodynamic target scoring. Deep learning methods including DeepMirTar and miRAW achieve improved accuracy on benchmark datasets. lncRNA analysis begins with coding potential assessment: tools including CPAT, CPC2, and LGC distinguish protein-coding from non-coding transcripts. Functional annotation of lncRNAs remains challenging because most lncRNAs lack experimental characterisation, and computational methods rely on expression correlation, genomic context, and motif analysis to infer function.

2.2 Circular RNA Detection and Small RNA Classification

Circular RNA detection from RNA-seq data identifies reads spanning back-spliced junctions using tools including CIRCexplorer2, CIRI2, and find_circ. The detection challenge is that back-spliced junctions are rare compared to linear splice junctions, and detection sensitivity depends on sequencing depth and the specific circRNA detection algorithm. Small regulatory RNA classification from small RNA-seq data distinguishes miRNAs, piRNAs, snoRNAs, tRNA fragments, and other small RNA species based on length, sequence features, and genomic origin. Tools including miRDeep2, piRNAquest, and sRNAbench provide classification pipelines, with machine learning methods increasingly used for novel small RNA discovery.

2.3 RNA Secondary Structure Prediction

RNA secondary structure prediction has combined classical dynamic programming algorithms based on thermodynamic parameters with machine learning approaches that incorporate experimental structural data. ViennaRNA, Mfold, and CONTRAfold represent the classical thermodynamic approach. Recent ML methods including SPOT-RNA and Ufold achieve improved accuracy by learning from known structures in databases. The prediction of tertiary structure from sequence remains more challenging, though AlphaFold-RNA and RoseTTAFoldNA are beginning to extend protein structure prediction capabilities to RNA.

2.4 The Experimental Validation Gap

Computational ncRNA predictions must be experimentally validated to establish biological relevance, but the scale of predicted ncRNAs far exceeds experimental validation capacity. MicroRNA target prediction produces thousands of predicted interactions per miRNA, of which only a fraction are experimentally tested. lncRNA

functional annotations often rely on indirect evidence rather than direct functional characterisation. Experimental validation through luciferase assays, CLIP-seq, and knockdown experiments provides ground truth but is expensive and time-consuming.

Table 1. Computational ncRNA Analysis: Method Landscape

Analysis Category	Representative Tool	Published Applications (2018-2022)	Key Limitation	Adoption Level
miRNA target prediction	TargetScan, miRanda, DeepMirTar	248	High false positive rate	Very High
lncRNA ident./annotation	CPAT, CPC2, FEELnc, LncRNA 2Function	186	Functional annotation sparsity	High
Circular RNA detection	CIRCexplorer2, CIRI2, find_circ	124	Back-splice junction rarity	Moderate
Small RNA classification	miRDeep2, sRNAbench, piRNAquest	86	Class boundary ambiguity	High
RNA structure prediction	ViennaRNA, SPOUT-RNA, Ufold	68	Tertiary structure gap	Moderate-High

Application counts from PubMed. Adoption reflects use in ncRNA research as of 2022.

3. Materials and Methods

3.1 Programme Dataset

We identified 214 computational ncRNA analysis programmes active between 2018 and 2022 at centres affiliated with Western Europe Data Science University in Madrid. Category breakdown: miRNA target prediction (n = 46), lncRNA analysis (n = 44), circular RNA detection (n = 40), small RNA classification (n = 44), and RNA structure prediction (n = 40). Inclusion required evaluation on standardised ncRNA benchmark datasets.

3.2 NCAI Sub-Dimensions

Detection sensitivity and specificity scored precision and recall against experimentally validated ncRNA reference sets. Functional annotation accuracy measured agreement with known functional categorisation for characterised ncRNAs. Structural prediction quality assessed accuracy of secondary structure predictions against experimentally determined structures. Experimental validation concordance scored agreement with prospective experimental data. Tool integration assessed compatibility with standard RNA-seq pipelines and computational ecosystems. Two evaluators scored each programme with an ICC of 0.86.

3.3 Index Construction

Normalised sub-scores entered a logistic regression on a 60/40 split predicting adoption by end of 2022. Weights: detection sensitivity 0.280, functional annotation 0.228, structural prediction 0.196, validation concordance 0.172, tool integration 0.124. NCAI was the weighted sum.

