RESEARCH ARTICLE

Meta-2OM: A multi-classifier meta-model for the accurate prediction of RNA 2′-O-methylation sites in human RNA

Md. Harun-Or-Roshid[1], **Nhat Truong Pham**[2], **Balachandran Manavalan**[2]*, **Hiroyuki Kurata**[1]*

[1] Department of Bioscience and Bioinformatics, Kyushu Institute of Technology, Iizuka, Fukuoka, Japan, [2] Department of Integrative Biotechnology, College of Biotechnology and Bioengineering, Sungkyunkwan University, Suwon, Republic of Korea

* kurata@bio.kyutech.ac.jp (HK); bala2022@skku.edu (BM)

OPEN ACCESS

Citation: Harun-Or-Roshid M., Pham NT, Manavalan B, Kurata H (2024) Meta-2OM: A multi-classifier meta-model for the accurate prediction of RNA 2′-O-methylation sites in human RNA. PLoS ONE 19(6): e0305406. https://doi.org/10.1371/journal.pone.0305406

Editor: Shahid Akbar, Abdul Wali Khan University Mardan, PAKISTAN

Received: January 17, 2024

Accepted: May 29, 2024

Published: June 26, 2024

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pone.0305406

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Abstract

2′-O-methylation (2-OM or Nm) is a widespread RNA modification observed in various RNA types like tRNA, mRNA, rRNA, miRNA, piRNA, and snRNA, which plays a crucial role in several biological functional mechanisms and innate immunity. To comprehend its modification mechanisms and potential epigenetic regulation, it is necessary to accurately identify 2-OM sites. However, biological experiments can be tedious, time-consuming, and expensive. Furthermore, currently available computational methods face challenges due to inadequate datasets and limited classification capabilities. To address these challenges, we proposed Meta-2OM, a cutting-edge predictor that can accurately identify 2-OM sites in human RNA. In brief, we applied a meta-learning approach that considered eight conventional machine learning algorithms, including tree-based classifiers and decision boundary-based classifiers, and eighteen different feature encoding algorithms that cover physicochemical, compositional, position-specific and natural language processing information. The predicted probabilities of 2-OM sites from the baseline models are then combined and trained using logistic regression to generate the final prediction. Consequently, Meta-2OM achieved excellent performance in both 5-fold cross-validation training and independent testing, outperforming all existing state-of-the-art methods. Specifically, on the independent test set, Meta-2OM achieved an overall accuracy of 0.870, sensitivity of 0.836, specificity of 0.904, and Matthew's correlation coefficient of 0.743. To facilitate its use, a user-friendly web server and standalone program have been developed and freely available at http://kurata35.bio.kyutech.ac.jp/Meta-2OM and https://github.com/kuratahiroyuki/Meta-2OM.

Introduction

Post-transcriptional chemical modification of ribonucleic acid (RNA) plays significant functions in cellular regulation and biological information in all three phylogenetic domains (i.e. eukaryotes, bacteria, and archaea) of the life [1, 2]. Among various RNA alterations, nucleotide

Funding: This work was supported by a Grant-in-Aid for Scientific Research (B) (22H03688 to H.K.) from the Japan Society for the Promotion of Science (JSPS), Japan and the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (2021R1A2C1014338 to B. M.), Republic of Korea, supported by the Department of Integrative Biotechnology, Sungkyunkwan University (SKKU) and the BK21 FOUR Project.

Competing interests: The authors have declared that no competing interests exist.

modification stands out for its evolutionarily conserved nature within the epitranscriptome [3]. With the advances in genomics and molecular biology, approximately 300 RNA modification types have been identified, playing a crucial role in regulating transcriptional processes [3–5]. One prevalent modification is 2'-*O*-methylation (2-OM), where a methyl group (-CH_3) is added to the 2′ hydroxyl (-OH) of the ribose moiety of a nucleoside. It is also known as the Nm modification of RNA where 'N' stands for any nucleotide. 2-OM nucleotides are predominantly found in ribosomal RNAs (rRNAs), small nuclear/nucleolar RNAs (snRNAs), transfer RNAs (tRNAs), piwi-interacting RNAs (piRNAs), microRNAs (miRNAs), and messenger RNAs (mRNAs) [6–12]. The 2-OM modification alters the RNA activities in different, epigenetic ways such as translation regulation and secondary structure stability, and provides a molecular signature for the discrimination of self and non-self mRNAs [13–16]. Furthermore, 2-OM links to various human diseases [3], including Prader–Willi syndrome, asthma, Alzheimer's disease [13], and breast cancer [17]. Interestingly, 2-OM is also being investigated as a potential target for drugs that could prevent the early innate immune evasion of the SARS-Coronavirus 2 viral RNA [18, 19]. In light of these, accurate detection of 2-OM sites is essential for developing a deeper understanding of their functional implications.

Several experimental methods have been developed for identifying 2-OM sites, both before and after the era of high-throughput sequencing technologies. Some studies have proposed identification based on molecular approaches, with the goal of precisely identifying Nm sites in the ncRNA [19]. Krogh et al. [20], Erales et al. [21], Sharma et al. [22], and Zhou et al. [23] proposed several *in-vitro* techniques to detect the 2-OM sites, but these techniques were only partially effective and were specifically sensitive to p53. RibOxi-seq [24] is another *in-vitro* method that can be used to discriminate 2-OM modifications in rRNA. Dai et al. [12] proposed a high-throughput experimental method named Nm-seq, which is used to detect Nm-modification sites in mRNA. Experimental methods can detect 2-OM sites in the transcriptome, but they are time-consuming, labor-exhaustive, and expensive. Computational methods, on the other hand, are efficient and effective, and they play an important role in many areas of bioinformatics. For example, in silico techniques are being used rapidly in research on disease-gene interactions [25, 26], protein structure prediction [27], peptide therapeutic function, gene editing experiments [28], meaningful pattern detection [29], and drug repurposing [30, 31]. Previously, researchers have been proposed a few computational models for predicting the 2-OM sites based on single machine learning (ML) and deep learning (DL) approaches [32–39].

A summary of these methods is provided in Table 1. Chen et al. [39] developed an support vector machine (SVM)-based prediction model trained on the benchmark dataset of 147 positive and 147 negative samples of Am modifications. Huang et al.'s [34], iRNA-PseKNC (2methyl) [35], iRNA-2OM [37], and NmRF [32] were also built using the same dataset as proposed by Chen et al. Although these methods achieved high prediction performance, they did not generalize well to other types of 2-OM (Gm, Cm, Um) sites. Deep-2'-O-Me [38] is a convolutional neural Network (CNN) with word2vec feature encodings that achieved area under the receiver operating characteristic curves (AUC) of 90% on an independent test for both balanced and unbalanced datasets. NmSEER V2.0 [36] is an updated version of NmSEER [40] that uses random forest (RF) with one-hot, position-specific nucleotide sequence profile (PSNSP), and *K*-nucleotide frequencies (KNF) feature encoding to achieve an AUC of 86.2% for predicting Nm sites. DeepOMe [33] is a CNN-bidirectional long short-term memory (BLSTM) hybrid method that achieved an accuracy (ACC) of 0.956 and AUC of 0.998 on an independent test dataset. NmRF [32] is an RF-based method trained on optimal mixed features to identify 2-OM sites in multiple species. It achieved ACCs of 0.890 and 0.939 for humans and yeast, respectively. The ML-oriented latest method, i2OM [41] is an SVM- and

Table 1. Summary of available methods to predict 2-OM sites.

Year	Method	Algorithm	Feature techniques	Sample size[a]	Testing method	ACC	Genomes
2023	H2Opred [42]	CNN, Bi-GRU	19 encodings	6091/6091	5-fold CV and Independent test	0.858 0.910	Human
2023	i2OM [41]	SVM, XGBoost	K-mer, NCP, ANF	6091/6091	5-fold CV and Independent test	0.863 0.843	Human
2022	NmRF [32]	RF	mixed	147/147	10-fold CV and Independent test	0.891 0.939	Human, Yeast
2021	DeepOMe [33]	CNN, BLSTM	One-hot	3052	10-fold CV	0.956	Human
2020	Huang et al. [34]	SVM	-	147/147	10-fold CV	0.765	Human
2019	iRNA-PseKNC (2methyl) [35]	CNN	One-hot	147/147	5-fold CV	0.983	Human
	NmSEER V2.0 [36]	RF	One-hot, PSDSP, KNF	1989/1989	5-fold CV	0.862	Human
2018	iRNA-2OM [37]	SVM	NCP, NC, PseKNC	147/147	5-fold CV	0.979	Human
	Deep-2'-O-Me [38]	CNN	word2vec	-/-	Independent test	AUC-0.900	Human
2016	Chen et al. [39]	SVM	NCP, NC	147/147	Jacknife CV	0.956	Human

[a]Sample size: positive /negative dataset

https://doi.org/10.1371/journal.pone.0305406.t001

the eXtreme gradient boosting (XGB)-based method that can predict the four types of nucleotide modification in 2-OM sites. It achieved AUCs of 0.920, 0.869, 0.933, and 0.936 for Am, Um, Gm, and Cm, respectively, on an independent test dataset. The DL-based method H2Opred [42] was designed to predict the 2-OM modification for human RNAs. The H2Opred method was trained using both generic and nucleotide-specific datasets. Surprisingly, the generic model outperformed the nucleotide-specific models, which achieved the AUCs of 0.954, 0.949, 0.958, and 0.928 for Am, Cm, Gm, and Gm, respectively, on the independent datasets.

Most previously published methods for predicting 2-OM sites are limited in scope, such as only predicting 2-OM sites in a single type of RNA dataset (e.g., mRNA or rRNA), or a specific type of nucleotide modification sites, which are built based on smaller training datasets. In this study, we developed a more powerful integrated predictor Meta-2OM that predicts all 2-OM sites (Am, Cm, Gm, Um) in human epitranscriptome sequencing data. Fig 1 illustrates the overall framework of Meta-2OM, which is based on a meta-learning approach that explores eight ML classifiers (RF, SVM, XGB, light gradient boosting machine (LGBM), Catboost classifier (CBC), Naïve Bayes (NB), K-nearest neighbor (KNN), Logistic regression (LR)) with 18 diverse RNA sequence-based feature encoding methods. Subsequently, the predicted probabilities of 2-OM sites from multiple baseline models were concatenated and trained with LR to generate the meta-classifier. This approach allows Meta-2OM to learn from the strengths and weaknesses of each model, resulting in a more accurate and robust prediction. Meta-2OM was evaluated on an independent test set and performed slightly better than the state-of-the-art methods, demonstrating its effectiveness as the most effective predictor of 2-OM sites to date.

Materials and methods

Benchmark dataset construction

In this study, we utilized a benchmark dataset specifically designed for the development and evaluation of predictive models for 2-OM RNA modification sites in humans (*Homo sapiens*) RNA sequences. The dataset, originally constructed by Yang et al. [41] was derived from sequence data from RMBase v2.0 [43] and experimental datasets from Nm-seq (GSE90164) [12]. It comprised a total of 7,597 positive samples (2-OM or Nm sites) across various RNA types, including tRNAs, rRNAs, scRNAs, snRNAs, snoRNAs, scaRNAs, lincRNAs, protein-coding genes, and pseudogenes. Each sample was represented by a sequence of 41 nucleotide

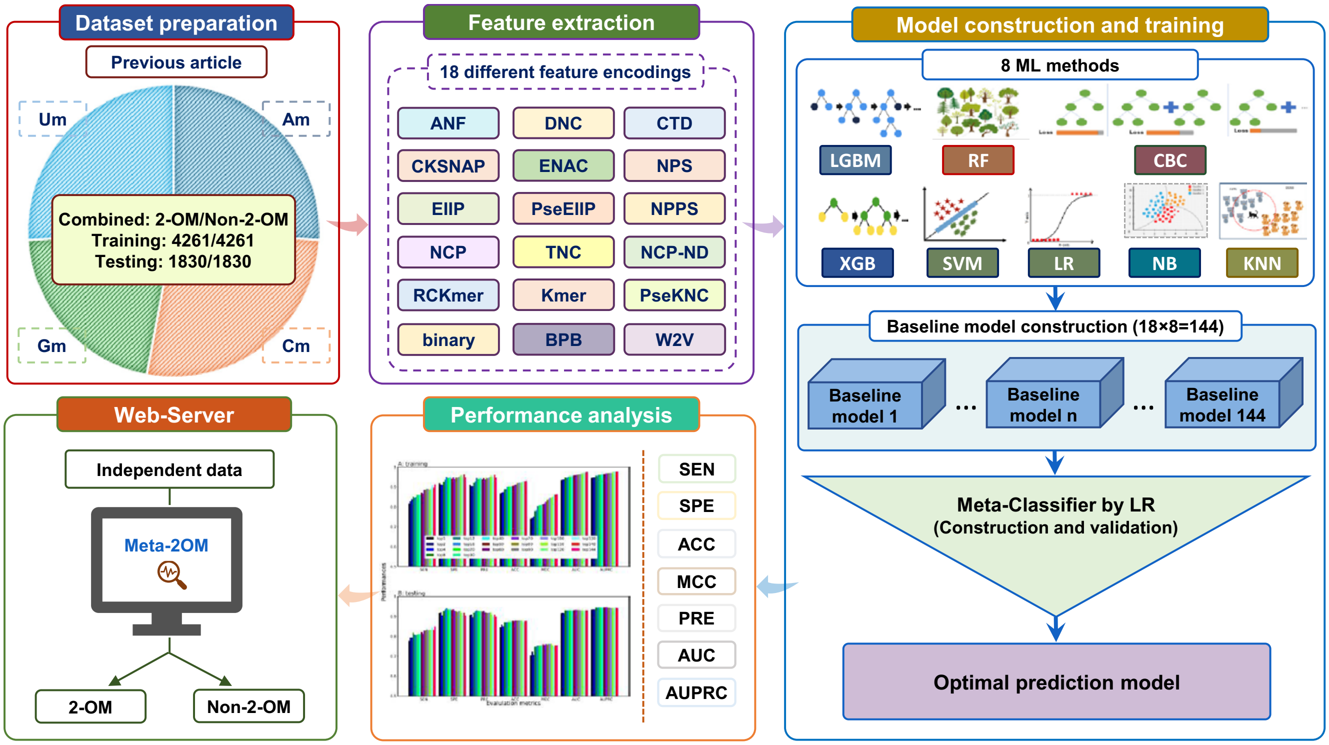


Fig 1. The development workflow of Meta-2OM. It consists of five major steps: (i) Data collection and preparation: collect and prepare 2-OM benchmark datasets from databases and split them into the training and independent test datasets. (ii) Feature extraction: extracting features using 18 different encoding methods. (iii) Model construction and training: constructing the baseline classifiers using eight different ML methods and meta-classifiers using LR for final prediction. (iv) Performance analysis: analyze and compare the performance of the meta-classifier, and (v) Web server construction: constructing a user-friendly web application.

https://doi.org/10.1371/journal.pone.0305406.g001

base pairs (bp), centered around the modification site. To remove sequence redundancy, the CD-HIT [44] was employed with a threshold of 0.80, yielding 6,091 positive samples. The dataset also included 21,520 negative samples, representing non-2-OM or non-Nm sites. They balanced the dataset to a 1:1 ratio of positive to negative samples to minimize training bias associated with imbalanced datasets. As a result, they developed four modification-specific (Am, Cm, Gm, Um) and generic (Nm) dataset, which were then divided into training and testing sets. The statistical summary of these datasets is provided in Table 2.

