frontiers
in Aging Neuroscience

ORIGINAL RESEARCH
published: 24 January 2022
doi: 10.3389/fnagi.2021.799500

An Iterative Method for Predicting Essential Proteins Based on Multifeature Fusion and Linear Neighborhood Similarity

Xianyou Zhu[1†], Yaocan Zhu[2†], Yihong Tan[2], Zhiping Chen[1,2] and Lei Wang[2*]*

[1] College of Computer Science and Technology, Hengyang Normal University, Hengyang, China, [2] College of Computer Engineering and Applied Mathematics, Changsha University, Changsha, China

OPEN ACCESS

Edited by:
Min Tang,
Jiangsu University, China

Reviewed by:
Weinan Zhou,
University of Illinois
at Urbana-Champaign, United States
Qianqian Song,
Wake Forest School of Medicine,
United States

***Correspondence:**
Yaocan Zhu
1930204639@qq.com
Lei Wang
wanglei@xtu.edu.cn

†These authors share first authorship

Specialty section:
This article was submitted to
Neurocognitive Aging and Behavior,
a section of the journal
Frontiers in Aging Neuroscience

Received: *21 October 2021*
Accepted: *02 December 2021*
Published: *24 January 2022*

Citation:
Zhu X, Zhu Y, Tan Y, Chen Z and Wang L (2022) An Iterative Method for Predicting Essential Proteins Based on Multifeature Fusion and Linear Neighborhood Similarity. Front. Aging Neurosci. 13:799500. doi: 10.3389/fnagi.2021.799500

Growing evidence have demonstrated that many biological processes are inseparable from the participation of key proteins. In this paper, a novel iterative method called linear neighborhood similarity-based protein multifeatures fusion (LNSPF) is proposed to identify potential key proteins based on multifeature fusion. In LNSPF, an original protein-protein interaction (PPI) network will be constructed first based on known protein-protein interaction data downloaded from benchmark databases, based on which, topological features will be further extracted. Next, gene expression data of proteins will be adopted to transfer the original PPI network to a weighted PPI network based on the linear neighborhood similarity. After that, subcellular localization and homologous information of proteins will be integrated to extract functional features for proteins, and based on both functional and topological features obtained above. And then, an iterative method will be designed and carried out to predict potential key proteins. At last, for evaluating the predictive performance of LNSPF, extensive experiments have been done, and compare results between LNPSF and 15 state-of-the-art competitive methods have demonstrated that LNSPF can achieve satisfactory recognition accuracy, which is markedly better than that achieved by each competing method.

Keywords: key protein, entropy, linear neighborhood similarity, iterative method, multi-feature fusion

INTRODUCTION

In the past few years, with the development of high-throughput and bioinformatics technologies, recognition of potential key proteins based on protein-protein interaction (PPI) networks has become a new research hotspot (Dai et al., 2021; Zhang et al., 2021). Essential proteins play an important role in cell growth and regulation, and researches on essential proteins can deepen the understanding of biological life processes. Existing key protein prediction methods can be roughly divided into two categories: one is based on the topological characteristics of PPI networks and the other is based on the fusion of topological structures of PPI networks and biological information of protein such as the gene expression data, the subcellular localization data, the homologous data, and the gene ontology of protein. For example, based on topological characteristics of PPI networks, Li et al. (2015) proposed a method called LAC, in which, the local average connectivity of nodes in the PPI network was adopted to estimate the essentiality of proteins. Qi and Luo (2016) introduced a model named LID by measuring the importance of proteins by the local interaction density between neighboring nodes in the PPI network. Lin designed two predictive models called MNC (maximum neighborhood connectivity) and DMNC (density of maximum neighborhood connectivity) based on the maximum neighborhood connectivity and density of maximum neighborhood connectivity of modes in the PPI network separately (Lin et al., 2011). In addition, researchers have proposed a series of methods to identify key proteins based on the centrality of nodes in PPI networks, such as DC (degree centrality) (Hahn and Kern, 2005), EC (eigenvector centrality) (Bonacich, 1987), CC (closeness centrality) (Wuchty and Stadler, 2003), IC (information centrality) (Stephenson and Zelen, 1989), SC (subgraph centrality) (Estrada and Rodríguez-Velázquez, 2005), BC (betweenness centrality) (Joy et al., 2005), and NC (neighbor centrality) (Wang et al., 2012). In all these methods, since only topological characteristics of PPI networks were considered, then unknown interactions between proteins might greatly affect the identification accuracy of potential key proteins. Hence, to improve the recognition accuracy, some other methods based on the fusion of biological information and topological features were proposed successively. For instance, Tang and Li proposed two methods called WDC (weighted degree centrality) (Tang et al., 2014) and PEC (integration ECC and Pearson correlation) (Li et al., 2012), respectively, by fusing topological features of PPI networks with gene expression information of proteins to measure the importance of proteins. Peng et al. (2012) designed two methods, namely, UDoNC (united the domain features and the normalized ECC) and ION (integration of the properties of orthologous and the features of neighbors) (Peng et al., 2015a), through combining homology and domain information of proteins with topological features of PPI networks separately. Zhang et al. (2013) introduced a prediction model called CoEWC by integrating topological characteristics of PPI networks with co-expression characteristics of proteins in gene expression profiles. Li et al. (2018) proposed a method named subnetwork partition and prioritization by fusing subcellular localization information of proteins with PPI networks. Zhao et al. (2019) designed an iterative computing method called RWHN by combining homology, domain, and subcellular localization information of proteins with topological features of PPI networks. Zhao et al. (2014) proposed a prediction method called POEM by integrating gene expression data of proteins and topology features of PPI networks. Lei et al. (2020) designed a method based on gene expression data and Drosophila optimization algorithm (FOCA), which combines PPI network, subcellular localization, gene ontology annotation, gene expression data, and artificial fish swarm optimization (AFSO) algorithm (Lei et al., 2016) to predict key proteins. In addition, a prediction method based on the combination of a learning system and specific scoring matrix was proposed by Wang (Wang et al., 2017), and a prediction method based on the deep learning model proposed by Chen (Chen et al., 2019). Chen et al. (2020) proposed an identification method called NPRI by integrating heterogeneous networks. Dai et al. (2020) identified key proteins based on PPI network embedding. Zhang et al. (2019) proposed a method by fusing dynamic PPI networks. Sun et al. (2021) designed an iterative method called IoMCD (iteration based on multiple characteristic differences) based on cross-entropy. Li et al. (2020) proposed an iterative method called CVIM (character vector iteration method) based on the fusion of topological structures of PPI networks and functional characteristics of proteins.

Experimental results show that the fusion of network topological features and biological information of proteins can improve the accuracy of identifying potential key proteins effectively. However, in most existing methods, due to the limited categories of topological structures of PPI networks and functional characteristics of proteins fused, the predictive performances of these methods are not satisfactory. Hence, in this study, through combining a series of topological features of PPI networks and abundant biological information of proteins, a new predictive method called LNSPF (linear neighborhood similarity-based protein multifeatures fusion) is proposed to identify potential key proteins. In LNSPF, an original PPI network will be constructed first based on known PPI data downloaded from benchmark databases, and then, topological features will be extracted from the original PPI network. Next, the protein nodes in the original PPI network are defined as data points, the protein gene expression data are defined as the characteristics of the corresponding data points, and the data points are reconstructed to calculate the linear neighborhood similarity between the data points in the feature space. After that, subcellular location and homologous information of proteins will be integrated to extract functional features for proteins. At last, based on both functional and topological features extracted above, an iterative method will be designed to predict key proteins. Experimental results show that LNSPF can achieve reliable prediction accuracies of 100%, 90%, and 87% in top 1%, 5%, and 10% ranked key proteins separately based on the GAVIN database, which is markedly superior to 15 state-of-the-art competitive methods, namely, DC (Hahn and Kern, 2005), CC (Wuchty and Stadler, 2003), IC (Stephenson and Zelen, 1989), SC (Estrada and Rodríguez-Velázquez, 2005), BC (Joy et al., 2005), NC (Wang et al., 2012), PEC (Li et al., 2012), LAC (Li et al., 2015), COEWC (Zhang et al., 2013), POEM (Zhao et al., 2014), ION (Peng et al., 2015a), TEGS (Li et al., 2018), RWHN (Zhao et al., 2019), IoMCD (Sun et al., 2021), and CVIM (Li et al., 2020) simultaneously.