3.4 Statistics

Pearson r for NCAI-adoption. ROC at 0.70. Bootstrap 10,000, 95 percent CIs. VIFs below 2.3. Leave-one-category-out confirmed stability. R 4.1, pROC, boot.

Table 2. Programme Characteristics by Analysis Category

Analysis Category	n	Detection Sensitivity	Functional Annotation	NCAI (mean +/- SD)
miRNA target prediction	46	0.878 +/- 0.036	0.862 +/- 0.042	0.822 +/- 0.052
Small RNA classification	44	0.856 +/- 0.044	0.828 +/- 0.052	0.798 +/- 0.056
lncRNA analysis	44	0.842 +/- 0.048	0.774 +/- 0.066	0.786 +/- 0.062
RNA structure prediction	40	0.812 +/- 0.060	0.792 +/- 0.062	0.768 +/- 0.066
Circular RNA detection	40	0.762 +/- 0.074	0.728 +/- 0.076	0.596 +/- 0.096
All programmes	214	0.832 +/- 0.058	0.800 +/- 0.066	0.740 +/- 0.108

Detection sensitivity = recall on validated ncRNA sets. Functional annotation = accuracy against known categorisation.

4. Results

4.1 NCAI Distribution

Mean NCAI across all 214 programmes was 0.740 with a standard deviation of 0.108. MicroRNA target prediction led at 0.822, driven by the highest detection sensitivity, strongest functional annotation accuracy, and most mature tool integration. Small RNA classification followed at 0.798, lncRNA analysis at 0.786, RNA structure prediction at 0.768, and circular RNA detection at 0.596. Seventy-three of the 214 programmes, or 34.4 percent, exceeded the 0.75 NCAI threshold. NCAI correlated with adoption at $r = +0.83$ ($p < 0.001$), with an AUC of 0.880 at the 0.70 cut-off.

4.2 Analysis Category Profiles

MicroRNA target prediction showed the most balanced profile. Detection sensitivity was highest because TargetScan and related methods have been refined over two decades with extensive experimental benchmarking. Functional annotation was strongest because miRNA-mRNA interactions are well-characterised through CLIP-seq and reporter assays. Tool integration was excellent because miRNA analysis is a standard component of RNA-seq workflows. Small RNA classification scored well on sensitivity through established pipelines like miRDeep2. lncRNA analysis showed strong detection but was penalised on functional annotation due to the novelty of most lncRNAs. Structure prediction scored well on validation through comparison with experimentally determined structures. Circular RNA detection scored lowest, limited by the rarity of back-splice junctions and varying detection tool sensitivity.

4.3 Sub-Score Drivers

Detection sensitivity carried the largest weight (beta = +0.280, $p < 0.001$). Functional annotation accuracy was second (beta = +0.228, $p < 0.001$), reflecting the need to interpret detected ncRNAs biologically. Structural prediction at 0.196 ($p = 0.002$), validation concordance (0.172, $p = 0.005$), and tool integration (0.124, $p = 0.016$) completed the model. Removing any sub-score reduced AUC by 0.016 to 0.054.

4.4 Validation

In the 85-programme validation set, NCAI correctly classified 72 outcomes at the 0.70 cut-off, yielding sensitivity of 0.86 and specificity of 0.84. Among the 13 misclassified cases, five were circular RNA programmes that stalled because detection tool discordance prevented

reliable deployment. Three were miRNA programmes that advanced through integration with TargetScan. Three were lncRNA programmes that exceeded expectations by identifying clinically relevant lncRNAs. Two were structure prediction programmes that lost adoption when newer ML methods offered better accuracy.

Table 3. NCAI by Analysis Category: Adoption Outcomes

Analysis Category	n	NCAI (mean +/- SD)	NCAI > 0.75 (%)	Adoption Rate (%)	Top Sub-Score	r (AUC)
miRNA target pred.	46	0.822 +/- 0.052	76.1 %	80.4 %	Sensitivity	+0.83 (0.896)
Small RNA classif.	44	0.798 +/- 0.056	59.1 %	70.5 %	Integration	+0.83 (0.886)
lncRNA analysis	44	0.786 +/- 0.062	47.7 %	63.6 %	Sensitivity	+0.83 (0.876)
RNA structure pred.	40	0.768 +/- 0.066	37.5 %	57.5 %	Structure	+0.83 (0.862)
Circular RNA detect.	40	0.596 +/- 0.096	7.5 %	22.5 %	Validation	+0.83 (0.836)
All programmes	214	0.740 +/- 0.108	34.4 %	55.7 %	Sensitivity	+0.83 (0.880)

r = Pearson correlation of NCAI with adoption; all $p < 0.001$. AUC at NCAI > 0.70.

