Biophysics 生物物理学, 2025, 13(3), 27-44
Published Online August 2025 in Hans. https://www.hanspub.org/journal/biphy
https://doi.org/10.12677/biphy.2025.133003

Hans汉斯

多源特征结合机器学习算法预测钾离子(K^+)与钠离子(Na^+)的结合位点

刘　玮*，冯永娥#

内蒙古农业大学理学院，内蒙古　呼和浩特

收稿日期：2025年12月19日；录用日期：2026年1月12日；发布日期：2026年1月19日

摘　要

钾离子(K^+)与钠离子(Na^+)是生物体内重要的电解质，在维持细胞渗透压平衡、调节神经信号传导以及参与酶促反应调控中发挥关键作用。准确识别蛋白质中的金属离子结合位点，对于深入理解离子调控机制及相关疾病的分子基础具有重要意义。本文基于BioLiP数据库获取K^+和Na^+结合蛋白序列，利用CD-HIT进行序列去冗余处理，并按5:1的比例划分为训练集和测试集。采用SMOTEENN算法对训练集进行类别平衡处理，从序列、结构与能量三个层面共提取9类特征(PSSM、氨基酸组分、密码子频率、相对可及表面积、SASA-RASA、疏水性、二级结构、结合能和图能量)，并分别使用7种机器学习算法(Logistic Regression, SVM, KNN, Random Forest, Gradient Boosting, XGBoost, LightGBM)进行模型构建与性能评估。结果表明，单特征PSSM在K^+和Na^+结合位点的预测中均表现最优，其中K^+结合位点预测的敏感性Sn = 100%，特异性Sp = 85.3%，总精度Acc = 85.6%，AUC值达到0.984；Na^+结合位点预测的敏感性Sn = 100%，特异性Sp = 86.5%，总精度Acc = 86.6%，AUC值达到0.978。鉴于梯度提升算法在处理非线性关系的能力较强，同时对特征交互的捕捉更高效等优点，随后在LightGBM算法下，采用最优特征PSSM与其他8种特征作逐一融合，结果发现：特征融合后K^+和Na^+结合位点的预测精度的各项指标都有一定的提高；同时也发现特征融合不是越多越好，部分特征间存在一定信息冗余，故合理的特征选择与融合策略对模型优化至关重要。本研究对于离子通道蛋白功能解析，靶向药物研发等方面具有一定的生物学意义。

关键词

(K^+/Na^+)离子结合位点，位置特异性打分矩阵(PSSM)，机器学习，特征融合，SMOTEENN算法

Prediction of Potassium (K^+) and Sodium (Na^+) Ion Binding Sites Using Multi-Source Features and Machine Learning Algorithms

Wei Liu*, Yong'e Feng#

*第一作者。
#通讯作者。

文章引用: 刘玮, 冯永娥. 多源特征结合机器学习算法预测钾离子(K^+)与钠离子(Na^+)的结合位点[J]. 生物物理学, 2025, 13(3): 27-44. DOI: 10.12677/biphy.2025.133003

College of Science, Inner Mongolia Agriculture University, Hohhot Inner Mongolia

Received: December 19, 2025; accepted: January 12, 2026; published: January 19, 2026

Abstract

K^+ and Na^+ are important electrolytes in organisms, which play a key role in maintaining cell osmotic pressure balance, regulating nerve signaling, and participating in the regulation of enzymatic reactions. Accurate identification of K^+ and Na^+ binding sites in proteins is of great significance for in-depth understanding of ion regulation mechanisms and the molecular basis of related diseases. In this paper, the sequences of K^+ and Na^+ binding proteins were selected from the BioLiP database, and the sequence redundancy was removed by CD-HIT. The sequence is divided into training and test sets according to a 5:1 ratio. Balance training data was the SMOTEENN algorithm, nine types of features (PSSM, amino acid components, codon frequency, relative accessible surface area, SASA-RASA, hydrophobicity, secondary structure, binding energy and graph energy) from three levels (sequence, structure and energy information) were extracted, and seven machine learning algorithms (Logistic Regression, SVM, KNN, Random Forest, Gradient Boosting, XGBoost, and LightGBM) were used to build models and evaluate performance. The results showed that the single-feature PSSM performed the best in the prediction of K^+ and Na^+ binding sites, among which the Sn = 100%, Sp = 85.3%, Acc = 85.6%, and AUC value reached 0.984, and the Sn = 100%, Sp = 86.5%, Acc = 86.6%, and AUC value of Na^+ binding site prediction reached 0.978. In view of the advantages of the gradient algorithm in processing nonlinear relationships and more efficient capture of feature interactions, the optimal feature PSSM was used to fuse with 8 other features one by one under the LightGBM algorithm, and the results showed that the prediction accuracy of K^+ and Na^+ binding sites was improved to a certain extent after feature fusion. At the same time, it is also found that using more features does not yield better results, due to information redundancy among some features. So reasonable feature selection and fusion strategies are very important for model optimization. This study has certain biological significance for the functional elucidation of ion channel proteins and the development of targeted drugs.

Keywords

(K^+/Na^+) Ion Binding Sites, Position-Specific Scoring Matrix (PSSM), Machine Learning, Feature Fusion, SMOTEENN Algorithm

1. 引言

金属离子在生物体内参与多种生命活动，其中钾离子(K^+)和钠离子(Na^+)在维持电化学平衡、调节渗透压及信号传导过程中发挥关键作用[1]。蛋白质与金属离子的结合位点是其功能实现的核心环节，准确识别这些结合位点对于深入理解蛋白质功能及离子调控机制具有重要意义[2]。传统的实验方法(如 X 射线晶体衍射和核磁共振谱 NMR)虽然精度较高，但实验操作复杂、成本高且难以实现大规模筛查，因此，基于计算的预测方法逐渐成为研究主流[3]。早期的研究多基于序列特征，如氨基酸组分(AAC)、密码子使用频率(CUF)以及位置特异性打分矩阵(PSSM)等，其中 PSSM 因能够捕捉蛋白质序列的进化保守性信

息而被广泛应用[4] [5]。随着蛋白质结构数据的积累，研究者逐渐引入了结构特征，如相对可及表面积(rASA)和溶剂可及表面积(SASA)，用于描述结合位点的空间暴露程度与可接触性[6]。进一步地，理化与能量类特征(如疏水性、结合能、图能量)被用于刻画残基间的能量关系及其在结合过程中的相互作用稳定性，从而有效提升模型对结合区域的区分能力。在算法方面，传统的机器学习方法如逻辑回归(Logistic Regression, LR)和支持向量机(Support Vector Machine, SVM)在早期的蛋白质结合位点预测研究中应用较多，但其线性或低阶核函数形式难以充分表征复杂的非线性特征关系。近年来，机器学习算法(如随机森林 Random Forest、梯度提升树 Gradient Boosting、极端梯度提升 XGBoost 及轻量梯度提升 LightGBM)凭借在特征表达能力与泛化性能上的优势，已成为蛋白质功能与结构预测研究的主流方法[7]-[9]。此外，多特征融合策略逐渐成为提高预测精度的重要方向，通过在序列、结构及能量等多维层面融合信息，能够在不同尺度上捕获结合位点的差异特征，从而显著提升模型的稳健性与准确性[10]。基于上述研究进展，本文以 BioLiP 数据库中钾离子(K^+)与钠离子(Na^+)结合蛋白为研究对象，利用 CD-HIT 工具对序列进行去冗余处理，并按 5:1 比例划分训练集与测试集。在特征构建阶段，从序列、结构及能量三个层面提取九类特征，包括位置特异性打分矩阵(PSSM)、二级结构、氨基酸组分(AAC)、密码子频率(CUF)、疏水性、图能量、结合能、相对可及表面积(rASA)与综合表面积特征(SASA-rASA)。为缓解样本不平衡问题，本文采用 SMOTEENN 方法对训练集进行平衡处理。随后，分别基于七种机器学习算法(逻辑回归、支持向量机、K 近邻[11]、随机森林[12]、梯度提升树[13]、XGBoost [14]与 LightGBM [15])建立预测模型。实验结果表明，PSSM 特征在两类离子结合位点预测中表现最佳。进一步的特征融合分析结果显示，融合序列、结构与能量的部分特征可显著提升预测精度与模型稳健性。

