WASSERSTEIN GENERATIVE ADVERSARIAL NETWORKS FOR BACTERIAL HEMOGLOBIN-LIKE PROTEINS PREDICTION

Soumiya Hamena

Department of Computer Science and its Applications, Faculty of NTIC, Constantine2 University, Abdelhamid Mehri, Constantine, Algeria

ABSTRACT

In the event of decreased oxygen or oxidative and nitrosative stress, bacteria express three distinct structures of hemoglobin proteins which are flavohemoglobins (FlavoHb), truncated hemoglobins (TruncHb) and single domain hemoglobin proteins (SingHb). These proteins were expressed in different heterologous hosts and have been shown to enhance growth and productivity, making them attractive to scientific researchers. At present, only a small number of bacterial hemoglobin-Like proteins have been experimentally annotated. Therefore, it is beneficial to develop a data augmentation method capable of generating high quality of new synthetic sequences. Hence, we propose in this study a model that combines Wasserstein Generative Adversarial Network (WGAN) to generate novel bacteria hemoglobins sequences and Support Vector Machine (SVM) method to predict and classify these proteins. The performance measure comparison of the proposed model with the existing method by the fivefold cross-validation technique has demonstrated the efficiency and the effectiveness of our model. The experiment results were obtained with the evaluation metrics scores of Accuracy (Acc), Precision, Recall, F1_score, Cohen's Kappa (Kappa) and Matthews Correlation Coefficient (Mcc). Further, we have also plotted the learning and Receiver Operating Characteristic (Roc) curves. All experimental results indicate that the proposed model outperforms the existing method.

KEYWORDS

Bacterial Hemoglobin-Like proteins, FlavoHb, TruncHb, SingHb, WGAN, SVM, Prediction, Classification

1. INTRODUCTION

The hemoglobin protein was long thought to be reserved for mammals, but recent results indicate that this proteins is found almost everywhere in mammals, vertebrates, plants and bacteria (1-2). The function of hemoglobins has arguably become the transport and the storage of oxygen only in complex multicellular organisms (3-4). However, in unicellular organisms and invertebrates, the transport and the physiological diffusion of oxygen is less important than protection from the toxic effects of carbon dioxide and nitrogen oxide (5-6). In response to the reduction of oxygen or oxidative and nitrosative stress, bacteria express based on their structure three categories of hemoglobin proteins which are FlavoHb, TruncHb and SingHb (7). Indeed, the first hemoglobin discovered in bacteria was isolated from Vitreoscilla stercoraria (VHb) (8). Further experiments showed that the effects of VHb were not only impacted enhancements in growth, but also improved protein productions (9-11). Thus, this knowledge can be applied to enhance the metabolism of antibiotic-producing strains and a myriad of microorganisms (12-13).At present, only a very small and limited number of bacterial hemoglobins encoding genes have been experimentally validated and several of these proteins that could potentially be used in

biotechnology production process to enhance cell growth and metabolic properties remain unidentified. Therefore, hemoglobin biosynthesis has become an interesting solution to meet the growing demands and overcome the disadvantages of chemical extraction (14). Recently, different microorganisms using metabolic engineering and synthetic biology have synthesized many hemoglobins (15). However, the currents strategies have been applied for a limited amount of hemoglobin, which has led to a slow progression in hemoglobin production.

In this context, Deep Learning, and specially Generative Adversarial Networks (GANs), have great potential to improve and accelerate the discovery of unknown proteins function without the need for a detailed model of the underlying physics or biological pathways(16-22). GANs are a deep learning architecture for training strong generative models, which have the ability to learn and generate new artificial samples that come from an existing distribution of samples (23). GANs are an emerging technique allowing us deep representations without the need for important annotated training data. Indeed, GANs have been widely applied and have enabled significant advancements and excellent performance in many fields such as semantic image editing, style transfer, image synthesis, image super-resolution, classification, cyber security, biomedical informatics and many more (24-30). Recently, GAN has been proposed to protein function prediction by discovering gene ontology term correlations (31). Moreover, a novel method FFPred-GAN has been successfully applied for learning accurately the high dimensional distributions of protein sequence based biophysical features (32). GAN has been also used in protein contact map refinement for improving structure prediction and has achieved an extra gain in contact prediction accuracy (33). A novel data augmentation algorithm ProGAN was developed to improve the prediction of protein solubility (34). In addition, another study was presented to enhance performance in protein classification using synthetic image augmentation with GAN (35). As well, GAN was applied to define correlations between genes, diseases and drugs using biomedical abstracts (36).

Moreover, the Wasserstein generative adversarial network (WGAN) is an extension of GAN, which enhances the stability during model training and provides a loss function correlating with the quality of samples generated (37-38). Indeed, WGAN has received a lot of attention and are currently widely used for many modern machine-learning applications, as it showed good potential in generating artificial data that is close to real world data. WGAN was used for the design of Mont Carlo simulations (39). As well, WGAN was applied for a panchromatic image super-resolution (40). Also, WGAN has been proposed to a data-driven event generator for Hadron Colliders (41).WGAN was moreover employed for data augmentation in fault diagnosis with gradient penalty (42). Nevertheless, only a limited number of projects have been completed based on WGAN for the protein function prediction. De novo protein design for new folds was developed to overcome training difficulties and improve design qualities, this model that uses WGAN discovers an unexplored sequence area to design proteins by analyzing generalizable properties on the basis of available informations on sequence structure (43). As well, a new multi-classification machine-learning pipeline (MultiLyGAN) was developed to predict and analysis multiple protein lysine modified sites (44). In addition, a method has been introduced to generate new synthetic protein sequences from antibiotic resistance genes using WGAN (45).In addition, a novel feedback-loop architecture (FBGAN) was applied to generate new genes coding for antimicrobial peptides and optimize synthetic genes for secondary structure of generated peptides(46). Further, a novel study demonstrates that WGAN strikes a good balance and manage to capture both local and distal patterns of protein tertiary structures (47). Moreover, ProteinSeqGAN is a model that has been developed to generate bio-informed protein sequences for predicting multiclass virus mutation (48).

Given the importance of bacterial hemoglobin proteins in biotechnology process, the current study is carried out to improve the prediction performance of these proteins. In this regard, the

key idea behind this work is to study the effect of using WGAN to achieve better prediction and classification of bacterial hemoglobin-like proteins. To the best of our knowledge, BacHbpred is the first and only framework proposed to address this problem based on support vector machine (SVM) methods by using amino acid composition (AC), dipeptide composition (DC), hybrid method (AC-DC), position specific scoring matrix (PSSM) and max to min amino acid residue profiles (MM)(49).The experimental results obtained by BacHbpred have shown that the hybrid method, which combines AC and DC compositions, has allowed a better prediction, the reason for which we investigate the use of it to describe a protein sequences. In this context, we propose an approach that combines WGAN and SVM methods to predict firstly bacterial hemoglobin from no-bacterial hemoglobin proteins and to determine secondly the class of the sequences predicted as bacterial hemoglobin proteins to be whether FlavoHb, TruncHb or SingHb. In this study, we developed WGAN to synthesize novel sequences of bacterial hemoglobin proteins. Then, we applied SVM method for predicting and classifying the sequences. In fact, we used SVM method as a classifier to analyze and demonstrate the performance of our proposed model by comparing it with BacHbpred.

This manuscript is organized into six sections. The first section is the introduction. The second section explains the different material and methods applied in this study. The third section describes the proposed approach. The fourth section shows the techniques employed to evaluate our approach. The fifth section presents the experimental results and finally a conclusion.

2. MATERIAL AND METHODS

In this section, we describe firstly the dataset used in this study. Then, we present a feature descriptor methods applied to represent protein sequences. Finally, we introduce the different techniques developed in this work, which are WGAN, Multi-Layer Perceptron, and SVM.

2.1. Dataset

The dataset used to demonstrate the power of our proposed approach that consists of protein sequences in the standard format FASTA, was collected from UniProt/Swiss-Prot (50). This dataset contains 139 sequences of bacterial hemoglobin proteins. In addition, 197 sequences of no-bacterial hemoglobin proteins were retrieved randomly. The bacterial hemoglobin proteins contain three different classes distributed as follows: 71 FlavoHb, 36 TruncHb and 32 SingHb.