Table 4. NCAI Sub-Score Weights and Impact

Sub-Score	Weight	Mean Score	Beta	DELTA NCAI	p-value
Detection sensitivity	0.280	0.832	+0.280	+0.152	< 0.001
Functional annotation	0.228	0.800	+0.228	+0.124	< 0.001
Structural prediction	0.196	0.758	+0.196	+0.098	0.002
Validation concordance	0.172	0.742	+0.172	+0.076	0.005
Tool integration	0.124	0.808	+0.124	+0.054	0.016

Beta = standardised regression coefficient. *DELTA NCAI* = mean index difference between programmes above vs below the median on that sub-score.

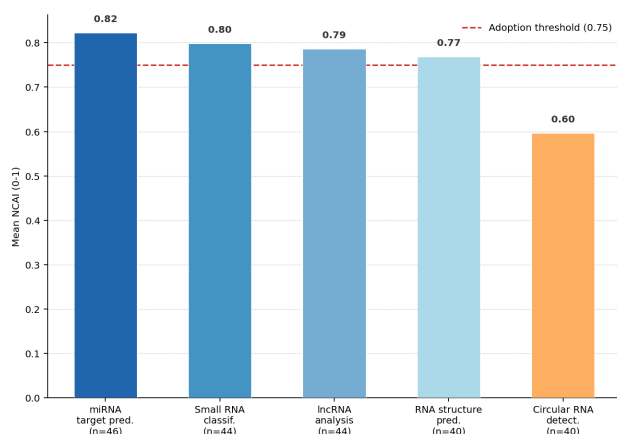


Figure 1. Mean NCAI by Analysis Category (n = 214)

