**PREDICTING BITTER PEPTIDES FROM SEQUENCE USING MACHINE LEARNING TECHNIQUES**

A Thesis

by

Nishitha Yendapally

Submitted to the College of Graduate Studies

Texas A&M University-Kingsville

in partial fulfilment of the requirements for the degree of

MASTER OF SCIENCE

August 2021

Major Subject: Computer Science

**ABSTRACT**

Predicting Bitter Peptides from Sequence using Machine Learning Techniques

(August 2021)

Nishitha Yendapally, Bachelor of Technology in Computer Science

Jawaharlal Nehru Technological University - Hyderabad

Chairman of the Advisory Committee: Dr. Avdesh Mishra

In this post-genomic era prediction of peptides bitterness is an important element in developing drugs and nutritional studies. Identifying bitter peptides using an experimental method such as *in-vivo* and *in-vitro* approaches that include “human taste panel studies” can be arduous, tiresome, time-consuming, and costly. In contrary to the state-of-the-art approaches, this thesis proposes to explore various machine learning techniques specifically, a stacking-based approach for an effective prediction of bitter peptides directly from the peptide sequence. In addition, this work will explore various feature encoding and transformation techniques including term frequency-inverse document frequency, Pep2Vec, FastText, and amino acid-wise contact energy matrix transformation. Finally, to validate the robustness, the proposed approach will be compared with the state-of-the-art approaches utilizing the standard performance evaluation metrics such as Mathew’s Correlation Coefficient and Accuracy obtained through cross-validation and independent testing.

# **INTRODUCTION**

Peptides are small chained amino acids. They are found to be useful in the development of drugs to combat critical diseases. Many of these drugs have a bitter taste by nature. Significant efforts are being made to reduce the bitterness of these medicines to enhance taste and, as a result, increase drug compliance [1] [2] [3]. The experimental methods such as *in-vivo* and *in-vitro* approaches that include “human taste panel studies” [4] utilized to identify bitter peptides (BPs) are tedious, expensive, and time-consuming.

## **1.1 LITERATURE REVIEW**

The literature presents several remarkable research on the identification of BPs. Several authors have used machine learning (ML) and statistical approaches for identifying the bitterness in the peptides [5] [6] [7]. The first method that utilizes a ML approach to predict bitterness in peptides is, iBitter-SCM [8]. It uses the scoring card approach with propensity ratings for 20-residues and 400-dipeptides to differentiate the bitterness and non-bitterness of the peptides. In their work, 10 fold CrossValidation (CV) and an independent testing method to examine five ML classifiers: K-Nearest Neighbour (KNN), Random Forest (RF), Nave Bayes (NB), Support Vector Machine (SVM), and Decision Tree (DT) are utilized. As reported, iBitter-SCM [8], obtained a 10-fold CV and accuracies of independent test are 0.871 and 0.844, respectively. Although iBitter-SCM produced relatively high forecast accuracies as mentioned above, its prediction performance should be further improved before it can be utilized in real-world applications. One of the shortcomings of iBitter-SCM is that it utilizes a single feature descriptor.

To develop traditional ML predictors, it is well acknowledged that combining several feature descriptors may significantly enhance prediction performance when compared to a single feature descriptor. However, combining multiple feature descriptors may complicate the problem further as it leads to the problem of the curse of dimensionality. Moreover, combining feature descriptors may involve the inclusion of redundant and noisy information, resulting in poor prediction outcomes. While the identification of useful features using a feature selection algorithm can be employed to address the challenge of high-dimensional feature space, this procedure is tedious since it involves numerous manual, laborious, and trial-and-error attempts. Moreover, the development of traditional ML algorithms is rather difficult, requiring to extract features, feature significance identification, and prediction model optimization.

To overcome the challenges of traditional ML algorithms, the BERT4Bitter [5] method was introduced in the recent past. BERT4Bitter applies bidirectional encoder representation from transformers (BERT) technique for predicting BPs. As BERT is based on the NLP approach, peptide sequence is considered as a sentence, with each of the 20 amino acids regarded as a word. Three NLP based feature encoding methods, such as Term frequency-inverse document frequency (TF-IDF) [9], Pep2Vec, and FastText, are used to encode peptide sequences, which are then fed to ML (ML)-based models as input. In comparision to the iBitter-SCM, BERT4Bitter was observed to be more effective and outperformed the iBitter-SCM with an Mathew’s Correlation Coefficient (MaCC) and Accuracy (Acy) of 0.844 and 0.922, respectively.

