Analytical Biochemistry 631 (2021) 114358

Contents lists available at ScienceDirect

Analytical Biochemistry

journal homepage: www.elsevier.com/locate/yabio

Improved protein relative solvent accessibility prediction using deep multi-view feature learning framework

Xue-Qiang Fan [a], Jun Hu [a,*], Ning-Xin Jia [a], Dong-Jun Yu [b,**], Gui-Jun Zhang [a,***]

[a] College of Information Engineering, Zhejiang University of Technology, Hangzhou, 310023, China
[b] School of Computer Science and Engineering, Nanjing University of Science and Technology, Xiaolingwei 200, Nanjing, 210094, China

ARTICLE INFO

Keywords:
Protein relative solvent accessibility prediction
Bioinformatics
Multi-view feature learning
Bidirectional long short-term memory recurrent neural networks
Sequence-based feature

ABSTRACT

The accurate prediction of the relative solvent accessibility of a protein is critical to understanding its 3D structure and biological function. In this study, a novel deep multi-view feature learning (DMVFL) framework that integrates three different neural network units, i.e., bidirectional long short-term memory recurrent neural network, squeeze-and-excitation, and fully-connected hidden layer, with four sequence-based single-view features, i.e., position-specific scoring matrix, position-specific frequency matrix, predicted secondary structure, and roughly predicted three-state relative solvent accessibility probability, is developed to accurately predict relative solvent accessibility information of protein. On the basis of this newly developed framework, one new protein relative solvent accessibility predictor was proposed and called DMVFL-RSA, which employs a customized multiple feedback mechanism that helps to extract discriminative information embedded in the four single-view features. In benchmark tests on TEST524 and CASP14-derived (CASP14set) datasets, DMVFL-RSA outperforms other existing state-of-the-art protein relative solvent accessibility predictors when predicting two-state (exposure threshold of 25%), three-state (exposure thresholds of 9% and 36%), and four-state (exposure thresholds of 4%, 25%, and 50%) discrete values. For real-valued prediction on TEST524 and CASP14set, DMVFL-RSA has also gained high Pearson correlation coefficient values, indicating a positive correlation between the predicted and native relative solvent accessibility. Detailed analyses show that the major advantages of DMVFL-RSA lie in the high efficiency of the DMVFL framework, the applied multiple feedback mechanism, and the strong sensitivity of the sequence-based features. The web server of DMVFL-RSA is freely available at https://jun-csbio.github.io/DMVFL-RSA/for academic use. The standalone package of DMVFL-RSA is downloadable at https://github.com/XueQiangFan/DMVFL-RSA.

1. Introduction

The relative solvent accessibility (RSA) of protein is closely related to the spatial arrangement and packing of amino acid residues and thus is an important local structural characteristic for studying protein folding and functions [1–4]. Although wet-laboratory experiments can be used to measure the protein RSA, these tests are expensive and time consuming. In the post-genomic era with an avalanche of newly emerging protein sequences, developing a computationally efficient method that accurately predicts the RSA information of any protein is highly necessary. In view of this situation, a number of protein RSA predictors are continuously being designed.

Most existing predictors use statistical and machine-learning algorithms based on protein sequence information to predict the protein RSA information. According to the prediction modes, the existing predictors can be roughly divided into two categories, i.e., discrete-valued and real-valued predictors. For discrete-valued predictors, the RSA information of each residue is labeled as one of the multiple classified states using different exposure thresholds. For instance, in two-state discrete-valued predictors, the two states, i.e., exposed and buried, are generally sorted by an exposure threshold of 25%. In three-state discrete-valued predictors, the three states, i.e., exposed, intermediate, and buried, are classified by two exposure thresholds. In the early stage, the discrete-valued predictors, including SVMpsi [5], RaptorX [6], ACCpro5 [7],

* Corresponding author.
** Corresponding author.,
*** Corresponding author.
E-mail addresses: hujunum@zjut.edu.cn (J. Hu), njyudj@njust.edu.cn (D.-J. Yu), zgj@zjut.edu.cn (G.-J. Zhang).

https://doi.org/10.1016/j.ab.2021.114358
Received 8 June 2021; Received in revised form 22 August 2021; Accepted 25 August 2021
Available online 31 August 2021

BMRSA [8], PaleAle 4.0 [9], and PaleAle 5.0 [10], dominated the field of solvent accessibility prediction. However, their common drawback is the inability to accurately predict the real number value of RSA for each residue, thus seriously restricts their applicability.

Real-valued predictors can predict the discrete state and the real number value of RSA for each residue in a protein, thus overcoming the defects of discrete-valued predictors. To name a few: NETASA [11], NetSurfp [12], SVM-Cabins [13], SANN [14], PredRSA [15], PSO-SVR [16], QBES [17], SPIDER2 [18], SPIDER3 [19], SPIDER3-Single [20], NetSurfP-2.0 [21], and SPOT-1D [22]. These predictors generally utilize the protein sequence information and appropriate machine-learning algorithms to predict the RSA information of each protein residue. For example, in NETASA [11], a neural network-based method, which only takes the simple binary encoding of amino acid sequence information as its input features, is used to predict the real-valued protein RSA. In SANN [14], PSI-BLAST [23] is employed to generate sequence profiles that are fed into the algorithm of *k*-nearest neighbor to solve real-valued prediction and two-state and three-state discrete-valued predictions. In SPIDER2 [18], position-specific scoring matrix profile with a sliding window size of 17 (8 on either side of each query residue) is exploited to train an iterative deep-learning neural network for the real-valued prediction of protein RSA. In SPIDER3 [19], the bidirectional long short-term memory recurrent neural network (BiLSTM) [24] algorithm is applied to extract discriminative information from three different feature views, i.e., position-specific scoring matrix, physiochemical properties [25], and hidden Markov Model profiles [26], and accurately predict the real-valued protein RSA. In SPIDER3-Single [20], only one-hot vector is fed into the bidirectional long short-term memory recurrent neural networks for predicting the real-valued protein RSA. In NetSurfP-2.0 [21], two different sequence profiles generated using the HH-suite and MMseqs2 tools are employed as features to train the solvent accessibility prediction model using an architecture composed of convolutional and long short-term memory neural networks. In SPOT-1D [22], six discriminative feature views, i.e., position-specific scoring matrix, hidden Markov Model profiles [26], SPIDER3-predicted RSA [19], SPOT-Contact-predicted contact map [27], CCMpred-predicted contact map [28], and evolutionary coupling information [29,30], are serially combined to feed into the prediction model, which is trained via an ensemble of recurrent and residual convolutional neural networks. Although these predictors have achieved considerable progress in predicting the real number value of protein RSA, their accuracy is still not satisfactory. The main reason for this lies in that most of existing real-valued predictors, using one single-view feature or simply concatenating multiple single-view features in series as input, could not provide sufficiently discriminative information. There is an urgent need for designing new high-performance real-valued predictors.

In this study, a novel real-valued predictor called DMVFL-RSA was developed to further improve the performance of protein RSA prediction. This predictor is based on our newly established customized deep multi-view feature learning (DMVFL) framework that can extract discriminative information embedded in multiple single-view features. The DMVFL framework employs four sequence-based single-view features, i.e., position-specific scoring matrix, position-specific frequency matrix, predicted secondary structure, and roughly predicted three-state RSA probability, to effectively learn feature information. For each single-view feature, a sub neural network module, which is composed of two layers of bidirectional long short-term memory recurrent neural networks (BiLSTM), two layers of squeeze-and-excitation (SENet), and three fully-connected hidden layers (FC), is used to transform a single-view feature to a more discriminative feature. The four transformed single-view features are then coalesced and fed into a new sub neural network module, which is composed of two layers of SENet and three layers of FC (refer to the section of "Architecture of the DMVFL-RSA Framework" for detail). In addition, the predictor DMVFL-RSA uses a newly designed multiple feedback mechanism to improve the accuracy and generalization ability. Experimental results show that DMVFL-RSA outperforms other existing state-of-the-art discrete-valued and real-valued predictors, which is attributed to the proposed DMVFL framework, the applied multiple feedback mechanism, and the sensitivity of the sequence-based single-view features. The web server of DMVFL-RSA can be freely accessed for academic use at https://jun-csbio.github.io/DMVFL-RSA/.

2. Materials and methods

2.1. Benchmark data sets

A new dataset containing a training set and an independent validation set, denoted as TR10310 and TEST524, is constructed to evaluate the performance of RSA prediction. First, all protein chains released by RCSB Protein Data Bank (PDB) before November 10, 2019 are collected. After the exclusion of proteins with less than 30 residues or discontinuous 3D structure information, a dataset of 344,080 protein chains is obtained. The maximal pairwise sequence identity of the dataset protein chains is culled to 25% by using CD-HIT [31] program, and the obtained 10,310 protein chains are employed to constitute a training dataset called TR10310. To collect the independent validation set, we extract all protein chains deposited into PDB after November 10, 2019. The proteins with less than 30 residues or discontinuous 3D structure information are also removed. Again, the maximal pairwise sequence identity of the extracted protein chains is reduced to 25%. Furthermore, we also remove these proteins that each of them shares >25% identity to a protein chain in the training data set, i.e., TR10310. The remaining 524 protein chains constitute the independent validation set, called TEST524.

To further evaluate the performance of RSA prediction, the following 34 available target protein data are collected from the 14th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction (CASP14) to form an independent validation data set (called CASP14set) and further verify the performance of DMVFL-RSA and other existing predictors: T1024, T1025, T1026, T1027, T1029, T1030, T1031, T1032, T1033, T1035, T1036s1, T1037, T1038, T1039, T1040, T1041, T1042, T1043, T1044, T1046s1, T1046s2, T1049, T1050, T1054, T1056, T1064, T1067, T1073, T1074, T1079, T1080, T1082, T1090, and T1099. In CASP14set, there are 19 free modelling (FM) target proteins and 15 template-based modelling target proteins. All FM targets are selected to constitute one subset called CASP14set-Hard and the other targets are selected to compose another subset called CASP14set-Easy. Supplemental Table S1 lists the target proteins in the two subsets.

For each protein in the above datasets, the ground-truth solvent accessible surface area (ASA_i) of its ith residue is first calculated by DSSP program (Version 2.0.4) [32] based on its experimental 3D structure. The RSA value (RSA_i) of the ith residue, which is the true label value of interest in this study, is then obtained via dividing ASA_i by the maximum solvent accessible surface area (ASA_i^{MAX}) of the amino acid type of the ith residue. However, there is no standard for maximum solvent accessibility value. GLY-X-GLY extended tripeptides have been used in previous works [14,15,33]. Simply, the RSA value (RSA_i) of each residue could be calculated as:

$$RSA_i = \frac{ASA_i}{ASA_i^{MAX}} \times 100\%. \quad (1)$$

2.2. Feature representation

In this study, to effectively predict the protein RSA from protein sequence information, four feature views, i.e., position-specific scoring matrix, position-specific frequency matrix, predicted secondary structure, and roughly predicted three-state RSA probability, are employed to encode the feature representation of each residue.

2.2.1. Position-specific scoring matrix (PSSM)

The PSSM of a protein, which is achieved by PSI-BLAST [23] program, contains the important evolutionary information that implies whether one residue is conserved in its family of related proteins. PSSM can substantially improve the overall performance for protein RSA prediction [14,15,18,19,34]. In this study, the PSSM profile for protein sequence is built by using the PSI-BLAST [23] to search the non-redundant database through three iterations with 0.001 as the *E*-value cutoff for multiple sequence alignment against the query sequence. The obtained PSSM is further normalized with the logistic function:

$$f(x) = \frac{1}{1 + \exp(-x)}, \tag{2}$$

where x is the score derived from the PSSM profile.

