RF-GlutarySite: Random Forest based predictor for Glutarylation sites

Journal:	*Molecular Omics*
Manuscript ID	MO-RES-02-2019-000028.R1
Article Type:	Research Article
Date Submitted by the Author:	11-Apr-2019

Complete List of Authors:	Albarakati, Hussam; North Carolina Agricultural and Technical State University, CSE Saigo, Hiroto; Kyushu University, Informatics Newman, Robert; North Carolina Agricultural and Technical State University, Biology Dukka, B.; North Carolina Agricultural and Technical State University, Computational Science and Engineering;

Molecular Omics

ARTICLE

RF-GlutarySite: Random Forest based predictor for Glutarylation sites

Received 00th January 20xx,
Accepted 00th January 20xx

DOI: 10.1039/x0xx00000x

www.rsc.org/

Hussam J. AL-barakati[a], Hiroto Saigo[b], Robert H. Newman[c] and Dukka B. KC[a,*]

Glutarylation, which is a newly identified posttranslational modification that occurs on lysine residues, has recently emerged as an important regulator of several metabolic and mitochondrial processes. However, the specific sites of modification on individual proteins, as well as the extent of glutarylation throughout the proteome, remain largely uncharacterized. Though informative, proteomic approaches based on mass spectrometry can be expensive, technically challenging and time-consuming. Therefore, the ability to predict glutarylation sites from protein primary sequences can complement proteomics analyses and help researchers study the characteristics and functional consequences of glutarylation. To this end, we used Random Forest (RF) machine learning strategies to identify the physiochemical and sequence-based features that correlated most substantially with glutarylation. We then used these features to develop a novel method to predict glutarylation sites from primary amino acid sequences using RF. Based on 10-fold cross-validation, the resulting algorithm, termed 'RF-GlutarySite', achieved efficiency scores of 75%, 81%, 68% and 0.50 with respect to accuracy (ACC), sensitivity (SN), specificity (SP) and Matthew's correlation coefficient (MCC), respectively. Likewise, using an independent test set, RF-GlutarySite exhibited ACC, SN, SP and MCC scores of 72%, 73%, 70% and 0.43, respectively. Results using both 10-fold cross validation and an independent test set were on par with or better than those achieved by existing glutarylation site predictors. Notably, RF-GlutarySite achieved the highest SN score among available glutarylation site prediction tools. Consequently, our method has the potential to uncover new glutarylation sites and to facilitate the discovery of relationships between glutarylation and well-known lysine modifications, such as acetylation, methylation and SUMOylation, as well as a number of recently identified lysine modifications, such as malonylation and succinylation.

1. Introduction

Post-translational modifications (PTMs) play a critical role in regulating nearly all biological processes.[1,2] For instance, inside the cell, both enzyme-mediated PTMs, such as protein phosphorylation, acetylation and SUMOlyation, and non-enzymatic PTMs, such as oxidation and succinylation, can alter the stability,[3] subcellular localization,[4] interaction profiles and/or enzymatic activity of cellular proteins.[5,6] As a consequence, dynamic changes in PTM profiles on select amino acid residues regulate information flow within many cellular signaling networks.[7,8] Among the twenty canonical amino acids, lysine is subject to the most diverse range of PTMs.[9] Indeed, so-called protein lysine modifications, including acetylation, methylation, ubiquitylation, SUMOylation, and various other types of acyl modifications, coordinate a wide range of biological functions.[10-20] Recently, Tan and colleagues identified a novel lysine modification, termed glutarylation, that is found in both eukaryotes and prokaryotes.[21] Similar to other recently identified lysine acyl modifications, such as succinylation and malonylation, glutarylation changes both the size and the charge state of the modified lysine residue (i.e., from +1 in the unmodified state to -1 following acylation). However, due to the length of the glutarate moiety (which consists of a 5-carbon chain with carboxyl groups on either end), glutarylation results in the largest change in molecular mass among these modifications (**Fig. 1).** Importantly, dysregulation of glutarylation and related lysine acylations has recently been implicated in the etiology of a number of pervasive metabolic disorders,[20] including diabetes, neurodegenerative diseases, cancer and type 1 glutaric aciduria. [21,22]

Several studies have recently used proteomics approaches to identify glutarylation sites in cellular proteins. For instance, Tan et al. originally detected 23 glutarylation sites in 13 unique proteins isolated from *E. coli* and 10 sites in 10 glutarylated proteins from HeLa cells.[21] More recently, the same group identified 683 lysine glutarylation sites in 191 distinct proteins in mouse liver carrying a deletion of the gene encoding the surtuin 5 (SIRT5) deacetylase, which has recently been shown to remove glutarate moieties from proteins.[21] These studies revealed that lysine glutarylation is enriched on proteins associated with mitochondrial functions.

[a.] *Department of Computational Science and Engineering, North Carolina Agricultural & Technical State University, Greensboro NC 27411.*
[b.] *Department of Informatics, Kyushu University, Fukuoka 819-0395, Japan.*
Department of Biology, North Carolina Agricultural & Technical State University, Greensboro NC 27411.
**: Corresponding Author*

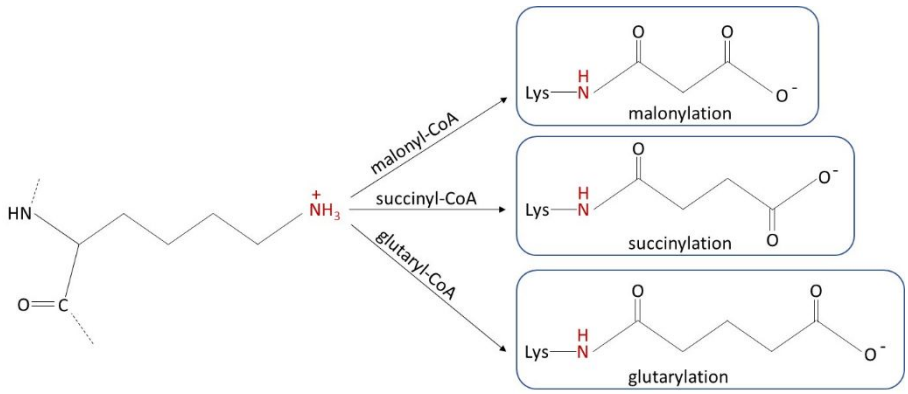

Fig 1. Structures of succinylation, malonylation and glutarylation. The site of conjugation is highlighted in red font.

In another study, Xie et al. used global proteomics approaches to detect 41 glutarylation sites in 24 cellular proteins from the pathogenic bacteria, *Mycobacterium tuberculosis*.[23] Finally, the Braulke group identified 73 glutarylation sites in 37 mitochondrial proteins involved in various metabolic processes in the brain and liver.[24]

Though unparalleled in their ability to experimentally identify sites of modification across entire proteomes, proteomics approaches are often time-consuming, technically demanding and expensive. Therefore, to complement and extend proteomics studies, researchers have developed computational methods to predict PTMs in silico.[25-29] For instance, computational methods have been developed to predict several PTMs, including protein lysine modifications such as acetylation,[25] succinylation,[30] malonylation[26] and propionylation.[31] These methods, which are typically trained to distinguish which residues are most likely to be modified, can help researchers better understand relationships between various PTMs and can offer insights into crosstalk between signaling pathways. Recently, Ju and Jian established the first computational method to predict sites of glutarylation.[32] Their method, which is termed 'GlutPred', uses Support Vector Machine (SVM) machine-learning strategies together with three features—amino acid factor, binary encoding, and k-space encoding—to predict glutarylation sites in proteins based on the primary amino acid sequence. Though GlutPred performed well with respect to specificity (SP) and accuracy (ACC), it struggled with regard to sensitivity (SN), where it achieved a SN score of 51.8% using an independent test set. Similarly, iGlu-Lys, which is a recently described SVM-based glutarylation site prediction tool developed by Xu et al., achieved high ACC and SP scores but a relatively low SN score of only 51.4% using an independent set.[33] Since SN describes a method's ability to correctly predict positive sites of modification, a method with improved SN scores would complement existing methods and extend our ability to predict novel sites of glutarylation throughout the proteome. Moreover, because both GlutPred and iGlu-Lys were trained using SVM-based machine learning strategies, they offer little information about the relative contribution of each feature to overall method performance. This information, which can offer insights into the biochemical and biophysical parameters that help determine whether a given lysine residue is likely to be glutarylated in situ, could be valuable for the development of future glutarylation site predictors. Here, we sought to develop a glutarylation site predictor with enhanced SN compared to existing methods. The resulting method, which we termed RF-GlutarySite, uses Random Forest (RF) and a series of complementary feature vectors to distinguish glutarylation sites from lysine residues that are not likely to be glutarylated inside the cell. Analysis of our method, which achieved SN scores of 77% using both 10-fold cross-validation and an independent test set, suggests that features describing amino acid composition, transition and distribution (CTD) and pseudoamino acid composition (PAAC), as well as their local spatial distribution (e.g., Geary autocorrelation), contribute most substantially to overall method performance. Together, these studies promise to offer important insights into the sites of glutarylation across various proteomes while providing information about the biochemical/biophysical features underlying glutarylation site selection. This information may also facilitate the investigation of other lysine modifications that share similar characteristics, such as succinylation and malonylation.

2. Material and methods

2.1. Datasets

To build our training and independent test sets, we first obtained sequences for 211 proteins containing at total of 716 glutarylation sites from the Protein Lysine Modification Database (PLMD).[34] The glutarylated proteins in the PLMD, which were identified in *Mus musculus* and *M. tuberculosis*, were obtained from two previous studies.[21,23] In addition, sequences for 13 proteins from *E. coli* (containing 23 unique glutarylation sites) and 10 protein sequences from human HeLa cells (containing 10 unique glutarylation sites) were retrieved from the National Center for Biotechnology (NCBI) database and the SWISS-PROT database, respectively.[21]

Table 1. Summary of datasets or resources used to retrieve glutarylation sites across different species. The number of sequences and sites from each species are shown. PLMD: Protein Lysine Modification Database.

References	Protein sequences	Species	Sites
PLMD	187	M. musculus	674
PLMD	24	M. tuberculosis	42
Tan M et al., 2014 [21]	13	E. coli	23
Tan M et al., 2014 [21]	10	HeLa cells	10
Combined	**234**	**All species**	**749**

Table 2. Number of positive and negative sites in the training and test sets before (left) and after (right) balancing.

Dataset	Positive sites (before/after)	Negative sites (before/after)
Training	400/400	1703/400
Test	44/44	203/44

The combined initial dataset, which was composed of 234 proteins containing a total of 749 glutarylation sites, is summarized in **Table 1**.