2.2. Feature Representation Method

Several feature extraction methods can be used to describe and represent a protein from amino acid sequences, for predicting the structural, functional and interaction profiles of proteins to assist in the annotation of proteomic data (51-53). Amino acid and dipeptide composition have been widely employed successfully for this task (54-56). Indeed, a protein sequence composed of N amino acid residues that can be represented by the set {A1, A2, A3,…, An}, where Ai is the amino acid at the i-th position in the sequence of protein. We denote in the following the position of an amino acid by i and j, and its type by r and s. In fact, amino acid composition (AC) is the most popular and simplest method to describe protein sequences, which is defined as the portion of each amino acid type in a protein sequence (57). The fractions of the 20 amino acid are calculated as follows:

$$f(r) = \frac{N_r}{N}, r = 1,2, \dots ,20 \quad (\mathbf{1})$$

The dipeptide composition (DC) is a feature extraction method that considers neighbourhood information or sequence order (58). This later is defined as follows:

$$f(r,s) = \frac{N_{rs}}{N-1}, r, s = 1,2, \ldots,20 \quad (\mathbf{2})$$

Where, *Nrs* define the number of dipeptide described by AC type r and s. thus, the DC represents 400 descriptor values that are calculated for the 20×20 AC combinations.

2.3. Multi-Layer Perceptron (MLP)

MLP is a form of Artificial Neural Networks (ANNs)(59).ANNs are a powerful machine learning techniques that originated from the idea of simulating the human brain(60). These techniques have been applied to solve several complex real-world problems and have been fruitfully applied to the study of neuroscientific questions (61-63). Indeed, ANN is a supervised learning system composed of a set of nodes interconnected. A node can make a decision and transfer it to the other nodes. Each node j receives input values x_i associated with weights w_j that evaluate the importance of the inputs and generates an output that can be transferred to other nodes. Then, the transfer function sums the responses and sends the signal to the activation function. Thus, the output is determined based on the threshold of the activation function. In a simpler way, the transfer function computes the weighted sum of the inputs by adding a threshold b_j. The activation equation is defined as follows:

$$t_j = \sum x_i \, w_{i,j} + b_j \quad (\mathbf{3})$$

Next, an activation function is used to produce the output value y_j. The different types of nodes are distinguished by the nature of their activation functions such as sigmoid, tanh, relu and softmax. Further, MLP consists of an input layer, one or multiple hidden layers and an output layer. Further, MLPs with more than one hidden layer are called Deep Neural Networks (DNNs). The learning is performed by minimizing the loss function by adjusting the parameters *w* and *b* of the model. The most used techniques for parameters learning are stochastic gradient descend (SGD)(64), Root Mean Square Propagation (RMSProp)(65), Adaptive Gradient Algorithm (AdaGrad) (66) and Adam (67).

2.4. WGAN

The WGAN is an additive to the GAN that reinforces model stability of the training and provides a loss function, which is related to the quality of the generated samples. GANs are a strong extension of deep neural networks architectures, which consist of two opposing neural networks, the generator and the discriminator models. The generator creates new samples from the domain, and the discriminator distinguishes the samples to determine if they are real or fake. Indeed, the discriminator model performance is employed to update the weights of the discriminator model itself and the generator model, which means that the generator does not see examples in the domain and is updated according to the discriminator performance. The generator and the discriminator model are denoted G and D, respectively. In fact, the training of GANs includes determining the parameters of D that maximize its accuracy classification and determining the parameters of G that make D as confusing as possible and make it incapable to differentiate between real and fake samples. Thus, the cost of training depends on G and D, which involves finding: $max_D min_G V(G,D)$, where

$$V(G,D) = E_{p_{data(x)}} \log D(x) + E_{p_{g(x)}} \quad (\mathbf{4})$$

E denotes the expectation operator. *pdata(x)* represents the probability density function of data distribution space. *pg(x)* denotes the distribution of the data produced by G. The WGAN architecture replaces the Jensen-Shannon divergence from the original GAN with the Wasserstein distance or also known as the Earth Mover distance. The Wasserstein distance is the minimal cost of moving or converging the model distribution to the real distribution.

$$max_{d\in D} min_G E_{p_{data(x)}}[d(x)] - E_{p_{g(x)}}[d(x)] \quad (\mathbf{5})$$

Moreover, the main idea of the WGAN model is to apply the Wasserstein distance to allow the discriminator to predict a score of how real or fake an input sample. Thus, the task of the discriminator changes by becoming a critic to decide the realness or the fakeness of a sample, where the difference between the scores is as high as possible.

2.5. SVM

The Support Vector Machine (SVM) is a popular supervised machine-learning technique that can be applied for both classification and regression problems (68). This technique has been applied extensively for solving a wide range of challenges in bioinformatics, and in particular for the prediction of protein function (69-74). In the SVM algorithm, samples labelled positives or negatives are projected into high dimensional feature space using a kernel, and the hyper-plane in the feature space is optimized to maximize the margin of positive and negative samples. Given a training dataset of labelled samples pairs $\{x_i, y_i\}$, i=1, 2,…, N, where x_i are the feature vectors or input data and $y_i \in \{+1, -1\}$ are the output label. The classification decision function developed by SVM is denoted as follows:

$$\gamma(x) = sign\left[\sum_{i=1}^{N} \gamma_i \alpha_i . K(x, x_i) + b\right] \quad (6)$$

Where α_i is obtained by solving the quadratic programming problem denoted as follows:

$$max \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j . \gamma_i \gamma_j . K(x_i, x_j) \quad (7)$$

Where $0 \leq \alpha_i \leq C$

$$\sum_{i=1}^{N} \alpha_i \gamma_i = 0, i = 1,2,..,N \quad (8)$$

Where x_i is named support vector, only if the corresponding $\alpha_j > 0$, C is a regularization parameter that adjust the regulation between margin and classification error.

$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2)$, referred to the Radial Basis Functions (RBF) kernel, has a better response to the limits where a most high dimensional data are approximated by Gaussian-like distributions (75-76).

3. PROPOSED MODEL

In order to solve the problem of the lack of annotated bacterial hemoglobin-like proteins, we propose in this work a classifier model that combines WGAN and SVM. The overall workflow of

our proposed approach can be observed in Figure1.Indeed, we considered mainly two tasks: the first one consists of generating new bacterial hemoglobin sequences using WGAN, to enhance the training data. The second one consists of predicting the bacterial hemoglobin sequences, then determining the class of the sequences predicted to be whether FlavoHb, TruncHb or SingHb. As can be seen, a data collection process is performed firstly for gathering sequences from UniProt/Swiss-Prot database. Secondly, a feature extraction stage follows to generate the descriptors applied in this study, which are the compositions AC and DC. In fact, the set of the proposed features was derived by applying the web server ProtrWeb, which is a very comprehensive, flexible and integrated toolkit for protein sequences extracted structural and physiochemical descriptor computation (77). Overall, 420 dimensional features were obtained by combining AC and DC descriptors. Then, WGAN is applied to enhance the data of the original training sequences by generating a large number of simulated samples. The fact that WGAN allows high quality results to be obtained without adjustment hyperparameters, make it a promising solution to many problems. Further, the sequences mixed by the original sequences and the synthetic sequences, are used to train the classifier based on SVM method.

The critic model is implemented as a fully connected MLP where it takes as input the set of features extracted and outputs a score for the realness or fakeness of the protein sequence. The rectified Linear Unit (RELU) activation function is applied to activate the hidden layer, which is defined as follows (78):

$$f_{RELU}(x) = \max(0, x) \quad (9)$$

Then, we employed the batch normalization technique for improving the speed, performance and stability of the model by normalizing the input layer, where the activations are adjusting and scaling (79). Next, the weighted outputs of the hidden layer are applied as input into the hyperbolic tangent (tanh) activation function in the output layer (80).

$$f_{tanh}(x) = \frac{2}{1 + e^{-2x}} - 1 \quad (10)$$

In addition, the critic model uses the weight constraint to reduce model weights after each mini-batch updates between [-0.01, 0.01]. Finally, the critic model is optimized using the Wasserstein distance function, and the RMSProp optimizer with a learning rate of 0.00005.