2.2.2. Position-specific frequency matrix (PSFM)

Besides PSSM, position-specific frequency matrix (PSFM), which is proven to increase the accuracy of predicting protein 1D properties [10,35,36], is also employed to dig out the protein evolutionary information for reflecting the residue conservation and improving the performance of protein RSA prediction. To obtain the PSFM profile, for each query protein sequence with L^Q amino acid residues, HHblits [37], a fast and accurate sequence alignment tool, is utilized to search against Uniclust30 [38] database through three iterations, with 0.001 used as the *E*-value cutoff for generating the multiple sequence alignment (MSA) profile, which contains N sequences aligned to the query sequence. Based on the MSA profile, the corresponding PSFM profile with a size of $L^Q \times 21$ is calculated as follows:

$$P_{i,j} = \frac{1}{N}\sum_{n=1}^{N} \sigma\left(MSA_i^n, R_j\right), \tag{3}$$

where $P_{i,j}$ is the ith row and jth column element of PSFM profile; MSA_i^n represents the residue type (including gap type) at the ith position of the nth aligned sequence in the MSA profile; $i = 1,2, \ldots, L^Q$ and $n = 1,2, \ldots, N$, R_j is the type of the jth element of the set of 20 naturally-occurring residue types and one gap type, $j = 1,2, \ldots, 21$; and $\sigma(MSA_i^n, R_j) = 1$ if MSA_i^n is same as R_j, otherwise, $\sigma(MSA_i^n, R_j) = 0$.

2.2.3. Predicted secondary structure (PSS)

Protein secondary structure (PSS) information can help improve the performance for protein RSA prediction [8,15]. In this study, PSS information is also employed and obtained using PSIPRED (Version 3.2.1) [39] software, which predicts the probability that each residue in a protein sequence belongs to three secondary structure classes (coil (C), helix (H), and strand (E)). For a protein sequence with L^Q residues, the PSIPRED outputs an $L^Q \times 3$ probability matrix, which represents the predicted secondary structure information of the protein.

2.2.4. Roughly predicted three-state RSA probability (RPRSA)

To dig out more useful feature information, a new threading-based predictor, called TBP, is designed and employed to roughly predict the three-state RSA (RPRSA), i.e., exposed (*E*), intermediate (*I*), and buried (*B*), probability information of protein residues. For each query protein with L^Q residues, in TBP, its PSFM profile, which is denoted by $\boldsymbol{P^Q} = \{P_{i,j}^Q\}_{i=1,j=1}^{L^Q,21}$ and generated in the section of "Position-Specific Frequency Matrix," is employed to search against a newly collected database called PRSA-DB to search for distantly homologous proteins. In PRSA-DB, the PSFM profile ($\boldsymbol{P^{DB}} = \{P_{k,j}^{DB}\}_{k=1,j=1}^{L^{DB},21}$) and three-state RSA ($\boldsymbol{Sol^{DB}} = \{Sol_k^{DB}\}_{k=1}^{L^{DB}}$, $Sol_k^{DB} \in \{E, I, B\}$) of each database protein are previously generated. The details of generating PRSA-DB could be found in Supplemental Text S1.

To detect distantly homologous proteins, inspired by S-SITE [40], the query PSFM profile, i.e., $\boldsymbol{P^Q}$, is compared with the PSFM profile, i.e., $\boldsymbol{P^{DB}}$, of each database protein in PRSA-DB by using the Needleman-Wunsch dynamic programming algorithm [41] to obtain the optimal residue alignment based on the residue similar matrix ($\boldsymbol{SM} = \{SM_{i,k}\}_{i=1,k=1}^{L^Q,L^{DB}}$). The similar score ($SM_{i,k}$) of aligning the ith residue in the query to the kth residue in the database protein is simply calculated as

$$SM_{i,k} = \sum_{j}^{21} P_{i,j}^Q P_{k,j}^{DB} \Big/ \sqrt{\left(\sum_{j}^{21} P_{i,j}^Q P_{i,j}^Q\right) \times \left(\sum_{j}^{21} P_{k,j}^{DB} P_{k,j}^{DB}\right)}, \tag{4}$$

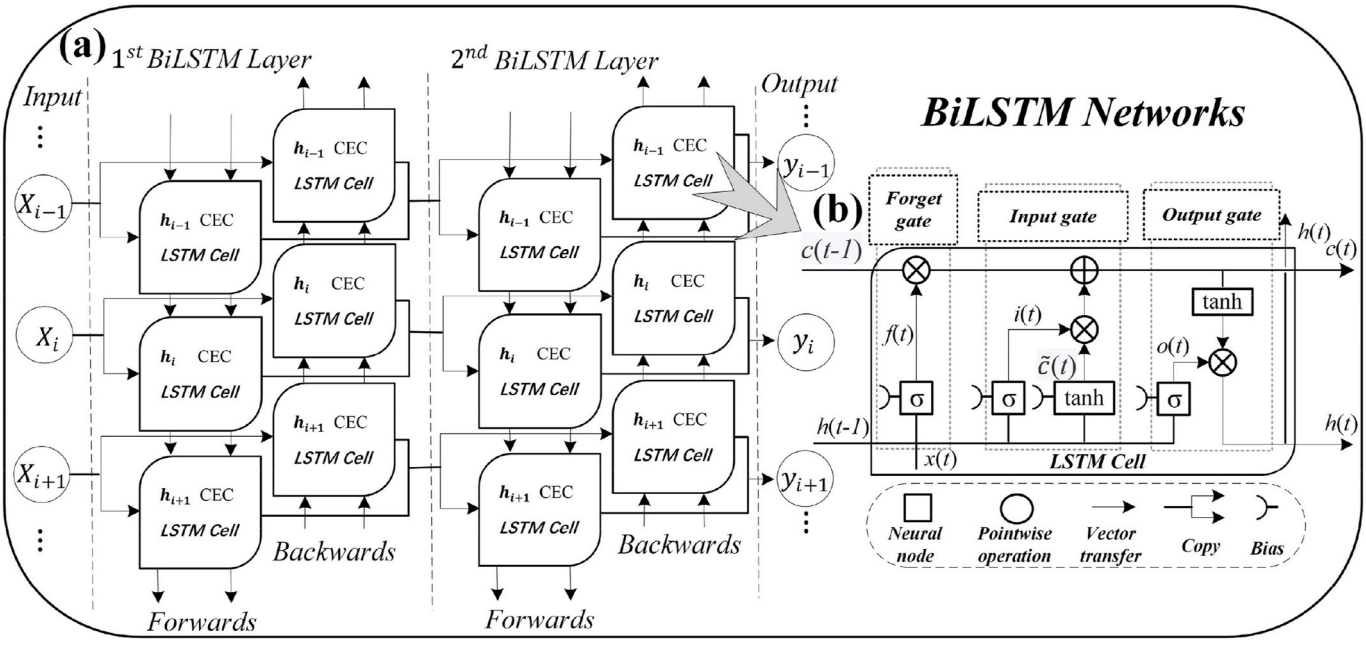

Fig. 1. Module of two-layer bidirectional long short-term memory recurrent neural networks. (a) Internal connections of two-layer BiLSTM networks. (b) Inner connections of long short-term memory neural network cell.

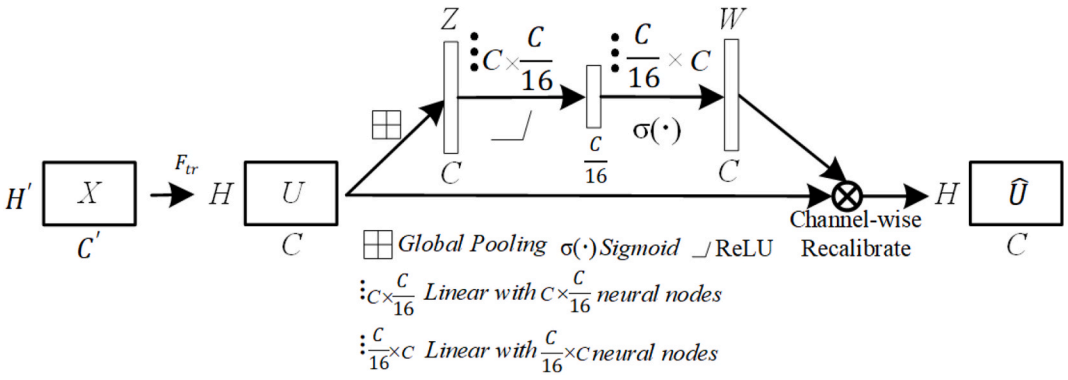

Fig. 2. Module of squeeze-and-excitation network.

where $P_{i,j}^{Q}$ denotes the ith row and jth column element in $\boldsymbol{P^{Q}}$, and $P_{k,j}^{DB}$ represents the kth row and jth column element in $\boldsymbol{P^{DB}}$. The alignment quality (AQ) between the query protein and the database protein is calculated by

$$AQ = \sum_{n}^{N_{ali}} SM_{ali_n^Q, ali_n^{DB}} \Big/ \sqrt{L^Q L^{DB}}, \tag{5}$$

where N_{ali} is the number of the aligned residue pairs, ali_n^Q and ali_n^{DB} represent the indexes of two residues, which form the nth aligned pair, in the query and database proteins.

Each database protein in PRSA-DB with a AQ value above a threshold (T_{AQ}) and a sequence identity <30% to the query protein is selected as one distantly homologous protein of the query. T_{AQ} is set to 0.5 in this study. Suppose there are D eligible database proteins, the roughly predicted three-state relative solvent accessibility probabilities, i.e., RP_i^E, RP_i^I, and RP_i^B) of the ith residue of the query protein could be easily calculated as

$$RP_i^S = \sum_{d}^{D} AQ^d \times \sum_{n}^{N_{ali}^d} \sigma\left(i, ali_{d,n}^{Q}\right) \times \sigma\left(S, Sol_{d,ali_{d,n}^{DB}}^{DB}\right) \Big/ \sum_{d}^{D} AQ^d, S \in \{E,\ I,\ B\}, \tag{6}$$

where $AQ^d$is the alignment quality between the query and the dth selected database proteins; N_{ali}^d is the number of the aligned residue pairs; $ali_{d,n}^{Q}$ and $ali_{d,n}^{DB}$ represent the indexes of two residues forming the nth aligned pair in the query and the dth selected database proteins, respectively; and $Sol_{d,ali_{d,n}^{DB}}^{DB}$ is the solvent accessibility state belonging to the set of E, I, and B of the $ali_{d,n}^{DB}$-th residue of the dth selected database protein. $\sigma(a,b) = 1$, if $a = b$, otherwise, $\sigma(a,b) = 0$.

According the above procedures, for each protein with L^Q residues, the TBP could easily generate an $L^Q \times 3$ probability matrix, which represents the RPRSA probability information of this protein.

