The exploration of DNA-binding proteins (DBPs) is an important aspect of studying biological life activities. Research on life activities requires the support of scientific research results on DBPs. The decline in many life activities is closely related to DBPs. Generally, the detection method for identifying DBPs is achieved through biochemical experiments. This method is inefficient and requires considerable manpower, material resources and time. At present, several computational approaches have been developed to detect DBPs, among which machine learning (ML) algorithm-based computational techniques have shown excellent performance. In our experiments, our method uses fewer features and simpler recognition methods than other methods and simultaneously obtains satisfactory results. First, we use six feature extraction methods to extract sequence features from the same group of DBPs. Then, this feature information is spliced together, and the data are standardized. Finally, the extreme gradient boosting (XGBoost) model is used to construct an effective predictive model. Compared with other excellent methods, our proposed method has achieved better results. The accuracy achieved by our method is 78.26% for PDB2272 and 85.48% for PDB186. The accuracy of the experimental results achieved by our strategy is similar to that of previous detection methods.

Keywords: DNA-binding protein prediction, machine learning, feature extraction, dimensionality reduction, XGBoost model

INTRODUCTION

Organisms contain many macromolecular substances, such as DNA and proteins, which contain the genetic information of organisms and are important components of all cells and tissues that make up an organism. To study the life activities of cells, it is necessary to study DNA and proteins and the interaction between them. Research on DBPs has an extremely important status and significance in related life sciences and plays an important role in DNA replication and recombination, virus infection and proliferation. It is necessary to study the combination of DNA and protein to study the gene expression of organisms at the molecular level. Researchers are paying increasing attention to DBP studies. DBPs are a kind of protein that binds to DNA, and it is critical to determine which of the numerous proteins can attach to DNA (Liu et al., 2019a; Li et al., 2019; Li et al., 2020). However, the traditional use of biochemical methods to find DBP consumes considerable time and money. Based on the above requirements and the development of computer science and ML (Zheng et al., 2019; Zheng et al., 2020; Wang et al., 2021a), relevant researchers have developed many detection methods based on ML algorithms in the hopes of improving the efficiency of detecting DBP and saving manpower and material resources.
ML is frequently utilized in the fields of computational biology (Jiang et al., 2013a; Cheng et al., 2019a; Liu et al., 2019b; Wang et al., 2019; Liu et al., 2020a; Tao et al., 2020a; Wang et al., 2020a; Zhang et al., 2020a; Zhao et al., 2020a; Zhu et al., 2020; Wang et al., 2021b; Wang et al., 2021c; Dao et al., 2021; Yu et al., 2021) to analyze brain disease (Li et al., 2018a; Cheng et al., 2019b; Bi et al., 2020; Iqubal et al., 2020; Zhang et al., 2021a), IncRNA-miRNA interactions (Cheng et al., 2016; Liu et al., 2020b), and protein-protein interconnections (Ding et al., 2016a; Ding et al., 2016b; Yu et al., 2020b).

The protein sequence is very sizeable, and its number far exceeds the number of structures known to researchers (Zuo et al., 2017). Therefore, ML is used in various computer programs that predict DBP. The model IDNA-Prot[dis] (Liu et al., 2014) was proposed by Liu et al. and is used to detect DBP based on the pseudo amino acid composition (PseAAC), and it can accurately extract the characteristics of DNA binding proteins. There are two models that use PseAAC and physical-chemical distance transformation and support vector machine (SVM) algorithms, named PseDNA-Pro (Liu et al., 2015a) and iDNAPro-PseAAC (Liu et al., 2015b). Lin et al. developed the IDNA-Prot (Lin et al., 2011) prediction model based on the random forest (RF) algorithm through the PseACC feature. Kumar et al. developed two models based on RF and SVM classifiers called DNA-Prot (Kumar et al., 2009) and DNAbininder (Kumar et al., 2007). Dong et al. proposed the Kmer1+ACC (Liu et al., 2016) model based on the SVM algorithms Kmer composition and autocross covariance transformation. The position-specific scoring matrix (PSSM) can be obtained by calculating the protein sequence’s position frequency matrix, which has evolutionary information on the protein (Shao et al., 2021). The Local-DPP (Wei et al., 2017) uses the local pseudo position-specific scoring matrix (Pse-PSSM) and random forest algorithm to detect DBPs. Multiple kernel SVM is a DBP predictor from heuristically kernel alignment, and it is also named MKSVM-HKA (Ding et al., 2020a), which includes a variety of characteristics and was developed by Ding et al. The MSFBinder (Liu et al., 2018b) model proposed by Liu et al. is based on multiview features as well as classifiers. DPP-PseAAC (Rahman et al., 2018) is a model based on Chou’s general PseAAC, and it is used to detect DBPs. Methods have also been developed that combine multiscale features and deep neural networks to predict DBPs, such as MsDBP (Du et al., 2019). Adilina et al. (2019) analyzed protein sequence characteristics and implemented two different feature selection methods to build a DBP predictor.