MATERIALS AND METHODS

As shown in **Figure 1**, the process of LNSPF consists of the following four main steps:

Step 1: First, based on known PPI data downloaded from the benchmark database, an original PPI network is constructed, from which, topological features, namely,

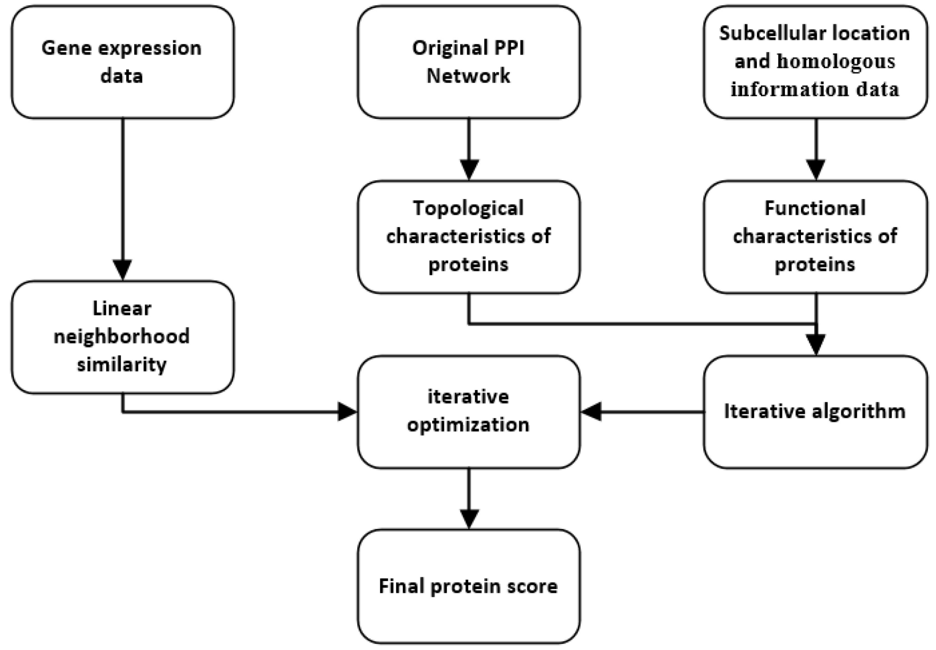

Gene expression data
Original PPI Network
Subcellular location and homologous information data
Topological characteristics of proteins
Functional characteristics of proteins
Linear neighborhood similarity
iterative optimization
Iterative algorithm
Final protein score

FIGURE 1 | Flowchart of the LNSPF.

degree, two hops degree, and triangle are extracted successively.

Step 2: Next, subcellular location and homologous information of proteins will be integrated to extract functional features for proteins.

Step 3: Moreover, based on the topological and biological properties obtained above, an iterative method is designed to estimate the importance of proteins.

Step 4: At last, based on the gene expression data downloaded from the benchmark database, the score was further optimized by using linear neighborhood similarity.

Extraction of Functional Features for Proteins

Let $G = (V, E)$ denote the original PPI network constructed from a dataset of known PPIs downloaded from any given benchmark database D, $V = \{p_1, p_2, \cdots p_N\}$ represent a set of different proteins, and $E = \{e(p_i, p_j) \mid p_i, p_j \in V\}$ represent a collection of edges between proteins in G. Here, if and Based a known interaction between any two given proteins in V, there is a side $e(p_i, p_j)$ between them. Obviously, based on the original PPI network G, we can obtain a $N \times N$ dimensional adjacency matrix $A = (a_{ij})_{N\times N}$,where there is $a_{ij} = 1$, if and only if there is an edge $e(p_i, p_j)$ between p_i and p_j, otherwise, there is $a_{ij} = 0$.

For any given protein p_i in G, let $NG(p_i)$ denote the set of nodes neighboring to p_i in G, then it is obvious that there is:

$$NG(p_i) = \{p_j \mid \exists\, e(p_i, p_j) \in E, p_j \in V\} \tag{1}$$

According to Equation 1, it is easy to know that the nodes in $NG(p_i)$ are one-hop from p_i in G, for convenience, we define $NG(p_i)$ as the set of one-hop neighbors of p_i in G, based on which, we can obtain a new set of two-hops neighbors of p_i in G as follows:

$$THNG(p_i) = \{p_j \mid \exists\, e(p_j, p_k) \in E, p_k \in NG(p_i)\} \tag{2}$$

Where $|NG(p_i)|$ denotes the number of different nodes in the set $NG(p_i\ )$.

According to Equations 1, 2, based on the fact that key proteins and their neighbors often form tight junction clusters (Li et al., 2015; Peng et al., 2015a), we can define two kinds of topological properties for any given protein p_i in G as follows:

$$TP_1(p_i) = \sum_{p_j \in NG(p_i)} TZ_1(p_i, p_j) \tag{3}$$

$$TP_2(p_i) = \sum_{p_j \in NG(p_i)} TZ_2(p_i, p_j) \tag{4}$$

Where,

$$TZ_1(p_i, p_j) = \begin{cases} \dfrac{|NG(p_i) \cap NG(p_j)|}{|NG(p_i)|}; & p_j \in NG(p_i) \\ 0; & \text{otherwise} \end{cases} \tag{5}$$

$$TZ_2(p_i, p_j) = \begin{cases} \dfrac{|THNG(p_i) \cap NG(p_j)|}{|THNG(p_i)|}; & p_j \in NG(p_i) \\ 0; & \text{otherwise} \end{cases} \tag{6}$$

From observing Equations 3, 4, it can be seen that, for any two given proteins p_i and p_j in G, the more the number of common one-hop or two-hops neighboring nodes between them, the bigger the values of $TZ_1\left(p_i, p_j\right)$ and $TZ_2(p_i, p_j)$ will be. Hence, it is obvious that $TZ_1\left(p_i, p_j\right)$ and $TZ_2(p_i, p_j)$ can to a certain extent reflect the tightness and the aggregation degree between p_i and p_j, respectively.

Extraction of Functional Features for Proteins

Key proteins tend to connect with each other rather than exist independently, and the key of proteins is usually expressed through protein complexes or functional modules, rather than a single protein (Min et al., 2017). Existing studies have shown that key proteins are closely related to the subcellular structures of proteins (Peng et al., 2015b; Li et al., 2016; Fan et al., 2017). In this section, we will adopt the subcellular locations to extract functional features for proteins. First, for any given protein p_i, let $Sub\left(p_i\right)$ denote the set of different subcellular locations relating to p_i, and $\left|Sub\left(p_i\right)\right|$ represent the number of different elements in $Sub\left(p_i\right)$, then, we can calculate one kind of functional property for $\mathrm{p_i}$ as follows:

$$FP_1(p_i) = \frac{\sum_{p_j \in NG(p_i)} TZ_3\left(p_i, p_j\right)}{\left|NG\left(p_i\right)\right|} + 1 \tag{7}$$

Where,

$$TZ_3(p_i, p_j) = \begin{cases} \dfrac{\left|Sub\left(p_i\right) \cap Sub\left(p_j\right)\right|^2}{\left|Sub\left(p_i\right)\right| * \left|Sub(p_j)\right|}; \\ \qquad\qquad \left|Sub\left(p_i\right)\right| * \left|Sub(p_j)\right| > 0 \\ 0; \qquad otherwise \end{cases} \tag{8}$$

In addition, in the study of Peng et al. (2012), key proteins were proved to be relatively conserved. Through whether each protein has homology, the homology score of each protein is obtained to indicate the degree of conservation of each protein. Based on the homology information of proteins, for any given protein p_i, let $os(p_i)$ denote the homology fraction of p_i, then we can obtain another kind of functional property for $\mathrm{p_i}$ as follows:

$$FP_2(p_i) = \frac{os(p_i)}{\underset{p_j \in V}{max}\left\{os(p_j)\right\}} \tag{9}$$

Construction of Linear Neighborhood Similarity-Based Protein Multifeatures Fusion

Initial Iteration

For generality, supposing that we have extracted M_1 different topological features (such as TP_1, TP_2,. . ., TP_{M_1}) and M_2 different functional features (such as FP_1, FP_2,. . .,FP_{M_2}), moreover, there is $M_1 + M_2 = M$, then, for any given protein p_i, we can construct a feature vector for it as follows:

$$\begin{aligned} V_i &= < TP_1, TP_2, \ldots, TP_{M_1}, FP_1, FP_2, \ldots, FP_{M_2} > \\ &= < P_1, P_2, \ldots, P_M > \end{aligned} \tag{10}$$

Based on Equation 10, we can further obtain a feature matrix for all N proteins in G as follows:

$$Z = [\, V_1 \cdots V_N \,]^T = \left[z_{ij}\right]_{N \times M} \tag{11}$$

Based on Equation 11, it is obvious that we can adopt entropy to measure the weight of each feature in all M different features as follows:

$$w_j = (1 - e_j) / \sum_{i=1}^{M} (1 - e_i) \tag{12}$$

Where,

$$e_j = -\sum_{i=1}^{N} z_{ij} \ln z_{ij} / \ln N \tag{13}$$

TABLE 1 | A brief description of the existing representative prediction models.