Encoding and feature extraction strategy

Feature extraction methods are crucial for developing sequence-based ML methods to detect not only RNA modification sites [45, 46], but also other function prediction problems. Therefore, we explore 18 different feature encoding methods, which are briefly described below.

Table 2. Description of benchmark datasets.

Datasets (*H. sapiens*)	Total sample	Training		Independent	
		Positive	Negative	Positive	Negative
Nm (2-OM)	12,182	4,261	4,261	1,830	1,830
Am	3,110	1,088	1,088	467	467
Cm	3,274	1,138	1,138	498	498
Gm	2,620	916	916	394	394
Um	3,196	1,118	1,118	480	480

https://doi.org/10.1371/journal.pone.0305406.t002

Binary. Binary encoding converts a single nucleotide (A, C, G, U) into a four-dimensional binary vector (0, 1) [47]. For example, the nucleotides A, C, G, and U of are represented as (1, 0, 0, 0), (0, 1, 0, 0), (0, 0, 1, 0) and (0, 0, 0, 1), respectively. Therefore, a 41-bp nucleotide sequence can be represented as a 164-dimensional feature vector.

Enhanced Nucleic Acid Composition (ENAC). ENAC encodes sequences based on the local nucleic acid composition (NAC) with a fixed-length window of size k, starting at the 5′ end and moving to the 3′ end. The default k is 5 [48]. For a 41-bp sequence, ENAC produces 148 dimensional feature vector, which is defined as $E = (b_1, b_2, \ldots, b_n)$ and

$$b_j = \frac{N_i}{k}, \; i \in \{A, C, G, U\} \tag{1}$$

where N_i represents the number of i-th nucleotides in j-th window and n is (sequence length − k + 1).

Nucleotide Chemical Property (NCP). NCP converts RNA nucleotides into three-dimensional vectors based on ring structure, functional group, and hydrogen bond properties. NCP encodes the nucleotides, A, C, G, and U as (1, 1, 1), (0, 1, 0), (1, 0, 0), and (0, 0, 1), respectively. For a given 41-bp RNA sequence, NCP produces a 3×41 (= 123)-dimensional feature vector.

Accumulated Nucleotide Frequency (ANF). ANF extracts numeric information from nucleotide sequences by analyzing the distribution of nucleotide types in the RNA [49]. The mathematical function of the ANF feature for a specific nucleotide n_i at the i-th position in the RNA sequence is as follows:

$$d_i = \frac{1}{|N_i|} \sum_{j=1}^{i} f(n_j), \; f(n_j) = \begin{cases} 1 \;\; if \; n_j = n_i \\ 0, otherwise \end{cases} \tag{2}$$

where $|N_i|$ represents the length string up to the i-th position $\{n_1, n_2, \ldots, n_i\}$ in the sequence. Therefore, the ANF converts 41-bp RNA sequences into 41-dimensional feature vectors.

Electron-Ion Interaction Pseudopotentials (EIIP). EIIP encoding converts RNA sequence data into numeric feature vectors by replacing each nucleotide with its corresponding electron energy. The electron energies for A, C, G, and U, are 0.1260, 0.1340, 0.0806 and 0.1335, respectively [50]. This method generates a 64-dimensional feature vector for a sequence. It is computed as follows:

$$D = [EIIP_{AAA} \times f_{AAA}, EIIP_{AAC} \times f_{AAC}, EIIP_{AAG} \times f_{AAG}, \ldots, EIIP_{UUU} \times f_{UUU}] \tag{3}$$

where $EIIP_{mno} = EIIP_m + EIIP_n + EIIP_o$, and $m, n, o \in$ {A, C, G, U}; f_{mno} represents the trinucleotide frequency.

Di-Nucleotide Composition (DNC). DNC encoding generates a 16-dimensional feature vector for an RNA sequence by counting the frequencies of all possible dinucleotide combinations. DNC can be extracted as:

$$f(m, n) = \frac{K_{m,n}}{K - 1}, \; m, n \in \{A, C, G, U\} \tag{4}$$

where $K_{m,n}$ is the dinucleotide combination frequencies by m and n nucleotides.

Tri-Nucleotide Composition (TNC). TNC encoding generates a 64-dimensional feature vector for an RNA sequence by counting the frequencies of all possible trinucleotide combinations. TNC can be expressed as:

$$f(m, n, o) = \frac{K_{m,n,o}}{K - 2}, \; m, n, o \in \{A, C, G, U\} \tag{5}$$

where $K_{m,n,o}$ represent the frequencies of m, n, and o trinucleotide combinations, and m, n, and o can make any set among 'AAA', 'AGA', 'ACA', . . ., and 'UUU'.

Composition of *k*-spaced Nucleic Acid Pairs (CKSNAP). CKSNAP encoding calculates the frequencies of k-spaced nucleotides pairs in an RNA sequence, where k (0, 1, 2, 3, 4, 5) is the distance between nucleotides in the pair [51]. If we consider $k = 0$, then we get 16 k-spaced nucleotide pairs: 'AA', 'AC', 'AG', 'AU', 'CA', 'CC', 'CG', 'CU', 'GA', 'GC', 'GG', GU', 'UA', 'UC', 'UG', 'UU. The CKSNAP encoding can be expressed as follows:

$$\left(\frac{N_{AA}}{N_{total}}, \frac{N_{AC}}{N_{total}}, \frac{N_{AG}}{N_{total}}, \frac{N_{AG}}{N_{total}}, \frac{N_{CA}}{N_{total}}, \ldots, \frac{N_{UU}}{N_{total}}\right)_{16} \tag{6}$$

where N_{AA} represents the total number of nucleotide pairs in the entire RNA sequences and N_{total} is the total number of pairs in the RNA sequences with the gap space k.

Reverse Complement Kmer (RCKmer). RCKmer is a concise variant of Kmer that works by removing reverse complement pairs. In Kmers, at $k = 2$, we obtain 16 Kmer nucleotides (i.e., 'AA', 'AC', 'AG', 'AU', 'CA', 'CC', 'CG', 'CU', 'GA', 'GC', 'GG', GU', 'UA', 'UC', 'UG', 'UU'). However, after removing reverse Kmer (e.g., 'UU' is the reverse complement of 'AA'), we are left with 10 unique Kmers: 'AA', 'AC', 'AG', 'AU', 'CA', 'CC', 'CG', 'GA', 'GC' and 'UA'. These 10 Kmers can be then used to create a feature vector.

Pseudo Electron-Ion Interaction Pseudopotentials (PseEIIP). PseEIIP is a feature encoding method that calculates the average values across trinucleotide sequences to generate a 64-dimensional feature vector, consistent with the dimensionality of the original EIIP model. By focusing on the mean EIIP values, PseEIIP captures the overall electron-ion interaction tendencies within trinucleotide segments, providing a nuanced yet comprehensive feature set.

BPB approach. The Bi-profile Bayes (BPB) feature encoding is a technique that encodes the sequence data using the posterior probability [52]. To calculate the posterior probability of each position of the sequence samples, we consider the following Baye's theorem:

$$P(f^{+}|S) = \frac{P(S|f^{+})P(f^{+})}{P(S)} \tag{7}$$

$$P(f^{-}|S) = \frac{P(S|f^{-})P(f^{-})}{P(S)} \tag{8}$$

where $S = \{s_1, s_2, \cdots, S_L\}$ is the sequence sample, f^+ is the positive dataset, f^- is the negative dataset, $P(f^+)$ is the prior probability, $P(f^{+}|S) = \{p_1^{+}, p_2^{+}, \cdots, p_L^{+}\}$ and $P(f^{-}|S) = \{p_1^{-}, p_2^{-}, \cdots, p_L^{-}\}$ are the posterior probabilities for positive and negative samples, respectively. L is the length of sequences. The BPB method generates a numeric vector for a sequence sample $\mathbf{D}_{\mathrm{BPB}}$ with dimension L×2 by combining their bi-profiles posterior probabilities.

CTD approach. The CTD feature encoding method is based on three sequence information descriptors: C (nucleotide composition), T (nucleotide transition), and D (nucleotide distribution) [53]. Nucleotide composition defines the percentage of each nucleotide within a sequence. Nucleotide transition describes the frequency percentage of transition between the four nucleotides at adjacent positions. The third descriptor, nucleotide distribution was calculated based on five relative positions: 0 (first one), 25%, 50%, 75%, and 100% (last one) for each nucleotide within the sequence. This method results in the generation of a 30-D feature vector.

Kmer. The Kmer encoding method constructs a feature vector based on the nucleic acid composition within segments of k nucleotides. The value of k can vary corresponding mono-, di-, tri-, tetra-, and pentanucleotides. This technique can generate a 4^k dimensional feature vectors. For instance, when $k = 2$, the method generates $4^k = 16$ dimensional feature vector for a given RNA sequence by counting the frequencies of all possible dinucleotide combinations. The mathematical representation is as follows:

$$f(m,n) = \frac{K_{m,n}}{K-1},\ m,n \in \{A,C,G,U\} \tag{9}$$

where $K_{m,n}$ is the dinucleotide combination frequencies by m and n nucleotides. Here, we set k to 4.

NCP-ND approach. The nucleotide chemical properties and nucleotide distribution (NCP-ND) encoding utilizes both the chemical property of nucleotides and their specific distribution at given positions to generate a combined features [54]. In the NCP component, nucleotides A, C, G, and U are encoded as (1, 1, 1), (0, 1, 0), (1, 0, 0), and (0, 0, 1), respectively, based on the properties of ring structure, hydrogen bond, and functional group. The ND calculates the distribution d_i of the i-th nucleotide m_i as follows:

$$d_i = \frac{1}{i}\sum_{j=1}^{L} f(m_j), \quad f(q) = \begin{cases} 1, & if\ m_j = q \\ 0, & otherwise \end{cases} \tag{10}$$

where $q \in \{A,C,G,U\}$ and L is the sequence length. Therefore, each of the sample sequences becomes an $L\times4$ dimensional numerical vector $\mathbf{D}_{\text{NCP\_ND}}$.

NPS approach. The nucleotide pair spectrum (NPS) is a sequence encoding method that generates an occurrence frequency vector of an RNA sequence by counting the k-spaced nucleotide pairs [55]. The k-spaced nucleotide pair $n_1\{k\}n_2$ indicates the nucleotide pair of n_1 and n_2 excluding k-spaces. The occurrence frequency can be generated as follows:

$$F_{n_1\{k\}n_2} = \frac{C(n_1\{k\}n_2)}{L-k-1} \tag{11}$$

where $C(n_1\{k\}n_2)$ is the number of nucleotide pair $n_1\{k\}n_2$ in a sequence window with length $L-k-1$. The parameter k comes from 0 to d_{max}, and $d_{max} = 3$ and each of the RNA sequences can be converted as a numeric vector $\boldsymbol{D}_{NPS}$ with a dimension of $4\times4\times(d_{max}+1)$.

NPPS approach. The nucleotide pair position specificity (NPPS) encoding method quantifies the statistical information based on the position-specific single nucleotide and k-spaced nucleotide pairs [56]. For a given sequence, it calculates the frequency matrix as follows:

$$F_S^+ = \begin{bmatrix} f^+_{S(A,1)} & f^+_{S(A,2)} & \cdots & f^+_{S(A,L)} \\ f^+_{S(C,1)} & f^+_{S(C,2)} & \cdots & f^+_{S(C,L)} \\ f^+_{S(G,1)} & f^+_{S(G,2)} & \cdots & f^+_{S(G,L)} \\ f^+_{S(U,1)} & f^+_{S(U,2)} & \cdots & f^+_{S(U,L)} \end{bmatrix} \tag{12}$$

$$F_d^+ = \begin{bmatrix} f^+_{d(AA,1)} & f^+_{d(AA,2)} & \cdots & f^+_{d(AA,L-k-1)} \\ f^+_{d(AC,1)} & f^+_{d(AC,2)} & \cdots & f^+_{d(AC,L-k-1)} \\ \vdots & \vdots & \ddots & \vdots \\ f^+_{d(UU,1)} & f^+_{d(UU,2)} & \cdots & f^+_{d(UU,L-k-1)} \end{bmatrix} \tag{13}$$

where F_s^+ is the $(4\times L)$ dimensional probability matrix of the single nucleotides occurring in the specific position of positive data samples, F_d^+ is $(16\times(L-k-1))$ dimensional probability matrix of the k-spaced nucleotide pairs occurring in the specific position of positive data sample.

Similarly, the probability matrix F_s^- and F_d^- are generated for the negative data samples. Then the conditional probability between a single nucleotide (A) and k-spaced nucleotide pair (AG) can be calculated as:

$$p_i^+ = \frac{f_{d(AG,i)}^+}{f_{s(G,i+k)}^+} \text{ and } p_i^- = \frac{f_{d(AG,i)}^-}{f_{s(G,i+k)}^-} \tag{14}$$

Finally, the NPPS encoding converts a sequence to an (L–k–1) dimensional vector:

$$\boldsymbol{D}_{NPPS} = P^+ - P^- \tag{15}$$

where $P^+ = [p_1^+, \cdots, p_{L-k-1}^+]$ and $P^- = [p_1^-, \cdots, p_{L-k-1}^-]$. In this study, we set k to 0.

PseKNC approach. Pseudo k-nucleotide component (PseKNC) [57] converts the sequence data into the feature vector using the local and global information of RNAs as follows:

$$D = [d_1, d_2, \cdots, d_{4^k}, d_{4^k+1}, \cdots, d_{4^k+\lambda}]^T \tag{16}$$

where

$$d_u = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{4^k} f_i + w\sum_{j=1}^{\lambda} \theta_j} (1 \le u \le 4^k) \\ \dfrac{w\theta_{u-4^k}}{\sum_{i=1}^{4^k} f_i + w\sum_{j=1}^{\lambda} \theta_j} (4^k \le u \le 4^k + \lambda) \end{cases} \tag{17}$$

d_u (u = 1,2,· · ·,4^k) is the occurrence frequency of k-tuple nucleotide compositions, $\boldsymbol{w}$ is the weight factor, λ is the number of the counted compositions correlated along RNA sequences, and

$$\theta_j = \frac{1}{L-j-1} \sum_{i=1}^{L-j-1} \Theta(R_i R_{i+1}, R_{i+j} R_{i+j+1}), \quad (j = 1, 2, \ldots, \lambda; \lambda < L) \tag{18}$$

$$\Theta\left(R_i R_{i+1}, R_{i+j} R_{i+j+1}\right) = \frac{1}{\mu} \sum_{\upsilon=1}^{\mu} [P_\upsilon(R_i R_{i+1}) - P_\upsilon(R_{i+j} R_{i+j+1})]^2 \tag{19}$$

where μ is the number of RNA physiochemical properties used, $R_i R_{i+1}$ is the i-th position dinucleotide, and $P_\upsilon(R_i R_{i+1})$ is the corresponding standardized value of ν-th RNA local structural property.

