DOI: 10.23977/medbm.2024.020218
ISSN 2616-2091 Vol. 2 Num. 2

SVM-Based Prediction of Protein Methylation Sites: A Comprehensive Analysis of 553 Properties from the AAindex Database

Xi Su, Mingjun Tang, Zipin Zhao, Ning Zhang[a,*]

Tianjin Key Laboratory of Brain Science and Neuroengineering, Department of Biomedical Engineering, Medical School of Tianjin University, Tianjin, 300072, China
[a]*zhni@tju.edu.cn*
[*]*Corresponding author*

Keywords: Protein Methylation; Support Vector Machine (SVM); AAindex Database; Feature Selection

Abstract: Identifying protein methylation sites experimentally is a challenging and costly task, leading to increased reliance on machine learning-based computational predictors to enhance efficiency. This study aims to improve these predictors through a comprehensive analysis of 553 properties from the AAindex database. We employed support vector machine (SVM) models and utilized 10-fold cross-validation for model evaluation to identify optimal feature combinations for predicting lysine and arginine methylation. The results indicate that the feature set "RACS820104+FUKS010109" yielded the highest performance for lysine methylation, with a *Recall* (*Re*) of 71.11%, *Precision* (*Pre*) of 75.68%, *Accuracy* (*Acc*) of 74.12%, and a *Matthews Correlation Coefficient* (*MCC*) of 0.48. For arginine methylation, the feature set "BAEK050101+CHAM810101" achieved a *Recall* (*Re*) of 74.60%, *Precision* (*Pre*) of 81.08%, *Accuracy* (*Acc*) of 78.60%, and an *MCC* of 0.57. Furthermore, this study explores hydrophobicity as a potentially valuable property for distinguishing methylation from malonylation. This thorough analysis enhances our understanding of the available physicochemical properties, which could lead to the development of more accurate and reliable prediction models.

1. Introduction

Protein post-translational modifications (PTMs) are essential chemical changes that significantly enhance the structural and functional diversity of proteins. Among these modifications, protein methylation—primarily targeting arginine or lysine residues—is a reversible PTM crucial for various biological functions including signal transduction and gene expression regulation [1]. Identifying protein methylation sites is vital for understanding the molecular mechanisms and implications of these modifications in related pathological pathways. Traditional experimental methods for site identification are often labor-intensive and time-consuming. This limitation has sparked a growing interest in computational approaches that leverage machine learning, which have shown remarkable performance in recent years.

Feature extraction is a pivotal aspect of machine learning predictions, significantly influencing their accuracy. Some studies have underscored the importance of physicochemical properties from the amino acid index (AAindex) database [2] as informative features for predictive models. Li et al. [3] introduced the Methy_SVMIACO predictor, which encoded sequence fragments using 183 AAindex indices. The PSSMe model [4] evaluated 544 properties from the AAindex database (version 9.2) and selected the top 3 properties that yielded the highest prediction accuracy for subsequent studies. However, despite evaluating various properties from the AAindex database, the specific AAindex IDs—unique identification numbers assigned to each physicochemical property—used in these analyses are often not provided in detail.

In this study, we thoroughly explored the potential utility of diverse physicochemical properties in predicting methylation sites. Our findings reveal the most effective properties for predicting lysine and arginine methylation, offering valuable resources for future research to directly integrate these specific physicochemical properties using their corresponding AAindex IDs.

2. Methods

2.1 Datasets

We derived the positive and negative datasets for methylation from the benchmark dataset provided by the dbPTM[5] database (Version 9.2). In our preliminary analysis, the positive dataset comprised 7,281 lysine methylation sites and 7,263 arginine methylation sites, while the negative dataset included 18,094 non-methylated lysine sites and 17,970 non-methylated arginine sites. To ensure unbiased prediction outcomes, we randomly sampled approximately 40% of non-methylated sequences from the negative dataset and combined them with all methylated sequences to create a balanced dataset. For optimal feature subset selection, we utilized 80% of the balanced dataset as the training set, reserving the remaining 20% as the independent test set. Additionally, to explore distinguishing properties for different PTMs, we gathered datasets for other PTMs from the dbPTM database, including 7,634 lysine malonylation sites, which closely matched the ratio of 1:1 with methylation sites. Consequently, we downloaded the positive dataset for lysine malonylation to facilitate subsequent analysis. Detailed statistics of the datasets are presented in Table 1.