This research proposes to explore various ML techniques specifically, stacking-based approaches for an effective prediction of BPs directly from the peptide sequence. In addition, this work will explore various feature encoding and transformation techniques such as term frequency inverse document frequency, Pep2Vec ( is Word2Vec-inspired technique), FastText, and amino acid-wise contact energy matrix transformation.

# **METHODOLOGY**

This chapter explains the proposed dataset, feature extraction techniques, evaluation metrics, and prediction models in more detail.

# **2.1 DATASET**

In the proposed study, the BTP640 dataset that was previously established by Charoenkwan et al. [5] is utilized. It contains of 320-bitter and an equal number of non-bitter non redundant peptides. In order to develop the prediction method and asses its generalization capacity [7], [8], [11],[12], the BTP640 dataset was randomly divided into a training/validation (named BTP-CV) and an independent test set (called BTP-TS) with a ratio of 0.8:0.2. The BTP-CV dataset contains 256 BPs and an equal number of non-BPs [10] that are utilized for validation and training. On the other hand, BTP-TS contains 64 BPs and an equal number of non-BPs that are utilized for independent testing. The previous researches contains more details on the benchmark and independent test dataset [11] [7] [12] [6].

# **2.2 FEATURE EXTRACTION AND SELECTION**

In this research, various feature extraction techniques for accurate prediction of BPs from the peptide sequences are explored. List of features that will be explored in this project includes TF-IDF, Pep2Vec, FastText, RCEM, and more.

**2.2.1 TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY (TF-IDF)**

This is a popular method for document representation in NLP that is based on information retrieval techniques [13]. This method is frequently used to assist model developers in representing documents [13] [14]. In TF-IDF, TF indicates the number of occurrences of word *i* (*ti*) in a particular document *j* (*dj*), and IDF represents the inverse document frequency for words of interest.

**2.2.2 PEP2VEC**

Pep2Vec is a Word2Vec motivated feature generation approach for biological sequences [15]. Google created and developed the Word2Vec technique, which has been effectively utilized in numerous NLP issues and other relevant research applications [16] [17]. In 2016, Aggarwala and Voight [18] suggested a windowing strategy for encoding each sequence as n-dimensional vector of words by splitting it into k-grams [19], which might represent its biophysical and biochemical features achieved through a hierarchical learning process. In Word2Vec, primarly two neural network architectures are utilized to generate word embedding vectors: (i) Continuous bag-of-words (CBOW) and (ii) Continuous skip-gram (skip-gram). CBOW is an unsupervised technique that describes prediction of target word depending on the surrounding words. On the other hand, skip-gram model learns to anticipate a word based on an adjacent word. This model is reverse to CBOW.

* + 1. **FASTTEXT**

Facebook developed a FastText algorithm for NLP applications. However, Asgari and Mofrad published this approach originally [20]. The successful version of FastText was created and introduced by earlier studies [21] [22]. In general, each peptide sequence was regarded as a multi-word phrase. In addition, each word is composed of a bag of n-gram characters. This technique, as reported by Asgari and Mofrad [20] and similar research [21] [22], uses a special symbol (‘and') by placing it at the border of words. This technique can help enhance the ability to distinguish prefixes and suffixes from other character sequences.

# **2.2.4 AMINO ACID-WISE CONTACT ENERGY MATRIX (AACEM):**

Stability of the protein structure is the product of interaction between residues within and outside the structure of the protein. Energy functions derived from known structures may be used to estimate the total energy of these interactions [23] [24] [25]. However, to determine the total energy contribution, the energy potential require knowledge of the 3-dimentional structure. As a result, they are inapplicable to proteins whose structures are unknown, as well as intrinsically disordered proteins (IDPs) [26]. Thus, this thesis work proposes to utilize residue-wise contact energies as a feature to automatically include essential information about amino acid interactions and the possible presence of intrinsically disordered regions (IDRs). These contact energies were calculated [27] by fitting 674 protein primary sequences with contact energy collected from 785 protein tertiary structures using least square fitting. RCEM is 20x20 matrix which represents 20 different amino acids.

# **2.3 METRICS TO EVALUATE PERFORMANCE:**

In this research, a 10-fold CV will be utilized for comparing the outcomes of the suggested approach with the state-of-the-art approaches. First, the data is splitted into ten equal sized folds, where 9-folds are utilized to train and the remaining one fold is utilized to test the method. This procedure will be continued until each fold is tested at least once. In addition to a 10-fold CV, a train-test split method will also be utilized, where the dataset will be segregated into 80% training and 20% test. Table 1 shows the performance evaluation metrics (and the corresponding definitions) that will be considered for an extensive evaluation of the proposed framework.