Algorithm	Network topology	Biological information	Particular year
DC (Hahn and Kern, 2005)	Degree centrality	NO	2005
EC (Bonacich, 1987)	Eigenvector centrality	NO	1987
CC (Wuchty and Stadler, 2003)	Closeness centrality	NO	2003
IC (Stephenson and Zelen, 1989)	Information centrality	NO	1989
SC (Estrada and Rodríguez-Velázquez, 2005)	Subgraph centrality	NO	2005
BC (Joy et al., 2005)	Betweenness centrality	NO	2005
NC (Wang et al., 2012)	Neighbor centrality	NO	2012
PEC (Li et al., 2012)	Edge clustering coefficient	Gene expression data	2012
LAC (Li et al., 2015)	Degree centrality, common neighbor node	NO	2011
CoEWC (Zhang et al., 2013)	Clustering coefficient	Gene expression data	2013
POEM (Zhao et al., 2014)	Degree centrality, subgraph, edge clustering coefficient, closeness centrality	Gene expression data	2014
ION (Peng et al., 2015a)	Edge clustering coefficient	Orthologous data	2012
TEGS (Li et al., 2018)	Subnetwork partition and prioritization	subcellular localization data	2018
RWHN (Zhao et al., 2019)	Degree centrality, protein-domain	Orthologous data, subcellular localization	2019
IoMCD (Sun et al., 2021)	Common neighbor node, degree Centrality	Gene expression data, orthologous data	2021
CVIM (Li et al., 2020)	Degree centrality, common neighbor node	Gene expression data, orthologous data	2020

Moreover, according to Equation 13, we can further calculate the feature-based score of p_i for any given protein as follows:

$$CScore(p_i) = \sum_{j=1}^{M} w_j Z_{ij} \tag{14}$$

Based on Equation 14, we can construct a new matrix H as follows:

$$H_{ij} = \begin{cases} \dfrac{CScore(p_i)}{\sum_{l=1}^{N} CScore(p_l)}; & if\ i = j \\ \dfrac{min\left\{CScore(p_i), CScore(p_j)\right\}}{\sum_{l=1}^{N} CScore(l)}; & else \end{cases} \tag{15}$$

Hence, according to Equation 15, we can obtain stable scores for all proteins in an iterative way as follows:

$$Y^{t+1} = \alpha H Y^t + (1 - \alpha)\, Y^0 \tag{16}$$

Where the parameter $\alpha \in (0,\ 1)$and $Y^0 =< FP_2\left(p_1\right), FP_2\left(p_2\right), \ldots, FP_2(p_N) >$ is the vector consisting of initial scores of all proteins. Moreover, for convenience, we define the final stable scores obtained by Equation 16 as Y^{Final}.

Further Optimization

Proteins can be considered as data points in the feature space, and how to predict the similarity between potential essential proteins in the feature space is very important for the prediction of essential proteins. Wang and Zhang (2008) found that every data point in a high-dimensional space can be reconstructed by its neighbors. Zhang et al. (2017) proposed a new similarity measure to predict drug side effects based on characteristics of drugs. Hence, based on above concepts, in this section, we will first define protein nodes in the original PPI network as data points, and the gene expression data of proteins as features of corresponding data points. And for convenience, for any given protein p_i, let $g_i =< g_{i1}, g_{i2}, \ldots, g_{i36} >$ represent its gene expression data, where g_{it} represents the gene expression level of p_i at the tth time point, then, we can further reconstruct each data point p_i based on features of its neighbors by minimizing the following reconstruction error ε_i:

$$\begin{aligned} \varepsilon_i &= \left\| g_i - \sum_{p_j \in NG(p_i)} s_{i,j} g_j^2 \right\| + \|s_i^2\| \\ &= \left\| \sum_{p_j \in NG(p_i)} s_{i,j}\left(g_i - g_j\right)^2 \right\| + \sum_{p_j \in NG(p_i)} \left(s_{i,j}\right)^2 \\ &= \sum_{p_j, p_k \in NG(p_i)} s_{i,j} s_{i,k}\left(g_i - g_j\right)^T\left(g_i - g_j\right) + \sum_{p_j \in NG(p_i)} \left(s_{i,j}\right)^2 \\ &= \sum_{p_j, p_k \in NG(p_i)} s_{i,j}\left(G^i + I\right) s_{i,k} \\ &= s_i^T\left(G^i + I\right) s_i \\ s.t. \quad & \sum_{p_j \in NG(p_i)} s_{i,j} = 1, s_{i,j} \geq 0 \end{aligned} \tag{17}$$

Here, $G^i = (g_i - g_j)^T(g_i - g_j)$, $s_i = (s_{i,1},\ s_{i,2} \cdots s_{i,k})^T$, $\|g_i - \sum_{p_j \in NG(p_i)} s_{i,j} g_j^2\|$ is the item of reconstruction error, $\|s_i^2\|$ is used for regularization and I is the identity matrix.

Obviously, according to Equation 17, let $S_{i,j} = \begin{cases} s_{i,j}: & if\ p_j \in NG\left(p_i\right) \\ 1: & i = j \\ 0: & otherwise \end{cases}$, then we can obtain a $N \times N$-dimensional similarity matrix S as follows:

$$S = \begin{bmatrix} S_{11} & \cdots & S_{1N} \\ \vdots & \ddots & \vdots \\ S_{N1} & \cdots & S_{NN} \end{bmatrix} \tag{18}$$

In addition, for any given protein node p_i in G, we can calculate the similarity $s_{i,j}$ between it and its neighboring node

TABLE 2 | Influence of parameter α on the effect of initial iteration algorithm in Gavin database.

	α								
Rank	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
Top1% (19)	16	17	18	18	18	**18**	17	15	15
Top5% (93)	75	80	83	83	82	80	80	78	79
Top10% (186)	147	155	156	159	162	162	**163**	160	161
Top15% (278)	198	205	213	219	218	**220**	220	219	217
Top20% (371)	249	259	264	268	271	267	**274**	278	272
Top25% (464)	303	306	309	314	317	**322**	322	320	321

The bold values represent the best predictive performance achieved by LNSPF under different conditions.

TABLE 3 | Effect of parameter β on prediction performance of LNSPF in Gavin database.