Table 1: Statistics on the number of lysine and arginine sites in the study

PTM Type	Residue type	Number of Positive sites	Number of Negative sites (all/randomly extract)	Training (80%)	Test (20%)
Methylation	Lysine	7,281	18,094/7,281	5,825	1,456
Methylation	Arginine	7,263	17,970/7,263	5,810	1,453
Malonylation	Lysine	7,634	—	—	—

2.2 AAindex database

The AAindex database [2] provides a comprehensive collection of numerical indices that reflect various physicochemical properties of amino acids and their pairs. The database is organized into three distinct sections: AAindex1, which contains 20 numerical indices for amino acids; AAindex2, which consists of an amino acid mutation matrix; and AAindex3, which offers statistical protein contact potentials. For the purpose of this study, we focused on utilizing AAindex1, which is based on data derived from published literature. The latest version of the AAindex database comprises a total of 566 distinct properties, and we excluded any properties that contained missing values ("NA"). As a result, we retained 553 properties for subsequent analysis, ensuring a robust foundation for our research.

2.3 Feature extraction

In this study, we utilized the physicochemical properties extracted from the AAindex database to transform sequence fragments, each consisting of 21 residues, into 21-dimensional numerical vectors. For illustration, we selected the property "ANDN920101" and applied it to the lysine methylation-positive peptide sequence "ELDTLSEESYKDSTLIMQLLR". The corresponding values for the 20 amino acids associated with the "ANDN920101" property in the AAindex database are as follows: (A: 4.35, C: 4.65, D: 4.76, E: 4.29, F: 4.66, G: 3.97, H: 4.63, I: 3.95, K: 4.36, L: 4.17, M: 4.52, N: 4.75, P: 4.44, Q: 4.37, R: 4.38, S: 4.50, T: 4.35, V: 3.95, W: 4.70, Y: 4.60). The peptide "ELDTLSEESYKDSTLIMQLLR" was then converted into the following 21-dimensional vector representation: [4.29, 4.17, 4.76, 4.35, 4.17, 4.50, 4.29, 4.29, 4.50, 4.60, 4.36, 4.76, 4.50, 4.35, 4.17, 3.95, 4.52, 4.37, 4.17, 4.17, 4.38]. We applied this feature extraction process to all peptide sequences in the dataset, resulting in a feature matrix with 7,281 rows (corresponding to sequence fragments in the dataset) and 21 columns (representing the 21-dimensional numerical vectors).

Following the application of this feature extraction process across all properties in the AAindex database, we generated a total of 553 matrices specifically for the lysine methylation-positive dataset. Considering the inclusion of five different types of datasets, we ultimately produced a cumulative total of 2,765 matrices (calculated as 553×5). These 21-dimensional numerical vectors were subsequently input into the selected machine learning algorithms for model training and further analysis, facilitating our exploration of the relationship between physicochemical properties and methylation.

2.4 Evaluation

In this study, we employed a 10-fold cross-validation strategy for our analysis. The dataset was divided into 10 equal subsets, with 9 subsets used for training and 1 subset used for testing in each iteration. This process was repeated 10 times, and the average performance across all iterations was calculated. The evaluation metrics used in this study included *Accuracy* (*Acc*), *Recall* (*Re*), *Precision* (*Pre*), and *F1 Score*. In addition to these metrics, we also presented *the Receiver Operating Characteristic* (*ROC*) curve and *the area under the ROC curve* (*AUC*). Furthermore, *the Matthews Correlation Coefficient* (*MCC*) was incorporated as an additional evaluation metric. Details of these metrics are provided in Table 2.

Table 2: Evaluation metrics.

Evaluation metric	Abbreviation	Formula
Accuracy	*Acc*	$\frac{TP+TN}{TP+FP+TN+FN}$
Recall	*Re*	$\frac{TP}{TP+FN}$
Precision	*Pre*	$\frac{TP}{TP+FP}$
F1 score	—	$2\times\frac{\mathrm{Pre}\times\mathrm{Re}}{\mathrm{Pre}+\mathrm{Re}}$
Matthews Correlation Coefficient	*MCC*	$\frac{(TP\times TN)-(FP\times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$

Notes: TP, TN, FP, and *FN* denote the number of true positives, true negatives, false positives, and false negatives, respectively.