Moreover, the generator model is implemented also as a fully connected MLP network. It takes as input a point in latent space and outputs the set of the proposed features. RELU activation function is also is used to activate the hidden units and the batch normalization is applied. The output layer is worked using the tanh function. Then, a WGAN model is achieved by combining both the generator and the critic model into one larger model, which takes as input a point in the latent space, specifically Gaussian distributed random variables which is described as follows(81):

$$p(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \quad (11)$$

Then, it generates a new protein sequence using the generator that is fed then as input to the critic model, which next scoring it as real or fake. The model is optimized using RMSProp optimizer with the Wasserstein loss function. Furthermore, in the WGAN model, the critic model is updated more than the generator model. Thus, a hyperparameter named n_critic, which is used to control the number of times that the critic model is updated for each update of the generator model, is set to five.

Finally, SVM method is developed as a classifier to predict if the given protein sequence is bacterialHb or no-bacterialHb, where the input is composed of both the original sequences and the generated sequences by the proposed mode. Then, if the protein sequence is predicted as bacterialHb, it will put in another model, which will classify it into one of the three sub-classes: FlavoHb, TruncHb or SingHb.In fact, SVM methods are configured using the same hyperparameters of BacHbpred, which are:$\gamma = 0.1, C = 375$, for predicting bacterialHb. $\gamma = 1, C = 150$, for both FlavoHb and TruncHb classification. $\gamma = 0.1, C = 350$, for SingHb classification.

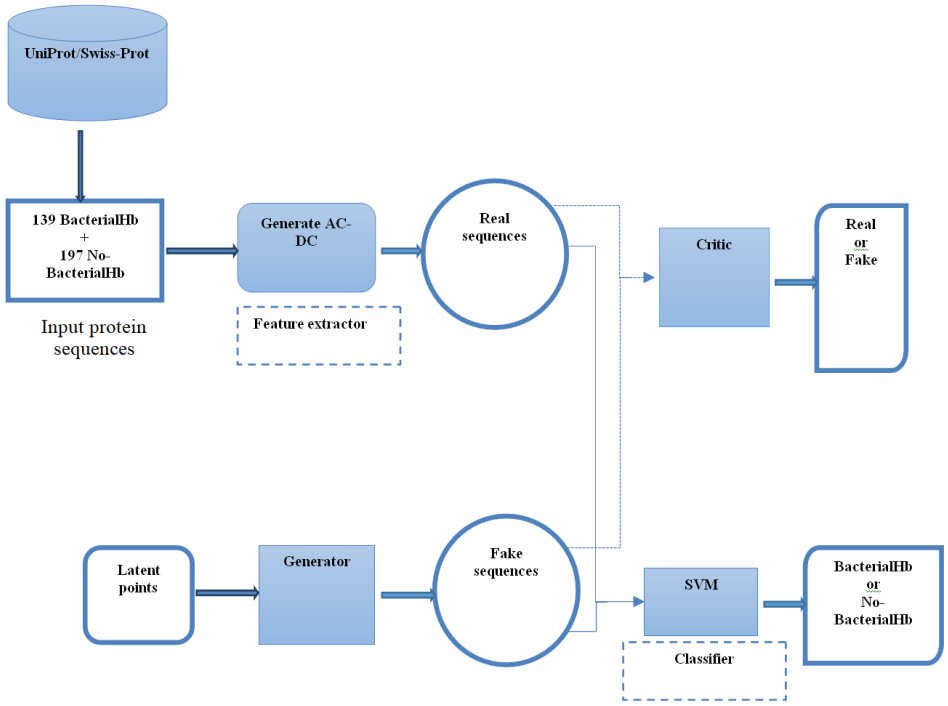

Figure 1.Workflow of the proposed approach

4. MODEL EVALUATION

To train and test our proposed model, we have divided our data into 80% for training using five-fold cross validation technique and 20% for testing. Moreover, in order to evaluate the prediction and the classification performance we have employed the score measurements of Accuracy (Acc), Precision, Recall, F1_score, Cohen Kappa (Kappa)and Matthews Correlation Coefficient (Mcc)(82), which their formulas are described as follows:

$$Acc = \frac{TP + TN}{TP + FN + FP + TN} \quad (12)$$

Where, TP is the number of real data of the positive class, which the model predicts correctly. TN is the number of real data of the negative class, which the model predicts correctly. FN is the number of real data of the negative class, which the model incorrectly predicts and FP is the number of real data of the positive class, which the model incorrectly predicts.

$$Precision = \frac{TP}{TP + FP} \quad (13)$$

$$Recall\ (Sensitivity) = \frac{TP}{TP + FN} \quad (14)$$

$$Specificity = \frac{TN}{TN + FP} \quad (15)$$

$$F1_{score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (16)$$

$$Kappa = \frac{Acc - P_e}{1 - P_e} \quad (17)$$

Where P_e presents the random accuracy or the hypothetical probability of chance agreement to estimate the probabilities of randomly choose each class, which is defined as (83):

$$P_e = \frac{((FP + TN) \times (TN + FN) \times (FN + TP) \times (FP + TP))}{(TN + TP + FN + FP)^2} \quad (18)$$

$$Mcc = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FN) \times (TP + FP) \times (TN + FP) \times (TN + FN)}} \quad (19)$$

5. RESULTS AND DISCUSSION

The proposed models were implemented in Python where the sequential model of Keras package with Theano backend was used to develop the WGAN model. In addition, we have selected a subset of 100 sequences that belong to the bacterial hemoglobin proteins class. Then, the model was fit for 10 training epochs and we have used a batch size of 20 sequences for all experiments. Thus, each training epoch includes 139/20, or about 6 batches of real and fake sequences, which are updated to the model. Therefore, the model was trained for 10 epochs of 6 batches, or 60 iterations. Indeed, we have created the line plots of the loss for real and fake sequences, as the loss for the generator for each model update. See Figure2 for details. As we can see, the loss value correlates with the generated sequences quality after a few epochs training.

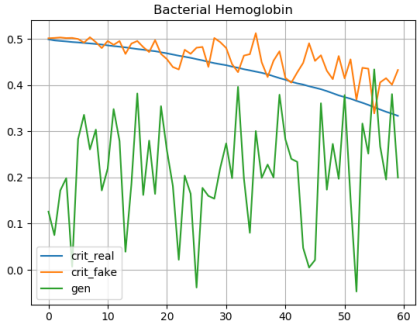

Figure 2. Line plots of loss for WGAN to BacterialHb prediction

Furthermore, we have applied the SVM classifier that is training using the mixed sequences or the original and the generated sequences (84). In addition, to achieve a comparative study, we have applied the SVM method on the original used dataset. The detailed results, which include the scores of Acc, Precision, Recall, F1_score, Kappa and Mcc for each fold, are listed in Table1.

Table 1. . Performance comparison of Bacterial Hemoglobin prediction.

		Acc	Precision	Recall	F1_score	Kappa	Mcc
SVM	**Fold1**	0.8971	1.0000	0.8000	0.8888	0.7952	0.8124
	Fold2	0.9265	1.0000	0.8148	0.8980	0.8414	0.8522
	Fold3	0.8971	0.9474	0.7500	0.8372	0.7634	0.7746
	Fold4	0.9118	1.0000	0.7931	0.8846	0.8147	0.8291
	Fold5	0.8971	1.0000	0.8000	0.8888	0.7952	0.8124
	Mean	0.9059	0.9895	0.7916	0.8795	0.8020	0.8161
WGAN-SVM	**Fold1**	0.9318	1.0000	0.8888	0.9412	0.8608	0.8692
	Fold2	0.9432	1.0000	0.8810	0.9367	0.8855	0.8914
	Fold3	0.8977	1.0000	0.8269	0.9053	0.7963	0.8134
	Fold4	0.9205	1.0000	0.8654	0.9278	0.8402	0.8512
	Fold5	0.9432	1.0000	0.9000	0.9474	0.8860	0.8918
	Mean	0.9273	1.0000	0.8724	0.9317	0.8538	0.8634

As shown in Table1, our proposed model WGAN-SVM achieved the best performance, where the Acc increases from 0.9059 to 0.9273. The Precision increases from 0.9895 to 1.0000. The Recall increases from 0.7916 to 0.8724.The F1_score increases from 0.8795 to 0.9317. The Kappa score increases from 0.8020 to 0.8538 and the Mcc increases from 0.8161 to 0.8634, which demonstrates the power and the effectiveness of the proposed WGAN model for synthesizing new bacterial hemoglobin proteins to improve the prediction performance process.