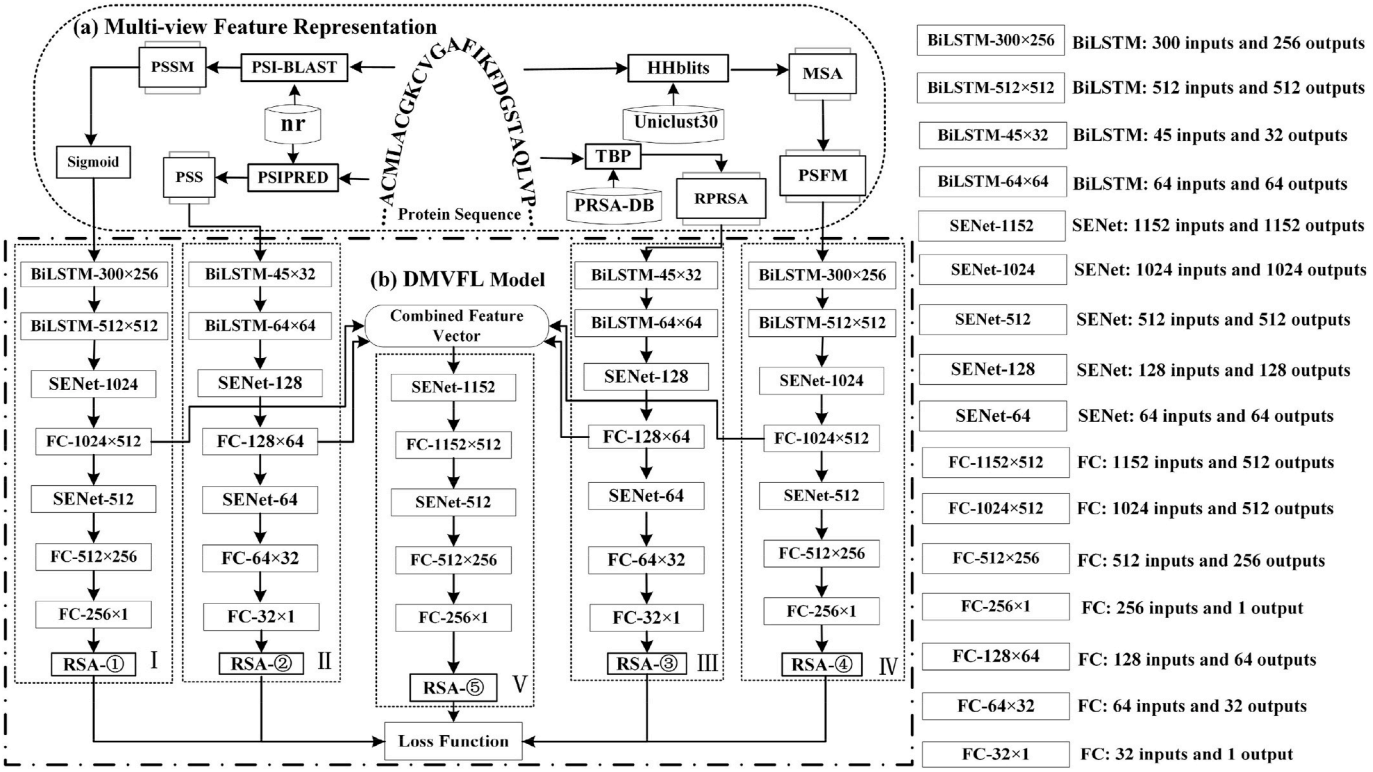

Fig. 3. Architecture of the proposed DMVFL-RSA. (a) Multi-view Feature Representation; (b) Architecture of the DMVFL framework.

2.3. Bidirectional long short-term memory recurrent neural networks (BiLSTM)

Given that the RSA information of each target residue is related to itself and to other residues in the same protein, extracting available feature information from the whole protein is important for accurately predicting the RSA of each residue. Sliding window technique is usually employed to extract the feature information [14,15,18] of a residue. However, the sliding window technique can only capture the information of the target residue and its neighbor residues, but cannot capture the information related to the target residue on the whole protein. In this study, BiLSTM [42] is employed to extract discriminative information related to each target residue on the whole protein. Inspired by the fact that a two-layer BiLSTM can learn context-dependent information more effectively [43,44], we use a two-layer BiLSTM framework in this study, as shown in Fig. 1a. Each layer contains several long short-term memory neural network (LSTM) cells [24]. Fig. 1b demonstrates the inner connections of each LSTM cell. The LSTM cell mainly includes three basic gates, i.e., an input gate, a forget get, and an output gate. The forget gate can decide what information will be thrown away from the cell state. When the cell state is updating, the input gate can decide what new information can be stored in the cell state, and the output gate decides what information can be output based on the cell state. The detail description of LSTM could be found in Supplemental Text S2.

2.4. Squeeze-and-Excitation Network (SENet)

Four single-view features, i.e., PSSM, PSFM, PSS, and RPRSA, are used to predict the protein RSA. SENet [45] is used to recalibrate the original feature representation during the training phase and extract discriminative information from the above single-view features. The architecture of the SENet is shown in Fig. 2. For a given input feature X, a linear transformation operator F_{tr} is first used to transform X to a new feature representation U. A squeeze operation is then employed to transform U as a channel descriptor via aggregating features across their spatial dimension. This descriptor produces an embedding of the global distribution of channel-wise feature responses, thus allowing information from the global receptive field of the network to be used by all of its layers. The squeeze operation is followed by an excitation operation, a simple self-gating mechanism that takes the embedding as the input and produces a collection of per-channel modulation weights. These weights are then applied to the feature U to generate the output feature $\widehat{U}$ of the SENet block [45], and this output be fed directly into the subsequent layers of the network.

2.5. Architecture of the DMVFL-RSA framework

In this study, a customized deep multi-view feature learning framework (DMVFL) is designed to extract the discriminative information from the above four single-view features (see Fig. 3a), i.e., PSSM, PSS, RPRSA, and PSFM. As shown in Fig. 3b, DMVFL contains five sub-pipelines named I, II, III, Ⅳ, and Ⅴ.

In the architecture of the DMVFL framework, the sub-pipelines I and Ⅳ correspond to the input features of the same shape, are set the same model hyperparameters, including two BiLSTM layers (denoted as "BiLSTM-300 × 256" and "BiLSTM-512 × 512"), two SENet layers (denoted as "SENet-1024" and "SENet-512"), and three fully-connected (FC) layers (denoted as "FC-1024 × 512," "FC-512 × 256," and "FC-256 × 1"). Similarly, the sub-pipelines II and III are also set the same model hyperparameters, including two BiLSTM layers (denoted as "BiLSTM-45 × 32" and "BiLSTM-64 × 64"), two SENet layers (denoted as "SENet-128" and "SENet-64"), and three FC layers (denoted as "FC-128 × 64," "FC-64 × 32," and "FC-32 × 1"). Besides, to dig out the fusion feature of the four single-view features above, we also design a sub-pipeline, called pipeline Ⅴ. The output features of the first FC layers of sub-pipelines I, II, III, and Ⅳ are fused as the input of sub-pipeline Ⅴ. The sub-pipeline Ⅴ includes two SENet layers (denoted as "SENet-1152" and "SENet-512") and three FC layers (denoted as "FC-1152 × 512," "FC-512 × 256," and "FC-256 × 1"). It is noted that, although the output of sub-pipeline V is employed as the final prediction result, the other four outputs of sub-pipeline I, II, III, and IV are also employed to calculate the loss value during the training procedure for extracting more discriminative feature information (see Section 2.6).

All nodes of the above-mentioned FC layers are activated by using hyperbolic tangent function. The output of each FC layer is normalized with batch instance normalization [46]. In the training phase, a multiple feedback mechanism is newly designed and applied to tune the parameter values of the DMVFL framework. The detail description of multiple feedback mechanism could be found in Section 2.6. To reduce network overfitting, a dropout strategy [47] is utilized. In this study, we use the strategy of grid search and adjust the network's hyper-parameters, i.e., learning rate and dropout ratio, by observing the model performance on the training dataset TR10310 over five-fold cross--validation tests. Finally, the optimal/sub-optimal values of learning rate and dropout ratio are tuned to be 0.001 and 0.5, respectively. The software of Pytorch (version 1.3.1) [48] is adopted to implement and tune the DMVFL framework on this cluster with one NVIDIA Titan RTX GPU.

2.6. Multiple feedback mechanism

The training of the DMVFL framework includes three cascaded phases, i.e., forward propagation, backward propagation, and gradient application. The backward propagation, which is a dynamic process, determines the optimal quality of the model parameters. In this study, to improve the generalization ability and prediction performance of the DMVFL model, a customized multiple backward propagation mechanism, which considers the combination of multiple factors to determine the optimal model parameters, is designed, named multiple feedback mechanism.

Different single-view features should contain varying impact factors to affect the prediction performance of protein RSA information. In multiple feedback mechanism, five output results of predicted RSA values (see Fig. 3b), i.e., RSA-①, RSA-②, RSA-③, RSA-④, and RSA-⑤, are first calculated by the five sub-pipelines during training to fully utilize the impact factors of different feature views. The mean squared error function is then adopted to compute five loss values i.e., loss-①, loss-②, loss-③, loss-④, and loss-⑤ corresponding to RSA-①, RSA-②, RSA-③, RSA-④, and RSA-⑤. The minimizing loss function for each back propagation is defined as:

$$Minimize \sum_{t=1}^{T} \frac{1}{2}(y - y_t)^2, \tag{7}$$

where y_t is the predicted RSA value, y is the actual RSA value of one target residue, and T is the number of sub-pipelines ($T = 5$). Back propagation is then utilized to minimize the loss function. Finally, the optimization method of Adam algorithm [49] is employed to estimate the optimal parameters of the DMVFL model.

2.7. Assessment metrics

Two widely used evaluation indexes, i.e., mean absolute error (MAE) and Pearson correlation coefficient (PCC), are employed to assess the performance for real-valued RSA prediction. MAE is used to quantitatively measure the deviation between the predicted and actual RSA values of each protein. PCC is applied to quantify the relationship between the predicted and actual RSA values of each protein, and its value was between −1 and 1. The two indexes can be calculated by following equations:

Table 1
Performance comparison of different features on TR10310 over five-fold cross-validation tests and on TEST524 over independent validation tests using the single-pipeline learning framework.