2.5 Two-step feature selection

In many cases, single feature-based models may not effectively identify methylation sites. Feature selection in this field can be viewed as a combinatorial optimization problem aimed at determining the feature set that maximizes the predictive model's performance. To address this challenge, a two-step feature selection strategy is implemented in this study. This approach is designed to extract the most optimal and significant features from the 553 properties.

Step 1: Feature Ranking and Incremental Feature Selection (IFS). In the first step, the preliminary analysis generates four feature catalogs. Each catalog ranks features based on their performance across four evaluation metrics: *Recall*, *Accuracy*, *Precision*, and *F1 score*, as assessed in prediction tasks. Subsequently, the Incremental Feature Selection (IFS) method [6] is employed. From the top 10 features identified within these catalogs—selected to ensure the inclusion of significant attributes while also reducing computational costs—10 unique feature sets are systematically constructed. For each feature set, a prediction model is constructed, and 10-fold cross-validation is employed to assess its performance. The feature set that yields the best performance based on the IFS approach is then determined.

Step 2: Combining Prominent Features and Model Redevelopment. In the second step, prominent features from different evaluation metrics identified in the first step are combined. This process begins by selecting the best feature sets based on individual evaluation metrics, such as *Recall*, *Accuracy*, *Precision*, and *F1 score*. These prominent features are then systematically combined to form new feature sets. Subsequently, models are redeveloped using these combined feature sets. Each model is subjected to 10-fold cross-validation to assess its performance. The optimal feature set is ultimately selected based on the highest predictive performance achieved through this integrated approach.

2.6 Model Building

Support Vector Machine (SVM), a supervised machine learning method for binary classification introduced by Vapnik et al. [7], is widely used in methylation site prediction studies. It can map input samples to a high-dimensional space through a kernel function. Then, the hyperplane with the maximum classification margin and minimum error is found in that space to classify the samples into two classes. We adopted the Radial Basis Function (RBF) as the kernel function, which is a commonly used and effective option. In this work, the *svm.SVC* implementation from the scikit-learn [8] machine learning library in Python was utilized. The kernel parameter γ and the penalty parameter C were optimized within predefined ranges. Additional details are provided below:

1) Several models were developed based on the 553 properties, with the kernel function set to RBF and default values used for other parameters to ensure an objective and unbiased comparison.

2) In the initial feature selection step, models were reconstructed according to the ranked catalogs. The parameter settings were kept consistent with those in 1.

3) During the second feature selection phase, the dataset was partitioned into training and test sets in an 8:2 ratio. Model parameters were firstly optimized within specified ranges, the test set was then employed to evaluate the generalization performance of the top 10 models, ensuring that selected parameters did not lead to overfitting on the training data. By comparing the *AUC* values of different models with the optimal parameters on the test set, the optimal feature sets were determined.

3. Results and Discussion

3.1 Lysine Methylation

For lysine methylation prediction, the top 10 AAindex properties utilized by models evaluated using the *Re* metric are detailed in Table 3. It is noteworthy that achieving an 80% *Re* score with just the "RACS820104" property is exceptional. When Incremental Feature Selection (IFS) method was applied, the improvement was not substantial. Ultimately, in the initial feature selection phase, only the "RACS820104" property was selected. This property represents the average relative fractional occurrence in EL(i).

In Table 3, the "FUKS010109" property demonstrated the best performance among the models evaluated using *Acc* as the metric. This property is associated with the amino acid composition of intracellular proteins in thermophiles. The *Pre* metric yielded results similar to those of *Acc*, to enhance the diversity of the feature set, the features assessed under the *Pre* metric were "FUKS010109" and "FUKS010112".

For the *F1 score* evaluation metric, the model utilizing "FUKS010101" exhibited the highest performance and ranked second in *Acc*. Following the incremental improvement observed in IFS, it was decided to retain the top 2 properties: "FUKS010101" and "NAKH900101". This decision was influenced by the notable similarity between "FUKS010101" and "FUKS010109", as well as the need to ensure a diverse feature set compared to the *Acc* metric. "NAKH900101" signifies the amino acid composition of total proteins, providing insights into the sequence composition of proteins.

Table 3: Results of lysine methylation prediction using top 10 AAindex properties and 10-fold cross-validation.