2. 材料与方法

2.1. 数据库的构建

本文基于在 Biolip 数据库获取蛋白质与 K^+、Na^+的结合位点序列[16]。为避免重复数据的干扰，用 CD-HIT 工具对序列集进行去冗余，保留序列一致性 ≤30%的序列[17]。对这些序列的结合位点进行标注，按 5:1 比例将序列分为训练集与检验集，以中心残基为结合位点或非结合位点，左右各取等长残基片段，构建结合位点数据集和非结合位点数据集。具体数据如表 1 所示。

Table 1. Datasets of binding and non-binding sites
表 1. 结合位点和非结合位点数据集

类型	结合位点数	结合位点数据集	非结合位点数据集
K^+	382	376	24,180
Na^+	478	475	37,588

2.2. 数据的平衡

在金属离子结合位点预测任务中，数据集往往存在显著的不平衡问题，即负样本(非结合位点)数量远高于正样本(结合位点)。直接进行模型训练，分类器容易偏向于学习多数类(负样本)特征，从而导致预测性能下降，尤其是对结合位点识别的敏感性较低。

本研究中，K^+与 Na^+结合数据的正负样本比例约为 1:100，若不进行平衡处理，模型将倾向于输出负集结果。为缓解样本不平衡问题，本文在训练阶段采用 SMOTEENN (SMOTE + Edited Nearest Neighbours)混合采样策略对训练集进行平衡化处理[18]。其中，SMOTE 通过在少数类样本之间插值生成新样

本，有效扩展了潜在结合位点的特征空间；ENN 进一步删除边界处的噪声与重叠样本，提升数据的纯净度与可分性。该混合方法综合了过采样与欠采样的优点，使模型能够更好地学习结合与非结合位点的特征差异。

2.3. 特征参数的选取

2.3.1. 氨基酸组分(AAC)

氨基酸组分反映蛋白质序列中 20 种标准氨基酸的相对比例，是最基础的序列特征[19]。该特征能揭示结合位点周围的残基组成差异，反应序列的最基本组成，定义如下：

$$\mathrm{AAC}\left(a_i\right)=\frac{N\left(a_i\right)}{L} \tag{1}$$

其中，$N\left(a_i\right)$ 为氨基酸 a_i 的出现次数，L 为序列总长度。

2.3.2. 疏水性

氨基酸的亲疏水性是决定蛋白质空间构象、稳定性及其与配体相互作用的重要理化性质。根据 Eidhammer 等人提出的亲疏水性分类原则[20]，将 20 种标准氨基酸按照亲水或疏水特性划分为六类，如表 2 所示。包括强亲水、强疏水、弱亲水、弱疏水、半胱氨酸、甘氨酸和脯氨酸六组。本文采用六类亲疏水性氨基酸的组成比例作为序列特征。

Table 2. Classification of amino acid hydrophobicity
表 2. 氨基酸亲疏水性分类表

氨基酸特性类别	包含的氨基酸	类别符号
强亲水性氨基酸	R, D, E, N, Q, K, H	Q
强疏水性氨基酸	L, I, V, A, M, F	A
弱亲水或弱疏水性氨基酸	S, T, Y, W	S
半胱氨酸	C	C
甘氨酸	G	G
脯氨酸	P	P

2.3.3. 结合能

结合能反映蛋白质与配体相互作用的能量稳定性，是决定结合强度的重要物理量。本文基于蛋白质残基间拉普拉斯矩阵特征值计算能量分布[21]，将 DNA 序列表示用四种核苷酸(A, C, G, T)映射为二维单位向量：$A=(0,1)$，$C=(-1,0)$，$G=(1,0)$，$T=(0,-1)$。通过“游走”方式构建 DNA 序列的图形(起点为原点，按序列顺序移动对应向量)。氨基酸的图形表示基于密码子，每个氨基酸由 1~6 个密码子(三联体核苷酸)编码(61 个有效密码子对应 20 种氨基酸)。对每个氨基酸的所有密码子按字母顺序排列并拼接成 DNA 序列，再通过上述方法构建其图形(忽略少数起始密码子的例外翻译)。考虑到密码子的简并性，20 种氨基酸的结合能是归一化后的结果，列在附录中[22]，本文依据 20 种氨基酸的能量区间将其划分为 5 类，如表 3 所示，我们根据 5 类氨基酸的组分构建的特征向量。

2.3.4. 图能量

图能量(Graph Energy)用于刻画蛋白质三维结构中残基之间的相互作用模式及其稳定性[23]。通过将氨基酸残基视为节点、相互作用为边，构建加权无向图，并计算其谱能量值。20 种氨基酸每个氨基酸的

图能量列在附录中，本文依据能量区间将 20 种氨基酸划分为 5 类，如表 4 所示，我们根据 5 类氨基酸的组分构建特征向量。该特征能揭示结合位点倾向出现在结构稳定的拓扑区域。

Table 3. Classification table of amino acid binding energy
表 3. 氨基酸结合能分类表

能量类别	能量区间	包含的氨基酸
X_1	4.0339~4.5491	C, E, Q, S, T
X_2	4.6337~4.7079	D, H, V
X_3	4.7973~5.0336	A, L, M, W
X_4	5.1529~5.2015	F, N, R, Y
X_5	5.3044~5.4214	G, I, P, K

Table 4. Classification table of amino acid diagram energy
表 4. 氨基酸图能量分类表

能量类别	能量区间	包含的氨基酸
X_1	4.0339~4.5491	E, H, L, M, Q, V
X_2	4.6337~4.7079	R
X_3	4.7973~5.0336	F, G, K, N, P, Y
X_4	5.1529~5.2015	I
X_5	3.9569~4.4722	A, C, D, S, T, W

2.3.5. 密码子频率(CUF)

密码子频率表示同义密码子在编码序列中的使用偏好[24]。不同密码子的使用与蛋白质的翻译效率、稳定性及结构形成有关。定义如下：

$$\mathrm{CUF}_i = \frac{N_i}{N_{total}} \tag{2}$$

其中 N_i 为第 i 个密码子的出现次数， N_{total} 为密码子总数。

2.3.6. 位置特异性打分矩阵(PSSM)