Data set	Feature	PCC		MAE	
		value	*p-value*	value	*p-value*
TR10310	PSSM	0.56	6.7×10^{-3}	21.2	1.4×10^{-4}
	PSFM	0.54	5.3×10^{-4}	22.3	7.7×10^{-3}
	PSS	0.45	2.4×10^{-6}	26.7	2.9×10^{-9}
	RPRSA	0.48	5.3×10^{-4}	23.1	3.2×10^{-7}
	PSSM + PSFM	0.60	4.3×10^{-2}	20.9	3.1×10^{-3}
	PSSM + PSFM + PSS	0.61	2.1×10^{-2}	20.4	8.4×10^{-3}
	PSSM + PSFM + PSS + RPRSA	0.62		17.2	
TEST524	PSSM	0.57	3.6×10^{-3}	17.3	1.9×10^{-1}
	PSFM	0.56	2.1×10^{-3}	18.6	5.1×10^{-1}
	PSS	0.41	7.6×10^{-3}	20.7	8.2×10^{-3}
	RPRSA	0.43	2.2×10^{-3}	18.4	7.5×10^{-2}
	PSSM + PSFM	0.61	1.1×10^{-2}	16.7	2.6×10^{-1}
	PSSM + PSFM + PSS	0.62	1.7×10^{-2}	16.9	4.5×10^{-1}
	PSSM + PSFM + PSS + RPRSA	0.64		16.6	

The *p*-values in Student's *t*-test are calculated for the differences between PSSM + PSFM + PSS + RPRSA and other features. Bold fonts highlight the best value in each category.

$$MAE = \frac{\sum_{i=1}^{N} \left| RSA_{pi} - RSA_{ri} \right|}{N} \times 100, \tag{8}$$

$$PCC = \frac{\sum_{i=1}^{N} \left(RSA_{pi} - \overline{RSA_p} \right) \left(RSA_{ri} - \overline{RSA_r} \right)}{\sqrt{\sum_{i=1}^{N} \left(RSA_{pi} - \overline{RSA_p} \right)^2 \sum_{i=1}^{N} \left(RSA_{ri} - \overline{RSA_r} \right)^2}}, \tag{9}$$

where N is the length of the query protein sequence; RSA_{pi} and RSA_{ri} are the predicted and actual RSA values of the ith residue in the query protein, respectively; and $\overline{RSA_p}$ and $\overline{RSA_r}$ are the corresponding average values of the entire query protein, respectively. The correlation between RSA_{pi} and RSA_{ri} increases with the PCC value. When PCC is -1, the correlation is fully negative. When PCC is 1, the correlation is fully positive.

Two evaluation indexes, i.e., accuracy score Q_i(i is 2, 3, and 4 for two-state, three-state, and four-state predictions, respectively) and Matthew's correlation coefficient (MCC), are adopted to measure the performance for discrete-valued RSA prediction. These indexes could be calculated by:

$$Q_i = \frac{N_{cor}}{N} \times 100, \tag{10}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}, \tag{11}$$

where N is the total number of residues in a protein chain; N_{cor} is the number of correctly predicted residues; and TP, TN, FP, and FN denote true positive, true negative, false positive, and false negative, respectively.

Table 2
Performance comparison between single-pipeline and multi-pipeline learning frameworks on TR10310 over five-fold cross-validation tests and on TEST524 over independent validation tests.

Data set	Learning Framework	PCC		MAE	
		Value	p-value	Value	p-value
TR10310	Single-pipeline	0.62	1.1×10^{-2}	17.2	3.7×10^{-4}
	Multi-pipeline	0.64		16.9	
TEST524	Single-pipeline	0.64	5.1×10^{-3}	16.6	4.6×10^{-4}
	Multi-pipeline	0.67		16.2	

The *p*-values in Student's *t*-test are calculated for the differences between Multi- and Single-pipeline learning frameworks. Bold fonts highlight the best value in each category.

3. Results and discussions

3.1. Prediction performance of different features

The real-valued RSA prediction performance of the four single-view features (i.e., PSSM, PSFM, PSS, and RPRSA) and three serially combined features (i.e., PSSM + PSFM, PSSM + PSFM + PSS, and PSSM + PSFM + PSS + RPRSA) is investigated in this section. Each feature is evaluated by performing five-fold cross-validation tests on TR10310 and independent validation tests on TEST524. Note that, in this section, each prediction model is trained using a specially designed single-pipeline learning framework (see Supplemental Text S3), which simply integrates three neural network units, i.e., BiLSTM, SENet, and FC. Table 1 summarizes the average PCC and MAE values of different features on TR10310 and TEST524.

From Table 1, it is easy to see that the four single-view features, i.e., PSSM, PSFM, PSS, and RPRSA, are all useful for predicting the real-valued RSA. The average PCC values of the four single-view features are all greater than 0.40 on both two data sets. It is also easily found that PSSM + PSFM + PSS + RPRSA achieves the best performance. The average PCC and MAE values of PSSM + PSFM + PSS + RPRSA are 0.62 and 17.2 on TR10310, which are 1.61% and 15.6% better than those of the second-best feature, i.e., PSSM + PSFM + PSS, and 10.7% and 18.8% better than those of the best single-view feature, i.e., PSSM, respectively. The difference between PSSM + PSFM + PSS + RPRSA and PSSM in the PCC values is statistically significant which has a p-value <0.01 in the Student's t-test. The similar comparison results could also be found on TEST524. These experimental results demonstrate that the four single-view features contain complementary information.

3.2. Enhancing performance by usage of multi-pipeline learning framework

To extract more discriminative information from the four used single-view features, i.e., PSSM, PSFM, PSS, and RPRSA, in this study, a new multi-pipeline learning framework is designed to learn the RSA prediction model rather than the above simple single-pipeline learning framework. The detail description of the multi-pipeline learning framework could be found in Supplemental Text S4. Actually, the multi-pipeline learning framework can be seen as a simplified version of DMVFL. To evaluate the efficacy of this multi-pipeline learning framework, the single-pipeline learning framework is employed as a control method. Based on the four single-view features, the RSA prediction performance of single- and multi-pipeline learning frameworks are evaluated on TR10310 over five-fold cross-validation tests and on TEST524 over independent validation tests. Table 2 illustrates the average values of PCC and MAE of single- and multi-pipeline learning frameworks.

From Table 2, we can find that the multi-pipeline learning framework is superior to single-pipeline learning framework with regard to the two evaluation indexes, i.e., PCC and MAE, on both TR10310 and

Table 3
Performance comparison between the DMVFL framework with and without using multiple feedback mechanism (MFM) on TR10310 over five-fold cross-validation tests and on TEST524 over independent validation tests.

Data set	With/without using MFM	PCC		MAE	
		value	p-value	Value	p-value
TR10310	Without	0.64	3.8×10^{-2}	16.9	5.2×10^{-2}
	With	0.66		15.7	
TEST524	Without	0.67	1.1×10^{-3}	16.2	8.6×10^{-2}
	With	0.71		14.0	

The *p*-values in Student's *t*-test are calculated for the differences between with and without using MFM. Bold fonts highlight the best value in each category.

TEST524. The average PCC values of multi-pipeline learning framework are 0.64 and 0.67 on TR10310 and TEST524, which are 3.2% and 4.7% higher than those of single-pipeline learning framework, respectively. The average MAE values of multi-pipeline learning framework are 16.9 and 16.2 on TR10310 and TEST524, which are 1.7% and 2.4% better than those of single-pipeline learning framework, respectively. The differences in PCC and MAE values are statistically significant which have *p*-values<0.02. The above comparison results demonstrate that using multi-pipeline learning framework could extract more discriminative information from multi-view features to enhance the RSA prediction performance.

3.3. Enhancing performance by usage of multiple feedback mechanism

To evaluate the efficiency of multiple feedback mechanism, we compare the real-valued RSA prediction performance of the DMVFL framework with and without using multiple feedback mechanism on TR10310 over five-fold cross-validation tests and on TEST524 over independent validation tests. The DMVFL framework without using multiple feedback mechanism is essentially the same as the multi-pipeline learning framework used in Section 3.2. Table 3 presents the performance comparison results.

From Table 3, it is easy to find that the DMVFL framework with using multiple feedback mechanism is superior to without using multiple feedback mechanism concerning the two evaluation indexes, i.e., PCC and MAE, on both TR10310 and TEST524. The average PCC values of with using multiple feedback mechanism are 0.66 and 0.71 on TR10310 and TEST524, which are 3.1% and 6.0% higher than those of without using multiple feedback mechanism, respectively. The average MAE values of with using multiple feedback mechanism are 15.7 and 14.0 on TR10310 and TEST524, which are 7.1% and 13.6% better than those of without using multiple feedback mechanism, respectively. The difference in PCC values is statistically significant which has *p*-values<0.05. The above comparison results demonstrate that using multiple feedback mechanism could help the DMVFL framework to enhance the RSA prediction performance.

3.4. Performance comparison with other existing methods

3.4.1. Performance comparison on real-valued RSA prediction

In this section, the real-valued RSA prediction efficiency of DMVFL-RSA is experimentally verified by comparing it with other existing state-of-the-art solvent accessibility predictors, including, SANN [14], SPIDER2 [18], SPIDER3 [19], SPIDER3-Single [20], NetSurfP-2.0 [21], and SPOT-1D [22] on two independent validation data sets, i.e., TEST524 and CASP14set. In order to obtain the prediction results of the other existing predictors quickly, the standalone programs of them, i.e., SANN, SPIDER2, SPIDER3, SPIDER3-Single, NetSurfP-2.0, and SPOT-1D, are downloaded from the corresponding websites, i.e., https://github.com/newtonjoo/sann, https://servers.sparks-lab.org/downloads/SPIDER2_local.tgz, https://servers.sparks-lab.org/downloads/SPIDER3_local.tgz, https://servers.sparks-lab.

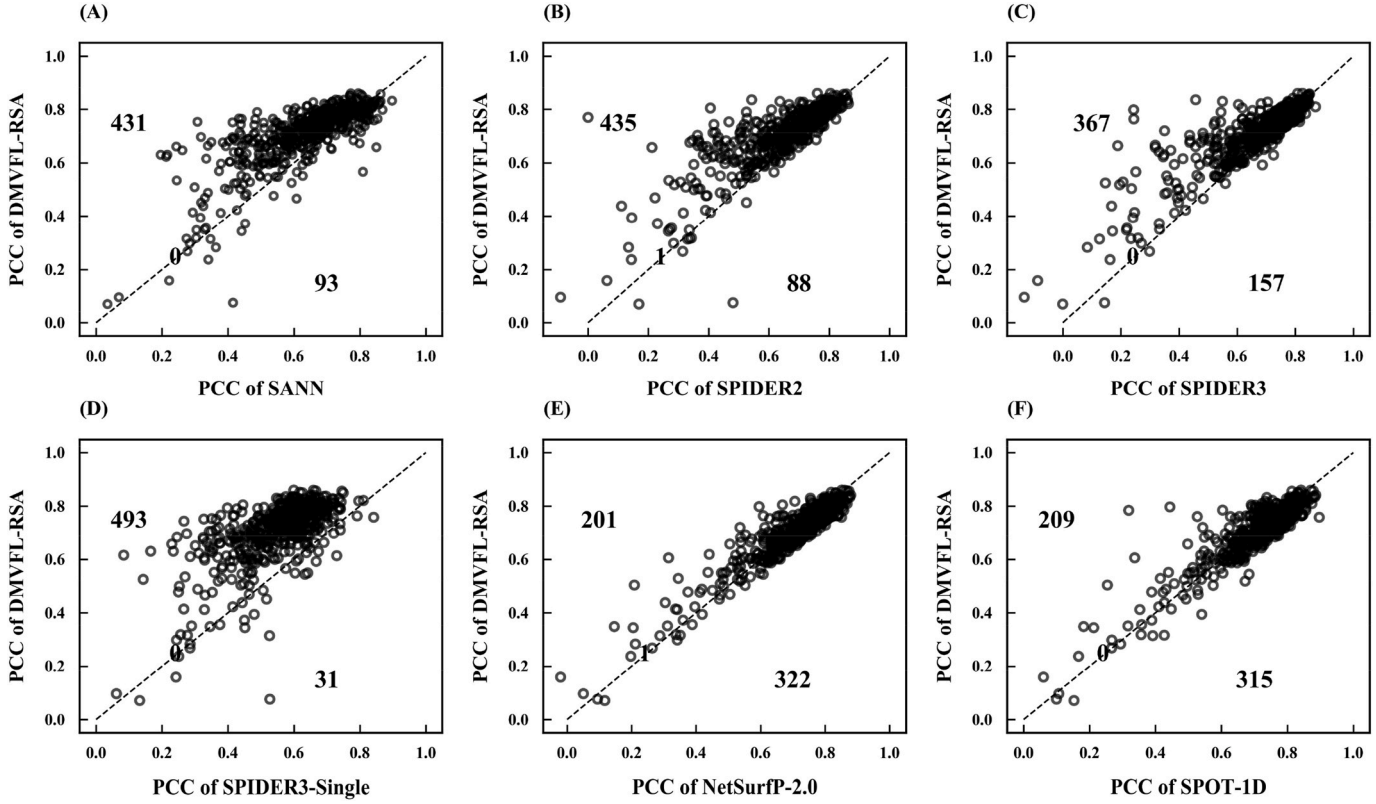

Fig. 4. Head-to-head comparisons of PCC values between DMVFL-RSA and other state-of-the-art predictors on TEST524. The numbers in each panel represent the number of points in the upper and lower triangles.