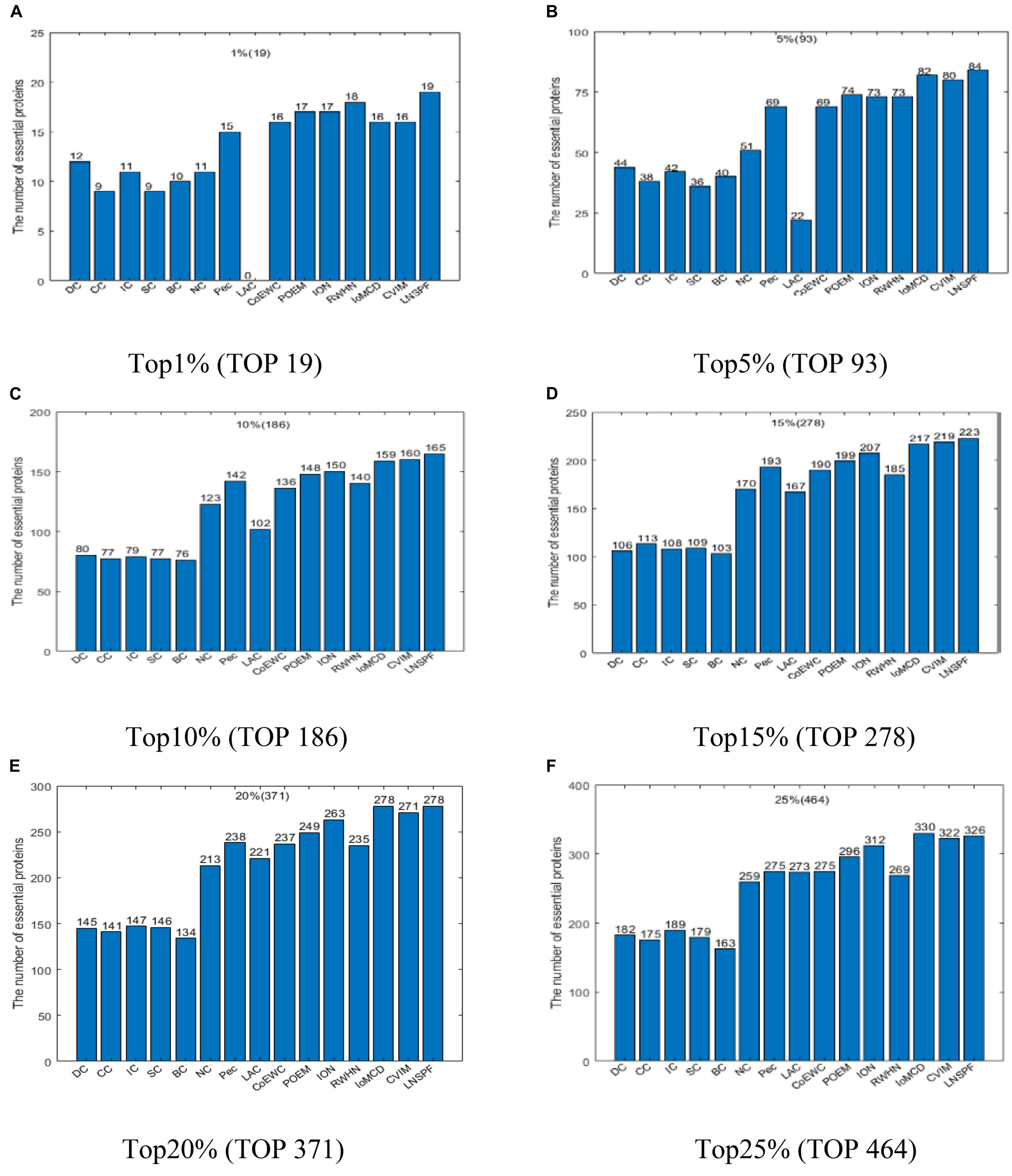
	β								
Rank	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
Top1% (19)	18	**19**	**19**	**19**	18	18	18	18	17
Top5% (93)	81	83	83	**84**	82	82	82	81	78
Top10% (186)	164	**165**	164	163	164	166	164	163	161
Top15% (278)	221	**223**	221	**223**	222	219	220	219	210
Top20% (371)	271	274	274	**278**	274	272	272	270	262
Top25% (464)	324	324	325	**326**	325	321	319	314	310

The bold values represent the best predictive performance achieved by LNSPF under different conditions.

TABLE 4 | Effect of parameter β on prediction performance of LNSPF based on DIP database.

	β								
Rank	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
Top1% (51)	46	**47**	47	46	46	46	44	44	43
Top5% (255)	203	**208**	205	203	203	200	198	197	189
Top10% (510)	347	**352**	350	**352**	**352**	349	342	334	330
Top15% (764)	468	468	467	**469**	467	459	457	458	429
Top20% (1019)	547	546	544	**548**	547	542	542	535	519
Top25% (1274)	626	**630**	628	625	622	622	623	615	608

The bold values represent the best predictive performance achieved by LNSPF under different conditions.



Top1% (TOP 19)

Top5% (TOP 93)

Top10% (TOP 186)

Top15% (TOP 278)

Top20% (TOP 371)

Top25% (TOP 464)

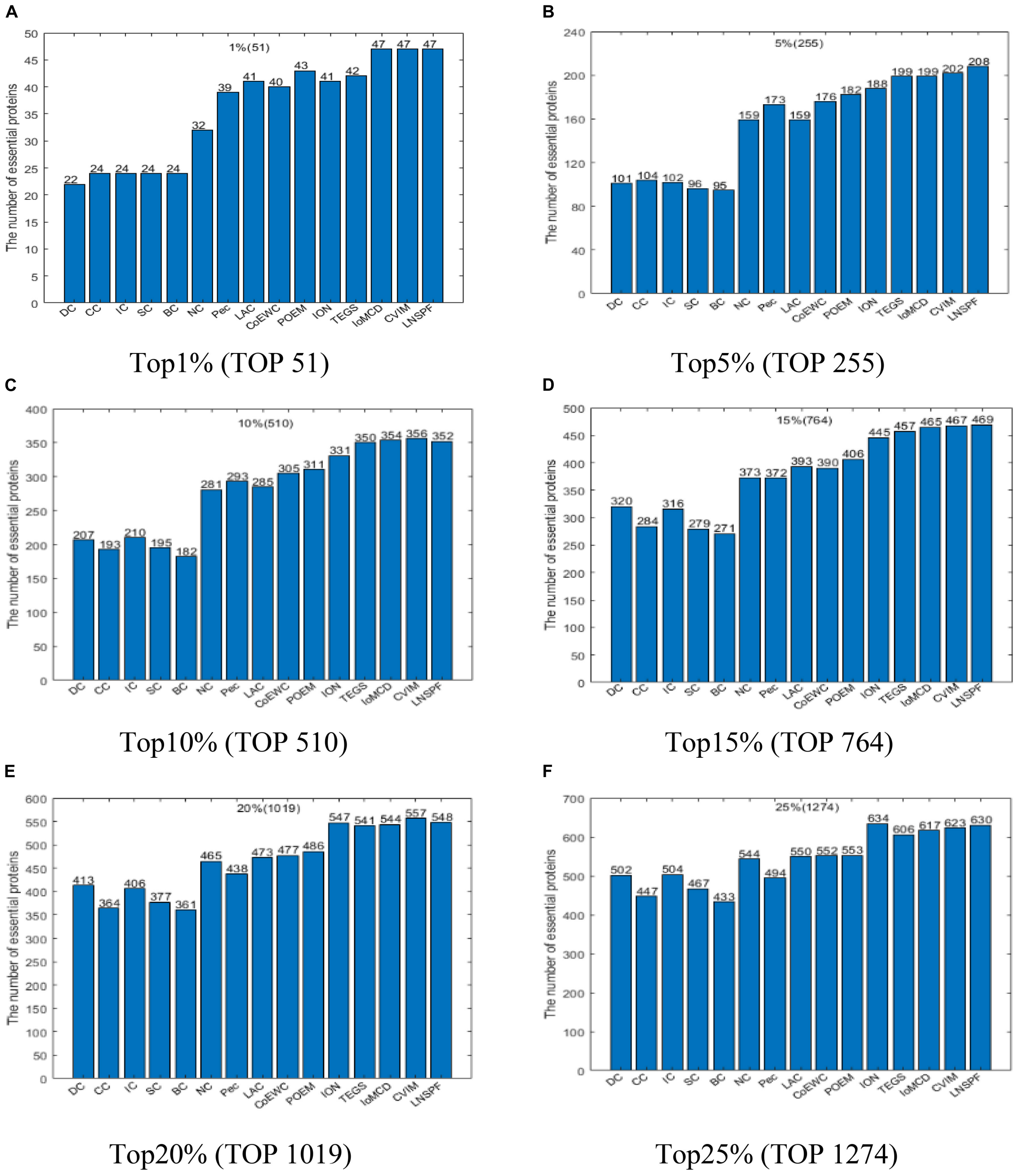
FIGURE 2 | Comparison results of the numbers of real key proteins predicted by LNSPF, DC, CC, IC, SC, BC, NC, PEC, LAC, CoEWC, POEM, ION, RWHN, IoMCD, and CVIM based on the GAVIN database. **(A)** Top 1% ranked proteins. **(B)** Top 5% ranked proteins. **(C)** Top 10% ranked proteins. **(D)** Top 15% ranked proteins. **(E)** Top 20% ranked proteins. **(F)** Top 25% ranked proteins.

$p_j \in NG\left(p_i\right)$ as follows:

$$\begin{aligned} & min \quad s_i^T \left(G^i + \mu I\right) s_i \\ & s.t. \quad \sum_{p_j \in NG\left(p_i\right)} s_{i,j} = 1 \text{ , } s_{i,j} \geq 0 \end{aligned}$$

Thereafter, let $T^0 = Y^{Final}$, based on above newly obtained matrix S, we can further optimize the scores for all proteins in an iterative way as follows:

$$T^{\sigma+1} = \beta S T^{\sigma} + (1-\beta) T^0 \tag{19}$$



Top1% (TOP 51)

Top5% (TOP 255)

Top10% (TOP 510)

Top15% (TOP 764)

Top20% (TOP 1019)

Top25% (TOP 1274)

FIGURE 3 | Comparison results of the numbers of real key proteins predicted by LNSPF, DC, CC, IC, SC, BC, NC, PEC, LAC, CoEWC, POEM, ION, RWHN, IoMCD, and CVIM based on the DIP database. **(A)** Top 1% ranked proteins. **(B)** Top 5% ranked proteins. **(C)** Top 10% ranked proteins. **(D)** Top 15% ranked proteins. **(E)** Top 20% ranked proteins. **(F)** Top 25% ranked proteins.