**Table 1**: Evaluation metrics and their definition.

|  |  |
| --- | --- |
| **Metric** | **Metric Definition** |
| Sensitivity (Sny) |  |
| Specificity (Spy) |  |
| Accuracy (Acy) |  |
| F1 score |  |
| Mathew’s correlation coefficient (MaCC) |  |

# **2.4 ARCHITECTURE**

# **2.4.1 CONVOLUTIONAL NEURAL NETWORK (CNN):**

CNN is initially created to classife the images, but it performs well in natural language processing [28]. In particular, it is taught by detecting low-level characteristics and then utilizing a sequence of convolutional layers to create high-level features. The CNN-based model in general has six layers. They are 1-embedding layer, 1-1D convolutional layer, 1- global max-pooling layer, 1-dropout layer and 2-dense layers. LetNet-5, VGG-16 Net, Inception Net, ResNet, AlexNet, and other CNN architectures will be explored, adopted, and extended for better prediction of BPs from peptide sequences.

# **2.4.2 LONG SHORT-TERM MEMORY (LSTM) NETWORK:**

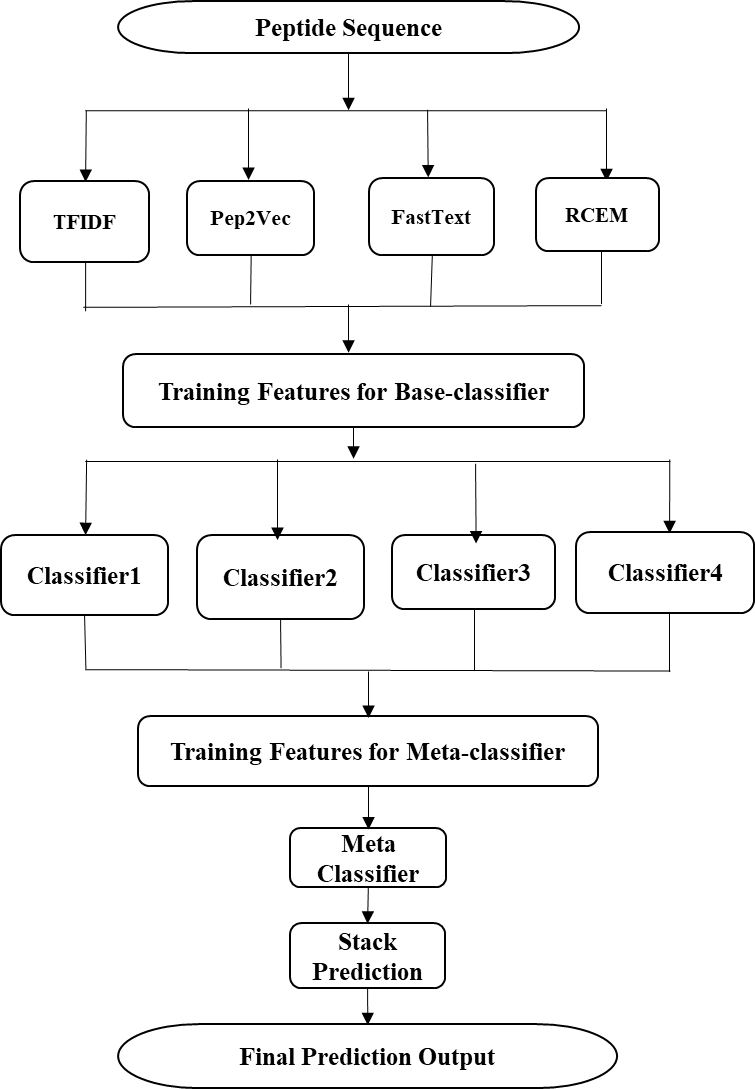
Schmidhuber and Hochreiter originally suggested and proposed LSTM method in 1997. It is an adaptive gating mechanism to overcome the vanishing gradient problem in network topology. The LSTM is a more advanced version of the recurrent RNN for learning long-term dependence information. This approach has already been utilized effectively in a variety of fields, including NLP and related problems [29]. The traditional LSTM design consists of an input, forget,update gates, and a memory block. This architecture will be explored, adopted, and extended for the robust prediction of BPs.