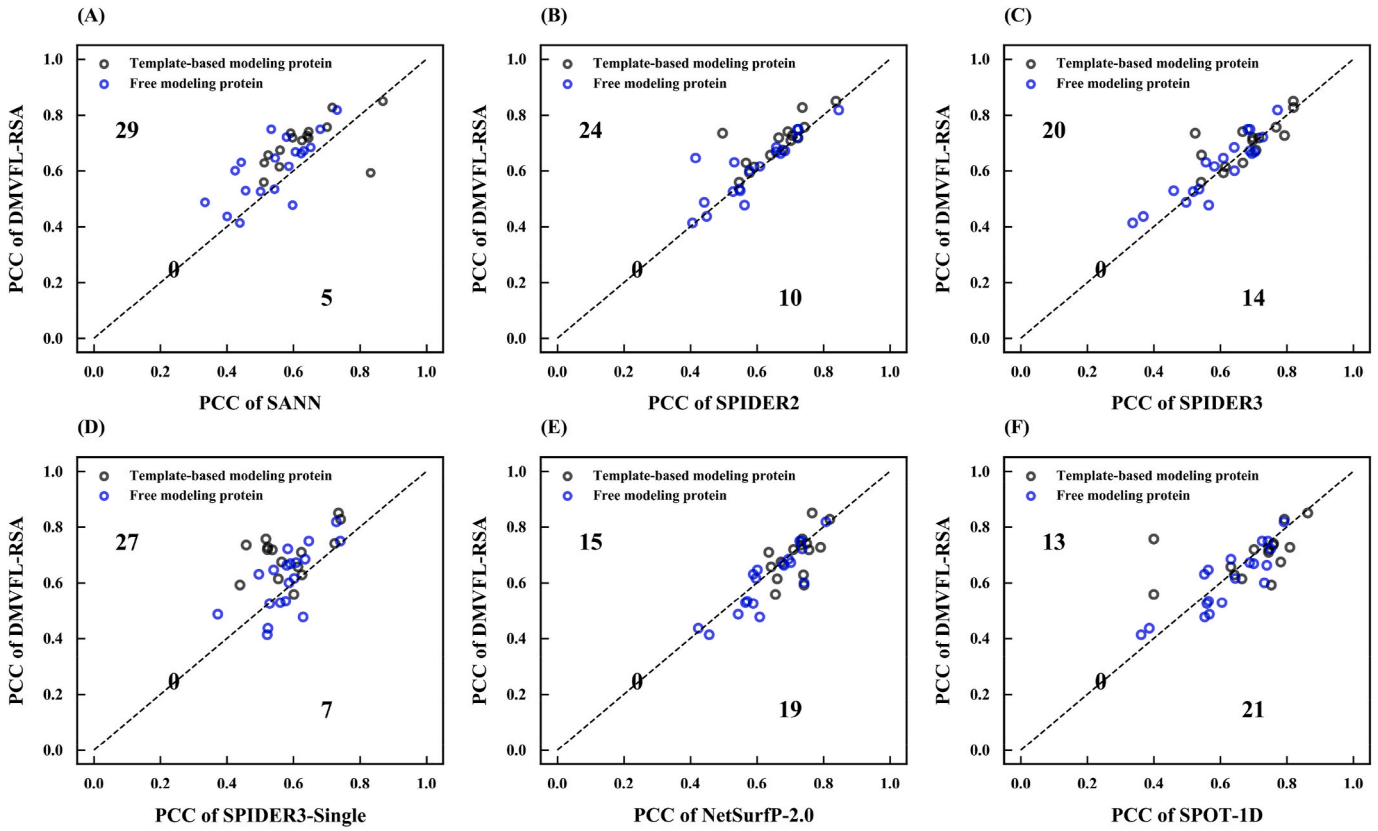

Fig. 5. Head-to-head comparisons of PCC values between DMVFL-RSA and other state-of-the-art predictors on CASP14set and its subsets, i.e., CASP14set-Hard and CASP14set-Easy. Each black (or blue) circle means one template-based modelling (or free modelling) protein target in the CASP14set. The numbers in each panel represent the number of points in the upper and lower triangles. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

org/downloads/SPIDER3-Single_np.tgz, http://www.cbs.dtu.dk/services/NetSurfP-2.0/, and https://servers.sparks-lab.org/downloads/SPOT-1D-local.tar.gz, respectively. The prediction model of DMVFL-RSA is learned on the training data set, i.e., TR10310, using the parameters tuned on TR10310 over five-fold cross-validation tests. The performance comparison results of DMVFL-RSA and other existing state-of-the-art predictors on TEST524 and CASP14set are separately demonstrated in Figs. 4 and 5. It is noted that, on CASP14set, the full chain sequences of all protein targets are employed as inputs to feed into each of the above predictors, rather than the single domain sequence to obtain the RSA prediction results, although the prediction performance is solely evaluated on the single domain regions with known label information.

By observing Fig. 4, we can easily find that, out of the 524 test proteins, there are 431, 436, 367, 493, 201, and 209 cases where DMVFL-RSA has equal or higher PCC than SANN, SPIDER2, SPIDER3, SPIDER3-Single, NetSurfP-2.0, and SPOT-1D, respectively. Although there are 322 and 315 cases where DMVFL-RSA has lower PCC than NetSurfP-2.0 and SPOT-1D, respectively, DMVFL-RSA could achieve a significantly higher PCC values on several cases, e.g., 6yj4J. Supplementary Table S2 also demonstrates the performance comparison between DMVFL-RSA and other existing state-of-the-art predictors on TEST524. From Table S2, it is easy to see that DMVFL-RSA outperforms other existing predictors concerning two evaluation indexes, i.e., PCC and MAE. The PCC and MAE values of DMVFL-RSA are 0.71 and 14.0, which are 10.9% and 19.5% better than those of SANN, 12.7% and 14.1% better than those of SPIDER2, 7.6% and 6.0% better than those of SPIDER3, 29.1% and 23.1% better than those of SPIDER3-Single, 1.4% and 0.7% better than those of NetSurfP-2.0, and 2.9% and 1.4% better than those of SPOT-1D, respectively. The differences between DMVFL-RSA and the existing predictors, i.e., SANN, SPIDER2, SPIDER3-Single, NetSurfP-2.0, and SPOT-1D, in MAE are statistically significant which have p-values$<10^{-5}$.

By visiting Fig. 5, it is easy to find that, among the 34 target proteins on CASP14set, DMVFL-RSA has 29, 24, 20, 27, 15, and 13 cases with higher PCC values than SANN, SPIDER2, SPIDER3, SPIDER3-Single, NetSurfP-2.0, and SPOT-1D, respectively. Out of the 19 target proteins in CASP14set-Hard, there are 16, 11, 9, 14, 8, and 5 cases where DMVFL-RSA has equal or higher PCC than SANN, SPIDER2, SPIDER3, SPIDER3-Single, NetSurfP-2.0, and SPOT-1D, respectively. Out of the 15 targets in CASP14set-Easy, there are 13, 13, 11, 13, 7, and 8 cases where DMVFL-RSA has equal or higher PCC than SANN, SPIDER2, SPIDER3, SPIDER3-Single, NetSurfP-2.0, and SPOT-1D, respectively. It is not escaped from our notice that DMVFL-RSA achieve a slightly lower overall performance than NetSurfP-2.0 and SPOT-1D on CASP14set (see Supplementary Table S3). The potential reason is that the other prediction tasks, i.e., protein secondary structure and backbone angles prediction, of NetSurfP-2.0 and SPOT-1D should help them to enhance the protein RSA prediction performance. Developing multi-task prediction method is a good way to further enhance the protein RSA prediction performance.

3.4.2. *Performance comparison on two-state RSA prediction*

In this section, the efficiency of the proposed DMVFL-RSA is evaluated on the two-state (i.e., exposed and buried) RSA prediction. To predict the two-state RSA value, DMVFL-RSA first predicts the real-valued RSA for each query protein residue. The predicted real-valued RSA is transformed into two states of seven different types using different predefined thresholds, i.e., 5%, 10%, 20%, 25%, 30%, 40%, and 50% [15,34]. Taking threshold of 5% as an example, if the real-valued RSA of a residue is no less than 5%, it is regarded as an exposed residue, otherwise it is classified into the buried class [15]. Six

Table 4
Two-state discrete-valued RSA prediction performance of DMVFL-RSA, SANN, SPIDER2, SPIDER3, SPIDER3-Single, NetSurfP-2.0, and SPOT-1D under seven different two-state thresholds, i.e., 5%, 10%, 20%, 25%, 30%, 40%, and 50%, on TEST524.

Predictor	Q_2						
	5%	10%	20%	25%	30%	40%	50%
SANN	84.0	80.9	76.3	74.3	72.9	72.2	74.5
SPIDER2	84.2	82.9	80.8	78.1	76.7	74.7	75.6
SPIDER3	85.0	82.9	80.8	80.4	79.6	78.8	78.0
SPIDER3-Single	80.3	75.6	73.9	74.0	73.6	73.3	74.2
NetSurfP-2.0	85.6	83.3	81.2	80.2	80.1	78.2	78.1
SPOT-1D	85.1	83.7	81.4	80.5	79.4	77.8	77.9
DMVFL-RSA	86.0	84.9	81.8	81.0	80.3	79.1	78.3

Bold fonts highlight the best value in each category.

state-of-the-art RSA predictors, i.e., SANN [14], SPIDER2 [18], SPIDER3 [19], SPIDER3-Single [20], NetSurfP-2.0 [21], and SPOT-1D [22] are used as control in this section. Tables 4 and 5 summarize the comparison results on TEST524 and CASP14set, respectively.

By observing Table 4, it is easy to find that DMVFL-RSA is consistently superior to the six control predictors with regard to the Q_2 evaluation indexes with the seven different two-state thresholds on TEST524. The Q_2 values of DMVFL-RSA for the seven different thresholds, i.e., 5%, 10%, 20%, 25%, 30%, 40%, and 50% are 86.0%, 84.9%, 81.8%, 81.0%, 80.3%, 79.1%, and 78.3%, respectively, which are 0.5%, 1.4%, 0.5%, 0.6%, 0.9%, 0.4%, and 0.3% higher than the second-best values, respectively. Supplementary Table S4 demonstrates the *p*-values in Student's *t*-test for the differences in Q_2 values with different thresholds between DMVFL-RSA and other six existing predictors on TEST524.