AAindex	***Re.mean***	**AAindex**	***Acc.mean***	**AAindex**	***Pre.mean***	**AAindex**	***F1.mean***
RACS820104	0.800	FUKS010109	0.733	FUKS010109	0.761	FUKS010101	0.711
AURR980102	0.792	FUKS010101	0.721	FUKS010112	0.740	NAKH900101	0.710
RACS820112	0.766	NAKH900101	0.720	FUKS010110	0.740	CEDJ970104	0.708
KARS160118	0.758	CEDJ970104	0.720	FUKS010101	0.737	FUKS010109	0.708
GEOR030104	0.734	FUKS010110	0.720	CEDJ970104	0.734	NAKH920106	0.703
GEOR030103	0.729	FUKS010112	0.717	NAKH900101	0.734	JOND920101	0.700
GRAR740101	0.726	NAKH920106	0.713	NAKH920106	0.727	FUKS010110	0.699
GARJ730101	0.725	JOND920101	0.712	JOND920101	0.725	FUKS010112	0.699
WILM950103	0.725	KUMS000101	0.710	KUMS000101	0.724	NAKH920101	0.693
KRIW710101	0.706	KUMS000102	0.707	KUMS000102	0.718	CEDJ970102	0.693

Initially, the optimal feature subset was identified by comparing 4 property sets: "RACS820104", "FUKS010109", "FUKS010109 + FUKS010112", and "FUKS010101 + NAKH900101". While the sets "FUKS010109 + FUKS010112" and "FUKS010101 + NAKH900101" showed high *Pre*, their *Re* rates were suboptimal, possibly due to the presence of noisy or redundant information in the sets. In contrast, "RACS820104" achieved a *Re* of 69.32%, *Pre* of 61.44%, and *Acc* of 62.90%, and "FUKS010109" had a *Re* of 57.38%, *Pre* of 80.46%, and *Acc* of 71.71%. Combining "RACS820104" and "FUKS010109" might enhance the identification of lysine methylation sites, but constructing a model using these two properties resulted in a decreased *Re* to 30.20%. Normalizing the values and rebuilding the model led to improved performance, with a *Re* of 71.11%, *Pre* of 75.68%, *Acc* of 74.12%, and *MCC* of 0.48 on the test set, surpassing the performance of other property sets.

Table 4 outlines the final selected optimal parameters and performance evaluation on the test set.

In selecting these parameters, we considered the generalization capability of the test set, which helped to mitigate the risk of overfitting. Additionally, the *ROC* curve was plotted to further assess the models' performance, with Fig. 1 displaying the *ROC* plot and the corresponding *AUC* value. The findings indicate that the feature subset "RACS820104 + FUKS010109*" is optimal for lysine methylation prediction.

Table 4: Optimal parameters and performance of models trained with different feature sets on the lysine methylation test set.

Feture Set	Optimal Parameters			Performance			
AAindex ID	Kernel	*C*	γ	*Re*(%)	*Pre*(%)	*Acc*(%)	*MCC*
RACS820104	RBF	1	0.85	69.32	61.44	62.90	0.26
FUKS010109	RBF	5	0.05	57.38	80.46	71.72	0.45
FUKS010109+FUKS010112	RBF	1	0.05	30.20	86.27	62.70	0.33
FUKS010101+NAKH900101	RBF	5	0.05	5.90	94.51	52.78	0.16
RACS820104+FUKS010109	RBF	1	0.05	53.81	83.40	71.55	0.46
RACS820104+FUKS010109*	RBF	5	0.05	71.11	75.68	74.12	0.48

Notes: The highest values are indicated in bold. '*' denotes the normalization operation.

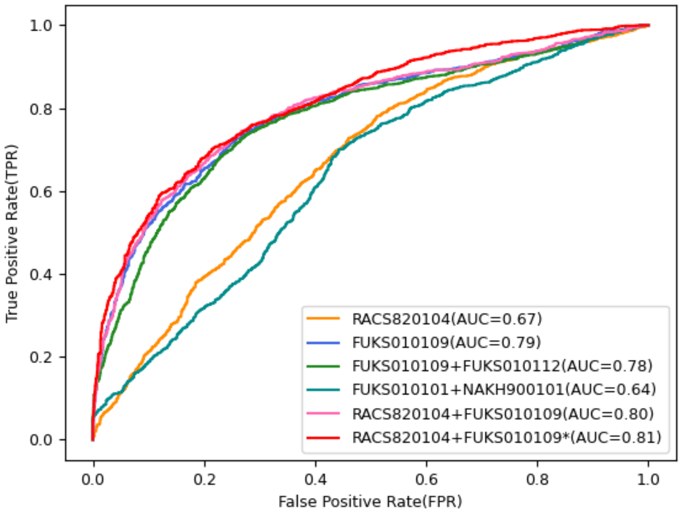

Figure 1: The ROC Curves for Evaluating the Performance of Lysine Prediction Models on the Test Set. '*' Indicates the Normalization Operation.