位置特异性打分矩阵(position-specific score matrix, PSSM)能反映蛋白质在进化过程中对特定残基的选择压力，记录序列中每个位点与不同氨基酸匹配的保守性得分[25]。本文利用了 BLAST 软件包中的 PSI-BLAST 来对 swissport 数据库进行搜索对比，迭代次数设置为 2，期望值设置为 $1e^{-5}$，其他参数设为默认值来生成 PSSM 文件。对于每一条序列，生成其 PSSM 可表示为

$$P = A_1 A_2 \cdots A_L$$

$$P_{PSSM} = \begin{bmatrix} A_{1,1} & A_{1,2} & \cdots & A_{1,20} \\ A_{2,1} & A_{2,2} & \cdots & A_{2,20} \\ \vdots & \vdots & \ddots & \vdots \\ A_{L,1} & A_{L,2} & \cdots & A_{L,20} \end{bmatrix} \tag{3}$$

其中，A_1 表示蛋白质序列的第一个氨基酸残基，A_2 表示第二个氨基酸残基，以此类推，A_L 表示蛋白质序列第 L 个氨基酸。PSSM 为一个 $L \times 20$ 的矩阵，其中 20 表示标准氨基酸的数量，L 是蛋白质序列的长度，

矩阵中元素 $A_{i,j}$ 表示序列上第 j 个氨基酸突变为第 i 个氨基酸的得分，得分值越低说明突变概率越小，得分值越高表示突变概率越高[4]。

2.3.7. 二级结构组分

二级结构描述蛋白质中局部区域的构象类型，主要包括 α-螺旋(Helix)、β-折叠(Sheet)和无规卷曲(Coil)。本文通过预测工具 PSIPRED 基于氨基酸序列预测二级结构[26]，然后计算 3 种类型的构象的组分。

2.3.8. 相对可及表面积(RASA)

溶剂可及表面积 SASA (Solvent Accessible Surface Area)表示蛋白质中每个氨基酸残基暴露在溶剂中的表面积，通常以 A^2 为单位[27]。其计算基于蛋白质的三维结构模型，本文基于使用 AlphaFold2 预测结构[28]，得到每条蛋白质序列的 phd 文件，采用 Biopython 的 ShrakeRupley 算法计算得到每个残基的溶剂可及表面积。RASA 是 SASA 通过归一化后可比较不同残基的可接触性，其公式为

$$\mathrm{RASA}=\frac{\mathrm{SASA}_i}{\mathrm{SASA}_{\max}} \tag{4}$$

其中 SASA_i 为第 i 个残基的实际 SASA 值，$\mathrm{SASA}_{\max}$ 为该残基在完全暴露状态下的理论最大可及表面积，参考标准值如表 5 所示。RASA 的取值范围在 0 到 1 之间，值越大代表残基越暴露。

Table 5. Values of solvent accessible surface area of 20 amino acid
表 5. 氨基酸溶剂可及表面积值

氨基酸	MaxASA	氨基酸	MaxASA
A	121	L	191
R	265	K	230
N	187	M	203
D	187	F	228
C	148	P	154
Q	214	S	143
E	214	T	163
G	97	W	264
H	216	Y	255
I	195	V	165

2.3.9. 综合表面积特征(SASA-RASA)

离子结合位点多位于“部分暴露的凹槽区域”，此类区域具有适中的表面积和良好的几何限制，有助于离子的稳定结合[29]。本文将 SASA 与 RASA 拼接为综合暴露性特征，以同时刻画绝对暴露量与相对暴露比例，其公式为

$$\mathrm{SASA\text{-}RASA}=\frac{1}{n}\sum_{i=1}^{n}\mathrm{SASA}_i+\frac{1}{n}\sum_{i=1}^{n}\mathrm{RASA}_i \tag{5}$$

2.4. 评价指标

目前，预测算法性能检验常用的方法有独立检验(independent test) k-折交叉检验(k-fold cross-validation

test) [30]。本文采用了独立检验的方法，并采用了机器学习算法中常用的 5 种评估指标来评估模型的性能。5 种评估指标分别是准确率(Accuracy)，马修斯相关系数(MCC)，敏感性(Sn)，特异性(Sp)，ROC 曲线下面积(AUC)。

$$\mathrm{Acc}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \tag{6}$$

$$\mathrm{MCC}=\frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}} \tag{7}$$

$$\mathrm{Sn}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{8}$$

$$\mathrm{Sp}=\frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} \tag{9}$$

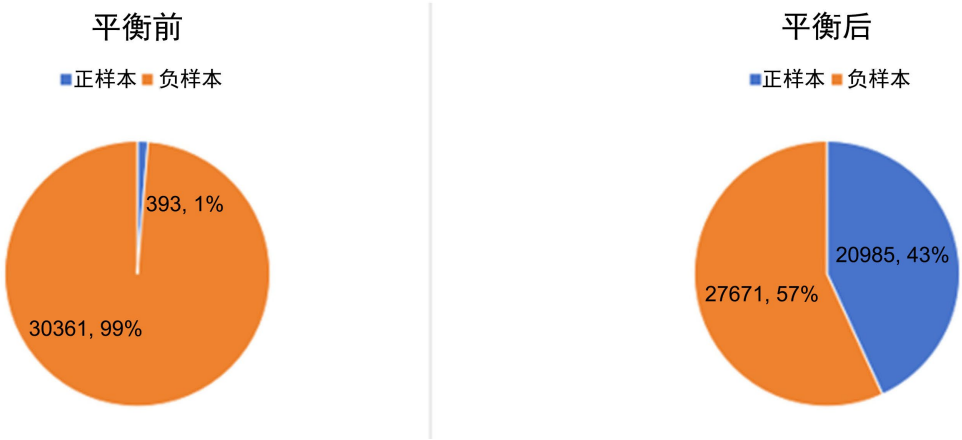
其中，TP 表示正样本预测正确的数量，TN 表示负样本预测正确的数量，FP 表示实际不是正样本，却被预测为正样本的数量，FN 表示实际不是负样本，却被预测为负样本的数量。AUC (ROC-AUC)常用于二分类问题，ROC 曲线是假阳性和召回率之间的标准权衡表示，AUC 是 ROC 曲线下面积，衡量模型区分正负样本的能力。

3. 结果与讨论

3.1. 平衡数据库

鉴于 K^+与 Na^+结合数据的正负样本比例悬殊，故本文采用 SMOTEENN (SMOTE + Edited Nearest Neighbours)混合采样策略，对训练集进行平衡化处理。本研究对于分布相对平滑的特征(如 PSSM、二级结构、相对可及表面积及融合特征)，采用默认参数的 SMOTEENN 方法(random_state = 42)，以在保持样本多样性的同时去除噪声样本。而对于分布高度集中的特征(如图能量、结合能、疏水性、密码子频率和氨基酸组分)，在初步实验中发现默认参数易引入噪声样本，因此采用较为保守的 SMOTE 参数(sampling_strategy = 0.4, k_neighbors = 3)，并结合较大邻域范围的 ENN (n_neighbors = 22)以增强对边界噪声的清除能力。图 1 显示了 Na^+/K^+训练集在 SMOTEENN 平衡前后正负集样本占比图，由图可见平衡后正负集样本数差别不大，从而缓解类不平衡的问题。这种方法可有效提升了模型区分类别边界的能力，并改善了分类性能。