Furthermore, we have plotted the learning curves of both the proposed WGAN-SVM and SVM method. As shown in Figure3, in the first row, the learning curves of our proposed WGAN-SVM outperforms the SVM method for both the training and the cross-validation scores. We note that the learning curves of SVM method decrease at the end. However, the training curve of our proposed WGAN-SVM still around the maximum and the cross-validation score could be increased with more training sequences. In the second row, the plots illustrate the time required by the models for training process with various sizes of training sequences. Finally, in the third row, the plots display the time taken to train the models for each training sizes.

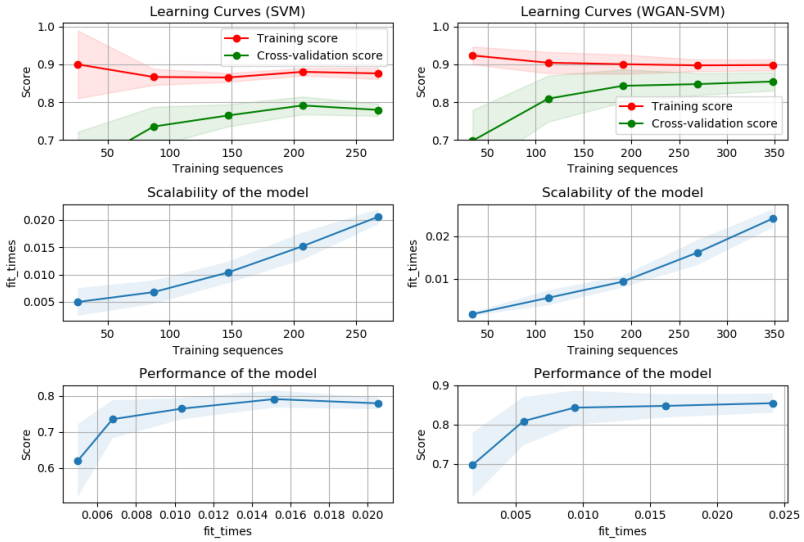

Figure 3.Plot of Learning curves performance comparison of bacterial hemoglobin prediction

In addition, we have plotted the area under the ROC (Receiver Operating Characteristic) or ROC AUC curves for both models. ROC curves summarize the trade-off between the true positive rate

(Sensitivity) and the false positive rate (1-Specificity) for the predictive model. As illustrated in Figure 4, our proposed WGAN-SVM model outperforms SVM method, where have obtained a ROC AUC of 0.9318.

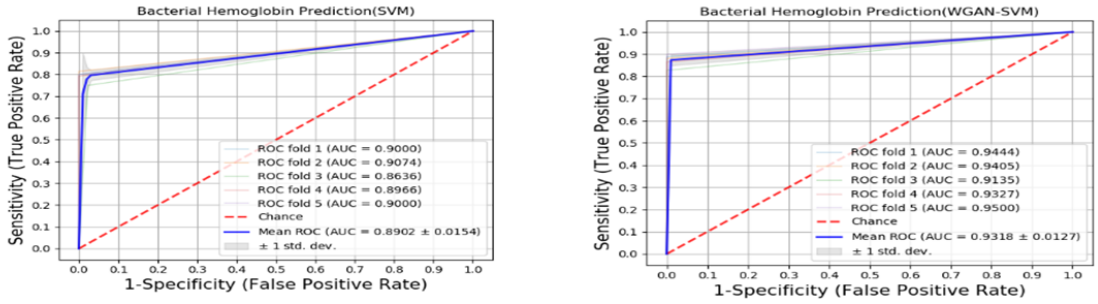

Figure 4. ROC curves performance comparison of bacterial hemoglobin prediction

Next, three additional WGAN-SVM models were developed to classify the bacterial hemoglobin proteins in each of the three subclasses, which are FlavoHb, TruncHb and SingHb. For the FlavoHb class, we have selected randomly 50 sequences where the model was fit for 10 training. For the TruncHb class, we have selected randomly 20 sequences and the model was fit for 75 training epochs. We have randomly also selected 15 SingHb sequences, where the model was fit for 100 epochs. The line plots of the loss for real and fake FlavoHb, TruncHb and SingHb classes, as the loss for the generator for each model update, are illustrated in Figure 5, respectively.

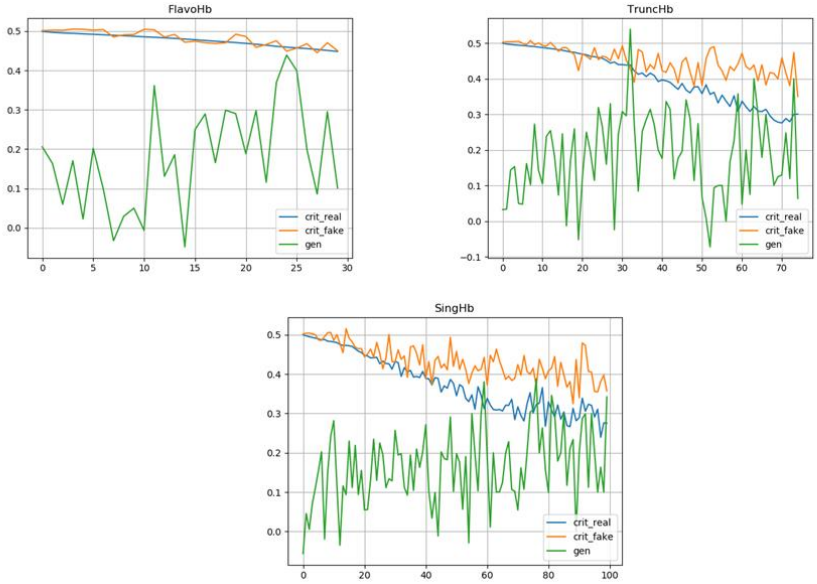

Figure 5. Line plots of loss for WGAN of FlavoHb, TruncHb and SingHb classification, respectively

Moreover, we have also achieved a comparative study for the classification tasks of FlavoHb, TruncHb and SingHb, respectively. The results obtained are listed in Table2. As we can see, our proposed WGAN-SVM outperforms the SVM method, where it achieved the Acc of 0.9897, the Precision of 1.0000, the Recall of 0.9695, the F1_score of 0.9844, the Kappa of 0.9766 and the Mcc of 0.9769 for FlavoHb classification. For TruncHb classification, our proposed model

achieved the Acc of 0.9611, the Precision of 1.0000, the Recall of 0.7239, the F1_score of 0.8238, the Kappa of 0.8041 and the Mcc of 0.8257. Finally, WGAN-SVM achieved the Acc of 0.9465, the Precision of 0.6694, the Recall of 0.9206, the F1_score of 0.7619, the Kappa of 0.7334 and the Mcc of 0.7527 for SingHb classification.

Table 2. Performance comparison of bacterial hemoglobin classification.