Here, there is $\beta \in (0, \quad 1)$.

Based on the above descriptions, the process of LNSPF can be described in detail as follows:

Algorithm: LNSPF.

Input: Original PPI network, gene expression data, subcellular location data and homologous data, parameters δ and K.

Output: Rank the proteins in descending order according to T^{Final} value, and output TOP K%.

Step 1: According to Equations 3, 4, an original PPI network $G = (V, E)$ is generated, based on which, topological features are extracted;

Step 2: According to Equations 7, 9, functional characteristics are extracted from the subcellular location data and homologous data, respectively.

Step 3: According to Equation 15, the matrix H is obtained;

Step 4: let $t = t + 1$; calculate Y^{t+1} according to Equation 16;

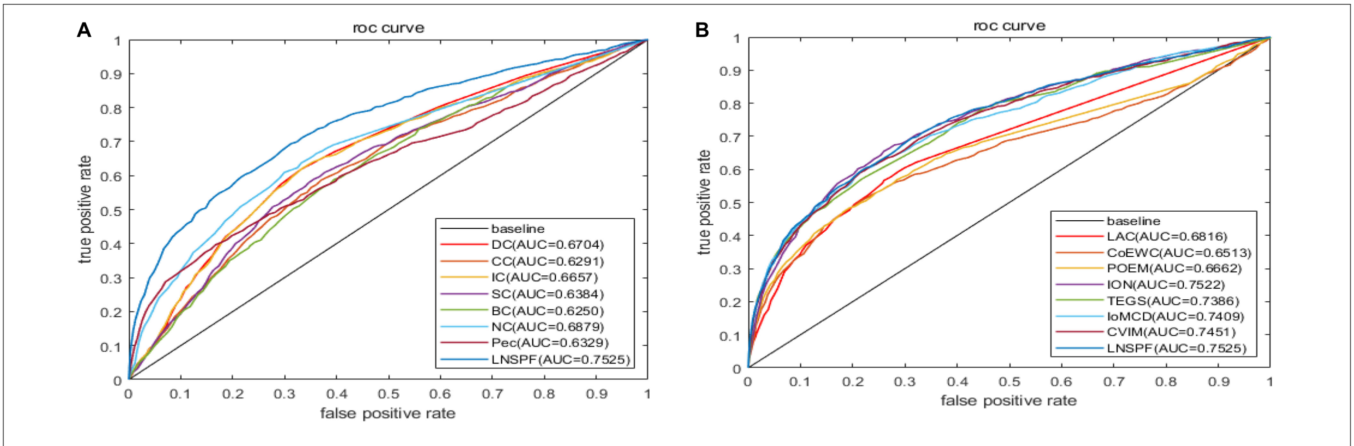


FIGURE 4 | The ROC curves of LNSPF method based on DIP dataset and DC, CC, IC, SC, BC, NC, Pec, and LAC CoEWC, POEM, ION, TEGS, IoMCD, and CVIM 14 prediction methods. **(A)** Comparison between LNSPF and DC, CC, IC, SC, BC, NC, PEC. **(B)** Comparison between LNSPF and LAC, CoEWC, POEM, ION, TEGS, IoMCD, CVIM.

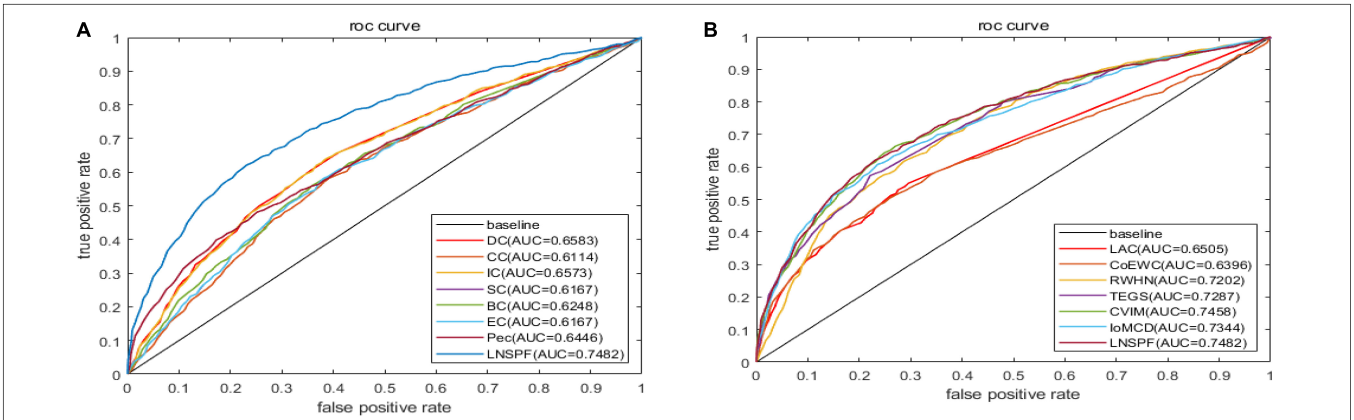


FIGURE 5 | The ROC curves of LNSPF method based on Krogan dataset and DC, CC, IC, SC, BC, EC, PEC, and LAC, CoEWC, RWHN, TEGS, CVIM, and IOMCD 13 prediction methods. **(A)** Comparison between LNSPF and DC, CC, IC, SC, BC, EC, PEC. **(B)** Comparison between LNSPF and LAC, CoEWC, RWHN, TEGS, CVIM, IOMCD.

Step 5: Repeat step 4 until $||Y^{t+1} - Y^{t}|| < \delta$, the matrix Y^{Final} is obtained;

Step 6: According to Equation 18, the similarity matrix S is obtained;

Step 7: let $T^{0} = Y^{Final}$ and $\sigma = \sigma + 1$, the matrix Y^{Final} is further optimized according to Equation 19;

Step 8: Repeat step 7 until $||T^{\sigma+1} - T^{\sigma}|| < \delta$, the matrix T^{Final} is obtained;

Step 9: The values of T^{Final} are sorted in descending order, and the top K% proteins with the highest final scores are output.

EXPERIMENTAL RESULTS

Experimental Data

During experiments, we first downloaded known PPIs from three different databases such as the Gavin (Gavin et al., 2006) database, the DIP (Xenarios et al., 2002) database, and the Krogan (Cherry, 1998) database, and then, after filtering repeated interactions and self-interactions, we finally obtained 24,743 interactions between 5,093 proteins based on the DIP database, 7,669 interactions between 1,855 proteins based on the Gavin database, and 14,317 interactions between 3,672 proteins based on the Krogan database, respectively. Moreover, we obtained a group of 1,285 essential proteins in Saccharomyces cerevisiae from the databases of SGDP (Holman et al., 2009), SGD (Holman et al., 2009), DEG (Zhang and Lin, 2009), and MIPS (Bruno et al., 2012) as well. Furthermore, we

TABLE 5 | Based on DIP database, LNSPF and AUC of 14 competitive methods.

Method	LNSPF	DC	CC	IC	SC	BC	NC	Pec
AUC	0.7525	0.6704	0.6293	0.6657	0.6384	0.6250	0.6879	0.6329
Method	**LNSPF**	**LAC**	**CoEWC**	**POEM**	**ION**	**TEGS**	**IoMCD**	**CVIM**
AUC	0.7525	0.6816	0.6513	0.6662	0.7522	0.7386	0.7409	0.7451

TABLE 6 | AUC values of LNSPF and 13 competing methods based on Krogan dataset.