In recent years, an increasing number of researchers have adopted complex feature extraction methods (Fu et al., 2020; Jin et al., 2021) and classification models to identify DBPs. It is critical to develop a method that uses as few DBP features as possible and includes a simple classification model while also ensuring a good ability to detect DBP. According to previous work, we proposed a DBP identification method based on the XGBoost model. First, several features were extracted from the protein sequence. Second, the features of these sequences were spliced. Third, the dimension of the data was standardized and reduced. Finally, the XGBoost model was used to detect DBPs. We have evaluated the effectiveness of our method on some benchmark data sets. Compared with some current experimental methods, our method achieves a better Matthew’s correlation coefficient (MCC), with a value of 0.713 for PDB186 and 0.5652 for PDB2272.

**METHODS**

Identifying DBPs is a common dichotomy problem. First, we used six different feature extraction models for DBPs sequences to extract the corresponding sequence feature information. Then, the sequence feature information was spliced. Next, dimensionality reduction was performed on the spliced sequence feature information. Finally, the XGBoost model was utilized to identify DBPs. Figure 1 depicts the flowchart of our adopted technique.

**Extracting Features**

To recognize DBPs, the corresponding features must be extracted. We adopt six feature extraction methods to obtain sequence information: global encoding, GE (Li et al., 2009); multi-scale continuous as well as discontinuous descriptor, MCD (You et al., 2014); normalized Moreau-Broto auto correlation, NMBAC (Ding et al., 2016b; Feng and Zhang, 2000); position specific scoring matrix-based average blocks, PSSM-AB (Jeong et al., 2011; Zhu et al., 2019); position speciﬁc scoring matrix-based discrete cosine transform, PSSM-DCT (Huang et al., 2015); and PSSM-based discrete wavelet transform, PSSM-DWT (Nanni et al., 2012). The abovementioned feature extraction models are all well-known protein sequence extraction algorithms and commonly used, which could be described in related works (Zou et al., 2021). Table 1 shows the feature dimensions derived by various feature extraction methods. After completing the above work, we used MATLAB to horizontally stitch together (Ding et al., 2020c; Ding et al., 2020d; Yang et al., 2021a) the features extracted from the same protein sequence using different feature extraction methods. The spliced features are represented by $Z'$. After splicing, the dimensions of PDB14189 and PDB2272 are 2692, and the dimensions of PDB1075 and PDB186 are 3092.

**Standardize the Data**

To make the data more standardized and unified and to strengthen the relationship between the characteristics of the data and the labels of the data, we use Z-score standardization to process the data. Z-score standardization is defined as follows:

$$M^* = \frac{Z^* - \bar{Z}}{\sigma}$$  \hspace{1cm} (1A)

$$\bar{Z} = \frac{\sum_{i=1}^{N} Z^*_i}{N}$$  \hspace{1cm} (1B)

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (Z^*_i - \bar{Z})^2}{N}}$$  \hspace{1cm} (1C)

$$i = 1, 2, \ldots, N$$  \hspace{1cm} (1D)
where N is the total number of samples and σ is the standard deviation.

The DBP sequence was processed in three stages: feature extraction, feature information splicing, and data standardization. Following the aforementioned three stages, we can obtain the sequence feature information $M^*$.