Word2vec. Word2vec (W2V), an essential feature embedding method based on natural language processing (NLP), is extensively employed in text data analysis and has proven particularly effective in various bioinformatics pattern recognition tasks involving sequence data [58–61]. W2V utilizes two algorithms: continuous bag-of-words (CBOW) and continuous skip-gram. While CBOW predicts the current words from its context, skip-gram predicts the context from the neighboring words. In this study, we employed the skip-gram algorithm to train a W2V model on 162 RNA sequences retrieved from RNAcentral [62] using the key words "Human", "Rfam", and "non-coding RNA". This resulted in 64-dimensional feature vectors for each nucleotide.

Implemented ML classifiers

We implemented eight ML methods to develop a predictor for identifying 2-OM RNA modification sites. These methods included four tree-based algorithms (RF, XGB, LGBM, and CBC), two decision boundary-based classifiers (SVM, LR), and other classifiers (NB, KNN). It should be noted that all ML models were implemented with the following libraries and packages, including the scikit-learn (https://scikit-learn.org/), XGBoost (https://xgboost.readthedocs.io/), CatBoost (https://catboost.ai/) and lightGBM (https://lightgbm.readthedocs.io/). A brief description of these methods is provided as follows:

RF. RF is one of the most popular supervised ML algorithms, which is broadly used in bioinformatics for wide-range of problems [63–69]. RF combines numerous decision trees trained on the different training samples and predicts the class of new samples based on the majority voting techniques. Thus, the RF method delivers a highly powerful classification decision.

SVM. SVM is a robust supervised learning algorithm that are widely used for classification problems. SVMs work by finding a hyperplane, which is a flat surface in a higher-dimensional space, that best separates the data points into two classes. To improve their ability to classify complex data, SVMs use a kernel function to transform the low-dimensional data into high-dimensional space. This makes SVMs as a powerful tool for bioinformatics applications, where the data if often high-dimensional and complex [45, 68–70].

XGB. The XGB is a popular ensemble-based ML algorithm that is widely used in regression and classification tasks, especially on large datasets, which is common in bioinformatics [41, 45]. XGB achieves high predictive ACC by combining the predictions of multiple individual decision trees, which are trained using a gradient boosting approach. One of the key strengths of XGB is its ability to control overfitting, which is a common problem in ML. XGB does this by using L1 (Lasso regression) and L2 (ridge regression) regularization methods to penalize the weights of the individual decision trees. This helps to reduce the complexity of the ensemble model and improve its robustness. XGB is also designed to efficiently handle sparse datasets, which are common in high-dimensional biological data. XGB's architecture is optimized for scalability and parallel processing, allowing it to perform computations on multiple cores, which significantly reduces memory usage and computational time.

LGBM. LGBM is a highly efficient ML algorithm that is known for its fast training times, high efficiency, low memory usage, and ability to handle large datasets without sacrificing accuracy on a variety of tasks, including classification, regression, and ranking [71]. Like XGB, LGBM is based on decision trees, but it incorporates several advanced computing techniques, such as optimizations for sparse data structures, parallel processing, multiple loss functions, regularization techniques, bagging, and early stopping to prevent overfitting. What particularly sets LGBM apart are two innovative methodologies: Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB). GOSS accelerates the training process by keeping the most informative instances while filtering out less important ones, resulting in more focused and efficient learning. EFB, on the other hand, improves efficiency by grouping mutually exclusive features, effectively reducing the dimensionality of the feature space without significant information loss.

CBC

CBC represents an advanced ML-based algorithm, which is based on gradient-boosted decision trees methodology [72, 73]. This algorithm is particularly effective in handling both classification and regression tasks that incorporate categorical features. CBC employs a combination of ordered boosting, random permutations, and gradient-based optimization to achieve superior classification performance on large and complex datasets featuring categorical variables. Its robustness and efficiency have led to widespread adoption in various bioinformatics domains, notably in RNA modification identification task [74].

NB

A NB classifier is the probabilistic ML model built based on the principles of Bayes' theorem, with the foundational assumption of conditional independence among all pairs of input features. This model is extensively applied across various bioinformatics challenges, especially sequence-based function prediction [75]. Among the different variants of Naive Bayes classifiers, we implemented here the Gaussian Naive Bayes classifier.

KNN

The KNN algorithm is a popular ML technique used for classification and regression tasks. It relies on the idea that similar data points tend to 'K' most similar data points in the training dataset. The homogeneity of different data points is measured by the Euclidean distance method to find the neighbors. Moreover, the right choice of 'K' is crucial to better performance of the KNN method. It is an effective method in classification tasks of bioinformatics fields [76, 77].

LR

LR is a well-known generalized linear ML technique. It is used to classify binary classification problems using datasets without multi-collinearity. LR produces a probability value that lies between 0 to 1 for each data sample, which can be used to predict the likelihood of a sample belonging to one of the two classes. This makes LR a useful technique for bioinformatics applications, such as DNA, RNA, and protein modification site prediction [78, 79]. Additionally, the search range for the hyperparameters of ML classifiers is given in the S1 Table.

Meta-learning approach

We generated baseline models (BMs) by combining eight ML algorithms with 18 feature encoding methods in a one-to-one manner. This resulted in a total of 144 BMs, each of which produced prediction probability scores (PBS) in 5-fold cross-validation (CV) trial. To develop the final meta-prediction model, we used the LR model as a meta-learning method. The meta-predictor generates the final probability scores by analyzing the PBS of BMs. The meta-predictor model is as follows:

$$\log\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n \tag{20}$$

where β_o and β_i are the regression coefficients, X_i is the probability scores generated by the i-th BMs and $n = 144$ is the total number of the BMs. The meta-predictor generates probability scores between 0 and 1, with a score of 0.5 or higher indicating a 2-OM site and a score below 0.5 indicating a non-2-OM site.

Performance evaluation

To assess the performance of the prediction models, we used seven statistical measures: sensitivity (SEN), specificity (SPE), precision (PRE), ACC, Matthew's correlation coefficient (MCC), AUC, and area under the precision-recall curve (AUPRC). Most compatible measures were defined as

$$SEN = \frac{TP}{TP + FN} \tag{21}$$

$$SPE = \frac{TN}{TN + FP} \tag{22}$$

$$PRE = \frac{TP}{TP + FP} \tag{23}$$

$$ACC = \frac{TP + TN}{TP + FN + FP + TN} \tag{24}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TN + FN) \times (TP + FP) \times (TN + FP) \times (TP + FN)}} \tag{25}$$

where TP is the number of true positives, FP is the number of false positives, TN is the number of true negatives, and FN is the number of false negatives. The receiver operating characteristic curve (ROC) is also utilized to visualize the prediction performance of the model according to 5-fold CV.

Results

Nucleotide preference analysis

The nucleotide patterns surrounding the central position in sequence data play a pivotal role in classification tasks. We used the kpLogo command-based tool [80] to perform sequence logo analysis to examine the nucleotide distribution among the 2-OM (positive) and non-2-OM (negative) samples. Fig 2 shows the probability logo (pLogo) and *k*-mer probability logo (*k*pLogo) for k-mer of 1 to 4, highlighting the most significantly enriched or depleted base pairs and sequence motifs, respectively. The pLogo result suggests that the base pairs A, G, A, and U were highly enriched at positions 22, 23, 24, and 25, respectively, for 2-OM sites. Conversely, the base pairs G, A, G, and G are depleted at the same positions 22, 23, 24, and 25, respectively, for non-2-OM sites (Fig 2A). The *k*pLogo results show that the 2-OM samples presented the most significant motifs of ACAG, AGAU, GAUC, AUCG, UCGG, CGG, GGAA, and GAAG, which occurred at positions 20, 22, 23, 24, 25, 26, 27, and 28, respectively (Fig 2B). The depletion motifs were the same as the pLogo findings. Overall, the enrichment and depletion results mostly occurred on the upstream side of positive and negative samples. These results suggested that the distinct nucleotide preferences and unique motifs made it possible to differentiate RNA 2-OM modification/non-modification sites.

Baseline models: Construction and evaluation

ML classifiers effectively leverage feature information from the training dataset to categorize the class. However, the performance of each classifier can vary depending on the feature encodings used, as each encoding has unique characteristics [69, 77]. In this study, we employed eighteen different feature encoding algorithms (Binary, DNC, TNC, RCKmer, ENAC, CKSNAP, ANF, NCP, EIIP, PseEIIP, Kmer, BPB, CTD, NCP-ND, NPS, NPPS, PseKNC, W2V), and assessed their inherent patterns between 2-OM and non-2-OM sites using eight ML classifiers, including four tree-based classifiers (RF, LGBM, XGB, CBC), two decision boundary-based classifier (SVM, LR), and others (NB, KNN). Notably, we trained each model via a 5-fold CV and subsequently evaluated the optimal model using an independent dataset. Firstly, we conducted the performance analysis by using generic datasets. Detailed performance of the 144 baseline models via CV and their transferability on the

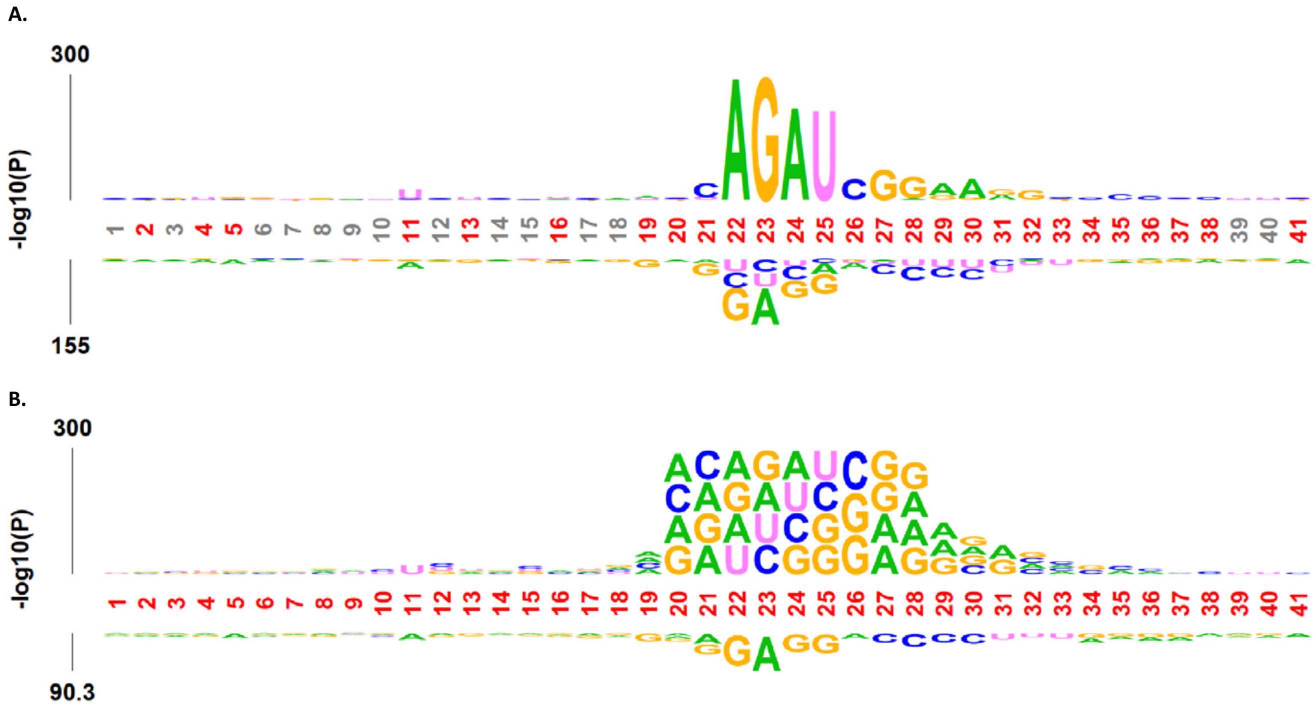
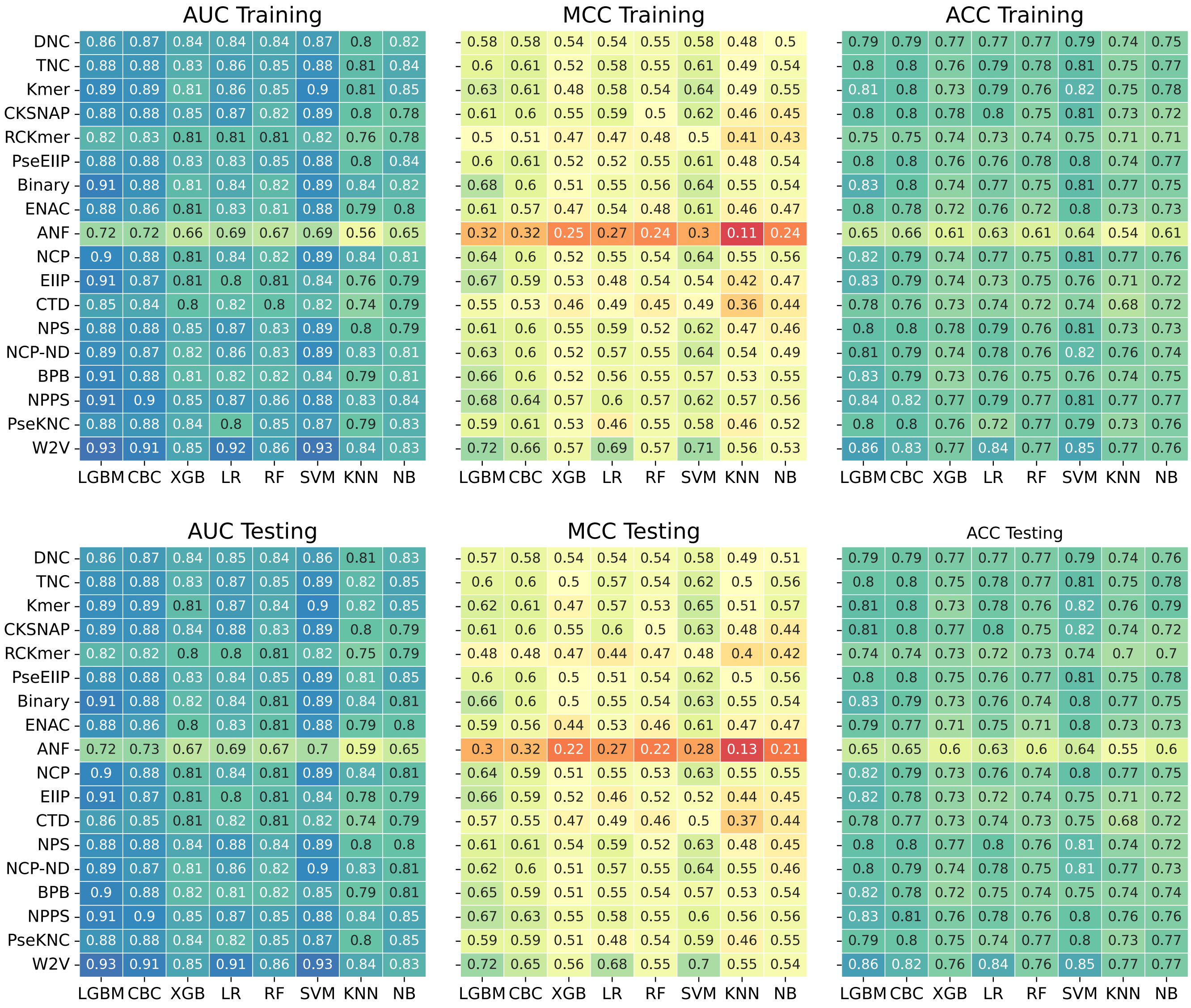


Fig 2. Nucleotide preference pattern analysis of 2-OM RNA modification sites to understand the effect of base pairs and motifs of sequence data. (A) pLogo, and (B) kpLogo.

https://doi.org/10.1371/journal.pone.0305406.g002

independent dataset are shown in S1 and S2 Figs in S1 File and S2 Table. We observed that the overall performance trend of each classifier with 18 different encodings was similar between the training and independent datasets. We observed that W2V, NPPS, and Binary were the top encoding algorithms. The LGBM classifiers achieved very high performance with an average MCC of 0.604 across 18 encodings, SVM and LR provided MCCs of 0.585 and 0.535. CBC, RF, NB, KNN and XGB MCCs of 0.580, 0.516, 0.491, 0.467 and 0.505, respectively. These results indicate that all eight ML algorithms worked well on the given task.