Next, we applied CD-hit[35] to remove homologous sequences that exhibited ≥ 40% sequence identity. This left us with a total of 204 non-redundant proteins sequences. We then used a sliding window to generate peptides based on experimentally-identified sites of glutarylation. To this end, we first generated positive windows based on glutarylation sites collected from different resources. The length of each sequence was 21, with 10 residues upstream and downstream from a central lysine residue. Negative sequences were generated in a similar manner, except the central lysine residue was not known to be glutarylated. Together, this led to a total of 626 positive sites and 4,201 negative sites. To avoid overfitting, we removed homologous peptides that exhibited 100% identity within each set. Likewise, because the glutarylation status of a protein may depend on the cellular context from which it was obtained, we crosschecked the sequences in the positive set with those in the negative set. If a peptide in the negative set exhibited 100% identity with a sequence in the positive set, it was removed from the negative set and kept only in the positive set. Next, we applied CD-hit individually on peptides in the positive set to remove homologous fragments with more than 40% identity. The same procedure was used to remove redundant sequences from the negative set. In this way, CD-hit was used as a tool to avoid overestimation of our predictor caused by a large number of homologous peptides in the positive and negative sets. The final, non-redundant datasets contained 453 positive sites and 2,043 negative sites. Finally, we randomly split the data into the training and test sets. Specifically, 90% of the sequences from the positive and negative sets were used for training while the remaining 10% were used for the independent test set (**Table 2**).

2.2. Feature extraction and encoding

Most of the features used in this study were retrieved from the Features Extraction from Protein Sequence (FEPS) web application,[36] which, in its original iteration, contained 2,755 features. The FEPS application utilizes 48 published feature extraction approaches and has been effectively employed in a variety of computational problems, including the prediction/classification of nuclear receptors,[37] the prediction of phosphorylation sites,[27] and the prediction of hydroxylation sites.[28] In this study, we used an updated version of the FEPS web application to extract 12,249 features for method development. To these features, we added binary encoding (BE) features, which had vector length of 400, and two types of physiochemical properties, AAindex features and Afactor, that contain 80 and 100 length vectors, respectively. In total, we used 12,829 features during initial method development (**Table 3**). Each feature class is described below.

2.2.1- Pseudo-amino acid composition.

Pseudo amino acid composition (PseAAC) combines discrete attributes with consecutive attributes.[29,38] The initial twenty attributes denote amino acid composition while the other attributes represent sequence order data based on physicochemical properties, including hydrophobicity (H1), hydrophilicity (H2) and side-chain mass (M) of amino acids. The formula used to normalize H1, H2 and M is as follows:

$$\dot{E}(i)= \frac{E(i) - \sum_{i=1}^{20} \frac{E(i)}{20}}{\sqrt{\sum_{i=1}^{20} [E(i) - \sum_{i=1}^{20} \frac{E(i)}{20}}} \qquad (1)$$

where E(i) represents the property values of H1, H2 and M whereas È(i) represents the property value of three properties after standardization. Meanwhile, the sequence order correlated factor can be defined as:

$$J_\lambda = \frac{1}{N-\lambda} \sum_{i=1}^{L-\lambda} \theta(R_i, R_{i+\lambda}) \qquad (2)$$

where J_λ represents the first-tier correlation, which specifies the sequence order between all of the λ maximum nearest residues in the protein sequence, with λ_{max} = S. For instance, if N denotes the length of the sequence, then it must be > λ. Meanwhile, $\theta(R_i, R_{i+\lambda})$ is the correlation factor, which can be calculated as:

$$\theta(R_i, R_{i+\lambda}) = \frac{1}{3}\{[H_1(R_i) - H_1(R_j)]^2 + [H_2(R_i) - H_2(R_j)]^2 + \{[M(R_i) - M(R_j)]^2\} \qquad (3)$$

where $H_1(R_i)$, $H_2(R_i)$ *and* $M(R_i)$ represent the original values of hydrophobicity, hydrophilicity, and side-chain mass before

normalization while $H_1(R_j), H_2(R_j)\ and\ M(R_j)$ correspond to the values for these parameters after normalization.

The initial twenty attributes of PseAAC produce frequencies of amino acid composition that are specified by:

$$X_a = \frac{f_t}{\sum_{t=1}^{20} f_t + w\sum_{d=1}^{30} J_i} \quad (4)$$

where f_t represents the frequency of an amino acid of type t, w denotes the weight factor (with a default value is 0.1) and J represents first-tier correlation factor.

The remaining attributes of PsAAC reproduce sequence order and are given by:

$$X_b = \frac{wJ_{r-20}}{\sum_{t=1}^{20} f_t + w\sum_{d=1}^{30} J_i} \quad r = 21,22,\ldots S \quad (5)$$

where S is maximum number of λ.[39,40] PsAAC has been widely used in various problems in bioinformatics, such as to enhance the quality of subcellular localization predictions[41] and to predict phosphorylation sites.[42] The total number of features used for PsAAC was 47.

Table 3. Features used for method development.

Number	Name of features	Length vectors
1	Pseudo-amino acid composition Type I Pseudo amino acid composition (20 + Defalut lamda =7) => 27 Type II Pseudo amino acid composition (20 + Defalut lamda =0) => 20	47
2	Conjoint triad	512
3	Entropy, relative entropy and gain	3
4	Composition => 21 Transition => 21 Distribution => 105	147
5	Amino acid composition	20
6	Dipeptide composition	400
7	Tripeptide composition	8000
8	Geary autocorrelation (Default lambda = 30)	240 (8* lambda)
9	Moran autocorrelation (Default lambda = 30)	240 (8*lambda)
10	Normalized Moreau–Broto autocorrelation (Default lambda = 30)	240 (8*lambda)
11	k-Spaced Amino Acid Pairs	2400 (20*20*6)
12	Binary encoding	400 (20*L, where L= window-1)
13	AAindex features => 4	80
14	Afactor feature => 5	100
	All features	**12,829**

2.2.2- Conjoint Triad.

Conjoint triad (CT) is a feature originally established by Shen et al. to predict protein-protein interactions.[43,44] It has also been successfully applied to predict enzyme function[45] and subfamilies of nuclear receptors.[46] To calculate the CT, the twenty canonical amino acids were first subdivided into seven groups based on their dipoles and side-chain volumes (**Table 4**, **rows 1-7**).

Next, any three continuous neighboring amino acid residues were considered as one component. Finally, components were divided into their corresponding groups and the frequency of each group or class was counted. For instance, if we have two components related to the same group, such as "VKS" and "ART", these can defined as identical due to the similarities in their physiochemical properties. Recently, Yin and Tan developed an enhanced version of CT that contains a new group for "dummy" residues, which they termed "O" (**Table 4, row 8**).[47] This approach helped capture each residue in the peptide without sacrificing information near the termini. In our study, we used the enhanced version of CT, yielding a 512-dimensional vector.

2.2.3- Shannon entropy.

Shannon entropy (SE) was first time introduced in 1984.[48] It is a metric used to measure the uncertainty of a set of residues in protein sequences.[49,50] The SE can be computed by:

$$\mathrm{SE} = -\sum_{d=1}^{20} p_d \log_2 (p_d) \quad (6)$$

where p_d is the probability of an amino acid of type d in the protein sequence or peptide. It can be calculated by determining the number of amino acids of type d in the sequence and subtracting by the length of the protein sequence or peptide. Those amino acids that are not present in the sequence or peptide are assigned a probability of 0. Finally, for each sequence, we summed the result for each type of amino acid. The length of this feature is 1.

Table 4. Dipole and side-chain volume classifications of the 20 canonical amino acids (plus a "dummy" residue, O) used to determine conjoint triad (CT) features. See text for details.

Number	Groups	Dipole scale	Volume scale
1	[A,G,V]	-	-
2	[I,L,F,P]	-	+
3	[Y,M,T,S]	+	+
4	[H,N,Q,W]	++	+
5	[R,K]	+++	+
6	[D,E]	+++	+
7	[C]	+	+
8	[O]	+	+

2.2.4- Relative entropy.

Relative entropy (RE), also is known as the Kullback-Leibler distance, measures the frequency of a given amino acid divided by the background distribution.[51] It can be computed according to the following relationship:

$$\mathrm{RE} = -\sum_{d=1}^{20} p_d \log_2 \left(\frac{p_d}{p_a}\right) \quad (7)$$

where p_d denotes the frequency distribution of each of the twenty amino acids in the peptide and p_a represents an equal number of frequency for each amino acid in peptide (n_c), which can be defined by:

$$p_a = \left(\frac{1}{n_c}\right) \quad (8)$$

RE must be a non-negative value that converts to a zero value when $p_d = p_a$ .The length of this feature is 1. RE has been applied in previous studies to determine the conserved position.[52-54]

2.2.5- Information gain.

Information gain (IG) quantifies the change of data in a peptide sequence impacted by a gathering factor. It is simply the difference between SE and RE and is given by:

$$\mathrm{IG} = \mathrm{SE} - \mathrm{RE} \quad (9)$$

The length of the IG vector is 1.[27]

2.2.6- Composition, Transition and Distribution.

Composition, transition and distribution (CTD) featurs have been widely applied to many computational problems.[55-57] The initial phase is to extract information about the composition, transition and distribution of amino acids in the sequence. To this end, each of the twenty amino acids are classified into one of three classes for seven physiochemical properties, as shown in **Table 5**.[58]

2.2.6.1. Composition.

The composition (CP) refers to the number of amino acids in a peptide that can be encoded into each of the three classes, divided by the length of the peptide.[58,59] It can be computed according to the equation:

$$CP_t = \frac{O_c}{M} \quad (10)$$

where O_c denotes the number of occurrences of class of type c (where c represents one of the three classes), M represents the length of the peptide and CP_t represents composition of type t (where t is describes one of the seven types of physicochemical properties in **Table 5**). For example, if we want to determine the composition with respect to polarity for the fragment 'MTEMHMPDF', we first assign each residue to the required class, yielding '123131231'. Then, we count the number of each class. Finally, we divide the occurrence of each class by the length of the peptide (i.e., 9 residues). Therefore, for this sequence, the final result would be: class 1 = 0.44 (or 4/9), class 2 = 0.22 (or 2/9), and class 3 = 0.33 (or 3/9). The total feature length is 21.

2.2.6.2. Transition.

Transition (Tr) is defined as the frequency with which a given class (i) is followed by another class (j). Since there are three classes in each property, the possible transitions for a particular property are (1, 2), (1, 3), and (2, 3). The Tr is given by:

$$Tr_{ij} = \frac{O_{ij} + O_{ji}}{M - 1} \quad (11)$$

where O_{ij} denotes number of occurrences of a transition from class i to class j and O_{ji} represents the number of occurrences of a transition from class j to class i. The number of features is 21 (3 features for each of the seven physicochemical properties). For example, suppose we want to find the transition of the fragment 'HKVIRWPS' with respect to polarity. We would first encode each residue to its respective class (i.e., '33113122'). Then, we would count the number of transitions from class 1 to class 2 (class 1 → class 2), from class 1 to class 3 (class 1 → class 3) and from class 2 to class 3 (class 2 → class 3). Finally, we sum the occurrences of class i followed by class j and then divide the result by one less than the length of the peptide. Thus, for our example, the final result would be: class 1 = 0.43, class 2 = 0.14, and class 3 = 0.