**Dimensionality Reduction by Max-Relevance-Max-Distance**

Zou et al. (Quan et al., 2016; Niu et al., 2020) developed a dimensionality reduction method in 2015 named Max-Relevance-Max-Distance (MRMD), and the user guide and complete runtime program can be obtained and downloaded from the following URL: https://github.com/heshida01/MRMD3.0. It judges data independence through a distance function and completes the dimensionality reduction operation in three steps (Tao et al., 2020b). It first evaluates each feature’s contribution to the classification and then quantifies each feature’s contribution to the classification. Second, the weights of different features are calculated for classification and the selected features are sorted accordingly. Third, the different numbers of features are filtered and classified and the results are recorded. We analyze and compare the results of the previous step to select the most effective group and use the sequence features chosen from this group as the result of dimensionality reduction.

The maximum correlation and the maximum distance are the main bases for the MRMD algorithm to judge the weight of each feature to the prediction result. The Pearson correlation coefficient can be used to quantify the degree of correlation between features and cases, and it can be calculated by the maximum relevance (MR).

The Pearson correlation coefficient is defined as follows:

$$\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$$ (2)

The $i_{th}$ characteristic from the sequence and the category label to which those sequences belong make up the vectors X and Y.

---

**TABLE 1 | Dimensional information about the features.**

| Model          | Dimensionality |
|----------------|----------------|
| GE             | 150            |
| MCD            | 882            |
| MNBAC          | 200            |
| PSSM-AB        | 200            |
| PSSM-DCT       | 399            |
| PSSM-DWT       | 1,040          |

**FIGURE 1 | Process of predicting DBPs.**
The maximum distance (MD) is used to assess feature redundancy. We calculate the three indices between characteristics in total.

$$ED(X, Y) = \sqrt[2]{\sum_{i=0}^{N} (x_i - y_i)^2} \quad (i = 1, 2, \ldots, N) \quad (3A)$$

$$\cos(X, Y) = \frac{X \cdot Y}{\|X\|\|Y\|} \quad (3B)$$

$$TC(X, Y) = \frac{X \cdot Y}{\|X\|^2 + \|Y\|^2 - X \cdot Y} \quad (3C)$$

Equations 3A, 3B, 3C represent Euclidean distance, cosine similarity and Tanimoto coefficient, respectively. We can obtain the MD value by calculating the three indicators. Finally, the classification contribution value of each feature is calculated by combining MR and MD in a specific ratio.

After dimensionality reduction, the dimensions of PDB14189 and PDB2272 are 379, and the dimensions of PDB1075 and PDB186 are 1460.

Based on the three steps of feature extraction and splicing, data standardization and dimensionality reduction operations, we obtain the final sequence features.

**Extreme Gradient Boosting Algorithm**

In 2011, Tianqi Chen and Carlos Guestrin (Chen and Guestrin, 2016) first proposed the XGBoost algorithm, or the extreme gradient boosting algorithm. It is a machine learning model that achieves a stronger learning effect by integrating multiple weak learners. The XGBoost model has many advantages, such as strong flexibility and scalability (Yang et al., 2021b; Zhang et al., 2021b).

Generally, most boosting tree models have difficulty implementing distributed training because when training $n$ trees, they will be affected by the residuals of the first $n-1$ trees and only use first-order derivative information. The XGBoost model is different. It performs a second-order Taylor expansion of the loss function and uses a variety of methods to prevent overfitting as much as possible. XGBoost can also automatically use the CPU’s multithreaded parallel computing to speed up the running speed. This feature represents a great advantage of XGBoost over other methods. XGBoost has improved significantly in terms of effect and performance.

The XGBoost algorithm is described in detail as follows:

\[
\hat{y}_i = \sum_{m=1}^{M} f_m(x_i), f_m \in F
\]

where $M$ is the number of trees and $F$ represents the basic model of the trees.

The objective function is defined as follows:

\[
L = \sum_i I(\hat{y}_i, y_i) + \sum_m \Omega(f_m)
\]

The error between the predicted value and the true value is represented by the loss function $l$, and the regularized function $\Omega$ to prevent overfitting is defined as follows:

\[
\Omega(f) = \gamma T + \frac{1}{2} \lambda \|w\|^2
\]
\[ \text{Spec} = \frac{TN}{TN + FP} \]  
\[ \text{SN} = \frac{TP}{TP + FN} \]  
\[ \text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}} \]  
\[ \text{ACC} = \frac{TP + TN}{TP + TN + FP + FN} \]

Among them, TN, TP, FP and FN reflect the values of true negatives, true positives, false positives, and false negatives, respectively.