(a) K^+



(b) Na^+

Figure 1. Sample proportions of positive and negative set before and after SMOTEENN in Na^+/K^+ training sets
图 1. Na^+/K^+训练集在 SMOTEENN 平衡前后正负集样本比例图

3.2. 筛选最佳窗口

本文以 K^+、Na^+的结合位点序列进行展开分析，以中心残基为结合位点或非结合位点构建滑动窗口大小为 15，17，19，21 的结合位点数据集和非结合位点数据集，为了排除窗口大小对特征选择的干扰，针对同一特征同一算法下，对不同窗口大小(15、17、19 和 21)进行比较，结果发现：综合 5 个评价指标，基本上窗口大小为 19 时在 Na^+和 K^+结合位点预测中性能最好，因此 w = 19 作为后续模型构建的固定窗口大小。图 2 展示了氨基酸组分特征，在不同窗口大小(15、17、19 和 21)预测的 AUC 指标(注：其他指标的对比结果此处省略)。

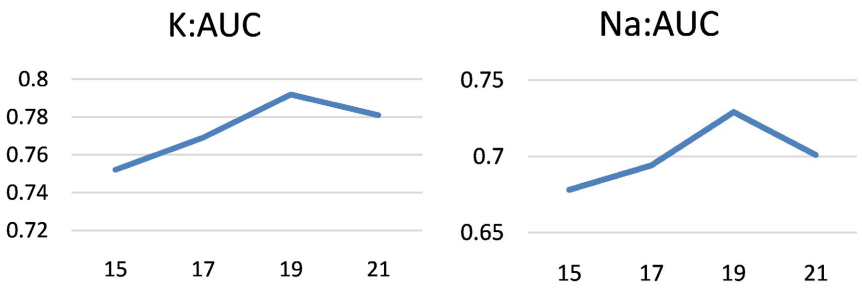


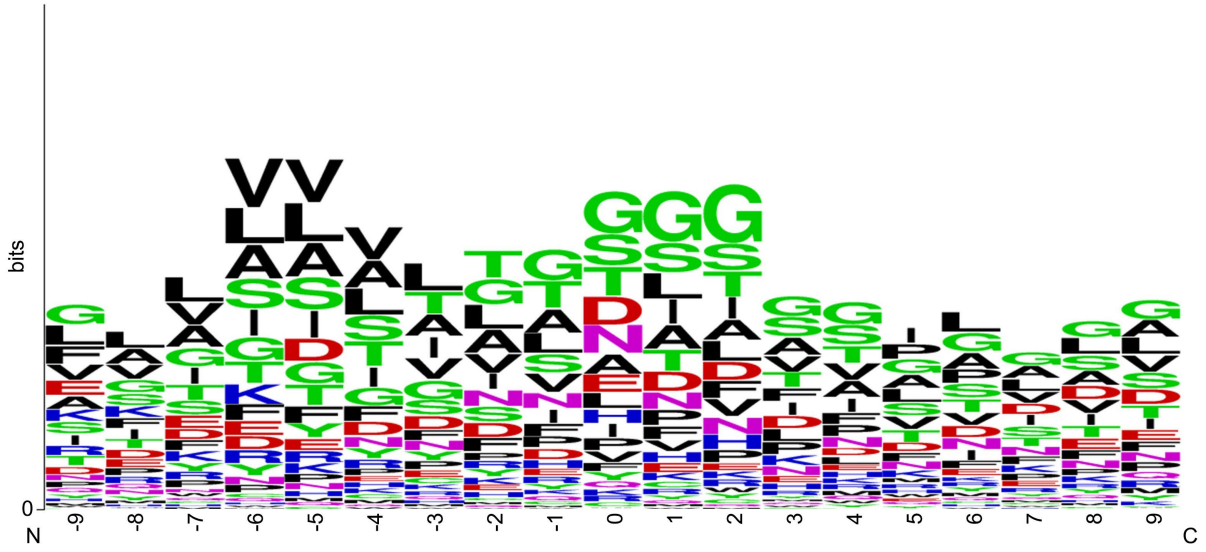
Figure 2. Comparison of AUC performance by different window
图 2. 不同窗口 AUC 性能比较

3.3. 结合/非结合位点保守性分析

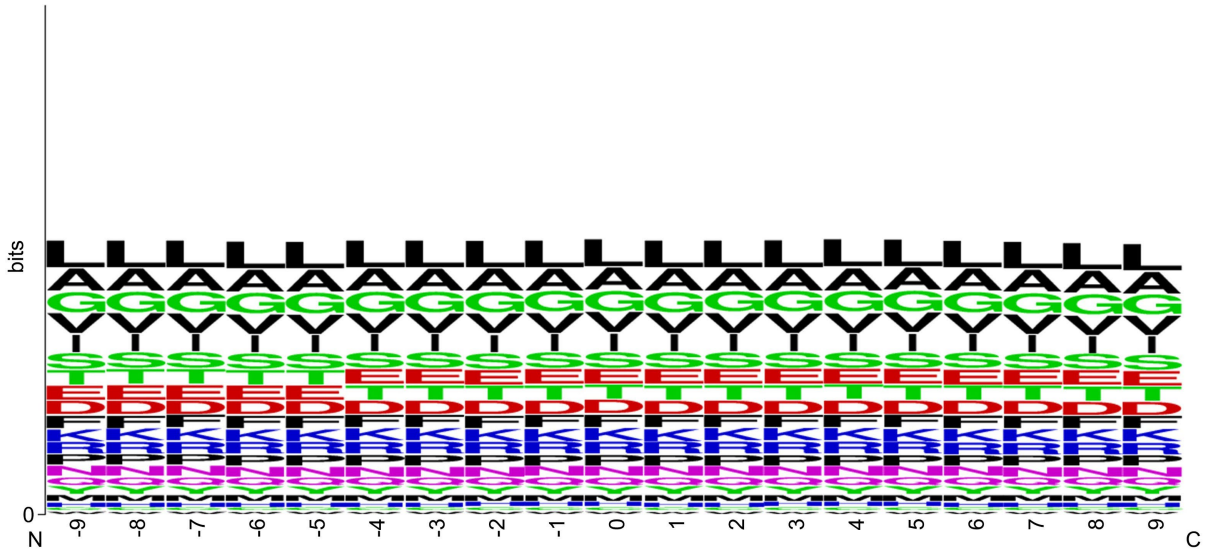
本文利用 WebLogo 在线工具对 K^+与 Na^+结合位点及其对应的非结合位点序列进行了保守性分析[31]。图 3 与图 4 分别展示了钾离子和钠离子结合位点(a)与非结合位点(b)的序列保守性分布情况。

从 K^+结合位点(图 3(a))可见，序列中心残基处的氨基酸字母堆叠高度显著高于两端，说明该位置具有较强的序列保守性，可能是离子直接参与结合的关键区域。整体上，以结合位点为中心的窗口呈现出明显的“山峰”分布特征，表明核心残基在进化过程中被高度保留。结合位点周围(±3 位点)仍存在多个次高峰，提示这些位置的氨基酸在稳定离子结合构象中具有辅助作用。而在非结合位点(图 3(b))中，各位置字母堆叠高度较低，整体保守性较差，说明这些区域在进化过程中变异较快，与离子结合的功能关联