Table 5. Three classes with seven physiochemical properties.

N	Types	Class 1	Class 2	Class 3
1	Hydrophobicity	Polar R,K,E,D,Q,N	Neural G,A,S,T,P,H,Y	Hydrophobicity C,L,V,I,M,F,W
2	Normalized van der Waal	0-2.78 G,A,S,T,P,D	2.95-4.0 N,V,E,Q,I,L	4.03-8.08 M,H,K,F,R,Y,W
3	Polarity	4.9-6.2 L,I,F,W,C,M,V,Y	8.0-9.2 P,A,T,G,S	10.4-13.0 H,Q,R,K,N,E,D
4	Polarizability	0-1.08 G,A,S,D,T	0.128-0.186 C,P,N,V,E,Q,I,L	0.219-0.409 K,M,H,F,R,Y,W
5	Charge	Positive K,R	Neutral A,N,C,Q, G,H,I,L, M,F,P,S,T, W,Y, V	Negative D,E
6	Secondary structure	Helix E,A,L,M,Q,K,R,H	Strand V,I,Y,C,W,F,T	Coil G,N,P,S,D
7	Solvent accessibility	Buried A,L,F,C,G,I,V,W	Exposed R,K,Q,E,N,D	Intermediate M,S P,T,H,Y

Table 6. Distribution values of each class for specific piptide

Type	Class type	Value 1	Value 2	Value 3	Value 4	Value 5
Polarity	Class1	0	15.38%	46.15%	73.1%	100%
	Class2	7.69%	23.07%	53.8%	73.1%	92.3%
	Class3	0	0	0	0	0

2.2.6.3. Distribution.

Distribution (Dr) was first proposed by Dubchak et al.[58-60] It determines the distribution of each class along the protein sequence. It contains five values. These values denote the location of each class along the sequence that can be defined as the first amino acid, 25% of the amino acids in the sequence, 50% of the amino acids in the sequence, 75% of amino acids in the sequence and 100% of amino acids in the sequence, respectively. For a single property, the feature length is 15 (5 values for each class); therefore, since there are a total of seven types of physiochemical properties with five values for each class, the total length of features is 105. This feature can be computed as follows:

$$Ds_t = \frac{O_c}{M} * 100 \quad (12)$$

where O_c denotes the location of five values of each class of type c for a peptide of length M. For example, suppose we want to find the distribution with respect to polarity for the fragment 'MPPMPPPPPMMMMPPPMPMPPPPMMM'. First, we would encode each residue to its respective class, yielding '12212222211112221212222111'. Then, we would find the distribution of each coding class along the first, second, third, fourth and fifth quintiles. Finally, we would divide by the coding class stored in these locations and multiply by 100. The final results for our sample peptide are shown in **Table 6.**

2.2.7- Amino acid composition.

Amino acid composition (AAC), which was originally developed to identify the subcellular localization of proteins, was recently shown to be the primary feature during the identification of bacterial toxin proteins.[61-63] AAC, which is the proportion of each of the twenty amino acids along the peptide sequence, can be expressed as:

$$A_x = \frac{L_x}{M} \quad (13)$$

where L_x denotes the number of instances of "amino acid x" divided by the length of the peptide, M.[64,65] The length of this feature is 20.

2.2.8- Dipeptide composition.

Dipeptide composition (E) was first proposed by Reczko and Bohr to predict protein classes[66] and has been applied successfully to identify several protein families and subfamilies, including those of G-proteins.[67] E, which is defined as the frequency of two pairs of amino acids along the peptide,[40] is given by:

$$\mathrm{E_{x,y}} = \frac{\mathrm{L_{xy}}}{\mathrm{M}-1} * 100 \quad (14)$$

where $\mathrm{L_{xy}}$ represents the frequency of each amino acid pair and M is the length of peptide. For example, to determine E for the fragment 'FDPFDRR', we would will first divide the peptide into each of the possible amino acid pairs. In this example, the possible dipeptide combinations are: FD, DP, PF, FD, RR, with a frequency of 2 for the FD pair, 1 for the DP pair, 1 for the PF pair, and 1 for the RR pair. Therefore, according to equation 14, EFD = 2/(7-1)*100 = 33 while EDP = EPF = ERR=16.5. The total length of E is 20 * 20 = 400 features.

2.2.9- Tripeptide composition.

Tripeptide composition (T), which is conceptually similar to E, has been implemented in different web servers to extract biological information from the protein primary sequence[36,39,40] and has been applied successfully in several bioinformatics tools.[68,69] It can be computed as [40]:

$$T_{x,y,z} = \frac{L_{xyz}}{M-1} \quad (15)$$

where L_{xyz} represents the frequency of each tripeptide composed of amino acids x, y, and z along a peptide of length M. The length of this feature is 20*20*20 is 8,000.

2.2.10- Autocorrelation.

Autocorrelation (R_t) is the connection between the values of a solitary variable. Autocorrelation attributes depict connections

between two items, such as protein or peptide sequences, based on their particular construction or physicochemical properties.[70,71] Over the past decade, several web servers have been developed to extract autocorrelation features from protein and peptide sequences.[36,39,40,72] Most of these methods attempt to discover a connection between two residues in a given peptide sequence in a similar manner. Specifically, the correlation between two residues is determined by values derived from eight physicochemical properties.[59] The primary physicochemical property for each residue is the hydrophobicity[73] followed by average flexibility,[74] polarizability,[75] free energy in water,[75] accessible surface area (ASA),[76] amino acid volume,[77] steric parameters,[78] and relative mutability.[79] Before incorporating any of the physicochemical properties into the autocorrelation formula, they must be normalized using the following relationship[59]:

$$\acute{R}_t = \frac{R_x - \acute{R}}{S} \quad (16)$$

where R_x represents the physicochemical property values of amino acid, x. Meanwhile, $\acute{R}$ is the mean of the eight physicochemical properties and S represents the standard deviation, which can be defined according to equations 17 and 18, respectively:

$$\acute{R} = \frac{\sum_{x=1}^{20} R_x}{20} \quad (17)$$

and

$$S = \sqrt{\frac{1}{20}\sum_{x=1}^{20}(R_x - \acute{R})^2} \quad (18)$$

There are three types of autocorrelation. The first type, known as the Moreau-Broto autocorrelation (A_k),[71,80] can be calculated by:

$$A_k = \sum_{x=1}^{M-k} R_x * R_{x+k} \quad (19)$$

where R_x is the amino acid property at position x, R_{x+k} is the amino acid property at position x + k and M is the peptide length. Here, k represents the autocorrelation along the protein sequence, which we initialized with 30 as the default value. Finally, by rearranging Eq 19, A_k can be normalized based on peptide length to yield the normalized Moreau-Broto autocorrelation[39]:

$$A_k = \frac{\sum_{x=1}^{M-k} R_x * R_{x+k}}{M-k} \quad (20)$$

where R_x , R_{x+k} , M and k are as above. The total length of this feature class is 240.

The second type of autocorrelation, named the Moran autocorrelation (Bk),[81] can be computed as:

$$B_k = \frac{\frac{1}{M-k}\sum_{x=1}^{M-k}(R_x - \acute{R})(R_{x+k} - \acute{R})}{M-k} \quad (21)$$

where R_x , R_{x+k} , M and k are as defined in Eq. 19 and $\acute{R}$ is the mean of R_x across the sequence, which is defined as:

$$\acute{R} = \frac{\sum_{x=1}^{M} R_x}{M} \quad (22)$$

The main difference between the Moreau-Broto autocorrelation and the Moran autocorrelation methods is that, unlike the Moreau-Broto autocorrelation, the Moran autocorrelation uses the average value of a given physiochemical property instead of the actual value of the property. The total length of the Moran autocorrelation is 240.

The last type of autocorrelation, known as the Geary autocorrelation,[82] can be calculated according to:

$$C_k = \frac{\frac{1}{2(M-k)}\sum_{x=1}^{M-k}(R_x - R_{x+k})^2}{\frac{1}{M-1}\sum_{x=1}^{M}(R_x - \acute{R})^2} \quad (23)$$

where R_x , R_{x+k} , M and k are as defined in Eq. 19 and $\acute{R}$ is the mean of Rx, as described by Eq. 22.

The primary difference between Geary autocorrelation and the other two types of autocorrelation is that the Geary autocorrelation uses the square-difference of property values.[39] The total length of Geary autocorrelation is 240.

2.2.11- Binary encoding.

Binary encoding (BE) is used to transform each residue in a peptide into 20 coding values. For instance, Ala is represented as (10000000000000000000) while Cys is represented as (01000000000000000000), etc. This feature has been widely applied in different contexts, including the prediction of 1) conformational epitopes in B-cells;[83] 2) the subcellular localization of proteins;[84] and 3) sites of post-translation modification, such as SUMOylation and acetylation.[64,85] In our study, we applied a BE scheme similar to that used to predict acetylation sites.[90] Importantly, once each residue in a given peptide had been transformed into the coding values, we removed the central lysine from all windows. The total length of this feature is 20 * 20 = 400 vectors.

2.2.12- Amino acid index.

Amino acid index (AAindex) is a database of values corresponding to different types of physicochemical properties for the twenty amino acids.[86] In our study, we used four physicochemical properties previously found to be beneficial during the prediction of succinylation sites. These values, which were retrieved from Hasan et al,[65] were: normalized frequency of alpha-helix (PALJ810101), weights for coil at a window position of -4 (QIAN880129), slope in regression analysis x 1.0E1 (PRAM820102), and normalized frequency of turn in all-alpha class (PALJ810113). The length of the feature was 80.

2.2.13- Amino acid factors.

As alluded to above, the AAIndex database contains several types of physicochemical properties for amino acids with their

corresponding values.[86] It has been successfully implemented in many computational biology studies.[87-89] The amino acid factors (AAfactor) are polarity, secondary structure, molecular volume, codon diversity, and electrostatic charge.[90] We used these factors in our study because, when used in conjunction with other features, they appeared to play an important role during the prediction of both propionylation[31] and glutarylation sites.[32]

2.2.15- K-spaced amino acid pairs.

The k-spaced amino acid pairs (KSAAP) feature, which has been successfully used to predict O-glycosylation sites,[91] succinylation sites,[92] and phosphorylation sites,[93] describes the number occurrences of all possible adjacent residues in a protein sequence. The KSAAP feature can be expanded by separating two adjacent residues by a distance of k, which can be any number of residues up to two less than the length of the peptide.[94,95] For instance, **Table 7** illustrates the results of using k-spaced features with various values of k for the peptide 'AAAD'. In our study, we chose to use k = 6; therefore, the total length of the feature is 20 * 20 * 6 = 2,400 attributes.