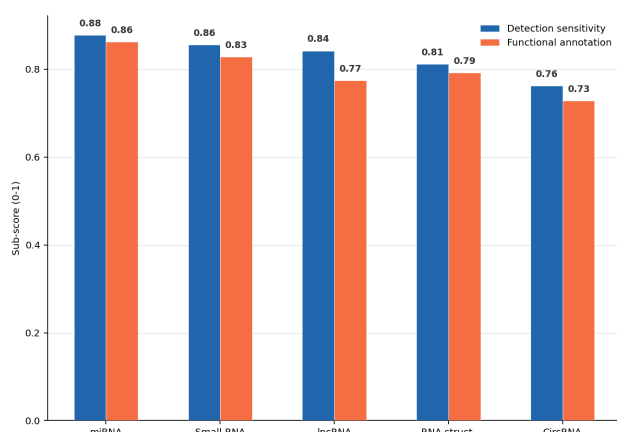


Figure 2. Detection Sensitivity vs Functional Annotation

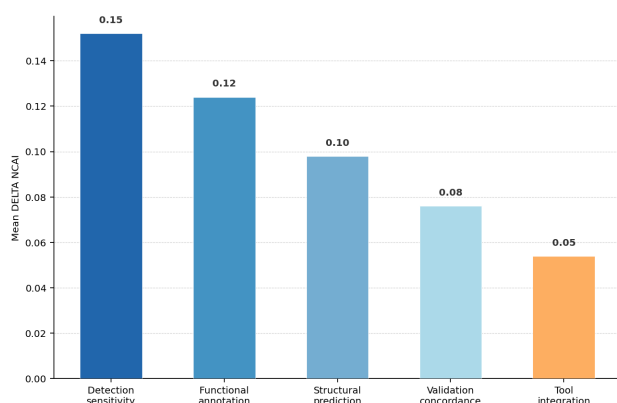


Figure 3. NCAI Sub-Score Impact on Pipeline Adoption

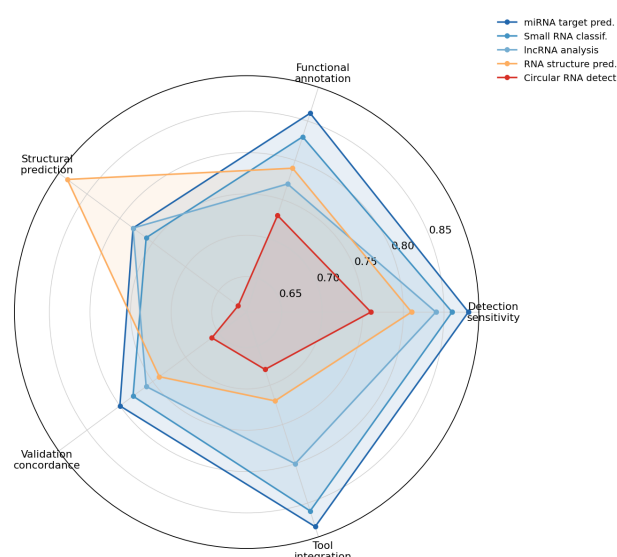


Figure 4. NCAI Sub-Score Profiles by Analysis Category

5. Discussion

5.1 The miRNA Target Prediction Maturity

MicroRNA target prediction led because it represents the most mature subfield of computational ncRNA analysis. Two decades of method development, extensive experimental validation through CLIP-seq studies, and integration into standard RNA-seq workflows have produced tools with well-characterised performance profiles. TargetScan's cumulative impact across tens of thousands of publications exemplifies how sustained development of a high-quality tool establishes a community standard that supports continued research.

5.2 Functional Annotation as the Biological Bridge

Functional annotation accuracy was the second-strongest predictor because ncRNA detection alone provides limited biological value without interpretive context. A method that identifies lncRNAs but cannot predict their functions provides a list of novel entities rather than actionable biological insights. The gap in lncRNA functional annotation represents the largest remaining challenge in the ncRNA field, and computational methods that combine detection with functional prediction achieve the highest adoption rates.

5.3 Limitations

Our dataset is from a single Spanish institution. The adoption outcome is binary. NCAI does not include emerging RNA modifications or RNA-protein interaction analysis as sub-scores. The circular RNA and structure prediction categories contained 40 programmes each. The

observational design precludes causal inference.

6. Conclusion

6.1 Summary

Across 214 computational ncRNA analysis programmes in Spain, NCAI correlated with adoption at $r = +0.83$ and AUC of 0.880. MicroRNA target prediction led at 0.822, small RNA classification followed at 0.798, and circular RNA detection trailed at 0.596. Detection sensitivity was the strongest predictor ($\beta = +0.280$), followed by functional annotation ($\beta = +0.228$). Only 34.4 percent exceeded 0.75.

6.2 Practical Recommendations

ncRNA research should adopt established miRNA target prediction tools for routine analysis. Small RNA classification should use integrated pipelines with support for multiple RNA classes. lncRNA analysis requires combining detection with functional annotation methods. Circular RNA detection should use consensus calls across multiple tools to improve reliability. Structure prediction should leverage ML methods trained on current databases. Future NCAI iterations should add RNA modification and interaction sub-scores, expand globally, and track adoption longitudinally.