			Acc	Precision	Recall	F1_score	Kappa	Mcc
FlavoHb	**SVM**	**Fold1**	0.9853	1.0000	0.9500	0.9744	0.9641	0.9647
		Fold2	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
		Fold3	0.9853	1.0000	0.9333	0.9655	0.9562	0.9571
		Fold4	0.9853	1.0000	0.9167	0.9565	0.9477	0.9490
		Fold5	0.9706	1.0000	0.9091	0.9524	0.9312	0.9333
		Mean	0.9853	1.0000	0.9418	0.9698	0.9598	0.9608
	WGAN-SVM	**Fold1**	0.9872	1.0000	0.9565	0.9778	0.9688	0.9692
		Fold2	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
		Fold3	0.9872	1.0000	0.9714	0.9855	0.9740	0.9743
		Fold4	0.9872	1.0000	0.9630	0.9811	0.9714	0.9718
		Fold5	0.9872	1.0000	0.9565	0.9778	0.9688	0.9692
		Mean	0.9897	1.0000	0.9695	0.9844	0.9766	0.9769
TruncHb	**SVM**	**Fold1**	0.9118	1.0000	0.3333	0.5000	0.4646	0.5501
		Fold2	0.9706	1.0000	0.6667	0.8000	0.7848	0.8036
		Fold3	0.9706	1.0000	0.6000	0.7500	0.7354	0.7626
		Fold4	0.9265	1.0000	0.5000	0.6667	0.6304	0.6785
		Fold5	0.9559	1.0000	0.5714	0.7273	0.7052	0.7380
		Mean	0.9471	1.0000	0.5343	0.6888	0.6641	0.7065
	WGAN-SVM	**Fold1**	0.9167	1.0000	0.4000	0.5714	0.5345	0.6039
		Fold2	0.9861	1.0000	0.9333	0.9655	0.9568	0.9577
		Fold3	0.9861	1.0000	0.9000	0.9474	0.9394	0.9411
		Fold4	0.9722	1.0000	0.7500	0.8571	0.8421	0.8528
		Fold5	0.9444	1.0000	0.6364	0.7778	0.7478	0.7728
		Mean	0.9611	1.0000	0.7239	0.8238	0.8041	0.8257
SingHb	**SVM**	**Fold1**	0.9118	0.5000	1.0000	0.6667	0.6222	0.6720
		Fold2	0.9559	0.7143	0.8333	0.7692	0.7450	0.7477
		Fold3	0.9412	0.5000	1.0000	0.6667	0.6383	0.6847
		Fold4	0.9265	0.5833	1.0000	0.7368	0.6975	0.7318
		Fold5	0.9412	0.6000	1.0000	0.7500	0.7190	0.7492
		Mean	0.9353	0.5795	0.9667	0.7179	0.6844	0.7171
	WGAN-SVM	**Fold1**	0.9296	0.6667	0.8889	07619	0.7216	0.7319
		Fold2	0.9296	0.5000	1.0000	0.6667	0.6321	0.6798
		Fold3	1.000	1.0000	1.0000	1.0000	1.0000	1.0000
		Fold4	0.9296	0.6250	0.7143	0.6667	0.6275	0.6292
		Fold5	0.9437	0.5556	1.0000	0.7143	0.6858	0.7224
		Mean	0.9465	0.6694	0.9206	0.7619	0.7334	0.7527

Next, we have plotted the learning curves of each classification models. As we can see in Figure 6 that presents the learning curves of the FlavoHb classification, the cross-validation score of our proposed model WGAN-SVM could be increased with more training sequences. However, we can see clearly that the validation score of the SVM method decreased at the end. The training score of both models is high at the beginning and still around the maximum.

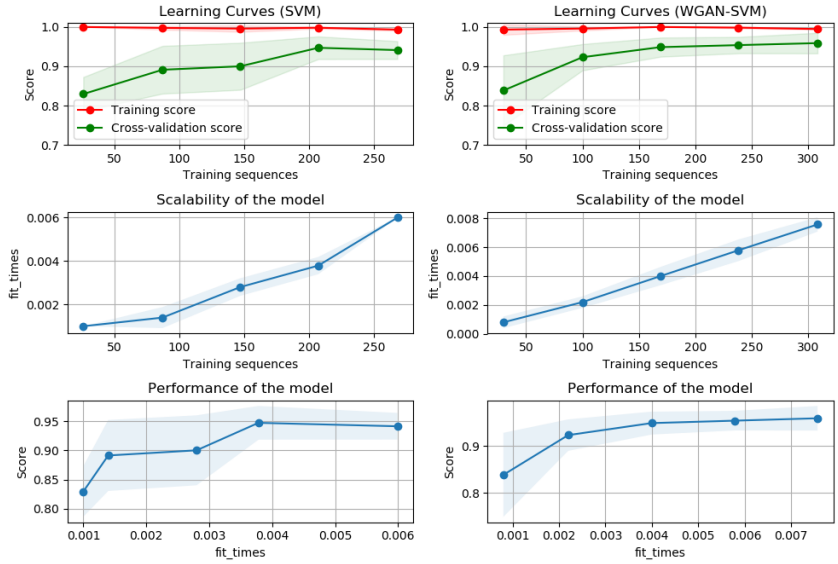

Figure 6.plot of Learning curves performance comparison of FlavoHb classification

Moreover, the ROC curves demonstrate also that our proposed model achieved the best mean with ROC AUC of 0.9798.

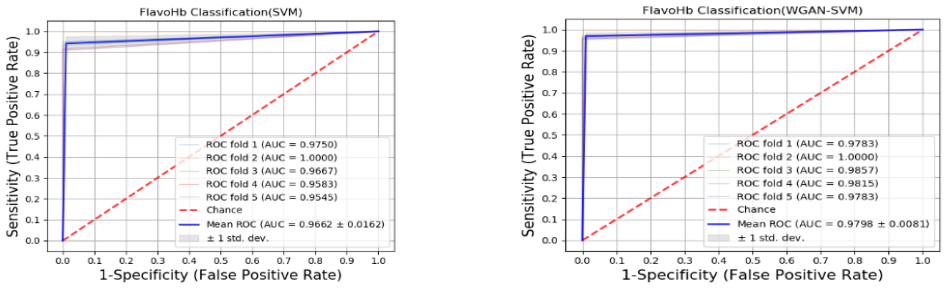

Figure 7.ROC curves performance comparison of FlavoHb classification.

Figure 8 illustrates the learning curves of TruncHb classification performance comparison. As we can see, the cross-validation score of our proposed model outperforms the SVM method at the beginning and increases at the end.

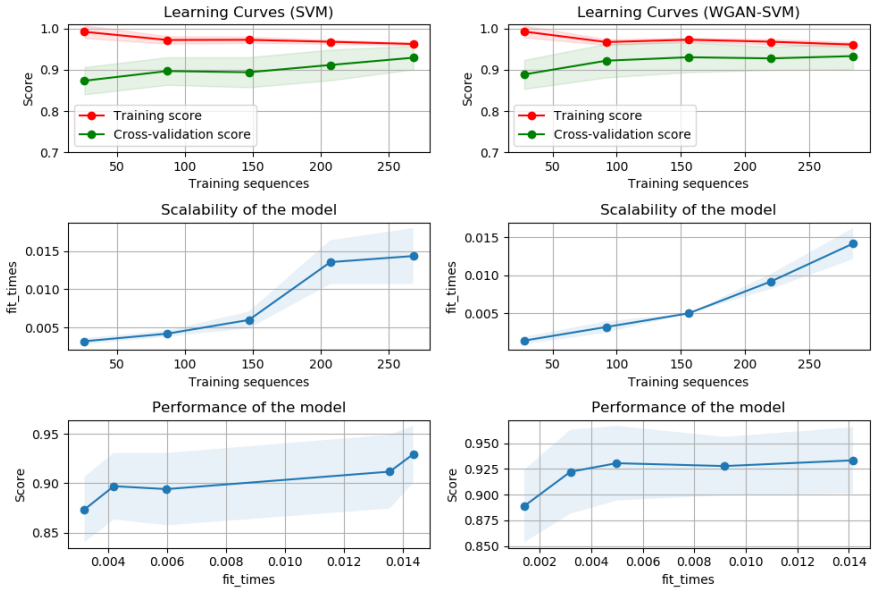

Figure 8. plot of Learning curves performance comparison of TruncHb classification

Figure 9 shows the performance comparison of TruncHb classification where our proposed model achieved the best results with mean ROC AUC of 0.8583.

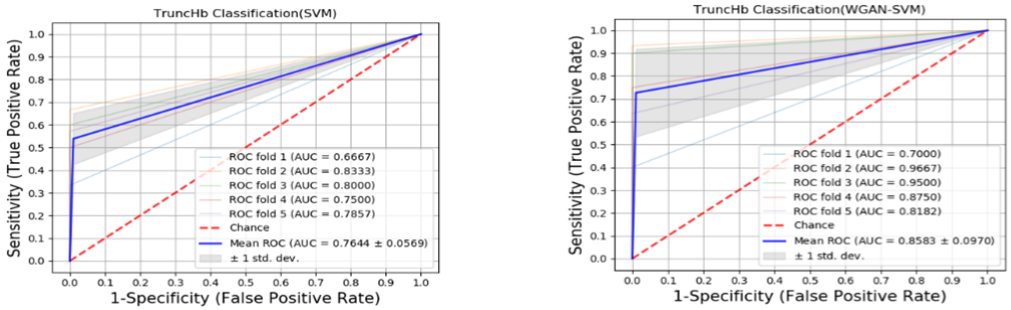

Figure 9. ROC curves performance comparison of TruncHb classification

In addition, Figure 10 illustrates the learning curves of performance comparison of SingHb classification. We can see clearly that our proposed model achieve the best validation score at the beginning and the end of training sequences.