# **2.4.3 BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS (BERT):**

LSTM-based architecture have produced excellent results in a variety of areas, although they have a few drawbacks. To address these issues, Devlin et al. [30] created and presented BERT, a representation of novel language model that overcomes the unidirectional nature of traditional language models. The BERT is a technique based on natural language processing that uses a global receptive field to capture extended global contextual information than CNN-based models. In this study, BERT architecture will also be explored, adopted, and extended for the accurate prediction of BPs.

# **2.4.4 STACKING TECHNIQUE :**

Stacking is an ensemble approach for combining different prediction algorithms to create a new model. It helps reduce the generalization error associated with individual traditional predictive algorithms and is victoriously employed in a variety of ML studies [5] [31].



**Figure 1:** Architecture of the proposed framework.

There are two phases of learning in the stacking framework. The predictor in the first and second stages are referred to as the base-predictor and meta-predictor, respectively. In the first stage, a pool of base-predictors are used. The outputs of the base-predictors are then merged with original feature set in the second stage, to train the meta-predictor with the goal of decreasing the generalization error. To provide additional information about the solution space to the meta-classifier, it is preferable to utilize base-classifiers that differ from each other depending on their fundamental operational principle.

**AIMS:**

* Explore novel features and prediction models for the prediction of BPs.
* Build the prediction model to accurately predict BPs.
* Develop an efficient model to accurately predict BPs.

**OBJECTIVES:**

* Implement ML models with the goals of improving accuracy over existing methods.
* Use appropriate feature extraction strategies to improve the outcomes of the proposed ML model.
* Run and test the suggested model on validation datasets and independent test datasets to assess the robustness of the proposed architecture.
* Ensure that the chosen classifier has a reduced generalization error.

**EXPECTED RESULTS:**

* A model with improved performance for predicting BPs.
* A robust tool that biologists and research experts can trust and utilize to advance the study of drug design and discovery.

# **CONCLUSIONS:**

This research proposes to explore, adapt, and extend ML techniques to predict BPs from the peptide sequence. First, feature extraction techniques such as TF-IDF, Pep2Vec, FastText, RCEM, and others will be explored. Then, the extracted features will be utilized to train different ML algorithms such as SVM, XG-Boost, CNN, RNN, and others. Additionally, a stacking-based technique will be implemented. The ultimate goal of the project will be to reduce generalization error and improve the outcomes of the proposed model over the state-of-the-art approaches. Lastly, different performance measures along with cross-validation and independent testing will be employed to introduce a robust model to the scientific community.