Method	LNSPF	DC	CC	IC	SC	BC	EC
AUC	0.7482	0.6583	0.6114	0.6573	0.6167	0.6248	0.6167
Method	**PEC**	**LAC**	**CoEWC**	**RWHN**	**TEGS**	**CVIM**	**IoMCD**
AUC	0.6446	0.6505	0.6396	0.7202	0.7287	0.7458	0.7344

downloaded the homology information of proteins from the Inparanoid database (Gabriel et al., 2010), the gene expression dataset composing of 6,776 proteins representing the gene expression level of proteins in continuous metabolic cycles from the database provided by Tu et al. (2005), and the dataset of subcellular location information from the part-means database (Binder et al., 2014) separately. Especially, the dataset of subcellular location information consists of 11 kinds of subcellular localization, namely, the extracellular, peroxisome, nucleus, plasma, endosome, mitochondrion, vacuole, cytosol, golgi, cytoskeleton, and endoplasmic, which are closely related to known key proteins. At last, to evaluate the recognition rate of true essential proteins predicted by LNSPF, we compared LNSPF with 16 representative predictive models, as shown in **Table 1**, namely, DC, EC, CC, IC, SC, BC, NC, Pec, LAC, CoEWC, POEM, ION, TEGS, RWHN, IoMCD, and CVIM.

Influence of Parameters on Linear Neighborhood Similarity-Based Protein Multifeatures Fusion Performance

In LNSPF, we set parameters α and β, the value ranges of both α and β are (0, 1), to adjust the final protein score. During experiments, we will set different values to the parameter α or β first based on the Gavin database and the DIP database, respectively, and then, the setting value with the highest

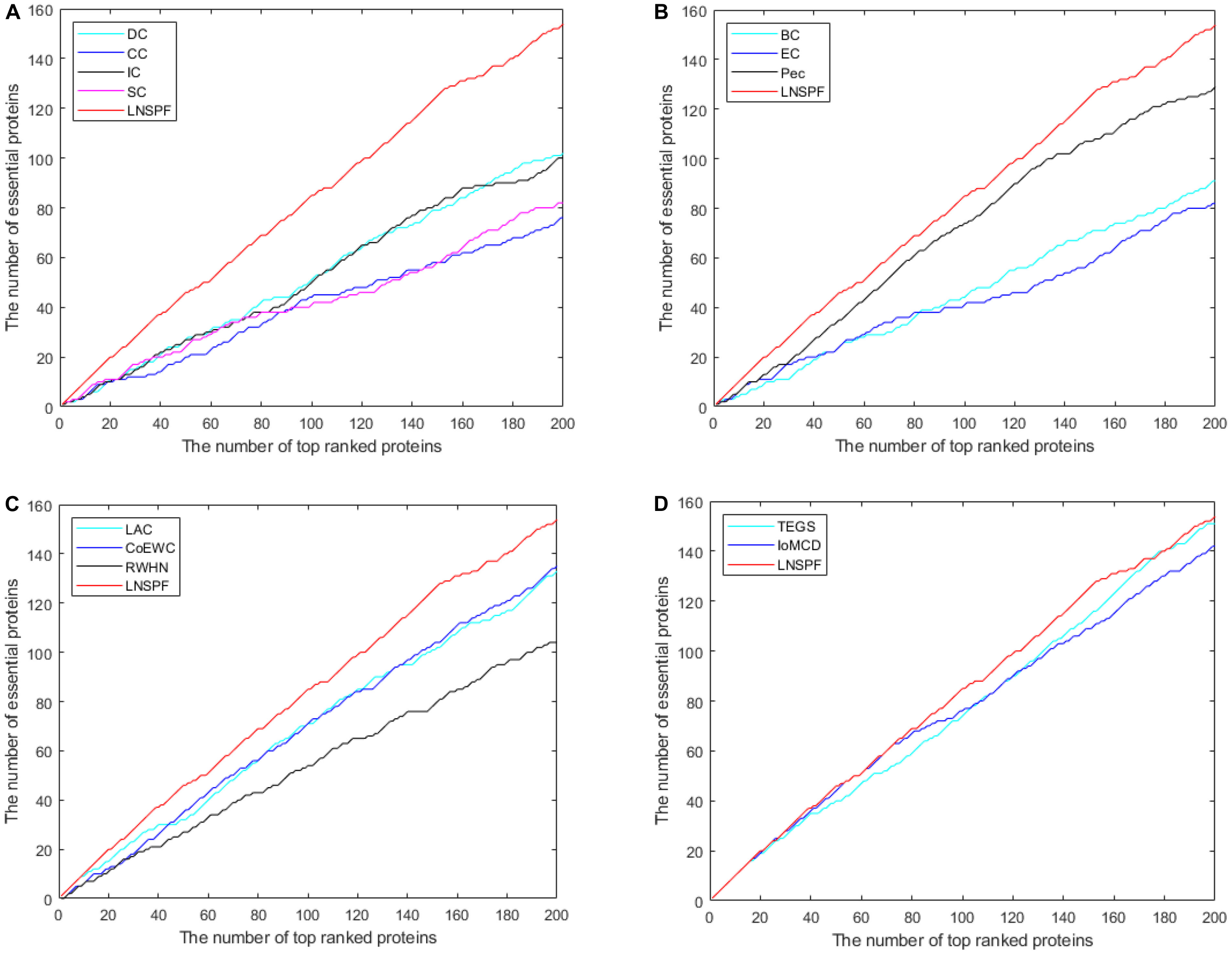


FIGURE 6 | The figure shows the Jackknife curves of LNSPF and DC, CC, IC, SC, BC, EC, and PEC based on Krogan dataset, and LAC, CoEWC, RWHN, TEGS, and IOMCD 12 prediction methods. The X-axis represents the number of potentially critical proteins ranked in the top 200, and the Y-axis represents the number of truly essential proteins identified by these models. **(A)** Comparison between LNSPF and DC, CC, IC, SC. **(B)** Comparison between LNSPF and BC, EC, PEC. **(C)** Comparison between LNSPF and LAC, CoEWC, RWHN. **(D)** Comparison between LNSPF and TEGS, IOMCD.

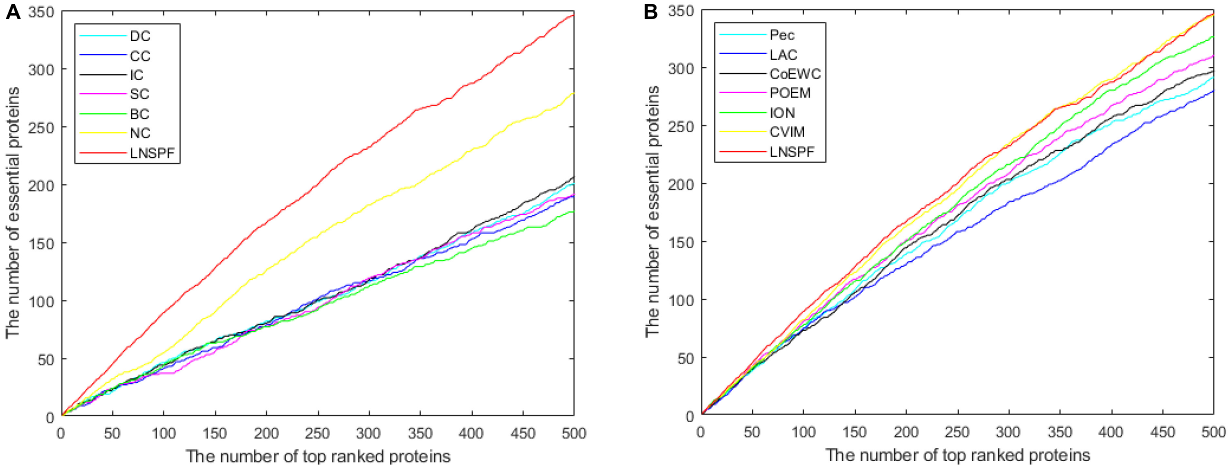


FIGURE 7 | The figure, respectively, shows the Jackknife curve of LNSPF and DC, CC, IC, SC, BC, NC, and PEC, LAC, COEWC, POEM, ION, and CVIM 12 prediction methods based on DIP data set. The X-axis represents the number of potentially critical proteins ranked in the top 500, and the Y-axis represents the number of truly essential proteins identified by these models. **(A)** Comparison between LNSPF and DC, CC, IC, SC, BC, NC. **(B)** Comparison between LNSPF and PEC, LAC, COEWC, POEM, ION, CVIM.

prediction accuracy of essential protein will be selected as the final value of parameter α or β. Based on the Gavin dataset, we set α to 0.1., 0.8, and 0.9 to predict the effect of the preliminary iterative algorithm. From observing **Table 2**, it is obvious that when α = 0.6, the protein score with obvious effect and the most stable one can be obtained. At this time, the setting value of α in Gavin dataset is 0.6 and that in DIP database is 0.8. β set 0.1, . . ., 0.8, 0.9. The prediction results based on Gavin data set (α = 0.6) and dip data set (α = 0.8) are shown in **Tables 3**, **4**, respectively. By observing **Table 3**, it is easy to see that the prediction performance of LNSPF is the highest at 1%, 5%, 15%, 20%, and 25% when β = 0.4 is used. Therefore, based on Gavin data set, it is appropriate to set β as 0.4. By observing **Table 4**, it is easy to see that the prediction performance of LNSPF is the highest at 1%, 5%, 10%, and 25% when β = 0.2 is used. Therefore, based on the DIP data set, it is more appropriate to set β as 0.2.