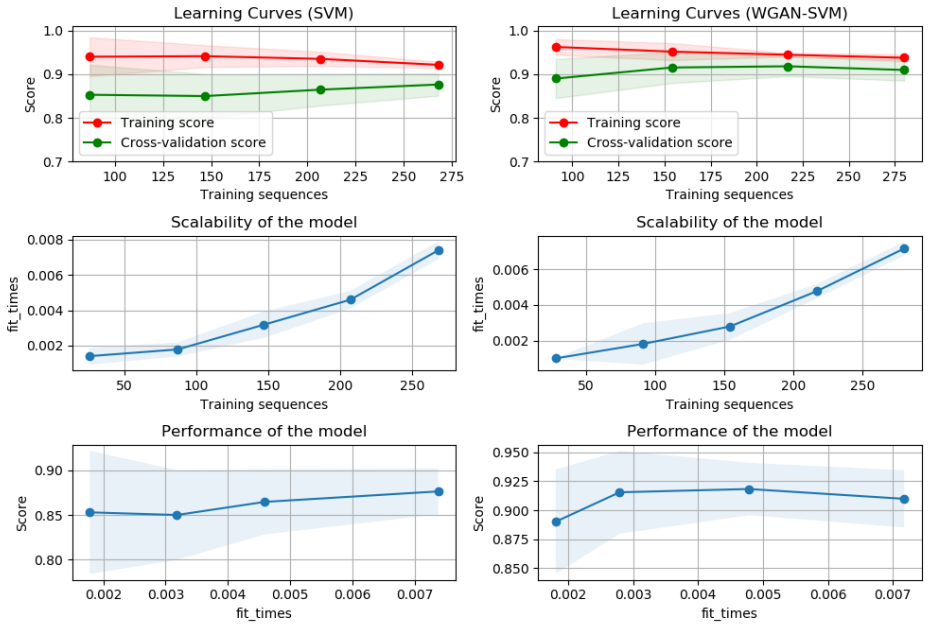

Figure 10. Plot of Learning curves performance comparison of SingHb classification

In addition, we can see in Figure 11 that our proposed model WGAN-SVM has the highest results where we have obtained a mean ROC AUC of 0.7912.

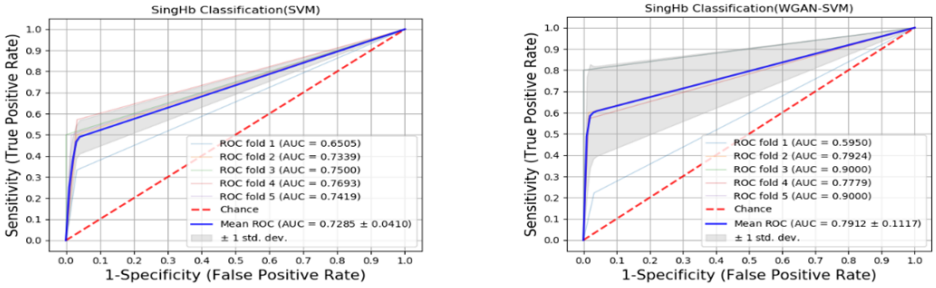

Figure 11.ROC curves performance comparison of SingHb classification

6. CONCLUSION

This paper proposes a model for bacterial hemoglobin proteins prediction and classification. This model combines Wasserstein Generative Adversarial Network (WGAN) and Support Vector Machine (SVM). In this study, we have shown the potential benefit of applying WGAN to generate high quality synthetic protein sequences for enhancing the capacity of a classifier on limited number of training samples. With the power of WGAN as an augmentation data method, there are vast opportunities for new applications to not only solve the problem of limited training data, but also to address the effects of balancing on the dataset.

REFERENCES

[1] HEMOGLOBINS FROM BACTERIA TO MAN: EVOLUTION OF DIFFERENT PATTERNS OF GENE EXPRESSION. HARDISON, ROSS. 1998, The Journal of Experimental Biology, Vol. 201, pp. 1099–1117.

[2] Thermoglobin, Oxygen-avid Hemoglobin in a Bacterial Hyperthermophile. JJ L. Miranda, David H. Maillett, Jayashree Soman. 44, 2005, THE JOURNAL OF BIOLOGICAL CHEMISTRY, Vol. 280, pp. 36754 –36761.

[3] The multiple functions of hemoglobin. B Giardina, I Messana, R Scatena, M Castagnola. 3, 1995, Crit Rev Biochem Mol Biol, Vol. 30, pp. 165-96.

[4] Oxygen transport by fetal bovine hemoglobin. M E Clementi, R Scatena, A Mordente, S G Condò, M Castagnola, B Giardina. 1, J Mol Biol, Vol. 255, pp. 229-34.

[5] Nonvertebrate Hemoglobins:. VINOGRADOV, ROY E. WEBER AND SERGE N. 2, 2001, PHYSIOLOGICAL REVIEWS, Vol. 81, pp. 569-628.

[6] Plant hemoglobins: a journey from unicellular green algae to vascular plants. Manuel Becana, Inmaculada Yruela Gautam Sarath Pilar Catalán Mark S. Hargrove. 2020, New Phytologist, Vol. 227, pp. 1618-1635.

[7] Bacterial hemoglobins and flavohemoglobins: versatile proteins and their impact on microbiology and biotechnology. Alexander D Frey, Pauli T Kallio. 4, 2003, FEMS Microbiol Rev, Vol. 27, pp. 525-45.

[8] Unusual structure of the oxygen-binding site in the dimeric bacterial hemoglobin from Vitreoscilla sp. Cataldo Tarricone, Alessandro Galizzi, Alessandro Coda, Paolo Ascenzi, Martino Bolognesi.4, 1997, Structure, Vol. 5, pp. 497-507.

[9] Expression of Vitreoscilla hemoglobin enhances production of arachidonic acid and lipids in Mortierella alpina. Huidan Zhang, Yingang Feng, Qiu Cui, Xiaojin Song. 68, 1017, BMC Biotechnology, Vol. 17.

[10] Enhancement of glucaric acid production in Saccharomyces cerevisiae by expressing Vitreoscilla hemoglobin. Xi Zhang, Chi Xu, YingLi Liu, Jing Wang,YunYing Zhao. 2020, Biotechnol Lett, Vol. 42, pp. 2169–2178.

[11] Expression of Intracellular Hemoglobin Improves Protein Synthesis in Oxygen-Limited Escherichia coli. Chaitan Khosla, Joseph E. Curtis, John DeModena, Ursula Rinas, James E. Bailey. 1990, Bio/Technology, Vol. 8, pp. 849–853.

[12] Genetic engineering of an industrial strain of Saccharopolyspora erythraea for stable expression of the Vitreoscilla haemoglobin gene (vhb). Peter Brünker, Wolfgang Minas, Pauli T Kallio, James E Baile. 9, 1998, Microbiology (Reading), Vol. 144, pp. 2441-2448.

[13] Novel hemoglobins to enhance microaerobic growth and substrate utilization in Escherichia coli. C J Bollinger, J E Bailey, P T Kallio. 5, 2001, Biotechnol Prog, Vol. 17, pp. 798-808.

[14] Heterologous overexpression of bacterial hemoglobin VHb improves erythritol biosynthesis by yeast Yarrowia lipolytica. Aleksandra M. Mirończuk, Katarzyna E. Kosiorowska, Anna Biegalska, Magdalena Rakicka-Pustułka, Mateusz Szczepańczyk,Adam Dobrowolski. 176, 2019, Microbial Cell Factories, Vol. 18.

[15] Recent Advances in the Microbial Synthesis of Hemoglobin. Xinrui Zhao, Jingwen Zhou,Guocheng Du,Jian Chen. 2020, Trends in Biotechnology.

[16] iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data. Zhen Chen, Pei Zhao, Fuyi Li, Tatiana T Marquez-Lago, André Leier, Jerico Revote, Yan Zhu, David R Powell, Tatsuya Akutsu, Geoffrey I Webb, Kuo-Chen Chou, A Ian Smith, Roger J Daly, Jian Li, Jiangning Song. 3, 2020, Briefings in Bioinformatics, Vol. 21, pp. 1047–1057.