2.3. Balancing the dataset.

Class unbalances occur when the sample data in one dataset is greater than that in the other set.[96] It is considered one of the greatest problems in the machine learning field.[97] Earlier studies have described many techniques to handle unbalanced datasets.[98-100] For instance, under-sampling strategies reduce the size of the largest class so that it is equal to that of the smallest class. The disadvantage of this strategy is that it has the potential to remove important data that could affect the model.[99] The second approach is over-sampling, which replicates data in a small class so that it has the same size as the largest class. The disadvantage of this strategy is that it can lead to over-fitting due to the large amount of replicated data (usually in the positive dataset). In this study, we implemented an under-sampling strategy to reduce computational time and to avoid over-fitting our model.

Table 7. Result of k-spaced feature for peptide AAAD with different lengths, along with corresponding length depending on the specified k.

k	k-space amino acid pairs	k-space encoding features	Length vectors
0	(AA,AC,AD,……YY)	(2,0,1,……,0)	400
1	(AXA,AXC,AXD,……YXY)	(1,0,0,……,0)	800
2	(AXXA,AXXC,AXXD,……YXXY)	(0,0,1,……,0)	1200

2.4. Feature selection.

Though, on the surface, it may seem counterintuitive, in many cases, using the entire feature set can actually be detrimental to model performance. For instance, if a large number of features contain irrelevant information, correlated knowledge between irrelevant features can adversely affect model performance. Likewise, larger dimension data can lead to more difficult tasks for many machine learning classifiers that are used to solve bioinformatics problems. Therefore, it is important to minimize the dimensionality of data (particularly irrelevant data) when implementing machine learning algorithms and visualization techniques.[101,102] To accomplish this, feature selection techniques are often employed to identify those features that contribute most substantially to model performance.[103] Not only can this information provide insights into the biochemical/biophysical parameters underlying the system-under-study, but it can also enhance the efficiency of machine learning classifiers by reducing the computational time and memory necessary for model deployment. Moreover, during development, feature selection can also improve the performance of models that are based on learning classifiers by preventing over-fitting of the training data.[104]

There are many approaches to select important features from the data. These approaches can be sub-divided into three types. For instance, filter methods use statistical techniques to assign scores to each feature, allowing the features to be ranked according to their respective scores.[105,106] Some examples of this approach include chi-squared analysis and information gain. This strategy has been used in many bioinformatics problems, such as cancer classification.[107,108] The second approach is the wrapper method, which searches for the best subset of features for a specific algorithm.[109] Some examples of this method include recursive feature elimination and genetic algorithms (GA). The third approach is the embedded method, which uses an intrinsic model building learning metric to optimize the performance of the model.[110] One example of embedded is Lasso.[111]

Because it is computationally less expensive than the wrapper and embedded methods, in this study we used a filtering strategy. Specifically, we applied the Gradient Boosted Trees technique, Xgboost, to identify non-linear correlations from the largest amount of data. Xgboost has been used to select the most efficacious features during the prediction of β-Lactamases and their classes.[112] It has also been used with deep learning methods for prediction of protein contacts.[113]

We implemented Xgboost in Python in a manner that structured the gradient boosted trees[114] to select important features from our training dataset and that enhanced the prediction quality of our method. The corresponding feature importance was then computed based on Gini impurity. The Gini impurity is a metric applied to determine the ability of a given attribute to efficiently divide the input information into the right label. Gini impurity can be defined as:

$$I = \sum_{j=1}^{m_a} k_j(1 - k_j) \quad (24)$$

where m_a is the number label and k_j is the fractional value of j. Using values for each node in the gradient boosted trees, we were able to measure the Gini Importance based ono Gini

impurities. The Gini Importance metric can be computed as following:

$$N = I_{parent} - I_{child\,1} - I_{child\,2} \quad (25)$$

Any feature with a relative importance value less than 0.002 was regarded as an irrelevant feature. Based on this cutoff, 128 of the initial 12,829 features were selected for method development while the remaining 12,701 features were discarded from the training set.

2.5. Random forest classifier.

Random forest (RF) is an ensemble supervised method composed of a combination of decision trees using the bagging algorithm.[115-117] It has been used in many computational biology problems, such as the prediction of residues important for DNA binding by transcription factors and other DNA-binding proteins,[118] the prediction of microRNA (miRNA) target sites,[119] and the prediction of various PTMs, including sites of phosphorylation,[27] hydroxylation,[28] succinylation[30] and glycosylation.[120] In this work, RF was used to categorize glutarylation sites and sites that are not glutarylated in peptides. The first stage involved a bootstrapping algorithm to create multiple sets of decision trees from the training set, where each decision tree has a subset of features, termed "v", and a random subset of samples, termed "s". In the second stage, the best node was designated amongst the v features. In the final stage, the majority vote from each decision tree and class label was assigned based on the highest number of votes. Python (v 3.6.0) with the Scikit-learn (v 0.19.0)[121] and pandas (v 0.20.3)[132] packages were implemented to create our method.

2.6. Evaluation of model performance.

To evaluate the performance of our method and the existing glutarylation site prediction tools, we used two assessment strategies. The first strategy was based on 10-fold cross-validation. During 10-fold cross-validation, we first split our training dataset into ten equal partitions. We then used nine partitions for training and the remaining partition for testing. This process was repeated ten times and the results averaged. Likewise, we also used an independent test set to assess model performance.

The performance of each method using either 10-fold cross-validation or the independent test set was measured using several common performance metrics, including accuracy (ACC), sensitivity (SN), specificity (SP), Matthew's correlation coefficient (MCC), F1-score (F1) and precision (PR). These metrics, which have been implemented to measure the quality of different methods in many studies,[27,61,91,103] are defined below:

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \times 100 \quad (26)$$

$$\mathrm{SN} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \times 100 \quad (27)$$

$$\mathrm{SP} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \times 100 \quad (28)$$

$$\mathrm{MCC} = \frac{(\mathrm{TP})(\mathrm{TN}) - (\mathrm{FP})(\mathrm{FN})}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}} \quad (29)$$

$$\mathrm{F1} = 2 * \frac{SN * \mathrm{PR}}{\mathrm{PR} + \mathrm{SN}} \quad (30)$$

$$\mathrm{PR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \times 100 \quad (31)$$

where TP represents the number of true positives (i.e., the number of known glutarylation sites that were classified correctly), TN represents the number of true negatives (i.e., the number of non-glutarylation sites that are classified correctly), FP denotes the number of false positives (i.e., the number of non-glutarylation sites that were incorrectly classified as glutarylation sites) and FN indicates the number of false negatives (i.e., the number of known glutarylation sites that were incorrectly classified as non-glutarylation sites). Because it accounts for TP, TN, FP and FN rates, the MCC value is commonly considered as surrogate for overall method performance.[62-64,122] Likewise, we measured the area under the receiver-operator characteristic (ROC) curve. The ROC, which plots sensitivity versus 1 – specificity with every possible threshold,[123] can be converted into a numerical value by calculating the area under the curve (AUC), where an AUC score of 0.5 denotes a random classifier and an AUC score of 1.0 denotes a perfect classifier.[124] Similarly, the precision-recall curve (PRC) is a plot of precision versus sensitivity.[125,126] It is the most informative and powerful plot for imbalanced datasets and, importantly, is able to explicitly reveal differences in early-retrieval performance.[127]

3. Results and Discussion.

3.1. Model development.

To develop a glutarylation site prediction tool, we first mined the Protein Lysine Modification Database (PLMD) and various literature resources to construct a set of 204 non-redundant protein sequences containing at least one experimentally-validated glutarylation site.[21,23] After removing sequences that shared >40% sequence identity, a total of 453 unique glutarylation sites (i.e., positive sites) and 2,043 lysine residues

not known to be glutarylated (i.e., negative sites) remained. We then randomly selected 10% of the positive sites and 10% of the negative sites to serve as the independent test set and used the remaining sites for training and method development (**Table 2**). The FEPS webserver was used to extract various features related to physiochemical properties (e.g., entropy, pseudoamino acid composition (PseAAC) and composition, transition & distribution (CTD)) and sequence distribution (e.g., k-spaced amino acid pairs (KSAAP), Geary autocorrelation and Moran autocorrelation) (**Fig. 2, Table 3**). In addition to the features extracted from the FEPS server, we also included other features, such as binary coding (BE), amino acid index (AAindex) and amino acid factors (AAfactor).[64,65] [85,90] In total, 12,829 features were used for method development (**Table 3**).

Next, we evaluated the fidelity of four supervised machine algorithms, namely SVM, Naïve Bayes (NB), k-nearest neighbour (KNN) and Random Forest (RF). To this end, we assessed the performance of each classifier with respect to ACC, SP, SN and Matthew's correlation coefficient (MCC) using both 10-fold cross validation and our independent test set (**Tables S1 and S2**). Using both 10-fold cross validation and the independent test set, RF performed the best with respect to almost all of the performance metrics. For instance, based on 10-fold cross-validation, RF exhibited efficiency scores that were an average of 11.9%, 18.5%, 5.7%, 67.1% and 15.4% higher with respect to ACC, SN, SP, MCC, and AUC respectively, than those achieved by the other classifiers (**Table S1 and Fig. S1**). Indeed, aside from SP, where the SVM classifier performed marginally better than RF, RF achieved the highest score by each metric. The disparity was even greater when using the independent test set, where RF achieved ACC, SN, SP and MCC scores that were, on average, 7.9%, 17.9%, 5.6% and 33.3% higher than those exhibited by the other classifiers, respectively (**Table S2 and Fig. S2**).