较弱。



(a)



(b)

Figure 3. Conservatism of K^+ binding sites/non-binding sites. (a) Conservatism of the K^+ binding site; (b) Conservation of the K^+ non-binding site

图 3. K^+结合位点/非结合位点的保守性分析。(a) K^+结合位点的保守性；(b) K^+非结合位点的保守性

对于 Na^+结合位点(图 4(a))，其中心区域同样存在明显峰值，其中氨基酸 S、D、G 的出现频率较高，推测这些残基可能与 Na^+离子形成直接相互作用。与 K^+相比，Na^+的保守性峰值更集中且幅度略高，说明钠离子结合位点在序列上具有更显著的保守性特征。相对地，非结合位点(图 4(b))中氨基酸分布较为分散，缺乏明显的高堆叠区域，整体序列保守性较低。综上所述，K^+与 Na^+的结合位点均表现出较高的进化保守性，而非结合位点保守性较低。这一现象说明，离子结合残基在蛋白质功能实现中具有关键作用，并在进化过程中受到强烈的保留选择压力。尤其是中心结合残基及其邻近区域的高保守性，提示这些区域可能是维持离子识别与结合稳定性的主要结构基础。

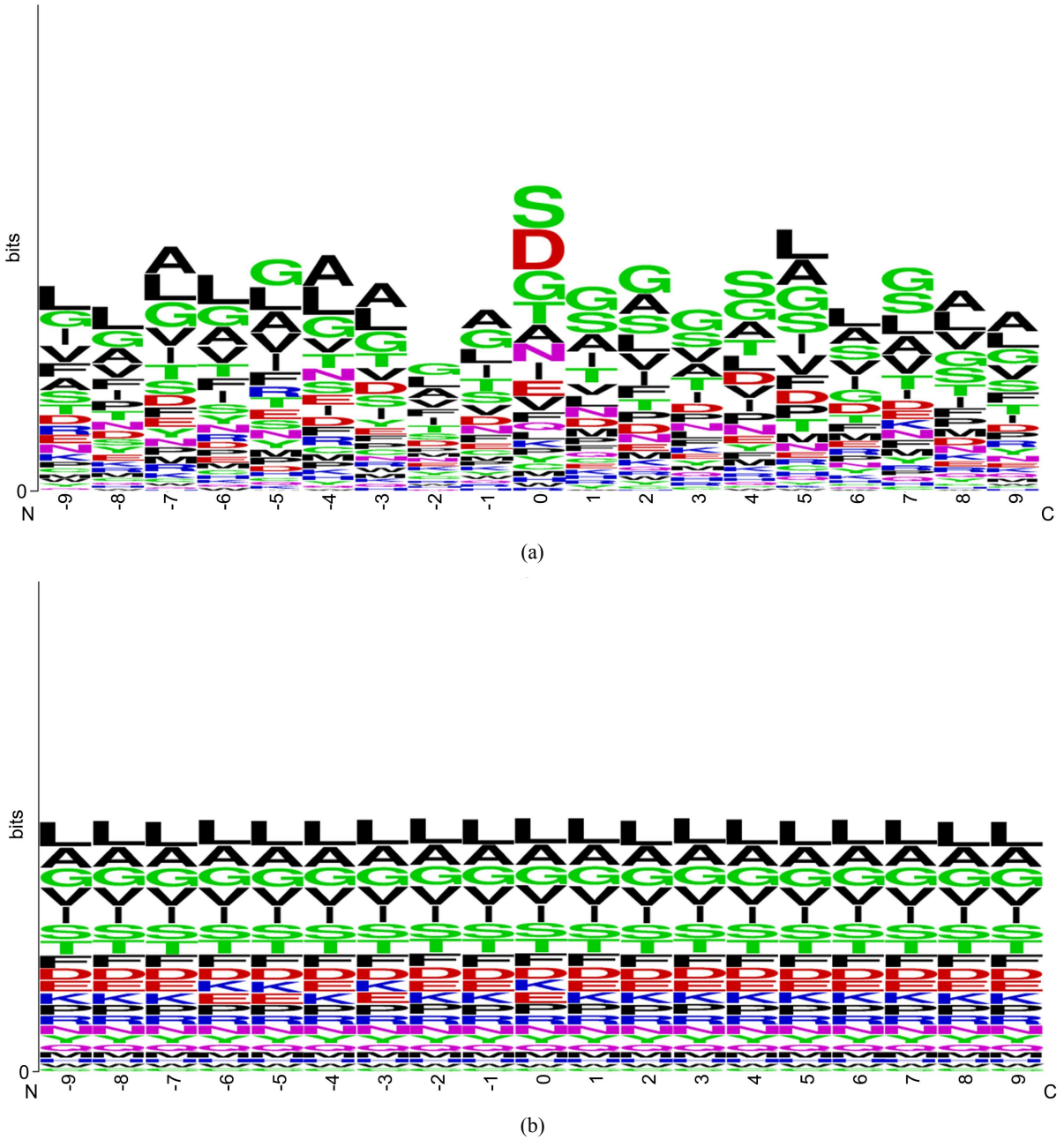


Figure 4. Conservative of Na^+ binding sites/non-binding sites. (a) Conservatism of Na^+ binding sites; (b) Conservatism of Na^+ non-binding sites

图 4. Na^+结合位点/非结合位点的保守性分析。(a) Na^+结合位点的保守性；(b) Na^+非结合位点的保守性

3.4. 单特征预测结果

在本研究中，我们从序列、结构和能量三个角度提取了九类特征：PSSM、氨基酸组分、密码子频率、相对可及表面积(RASA)、综合特征(SASA-RASA)、图能量、疏水性、二级结构和结合能。采用七种机器学习算法：逻辑回归(LR)、随机森林(RF)、梯度提升树(GB)、XGBoost(XB)、LightGBM(LG)、支持向量机(SVM)和 K 近邻算法(KNN)对钾离子(K^+)和钠离子(Na^+)结合位点进行了建模预测，结果如表 6 与表 7 所示。

Table 6. Single-feature prediction results of K^+
表 6. K^+的单特征预测结果

算法	特征参数	Sn	Sp	Acc	Mcc	Auc
LR	1	**1.000**	**0.853**	**0.856**	**0.318**	**0.984**
	2	0.710	0.784	0.782	0.154	0.792
	3	0.681	0.789	0.788	0.148	0.798
	4	0.507	0.727	0.724	0.068	0.633
	5	0.739	0.460	0.464	0.052	0.490
	6	0.725	0.630	0.631	0.095	0.706
	7	0.754	0.540	0.543	0.076	0.693
	8	0.493	0.647	0.645	0.038	0.559
	9	0.725	0.658	0.659	0.104	0.740
SVM	1	0.917	0.924	0.924	0.401	0.975
	2	0.130	0.948	0.934	0.045	0.690
	3	0.203	0.925	0.913	0.063	0.699
	4	0.377	0.863	0.854	0.089	0.627
	5	0.551	0.534	0.534	0.022	0.569
	6	0.580	0.604	0.603	0.049	0.620
	7	0.638	0.659	0.659	0.081	0.701
	8	0.362	0.638	0.633	0.000	0.475
	9	0.667	0.680	0.680	0.096	0.668
LG	1	0.500	0.981	0.971	0.398	0.968
	2	0.116	0.988	0.973	0.116	0.751
	3	0.145	0.984	0.969	0.124	0.711
	4	0.391	0.739	0.733	0.039	0.634
	5	0.377	0.681	0.676	0.016	0.531
	6	0.261	0.758	0.750	0.006	0.597
	7	0.565	0.694	0.691	0.073	0.651
	8	0.130	0.856	0.844	−0.005	0.498
	9	0.362	0.750	0.743	0.033	0.603
XB	1	0.972	0.850	0.852	0.305	0.971
	2	0.145	0.987	0.972	0.138	0.710
	3	0.130	0.991	0.976	0.148	0.732
	4	0.377	0.766	0.760	0.044	0.625
	5	0.333	0.620	0.615	−0.012	0.532