**Performance Analysis**

On the PDB 1075 data set, the performance of the spliced sequence features and single sequence features is evaluated by randomly extracting 30% of the data as a test set. Figure 2; Table 3 depict the experimental outcomes. PSSM-DWT (MCC: 0.4981) achieved better performance than other single sequence features. The spliced sequence features perform better than the single sequence feature on all parameters. The spliced sequence feature (ROC: 0.81) also gained the best ROC performance.

**Independent Data Set of PDB186**

In this experiment, different sequence features have different prediction performances. We use PDB1075 as the training set and PDB186 as the test set to evaluate our experimental method and compared the experimental findings of our approach to those of 13 other methods. Table 4 clearly shows the complete experimental outcomes.

The MCC values of the five methods are all above 0.6 for MSDBP, MSFBinder, Local-DPP MKSVM-HKA, and Adilina’s work (0.606, 0.616, 0.625, 0.648 and 0.670, respectively). Thus, these methods have excellent performance. Although Adilina’s work (SN: 95.0%) performs best in terms of the value of SN, the results of XGBoost achieve optimal ACC (85.48%), MCC (0.713) and Spec (80.6%). On PDB1075 and PDB186, XGBoost outperforms the other methods.

**Experimental Results With PDB2272 and PDB186 as Test Set**

We combined PDB14189 and PDB1075 as the training set, and combined PDB2272 and PDB186 as the test set. After normalization and dimensionality reduction operations, we got an accuracy of 79.09% and the MCC value was 0.5818. It can be seen that this result is between the previous two experimental results.

**DISCUSSION AND CONCLUSION**

This paper proposes a method of predicting DBPs using the XGBoost algorithm and by splicing sequence feature information. The final sequence feature is built from multiple sequence features and spliced by MATLAB. To make the data more standardized and strengthen the relationship between data characteristics and data tags, the data are processed using Z-Score standardization. During the experiment, we used MRMD to reduce the dimensionality of the data and thus reduce the characteristics of the data. We
performed experiments and compared the performance of XGBoost in terms of single sequence feature information and spliced sequence feature information. On the PDB 1075 data set, performance of the spliced sequence feature (MCC: 0.7272) is obviously better than that of the single sequence feature. To further assess our method, we applied the XGBoost model to the PDB186 and PDB2272 data sets. XGBoost produced superior results for PDB186 (MCC: 0.713) and PDB2272 (MCC: 0.5652) compared to available methods.

**DATA AVAILABILITY STATEMENT**

The original contributions presented in the study are included in the article/supplementary material. Further inquiries can be directed to the corresponding authors.

**AUTHOR CONTRIBUTIONS**

ZZ and WY designed, planned and implemented the experiment. ZZ also wrote the main part of the article, and YXZ wrote other parts of the article. YL and YMZ participated in the coordination of the study and reviewed the article. All authors read and approved the final article.

**FUNDING**

This work was supported by the National Natural Science Foundation of China (61971119), The Heilongjiang Postdoctoral Fund (LBH-Q20135). The National Natural Science Foundation of China (NSFC) is a sub ministerial institution in charge of NSFC. NSFC operates relatively independently, and is responsible for the organization and implementation of funding plans, project setting and evaluation, project approval, supervision, etc.

**REFERENCES**

Adilina, S., Farid, D. M., and Shatabda, S. (2019). Effective DNA Binding Protein Prediction by Using Key Features via Chou’s General PseAAC. J. Theor. Biol. 460, 64–78. doi:10.1016/j.jtbi.2018.10.027

Bi, X.-a., Liu, Y., Xie, Y., Hu, X., and Jiang, Q. (2020). Morbigenous Brain Region and Gene Detection with a Genetically Evolved Random Neural Network and spliced sequence feature information. On the PDB 1075 data set, performance of the spliced sequence feature (MCC: 0.7272) is obviously better than that of the single sequence feature. To further assess our method, we applied the XGBoost model to the PDB186 and PDB2272 data sets. XGBoost produced superior results for PDB186 (MCC: 0.713) and PDB2272 (MCC: 0.5652) compared to available methods.