One disadvantage of using very large feature sets for method development is that the inclusion of extraneous features that do not contribute substantially to method performance can markedly increase computational cost with little to no improvement in overall performance. In fact, in some cases, including irrelevant features can adversely affect model performance.[128] Therefore, to decrease computational cost and potentially increase model performance, we sought to identify an optimal feature set for model development. Since it is a decision tree matrix, RF allows the relative contribution of each feature to the overall method performance to be determined in a straightforward manner. To this end, we selected those features that exhibited a relative importance of at least 0.002 **(Fig. 3A**). This led to the selection of a total of 128 features that were included in our final method (**Fig. 3B**). Interestingly, though CTD represents only a very small fraction of the total features evaluated (147/12,829 = ~1.1%), 3 of the top 10 features fall within this feature class (**Fig. 3C**). In particular, the "charged" attribute within CTD is highly represented. For instance, "ChargeD1001" and "ChargeC1", which were the two top-ranked features, are distribution and composition features describing the extent of positive charge in the sequence, respectively. This suggests that the charge state of the residues surrounding a given lysine residue may be an important factor in determining whether it is glutarylated. For example, high charge density in the vicinity of Lys may lower the pK_a of its ε-amino group ($-NH_3$), stabilizing the amine ($-NH_2$) and facilitating nucleophilic attack on glutaryl-CoA.[129,130]

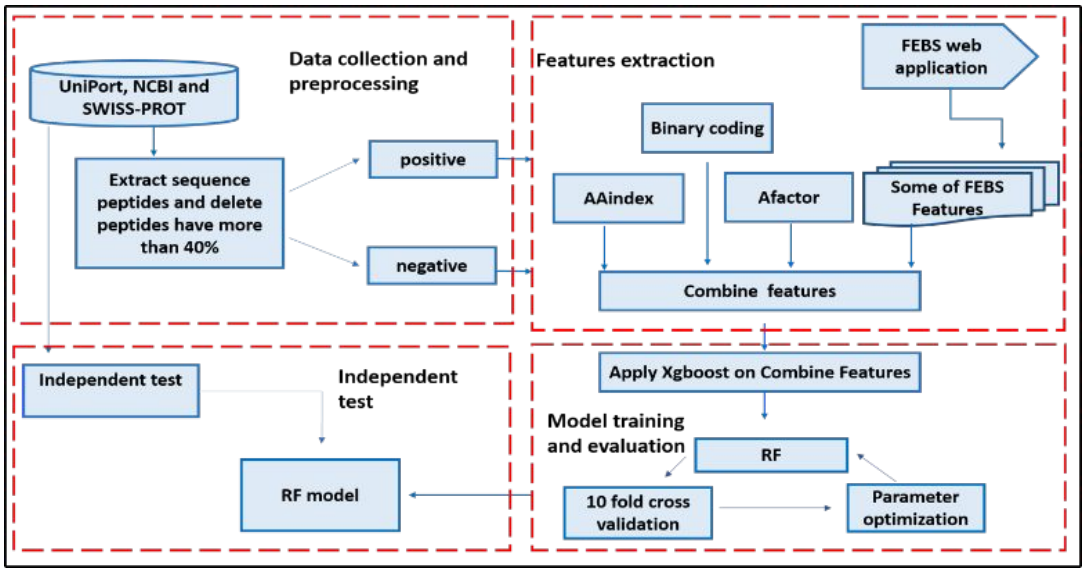

Fig 2. Flowchart of strategy used for the development and evaluation of RF-glutarysite. Positive and negative sets were generated from public databases and literature resources. Features were then extracted from the FEPS web server and combined with additional features, including amino acid index (AAindex), amino acid factor (Afactor) and binary encoding. Random forest (RF) classifier was then used to select the most important features and, after parameter optimization, the resulting algorithm was evaluated using both 10-fold cross validation and an independent test set

Consistently, "PAAC12", which is a pseduo-amino composition feature that combines conventional amino acid composition attributes with information about the sequence order, was one of the ten most important features (**Fig. 3C and Fig. S3**). Likewise, Moran autocorrelation, Geary autocorrelation and Normalized Morea-Broto autocorrelation, which assess local spatial patterns within a sequence, were each represented in the top 10 (**Fig. 3C**). More broadly, CTD, Moran autocorrelation, Geary autocorrelation, Normalized Morea-Broto autocorrelation and PseAAC appear to play the largest role in discriminating between sites that are glutarylated and those that are not (**Fig. 3B**).

With optimal features in hand, we next sought to develop a glutarylation site prediction method based on the optimal feature set. Since imbalanced datasets can affect the accuracy of various types of machine learning algorithms due to overfitting, during training we balanced the positive and negative datasets using an under-sampling strategy.[95,131]. We then evaluated the four machine learning classifiers described above using the optimal feature set. Similar to the results obtained when using the entire feature set, RF achieved the highest efficiency scores by each metric. For instance, based on 10-fold cross-validation, RF achieved ACC, SN, SP and MCC scores of 75%, 81%, 68% and 0.50, respectively (**Table 8**). Likewise, RF exhibited the highest area under the receiver-operator curve (AUC) of any of the classifiers tested (**Fig. 4**). In contrast, SVM, which was the next best classifier, exhibited efficiency scores of 67%, 71%, 63%, 0.34, and 0.71 for ACC, SN, SP, MCC and AUC, respectively (**Table 8; Fig. 4**). The overall hierarchy based on 10-fold cross-validation was RF, SVM, KNN and NB. A similar hierarchy was observed when the classifiers were evaluated using our independent test set, with RF yielding scores of 72%, 73%, 70%, 0.43, and 0.81 for ACC, SN, SP, MCC and AUC, respectively (**Table 9; Fig. 5**). Therefore, we chose to develop our method using RF. The resulting algorithm, which we termed RF-GlutarySite, is designed to predict putative sites of glutarylation based on a protein's primary amino acid sequence.

Notably, though RF-GlutarySite uses only ~1% of the original feature attributes (128 out of the initial 12,829 features), marked improvements in performance were observed compared to the full feature set (e.g., compare **Tables 8 & 9** and **Tables S1 & S2**). For instance, when evaluated by 10-fold cross-validation, the ACC, SN and SP scores exhibited by RF-GlutarySite increased by 8.7%, 5.2% and 9.7%, respectively, compared to those achieved using the entire feature set. This led to a 28.2% increase in MCC when using the optimal feature set. Similar gains were observed when the independent test set was used for evaluation, culminating in a 19.4% increase in

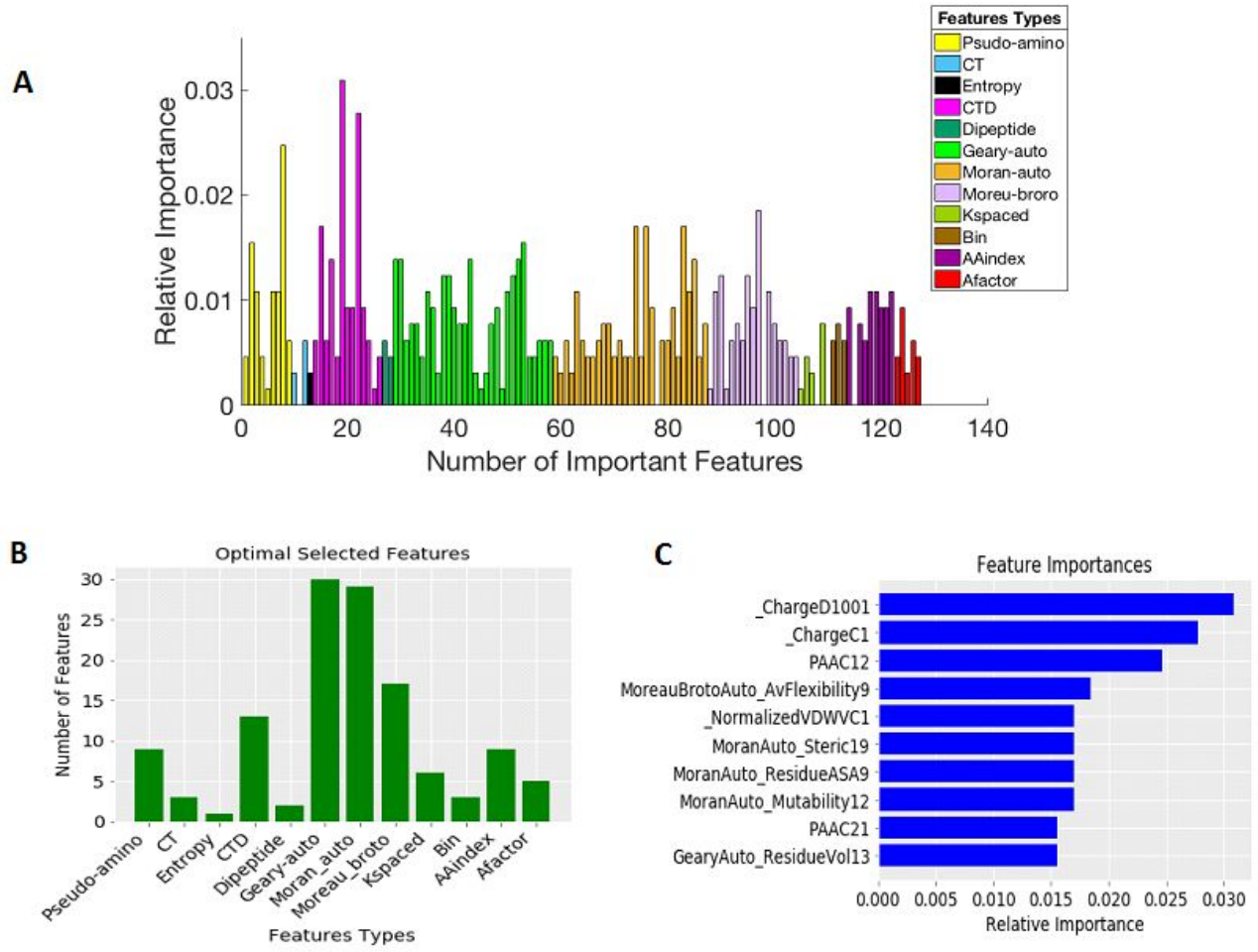

Fig. 3. Feature importance and selection of optimal features. **A.** Relative importance of the 335 features from among the 12,829 initial features that exhibited a relative importance of at least 0.002. These features were used for model development. **B.** Number of features within each feature class that were selected for model development in A. **C.** Relative feature importance of the top 10 features selected for model development.

MCC. Together, these data suggest that RF-GlutarySite is able to improve method performance while using only a fraction of the features in the initial feature set. As a result, RF-GlutarySite increases efficiency—and, by extension, decreases computational cost—without sacrificing performance.

3.2. Comparison with existing methods.

Next, we compared our method to the existing glutarylation site predictors, GlutPred[32] and iGlu-Lys.[33] To ensure that we did not bias the results by using our training set for evaluation, we retrieved the training and independent test sets from GlutPred's webserver and evaluated the performance of all three methods based on ACC, SN, SP and MCC.[32,‡] Though RF-GlutarySite exhibited lower ACC and SP scores than GlutPred and iGlu-Lys when assessed by 10-fold cross-validation, it achieved the highest SN score of all the methods tested (**Table 10**). Indeed, the SN score observed for RF-GlutarySite was ~16% higher than that observed for GlutPred and ~49% higher than that exhibited by iGlu-Lys. Similar results were observed using the independent test set, where RF-GlutarySite achieved an SN score that was >42% higher than those of GlutPred and iGlu-Lys (**Table 11**). In contrast, both GlutPred and iGlu-Lys outperformed our method with respect to ACC and SP using the independent test set from GlutPred's webserver. For instance, RF-GlutarySite exhibited an ACC score that was 5.4% lower than that for GlutPred and 19.4% lower than that for iGlu-Lys (**Table 11**). Likewise, RF-GlutarySite's SP score was 12.7% and 28.1% lower than those for GlutPred and iGlu-Lys. We suspect that these discrepancies may stem from the fact that our method was developed using only non-homologous fragments while the other methods did not remove homologous peptides from their training set, particularly from the negative dataset.