COMPARISON OF LNSPF WITH OTHER METHODS

Comparison of the Number of Real Essential Proteins Between Linear Neighborhood Similarity-Based Protein Multifeatures Fusion and 14 Representative Methods

According to above descriptions, it is easy to see that LNSPF can achieve it best predictive performance while we set α to 0.6 and β to 0.4 based on the Gavin database. Hence, in this section, in order to estimate the actual predictive performance of LNSPF, we will first compare it with 14 advanced predictive methods based on the Gavin database while setting α to 0.6 and β to 0.4, and the comparison results are shown in **Figure 2**. From observing the **Figure 2**, it is easy to see that, in the ranking of the number of true essential proteins inferred by these 15 predictive methods, LNSPF can achieve better predictive performance than all these competitive methods in top 1, 5, 10, 15, and 20% predicted key proteins simultaneously. For instance, from the top 1% to top 20% predicted key proteins, the predictive accuracies of LNSPF are 15.8, 4.3, 2.6, 1.4, and 1.8% higher than that of the method of CVIM, respectively.

Similarly, according to above descriptions, it is easy to see that LNSPF can achieve it best predictive performance while we set α to 0.6 and β to 0.2 based on the DIP database. Hence, in this section, in order to estimate the actual predictive performance of LNSPF, we will further compare it with 14 advanced predictive methods based on the DIP database while setting α to 0.6 and β to 0.2, and the comparison results are shown in **Figure 3**. From observing the **Figure 3**, it is easy to see that, the numbers of essential proteins detected by LNSPF in the top 1, 5, 10, 15, 20, and 25% ranked proteins are significantly better than that of all competitive methods as a whole.

Receiver Operating Characteristic Curve Verification

Receiver operating characteristic curve (ROC) is used to compare the prediction performance of LNSPF with DC, CC, IC, SC, BC, NC, PEC, LAC, CoEWC, POEM, ION, TEGS, IoMCD, and CVIM based on DIP data set. The larger the area of ROC curve, the better the performance of the model, it can be seen from **Figure 4** and **Table 5** that the performance of this model is significantly higher than that of the 14 competitive methods. The prediction performance of LNSPF method based on Krogan dataset compared with DC, CC, IC, SC, BC, EC, PEC, and LAC, CoEWC, RWHN, TEGS, CVIM, and IoMCD 13 competing methods. It can be seen from **Figure 5** and **Table 6** that the performance of this model is significantly higher than that of these 13 competing methods.

Verification of Jackknife Method

In this section, I'll use the Jackknife method to verify the performance of the LNSPF against the other models. The performance of LNSPF was compared with DC, CC, IC, SC, BC, EC, PEC, and LAC, CoEWC, RWHN, TEGS, and IOMCD based on Krogan data set. As shown in **Figure 6**. It is obvious that this method is superior to other models. The performance of LNSPF is compared with DC, CC, IC, SC, BC, NC, PEC, and LAC, COEWC, POEM, ION, and CVIM based on DIP data set, as shown in **Figure 7**.

DISCUSSION

Essential proteins play an important role in cell growth and regulation, for the past few years, accumulating computational methods have been proposed to detect potential key proteins, however, the predictive performances of these existing methods are not very satisfactory yet. In this study, a novel predictive model called LNSPF was designed by combining topological features of PPI networks with a series of biological characteristics of proteins to detect potential key proteins. In LNSPF, a new entropy-based method for feature fusion and a linear neighborhood similarity method for optimization were adopted. Comparing with traditional identification methods, LNSPF can achieve better predictive performance, which demonstrates that the method based on the fusion of biological information of proteins and topological features of PPI networks can improve the prediction accuracy of essential proteins effectively. In addition, there are some limitations in current version of LNSPF as well, for example, the loss of gene time expression data or homologous data of some proteins will affect the recognition accuracy of LNSPF to some degree.

CONCLUSION

In this paper, an iterative model of protein multifeature fusion based on linear neighborhood similarity (LNSPF) is proposed to predict essential proteins by fusing biological and topological information of proteins. In LNSPF, first, the topological features are extracted from the original PPI network, and then the functional features are extracted from the subcellular location data. Second, an entropy weight method is used to fuse the features, and then a stable protein score is obtained by an iterative method. At last, a linear neighborhood similarity method is used to optimize the score effectively. The experimental results show that based on Gavin data sets, the Krogan data sets, and DIP held several experimental data sets, through a variety of methods to verify the effectiveness of the new model LNSPF and stability. Compared with many advanced prediction models, the new model LNSPF has better prediction effect.

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 author/s.

AUTHOR CONTRIBUTIONS

XZ and YZ conceived the study, implemented the algorithms corresponding to the study, and wrote the manuscript. LW and ZC improved the study based on the original model. YT and LW supervised the study. XZ and YZ revised the manuscript. All authors reviewed and improved the manuscript.

FUNDING

This research is partly sponsored by the Research Foundation of Education Bureau of Hunan Province (No. 20B080), the Natural Science Foundation of Hunan Province (No. 2019JJ70010), the Hunan Provincial Natural Science Foundation of China (2020JJ4152), the Science and Technology Plan Project of Hunan Province (2016TP1020), the Hunan Province Science and Technology Project Funds (2018TP1036), and the National Scientific Research Foundation of Hunan Province Education Commission (18B367).

ACKNOWLEDGMENTS

The authors thank Shiyuan Li and the anonymous referees for suggestions that helped improve the manuscript substantially.

REFERENCES

Binder, J. X., Pletscher-Frankild, S., Tsafou, K., Stolte, C., O'Donoghue, S. I., Schneider, R., et al. (2014). COMPARTMENTS: unification and visualization of protein subcellular localization evidence. *Database* 2014:bau012. doi: 10.1093/database/bau012

Bonacich, P. (1987). ‘Power and centrality: a family of measures. *Am. J. Sociol.* 92, 1170–1182. doi: 10.1086/228631

Bruno, A., Jef, B., and Carla, C. (2012). *SGDP: Saccharomyces Genome Deletion Project [EB/OL]*. Available online at: http://yeastdeletion.stanford.edu/ (accessed June 20, 2012).

Chen, Z., Meng, Z., Liu, C., Wang, X., Kuang, L., Pei, T., et al. (2020). A novel model for predicting essential proteins based on heterogeneous protein-domain network. *IEEE Access* 8, 8946–8958. doi: 10.1109/access.2020.2964571

Chen, Z.-H., You, Z.-H., Li, L.-P., Guo, Z. H., Hu, P. W., Jiang, H. J., et al. (2019). "Combining LSTM network model and wavelet transform for predicting self-interacting proteins," in *Intelligent Computing Theories and Application. ICIC 2019. Lecture Notes in Computer Science*, eds D. S. Huang, V. Bevilacqua, and P. Premaratne (Cham: Springer).

Cherry, J. (1998). SGD: *Saccharomyces* genome database. *Nucleic Acids Res.* 26, 73–79. doi: 10.1093/nar/26.1.73

Dai, W., Chang, Q., Peng, W., Zhong, F., and Li, Y. (2020). Network embedding the protein-pro tein interaction network for human essential genes identification. *Genes* 11:153. doi: 10.3390/genes11020153

Dai, W., Chen, B., Peng, W., Li, X., Zhong, J., and Wang, J. (2021). A novel multi-ensemble method for identifying essential proteins. *J. Comp. Biol.* 28, 637–649. doi: 10.1089/cmb.2020.0527

Estrada, E., and Rodríguez-Velázquez, J. A. (2005). Subgraph centrality in complex networks. *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.* 71, 33–122.