To assess the performance of the baseline models, we evaluated three global statistical measures (MCC, ACC, AUC) on the training and independent datasets for generic cases (Fig 3). The performance of the baseline models in each ML showed a similar trend with respect to encodings. RFs with DNC, TNC, PseEIIP, Binary, NCP-ND, BPB, NPPS, PseKNC, and W2V encodings surpassed MCC of 0.550, ACC of 0.750, and AUC of 0.820 for both the training and independent datasets. The SVM with W2V exceeded the performance on MCC of 0.710, ACC of 0.850, and AUC of 0.930. In contrast, the XGB with CKSNAP, NPS, NPPS, and W2V attained performance exceeding MCC of 0.550, ACC of 0.770, and AUC of 0.840 during training and testing. Also, the W2V encoding revealed significant prediction performance with CBC and LR where overcoming the MCC of 0.650, ACC of 0.820, and AUC of 0.910. Importantly, the LGBM with Binary, NCP, EIIP, BPB, NPPS, and W2V showed high performance compared to other baseline models, where they produced a surpassed MCC of 0.640, ACC of 0.820, and AUC of 0.900 on training and independent datasets. Among the 144 baseline models, the best single model was LGBM_W2V. The common thought is that an optimal single-feature classifier is considered as the final prediction model. Generally, a single-feature classifier varies the performance, depending on the size of the datasets [81, 82]. For example, they sometimes fail to perform well on large-scale datasets due to the dimensionality of the data



AUC Training

	LGBM	CBC	XGB	LR	RF	SVM	KNN	NB
DNC	0.86	0.87	0.84	0.84	0.84	0.87	0.8	0.82
TNC	0.88	0.88	0.83	0.86	0.85	0.88	0.81	0.84
Kmer	0.89	0.89	0.81	0.86	0.85	0.9	0.81	0.85
CKSNAP	0.88	0.88	0.85	0.87	0.82	0.89	0.8	0.78
RCKmer	0.82	0.83	0.81	0.81	0.81	0.82	0.76	0.78
PseEIIP	0.88	0.88	0.83	0.83	0.85	0.88	0.8	0.84
Binary	0.91	0.88	0.81	0.84	0.82	0.89	0.84	0.82
ENAC	0.88	0.86	0.81	0.83	0.81	0.88	0.79	0.8
ANF	0.72	0.72	0.66	0.69	0.67	0.69	0.56	0.65
NCP	0.9	0.88	0.81	0.84	0.82	0.89	0.84	0.81
EIIP	0.91	0.87	0.81	0.8	0.81	0.84	0.76	0.79
CTD	0.85	0.84	0.8	0.82	0.8	0.82	0.74	0.79
NPS	0.88	0.88	0.85	0.87	0.83	0.89	0.8	0.79
NCP-ND	0.89	0.87	0.82	0.86	0.83	0.89	0.83	0.81
BPB	0.91	0.88	0.81	0.82	0.82	0.84	0.79	0.81
NPPS	0.91	0.9	0.85	0.87	0.86	0.88	0.83	0.84
PseKNC	0.88	0.88	0.84	0.8	0.85	0.87	0.79	0.83
W2V	0.93	0.91	0.85	0.92	0.86	0.93	0.84	0.83

MCC Training

	LGBM	CBC	XGB	LR	RF	SVM	KNN	NB
DNC	0.58	0.58	0.54	0.54	0.55	0.58	0.48	0.5
TNC	0.6	0.61	0.52	0.58	0.55	0.61	0.49	0.54
Kmer	0.63	0.61	0.48	0.58	0.54	0.64	0.49	0.55
CKSNAP	0.61	0.6	0.55	0.59	0.5	0.62	0.46	0.45
RCKmer	0.5	0.51	0.47	0.47	0.48	0.5	0.41	0.43
PseEIIP	0.6	0.61	0.52	0.52	0.55	0.61	0.48	0.54
Binary	0.68	0.6	0.51	0.55	0.56	0.64	0.55	0.54
ENAC	0.61	0.57	0.47	0.54	0.48	0.61	0.46	0.47
ANF	0.32	0.32	0.25	0.27	0.24	0.3	0.11	0.24
NCP	0.64	0.6	0.52	0.55	0.54	0.64	0.55	0.56
EIIP	0.67	0.59	0.53	0.48	0.54	0.54	0.42	0.47
CTD	0.55	0.53	0.46	0.49	0.45	0.49	0.36	0.44
NPS	0.61	0.6	0.55	0.59	0.52	0.62	0.47	0.46
NCP-ND	0.63	0.6	0.52	0.57	0.55	0.64	0.54	0.49
BPB	0.66	0.6	0.52	0.56	0.55	0.57	0.53	0.55
NPPS	0.68	0.64	0.57	0.6	0.57	0.62	0.57	0.56
PseKNC	0.59	0.61	0.53	0.46	0.55	0.58	0.46	0.52
W2V	0.72	0.66	0.57	0.69	0.57	0.71	0.56	0.53

ACC Training

	LGBM	CBC	XGB	LR	RF	SVM	KNN	NB
DNC	0.79	0.79	0.77	0.77	0.77	0.79	0.74	0.75
TNC	0.8	0.8	0.76	0.79	0.78	0.81	0.75	0.77
Kmer	0.81	0.8	0.73	0.79	0.76	0.82	0.75	0.78
CKSNAP	0.8	0.8	0.78	0.8	0.75	0.81	0.73	0.72
RCKmer	0.75	0.75	0.74	0.73	0.74	0.75	0.71	0.71
PseEIIP	0.8	0.8	0.76	0.76	0.78	0.8	0.74	0.77
Binary	0.83	0.8	0.74	0.77	0.75	0.81	0.77	0.75
ENAC	0.8	0.78	0.72	0.76	0.72	0.8	0.73	0.73
ANF	0.65	0.66	0.61	0.63	0.61	0.64	0.54	0.61
NCP	0.82	0.79	0.74	0.77	0.75	0.81	0.77	0.76
EIIP	0.83	0.79	0.74	0.73	0.75	0.76	0.71	0.72
CTD	0.78	0.76	0.73	0.74	0.72	0.74	0.68	0.72
NPS	0.8	0.8	0.78	0.79	0.76	0.81	0.73	0.73
NCP-ND	0.81	0.79	0.74	0.78	0.76	0.82	0.76	0.74
BPB	0.83	0.79	0.73	0.76	0.75	0.76	0.74	0.75
NPPS	0.84	0.82	0.77	0.79	0.77	0.81	0.77	0.77
PseKNC	0.8	0.8	0.76	0.72	0.77	0.79	0.73	0.76
W2V	0.86	0.83	0.77	0.84	0.77	0.85	0.77	0.76

AUC Testing

	LGBM	CBC	XGB	LR	RF	SVM	KNN	NB
DNC	0.86	0.87	0.84	0.85	0.84	0.86	0.81	0.83
TNC	0.88	0.88	0.83	0.87	0.85	0.89	0.82	0.85
Kmer	0.89	0.89	0.81	0.87	0.84	0.9	0.82	0.85
CKSNAP	0.89	0.88	0.84	0.88	0.83	0.89	0.8	0.79
RCKmer	0.82	0.82	0.8	0.8	0.81	0.82	0.75	0.79
PseEIIP	0.88	0.88	0.83	0.84	0.85	0.89	0.81	0.85
Binary	0.91	0.88	0.82	0.84	0.81	0.89	0.84	0.81
ENAC	0.88	0.86	0.8	0.83	0.81	0.88	0.79	0.8
ANF	0.72	0.73	0.67	0.69	0.67	0.7	0.59	0.65
NCP	0.9	0.88	0.81	0.84	0.81	0.89	0.84	0.81
EIIP	0.91	0.87	0.81	0.8	0.81	0.84	0.78	0.79
CTD	0.86	0.85	0.81	0.82	0.81	0.82	0.74	0.79
NPS	0.88	0.88	0.84	0.88	0.84	0.89	0.8	0.8
NCP-ND	0.89	0.87	0.81	0.86	0.82	0.9	0.83	0.81
BPB	0.9	0.88	0.82	0.81	0.82	0.85	0.79	0.81
NPPS	0.91	0.9	0.85	0.87	0.85	0.88	0.84	0.85
PseKNC	0.88	0.88	0.84	0.82	0.85	0.87	0.8	0.85
W2V	0.93	0.91	0.85	0.91	0.86	0.93	0.84	0.83

MCC Testing

	LGBM	CBC	XGB	LR	RF	SVM	KNN	NB
DNC	0.57	0.58	0.54	0.54	0.54	0.58	0.49	0.51
TNC	0.6	0.6	0.5	0.57	0.54	0.62	0.5	0.56
Kmer	0.62	0.61	0.47	0.57	0.53	0.65	0.51	0.57
CKSNAP	0.61	0.6	0.55	0.6	0.5	0.63	0.48	0.44
RCKmer	0.48	0.48	0.47	0.44	0.47	0.48	0.4	0.42
PseEIIP	0.6	0.6	0.5	0.51	0.54	0.62	0.5	0.56
Binary	0.66	0.6	0.5	0.55	0.54	0.63	0.55	0.54
ENAC	0.59	0.56	0.44	0.53	0.46	0.61	0.47	0.47
ANF	0.3	0.32	0.22	0.27	0.22	0.28	0.13	0.21
NCP	0.64	0.59	0.51	0.55	0.53	0.63	0.55	0.55
EIIP	0.66	0.59	0.52	0.46	0.52	0.52	0.44	0.45
CTD	0.57	0.55	0.47	0.49	0.46	0.5	0.37	0.44
NPS	0.61	0.61	0.54	0.59	0.52	0.63	0.48	0.45
NCP-ND	0.62	0.6	0.51	0.57	0.55	0.64	0.55	0.46
BPB	0.65	0.59	0.51	0.55	0.54	0.57	0.53	0.54
NPPS	0.67	0.63	0.55	0.58	0.55	0.6	0.56	0.56
PseKNC	0.59	0.59	0.51	0.48	0.54	0.59	0.46	0.55
W2V	0.72	0.65	0.56	0.68	0.55	0.7	0.55	0.54

ACC Testing

	LGBM	CBC	XGB	LR	RF	SVM	KNN	NB
DNC	0.79	0.79	0.77	0.77	0.77	0.79	0.74	0.76
TNC	0.8	0.8	0.75	0.78	0.77	0.81	0.75	0.78
Kmer	0.81	0.8	0.73	0.78	0.76	0.82	0.76	0.79
CKSNAP	0.81	0.8	0.77	0.8	0.75	0.82	0.74	0.72
RCKmer	0.74	0.74	0.73	0.72	0.73	0.74	0.7	0.7
PseEIIP	0.8	0.8	0.75	0.76	0.77	0.81	0.75	0.78
Binary	0.83	0.79	0.73	0.76	0.74	0.8	0.77	0.75
ENAC	0.79	0.77	0.71	0.75	0.71	0.8	0.73	0.73
ANF	0.65	0.65	0.6	0.63	0.6	0.64	0.55	0.6
NCP	0.82	0.79	0.73	0.76	0.74	0.8	0.77	0.75
EIIP	0.82	0.78	0.73	0.72	0.74	0.75	0.71	0.72
CTD	0.78	0.77	0.73	0.74	0.73	0.75	0.68	0.72
NPS	0.8	0.8	0.77	0.8	0.76	0.81	0.74	0.72
NCP-ND	0.8	0.79	0.74	0.78	0.75	0.81	0.77	0.73
BPB	0.82	0.78	0.72	0.75	0.74	0.75	0.74	0.74
NPPS	0.83	0.81	0.76	0.78	0.76	0.8	0.76	0.76
PseKNC	0.79	0.8	0.75	0.74	0.77	0.8	0.73	0.77
W2V	0.86	0.82	0.76	0.84	0.76	0.85	0.77	0.77

Fig 3. Performance analysis of the baseline classifiers. The classifiers were generated through eight different MLs with eighteen single-feature encoding methods. The MCC, ACC, and AUC are presented on the training (A, B, C) and independent (D, E, F) datasets.

https://doi.org/10.1371/journal.pone.0305406.g003

points. Therefore, a meta-learning ensemble strategy was considered to stack the baseline or single-feature models for enhanced prediction.

Next, we trained the meta-classifiers implementing multiple baseline models on the specific (Am, Cm, Gm, Um) datasets and evaluated their performance on the test dataset. Three statistical measures (MCC, ACC, AUC) were illustrated for respective specific nucleotide modifications in S3-S6 Figs in S1 File. The LGBM_W2V and SVM_W2V were regarded as the best baseline models for all nucleotide-specific datasets.

Impact of meta-learning on the development of Meta-2OM

The stacking approach represents an advanced ensemble method that enhances prediction performance by amalgamating the strengths of multiple models [45, 64, 83–87]. Unlike conventional ensemble approaches that primarily use averaging and voting, stacking employs a

Table 3. Ranking of the baseline models in the descending order of the AUC values.