Consequently, GlutPred and iGlu-Lys likely exhibited higher TN rates than our method, which improved their performance with respect to metrics that are heavily influenced by TN rate (e.g., ACC and SP) while having little effect on those that are not (e.g., SN). Consistent with this notion, RF-GlutarySite also performed markedly better than GlutPred with respect to precision (PR) and F1-score, which are both sensitive to TP rate

Table 8. Comparison between various machine learning algorithms using the optimal feature set based on 10-fold cross-validation.

Features	ACC(%)	SN(%)	SP(%)	MCC
RF-GlutarySite	75	81	68	0.50
Support vector machine (SVM)	67	71	63	0.34
Naïve Bayes (NB)	61	60	62	0.22
K-nearest neighbor(KNN)	64	65	64	0.29

Table 9. Comparison between various machine learning algorithms using the optimal feature set based on independent test set.

Features	ACC(%)	SN(%)	SP(%)	MCC
RF-GlutarySite	72	73	70	0.43
Support vector machine (SVM)	63	66	59	0.25
Naïve Bayes (NB)	52	43	61	0.05
K-nearest neighbor(KNN)	62	61	64	0.25

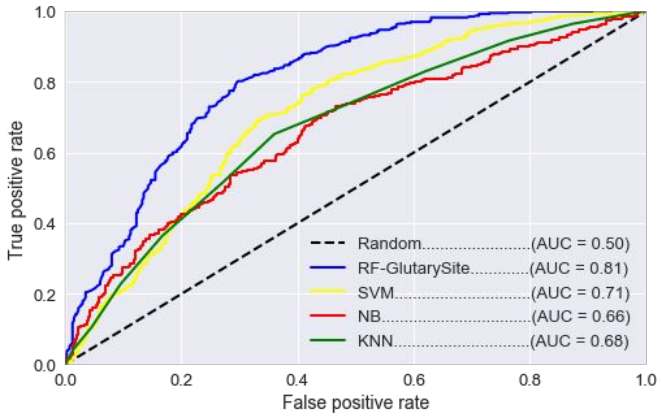

Fig. 4. Receiver operator characteristic (ROC) curves for optimal features used to develop our method as well as other machine learning algorithms using the same number of features based on 10-fold cross-validation. The area under the curve (AUC) for each algorithm is given in parentheses. SVM: Support vector machine; NB: Naïve Bayesian; KNN: k-nearest neighbour; RF: Random forest.

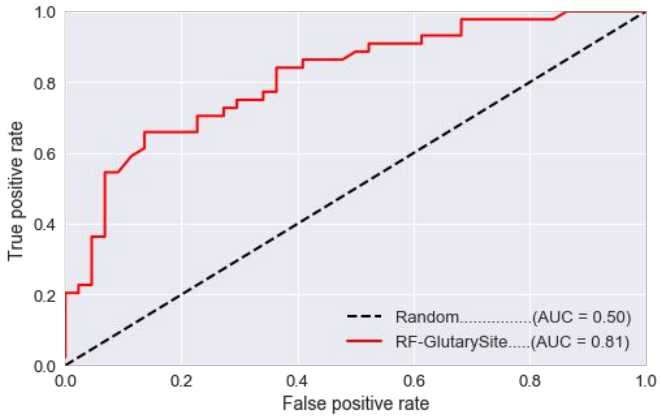

Fig. 5. Receiver operator characteristic (ROC) curves for optimal features used to develop our method based on the independent set. The area under the curve (AUC) for each algorithm is given in parentheses.

but not TN rate (**Tables 10 & 11**)[€]. For instance, RF-GlutarySite achieved PR and F1-scores that were 2.9- and 2.2-times higher, respectively, than those exhibited by GlutPred using the independent test set from the GlutPred webserver. Likewise, when using the independent test set from the GlutPred webserver, the precision-recall curve (PRC) for RF-GlutarySite increased more rapidly and remained higher as it approached 100% recall than did GlutPred's PRC (**Fig. 6**). As a result, the area under the PRC (AUC-PRC) for RF-GlutarySite was ~2.5-times higher than of GlutPred (**Fig. 6; Table 11**).

Together, these data suggest that, though RF-GlutarySite does not perform as well as existing methods with respect to metrics that are heavily influenced by TN rates (such as SP and ACC), it achieved the highest scores among all of the methods tested with respect to SN (and, compared to GlutPred, with respect to PR and F1-score). Since SN, PR and F1-score are more sensitive to TP rates than to TN rates, RF-GlutarySite may be more likely to identify positive sites than existing glutarylation site predictors. In support of this notion, Geng et al. recently found that high SN scores correlated most strongly with the ability to predict active sites based on protein-protein interactions profiles[61].

Table 10. Performance comparison of existing glutarylation site prediction methods based on 10-fold cross validation.

Predictor	ACC(%)	SN(%)	SP(%)	PR(%)	F1	MCC
RF-GlutarySite	72.3	74.9	69.7	71.2	0.73	0.45
GlutPred*[32]	74.9	64.8	76.6	31.8	0.43	0.32
iGlu-Lys*[33]	88.4	50.4	95.2	-	-	0.50

Table 11. Performance comparison of existing glutarylation site prediction methods based on independent test set.

Predictor	ACC(%)	SN(%)	SP(%)	PR(%)	F1	MCC
RF-GlutarySite	71.3	74.1	68.5	70.2	0.72	0.43
GlutPred*[32]	75.4	51.8	78.5	24.0	0.33	0.22
iGlu-Lys*[33]	88.5	51.4	95.3	-	-	0.52

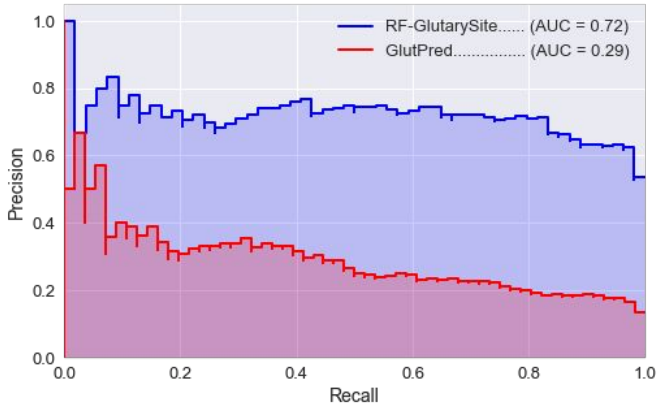

Fig. 6. Precision Recall Curves (PRC) for RF-GlutarySite (blue) and GlutPred (red) using the independent set from the GlutPred webserver. The area under the curve (AUC) for each method is given in parentheses

4. Conclusion.

In this study, we developed a novel method to predict glutarylation sites based on the primary amino acid sequence of proteins. This method, which is termed "RF-GlutarySite', uses a RF classifier together with Xgboost feature selection to identify important features and reduce dimensionality lengths. Based on evaluation using both 10-fold cross-validation and an independent test set, RF-GlutarySite outperforms existing glutarylation site predictors with respect to performance metrics that are most heavily influenced by TP rate (e.g., SN, PR, and F1-score). In contrast, it does not perform as well with respect to metrics that are more sensitive to TN rate (e.g., SP and ACC). In this regard, RF-GlutarySite can be considered complementary to existing glutarylation site prediction methods and may be useful in predicting residues that are most likely to be glutarylated (as opposed those that are most likely not to be modified). Furthermore, the ability of RF to identify features that contribute most substantially to method performance can provide clues about the physiochemical properties that underlie glutarylation site selection in cellular proteins. In the future, method performance, as well as the biochemical insights that can be gained from feature importance, will be improved as more experimentally validated glutarylation sites are identified. Finally, when used in conjunction with proteomics and other PTM prediction methods, RF-GlutarySite may offer insights into crosstalk between glutarylation with other lysine modifications, such as acetylation, methylation succinylation and malnolyation. Together, this information will facilitate a deeper understanding of glutarylation and its impact on cellular physiology. To facilitate its use by the signalling community and the broader scientific community, the RF-GlutarySite software, code and documentations are freely available in the GitHub repository (https://github.com/HussamAlbarakati/ RF-GlutarySite). We are also developing a web server for the RF-GlutarySite tool, which should be available shortly on the KC lab website (http://bcb.ncat.edu/softwares/).

Conflicts of interest.

There are no conflict to declare.

Acknowledgements.

This work was supported by National Science Foundation (NSF) grant nos. 1647884 and 1564606 (to DBK), DBI-0939454 (to RHN and DBK) and National Institutes of Health grant 5SC2GM113784 to RHN. Portion of the work was done by DBK while he was a JSPS visiting professor at Kyushu University.

Notes.

ǂ It should be noted that GlutPred and iGlu-Lys used the same training and independent datasets for model development and evaluation, respectively.

€ PR, F1-scores and PRC were not reported for iGlu-Lys, therefore no comparison could be made with respect to these parameters.

References.

1. Walsh, C.T., Garneau-Tsodikova, S. & Gatto, G.J., Jr. Protein posttranslational modifications: the chemistry of proteome diversifications. *Angew Chem Int Ed Engl* 44, 7342-7372 (2005).

2. Xu, Y., Ding, J. & Wu, L.Y. iSulf-Cys: Prediction of S-sulfenylation Sites in Proteins with Physicochemical Properties of Amino Acids. *PLoS One* 11, e0154237 (2016).

3. Maeda, A.*, et al.* Palmitoylation stabilizes unliganded rod opsin. *Proc Natl Acad Sci U S A* 107, 8428-8433 (2010).

4. Hunter, T. Tyrosine phosphorylation: thirty years and counting. *Curr Opin Cell Biol* 21, 140-146 (2009).

5. Newman, R.H., Zhang, J. & Zhu, H. Toward a systems-level view of dynamic phosphorylation networks. *Front Genet* 5, 263 (2014).

6. Kamynina, E. & Stover, P.J. The Roles of SUMO in Metabolic Regulation. *Adv Exp Med Biol* 963, 143-168 (2017).

7. Mann, M. & Jensen, O.N. Proteomic analysis of post-translational modifications. *Nat Biotechnol* 21, 255-261 (2003).

8. Wang, Y.-C., Peterson, S.E. & Loring, J.F. Protein post-translational modifications and regulation of pluripotency in human stem cells. *Cell research* 24, 143 (2014).

9. Lanouette, S., Mongeon, V., Figeys, D. & Couture, J.F. The functional diversity of protein lysine methylation. *Mol Syst Biol* 10, 724 (2014).

10. Shaid, S., Brandts, C.H., Serve, H. & Dikic, I. Ubiquitination and selective autophagy. *Cell Death Differ* 20, 21-30 (2013).

11. Choudhary, C., Weinert, B.T., Nishida, Y., Verdin, E. & Mann, M. The growing landscape of lysine acetylation links metabolism and cell signalling. *Nat Rev Mol Cell Biol* 15, 536-550 (2014).

12. Huang, H., Lin, S., Garcia, B.A. & Zhao, Y. Quantitative proteomic analysis of histone modifications. *Chem Rev* 115, 2376-2418 (2015).