[17] Machine Learning for Biologics: Opportunities for Protein Engineering, Developability, and Formulation. Harini Narayanan, Fabian Dingfelder,Alessandro Butté,Nikolai Lorenzen,Michael Sokolov,Paolo Arosio. 3, 2021, Trends in Pharmacological Sciences, Vol. 42, pp. 151-165.

[18] Artificial intelligence method to design and fold alpha-helical structural proteins from the primary amino acid sequence. Zhao Qin, Lingfei Wu,Hui Sun,Siyu Huo, Tengfei Ma,Eugene Lim,Pin-Yu Chen,Benedetto Marelli,Markus J. Buehler. 2020, Extreme Mechanics Letters, Vol. 36.

[19] Machine-learning-guided directed evolution for protein engineering. Kevin K. Yang, Zachary Wu, Frances H. Arnold. 2019, Nature Methods , Vol. 16, pp. 687–694.

[20] Expanding functional protein sequence spaces using generative adversarial networks. Donatas Repecka, Vykintas Jauniskis, Laurynas Karpus, Elzbieta Rembeza, Irmantas Rokaitis, Jan Zrimec, Simona Poviloniene, Audrius Laurynenas, Sandra Viknander, Wissam Abuajwa, Otto Savolainen, Rolandas Meskys, Martin K. M. Engqvist, Aleksej Zelezniak. 2021, Nature Machine Intelligence.

[21] Namrata Anand, Possu Huang. Generative modeling for protein structures. NEURIPS2018. 2018, Vol. 31.

[22] Deep Dive into Machine Learning Models for Protein Engineering. Yuting Xu, Deeptak Verma,Robert P. Sheridan,Andy Liaw,Junshui Ma,Nicholas M. Marshall, John McIntoshJohn McIntosh,Edward C. Sherer, Vladimir Svetnik,Jennifer M. Johnston. 6, 2020, J. Chem. Inf. Model, Vol. 60, pp. 2773–2790.

[23] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio. Generative Adversarial Nets. Advances in Neural Information Processing Systems 27. 2014.

[24] Applications of Generative Adversarial Networks (GANs): An Updated Review. Hamed Alqahtani, Manolya Kavakli-Thorne, Gulshan Kumar. 2021, Archives of Computational Methods in Engineering , Vol. 28, pp. 525–552.

[25] Accelerating Science with Generative Adversarial Networks: An Application to 3D Particle Showers in Multilayer Calorimeters. Michela Paganini, Luke de Oliveira, Benjamin Nachman. 4, 2018, Phys. Rev. Lett., Vol. 120.

[26] Ming-Yu Liu, Xun Huang,Jiahui Yu,Ting-Chun Wang,Arun Mallya. Generative Adversarial Networks for Image and Video Synthesis: Algorithms and Applications. in Proceedings of the IEEE.

[27] A review of generative adversarial networks and its application in cybersecurity. Chika Yinka-Banjo, Ogban-Asuquo Ugot. 2020, Artificial Intelligence Review, Vol. 53, pp. 1721–1736.

[28] Generative Adversarial Networks and Its Applications in Biomedical Informatics . Lan Lan, You Lei, Zhang Zeyang, Fan Zhiwei, Zhao Weiling, Zeng Nianyin, Chen Yidong, Zhou Xiaobo. 2020, Frontiers in Public Health , Vol. 8.

[29] Creating Artificial Images for Radiology Applications Using Generative Adversarial Networks (GANs) – A Systematic Review. Vera Sorin, Yiftach Barash, Eli Konen, Eyal Klang. 8, 2020, Academic Radiology, Vol. 27, pp. 1175-1185.

[30] Medical Image Generation Using Generative Adversarial Networks: A Review. Nripendra Kumar, SinghKhalid Raza. 2021, Studies in Computational Intelligence, Vol. 932, pp. 77-96.

[31] PFP-WGAN: Protein function prediction by discovering Gene Ontology term correlations with generative adversarial networks. Seyyede Fatemeh Seyyedsalehi, Mahdieh Soleymani,Hamid R. Rabiee,Mohammad R. K. Mofrad. 2, 2021, PLOS ONE, Vol. 16.

[32] Protein function prediction is improved by creating synthetic feature samples with generative adversarial networks. Cen Wan, David T. Jones. 2020, Nature Machine Intelligence, Vol. 2, pp. 540–550.

[33] Protein Contact Map Refinement for Improving Structure Prediction Using Generative Adversarial Networks. Sai Raghavendra Maddhuri Venkata Subramaniya, Genki Terashi, Aashish Jain, Yuki Kagaya, Daisuke Kihara. Bioinformatics.

[34] ProGAN: Protein solubility generative adversarial nets for data augmentation in DNN framework. Xi Han, Liheng Zhang, Kang Zhou, Xiaonan Wang. 2019, Computers & Chemical Engineering, Vol. 131.

[35] Synthetic image augmentation with generative adversarial network for enhanced performance in protein classification. Rohit Verma, Raj Mehrotra,Chinmay Rane, Ritu Tiwari, Arun Kumar Agariya. 2020, Biomedical Engineering Letters, Vol. 10, pp. pages443–452.

[36] Vijaya, S. Generative Adversarial Networks in Disease Gene Drug Relationships. IOP Conf. Ser.: Mater. Sci. Eng. 2021, Vol. 1055.

[37] Wasserstein Generative Adversarial Networks. Martin Arjovsky, Soumith Chintala, Léon Bottou. 2017. Proceedings of the 34th International Conference on Machine Learning. Vol. 70, pp. 214-223.

[38] Improved Training of Wasserstein GANs. Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, Aaron C. Courville. 2017. Advances in Neural Information Processing Systems 30.

[39] Using Wasserstein Generative Adversarial Networks for the design of Monte Carlo simulations. Susan Athey, Guido W.Imbens,Jonas Metzger,Evan Munro. 2021, Journal of Econometrics.

[40] Panchromatic Image Super-Resolution Via Self Attention-Augmented Wasserstein Generative Adversarial Network. Juan Du, Kuanhong Cheng ,Yue Yu,Dabao Wang,Huixin Zhou. 6, 2021, Sensors, Vol. 21.