**DATA AVAILABILITY STATEMENT**

The original contributions presented in the study are included in the article/supplementary material. Further inquiries can be directed to the corresponding authors.

**AUTHOR CONTRIBUTIONS**

ZZ and WY designed, planned and implemented the experiment. ZZ also wrote the main part of the article, and YXZ wrote other parts of the article. YL and YMZ participated in the coordination of the study and reviewed the article. All authors read and approved the final article.

**FUNDING**

This work was supported by the National Natural Science Foundation of China (61971119), The Heilongjiang Postdoctoral Fund (LBH-Q20135). The National Natural Science Foundation of China (NSFC) is a sub ministerial institution in charge of NSFC. NSFC operates relatively independently, and is responsible for the organization and implementation of funding plans, project setting and evaluation, project approval, supervision, etc.

**REFERENCES**

Adilina, S., Farid, D. M., and Shatabda, S. (2019). Effective DNA Binding Protein Prediction by Using Key Features via Chou’s General PseAAC. J. Theor. Biol. 460, 64–78. doi:10.1016/j.jtbi.2018.10.027

Bi, X.-a., Liu, Y., Xie, Y., Hu, X., and Jiang, Q. (2020). Morbigenous Brain Region and Gene Detection with a Genetically Evolved Random Neural Network

| TABLE 3 | Performance of PDB1075 using different feature extraction methods in XGBoost. |
| --- | --- | --- | --- | --- |
| Model name | Feature extraction method | ACC (%) | SN (%) | MCC | Spec (%) |
| GE | 66.87 | 71.17 | 0.3342 | 62.09 |
| MCD | 69.04 | 70.00 | 0.3975 | 67.97 |
| MMAC | 72.14 | 75.29 | 0.4404 | 68.62 |
| XGboost | PSSM-AB | 76.47 | 75.29 | 0.5300 | 77.77 |
| PSSM-Pse | 74.30 | 75.88 | 0.4845 | 72.54 |
| XGboost | The spliced sequence feature | 81.42 | 84.11 | 0.6272 | 78.43 |

Bold indicates that their experimental results are the best and the experimental values are the highest.

| TABLE 4 | Comparison between the XGBoost model and other methods on the PDB186 data set. |
| --- | --- | --- | --- | --- |
| Models | ACC (%) | SN (%) | Spec (%) | MCC |
| IDNA-Pro|dis | 72.0 | 79.5 | 64.5 | 0.445 |
| IDNA-Pro | 67.2 | 67.7 | 66.7 | 0.344 |
| DNA-Prot | 61.8 | 69.9 | 53.8 | 0.240 |
| DNA predictor | 60.8 | 57.0 | 64.5 | 0.216 |
| DBP|Pre | 76.9 | 79.6 | 74.2 | 0.538 |
| IDNAPro-PseAAC | 71.5 | 82.8 | 60.2 | 0.442 |
| Kmerl + ACC | 71.0 | 82.8 | 59.1 | 0.431 |
| Local-DPP | 79.0 | 92.5 | 65.6 | 0.625 |
| DPP-PseAAC | 77.4 | 83.0 | 70.9 | 0.550 |
| MSF|Binder | 79.6 | 93.6 | 65.6 | 0.616 |
| MsDBP | 80.1 | 86.0 | 74.2 | 0.606 |
| MKSVM-HKA | 81.2 | 94.6 | 67.7 | 0.648 |
| Adlina’s work | 82.3 | 95.0 | 69.9 | 0.670 |
| XGboost | 85.48 | 90.3 | 80.6 | 0.713 |

Bold indicates that their experimental results are the best and the experimental values are the highest.