By visiting Table 5, we can find that DMVFL-RSA obtains $Q_2 > 73\%$ for all the seven two-state thresholds on CASP14set and its subsets, i.e., CASP14set-Hard and CASP14set-Easy. On CASP14set, DMVFL-RSA achieves the highest values, i.e., 85.4, 80.8, 78.2, 76.8, and 75.5, on the Q_2 evaluation indexes for thresholds 5%, 10%, 20%, 25%, and 30%, although DMVFL-RSA achieves slightly lower Q_2 values (i.e., 73.6% and 75.4%) on the thresholds of 40% and 50%. On CASP14set-Hard, DMVFL-RSA achieves the best performance concerning the Q_2 evaluation index for thresholds 5% and obtains the second-best performance concerning Q_2 for 20% and 25%. On CASP14set-Easy, DMVFL-RSA gains the best performance concerning the Q_2 evaluation indexes for thresholds 10%, 20%, 25%, 30%, 40%, and 50% and obtains the second-best performance concerning Q_2 for 5%. Supplementary Table S5 lists the *p*-values in Student's *t*-test for the differences in Q_2 values with different thresholds between DMVFL-RSA and other six existing predictors on CASP14set and its subsets, i.e., CASP14set-Hard and CASP14set-Easy.

3.4.3. Performance comparison on three-state RSA prediction

In this section, the efficiencies of DMVFL-RSA, SANN, SPIDER2, SPIDER3, SPIDER3-Single, NetSurfP-2.0, and SPOT-1D on three-state (i.e., exposed, intermediate, and buried) RSA prediction are evaluated. Each predictor, e.g., DMVFL-RSA, first predicts the real-valued RSA for each query protein residue. The predicted real-valued RSA is then transformed into three states using a general pair of thresholds, i.e., 9% and 36% [14,50]. If the real-valued RSA of a residue is more than 36%, then it is regarded as the residue with exposed (*E*) state. If the real-valued RSA of a residue is less than 9%, then it is labeled as the state of buried (*B*). If the real-valued RSA of a residue is between 9% (inclusive) than 36% (exclusive), then it is classified as the state of intermediate (*I*). The three-state RSA prediction performances of DMVFL-RSA, SANN, SPIDER2, SPIDER3, SPIDER3-Single, NetSurfP-2.0, and SPOT-1D generated on TEST524 and CASP14set are shown in Fig. 6.

By observing Fig. 6, we can find that DMVFL-RSA outperforms other six state-of-the-art predictors concerning the Q_3 and MCCs of *B* and *E* evaluation indexes on both TEST524 and CASP14set. The values of Q_3 and MCCs of *B*, *I*, and *E* of DMVFL-RSA are 67.0 and 60.0, 0.51 and 0.41, 0.22 and 0.19, and 0.52 and 0.49 on TEST524 and CASP14set, respectively (see details in Supplementary Tables S6 and S7). On CASP14set-Hard, DMVFL-RSA achieves the highest value (0.36) of MCC of *B* and gains the second-best value (0.19) of MCC of *I* (see Supplementary Table S8). On CASP14set-Easy, DMVFL-RSA obtains the highest values (60.5 and 0.52) of Q_3 and MCC of *E* and gains the second-best value (0.46) of MCC of *B* (see Table S9). Supplementary Tables S6, S7, S8, and S9 present the *p*-values in Student's *t*-test for the differences between DMVFL-RSA and other six existing predictors on TEST524, CASP14set, CASP14set-Hard, and CASP14set-Easy, respectively.

3.4.4. Performance comparison on four-state RSA prediction

In this section, the performance of the proposed DMVFL-RSA is

Table 5
Two-state discrete-valued RSA prediction performance of DMVFL-RSA, SANN, SPIDER2, SPIDER3, SPIDER3-Single, NetSurfP-2.0, and SPOT-1D under seven different two-state thresholds, i.e., 5%, 10%, 20%, 25%, 30%, 40%, and 50%, on CASP14set and its subsets, i.e., CASP14set-Hard and CASP14set-Easy.

Data set	Predictor	Q_2						
		5%	10%	20%	25%	30%	40%	50%
CASP14set	SANN	83.4	78.7	75.1	74.4	73.1	72.6	74.0
	SPIDER2	83.8	80.1	76.3	75.3	74.5	73.6	75.8
	SPIDER3	84.2	80.2	76.8	75.6	75.1	74.4	75.5
	SPIDER3-Single	82.1	76.5	73.4	72.4	72.2	72.0	74.1
	NetSurfP-2.0	84.3	79.7	76.9	75.9	75.4	74.6	74.9
	SPOT-1D	85.0	80.4	77.4	76.3	75.2	75.0	75.6
	DMVFL-RSA	85.4	80.8	78.2	76.8	75.5	73.6	75.4
CASP14set-Hard	SANN	85.1	79.2	74.9	74.2	72.0	70.0	71.4
	SPIDER2	85.7	81.4	76.3	75.0	73.7	72.7	73.7
	SPIDER3	85.7	81.2	76.1	74.6	74.2	73.1	72.9
	SPIDER3-Single	85.1	79.4	74.7	73.2	72.3	70.8	72.1
	NetSurfP-2.0	84.9	79.1	76.2	75.3	75.7	74.2	75.1
	SPOT-1D	84.2	80.6	77.6	76.5	75.1	74.7	74.6
	DMVFL-RSA	86.4	81.0	77.5	75.7	74.0	71.8	73.3
CASP14set-Easy	SANN	81.5	77.8	75.1	74.2	74.4	75.3	77.6
	SPIDER2	82.1	78.6	76.5	75.0	74.0	75.1	77.9
	SPIDER3	82.5	79.1	77.4	77.4	76.1	75.2	77.5
	SPIDER3-Single	78.4	72.4	70.5	71.6	72.6	72.1	76.4
	NetSurfP-2.0	83.7	80.3	77.8	76.6	75.0	75.1	74.7
	SPOT-1D	85.9	80.2	77.1	76.1	75.5	75.3	76.6
	DMVFL-RSA	84.1	80.6	79.0	78.1	77.3	75.9	78.2

Bold fonts highlight the best value in each category.

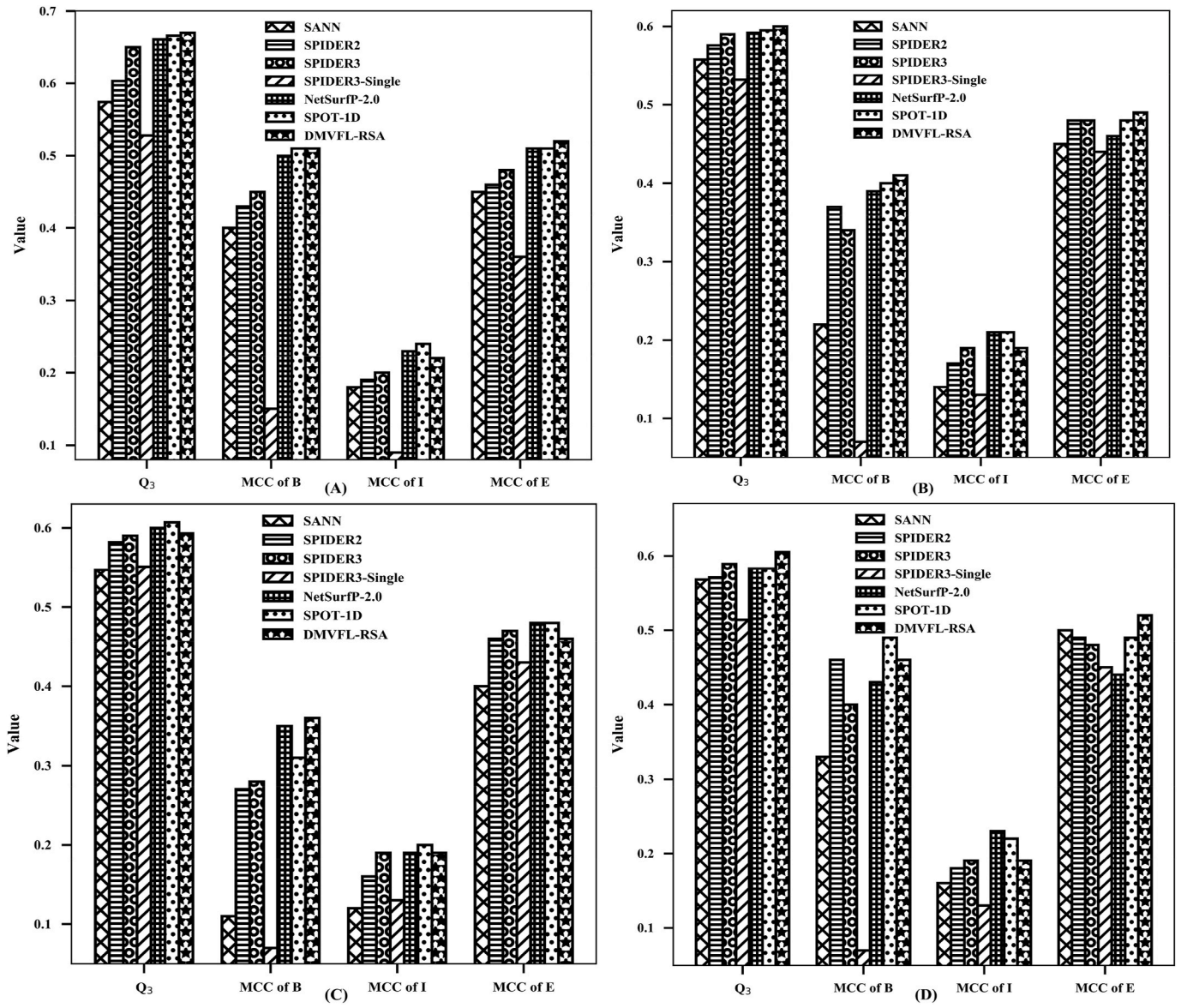

Fig. 6. Three-state discrete-valued prediction performance of DMVFL-RSA, SANN, SPIDER2, SPIDER3, SPIDER3-Single, NetSurfP-2.0, and SPOT-1D on TEST524, CASP14set, CASP14set-Hard, and CASP14set-Easy. (A) On TEST524. (B) On CASP14set. (C) On CASP14set-Hard. (D) On CASP14set-Easy.

Table 6
Four-state discrete-valued prediction performances of DMVFL-RSA, SANN, SPIDER2, SPIDER3, SPIDER3-Single, NetSurfP-2.0, and SPOT-1D on TEST524, CASP14set, CASP14set-hard, and CASP14set-Easy.