# **References**

|  |  |
| --- | --- |
| [1] | Dagan-Wiener,A. et al. (2017) Bitter or not? BitterPredict, a tool for predicting taste from chemical structure. Sci. Rep., 7, 1–13. |
| [2] | Huang,W. et al. (2016) BitterX: a tool for understanding bitter taste in humans. Sci. Rep., 6, 23450. |
| [3] | Pripp,A. and Ardo,Y. (2007) Modelling relationship between angiotensin-(I)-converting enzyme inhibition and the bitter taste of peptides. Food Chem., 102, 880–888. |
| [4] | Zheng, Suqing, et al. "e-Bitter: bitterant prediction by the consensus voting from the machine-learning methods." Frontiers in chemistry 6 (2018): 82. |
| [5] | Phasit Charoenkwan, Chanin Nantasenamat, Md Mehedi Hasan, Balachandran Manavalan, Watshara Shoombuatong, BERT4Bitter: a bidirectional encoder representations from transformers (BERT)-based model for improving the prediction of bitter peptides, Bioinformatics, 2021;, btab133, https://doi.org/10.1093/bioinformatics/btab133 |
| [6] | Phasit Charoenkwan, Janchai Yana, Nalini Schaduangrat, Chanin Nantasenamat, Md. Mehedi Hasan, Watshara Shoombuatong, iBitter-SCM: Identification and characterization of bitter peptides using a scoring card method with propensity scores of dipeptides Genomics,Volume 112, Issue 4,2020,Pages 2813-2822,ISSN 0888-7543, |
| [7] | Arroyo-Ferna´ndez,I. et al. (2019) Unsupervised sentence representations as word information series: revisiting TF–IDF. Comput. Speech Language, 56,107–129. |
| [8] | Charoenkwan,P. et al. (2020a) iAMY-SCM: improved prediction and analysis of amyloid proteins using a scoring card method with propensity scores of dipeptides. Genomics, 112, 2813–2822. |
| [9] | Charoenkwan,P. et al. (2020b) iDPPIV-SCM: a sequence-based predictor for identifying and analyzing dipeptidyl peptidase IV (DPP-IV) inhibitory peptides using a scoring card method. J. Proteome Res., 19, 4125–4136. |
| [10] | Charoenkwan,P. et al. (2020c) iUmami-SCM: a novel sequence-based predictor for prediction and analysis of umami peptides using a scoring card method with propensity scores of dipeptides. J. Chem. Inf. Model., 60,6666–6678 |
| [11] | Charoenkwan,P. et al. (2020d) iBitter-SCM: identification and characterization of bitter peptides using a scoring card method with propensity scores of dipeptides. Genomics, 112, 2813-2822 |
| [12] | Wei,L. et al. (2020) Computational prediction and interpretation of cell-specific replication origin sites from multiple eukaryotes by exploiting stacking framework. Brief. Bioinform., 2020, bbaa275. |
| [13] | A. Aizawa, “An information-theoretic perspective of tf–idf measures.,” *Information Processing & Management,* pp. 45-65, 2003. |
| [14] | Chen,K. et al. (2016) Turning from TF-IDF to TF-IGM for term weighting in text classification. Expert Syst. Appl., 66, 245–260. |
| [15] | Mikolov,T. et al. (2013) Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781. |
| [16] | Tahir,M. et al. (2020) Prediction of N6-methyladenosine sites using convolution neural network model based on distributed feature representations. Neural Netw., 129, 385–391. |
| [17] | Wu,C. et al. (2019) PTPD: predicting therapeutic peptides by deep learning and word2vec. BMC Bioinformatics, 20, 1–8 |
| [18] | Aggarwala,V. and Voight,B.F. (2016) An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. Nat. Genet., 48, 349–355 |
| [19] | Arroyo-Ferna´ndez,I. et al. (2019) Unsupervised sentence representations as word information series: revisiting TF–IDF. Comput. Speech Language, 56,107–129 |
| [20] | Asgari,E. and Mofrad,M.R. (2015) Continuous distributed representation of biological sequences for deep proteomics and genomics. PLoS One, 10,e0141287. |
| [21] | Habibi,M. et al. (2017) Deep learning with word embeddings improves biomedical named entity recognition. Bioinformatics, 33, i37–i48. |
| [22] | Hamid,M.-N. and Friedberg,I. (2019) Identifying antimicrobial peptides using word embedding with deep recurrent neural networks. Bioinformatics, 35, 2009–2016 |
| [23] | Hoque,M.T. et al. (2016) sDFIRE: sequence-specific statistical energy function for protein structure prediction by decoy selections. J Comput. Chem., 37, 1119–1124. |
| [24] | Iqbal,S. et al. (2015) Improved prediction of accessible surface area results in efficient energy function application. J. Theor. Biol., 380, 380–391. |
| [25] | Mishra,A. et al. (2016) Discriminate protein decoys from native by using a scoring function based on ubiquitous Phi and Psi angles computed for all atom. J. Theor. Biol., 398, 112–121. |
| [26] | Babu,M.M. et al. (2011) Intrinsically disordered proteins: regulation and disease. Curr. Opin. Struct. Biol., 21, 432. |
| [27] | Doszta´nyi,Z. et al. (2005) The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. J. Mol. Biol., 347, 827–839. |
| [28] | Xie,R. et al. (2020) DeepVF: a deep learning-based hybrid framework for identifying virulence factors using the stacking strategy. Brief. Bioinf.,2020,bbaa125 |
| [29] | Fischer,T. and Krauss,C., (2018)“ Deep learning with long short-term memory networks for financial market predictions,” Eur. J. Operat. Res, pp. 270,654-669. |
| [30] | Devlin,J.et.al (2018)“Bert: pre-training of deep bidirectional transformers for language understanding.,” arXiv preprint ., p. arXiv:1810.04805. |
| [31] | Hu,Q.et.al.(2015) “ A stacking-based approach to identify translated upstream open reading frames. In Arabidopsis Thaliana, International Symposium on Bioinformatics Research and Applications,,” Bioinformatics Research and Applications, , p. pp. 138–149. |
| [32] | Mishra,A.and Hoque,M.T. (2017)“ Three-dimensional ideal gas reference state based energy function,” Curr.Bioinformatics, pp. 12, 171–180. |
| [33] | Arroyo-Ferna´ndez,I. et al. (2019) Unsupervised sentence representations as word information series: revisiting TF–IDF. Comput. Speech Language, 56,107–129 |
| [34] | Asgari,E. and Mofrad,M.R. (2015) Continuous distributed representation of biological sequences for deep proteomics and genomics. PLoS One, 10,e0141287. |