13. Hendriks, I.A. & Vertegaal, A.C. A comprehensive compilation of SUMO proteomics. *Nat Rev Mol Cell Biol* 17, 581-595 (2016).

14. Liu, Z.*, et al.* CPLM: a database of protein lysine modifications. *Nucleic Acids Res* 42, D531-536 (2014).

15. Nishida, Y.*, et al.* SIRT5 regulates both cytosolic and mitochondrial protein malonylation with glycolysis as a major target. *Molecular cell* 59, 321-332 (2015).

16. Du, Y.*, et al.* Lysine malonylation is elevated in type 2 diabetic mouse models and enriched in metabolic associated proteins. *Molecular & Cellular Proteomics* 14, 227-236 (2015).

17. Zhao, S.*, et al.* Regulation of cellular metabolism by protein lysine acetylation. *Science* 327, 1000-1004 (2010).

18. Olsen, C.A. Expansion of the lysine acylation landscape. *Angew Chem Int Ed Engl* 51, 3755-3756 (2012).

19. Chen, Y.*, et al.* Lysine propionylation and butyrylation are novel post-translational modifications in histones. *Mol Cell Proteomics* 6, 812-819 (2007).

20. Hirschey, M.D. & Zhao, Y. Metabolic Regulation by Lysine Malonylation, Succinylation, and Glutarylation. *Mol Cell Proteomics* 14, 2308-2315 (2015).

21. Tan, M.*, et al.* Lysine glutarylation is a protein posttranslational modification regulated by SIRT5. *Cell metabolism* 19, 605-617 (2014).

22. Osborne, B., Bentley, N.L., Montgomery, M.K. & Turner, N. The role of mitochondrial sirtuins in health and disease. *Free Radic Biol Med* 100, 164-174 (2016).

23. Xie, L.*, et al.* Proteome-wide Lysine Glutarylation Profiling of the Mycobacterium tuberculosis H37Rv. *J Proteome Res* 15, 1379-1385 (2016).

24. Schmiesing, J.*, et al.* Disease-Linked Glutarylation Impairs Function and Interactions of Mitochondrial Proteins and Contributes to Mitochondrial Heterogeneity. *Cell Rep* 24, 2946-2956 (2018).

25. Hou, T.*, et al.* LAceP: lysine acetylation site prediction using logistic regression classifiers. *PloS one* 9, e89575 (2014).

26. Xu, Y., Ding, Y.X., Ding, J., Wu, L.Y. & Xue, Y. Mal-Lys: prediction of lysine malonylation sites in proteins integrated sequence-based features with mRMR feature selection. *Sci Rep* 6, 38318 (2016).

27. Ismail, H.D., Jones, A., Kim, J.H., Newman, R.H. & Kc, D.B. RF-Phos: a novel general Phosphorylation site prediction tool based on random Forest. *BioMed research international* 2016(2016).

28. Ismail, H.D. & Newman, R.H. RF-Hydroxysite: a random forest based predictor for hydroxylation sites. *Molecular BioSystems* 12, 2427-2435 (2016).

29. Chou, K.-C. Some remarks on protein attribute prediction and pseudo amino acid composition. *Journal of theoretical biology* 273, 236-247 (2011).

30. Jia, J., Liu, Z., Xiao, X., Liu, B. & Chou, K.C. pSuc-Lys: Predict lysine succinylation sites in proteins with PseAAC and ensemble random forest approach. *J Theor Biol* 394, 223-230 (2016).

31. Ju, Z. & He, J.-J. Prediction of lysine propionylation sites using biased svm and incorporating four different sequence features into chou's pseaac. *Journal of Molecular Graphics and Modelling* 76, 356-363 (2017).

32. Ju, Z. & He, J.-J. Prediction of lysine glutarylation sites by maximum relevance minimum redundancy feature selection. *Analytical biochemistry* 550, 1-7 (2018).

33. Xu, Y., Yang, Y., Ding, J. & Li, C. iGlu-Lys: A Predictor for Lysine Glutarylation Through Amino Acid Pair Order Features. *IEEE transactions on nanobioscience* 17, 394-401 (2018).

34. Xu, H.*, et al.* PLMD: An updated data resource of protein lysine modifications. *J Genet Genomics* 44, 243-250 (2017).

35. Huang, Y., Niu, B., Gao, Y., Fu, L. & Li, W. CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* 26, 680-682 (2010).

36. Ismail, H.D., Smith, M. & Dukka, B. FEPS: Feature Extraction from Protein Sequences webserver.

37. Ismail, H.D., Saigo, H. & KC, D.B. RF-NR: Random forest based approach for improved classification of Nuclear Receptors. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)* 15, 1844-1852 (2018).

38. Shen, H.B. & Chou, K.C. PseAAC: a flexible web server for generating various kinds of protein pseudo amino acid composition. *Anal Biochem* 373, 386-388 (2008).

39. Li, Z.R*., et al.* PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucleic Acids Res* 34, W32-37 (2006).

40. Cao, D.S., Xu, Q.S. & Liang, Y.Z. propy: a tool to generate various modes of Chou's PseAAC. *Bioinformatics* 29, 960-962 (2013).

41. Chou, K.C. & Elrod, D.W. Protein subcellular location prediction. *Protein Eng* 12, 107-118 (1999).

42. Lumbanraja, F.R*., et al.* Improved Protein Phosphorylation Site Prediction by a New Combination of Feature Set and Feature Selection. *Journal of Biomedical Science and Engineering* 11, 144 (2018).

43. Shen, J*., et al.* Predicting protein-protein interactions based only on sequences information. *Proc Natl Acad Sci U S A* 104, 4337-4341 (2007).

44. Xiao, N., Cao, D.S., Zhu, M.F. & Xu, Q.S. protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences. *Bioinformatics* 31, 1857-1859 (2015).

45. Wang, Y.C., Wang, Y., Yang, Z.X. & Deng, N.Y. Support vector machine prediction of enzyme function with conjoint triad feature and hierarchical context. *BMC Syst Biol* 5 Suppl 1, S6 (2011).

46. Wang, H. & Hu, X. Accurate prediction of nuclear receptors with conjoint triad feature. *BMC Bioinformatics* 16, 402 (2015).

47. Yin, Z. & Tan, J. New encoding schemes for prediction of protein Phosphorylation sites. in *Systems Biology (ISB), 2012 IEEE 6th International Conference on* 56-62 (IEEE, 2012).

48. Shannon, C.E. A mathematical theory of communication. *Bell system technical journal* 27, 379-423 (1948).

49. Szoniec, G. & Ogorzalek, M.J. Entropy of never born protein sequences. *SpringerPlus* 2, 200 (2013).

50. Raza, K. Protein features identification for machine learning-based prediction of protein-protein interactions. in *International Conference on Information, Communication and Computing Technology* 305-317 (Springer, 2017).

51. Nigatu, D., Sobetzko, P., Yousef, M. & Henkel, W. Sequence-based information-theoretic features for gene essentiality prediction. *BMC bioinformatics* 18, 473 (2017).

52. Johansson, F. & Toh, H. A comparative study of conservation and variation scores. *Bmc Bioinformatics* 11, 388 (2010).

53. Li, C., Wang, J. & Zhang, Y. Similarity analysis of protein sequences based on the normalized relative-entropy. *Combinatorial chemistry & high throughput screening* 11, 477-481 (2008).

54. Erill, I. & O'Neill, M.C. A reexamination of information theory-based methods for DNA-binding site identification. *BMC bioinformatics* 10, 57 (2009).

55. Cai, C., Han, L., Ji, Z.L., Chen, X. & Chen, Y.Z. SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence. *Nucleic acids research* 31, 3692-3697 (2003).

56. Thomas, S., Karnik, S., Barai, R.S., Jayaraman, V.K. & Idicula-Thomas, S. CAMP: a useful resource for research on antimicrobial peptides. *Nucleic acids research* 38, D774-D780 (2009).

57. Bhadra, P., Yan, J., Li, J., Fong, S. & Siu, S.W. AmPEP: Sequence-based prediction of antimicrobial peptides using distribution patterns of amino acid properties and random forest. *Scientific reports* 8, 1697 (2018).

58. Govindan, G. & Nair, A.S. Composition, Transition and Distribution (CTD)—a dynamic feature for predictions based on hierarchical structure of cellular sorting. in *India Conference (INDICON), 2011 Annual IEEE* 1-6 (IEEE, 2011).

59. Ong, S.A., Lin, H.H., Chen, Y.Z., Li, Z.R. & Cao, Z. Efficacy of different protein descriptors in predicting protein functional families. *Bmc Bioinformatics* 8, 300 (2007).

60. Dubchak, I., Muchnik, I., Holbrook, S.R. & Kim, S.-H. Prediction of protein folding class using global description of amino acid sequence. *Proceedings of the National Academy of Sciences* 92, 8700-8704 (1995).

61. Geng, H., Lu, T., Lin, X., Liu, Y. & Yan, F. Prediction of protein-protein interaction sites based on naive Bayes classifier. *Biochemistry research international* 2015(2015).

62. Šícho, M., de Bruyn Kops, C., Stork, C., Svozil, D. & Kirchmair, J. FAME 2: Simple and Effective Machine Learning Model of Cytochrome P450 Regioselectivity. *Journal of chemical information and modeling* 57, 1832-1846 (2017).

63. Chen, C.-W., Lin, J. & Chu, Y.-W. iStable: off-the-shelf predictor integration for predicting protein stability changes. in *BMC bioinformatics*, Vol. 14 S5 (BioMed Central, 2013).

64. Chen, Y.-Z., Chen, Z., Gong, Y.-A. & Ying, G. SUMOhydro: a novel method for the prediction of sumoylation sites based on hydrophobic properties. *PLoS One* 7, e39195 (2012).

65. Hasan, M.M., Yang, S., Zhou, Y. & Mollah, M.N.H. SuccinSite: a computational tool for the prediction of protein succinylation sites by exploiting the amino acid patterns and properties. *Molecular BioSystems* 12, 786-795 (2016).

66. Reczko, M. & Bohr, H. The DEF data base of sequence based protein fold class predictions. *Nucleic acids research* 22, 3616 (1994).

67. Bhasin, M. & Raghava, G. GPCRpred: an SVM-based method for prediction of families and subfamilies of G-protein coupled receptors. *Nucleic Acids Research* 32, W383-W389 (2004).

68. Chaudhuri, R., Ansari, F.A., Raghunandanan, M.V. & Ramachandran, S. FungalRV: adhesin prediction and

immunoinformatics portal for human fungal pathogens. *BMC genomics* 12, 192 (2011).

69. Gupta, S., Sharma, A.K., Shastri, V., Madhu, M.K. & Sharma, V.K. Prediction of anti-inflammatory proteins/peptides: an insilico approach. *Journal of translational medicine* 15, 7 (2017).

70. Bartholomew, D.J. Time series analysis forecasting and control. *Journal of the Operational Research Society* 22, 199-201 (1971).