The experimental results of other methods come from (Wei et al., 2017).

| TABLE 5 | Experimental findings for the independent data set PDB2272 using the XGBoost algorithm and other models. |
| --- | --- | --- | --- | --- |
| Methods | ACC (%) | MCC | SN (%) | Spec (%) |
| MK-FSVM-SVDD | 76.12 | 0.5476 | 91.50 | 60.41 |
| DPP-PseAAC | 58.10 | 0.1625 | 56.83 | 49.61 |
| PseDNA-Pro | 61.88 | 0.2430 | 75.28 | 48.08 |
| MK-SVM | 75.00 | 0.5264 | 91.41 | 58.99 |
| MsDBP | 66.99 | 0.2397 | 70.69 | 43.18 |
| XGboost | 78.26 | 0.5652 | 80.39 | 76.05 |

Bold indicates that their experimental results are the best and the experimental values are the highest.

The experimental results of other methods come from (Du et al., 2019; Zou et al., 2021).
Zhang, D., Chen, H. D., Zulfiquar, H., Yuan, S. S., Huang, Q. L., Zhang, Z. Y., et al. (2021). iBLP: An XGBoost-Based Predictor for Identifying Bioluminescent Proteins. Comput. Math. Methods Med. 2021, 6664362. doi:10.1155/2021/6664362

Zhong, J., Zhang, Z., Pu, L., Tang, J., and Guo, F. (2020). AEPrep: an Ensemble Predictive Model of Classifier Chain to Identify Anti-inflammatory Peptides. IEEE/ACM Trans. Comput. Biol. Bioinformatics 18, 1831. doi:10.1109/TCBB.2020.2968419

Zhong, Z., Ding, J., Xu, J., Tang, J., and Guo, F. (2021). Multi-Scale Time-Series Kernel-Based Learning Method for Brain Disease Diagnosis. IEEE J. Biomed. Health Inform. 25 (1), 209–217. doi:10.1109/jbhi.2020.2983456

Zhao, T., Hu, Y., Peng, J., and Cheng, L. (2020). DeepLGP: a Novel Deep Learning Method for Prioritizing lncRNA Target Genes. Bioinformatics 36, 4466. doi:10.1093/bioinformatics/btaa428

Zhao, X., Jiao, Q., Li, H., Wu, Y., Wang, H., Huang, S., et al. (2020). ECFS-DEA: an Ensemble Classifier-Based Feature Selection for Differential Expression Analysis on Expression Profiles. BMC Bioinformatics 21 (1), 43. doi:10.1186/s12859-020-3388-y

Zhao, X., Wang, H., Li, H., Wu, Y., and Wang, G. (2021). Identifying Plant Pentatricopeptide Repeat Proteins Using a Variable Selection Method. Front. Plant Sci. 12, 506681. doi:10.3389/fpls.2021.506681

Zheng, L., Huang, S., Mu, N., Zhang, H., Zhang, J., Chang, Y., et al. (2019). RAACBook: a Web Server of Reduced Amino Acid Alphabet for Sequence-dependent Inference by Using Chou’s Five-step Rule. Database (Oxford) 2019, baz131. doi:10.1093/database/baz131

Zheng, L., Liu, D., Yang, W., Yang, L., and Zuo, Y. (2020). RaaCLogo: a New Sequence Logo Generator by Using Reduced Amino Acid Clusters. Brief Bioinform 22, bbaa096. doi:10.1093/bib/bbaa096

Zhu, Y., Li, F., Xiang, D., Akutsu, T., Song, J., and Jia, C. (2020). Computational Identification of Eukaryotic Promoters Based on Cascaded Deep Capsule Neural Networks. Brief. Bioinform. 22, bbaa299. doi:10.1093/bib/bbaa299

Zou, Y., Wu, H., Guo, X., Peng, L., Ding, Y., Tang, J., et al. (2021). MK-FSVM-SVDD: A Multiple Kernel-Based Fuzzy SVM Model for Predicting DNA-Binding Proteins via Support Vector Data Description. Curr. Bioinformatics 16 (2), 274–283. doi:10.2174/1574893615999200607173829

Zhao, X., Li, Y., Chen, Y., Li, G., Yan, Z., and Yang, L. (2017). PseKRAAC: a Flexible Web Server for Generating Pseudo K-Tuple Reduced Amino Acids Composition. Bioinformatics 33 (1), 122–124. doi:10.1093/bioinformatics/btw564

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Zhao, Yang, Zhai, Liang and Zhao. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.