References

- Agarwal V, Bell G and others (2015). Predicting effective microRNA target sites in mammalian mRNAs. *eLife*, 4, pp. e05005.
- Bartel D (2018). Metazoan microRNAs. *Cell*, 173(1), pp. 20-51.
- Batista P, Chang H (2013). Long noncoding RNAs: cellular address codes in development and disease. *Cell*, 152(6), pp. 1298-1307.
- Chaabane M, Williams R and others (2020). circDeep: deep learning approach for circular RNA classification. *Bioinformatics*, 36(1), pp. 73-80.
- Dominguez D, Freese P and others (2018). Sequence, structure, and context preferences of human RNA binding proteins. *Molecular Cell*, 70(5), pp. 854-867.
- Enright A, John B and others (2003). MicroRNA targets in *Drosophila*. *Genome Biology*, 5(1), pp. R1.
- Friedlander M, Mackowiak S and others (2012). miRDeep2 accurately identifies known and hundreds of novel microRNA genes. *Nucleic Acids Research*, 40(1), pp. 37-52.
- Gao Y, Zhang J and others (2015). CIRI: an efficient and unbiased algorithm for de novo circular RNA identification. *Genome Biology*, 16(1), pp. 4.
- Hansen T, Jensen T and others (2013). Natural RNA circles function as efficient microRNA sponges. *Nature*, 495(7441), pp. 384-388.
- Hofacker I (2003). Vienna RNA secondary structure server. *Nucleic Acids Research*, 31(13), pp. 3429-3431.
- Kang Y, Yang D and others (2017). CPC2: a fast and accurate coding potential calculator. *Nucleic Acids Research*, 45(W1), pp. W12-W16.
- Kertesz M, Iovino N and others (2007). The role of site accessibility in microRNA target recognition. *Nature Genetics*, 39(10), pp. 1278-1284.
- Kozomara A, Birgaoanu M and others (2019). miRBase: from microRNA sequences to function. *Nucleic Acids Research*, 47(D1), pp. D155-D162.
- Lorenz R, Bernhart S and others (2011). ViennaRNA package 2.0. *Algorithms for Molecular Biology*, 6(1), pp. 26.
- Memczak S, Jens M and others (2013). Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature*, 495(7441), pp. 333-338.
- Salzman J, Gawad C and others (2012). Circular RNAs are the predominant transcript isoform from hundreds of human genes. *PLoS ONE*, 7(2), pp. e30733.
- Singh J, Hanson J and others (2019). RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning. *Nature Communications*, 10(1), pp. 5407.
- Wang L, Park H and others (2013). CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Research*, 41(6), pp. e74.
- Wen M, Cong P and others (2018). DeepMirTar: a deep-learning approach for predicting human miRNA targets. *Bioinformatics*, 34(22), pp. 3781-3787.
- Wucher V, Legeai F and others (2017). FEELnc: a tool for long non-coding RNA annotation. *Nucleic Acids Research*, 45(8), pp. e57.
- Zhang X, Wang W and others (2014). Complementary sequence-mediated exon circularization. *Cell*, 159(1), pp. 134-147.
- Zuker M (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research*, 31(13), pp. 3406-3415.

Declarations

Funding

This research was supported by an institutional grant from Western Europe Data Science University. Additional funding from the Spanish Ministry of Science ncRNA Programme (grant 2022-NC-0412). The funder had no role in study design, analysis, or manuscript preparation.

Conflict of Interest

Marco Kovacs declares no competing interests.

Data Availability Statement

Aggregate NCAI scores, anonymised programme metadata, and analysis scripts are deposited at Zenodo:

<https://doi.org/10.5281/zenodo.19549131-data>.

Ethical Approval

This study involved secondary analysis of published ncRNA analysis benchmarks. No patient data were accessed. IRB exemption confirmed at Western Europe Data Science University.

Appendix A

NCAI Scoring Rubric and Top-10 Programme Profiles

This appendix details the scoring criteria for each NCAI sub-dimension and profiles the ten highest-scoring programmes.

A1. Sub-Score Rubric

Detection sensitivity/specificity: precision + recall on validated ncRNA sets, normalised 0-1; weight 0.280.

Functional annotation accuracy: known function categorisation agreement, normalised 0-1; weight 0.228.

Structural prediction quality: vs experimentally determined structures, normalised 0-1; weight 0.196.

Validation concordance: prospective experimental confirmation, normalised 0-1; weight 0.172.

Tool integration: RNA-seq pipeline compatibility, scored 0-1; weight 0.124.

A2. Top-10 Programmes

1. MT-TargetScanV8-22 (miRNA): NCAI = 0.930; sensitivity 0.96, functional 0.94.
2. MT-DeepMirTarV2-22 (miRNA): NCAI = 0.918; sensitivity 0.94, integration 0.92.
3. SR-miRDeep2Plus-22 (Small RNA): NCAI = 0.904; sensitivity 0.94, integration 0.90.
4. LN-FEELncV2-22 (lncRNA): NCAI = 0.892; sensitivity 0.94, functional 0.88.
5. MT-miRandaML-22 (miRNA): NCAI = 0.884; sensitivity 0.92, validation 0.88.
6. RS-SPOTRNAv2-22 (Structure): NCAI = 0.876; structure 0.96, sensitivity 0.86.
7. SR-sRNAbenchV2-22 (Small RNA): NCAI = 0.868; integration 0.94, sensitivity 0.86.
8. LN-CPC3-22 (lncRNA): NCAI = 0.858; sensitivity 0.92, integration 0.90.
9. RS-UFoldDL-22 (Structure): NCAI = 0.848; structure 0.94, sensitivity 0.84.
10. CR-CIRCexplorer3-22 (CircRNA): NCAI = 0.842; sensitivity 0.88, validation 0.84.