Fan, Y., Tang, X., Hu, X., Wu, W., and Ping, Q. (2017). Prediction of essential proteins bases on subcellular localization and gene expression correlation. *BMC Bioinformatics* 18(Suppl. 13):470. doi: 10.1186/s12859-017-1876-1875

Gabriel, O., Thomas, S., Kristoffer, F., Köstler, T., Messina, D. N., Roopra, S., et al. (2010). InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res.* 38, D196–D203. doi: 10.1093/nar/gkp931

Gavin, A. C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., et al. (2006). Proteome survey reveals modularity of the yeast cell machinery. *Nature* 440:631. doi: 10.1038/nature04532

Hahn, M. W., and Kern, A. D. (2005). Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Mol. Biol. Evol.* 22, 803–806. doi: 10.1093/molbev/msi072

Holman, A. G., Davis, P. J., Foster, J. M., Carlow, C. K., and Kumar, S. (2009). Computational prediction of essential genes in an unculturable endosymbiotic bacterium. Wolbachia of Brugia Malayi. *BMC Microbiol.* 9:243. doi: 10.1186/1471-2180-9-243

Joy, M. P., Brock, A., Ingber, D. E., and Huang, S. (2005). High-betweenness proteins in the yeast protein interaction network. *J. Biomed. Biotechnol.* 2005, 96–103. doi: 10.1155/JBB.2005.96

Lei, X., Ding, Y., Fujita, H., and Aidong, Z. (2016). Identification of dynamic protein complexes based on fruit fly optimization algorithm. *Knowl Base Syst.* 105, 270–277. doi: 10.1038/s41598-018-28680-8

Lei, X., Yang, X., and Wu, F.-X. (2020). Artificial fish swarm optimization-based method to identify essential proteins. *IEEE/ACM Trans Comput Biol Bioinform.* 17, 495–495. doi: 10.1109/TCBB.2018.2865567

Li, G., Min, L., Wang, J., Wu, J., Wu, F. X., Pan, Y., et al. (2016). Predicting essential proteins based on subcellular localization, orthology and PPI networks. *BMC Bioinformatics* 17:279. doi: 10.1186/s12859-016-1115-5

Li, M., Li, W., Wu, F. X., Pan, Y., and Wang, J. (2018). Identifying essential proteins based on sub-network partition and prioritization by integrating subcellular localization information. *J. Theoretical Biol.* 447, 65–47. doi: 10.1016/j.jtbi.2018.03.029

Li, M., Lu, Y., Wang, J., Wu, F.-X., and Pan, Y. (2015). 'A topology potential-based method for identifying essential proteins from PPI networks. *IEEE/ACM Trans. Comput. Biol. Bioinf.* 12, 372–383. doi: 10.1109/TCBB.2014.2361350

Li, M., Zhang, H., Wang, J. X., and Pan, Y. (2012). A new essential protein discovery method based on the integration of protein-protein interaction and gene expression data. *BMC Syst. Biol.* 6:15. doi: 10.1186/1752-0509-6-15

Li, S., Chen, Z., He, X., Zhang, Z., Pei, T., Tan, Y., et al. (2020). An iteration method for identifying yeast essential proteins from weighted PPI network based on topological and functional features of proteins. *IEEE Access* 8, 90792–90804. doi: 10.1109/access.2020.2993860

Lin, C. Y., Chin, C. H., and Wu, H. H. (2011). Hubba: hub objects analyzer-a framework of interactome hubs identification for network biology. *Comp. Biol. Chem.* 35:143. doi: 10.1093/nar/gkn257

Min, L., Yu, L., Niu, Z., and Wu, F. X. (2017). United complex centrality for identification of essential proteins from PPI networks. *IEEE/ACM Trans. on Comp. Biol. Bioinform. (TCBB)* 14, 370–380. doi: 10.1109/TCBB.2015.2394487

Peng, W., Wang, J. X., Cheng, Y., Lu, Y., Wu, F., Pan, Y., et al. (2015a). UDoNC: an algorithm for identifying essential proteins based on protein domains and protein-protein interaction networks. *IEEE/ACM Trans. Comp. Biol. Bioinform.* 12, 276–288. doi: 10.1109/TCBB.2014.2338317

Peng, W., Wang, J. X., Wang, W., Liu, Q., Wu, F. X., Pan, Y., et al. (2012). Iteration method for predicting essential proteins based on orthology and protein-protein interaction networks. *BMC Syst. Biol.* 6:87. doi: 10.1186/1752-0509-6-87

Peng, X., Wang, J., Zhong, J., Luo, J., and Pan, Y. (2015b). "An efficient method to identify essential proteins for different species by integrating protein subcellular localization information," in *Proceedings of the IEEE International Conference on Bioninformatics and Biomedicine*, (Piscataway, NJ: IEEE).

Qi, Y., and Luo, J. (2016). Prediction of essential proteins based on local interaction density. *IEEE/ACM Trans. Comp. Biol. Bioinform.* 13, 1170–1182. doi: 10.1109/TCBB.2015.2509989

Stephenson, K., and Zelen, M. (1989). Rethinking centrality: methods and examples. *Soc. Netw.* 11, 1–37.

Sun, W., Wang, L., Peng, J., Zhang, Z., Pei, T., Tan, Y., et al. (2021). A cross-entropy-based method for essential protein identification in yeast protein-protein interaction network. *Curr. Bioinform.* 16, 565–575. doi: 10.2174/1574893615999201116210840

Tang, X., Wang, J., Zhong, J., and Pan, Y. (2014). Predicting essential proteins based on weighted degree centrality. *IEEE/ACM Trans. Comp. Biol. Bioinform.* 11, 407–418. doi: 10.1109/TCBB.2013.2295318

Tu, B. P., Kudlicki, A., Rowicka, M., and McKnight, S. L. (2005). Logic of the yeast metabolic cycle: temporal compartmentalization of cellular processes. *Science* 310, 1152–1158. doi: 10.1126/science.1120499

Wang, F., and Zhang, C. (2008). Label propagation through linear neighborhoods. *Knowledge Data Eng. IEEE Trans.* 20, 55–67. doi: 10.1109/tkde.2007.190672

Wang, J. X., Li, M., and Wang, H. (2012). Identifification of essential proteins based on edge clustering coefficient. *IEEE/ACM Trans. Comput. Biol. Bioinf.* 9, 1070–1080. doi: 10.1109/tcbb.2011.147

Wang, L., You, Z.-H., Xia, S.-X., Liu, F., Chen, X., Yan, X., et al. (2017). Advancing the prediction accuracy of protein-protein interactions by utilizing evolutionary information from position-specific scoring matrix and ensemble classifier. *J. Theor. Biol.* 418, 105–110. doi: 10.1016/j.jtbi.2017.01.003

Wuchty, S., and Stadler, P. F. (2003). Centers of complex networks. *J. Theor. Biol.* 223, 45–53. doi: 10.1016/s0022-5193(03)00071-7

Xenarios, I., Salwinski, L., Duan, X. J., Higney, P., Kim, S. M., Eisenberg, D., et al. (2002). DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.* 30, 303–305. doi: 10.1093/nar/30.1.303

Zhang, F., Peng, W., Yang, Y., Dai, W., and Song, J. (2019). A novel method for identifying essential genes by fusing dynamic protein–protein interactive networks. *Genes* 10:31. doi: 10.3390/genes10010031

Zhang, R., and Lin, Y. (2009). DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes. *Nucleic Acids Res.* 37, D455–D458. doi: 10.1093/nar/gkn858

Zhang, W., Xue, X., Xie, C., Li, Y., Liu, J., Chen, H., et al. (2021). CEGSO: boosting essential proteins prediction by integrating protein complex, gene expression, gene ontology, subcellular localization and orthology information. *Interdisciplinary Sci. Comp. Life Sci.* 13, 349–361. doi: 10.1007/s12539-021-00426-7

Zhang, W., Yue, X., Liu, F., Chen, Y., Tu, S., Zhang, X., et al. (2017). A unified frame of predicting side effects of drugs by using linear neighborhood similarity. *BMC Syst. Biol.* 11:101. doi: 10.1186/s12918-017-0477-472

Zhang, X., Xu, J., and Xiao, W. (2013). A new method for the discovery of essential proteins. *PLoS One* 8:e58763. doi: 10.1371/journal.pone.0058763

Zhao, B., Wang, J., Li, M., Wu, F.-X., and Pan, Y. (2014). Prediction of essential proteins based on overlapping essential modules. *IEEE Trans. Nanobiosci.* 13, 415–424. doi: 10.1109/TNB.2014.2337912

Zhao, B., Zhao, Y., Zhang, X., Zhang, Z., Zhang, F., and Wang, L. (2019). An iteration method for identifying yeast essential proteins from heterogeneous network. *BMC Bioinf.* 20:355. doi: 10.1186/s12859-019-2930-2

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.