3.2 Arginine Methylation

For arginine methylation, using 10-fold cross-validation with *Re* as the evaluation metric, the top 10 high-performance AAindex properties are listed in Table 5. Notably, the model constructed based on the "BAEK050101" property showed the highest performance. Previous research [9] suggests that the strength of hydrophobicity increases with more linkers connecting two domains. Therefore, this property is related to hydrophobicity.

Under the *Acc* evaluation metric, the model using the "HUTJ700103" property, related to entropy of formation, achieved the highest performance. The leading models based on the *F1 score* metric largely corresponded with those evaluated using the *Acc* metric. Post Incremental Feature Selection (IFS), a marginal decline in performance prompted the selection of only the "FUKS010101" property for further analyses.

In terms of *Pre*, structural information was identified as a pivotal feature in the high-performing models for arginine methylation prediction. Specifically, the single property "CHAM810101" exhibited a *Pre* surpassing 85%, highlighting the significance of structural insights in predictive models.

Table 5: Results of arginine methylation prediction using top 10 AAindex properties and 10-fold cross-validation

AAindex	***Re.mean***	**AAindex**	***Acc.mean***	**AAindex**	***Pre.mean***	**AAindex**	***F1.mean***
BAEK050101	0.729	HUTJ700103	0.773	CHAM810101	0.862	FUKS010101	0.758
VINM940103	0.721	FUKS010101	0.771	KIMC930101	0.860	HUTJ700103	0.754
FUKS010101	0.718	KIMC930101	0.769	LEVM760104	0.856	QIAN880105	0.746
FUKS010102	0.710	QIAN880105	0.769	TANS770109	0.854	FUKS010102	0.743
GARJ730101	0.710	CHAM810101	0.766	KARS160108	0.854	KIMC930101	0.735
CASG920101	0.710	ISOY800106	0.765	LEVM760103	0.850	NAKH920101	0.732
NADH010104	0.708	QIAN880106	0.759	ISOY800106	0.847	ISOY800106	0.732
KRIW790101	0.708	FAUJ880102	0.759	MAXF760105	0.846	QIAN880106	0.731
GRAR740101	0.707	MUNV940104	0.757	NAKH900113	0.846	FUKS010104	0.730
NAKH920101	0.704	LEVM760104	0.757	MAXF760104	0.843	CASG920101	0.730

After comparing the properties "BAEK050101", "HUTJ700103", "CHAM810101" and "FUKS010101", the optimal feature for arginine methylation prediction was identified. Subsequently, in the model construction phase, step 3 was executed to build various models based on these properties.

The model that utilized only the "BAEK050101" property showed promise for improvement. While the "HUTJ700103" model achieved a *Re* of 92.64% but a lower *Pre* of 58.14%, normalization significantly improved its performance. In contrast, the *Re* for the "FUKS010101" model was 39.37%. The combination and normalization of the properties of "HUTJ700103" and "FUKS010101" resulted in a model that outperformed the others. Notably, the model based on the "CHAM810101" property demonstrated superior performance compared to single-property models and was comparable to the normalized combination of "HUTJ700103 + FUKS010101". Additional models were constructed based on combinations such as "BAEK050101 + CHAM810101" and "BAEK050101 + CHAM810101 + HUTJ700103 + FUKS010101". The optimal parameters and performance of these models on the test set are presented in Table 6.

Table 6: Optimal parameters and performance of models trained with different feature sets on the arginine methylation test set.

Feture Set	Optimal Parameters			Performance			
AAindex ID	Kernel	*C*	*γ*	*Re*(%)	*Pre*(%)	*Acc*(%)	*MCC*
BAEK050101	RBF	5	0.15	64.35	65.25	65.04	0.30
HUTJ700103	RBF	0.1	0.05	92.64	58.14	62.97	0.32
HUTJ700103*	RBF	1	0.75	67.24	73.57	71.54	0.43
CHAM810101	RBF	1	0.95	64.50	85.96	76.98	0.56
FUKS010101	RBF	10	0.05	39.37	87.86	66.96	0.41
HUTJ700103+FUKS010101*	RBF	1000	0.15	74.40	83.35	79.77	0.60
BAEK050101+CHAM810101*	RBF	5	0.45	74.60	81.08	78.60	0.57
BAEK050101+CHAM810101+ HUTJ700103+FUKS010101*	RBF	1000	0.15	72.13	81.43	77.84	0.56

Notes: The highest values are indicated in bold. '*' denotes the normalization operation.