续表

	6	0.304	0.725	0.718	0.008	0.573
	7	0.565	0.691	0.689	0.072	0.646
	8	0.174	0.821	0.810	−0.002	0.463
	9	0.406	0.715	0.710	0.035	0.572
GB	1	1.000	0.743	0.748	0.231	0.911
	2	0.145	0.960	0.946	0.069	0.744
	3	0.275	0.943	0.931	0.118	0.730
	4	0.420	0.777	0.771	0.061	0.644
	5	0.406	0.698	0.693	0.029	0.564
	6	0.420	0.762	0.757	0.056	0.651
	7	0.565	0.670	0.668	0.065	0.630
	8	0.217	0.855	0.845	0.027	0.549
	9	0.420	0.771	0.765	0.059	0.659
RF	1	0.444	0.942	0.932	0.215	0.917
	2	0.000	1.000	0.982	−0.003	0.727
	3	0.145	0.989	0.974	0.151	0.728
	4	0.246	0.847	0.837	0.034	0.579
	5	0.565	0.580	0.579	0.038	0.533
	6	0.333	0.696	0.689	0.008	0.528
	7	0.478	0.723	0.719	0.058	0.613
	8	0.174	0.736	0.727	−0.027	0.429
	9	0.464	0.633	0.630	0.026	0.545
KNN	1	0.000	1.000	0.981	0.000	0.628
	2	0.406	0.850	0.843	0.092	0.625
	3	0.348	0.848	0.839	0.070	0.636
	4	0.261	0.864	0.854	0.047	0.583
	5	0.493	0.545	0.544	0.010	0.519
	6	0.420	0.622	0.619	0.011	0.515
	7	0.580	0.653	0.652	0.063	0.624
	8	0.304	0.569	0.564	−0.033	0.436
	9	0.464	0.580	0.578	0.011	0.526

注：表中“LR”为逻辑回归算法，“LG”为 LightGBM 算法，“XB”为 XGBoost 算法，“GB”为 Gradient Boosting 梯度提升算法，“RF”为随机森林算法，“1”为 pssm，“2”为氨基酸组分，“3”为密码子频率，“4”为相对可及表面积，“5”为二级结构，“6”为疏水性，“7”为 SASA-RASA，“8”为结合能，“9”为图能量。表中加粗数据代表最佳预测结果。

Table 7. Single-feature prediction results of Na^+
表 7. Na^+的单特征预测结果

算法	特征参数	Sn	Sp	Acc	Mcc	Auc
LR	1	**1.000**	**0.865**	**0.866**	**0.246**	**0.978**
	2	0.512	0.798	0.795	0.081	0.729
	3	0.451	0.802	0.798	0.067	0.723
	4	0.720	0.740	0.740	0.110	0.758
	5	0.793	0.083	0.091	−0.047	0.424
	6	0.610	0.663	0.663	0.061	0.677
	7	0.768	0.587	0.589	0.076	0.760
	8	0.146	0.852	0.844	−0.001	0.525
	9	0.634	0.543	0.544	0.038	0.632
SVM	1	0.970	0.924	0.924	0.319	0.974
	2	0.110	0.948	0.938	0.027	0.585
	3	0.159	0.919	0.910	0.030	0.594
	4	0.500	0.884	0.880	0.124	0.752
	5	0.744	0.287	0.292	0.007	0.487
	6	0.476	0.665	0.663	0.031	0.551
	7	0.683	0.762	0.761	0.109	0.768
	8	0.329	0.653	0.650	−0.004	0.505
	9	0.585	0.518	0.518	0.022	0.589
LG	1	0.606	0.956	0.953	0.258	0.955
	2	0.024	0.992	0.981	0.020	0.630
	3	0.024	0.985	0.974	0.008	0.612
	4	0.488	0.751	0.742	0.062	0.640
	5	0.524	0.441	0.442	−0.007	0.488
	6	0.159	0.834	0.827	−0.002	0.484
	7	0.573	0.694	0.692	0.061	0.679
	8	0.171	0.872	0.864	0.013	0.536
	9	0.512	0.639	0.637	0.033	0.594
XB	1	0.939	0.870	0.871	0.234	0.954
	2	0.024	0.990	0.979	0.014	0.597
	3	0.037	0.989	0.978	0.025	0.587
	4	0.488	0.751	0.748	0.058	0.646
	5	0.585	0.411	0.413	−0.001	0.484

续表

	6	0.183	0.790	0.783	−0.007	0.477
	7	0.524	0.705	0.703	0.053	0.665
	8	0.244	0.818	0.811	0.017	0.530
	9	0.537	0.586	0.586	0.026	0.591
GB	1	0.970	0.826	0.828	0.206	0.933
	2	0.037	0.982	0.972	0.015	0.672
	3	0.146	0.961	0.952	0.058	0.662
	4	0.415	0.788	0.784	0.052	0.635
	5	0.561	0.395	0.397	−0.009	0.469
	6	0.195	0.877	0.869	0.023	0.566
	7	0.598	0.710	0.709	0.071	0.702
	8	0.134	0.914	0.906	0.018	0.516
	9	0.402	0.723	0.719	0.029	0.602
RF	1	0.364	0.946	0.940	0.133	0.905
	2	0.000	1.000	0.988	−0.002	0.688
	3	0.024	0.995	0.984	0.026	0.649
	4	0.195	0.915	0.907	0.041	0.575
	5	0.598	0.373	0.376	−0.006	0.507
	6	0.183	0.739	0.733	−0.019	0.495
	7	0.439	0.760	0.757	0.049	0.616
	8	0.244	0.716	0.711	−0.009	0.532
	9	0.610	0.516	0.517	0.027	0.580
KNN	1	0.000	1.000	0.990	0.000	0.529
	2	0.183	0.861	0.853	0.013	0.547
	3	0.122	0.866	0.857	−0.004	0.508
	4	0.183	0.922	0.914	0.041	0.580
	5	0.488	0.529	0.529	0.004	0.510
	6	0.317	0.669	0.665	−0.003	0.494
	7	0.512	0.681	0.679	0.044	0.620
	8	0.439	0.585	0.583	0.005	0.514
	9	0.622	0.480	0.481	0.021	0.551

注：表中“LR”为逻辑回归算法，“LG”为 LightGBM 算法，“XB”为 XGBoost 算法，“GB”为 Gradient Boosting 梯度提升算法，“RF”为随机森林算法，“1”为 pssm，“2”为氨基酸组分，“3”为密码子频率，“4”为相对可及表面积，“5”为二级结构，“6”为疏水性，“7”为 SASA-RASA，“8”为结合能，“9”为图能量。表中加粗数据代表最佳预测结果。