Rank	Baseline model	Rank	Baseline model	Rank	Baseline model	Rank	Baseline model	Rank	Baseline model
1	LGBM_W2V	31	CBC_PseEIIP	61	NB_Kmer	91	LGBM_RCKmer	121	RF_CTD
2	SVM_W2V	32	CBC_TNC	62	RF_Kmer	92	SVM_CTD	122	KN_NPS
3	LR_W2V	33	LGBM_PseEIIP	63	LR_DNC	93	RF_BPB	123	LR_EIIP
4	LGBM_NPPS	34	LGBM_TNC	64	KN_W2V	94	NB_DNC	124	LR_PseKNC
5	CBC_W2V	35	CBC_NCP	65	CBC_CTD	95	LR_BPB	125	KN_CKSNAP
6	LGBM_Binary	36	CBC_PseKNC	66	NB_NPPS	96	NB_Binary	126	KN_PseKNC
7	LGBM_EIIP	37	LGBM_PseKNC	67	RF_DNC	97	XGB_NCP-ND	127	NB_EIIP
8	LGBM_BPB	38	CBC_NCP-ND	68	XGB_DNC	98	RF_Binary	128	NB_NPS
9	CBC_NPPS	39	LR_CKSNAP	69	NB_PseEIIP	99	RF_NCP	129	KN_ENAC
10	SVM_Kmer	40	CBC_EIIP	70	NB_TNC	100	LR_CTD	130	KN_BPB
11	LGBM_NCP	41	LR_NPPS	71	LR_Binary	101	XGB_Kmer	131	NB_CTD
12	SVM_NCP-ND	42	LR_NPS	72	LR_NCP	102	NB_NCP-ND	132	NB_CKSNAP
13	LGBM_Kmer	43	CBC_DNC	73	XGB_PseKNC	103	NB_NCP	133	NB_RCKmer
14	LGBM_NCP-ND	44	SVM_DNC	74	SVM_EIIP	104	XGB_Binary	134	KN_EIIP
15	SVM_NCP	45	SVM_PseKNC	75	SVM_BPB	105	NB_BPB	135	KN_RCKmer
16	SVM_Binary	46	CBC_ENAC	76	KN_Binary	106	XGB_BPB	136	KN_CTD
17	SVM_CKSNAP	47	LR_Kmer	77	KN_NCP	107	RF_ENAC	137	CBC_ANF
18	SVM_NPS	48	LGBM_DNC	78	KN_NPPS	108	KN_Kmer	138	LGBM_ANF
19	CBC_Kmer	49	LR_TNC	79	RF_NPS	109	RF_EIIP	139	SVM_ANF
20	SVM_NPPS	50	RF_W2V	80	RF_NCP-ND	110	XGB_NCP	140	LR_ANF
21	LGBM_ENAC	51	LR_NCP-ND	81	XGB_TNC	111	RF_RCKmer	141	RF_ANF
22	SVM_TNC	52	RF_NPPS	82	XGB_PseEIIP	112	KN_TNC	142	XGB_ANF
23	SVM_PseEIIP	53	LGBM_CTD	83	LR_PseEIIP	113	XGB_EIIP	143	NB_ANF
24	CBC_Binary	54	XGB_NPPS	84	NB_PseKNC	114	XGB_RCKmer	144	KN_ANF
25	SVM_ENAC	55	XGB_W2V	85	LR_ENAC	115	XGB_ENAC		
26	CBC_NPS	56	RF_TNC	86	KN_NCP-ND	116	LR_RCKmer		
27	LGBM_CKSNAP	57	RF_PseEIIP	87	NB_W2V	117	NB_ENAC		
28	LGBM_NPS	58	RF_PseKNC	88	CBC_RCKmer	118	KN_PseEIIP		
29	CBC_CKSNAP	59	XGB_CKSNAP	89	RF_CKSNAP	119	XGB_CTD		
30	CBC_BPB	60	XGB_NPS	90	SVM_RCKmer	120	KN_DNC		

https://doi.org/10.1371/journal.pone.0305406.t003

meta-model to adeptly combine the forecasts from its base models. This meta-model is trained using the base models' outputs, optimizing their inputs to generate a refined, often superior, final prediction [88, 89]. The notable ACC of stacking is attributed to its proficiency in maximizing the positives of different classifiers while offsetting their limitations. In our research, we harnessed stacking to amalgamate predictions from eight distinct models. Through a meta-learner, which identified patterns among individual predictions, we drew a diverse set of features derived from RNA sequences and leveraged the unique strengths of each classifier to elevate the overall model efficiency. Firstly, we ranked the 144 baseline models by their AUC (Table 3). We then used these rankings to stack the probability scores of the generic (Nm) baseline models using LR. Specifically, we developed 20 meta-classifiers, leveraging the top 1, 2, 4, 8, 12, 16, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120, 130, 140, and all 144 baseline models. Fig 4 shows that the performance of these meta-classifiers progressively improves as the number of probabilistic features (baseline models) increases, both in the training and independent datasets. Our analysis indicating that employing all baseline models, our meta-classifiers achieved the highest AUC on the training dataset as well as the independent dataset.

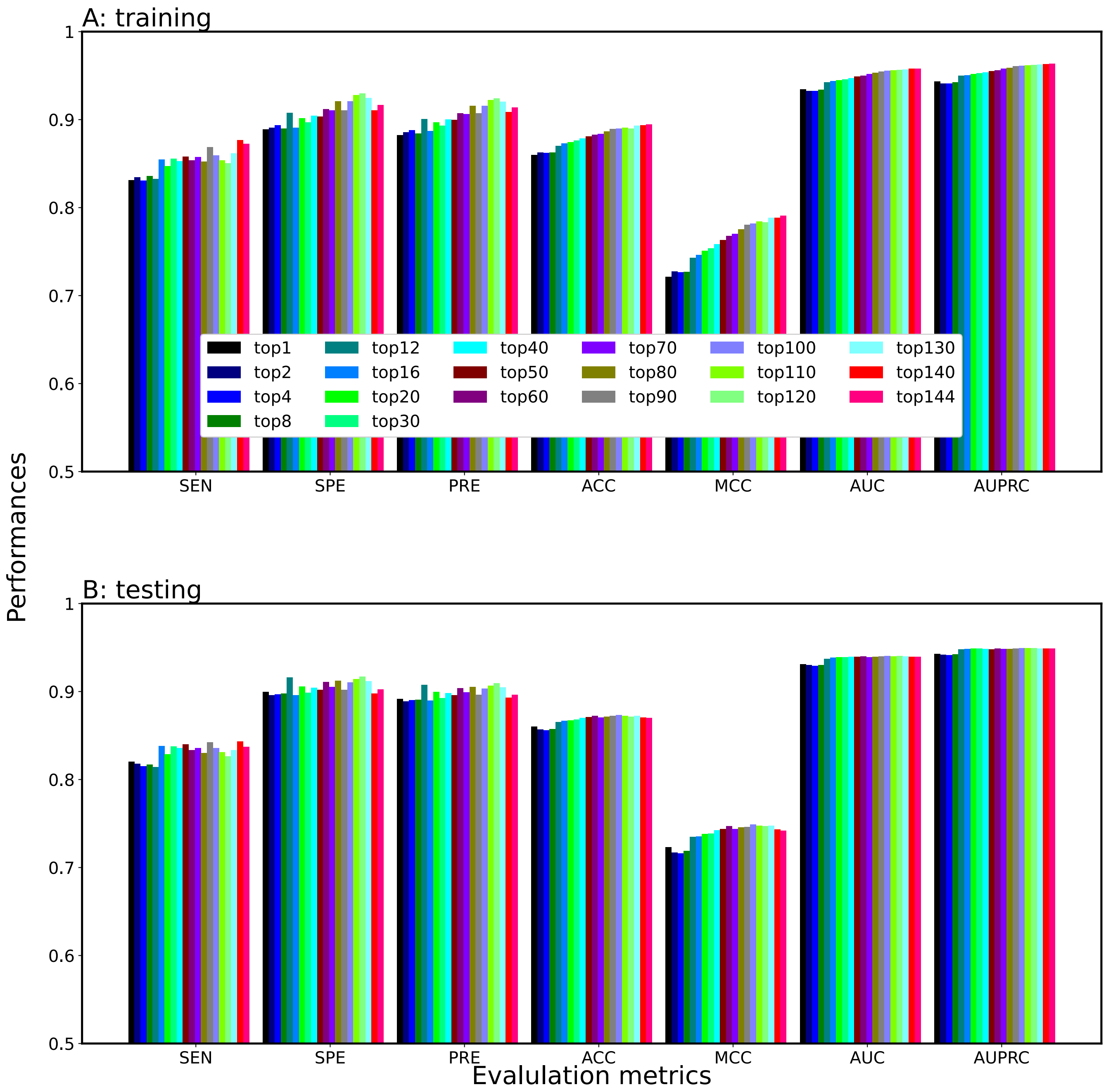


Fig 4. Effect of the employed feature sets on meta-classifiers. The 20 meta-classifiers were generated. The performance was compared based on (A) training and (B) independent datasets.

https://doi.org/10.1371/journal.pone.0305406.g004

To evaluate the effectiveness of various meta-classifiers, we employed six additional ML methods (LGBM, XGB, RF, SVM, KNN, NB) as a meta-classifier to stack baseline models. We then assessed and compared their performance metrics, as provided in Table 4. The results showed that the LR-based meta-classifier, which aggregated predictions from 144 baseline models, surpassed all other ML methods-based stacking models in terms of performance on independent test datasets. Conversely, the other meta-classifiers exhibited signs of overfitting,

Table 4. Performance comparison among seven meta-classifiers incorporating the 144 baseline models on the training and test datasets.

Meta-classifier	Training					Test				
	SEN	SPE	ACC	MCC	AUC	SEN	SPE	ACC	MCC	AUC
LR	0.871	0.918	0.894	0.791	0.958	**0.836**	0.904	**0.870**	**0.743**	**0.940**
LGBM	0.960	0.975	0.968	0.936	**0.996**	0.792	0.893	0.843	0.692	0.923
XGB	0.962	**0.980**	0.971	0.942	0.975	0.805	0.910	0.857	0.719	0.873
RF	0.962	0.977	0.970	0.940	0.994	0.803	0.898	0.850	0.704	0.922
SVM	**0.967**	0.978	**0.973**	0.945	**0.997**	0.820	0.903	0.862	0.727	0.933
KNN	0.964	**0.981**	**0.973**	**0.946**	0.976	0.814	**0.915**	0.864	0.733	0.879
NB	0.964	0.979	0.971	0.943	0.972	0.816	0.905	0.860	0.723	0.862

Bold indicates the highest score for the corresponding statistical measure

https://doi.org/10.1371/journal.pone.0305406.t004

demonstrated by superior performance on the training datasets compared to the test dataset. Therefore, we selected the LR-based meta-model incorporating all baseline models, as our final predictor, and named Meta-2OM. Meta-2OM achieved the best performance with SEN of 0.871, SPE of 0.918, ACC of 0.894, MCC of 0.791, PRE of 0.915, and AUC of 0.958 on the training dataset. On the independent datasets, it achieved a SEN of 0.836, SPE of 0.904, ACC of 0.870, MCC of 0.743, PRE of 0.898, and AUC of 0.940 (Table 4). Compared to the best-performing baseline model, LGBM_W2V, Meta-2OM demonstrated improvements of 9.5% in MCC, 4.1% in ACC, 2.6% in AUC on the training dataset, alongside increases of 2.8% in MCC, 1.3% in ACC, and 1.0% in AUC on the independent dataset, thus substantiating the advanced efficacy of our proposed method.

Comparison of performance based on generic and nucleotide-specific modifications

The 2-OM (Nm) datasets contained four distinct nucleotides (Am, Cm, Gm, Um) modifications. To ascertain the efficacy of Meta-2OM for each type of nucleotide modification, we employed LR-based meta-classifiers, previously demonstrated as the most effective in the generic model development, across all specific nucleotide modifications. Each nucleotide modification dataset was divided into the same 20 groups as in the generic case and corresponding the stacking models were developed. The performances of these models are shown in S7-S10 Figs in S1 File, with the meta-classifier aggregating all 144 baseline models achieved the highest AUC score across all nucleotide-specific cases. Moreover, the capability of the generically trained model (Meta-2OM) to distinguish between positive and negative modification sites was compared against that of the nucleotide-specific model using the independent dataset, as shown in Fig 5 and S3 Table. Remarkably, the generic model (Meta-2OM) more accurately classified the positive and negative sites than nucleotide-specific models. This suggests that Meta-2OM effectively identifies both collective (Nm) and individual nucleotide modifications (Am, Cm, Gm, Um) across comprehensive transcriptome-wide datasets, highlighting its robust applicability and superior predictive performance.

Comparison of Meta-2OM with existing methods

To demonstrate the superiority of Meta-2OM, we conducted a comparative analysis against existing state-of-the-art methods [32–39, 42]. Several methods were excluded from this comparison due to the use of different datasets, inactive web services, or the absence of standalone software packages. Additionally, we did not include methods such as NmRF [32], the approach

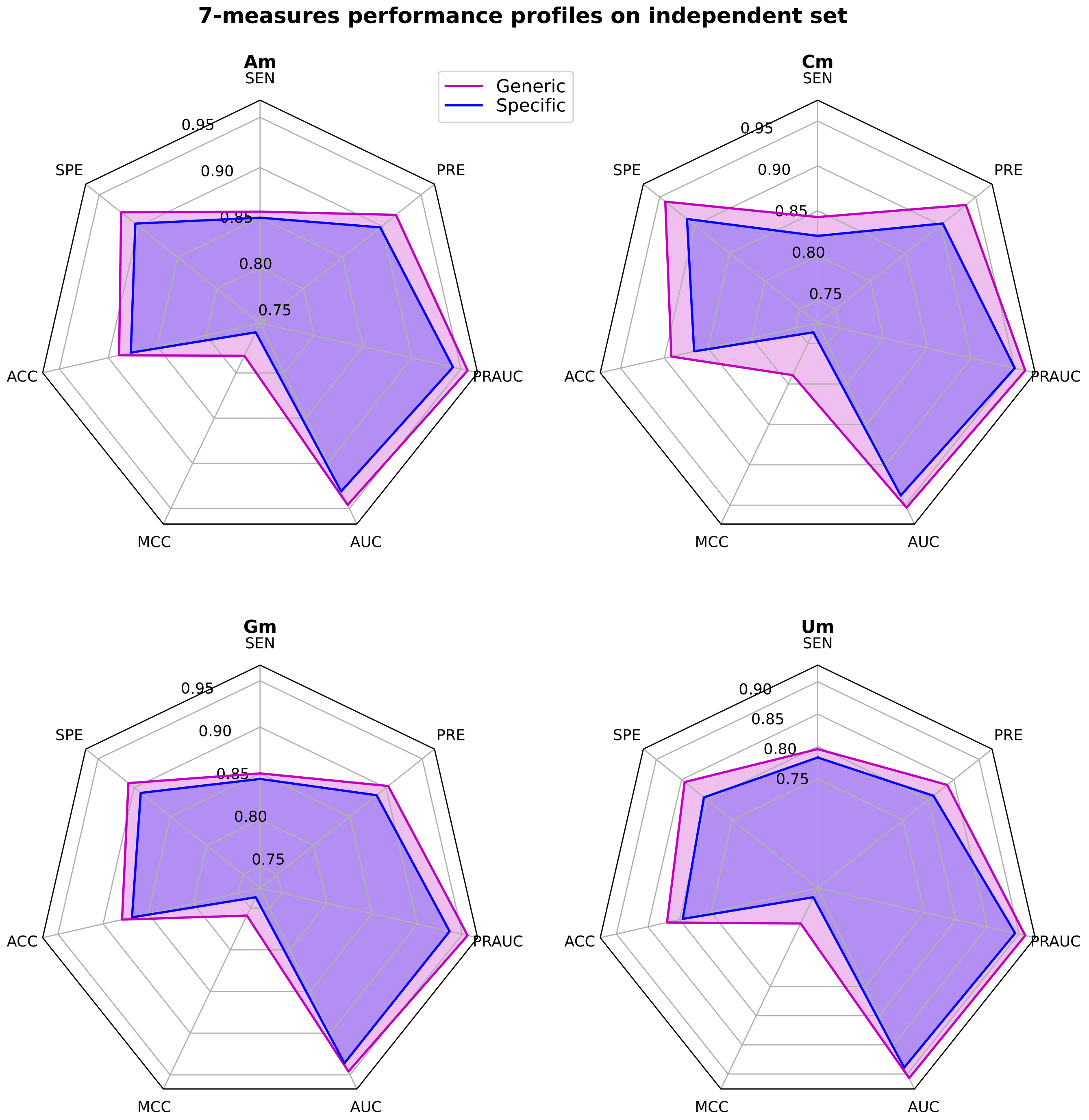


Fig 5. Comparison of prediction performance between Meta-2OM generic model and nucleotide-specific model under four different specific nucleotide modifications independent datasets.

https://doi.org/10.1371/journal.pone.0305406.g005

by Huang et al. [34], iRNA-PseKNC (2methyl) [35], iRNA-2OM [37], and Chen et al.'s method [39], as these were trained exclusively on datasets featuring adenine nucleotides (Am) at the central positions, unlike Meta-2OM which can predict Nm modification sites for all four nucleotides.