Fig. 2 illustrates the *ROC* plot and corresponding *AUC* values, indicating comparable performances among models based on the four property sets: "CHAM810101", "HUTJ700103 + FUKS010101*", "BAEK050101 + CHAM810101*", and "BAEK050101 + CHAM810101 + HUTJ700103 + FUKS010101*". Considering both feature dimensionality and model parameters, "BAEK050101 + CHAM810101*" emerges as the optimal feature subset for arginine methylation prediction.

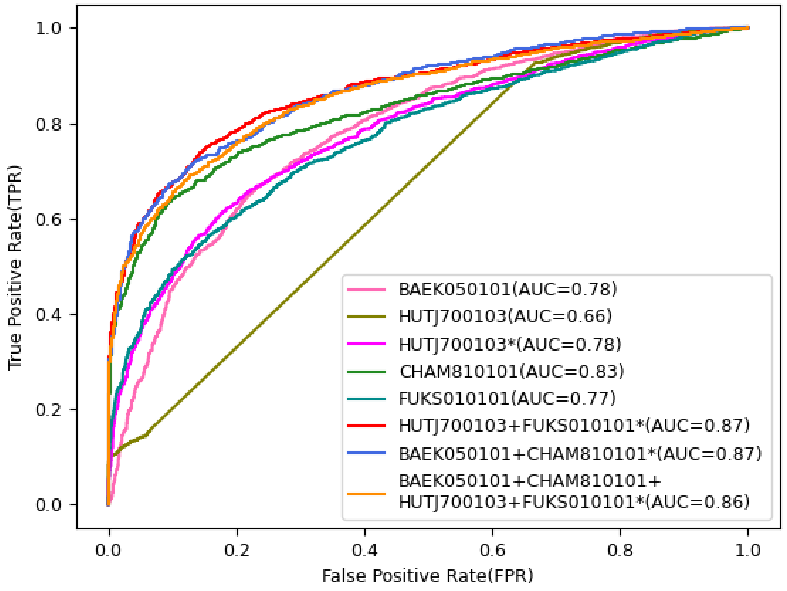

Figure 2: The ROC Curves for Evaluating the Performance of Arginine Prediction Models on the Test Set. '*' Indicates the Normalization Operation.

3.3 Key physicochemical properties for Distinguishing Methylation from Malonylation

The sequence fragment centered on the methylated lysine residue is used as the positive set, while the malonylated lysine residue serves as the negative set. Using 10-fold cross-validation with *Re*, *Acc*, *Pre* and *F1 score* as evaluation metrics, the performance rankings of multiple models based on 553 properties are presented in Table 7. The observed lower overall performance may be attributed to potential cross-talk between different PTMs.

Table 7: Results of lysine methylation and malonylation prediction using the top 10 AAindex properties and 10-fold cross-validation

AAindex	*Re.mean*	AAindex	*Acc.mean*	AAindex	*Pre.mean*	AAindex	*F1.mean*
RACS820104	0.586	ZIMJ680104	0.607	ZIMJ680104	0.623	RACS820104	0.558
AURR980102	0.553	KLEP840101	0.599	KLEP840101	0.610	EISD860102	0.555
EISD860102	0.540	FAUJ880111	0.593	FAUJ880111	0.601	ZIMJ680104	0.547
GEOR030104	0.537	WILM950101	0.583	WILM950101	0.594	WILM950103	0.544
YUTK870103	0.533	GUOD860101	0.581	GUOD860101	0.591	FAUJ880111	0.543
RACS820107	0.532	COWR900101	0.581	COWR900101	0.588	JANJ780101	0.542
FUKS010107	0.532	WILM950103	0.581	MEEJ810102	0.586	KLEP840101	0.540
YUTK870104	0.531	EISD860102	0.579	FINA910103	0.577	FUKS010107	0.540
RICJ880109	0.530	MEEJ810102	0.576	CIDH920102	0.574	JANJ790102	0.538
KARS160106	0.526	CIDH920102	0.576	YUTK870101	0.574	JANJ780103	0.538