从整体结果来看，两类离子在不同算法下的预测规律表现出较高一致性，说明模型在特征分布上具有一定的泛化性。在 K^+结合位点的预测中，PSSM 作为特征取得最优预测精度，其结果明显优于其他特征的结果。Na^+结合位点预测中与 K^+的类似，各算法在特征(pssm)上表现突出，但部分指标的绝对数值略低于 K^+的。这表明进化保守性是识别金属离子结合位点最关键的特征信息来源。由于结合位点通常在进化过程中被高度保留，PSSM 特征能够有效区分结合与非结合区域。但是，两类离子在各算法中不同特征下的表现存在一定的差异。鉴于梯度提升算法在处理非线性关系的能力较强，同时对特征交互的捕捉更高效，对数据的适应性更好，后续选择 LightGBM 算法进行特征融合分析。

3.5. 特征融合的预测结果

在金属离子结合位点预测中，单一特征往往只能反映结合残基的某一方面特性，而离子结合是一个由多种因素共同决定的复杂过程。特征融合可在模型层面整合不同来源特征的互补信息，从而提升预测精度和泛化能力。因此，接下来以最优 PSSM 特征为核心，选用 LightGBM 算法，对九类特征进行逐步融合，结果如表 8 和表 9 所示。总体来看，随着融合特征数量的增加，模型性能呈现上升趋势，尤其在引入结构相关特征(相对可及表面积、溶剂可及表面积)后，AUC 与 MCC 均有所提高。

Table 8. Feature fusion results of K^+
表 8. K^+的特征融合结果

特征融合	Sn	Sp	Acc	Auc	Mcc
1	0.889	0.763	0.766	0.935	0.208
1 + 2	0.889	0.775	0.778	0.950	0.216
1 + 2 + 3	0.917	0.770	0.773	0.957	0.221
1 + 2 + 3 + 4	**0.972**	0.851	0.853	**0.961**	0.306
1 + 2 + 3 + 4 + 5	0.833	0.883	0.882	0.949	0.294
1 + 2 + 3 + 4 + 5 + 6	0.833	0.888	0.887	0.954	0.300
1 + 2 + 3 + 4 + 5 + 6 + 7	0.806	0.887	0.885	0.941	0.287
1 + 2 + 3 + 4 + 5 + 6 + 7 + 8	0.861	**0.891**	0.890	0.951	**0.315**
1 + 2 + 3 + 4 + 5 + 6 + 7 + 8 + 9	0.806	0.894	**0.892**	0.952	0.297

注：表中“1”为 pssm，“2”为氨基酸组分，“3”为密码子频率，“4”为相对可及表面积，“5”为二级结构，“6”为疏水性，“7”为 SASA-RASA，“8”为结合能，“9”为图能量。加粗数据表示特征融合后的最佳结果。

Table 9. Feature fusion results of Na^+
表 9. Na^+的特征融合结果

特征融合	Sn	Sp	Acc	Auc	Mcc
1	1.000	0.834	0.836	0.934	0.219
1 + 2	1.000	0.830	0.832	0.950	0.216
1 + 2 + 3	1.000	0.828	0.830	0.948	0.215
1 + 2 + 3 + 4	0.970	0.919	**0.919**	**0.982**	**0.309**
1 + 2 + 3 + 4 + 5	1.000	0.917	0.918	**0.983**	**0.316**
1 + 2 + 3 + 4 + 5 + 6	1.000	0.916	0.917	0.981	0.315
1 + 2 + 3 + 4 + 5 + 6 + 7	1.000	0.916	0.917	**0.984**	0.314
1 + 2 + 3 + 4 + 5 + 6 + 7 + 8	1.000	0.915	0.916	0.984	0.313
1 + 2 + 3 + 4 + 5 + 6 + 7 + 8 + 9	1.000	0.914	0.915	0.984	0.310

注：表中“1”为 pssm，“2”为氨基酸组分，“3”为密码子频率，“4”为相对可及表面积，“5”为二级结构，“6”为疏水性，“7”为 SASA-RASA，“8”为结合能，“9”为图能量。加粗数据表示特征融合后的最佳结果。

对于 K^+结合位点预测，当模型加入相对可及表面积特征后，AUC 由 0.935 提升至 0.961，性能显著改善；在融合至第 4 个特征时(PSSM + AAC + CUF + RASA)，模型取得最高敏感性(0.972)。在逐步融合的过程中发现特征融合不是特征越多越好。同样，对于 Na^+结合位点预测，模型在融合至第 5 个特征(PSSM、氨基酸组分、密码子频率、相对可及表面积、二级结构)时性能最优(AUC = 0.983, MCC = 0.316)，之后加入能量相关特征(疏水性、SASA-RASA、结合能、图能量)模型性能趋于稳定。

这些结果表明，特征融合能够在一定程度上整合进化、结构及理化层面的互补信息，使模型在综合性能与稳定性方面优于单一特征模型。但是随着多特征融合时，可能会出现信息冗余，使得预测精度降低。故合理的特征组合与选择对性能提升尤为关键。

4. 结论

本文基于多源特征与多种机器学习算法，对钾离子(K^+)和钠离子(Na^+)的蛋白质结合位点进行了系统预测与分析。首先，从 BioLiP 数据库获取 K^+与 Na^+结合蛋白序列，经过 CD-HIT 去冗余后按比例划分训练集与测试集；从序列、结构及能量三个层面共提取 9 类特征。随后，采用 SMOTEENN 方法平衡样本分布，并利用七种机器学习算法(LR, SVM, KNN, RF, GB, XGBoost, LightGBM)构建预测模型。结果表明，PSSM 特征在两类离子结合位点的预测中表现最优，明显优于其他单一特征，显示出进化保守性信息在离子结合位点识别中的重要作用。在此基础上，进一步以 PSSM 为核心进行特征融合分析，结果显示，融合多维特征虽能有效提升预测性能，但是特征数量的增加并非总能带来性能提升，部分特征间存在一定信息冗余，故合理的特征选择与融合策略对模型优化至关重要。综上所述，本文提出的机器学习算法结合多特征预测框架能够有效识别钾、钠离子结合位点，为进一步研究钠、钾离子通道相关的药物靶点识别提供了计算依据，也为阐明钠、钾离子在神经信号传导与细胞渗透压调控等生命过程中的分子机制提供了一些参考。

基金项目

国家自然科学基金(编号：62262050)。

参考文献

[1] 邹向辉, 冯永娥. 基于氨基酸理化特征识别疾病相关的蛋白质与金属离子配体的结合位点[J]. 内蒙古农业大学学报(自然科学版), 2024, 45(2): 78-85.

[2] Denesyuk, A.I., Permyakov, S.E., Johnson, M.S., Permyakov, E.A. and Denessiouk, K. (2017) Building Kit for M 等 Cation Binding Sites in Proteins. *Biochemical and Biophysical Research Communications*, **494**, 311-317. https://doi.org/10.1016/j.bbrc.2017.10.034

[3] 孙锴. 基于深度学习算法识别蛋白质-金属离子配体结合位点[D]: [硕士学位论文]. 呼和浩特: 内蒙古工业大学, 2021.