Table 5. Comparison of Meta-2OM with state-of-the-art methods.

Method	Am					Cm				
	SEN	**PRE**	**ACC**	**MCC**	**AUC**	**SEN**	**PRE**	**ACC**	**MCC**	**AUC**
i2OM	**0.908**	0.135	0.461	0.192	0.862	0.774	0.425	0.884	0.518	0.915
H2Opred	0.863	0.906	0.887	0.774	0.943	**0.849**	0.912	0.883	0.769	0.946
Meta-2OM	0.856	**0.918**	**0.889**	**0.781**	**0.946**	0.843	**0.936**	**0.892**	**0.789**	**0.953**
Method	**Gm**					**Um**				
	SEN	**PRE**	**ACC**	**MCC**	**AUC**	**SEN**	**PRE**	**ACC**	**MCC**	**AUC**
i2OM	0.830	0.431	0.885	0.545	0.934	0.597	0.227	0.779	0.266	0.774
H2Opred	**0.853**	**0.906**	**0.882**	**0.765**	0.946	**0.831**	0.809	0.818	0.636	0.898
Meta-2OM	0.850	0.903	0.879	0.759	**0.946**	0.796	**0.837**	**0.820**	**0.642**	**0.907**
Method	**Average value**									
	SEN	**PRE**	**ACC**	**MCC**	**AUC**					
i2OM	0.777	0.305	0.752	0.380	0.871					
H2Opred	**0.849**	0.883	0.867	0.736	0.933					
Meta-2OM	0.836	**0.899**	**0.870**	**0.743**	**0.938**					

Bold value indicates the highest score for the corresponding statistical measure; Average value indicates Mean(Am, Cm, Gm, Um) under respective measures

https://doi.org/10.1371/journal.pone.0305406.t005

Finally, we selected the i2OM method [41] and H2Opred [42], trained on larger datasets, that can predict both Nm modification sites and specific nucleotide-based sites. The results demonstrated that Meta-2OM slightly better than H2Opred and significantly better than i2OM in identifying Nm modification sites as well as individual types of nucleotide modification sites (Table 5 and Fig 6). Upon assessing the performance across individual nucleotide datasets, Meta-2OM demonstrated marginal enhancements in the Am, Cm, and Um datasets when compared to H2Opred. However, in the case of Gm, H2Opred exhibited superior performance relative to Meta-2OM. Additionally, Meta-2OM significantly surpassed the performance of i2OM across these assessments.

Furthermore, Meta-2OM also outperformed single ML-based methods, such as RF-based NmRF [32], RF-based NmSEER V2.0 [36], SVM-based Huang et al. [34], and CNN-based Deep-2'-O-Me [38], in predicting Nm sites on the independent dataset, achieving an AUC of 0.940. Overall, the meta-learning approach implemented in the Meta-2OM method is responsible for its improved performance compared to the existing predictors.

Probabilistic-feature analysis

Meta-2OM exhibited superior performance compared to the single-feature models on both the training and independent datasets. To delve deeper into how the model works, we conducted SHapley Additive exPlanation (SHAP) analysis on the probabilistic features generated by the 144 baseline models. The most influential 20 are shown in Fig 7. The importance ranking of the baseline models revealed that classifiers such as LGBM, CBC, LR, and SVM methods contributing more to the prediction than RF, NB, KNN and XGB. Notably, the W2V based LGBM_W2V, LR_W2V, SVM_W2V, and CBC_W2V among the top 20 made significant contribution, while the remaining models played a complementary role on Meta-2OM prediction. To further demonstrate the Meta-2OM prediction capability, we applied the t-distributed stochastic neighbor embedding (t-SNE; scikit-learn v.1.0.2) to compute the 2D feature based on the 144D probabilistic features, as shown in Fig 8. The results demonstrated a clear separation between positive and negative samples, indicating that the Meta-2OM effectively discerns the underlying patterns distinguishing 2-OM from non-2-OM modification sequences.

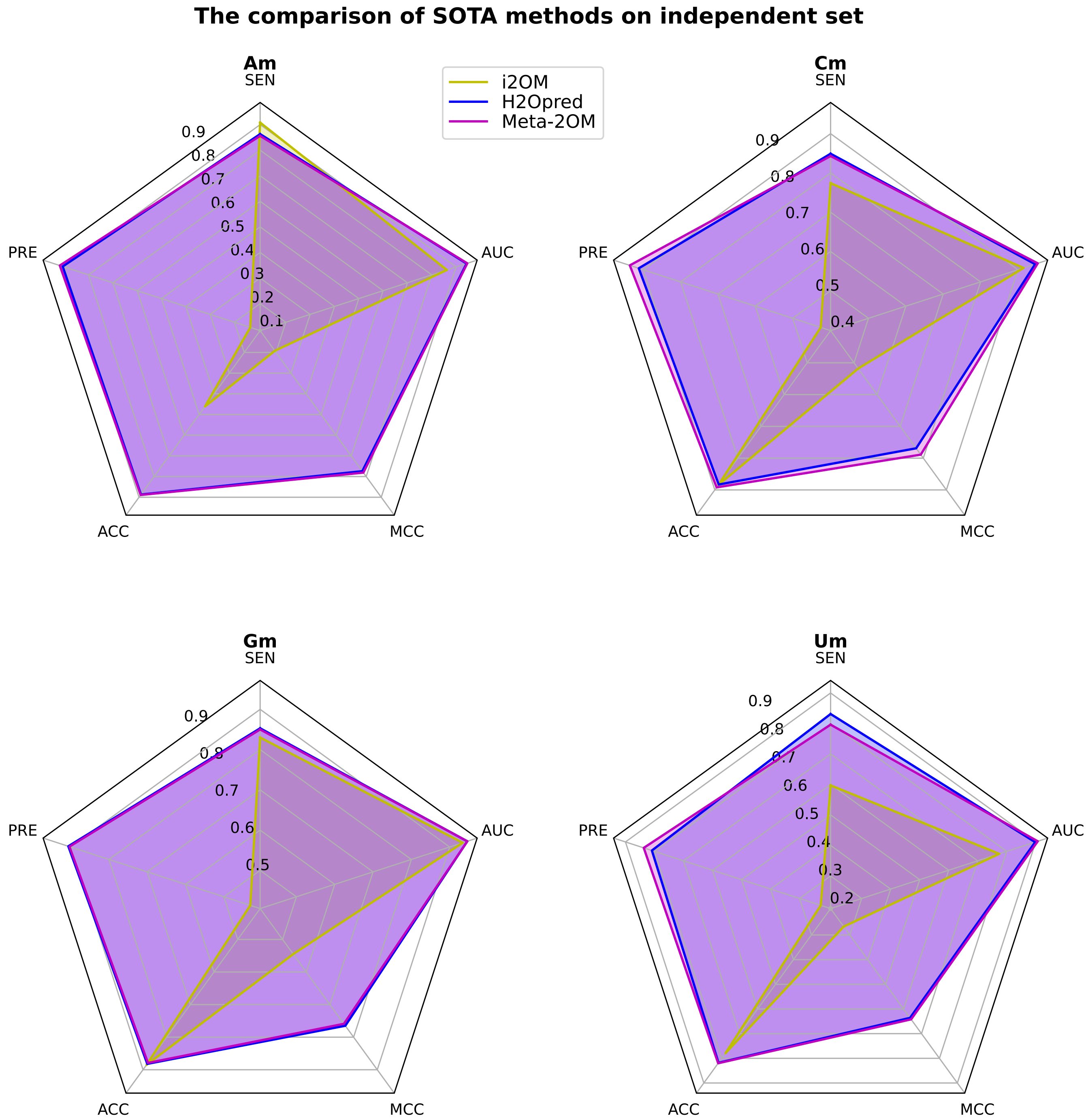


Fig 6. Parallel comparison of the Meta-2OM predictors with state-of-the-art methods based on the nucleotide-specific independent datasets.

https://doi.org/10.1371/journal.pone.0305406.g006

Web server implementation

We developed a user-friendly and freely accessible web server for Meta-2OM, which is available at http://kurata35.bio.kyutech.ac.jp/Meta-2OM/. The web server is implemented using Flask (2.2.2) in the Python program and Apache (2.4.52). Users can easily carry out the prediction by submitting FASTA sequences of 41 bp in length. The prediction results can be

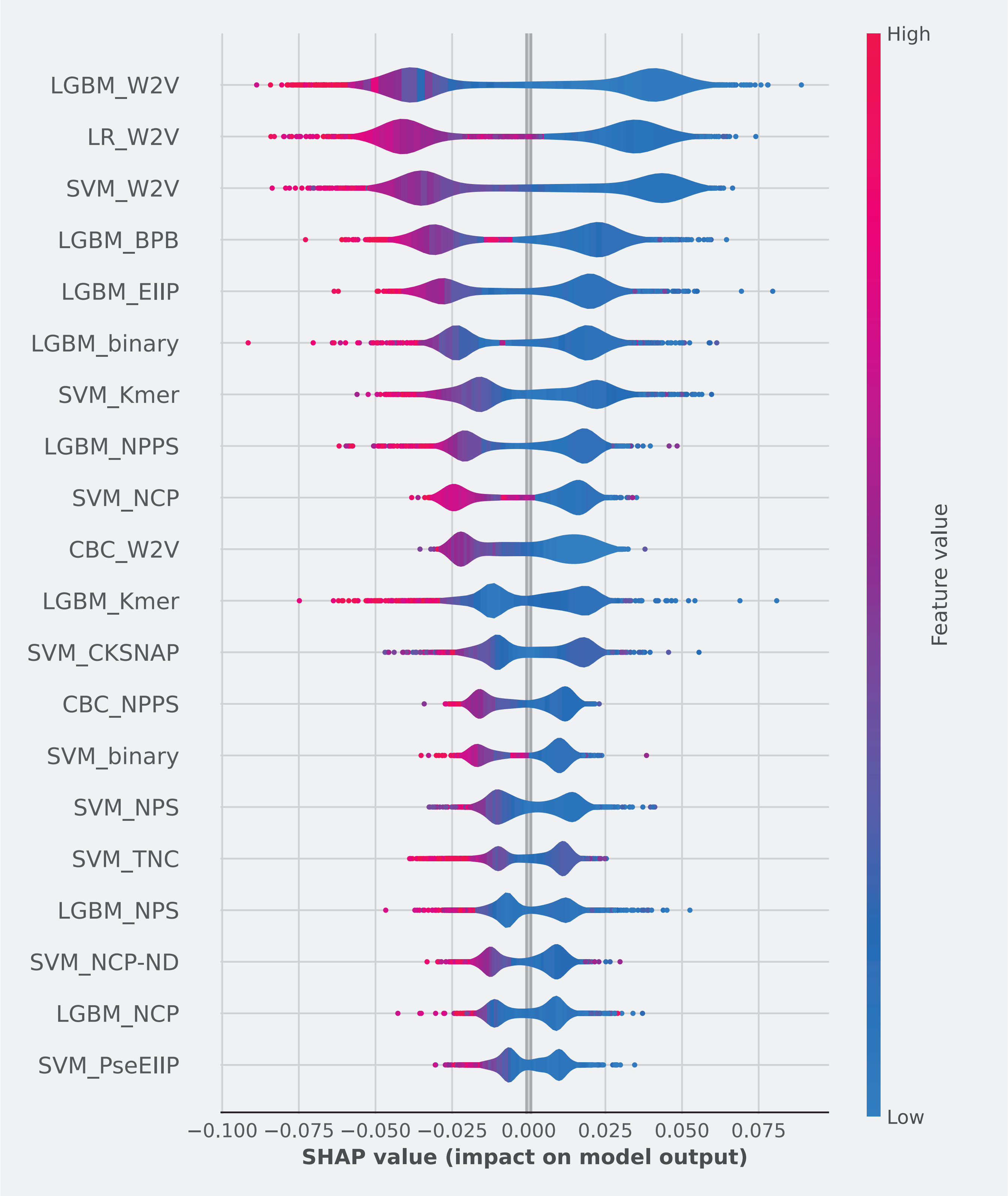


Fig 7. SHAP plot analysis using the probability features generated by 144 baseline models considered in Meta-2OM. The top 20 were illustrated. The position on the y-axis is determined by the feature and that on the x-axis by the SHAP value. The color (red to blue) represents the value of the feature from low to high. The SHAP's positive and negative values mean that they are associated with a higher and lower prediction, respectively. The features are ordered according to their importance.

https://doi.org/10.1371/journal.pone.0305406.g007

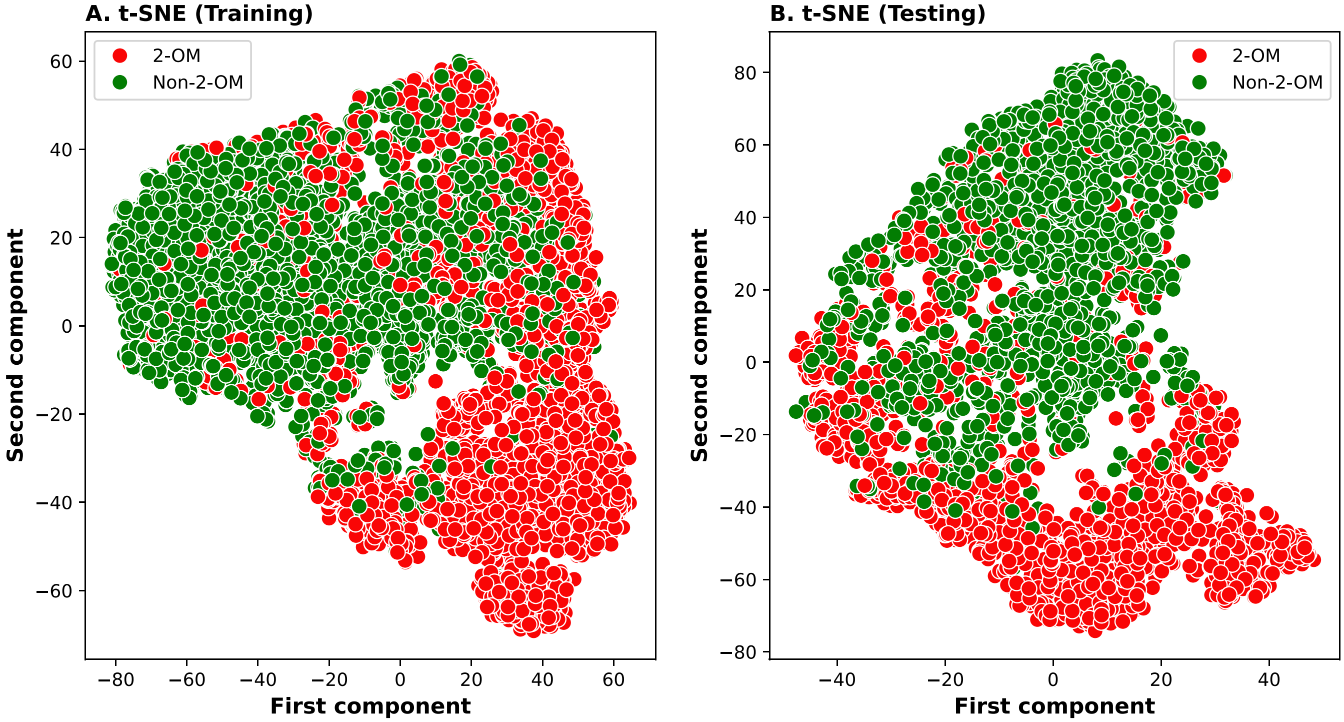


Fig 8. Analysis of the probability features generated by 144 baseline models for the negative and positive samples by t-distributed stochastic neighbor embedding (t-SNE). It represents the discriminative distribution of prediction probabilities. These results are shown for both (A) training and (B) independent datasets.

https://doi.org/10.1371/journal.pone.0305406.g008

downloaded as an Excel file (.csv) with three columns: predicted labels, predicted probability scores, and sample sequences. The threshold value for predicting Nm modification site is 0.5. Detailed user guidelines are available on the website's help page.