It is noteworthy that the top 2 properties ("RACS820104" and "AURR980102") evaluated in the

Re metric align with previous studies, indicating their sensitivity to methylation. Similarly, the top 6 properties evaluated for *Acc* and *Pre* share the same features, including descriptions such as "isoelectric point", "net charge", "positive charge". In the *F1 score* evaluation, the property "WILM950103" emerges as a high-performing feature, recognized for its role in representing residue hydrophobicity. This characteristic has been previously highlighted in distinguishing between methylation and non-methylation sites, underscoring the importance of hydrophobicity in predictive modeling.

4. Conclusion

In this study, we utilized 553 properties from the AAindex database as features to construct predictive models for methylation site prediction. Through rigorous performance comparisons, we identified the physicochemical properties most capable of distinguishing methylation, along with their exact AAindex IDs. The property set "RACS820104+FUKS010109" demonstrated exceptional predictive performance for lysine methylation, while "HUTJ700103+FUKS010101" exhibited relatively strong performance for arginine methylation in our study. In our investigations of lysine methylation and malonylation, we found that structure-based features and physical properties, such as charge equality and hydrophobicity, played significant roles in distinguishing between the two types of modifications. We also constructed models based on various window sizes ranging from 7 to 19, and present the performance rankings of multiple models based on 553 attributes of different window sizes in the Appendix.

Despite these findings, our method has limitations, primarily due to the relatively modest performance of predictive models constructed solely with physicochemical properties from the AAindex database. Future research will focus on integrating these optimal features with additional characteristics to enhance the model's predictive capability. In conclusion, by selecting optimal feature subsets from 553 properties through various combinations, our study contributes valuable insights to the field of methylation site prediction, paving the way for more accurate and robust predictive models in future research.

Acknowledgements

This work was supported by Funding Statement: This work was supported by grants from the Tianjin Applied Basic Research Foundation (21JCYBJC00630, 22JCQNJC01020, 23JCQNJC01610).

Availability of Data and Materials: The required dataset is sourced from a public database. The dataset can be downloaded directly from https://awi.cuhk.edu.cn/dbPTM/download.php.

Appendix: Ranking of 553 attributes under different windows.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] Murn, J., Shi, Y. (2017) The winding path of protein methylation research: milestones and new frontiers. Nat Rev Mol Cell Biol 18, 517-527.

[2] Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, Kanehisa M. (2008) AAindex: amino acid index database, progress report 2008. Nucleic Acids Res. 36(Database issue):D202-205.

[3] Li ZC, Zhou X, Dai Z, Zou XY. (2011) Identification of protein methylation sites by coupling improved ant colony optimization algorithm and support vector machine. Anal Chim Acta. 703(2):163-171.

[4] Wen PP, Shi SP, Xu HD, Wang LN, Qiu JD. (2016) Accurate in silico prediction of species-specific methylation sites based on information gain feature optimization. Bioinformatics. 32(20):3107-3115.

[5] Zhongyan Li, Shangfu Li, Mengqi Luo, Jhih-Hua Jhong, Wenshuo Li, Lantian Yao, Yuxuan Pang, Zhuo Wang, Rulan Wang, Renfei Ma, Jinhan Yu, Hsien-Da Huang and Tzong-Yi Lee. (2022) dbPTM in 2022: an updated database for exploring regulatory networks and functional associations of protein post-translational modifications. Nucleic Acids Research, Volume 50, Issue D1, Pages D471-D479.

[6] Huang, T., Cui, W., Hu, L., Feng, K., Li, Y. X., & Cai, Y. D. (2009) Prediction of pharmacological and xenobiotic responses to drugs based on time course gene expression profiles. PloS one, 4(12), e8126.

[7] Cortes, C., Vapnik, V. (1995) Support-Vector Networks. Machine Learning 20, 273-297.

[8] Pedregosa et al. (2011) Scikit-learn: Machine Learning in Python. JMLR 12, pp. 2825-2830.

[9] Chatterjee, P., Basu, S., Zubek, J., Kundu, M., Nasipuri, M., & Plewczyński, D. (2015) PDP-RF: Protein domain boundary prediction using random forest classifier. In (Ed.), Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (pp. 441–450).