Predictor	TEST524		CASP14set		CASP14set-Hard		CASP14set-Easy	
	Q_4	*p-value*	Q_4	*p-value*	Q_4	*p-value*	Q_4	*p-value*
SANN	45.9	7.5×10^{-7}	44.5	4.7×10^{-2}	43.2	2.1×10^{-1}	46.6	1.1×10^{-1}
SPIDER2	48.3	4.7×10^{-14}	48.0	8.8×10^{-1}	48.4	3.2×10^{-1}	47.7	1.6×10^{-1}
SPIDER3	52.8	3.9×10^{-6}	48.1	8.9×10^{-1}	47.7	5.1×10^{-1}	50.3	6.9×10^{-1}
SPIDER3-Single	40.6	6.5×10^{-42}	43.8	1.1×10^{-2}	42.6	1.7×10^{-1}	40.0	3.5×10^{-4}
PaleAle5.0	47.4	1.6×10^{-31}	45.9	5.1×10^{-2}	46.9	7.2×10^{-1}	44.8	3.1×10^{-3}
NetSurfP-2.0	55.8	7.4×10^{-3}	48.5	1.6×10^{-2}	49.6	2.7×10^{-2}	47.5	1.7×10^{-1}
SPOT-1D	55.3	6.7×10^{-3}	49.2	1.2×10^{-1}	49.4	2.0×10^{-1}	49.2	2.3×10^{-1}
DMVFL-RSA	55.9		49.5		47.2		51.4	

The *p*-values in Student's *t*-test are calculated for the differences between DMVFL-RSA and other predictors. Bold fonts highlight the best value in each category.

examined in terms of four-state (i.e., very exposed, somewhat exposed, somewhat buried, and very buried) RSA prediction. DMVFL-RSA first predicts the real-valued RSA value of each query protein residue. As described in PaleAle5.0 [10], which is a four-state discrete-valued RSA predictor, the above predicted real-valued RSA value is mapped into four states using three thresholds, i.e., 4%, 25%, and 50%. The performances of DMVFL-RSA, SANN [14], SPIDER2 [18], SPIDER3 [19], SPIDER3-Single [20], PaleAle5.0 [10], NetSurfP-2.0 [21], and SPOT-1D

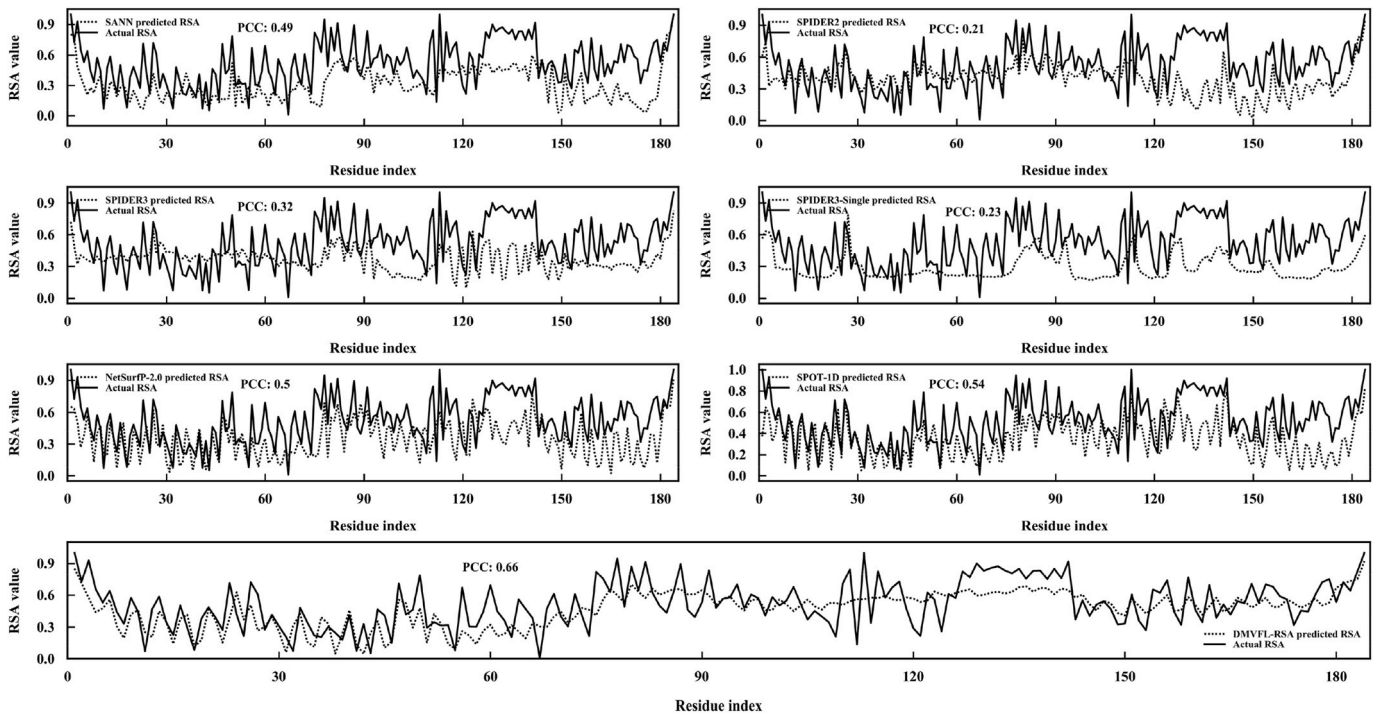

Figure 7. RSA prediction performance of different predictors, i.e., DMVFL-RSA, SANN, SPIDER2, SPIDER3, SPIDER3-Single, NetSurfP-2.0, and SPOT-1D, for 6yj4J. On each subgraph, the dotted line indicates the predicted RSA values, and the solid line indicates the actual RSA values.

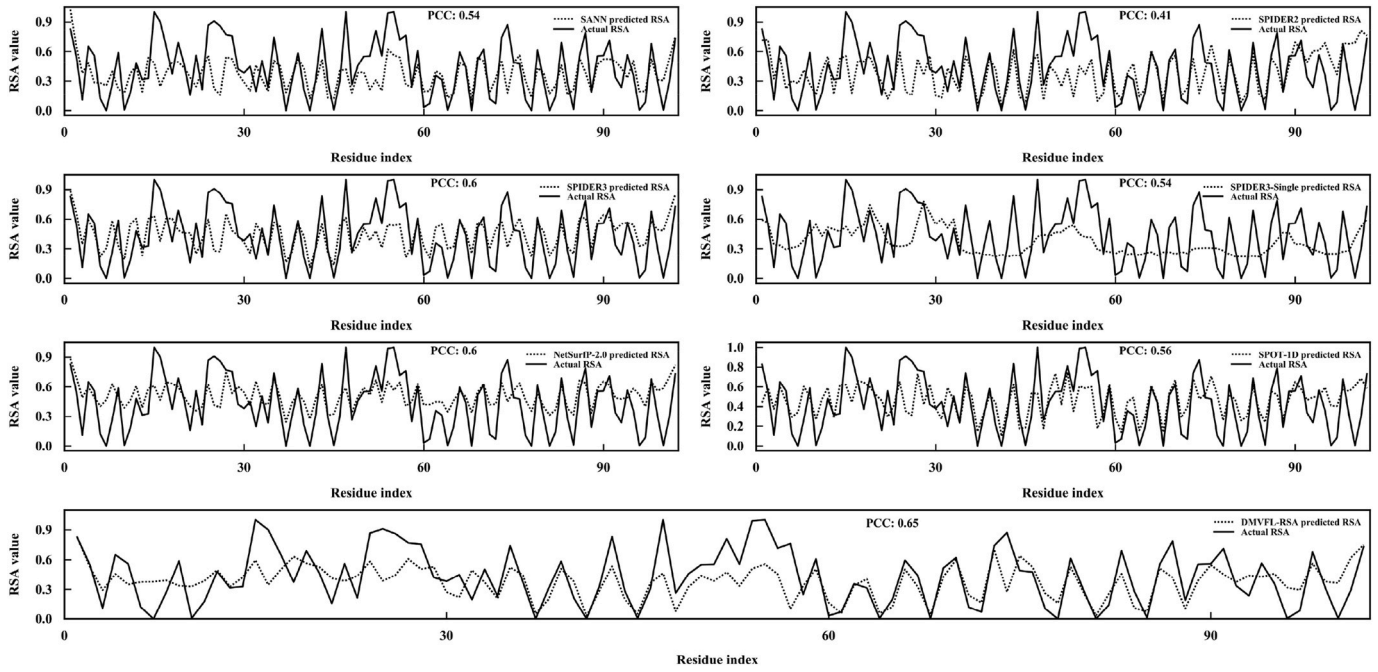

Fig. 8. RSA prediction performance of different predictors, i.e., DMVFL-RSA, SANN, SPIDER2, SPIDER3, SPIDER3-Single, NetSurfP-2.0, and SPOT-1D, for T1035. On each subgraph, the dotted line indicates the predicted RSA values, and the solid line indicates the actual RSA values.

[22] are generated on TEST524, CASP14set, CASP14set-Hard, and CASP14set-Easy. Their prediction results are summarized in Table 6.

By visiting Table 6, it is clearly demonstrated that DMVFL-RSA outperforms all of the seven existing state-of-the-art predictors, i.e., SANN, SPIDER2, SPIDER3, SPIDER3-Single, PaleAle5.0, NetSurfP-2.0, and SPOT-1D concerning Q_4 on three independent validation sets, i.e., TEST524, CASP14set, and CASP14set-Easy. Taking TEST524 as an example, the Q_4 value of DMVFL-RSA is 55.9%, which is 21.8%, 15.7%, 5.9%, 37.7%, 17.9%, 0.2%, and 1.2% higher than that of SANN, SPIDER2, SPIDER3, SPIDER3-Single, PaleAle5.0, NetSurfP-2.0, and SPOT-1D, respectively. It is not escaped from our notice that DMVFL-RSA achieve a lower Q_4 (47.2) on CASP14set-Hard. The potential reason should be that the pattern knowledge of the free modelling (FM) proteins dose not be completely learned by the prediction model of DMVFL-RSA trained on TR10310.

3.5. Case studies

In this section, two proteins with IDs 6yj4J and T1035, which are selected from TEST524 and CASP14set, respectively, are used for case studies. Figs. 7 and 8 show how well the real-valued RSA values are predicted by the seven predictors, i.e., DMVFL-RSA, SANN [14], SPIDER2 [18], SPIDER3 [19], SPIDER3-Single [20], NetSurfP-2.0 [21], and SPOT-1D [22]. The actual RSA values calculated by DSSP program [32] are fitted based on the experimental 3D structures of the two proteins.

Figs. 7 and 8 show that, in the two cases, the majority of DMVFL-RSA-predicted RSA values are in good agreement with the corresponding experimental RSA values. On 6yj4J and T1035, DMVFL-RSA achieves the PCC values of 0.66 and 0.65, which are 34.7% and 20.4%, 214.3% and 58.5%, 106.3% and 8.3%, 187.0% and 20.4%, 32.0% and 8.3%, and 22.3% and 16.1% higher than those of SANN [14], SPIDER2 [18], SPIDER3 [19], SPIDER3-Single [20], NetSurfP-2.0 [21], and SPOT-1D [22], respectively.

4. Conclusions

Accurately predicting the RSA is crucial in understanding the 3D structure and biological function of the protein. In this study, a new protein RSA predictor named DMVFL-RSA is designed and implemented to enhance the performance for protein RSA prediction. Experimental results show that the proposed DMVFL-RSA outperforms other existing state-of-the-art predictors. The superior performance of DMVFL-RSA can be attributed to several reasons, including a well-designed deep multi-view feature learning framework, an appropriate benchmark dataset, a multiple feedback mechanism, and a discriminative feature design. For ease of use, the proposed DMVFL-RSA has been implemented as a web server and is now available at https://jun-csbio.github.io/DMVFL-RSA/.