Discussion and conclusion

2-OM modification is a prevalent and crucial modification in RNA, necessary for various biological and functional mechanisms. Accurate identification of 2-OM sites is essential to further understanding its roles. The biological experimental methods that identify 2-OM sites are time-consuming, labor-intensive, and expensive [20–22, 24]. To complement such experimental methods, we have developed a computational model to predict 2-OM sites, using benchmark datasets covering all four types of nucleotide modification (Am, Cm, Um, Gm) in the human chromosomes. Specifically, a total of 12,182 positive and negative samples were employed to build the model, where each sample consisted of 41 nucleotide base pairs.

In this work, we developed Meta-2OM, a meta-learning-based ML approach for accurately identifying 2-OM RNA modification sites using sequence information. We first assessed the prediction performance of eight widely established ML classifiers, where each classifier trained with 18 types of RNA sequence features. Subsequently, we built a meta-model that integrates the predictions of the baseline models to enhance the overall prediction accuracy. Meta-2OM demonstrated superior performance compared to the existing predictors when tested on an independent dataset. The efficiency of Meta-2OM is attributed to the following factors: (i) utilization of 18 different feature encoding algorithms that capture various intrinsic capabilities to discriminate between 2-OM and non-2-OM sites; and (ii) construction of meta-model incorporating 144 baseline models. Therefore, we believe that the current approach can be extended to other sequence-based function prediction problems, including enhancer prediction, peptide therapeutic function prediction, replication original sites prediction, and post-translational modification sites prediction [90–93]. While Meta-2OM achieves an ACC exceeding 87% on both training and independent datasets, we expect that incorporating position-specific information or natural language processing-based features with variants DL algorithms may enhance the predictive performance. Currently, the proposed Meta-2OM predictors do not consider RNA tertiary structures, which are crucial for accurate RNA site prediction. By integrating information about RNA 3D structures into our model, we could significantly increase its predictive accuracy. We plan to explore these potential improvements in our future studies.

Supporting information

S1 Table. Hyperparameter tuning for each machine-learning method.
(PDF)

S2 Table. Sequential performance of 144 baseline classifiers through the training datasets.
(PDF)

S3 Table. The individual performance of the nucleotide-specific model and a comparison with the generic model under training and test datasets.
(PDF)

S1 File. The several performance analysis representations of 8 ML techniques with 18 different encoding models. Also, the stacking performance representation is based on nucleotide-specific datasets.
(PDF)

Author Contributions

Conceptualization: Md. Harun-Or-Roshid, Hiroyuki Kurata.

Data curation: Md. Harun-Or-Roshid.

Formal analysis: Md. Harun-Or-Roshid, Hiroyuki Kurata.

Funding acquisition: Balachandran Manavalan, Hiroyuki Kurata.

Investigation: Md. Harun-Or-Roshid.

Methodology: Md. Harun-Or-Roshid, Hiroyuki Kurata.

Project administration: Balachandran Manavalan, Hiroyuki Kurata.

Resources: Md. Harun-Or-Roshid, Hiroyuki Kurata.

Software: Md. Harun-Or-Roshid.

Supervision: Balachandran Manavalan, Hiroyuki Kurata.

Validation: Md. Harun-Or-Roshid.

Visualization: Md. Harun-Or-Roshid, Nhat Truong Pham.

Writing – original draft: Md. Harun-Or-Roshid.

Writing – review & editing: Nhat Truong Pham, Balachandran Manavalan, Hiroyuki Kurata.

References

1. Li S, Mason CE. The pivotal regulatory landscape of RNA modifications. Annu Rev Genomics Hum Genet. 2014;15. https://doi.org/10.1146/annurev-genom-090413-025405 PMID: 24898039
2. Sprinzl M, Vassilenko KS. Compilation of tRNA sequences and sequences of tRNA genes. Nucleic Acids Res. 2005;33. https://doi.org/10.1093/nar/gki012 PMID: 15608164
3. Arzumanian VA, Dolgalev GV., Kurbatov IY, Kiseleva OI, Poverennaya EV. Epitranscriptome: Review of Top 25 Most-Studied RNA Modifications. International Journal of Molecular Sciences. 2022. https://doi.org/10.3390/ijms232213851 PMID: 36430347
4. Behm-Ansmant I, Helm M, Motorin Y. Use of specific chemical reagents for detection of modified nucleotides in RNA. Journal of Nucleic Acids. 2011. https://doi.org/10.4061/2011/408053 PMID: 21716696
5. Shi H, Wei J, He C. Where, When, and How: Context-Dependent Functions of RNA Methylation Writers, Readers, and Erasers. Molecular Cell. 2019. https://doi.org/10.1016/j.molcel.2019.04.025 PMID: 31100245
6. Kiss T. Small nucleolar RNA-guided post-transcriptional modification of cellular RNAs. EMBO Journal. 2001. https://doi.org/10.1093/emboj/20.14.3617 PMID: 11447102
7. Somme J, Van Laer B, Roovers M, Steyaert J, Versées W, Droogmans L. Characterization of two homologous 2′-O-methyltransferases showing different specificities for their tRNA substrates. RNA. 2014;20. https://doi.org/10.1261/rna.044503.114 PMID: 24951554
8. Rebane A, Roomere H, Metspalu A. Locations of several novel 2′,-O-methylated nucleotides in human 28S rRNA. BMC Mol Biol. 2002;3. https://doi.org/10.1186/1471-2199-3-1 PMID: 11897011
9. Darzacq X, Jády BE, Verheggen C, Kiss AM, Bertrand E, Kiss T. Cajal body-specific small nuclear RNAs: A novel class of 2′-O-methylation and pseudouridylation guide RNAs. EMBO J. 2002;21. https://doi.org/10.1093/emboj/21.11.2746 PMID: 12032087
10. Li J, Yang Z, Yu B, Liu J, Chen X. Methylation protects miRNAs and siRNAs from a 3′-end uridylation activity in Arabidopsis. Curr Biol. 2005;15. https://doi.org/10.1016/j.cub.2005.07.029 PMID: 16111943
11. Yu B, Yang Z, Li J, Minakhina S, Yang M, Padgett RW, et al. Methylation as a crucial step in plant microRNA biogenesis. Science (80-). 2005;307. https://doi.org/10.1126/science.1107130 PMID: 15705854
12. Dai Q, Moshitch-Moshkovitz S, Han D, Kol N, Amariglio N, Rechavi G, et al. Nm-seq maps 2′-O-methylation sites in human mRNA with base precision. Nat Methods. 2017;14. https://doi.org/10.1038/nmeth.4294 PMID: 28504680
13. Dimitrova DG, Teysset L, Carré C. RNA 2′-O-Methylation (Nm) modification in human diseases. Genes (Basel). 2019;10. https://doi.org/10.3390/genes10020117 PMID: 30764532
14. Züst R, Cervantes-Barragan L, Habjan M, Maier R, Neuman BW, Ziebuhr J, et al. Ribose 2'-O-methylation provides a molecular signature for the distinction of self and non-self mRNA dependent on the RNA sensor Mda5. Nat Immunol. 2011;12. https://doi.org/10.1038/ni.1979 PMID: 21217758
15. Abou Assi H, Rangadurai AK, Shi H, Liu B, Clay MC, Erharter K, et al. 2′-O-Methylation can increase the abundance and lifetime of alternative RNA conformational states. Nucleic Acids Res. 2021;48. https://doi.org/10.1093/nar/gkaa928 PMID: 33104789
16. Drazkowska K, Tomecki R, Warminski M, Baran N, Cysewski D, Depaix A, et al. 2′-O-Methylation of the second transcribed nucleotide within the mRNA 5′ cap impacts the protein production level in a cell-specific manner and contributes to RNA immune evasion. Nucleic Acids Res. 2022;50. https://doi.org/10.1093/nar/gkac722 PMID: 36018811
17. Marcel V, Ghayad SE, Belin S, Therizols G, Morel AP, Solano-Gonzàlez E, et al. P53 Acts as a Safeguard of Translational Control by Regulating Fibrillarin and rRNA Methylation in Cancer. Cancer Cell. 2013;24. https://doi.org/10.1016/j.ccr.2013.08.013 PMID: 24029231
18. Encinar JA, Menendez JA. Potential drugs targeting early innate immune evasion of SARS-coronavirus 2 via 2'-O-Methylation of Viral RNA. Viruses. 2020;12. https://doi.org/10.3390/v12050525 PMID: 32397643
19. Paramasivam A. RNA 2′-O-methylation modification and its implication in COVID-19 immunity. Cell Death Discovery. 2020. https://doi.org/10.1038/s41420-020-00358-z PMID: 33298836
20. Krogh N, Jansson MD, Häfner SJ, Tehler D, Birkedal U, Christensen-Dalsgaard M, et al. Profiling of 2′-O-Me in human rRNA reveals a subset of fractionally modified positions and provides evidence for ribosome heterogeneity. Nucleic Acids Res. 2016;44. https://doi.org/10.1093/nar/gkw482 PMID: 27257078
21. Erales J, Marchand V, Panthu B, Gillot S, Belin S, Ghayad SE, et al. Evidence for rRNA 2′-O-methylation plasticity: Control of intrinsic translational capabilities of human ribosomes. Proc Natl Acad Sci U S A. 2017;114. https://doi.org/10.1073/pnas.1707674114 PMID: 29158377

22. Sharma S, Marchand V, Motorin Y, Lafontaine DLJ. Identification of sites of 2′-O-methylation vulnerability in human ribosomal RNAs by systematic mapping. Sci Rep. 2017;7. https://doi.org/10.1038/s41598-017-09734-9 PMID: 28904332

23. Zhou F, Liu Y, Rohde C, Pauli C, Gerloff D, Köhn M, et al. AML1-ETO requires enhanced C/D box snoRNA/RNP formation to induce self-renewal and leukaemia. Nat Cell Biol. 2017;19. https://doi.org/10.1038/ncb3563 PMID: 28650479

24. Zhu Y, Pirnie SP, Carmichael GG. High-throughput and site-specific identification of 2′-O-methylation sites using ribose oxidation sequencing (RibOxi-seq). RNA. 2017;23. https://doi.org/10.1261/rna.061549.117 PMID: 28495677

25. Kim Y, Park JH, Cho YR. Network-Based Approaches for Disease-Gene Association Prediction Using Protein-Protein Interaction Networks. Int J Mol Sci. 2022;23. https://doi.org/10.3390/ijms23137411 PMID: 35806415

26. Cai Y, Yang J, Huang T, Wang MW. Editorial: Computational Methods in Predicting Complex Disease Associated Genes and Environmental Factors. Frontiers in Genetics. 2021. https://doi.org/10.3389/fgene.2021.679651 PMID: 33968143

27. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596. https://doi.org/10.1038/s41586-021-03819-2 PMID: 34265844

28. Sledzinski P, Nowaczyk M, Olejniczak M. Computational Tools and Resources Supporting CRISPR-Cas Experiments. Cells. 2020. https://doi.org/10.3390/cells9051288 PMID: 32455882

29. Ibrahim OAS, Hamed BA, El-Hafeez TA. A new fast technique for pattern matching in biological sequences. J Supercomput. 2023;79. https://doi.org/10.1007/s11227-022-04673-3

30. Reza MS, Harun-Or-roshid M, Islam MA, Hossen MA, Hossain MT, Feng S, et al. Bioinformatics Screening of Potential Biomarkers from mRNA Expression Profiles to Discover Drug Targets and Agents for Cervical Cancer. Int J Mol Sci. 2022;23. https://doi.org/10.3390/ijms23073968 PMID: 35409328

31. Dalwadi SM, Hunt A, Bonnen MD, Ghebre YT. Computational approaches for drug repurposing in oncology: untapped opportunity for high value innovation. Frontiers in Oncology. 2023. https://doi.org/10.3389/fonc.2023.1198284 PMID: 37274281

32. Ao C, Zou Q, Yu L. NmRF: identification of multispecies RNA 2’-O-methylation modification sites from RNA sequences. Brief Bioinform. 2022;23. https://doi.org/10.1093/bib/bbab480 PMID: 34850821

33. Li H, Chen L, Huang Z, Luo X, Li H, Ren J, et al. DeepOMe: A Web Server for the Prediction of 2′-O-Me Sites Based on the Hybrid CNN and BLSTM Architecture. Front Cell Dev Biol. 2021;9. https://doi.org/10.3389/fcell.2021.686894 PMID: 34055810

34. Huang Q-L, Wang L, Han S-G, Tang H. Identification of 2’-O-methylation Site by Investigating Multi-feature Extracting Techniques. Comb Chem High Throughput Screen. 2020;23. https://doi.org/10.2174/1386207323666200425210609 PMID: 32334499

35. Tahir M, Tayara H, Chong KT. iRNA-PseKNC(2methyl): Identify RNA 2’-O-methylation sites by convolution neural network and Chou’s pseudo components. J Theor Biol. 2019;465. https://doi.org/10.1016/j.jtbi.2018.12.034 PMID: 30590059

36. Zhou Y, Cui Q, Zhou Y. NmSEER V2.0: A prediction tool for 2′-O-methylation sites based on random forest and multi-encoding combination. BMC Bioinformatics. 2019;20. https://doi.org/10.1186/s12859-019-3265-8 PMID: 31874624

37. Yang H, Lv H, Ding H, Chen W, Lin H. IRNA-2OM: A sequence-based predictor for identifying 2′-O-Methylation Sites in Homo sapiens. J Comput Biol. 2018;25. https://doi.org/10.1089/cmb.2018.0004 PMID: 30113871

38. Mostavi M, Salekin S, Huang Y. Deep-2’-O-Me: Predicting 2’-O-methylation sites by Convolutional Neural Networks. Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS. 2018. https://doi.org/10.1109/EMBC.2018.8512780 PMID: 30440889