71. Broto, P., Moreau, G. & Vandycke, C. Molecular structures: perception, autocorrelation descriptor and sar studies: system of atomic contributions for the calculation of the n-octanol/water partition coefficients. *European journal of medicinal chemistry* 19, 71-78 (1984).

72. Ren, X.-M. & Xia, J.-F. Prediction of protein-protein interaction sites by using autocorrelation descriptor and support vector machine. in *Advanced Intelligent Computing Theories and Applications. With Aspects of Artificial Intelligence* 76-82 (Springer, 2010).

73. Cid, H., Bunster, M., Canales, M. & Gazitúa, F. Hydrophobicity and structural classes in proteins. *Protein Engineering, Design and Selection* 5, 373-375 (1992).

74. Bhaskaran, R. & Ponnuswamy, P. Positional flexibilities of amino acid residues in globular proteins. *International Journal of Peptide and Protein Research* 32, 241-255 (1988).

75. Charton, M. & Charton, B.I. The structural dependence of amino acid hydrophobicity parameters. *Journal of Theoretical Biology* 99, 629-644 (1982).

76. Chothia, C. The nature of the accessible and buried surfaces in proteins. *Journal of molecular biology* 105, 1-12 (1976).

77. Bigelow, C.C. On the average hydrophobicity of proteins and the relation between it and protein structure. *Journal of Theoretical Biology* 16, 187-211 (1967).

78. Charton, M. Protein folding and the genetic code: an alternative quantitative model. *Journal of theoretical biology* 91, 115-123 (1981).

79. Dayhoff, M., Schwartz, R. & Orcutt, B. 22 a model of evolutionary change in proteins. *Atlas of protein sequence and structure*, 345-352 (1978).

80. Moreau, G. & Broto, P. Auto-correlation of molecular-structures, application to sar studies. *Nouveau Journal de Chimie-New Journal of Chemistry* 4, 757-764 (1980).

81. Moran, P.A. Notes on continuous stochastic phenomena. *Biometrika* 37, 17-23 (1950).

82. Geary, R.C. The contiguity ratio and statistical mapping. *The incorporated statistician* 5, 115-146 (1954).

83. Ansari, H.R. & Raghava, G.P. Identification of conformational B-cell Epitopes in an antigen from its primary sequence. *Immunome research* 6, 6 (2010).

84. Xiao, X., Shao, S., Ding, Y., Huang, Z. & Chou, K.-C. Using cellular automata images and pseudo amino acid composition to predict protein subcellular location. *Amino acids* 30, 49-54 (2006).

85. Xu, Y., Wang, X.-B., Ding, J., Wu, L.-Y. & Deng, N.-Y. Lysine acetylation sites prediction using an ensemble of support vector machine classifiers. *Journal of Theoretical Biology* 264, 130-135 (2010).

86. Kawashima, S.*, et al.* AAindex: amino acid index database, progress report 2008. *Nucleic acids research* 36, D202-D205 (2007).

87. Rubinstein, N.D., Mayrose, I. & Pupko, T. A machine-learning approach for predicting B-cell epitopes. *Molecular immunology* 46, 840-847 (2009).

88. Torkamani, A. & Schork, N.J. Accurate prediction of deleterious protein kinase polymorphisms. *Bioinformatics* 23, 2918-2925 (2007).

89. Marsella, L., Sirocco, F., Trovato, A., Seno, F. & Tosatto, S.C. REPETITA: detection and discrimination of the periodicity of protein solenoid repeats by discrete Fourier transform. *Bioinformatics* 25, i289-i295 (2009).

90. Atchley, W.R., Zhao, J., Fernandes, A.D. & Drüke, T. Solving the protein sequence metric problem. *Proceedings of the National Academy of Sciences* 102, 6395-6400 (2005).

91. Chen, Y.-Z., Tang, Y.-R., Sheng, Z.-Y. & Zhang, Z. Prediction of mucin-type O-glycosylation sites in mammalian proteins using the composition of k-spaced amino acid pairs. *BMC bioinformatics* 9, 101 (2008).

92. Xu, H.-D., Shi, S.-P., Wen, P.-P. & Qiu, J.-D. SuccFind: a novel succinylation sites online prediction tool via enhanced characteristic strategy. *Bioinformatics* 31, 3748-3750 (2015).

93. Zhao, X., Zhang, W., Xu, X., Ma, Z. & Yin, M. Prediction of protein phosphorylation sites by using the composition of k-spaced amino acid pairs. *PloS one* 7, e46302 (2012).

94. Chen, K., Kurgan, L.A. & Ruan, J. Prediction of flexible/rigid regions from protein sequences using k-spaced amino acid pairs. *BMC structural biology* 7, 25 (2007).

95. Chen, Z.*, et al.* Prediction of ubiquitination sites by using the composition of k-spaced amino acid pairs. *PloS one* 6, e22930 (2011).

96. Daskalaki, S., Kopanas, I. & Avouris, N. Evaluation of classifiers for an uneven class distribution problem. *Applied artificial intelligence* 20, 381-417 (2006).

97. He, H. & Garcia, E.A. Learning from imbalanced data. *IEEE Transactions on Knowledge & Data Engineering*, 1263-1284 (2008).

98. KrishnaVeni, C. & Sobha Rani, T. On the classification of imbalanced datasets. *IJCST* 2, 145-148 (2011).

99. Guo, X., Yin, Y., Dong, C., Yang, G. & Zhou, G. On the class imbalance problem. in *Natural Computation, 2008. ICNC'08. Fourth International Conference on*, Vol. 4 192-201 (IEEE, 2008).

100. Kotsiantis, S., Kanellopoulos, D. & Pintelas, P. Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering* 30, 25-36 (2006).

101. Barbu, A., She, Y., Ding, L. & Gramajo, G. Feature selection with annealing for computer vision and big data learning. *IEEE transactions on pattern analysis and machine intelligence* 39, 272-286 (2017).

102. Wang, R., Perez-Riverol, Y., Hermjakob, H. & Vizcaíno, J.A. Open source libraries and frameworks for biological data visualisation: A guide for developers. *Proteomics* 15, 1356-1374 (2015).

103. Wang, S., Wang, D., Li, J., Huang, T. & Cai, Y.-D. Identification and analysis of the cleavage site in a signal peptide using SMOTE, dagging, and feature selection methods. *Molecular omics* 14, 64-73 (2018).

104. Perez-Riverol, Y., Kuhn, M., Vizcaíno, J.A., Hitz, M.-P. & Audain, E. Accurate and fast feature selection workflow for high-dimensional omics data. *PloS one* 12, e0189875 (2017).

105. Soufan, O., Kleftogiannis, D., Kalnis, P. & Bajic, V.B. DWFS: a wrapper feature selection tool based on a parallel genetic algorithm. *PloS one* 10, e0117988 (2015).

106. Michalak, K. & Kwaśnicka, H. Correlation-based feature selection strategy in classification problems. *International Journal of Applied Mathematics and Computer Science* 16, 503-511 (2006).

107. Wang, Y.*, et al.* Gene selection from microarray data for cancer classification—a machine learning approach. *Computational biology and chemistry* 29, 37-46 (2005).

108. Wang, Y., Makedon, F. & Pearlman, J. Tumor classification based on DNA copy number aberrations determined using SNP arrays. *Oncology reports* 15, 1057-1059 (2006).

109. Kohavi, R. & John, G.H. Wrappers for feature subset selection. *Artificial intelligence* 97, 273-324 (1997).

110. Seo, M. & Oh, S. CBFS: High performance feature selection algorithm based on feature clearness. *PloS one* 7, e40419 (2012).

111. Usai, M.G., Goddard, M.E. & Hayes, B.J. LASSO with cross-validation for genomic selection. *Genetics research* 91, 427-436 (2009).

112. White, C., Ismail, H.D. & Saigo, H. CNN-BLPred: a convolutional neural network based predictor for β-lactamases (BL) and their classes. *BMC bioinformatics* 18, 577 (2017).

113. Stahl, K., Schneider, M. & Brock, O. EPSILON-CP: using deep learning to combine information from multiple sources for protein contact prediction. *BMC bioinformatics* 18, 303 (2017).

114. Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* 785-794 (ACM, 2016).

115. Breiman, L. Random forests. *Machine learning* 45, 5-32 (2001).

116. Polikar, R. Ensemble based systems in decision making. *IEEE Circuits and systems magazine* 6, 21-45 (2006).

117. Rokach, L. Ensemble-based classifiers. *Artificial Intelligence Review* 33, 1-39 (2010).

118. Ma, X., Guo, J., Liu, H.-D., Xie, J.-M. & Sun, X. Sequence-based prediction of DNA-binding residues in proteins with conservation and correlation information. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)* 9, 1766-1775 (2012).

119. Ding, J., Li, X. & Hu, H. TarPmiR: a new approach for microRNA target site prediction. *Bioinformatics* 32, 2768-2775 (2016).

120. Hamby, S.E. & Hirst, J.D. Prediction of glycosylation sites using random forests. *BMC bioinformatics* 9, 500 (2008).

121. Pedregosa, F.*, et al.* Scikit-learn: Machine learning in Python. *Journal of machine learning research* 12, 2825-2830 (2011).

122. Baldi, P., Brunak, S., Chauvin, Y., Andersen, C.A. & Nielsen, H. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* 16, 412-424 (2000).

123. Fawcett, T. An introduction to ROC analysis. *Pattern recognition letters* 27, 861-874 (2006).

124. Hanley, J.A. & McNeil, B.J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143, 29-36 (1982).

125. Davis, J. & Goadrich, M. The relationship between Precision-Recall and ROC curves. in *Proceedings of the 23rd international conference on Machine learning* 233-240 (ACM, 2006).

126. Bleakley, K., Biau, G. & Vert, J.-P. Supervised reconstruction of biological networks with local models. *Bioinformatics* 23, i57-i65 (2007).

127. Saito, T. & Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS one* 10, e0118432 (2015).

128. Bolon-Canedo, Veronica, Noelia Sanchez-Marono, and Amparo Alonso-Betanzos. "Feature selection and classification in multiple class datasets: An application to KDD Cup 99 dataset." Expert Systems with Applications 38.5 (2011): 5947-5957.

129. Highbarger, L.A., Gerlt, J.A. & Kenyon, G.L. Mechanism of the reaction catalyzed by acetoacetate decarboxylase. Importance of lysine 116 in determining the p K a of active-site lysine 115. *Biochemistry* 35, 41-46 (1996).

130. Harris, T.K. & Turner, G.J. Structural basis of perturbed pKa values of catalytic groups in enzyme active sites. *IUBMB life* 53, 85-98 (2002).

131. Hasan, M.M.*, et al.* Computational identification of protein pupylation sites by using profile-based composition of k-spaced amino acid pairs. *PloS one* 10, e0129635 (2015).

132. McKinney, Wes. "Data structures for statistical computing in python." *Proceedings of the 9th Python in Science Conference*. Vol. 445. 2010.