[41] A data-driven event generator for Hadron Colliders using Wasserstein Generative Adversarial Network. Suyong Choi, Jae Hoon Lim. 2021, Journal of the Korean Physical Society, Vol. 78, pp. 482–489.
[42] Data augmentation in fault diagnosis based on the Wasserstein generative adversarial network with gradient penalty. Xin Gao, Fang Deng,Xianghu Yue. 2020, Neurocomputing, Vol. 396, pp. 487-494.
[43] De Novo Protein Design for Novel Folds Using Guided Conditional Wasserstein Generative Adversarial Networks. Mostafa Karimi, Shaowen Zhu, Yue Cao, and Yang Shen. 12, 2020, J. Chem. Inf. Model, Vol. 60, pp. 5667–5681.
[44] Prediction and analysis of multiple protein lysine modified sites based on conditional wasserstein generative adversarial networks. Yingxi Yang, Hui Wang, Wen Li, Xiaobo Wang, Shizhao Wei, Yulong Liu, Yan Xu. 171, 2021, BMC Bioinformatics, Vol. 22.
[45] Generating protein sequences from antibiotic resistance genes data using Generative Adversarial Networks. Prabal Chhibbar, Arpit Joshi. 2019, ArXiv, Vol. abs/1904.13240.
[46] Feedback GAN for DNA optimizes protein functions. Anvita Gupta, James Zou. 2019, Nature Machine Intelligence, Vol. 1, pp. 105–111.
[47] Generative Adversarial Learning of Protein Tertiary Structures. Taseef Rahman, Yuanqi Du,Liang Zhao,Amarda Shehu. 5, 2021, Molecules, Vol. 26.
[48] Bio-informed Protein Sequence Generation for Multi-class Virus Mutation Prediction. Yuyang Wang, Prakarsh Yadav, Rishikesh Magar,Amir Barati Farimani. 2020, bioRxiv .
[49] BacHbpred: Support Vector Machine Methods for the Prediction of Bacterial Hemoglobin-Like Proteins. MuthuKrishnan Selvaraj, Munish Puri, Kanak L. Dikshit, Christophe Lefevre. 2016, Adv. Bioinformatics.
[50] UniProt: the universal protein knowledgebase in 2021. Consortium, The UniProt. 1, 2021, Nucleic Acids Research, Vol. 49, pp. D480–D489.
[51] PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. Z. R. Li, H. H. Lin, L. Y. Han, L. Jiang, X. Chen, Y. Z. Chen. 2, 2006, Nucleic Acids Research, Vol. 34, pp. W32–W37.
[52] iFeature: a Python package and web server for features extraction and selection from protein and peptide sequences. Zhen Chen, Pei Zhao, Fuyi Li, André Leier, Tatiana T Marquez-Lago, Yanan Wang, Geoffrey I Webb, A Ian Smith, Roger J Daly, Kuo-Chen Chou, Jiangning Song. 14, 2018, Bioinformatics, Vol. 34, pp. 2499–2502.
[53] DescribePROT: database of amino acid-level protein structure and function predictions. Bi Zhao, Akila Katuwawala, Christopher J Oldfield, A Keith Dunker, Eshel Faraggi, Jörg Gsponer, Andrzej Kloczkowski, Nawar Malhis, Milot Mirdita, Zoran Obradovic, Johannes Söding, Martin Steinegger, Yaoqi Zhou, Lukasz Kurgan. 1, 2021, Nucleic Acids Research, Vol. 49, pp. D298–D308.
[54] Systematic Comparison and Comprehensive Evaluation of 80 Amino Acid Descriptors in Peptide QSAR Modeling. Peng Zhou, Qian Liu, Ting Wu, Qingqing Miao, Shuyong Shang, Heyi Wang, Zheng Chen, Shaozhou Wang, Heyan Wang. 4, 2021, Journal of Chemical Information and Modeling, Vol. 61, pp. 1718-1731.
[55] Implementation protein sequence segmentation in AAC and DC as protein descriptors for improving a classification performance of acetylation prediction. A Rizqiana, M R Faisal,F R Lumbanraja. 2020. Journal of Physics: Conference Series, The 3rd International Conference on Applied Sciences Mathematics and Informatics. Vol. 1751.
[56] iAMY-SCM: Improved prediction and analysis of amyloid proteins using a scoring card method with propensity scores of dipeptides. Phasit Charoenkwan, Sakawrat Kanthawong, Chanin Nantasenamat, Md. Mehedi Hasan, Watshara Shoombuatong. 1, 2021, Genomics, Vol. 113, pp. 689-698.
[57] A novel approach to the recognition of protein architecture from sequence using Fourier analysis and neural networks. Shepherd, A.J., Gorse, D., Thornton. 2003, Proteins , Vols. 290 –302, pp. 290 –302.
[58] Classification of nuclear receptors based on amino acid composition and dipeptide composition . ManojBhasin, Gajendra P.S.Raghava. 22, 2004, J. Biol. Chem, Vol. 279, pp. 23262 –23266.
[59] Multilayer perceptrons for classification and regression . Murtagh, Fionn. 5-6, 1991, Neurocomputing, Vol. 2, pp. 183-197.
[60] Overview of Artificial Neural Networks. Zou J., Han Y., So SS. [ed.] Methods in Molecular Biology™. s.l. : Humana Press, 2008, In: Livingstone D.J. (eds) Artificial Neural Networks, Vol. 458, pp. 14-22.
[61] Artificial neural networks: fundamentals, computing, design, and application. I.A Basheer, M Hajmeer. 1, 2000, Journal of Microbiological Methods, Vol. 43, pp. 3-31.

[62] Supervised learning in spiking neural networks: A review of algorithms and evaluations. Xiangwen Wang, Xianghong Lin, Xiaochao Dang. 2020, Neural Networks, Vol. 125, pp. 258-280.
[63] Artificial Neural Networks for Neuroscientists: A Primer. Guangyu Robert Yang, Xiao-Jing Wang. 6, 2020, Neuron, Vol. 107, pp. 1048-1070.
[64] Stochastic Gradient Descent Tricks. Bottou, Léon. [ed.] Berlin, Heidelberg Springer. 2012, Neural Networks: Tricks of the Trade, Vol. 7700, pp. 421-436.
[65] Convergence of the RMSProp deep learning method with penalty for nonconvex optimization. Dongpo Xu, Shengdong Zhang, Huisheng Zhang, Danilo P. Mandic. 2021, Neural Networks, Vol. 139, pp. 17-23.
[66] An Improved Adagrad Gradient Descent Optimization Algorithm. N. Zhang, D. Lei and J. F. Zhao. 2018. Chinese Automation Congress (CAC). pp. 2359-2362.
[67] Improved Adam Optimizer for Deep Neural Networks. Zhang, Z. 2018. IEEE/ACM 26th International Symposium on Quality of Service (IWQoS). pp. 1-2.
[68] Support-vector networks. Corinna Cortes, Vladimir Vapnik. 1995, Machine Learning volume, Vol. 20, pp. 273–297.
[69] Protein function classification via support vector machine approach. C.Z Cai, W.L Wang, L.Z Sun, Y.Z Chen. 2, 2003, Mathematical Biosciences, Vol. 185, pp. 111-122.
[70] A novel fusion based on the evolutionary features for protein fold recognition using support vector machines. Refahi, M.S., Mir, A. & Nasiri, J.A. 14368, 2020, Sci Rep, Vol. 10.
[71] DeepSVM-fold: protein fold recognition by combining support vector machines and pairwise sequence similarity scores generated by deep learning networks. Bin Liu, Chen-Chen Li, Ke Yan. 5, 2020, Briefings in Bioinformatics, Vol. 21, pp. 1733–1741.
[72] Identification of DNA-Binding Proteins by Multiple Kernel Support Vector Machine and Sequence Information. Ding, Yijie, et al. 4, 2020, Current Proteomics, Vol. 17, pp. 302-310.
[73] SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence. C.Z. Cai, L.Y. Han, Z.L. Ji, X. Chen, Y.Z. Chen. 13, 2003, Nucleic Acids Research, Vol. 31, pp. 3692–3697.
[74] RBF Kernel Based Support Vector Machine with Universal Approximation and Its Application. Wang J., Chen Q., Chen Y. [ed.] Berlin, Heidelberg Springer. 2005, Advances in Neural Networks, Vol. 3173, pp. 512-517.
[75] Parameter selection in SVM with RBF kernel function. S. Han, Cao Qubo and Han Meng. 2012. World Automation Congress . pp. 1-4.
[76] protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences. Nan Xiao, Dong-Sheng Cao, Min-Feng Zhu, and Qing-Song Xu. 2015, Bioinformatics, Vol. 31, pp. 1857-1859.
[77] Deep sparse rectifier neural networks. Xavier Glorot, Antoine Bordes, Yoshua Bengio. 2011. Proceedings of the fourteenth international conference on artificial intelligence and statistics .
[78] Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. Sergey Ioffe, Christian Szegedy. 2015. Proceedings of the 32nd International Conference on Machine Learning. Vol. 37, pp. 448-456.
[79] Approximation of hyperbolic tangent activation function using hybrid methods. Silva, M. A. Sartin and A. C. R. da. 2013. 8th International Workshop on Reconfigurable and Communication-Centric Systems-on-Chip (ReCoSoC). pp. 1-6.
[80] Gaussian Distribution. X, Zhang. [ed.] Boston, MA Springer. 2011, Encyclopedia of Machine Learning.
[81] The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. Chicco, D., Jurman, G. 6, 2020, BMC Genomics, Vol. 21.
[82] Why Cohen's Kappa should be avoided as performance measure in classification. Rosario Delgado, Xavier-Andoni Tibau. 2019, PLOS ONE.
[83] Scikit-learn: Machine Learning in {Python}. Pedregosa, F. and Varoquaux, G. and Gramfort, A. and Michel, V.and Thirion, B. and Grisel, O. and Blondel, M. and Prettenhofer, P.and Weiss, R. and Dubourg, V. and Vanderplas, J. and Passos, A. andCournapeau, D. and Brucher, M. and Perrot, M. and Duchesna. 2011, Journal of Machine Learning Research, Vol. 12, pp. 2825--2830.