[4] Ahmad, S. and Sarai, A. (2005) PSSM-Based Prediction of DNA Binding Sites in Proteins. *BMC Bioinformatics*, **6**, Article No. 33. https://doi.org/10.1186/1471-2105-6-33

[5] Beckstette, M., Homann, R., Giegerich, R. and Kurtz, S. (2006) Fast Index Based Algorithms and Software for Matching Position Specific Scoring Matrices. *BMC Bioinformatics*, **7**, Article No. 389. https://doi.org/10.1186/1471-2105-7-389

[6] 陈梦淇. 蛋白质相对溶剂可及性与相互作用位点的计算建模研究[D]: [硕士学位论文]. 苏州: 苏州大学, 2024.

[7] 施绍萍. 基于支持向量机的蛋白质功能预测新方法研究[D]: [博士学位论文]. 南昌: 南昌大学, 2012.

[8] 王兵. 蛋白质相互作用及其位点的预测方法研究[D]: [博士学位论文]. 合肥: 中国科学技术大学, 2006.

[9] 魏志森, 杨静宇. 基于加权 PSSM 直方图和随机森林集成的蛋白质交互作用位点预测[J]. 南京理工大学学报, 2015, 39(4): 379-385.

[10] 安计勇. 基于相关向量机的蛋白质相互作用预测研究[D]: [博士学位论文]. 徐州: 中国矿业大学, 2018.

[11] Deen, A.J. and Gyanchandani, M. (2020) Pseudo Amino Acid Feature-Based Protein Function Prediction Using Support Vector Machine and K-Nearest Neighbors. *International Journal of Advanced Computer Science and Applications*, **11**, 187-195. https://doi.org/10.14569/ijacsa.2020.0110922

[12] 刘天宇. 基于集成支持向量机与随机森林的蛋白交互预测研究[D]: [硕士学位论文]. 长春: 东北师范大学, 2019.

[13] 周畅. 基于氨基酸序列多尺度编码的梯度提升树蛋白质交互作用预测算法研究[D]: [硕士学位论文]. 天津: 天津大学, 2018.

[14] 黄国华, 王攀, 张桂阳. 一种基于深度学习和 XGBoost 的蛋白质-蛋白质相互作用位点预测方法[P]. 中国专利, 113611360A. 2021-11-05.

[15] 陈焕超, 魏志森, 於东军, 等. 基于 LightGBM 的蛋白质类泛素化修饰位点预测[J]. 南京理工大学学报, 2022, 46(2): 156-163.

[16] Yang, J., Roy, A. and Zhang, Y. (2012) BioLiP: A Semi-Manually Curated Database for Biologically Relevant Ligand–protein Interactions. *Nucleic Acids Research*, **41**, D1096-D1103. https://doi.org/10.1093/nar/gks966

[17] Li, W. and Godzik, A. (2006) CD-Hit: A Fast Program for Clustering and Comparing Large Sets of Protein or Nucleotide Sequences. *Bioinformatics*, **22**, 1658-1659. https://doi.org/10.1093/bioinformatics/btl158

[18] Husain, G., Nasef, D., Jose, R., Mayer, J., Bekbolatova, M., Devine, T., *et al.* (2025) SMOTE vs. SMOTEENN: A Study on the Performance of Resampling Algorithms for Addressing Class Imbalance in Regression Models. *Algorithms*, **18**, Article 37. https://doi.org/10.3390/a18010037

[19] Cascarina, S.M. and Ross, E.D. (2025) Protein Activities Driven by Amino Acid Composition. *Journal of Biological Chemistry*, **301**, Article ID: 110640. https://doi.org/10.1016/j.jbc.2025.110640

[20] Alves, N.A., Aleksenko, V. and Hansmann, U.H.E. (2005) A Simple Hydrophobicity-Based Score for Profiling Protein Structures. *Journal of Physics*: *Condensed Matter*, **17**, S1595-S1606. https://doi.org/10.1088/0953-8984/17/18/015

[21] Zhang, S., Hahn, D.F., Shirts, M.R. and Voelz, V.A. (2021) Expanded Ensemble Methods Can Be Used to Accurately Predict Protein-Ligand Relative Binding Free Energies. *Journal of Chemical Theory and Computation*, **17**, 6536-6547. https://doi.org/10.1021/acs.jctc.1c00513

[22] Wu, H., Zhang, Y., Chen, W. and Mu, Z. (2015) Comparative Analysis of Protein Primary Sequences with Graph Energy. *Physica A*: *Statistical Mechanics and Its Applications*, **437**, 249-262. https://doi.org/10.1016/j.physa.2015.04.017

[23] Xu, D., Xu, H., Zhang, Y., Chen, W. and Gao, R. (2020) Protein-Protein Interactions Prediction Based on Graph Energy and Protein Sequence Information. *Molecules*, **25**, Article 1841. https://doi.org/10.3390/molecules25081841

[24] 吴宪明, 吴松锋, 任大明, 等. 密码子偏性的分析方法及相关研究进展[J]. 遗传, 2007(4): 420-426.

[25] Jeong, J.C., Lin, X. and Chen, X. (2011) On Position-Specific Scoring Matrix for Protein Function Prediction. *IEEE*/*ACM Transactions on Computational Biology and Bioinformatics*, **8**, 308-315. https://doi.org/10.1109/tcbb.2010.93

[26] 刘柏丽. 蛋白质二级结构预测 PSIPRED 方法的改进及其应用[D]: [硕士学位论文]. 长沙: 湖南大学, 2019.

[27] 张晓朦. 全原子编码方式预测蛋白质的溶剂可及性[D]: [硕士学位论文]. 大连: 大连理工大学, 2011.

[28] 张弘, 王慧洁, 鲁睿捷, 等. 蛋白质结构预测模型 αFold2 的应用进展[J]. 生物工程学报, 2024, 40(5): 1406-1420.

[29] 王攀文, 龚新奇, 李春华, 等. 蛋白质表面模块划分及其在结合位点预测中的应用[J]. 物理化学学报, 2012, 28(11): 2729-2734.

[30] 冯永娥, 孙鹏哲, 张强. 固有无序蛋白与结合配体作用位点的分析与预测[J]. 内蒙古大学学报(自然科学版), 2023, 54(4): 442-448.

[31] Crooks, G.E., Hon, G., Chandonia, J. and Brenner, S.E. (2004) WebLogo: A Sequence Logo Generator. *Genome Research*, **14**, 1188-1190. https://doi.org/10.1101/gr.849004

附　录

Table A1. Values of graph energy and binding energy of 20 amino acids
表 A1. 20 种氨基酸的图能与结合能的值

氨基酸	图能	结合能
A	3.9569	5.0336
C	3.8253	4.4658
D	3.9568	4.6337
E	4.123	4.5491
F	5.099	5.2015
G	5.0575	5.3044
H	4.1144	4.7079
I	5.3749	5.4214
K	5.099	5.3334
L	4.1147	4.883
M	4.4722	4.7975
N	5.099	5.2015
P	5.0575	5.3044
Q	4.123	4.5491
R	4.6685	5.1529
S	3.3573	4.5236
T	3.6742	4.0339
V	4.0372	4.6791
W	4.4722	4.7975
Y	5.099	5.2015