39. Chen W, Feng P, Tang H, Ding H, Lin H. Identifying 2′-O-methylationation sites by integrating nucleotide chemical properties and nucleotide compositions. Genomics. 2016;107. https://doi.org/10.1016/j.ygeno.2016.05.003 PMID: 27191866

40. Zhou Y, Cui Q, Zhou Y. NmSEER: A prediction tool for 2’-O-Methylation (Nm) sites based on random forest. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2018. https://doi.org/10.1007/978-3-319-95930-6_90

41. Yang YH, Ma CY, Gao D, Liu XW, Yuan SS, Ding H. i2OM: Toward a better prediction of 2′-O-methylation in human RNA. Int J Biol Macromol. 2023;239. https://doi.org/10.1016/j.ijbiomac.2023.124247 PMID: 37003392

42. Pham NT, Rakkiyapan R, Park J, Malik A, Manavalan B. H2Opred: a robust and efficient hybrid deep learning model for predicting 2'-O-methylation sites in human RNA. Brief Bioinform. 2024;25. https://doi.org/10.1093/bib/bbad476 PMID: 38180830

43. Xuan JJ, Sun WJ, Lin PH, Zhou KR, Liu S, Zheng LL, et al. RMBase v2.0: Deciphering the map of RNA modifications from epitranscriptome sequencing data. Nucleic Acids Res. 2018;46. https://doi.org/10.1093/nar/gkx934 PMID: 29040692

44. Huang Y, Niu B, Gao Y, Fu L, Li W. CD-HIT Suite: A web server for clustering and comparing biological sequences. Bioinformatics. 2010;26. https://doi.org/10.1093/bioinformatics/btq003 PMID: 20053844

45. Hasan MM, Basith S, Khatun MS, Lee G, Manavalan B, Kurata H. Meta-i6mA: An interspecies predictor for identifying DNA N6-methyladenine sites of plant genomes by exploiting informative features in an integrative machine-learning framework. Brief Bioinform. 2021;22. https://doi.org/10.1093/bib/bbaa202 PMID: 32910169

46. Xu ZC, Feng PM, Yang H, Qiu WR, Chen W, Lin H. IRNAD: A computational tool for identifying D modification sites in RNA sequence. Bioinformatics. 2019;35. https://doi.org/10.1093/bioinformatics/btz358 PMID: 31077296

47. Foster PG, Nunes CR, Greene P, Moustakas D, Stroud RM. The First Structure of an RNA m5C Methyltransferase, Fmu, Provides Insight into Catalytic Mechanism and Specific Binding of RNA Substrate. Structure. 2003;11. https://doi.org/10.1016/j.str.2003.10.014 PMID: 14656444

48. Chen Z, Zhao P, Li F, Leier A, Marquez-Lago TT, Wang Y, et al. IFeature: A Python package and web server for features extraction and selection from protein and peptide sequences. Bioinformatics. 2018;34. https://doi.org/10.1093/bioinformatics/bty140 PMID: 29528364

49. Chen Z, Zhao P, Li F, Marquez-Lago TT, Leier A, Revote J, et al. ILearn: An integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data. Brief Bioinform. 2020;21. https://doi.org/10.1093/bib/bbz041 PMID: 31067315

50. Lalović D, Veljković V. The global average DNA base composition of coding regions may be determined by the electron-ion interaction potential. BioSystems. 1990;23. https://doi.org/10.1016/0303-2647(90)90013-q PMID: 2322643

51. Cui X, Liang Z, Shen L, Zhang Q, Bao S, Geng Y, et al. 5-Methylcytosine RNA Methylation in Arabidopsis Thaliana. Mol Plant. 2017;10. https://doi.org/10.1016/j.molp.2017.09.013 PMID: 28965832

52. Shao J, Xu D, Tsai SN, Wang Y, Ngai SM. Computational identification of protein methylation sites through Bi-profile Bayes feature extraction. PLoS One. 2009;4. https://doi.org/10.1371/journal.pone.0004920 PMID: 19290060

53. Tong X, Liu S. CPPred: Coding potential prediction based on the global description of RNA sequence. Nucleic Acids Res. 2019;47. https://doi.org/10.1093/nar/gkz087 PMID: 30753596

54. Golam Bari ATM, Reaz MR, Choi HJ, Jeong BS. DNA encoding for splice site prediction in large DNA sequence. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2013. https://doi.org/10.1007/978-3-642-40270-8_4

55. Zhou Y, Zeng P, Li YH, Zhang Z, Cui Q. SRAMP: Prediction of mammalian N6-methyladenosine (m6A) sites based on sequence-derived features. Nucleic Acids Res. 2016;44. https://doi.org/10.1093/nar/gkw104 PMID: 26896799

56. Xing P, Su R, Guo F, Wei L. Identifying N6-methyladenosine sites using multi-interval nucleotide pair position specificity and support vector machine. Sci Rep. 2017;7. https://doi.org/10.1038/srep46757 PMID: 28440291

57. Guo SH, Deng EZ, Xu LQ, Ding H, Lin H, Chen W, et al. INuc-PseKNC: A sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition. Bioinformatics. 2014;30. https://doi.org/10.1093/bioinformatics/btu083 PMID: 24504871

58. Liao M, ping Zhao J, Tian J, Zheng CH. iEnhancer-DCLA: using the original sequence to identify enhancers and their strength based on a deep learning framework. BMC Bioinformatics. 2022;23. https://doi.org/10.1186/s12859-022-05033-x PMID: 36376800

59. Kurata H, Tsukiyama S, Manavalan B. IACVP: Markedly enhanced identification of anti-coronavirus peptides using a dataset-specific word2vec model. Brief Bioinform. 2022;23. https://doi.org/10.1093/bib/bbac265 PMID: 35772910

60. Zulfiqar H, Sun ZJ, Huang QL, Yuan SS, Lv H, Dao FY, et al. Deep-4mCW2V: A sequence-based predictor to identify N4-methylcytosine sites in Escherichia coli. Methods. 2022;203. https://doi.org/10.1016/j.ymeth.2021.07.011 PMID: 34352373

61. Adjuik TA, Ananey-Obiri D. Word2vec neural model-based techniqueto generate protein vectors for combating COVID-19: a machine learning approach. Int J Inf Technol. 2022;14. https://doi.org/10.1007/s41870-022-00949-2 PMID: 35611155

62. Petrov AI, Kay SJE, Kalvari I, Howe KL, Gray KA, Bruford EA, et al. RNAcentral: A comprehensive database of non-coding RNA sequences. Nucleic Acids Res. 2017;45. https://doi.org/10.1093/nar/gkw1008 PMID: 27794554

63. Breiman L. Random forests. Mach Learn. 2001;45. https://doi.org/10.1023/A1010933404324

64. Hasan MM, Alam MA, Shoombuatong W, Kurata H. IRC-Fuse: improved and robust prediction of redox-sensitive cysteine by fusing of multiple feature representations. J Comput Aided Mol Des. 2021;35. https://doi.org/10.1007/s10822-020-00368-0 PMID: 33392948

65. Khatun S, Hasan M, Kurata H. Efficient computational model for identification of antitubercular peptides by integrating amino acid patterns and properties. FEBS Lett. 2019;593. https://doi.org/10.1002/1873-3468.13536 PMID: 31297788

66. Rokach L. Ensemble-based classifiers. Artif Intell Rev. 2010;33. https://doi.org/10.1007/s10462-009-9124-7

67. Charoenkwan P, Shoombuatong W, Lee HC, Chaijaruwanich J, Huang HL, Ho SY. SCMCRYS: Predicting Protein Crystallization Using an Ensemble Scoring Card Method with Estimating Propensity Scores of P-Collocated Amino Acid Pairs. PLoS One. 2013;8. https://doi.org/10.1371/journal.pone.0072368 PMID: 24019868

68. Ali F, Akbar S, Ghulam A, Maher ZA, Unar A, Talpur DB. AFP-CMBPred: Computational identification of antifreeze proteins by extending consensus sequences into multi-blocks evolutionary information. Comput Biol Med. 2021;139. https://doi.org/10.1016/j.compbiomed.2021.105006 PMID: 34749096

69. Ali F, Ahmed S, Swati ZNK, Akbar S. DP-BINDER: machine learning model for prediction of DNA-binding proteins by fusing evolutionary and physicochemical information. J Comput Aided Mol Des. 2019;33. https://doi.org/10.1007/s10822-019-00207-x PMID: 31123959

70. Tang H, Zhao YW, Zou P, Zhang CM, Chen R, Huang P, et al. HBPred: A tool to identify growth hormone-binding proteins. Int J Biol Sci. 2018;14. https://doi.org/10.7150/ijbs.24174 PMID: 29989085

71. Kopitar L, Kocbek P, Cilar L, Sheikh A, Stiglic G. Early detection of type 2 diabetes mellitus using machine learning-based prediction models. Sci Rep. 2020;10. https://doi.org/10.1038/s41598-020-68771-z PMID: 32686721

72. Dorogush AV, Ershov V, Gulin A. CatBoost: gradient boosting with categorical features support. 2018; 1–7. Available: http://arxiv.org/abs/1810.11363

73. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. Catboost: Unbiased boosting with categorical features. Adv Neural Inf Process Syst. 2018; 2018-Decem: 6638–6648.

74. Wang Y, Guo R, Huang L, Yang S, Hu X, He K. m6AGE: A Predictor for N6-Methyladenosine Sites Identification Utilizing Sequence Characteristics and Graph Embedding-Based Geometrical Information. Front Genet. 2021;12. https://doi.org/10.3389/fgene.2021.670852 PMID: 34122525

75. Onah E, Uzor PF, Ugwoke IC, Eze JU, Ugwuanyi ST, Chukwudi IR, et al. Prediction of HIV-1 protease cleavage site from octapeptide sequence information using selected classifiers and hybrid descriptors. BMC Bioinformatics. 2022;23. https://doi.org/10.1186/s12859-022-05017-x PMID: 36344934

76. Wu H, Liu M, Zhang P, Zhang H. iEnhancer-SKNN: a stacking ensemble learning-based method for enhancer identification and classification using sequence information. Brief Funct Genomics. 2023;22. https://doi.org/10.1093/bfgp/elac057 PMID: 36715222

77. Akbar S, Hayat M. iMethyl-STTNC: Identification of N6-methyladenosine sites by extending the idea of SAAC into Chou's PseAAC to formulate RNA sequences. J Theor Biol. 2018;455. https://doi.org/10.1016/j.jtbi.2018.07.018 PMID: 30031793

78. Zhang X, Wang S, Xie L, Zhu Y. PseU-ST: A new stacked ensemble-learning method for identifying RNA pseudouridine sites. Front Genet. 2023;14. https://doi.org/10.3389/fgene.2023.1121694 PMID: 36741328

79. Du Z, Ding X, Hsu W, Munir A, Xu Y, Li Y. pLM4ACE: A protein language model based predictor for antihypertensive peptide screening. Food Chem. 2024;431. https://doi.org/10.1016/j.foodchem.2023.137162 PMID: 37604011

80. Wu X, Bartel DP. KpLogo: Positional k -mer analysis reveals hidden specificity in biological sequences. Nucleic Acids Res. 2017;45. https://doi.org/10.1093/nar/gkx323 PMID: 28460012

81. Hasan MM, Shoombuatong W, Kurata H, Manavalan B. Critical evaluation of web-based DNA N6-methyladenine site prediction tools. Brief Funct Genomics. 2021;20. https://doi.org/10.1093/bfgp/elaa028 PMID: 33491072

82. Basith S, Manavalan B, Hwan Shin T, Lee G. Machine intelligence in peptide therapeutics: A next-generation tool for rapid disease screening. Medicinal Research Reviews. 2020. https://doi.org/10.1002/med.21658 PMID: 31922268

83. Raza A, Uddin J, Almuhaimeed A, Akbar S, Zou Q, Ahmad A. AIPs-SnTCN: Predicting Anti-Inflammatory Peptides Using fastText and Transformer Encoder-Based Hybrid Word Embedding with Self-

Normalized Temporal Convolutional Networks. J Chem Inf Model. 2023; 63: 6537–6554. https://doi.org/10.1021/acs.jcim.3c01563 PMID: 37905969

84. Akbar S, Hayat M, Iqbal M, Jan MA. iACP-GAEnsC: Evolutionary genetic algorithm based ensemble classification of anticancer peptides by utilizing hybrid feature space. Artif Intell Med. 2017;79. https://doi.org/10.1016/j.artmed.2017.06.008 PMID: 28655440

85. Akbar S, Raza A, Shloul T Al, Ahmad A, Saeed AGhadi YY, et al. pAtbP-EnC: Identifying Anti-Tubercular Peptides Using Multi-Feature Representation and Genetic Algorithm-Based Deep Ensemble Model. IEEE Access. 2023; 11: 137099–137114. https://doi.org/10.1109/ACCESS.2023.3321100

86. Akbar S, Hayat M, Tahir M, Khan S, Alarfaj FK. cACP-DeepGram: Classification of anticancer peptides via deep neural network and skip-gram-based word embedding model. Artif Intell Med. 2022;131. https://doi.org/10.1016/j.artmed.2022.102349 PMID: 36100346

87. Harun-Or-Roshid M, Maeda K, Phan LT, Manavalan B, Kurata H. Stack-DHUpred: Advancing the accuracy of dihydrouridine modification sites detection via stacking approach. Comput Biol Med. 2024; 169: 107848. https://doi.org/10.1016/j.compbiomed.2023.107848 PMID: 38145601

88. Chen R, Li F, Guo X, Bi Y, Li C, Pan S, et al. ATTIC is an integrated approach for predicting A-to-I RNA editing sites in three species. Brief Bioinform. 2023;24. https://doi.org/10.1093/bib/bbad170 PMID: 37150785

89. Bupi N, Sangaraju VK, Phan LT, Lal A, Vo TTB, Ho PT, et al. An Effective Integrated Machine Learning Framework for Identifying Severity of Tomato Yellow Leaf Curl Virus and Their Experimental Validation. Research. 2023;6. https://doi.org/10.34133/research.0016 PMID: 36930763

90. MacPhillamy C, Alinejad-Rokny H, Pitchford WS, Low WY. Cross-species enhancer prediction using machine learning. Genomics. 2022;114. https://doi.org/10.1016/j.ygeno.2022.110454 PMID: 36030022

91. Wang L, Wang N, Zhang W, Cheng X, Yan Z, Shao G, et al. Therapeutic peptides: current applications and future directions. Signal Transduction and Targeted Therapy. 2022. https://doi.org/10.1038/s41392-022-00904-4 PMID: 35165272

92. Manavalan B, Basith S, Shin TH, Lee G. Computational prediction of species-specific yeast DNA replication origin via iterative feature representation. Brief Bioinform. 2021;22. https://doi.org/10.1093/bib/bbaa304 PMID: 33232970

93. Hasan MM, Tsukiyama S, Cho JY, Kurata H, Alam MA, Liu X, et al. Deepm5C: A deep-learning-based hybrid framework for identifying human RNA N5-methylcytosine sites using a stacking strategy. Mol Ther. 2022;30. https://doi.org/10.1016/j.ymthe.2022.05.001 PMID: 35526094