Our future work are directed toward the following four main directions to further improve the performance of protein RSA prediction: (1) developing useful strategies to extract discriminative single-view feature; (2) developing more accurate prediction models to predict the related feature source information, such as effective evolutionary information based on multiple sequence alignment [51]; (3) developing an accurate method by combining DMVFL-RSA and other state-of-the-art protein RSA prediction methods; (4) employing the multi-task learning algorithms [52] to predict protein solvent accessibility, protein secondary structure, protein backbone angles, and protein-protein interactions [53]. Furthermore, the predicted RSA information will be implemented to help enhance the accuracy of protein 3D structure prediction and the efficiency of protein design. Although the DMVFL-RSA still has room for optimization, we believe that it would be one of the most accurate tools for protein RSA prediction.

Declaration of competing interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

CRediT authorship contribution statement

Xue-Qiang Fan: designed research, performed research, Formal analysisanalyzed data, Writing – original draft. **Jun Hu:** designed research, performed research, Formal analysis. **Ning-Xin Jia:** Writing – original draft. **Dong-Jun Yu:** designed research. **Gui-Jun Zhang:** designed research.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 61902352, 62072243, 61772273, and 61773346), the Natural Science Foundation of Zhejiang (No. LY21F020025 and LZ20F030002), the Natural Science Foundation of Jiangsu (No. BK20201304), the Foundation of National Defense Key Laboratory of Science and Technology (JZX7Y202001SY000901), and the Fundamental Research Funds for the Provincial Universities of Zhejiang (No. RF-A20200012). J Hu, D.J. Yu, and G.J. Zhang are the corresponding authors for this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ab.2021.114358.

References

[1] H.S. Chan, K.A. Dill, Origins of structure in globular proteins, Proc. Natl. Acad. Sci. U. S. A. 87 (16) (1990) 6388–6392.
[2] C.B. Anfinsen, Principles that govern the folding of protein chains, Science 181 (4096) (1973) 223–230.
[3] S. Miller, A.M. Lesk, J. Janin, C. Chothia, The accessible surface area and stability of oligomeric proteins, Nature 328 (6133) (1987) 834–836.
[4] J. Janin, Surface and inside volumes in globular proteins, Nature 277 (5696) (1979) 491–492.
[5] H. Kim, H. Park, Prediction of protein relative solvent accessibility with support vector machines and long-range interaction 3D local descriptor, Proteins 54 (3) (2004) 557–562.
[6] S. Wang, W. Li, S. Liu, J. Xu, RaptorX-Property: a web server for protein structure property prediction, Nucleic Acids Res. 44 (W1) (2016) W430–W435.
[7] C.N. Magnan, P. Baldi, SSpro/ACCpro 5: almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity, Bioinformatics 30 (18) (2014) 2592–2597.
[8] W. Wu, Z. Wang, P. Cong, T. Li, Accurate prediction of protein relative solvent accessibility using a balanced model, BioData Min. 10 (2017) 1.
[9] C. Mirabello, G. Pollastri, Porter, PaleAle 4.0: high-accuracy prediction of protein secondary structure and relative solvent accessibility, Bioinformatics 29 (16) (2013) 2056–2058.
[10] M. Kaleel, M. Torrisi, C. Mooney, G. Pollastri, PaleAle 5.0: prediction of protein relative solvent accessibility by deep learning, Amino Acids 51 (9) (2019) 1289–1296.
[11] S. Ahmad, M.M. Gromiha, NETASA: neural network based prediction of solvent accessibility, Bioinformatics 18 (6) (2002) 819–824.
[12] B. Petersen, T.N. Petersen, P. Andersen, M. Nielsen, C. Lundegaard, A generic method for assignment of reliability scores applied to solvent accessibility predictions, BMC Struct. Biol. 9 (2009) 51.
[13] J.Y. Wang, H.M. Lee, S. Ahmad, SVM-Cabins: prediction of solvent accessibility using accumulation cutoff set and support vector machine, Proteins 68 (1) (2007) 82–91.
[14] K. Joo, S.J. Lee, J. Lee, Sann: solvent accessibility prediction of proteins by nearest neighbor method, Proteins-structure Function & Bioinformatics 80 (7) (2012) 1791–1797.
[15] C. Fan, D. Liu, R. Huang, Z. Chen, L. Deng, PredRSA: a gradient boosted regression trees approach for predicting protein solvent accessibility, BMC Bioinf. 17 (Suppl 1) (2016) 8.
[16] J. Zhang, W. Chen, P. Sun, X. Zhao, Z. Ma, Prediction of protein solvent accessibility using PSO-SVR with multiple sequence-derived features and weighted sliding window scheme, BioData Min. 8 (2015) 3.
[17] Z. Xu, C. Zhang, S. Liu, Y. Zhou, QBES: predicting real values of solvent accessibility from sequences by efficient, constrained energy optimization, Proteins 63 (4) (2006) 961–966.
[18] Y. Yang, R. Heffernan, K. Paliwal, J. Lyons, Y. Zhou, SPIDER2: A Package to Predict Secondary Structure, Accessible Surface Area, and Main-Chain Torsional Angles by Deep Neural Networks, 2017.
[19] R. Heffernan, Y. Yang, K. Paliwal, Y. Zhou, Capturing non-local interactions by long short-term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure, backbone angles, contact numbers and solvent accessibility, Bioinformatics 33 (18) (2017) 2842–2849.
[20] R. Heffernan, K. Paliwal, J. Lyons, J. Singh, Y. Yang, Y. Zhou, Single-sequence-based prediction of protein secondary structures and solvent accessibility by deep whole-sequence learning, J. Comput. Chem. 39 (26) (2018) 2210–2216.
[21] M.S. Klausen, M.C. Jespersen, H. Nielsen, K.K. Jensen, V.I. Jurtz, C.K. Sønderby, M. O.A. Sommer, O. Winther, M. Nielsen, B. Petersen, P. Marcatili, NetSurfP-2.0: improved prediction of protein structural features by integrated deep learning, Proteins 87 (6) (2019) 520–527.
[22] J. Hanson, K. Paliwal, T. Litfin, Y. Yang, Y. Zhou, Improving prediction of protein secondary structure, backbone angles, solvent accessibility and contact numbers by

using predicted contact maps and an ensemble of recurrent and residual convolutional neural networks, Bioinformatics 35 (14) (2019) 2403–2410.
[23] S.F. Altschul, T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D. J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (17) (1997) 3389–3402.
[24] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.
[25] C. Ao, S. Jin, Y. Lin, Q. Zou, Review of progress in predicting protein methylation sites, Curr. Org. Chem. 23 (15) (2019) 1663–1670.
[26] R. Sharma, S. Kumar, T. Tsunoda, T. Kumarevel, A. Sharma, Single-stranded and double-stranded DNA-binding protein prediction using HMM profiles, Anal. Biochem. 612 (2021) 113954.
[27] J. Hanson, K. Paliwal, T. Litfin, Y. Yang, Y. Zhou, Accurate prediction of protein contact maps by coupling residual two-dimensional bidirectional long short-term memory with convolutional neural networks, Bioinformatics 34 (23) (2018) 4039–4045.
[28] S. Seemayer, M. Gruber, J. Söding, CCMpred–fast and precise prediction of protein residue-residue contacts from correlated mutations, Bioinformatics 30 (21) (2014) 3128–3130.
[29] Y.-G. Wang, S.-Y. Huang, L.-N. Wang, Z.-Y. Zhou, J.-D. Qiu, Accurate prediction of species-specific 2-hydroxyisobutyrylation sites based on machine learning frameworks, Anal. Biochem. 602 (2020) 113793.
[30] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D.S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, M. Weigt, Direct-coupling analysis of residue coevolution captures native contacts across many protein families, Proc. Natl. Acad. Sci. U. S. A. 108 (49) (2011) E1293–E1301.
[31] W. Li, A. Godzik, Cd-hit, A fast program for clustering and comparing large sets of protein or nucleotide sequences, Bioinformatics 22 (13) (2006) 1658–1659.
[32] W. Kabsch, C. Sander, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, Biopolymers 22 (12) (1983) 2577–2637.
[33] C. Chothia, The nature of the accessible and buried surfaces in proteins, J. Mol. Biol. 105 (1) (1976) 1–12.
[34] B. Zhang, L. Li, Q. Lü, Protein solvent-accessibility prediction by a stacked deep bidirectional recurrent neural network, Biomolecules 8 (2) (2018).
[35] B. Rost, C. Sander, Combining evolutionary information and neural networks to predict protein secondary structure, Proteins 19 (1) (1994) 55–72.
[36] B. Rost, C. Sander, Conservation and prediction of solvent accessibility in protein families, Protns Structure Function & Bioinformatics 20 (3) (1994) 216–226.
[37] M. Remmert, A. Biegert, A. Hauser, J. Söding, HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment, Nat. Methods 9 (2) (2011) 173–175.
[38] M. Mirdita, L. von den Driesch, C. Galiez, M.J. Martin, J. Söding, M. Steinegger, Uniclust databases of clustered and deeply annotated protein sequences and alignments, Nucleic Acids Res. 45 (D1) (2017) D170–d176.
[39] L.J. McGuffin, K. Bryson, D.T. Jones, The PSIPRED protein structure prediction server, Bioinformatics 16 (4) (2000) 404–405.
[40] J. Yang, A. Roy, Y. Zhang, Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment, Bioinformatics 29 (20) (2013) 2588–2595.
[41] S.B. Needleman, C.D. Wunsch, A general method applicable to the search for similarities in the amino acid sequence of two proteins, J. Mol. Biol. 48 (3) (1970) 443–453.
[42] A. Graves, J. Schmidhuber, Framewise phoneme classification with bidirectional LSTM and other neural network architectures, Neural Network. 18 (5–6) (2005) 602–610.
[43] M.E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep Contextualized Word Representations, 2018 arXiv preprint arXiv: 1802.05365.
[44] Z. Lin, M. Feng, C.N.d. Santos, M. Yu, B. Xiang, B. Zhou, Y. Bengio, A Structured Self-Attentive Sentence Embedding, 2017 arXiv preprint arXiv:1703.03130.
[45] J. Hu, L. Shen, S. Albanie, G. Sun, E. Wu, Squeeze-and-Excitation networks, IEEE Trans. Pattern Anal. Mach. Intell. 42 (8) (2020) 2011–2023.
[46] S. Wu, G. Li, L. Deng, L. Liu, D. Wu, Y. Xie, L. Shi, L1 -norm batch normalization for efficient training of deep neural networks, IEEE Trans Neural Netw Learn Syst 30 (7) (2019) 2043–2051.
[47] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (1) (2014) 1929–1958.
[48] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, Pytorch: an Imperative Style, High-Performance Deep Learning Library, 2019 arXiv preprint arXiv:1912.01703.
[49] D.P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization, Arxiv, 2014.
[50] J. Sim, S.Y. Kim, J. Lee, Prediction of protein solvent accessibility using fuzzy k-nearest neighbor method, Bioinformatics 21 (12) (2005) 2844–2849.
[51] B. Liu, S. Jiang, Q. Zou, HITS-PR-HHblits: Protein Remote Homology Detection by Combining PageRank and Hyperlink-Induced Topic Search, Brief Bioinform, 2018.
[52] Y. Zhang, Q. Yang, A survey on multi-task learning, IEEE Trans. Knowl. Data Eng. (2021), https://doi.org/10.1109/TKDE.2021.3070203.
[53] S. Patel, R. Tripathi, V. Kumari, P. Varadwaj, DeepInteract: deep neural network based protein-protein interaction prediction tool, Curr. Bioinf. 12 (6) (2017) 551–557.