The Journal of Chemical Physics

RESEARCH ARTICLE | JULY 25 2024

Unraveling the role of physicochemical differences in predicting protein–protein interactions

Special Collection: Machine Learning for Biomolecular Modeling

Hamid Teimouri ; Angela Medvedeva; Anatoly B. Kolomeisky

J. Chem. Phys. 161, 045102 (2024)
https://doi.org/10.1063/5.0219501

AIP Publishing

Unraveling the role of physicochemical differences in predicting protein–protein interactions

Cite as: J. Chem. Phys. **161**, 045102 (2024); doi: 10.1063/5.0219501
Submitted: 17 May 2024 • Accepted: 9 July 2024 •
Published Online: 25 July 2024

Hamid Teimouri,[1,2,3] Angela Medvedeva,[1,2,3] and Anatoly B. Kolomeisky[1,2,3,a)]

AFFILIATIONS

[1] Department of Chemistry, Rice University, Houston, Texas 77005, USA
[2] Center for Theoretical Biological Physics, Rice University, Houston, Texas 77005, USA
[3] Department of Chemical and Biomolecular Engineering, Rice University, Houston, Texas 77005, USA

Note: This paper is part of the JCP Special Topic on Machine Learning for Biomolecular Modeling.
[a)] **Author to whom correspondence should be addressed:** tolya@rice.edu

ABSTRACT

The ability to accurately predict protein–protein interactions is critically important for understanding major cellular processes. However, current experimental and computational approaches for identifying them are technically very challenging and still have limited success. We propose a new computational method for predicting protein–protein interactions using only primary sequence information. It utilizes the concept of physicochemical similarity to determine which interactions will most likely occur. In our approach, the physicochemical features of proteins are extracted using bioinformatics tools for different organisms. Then they are utilized in a machine-learning method to identify successful protein–protein interactions via correlation analysis. It was found that the most important property that correlates most with the protein–protein interactions for all studied organisms is dipeptide amino acid composition (the frequency of specific amino acid pairs in a protein sequence). While current approaches often overlook the specificity of protein–protein interactions with different organisms, our method yields context-specific features that determine protein–protein interactions. The analysis is specifically applied to the bacterial two-component system that includes histidine kinase and transcriptional response regulators, as well as to the barnase–barstar complex, demonstrating the method's versatility across different biological systems. Our approach can be applied to predict protein–protein interactions in any biological system, providing an important tool for investigating complex biological processes' mechanisms.

I. INTRODUCTION

Protein–protein interactions (PPIs), which can be viewed as the result of various biochemical reactions and electrostatic attractions, play a critical role in many cellular processes by supporting a variety of crucial biological functions.[1,2] These functions range from signal transduction, such as stimulus–response coupling in bacteria,[3,4] and enzymatic regulation to the generation of immune responses.[5–7] Furthermore, certain protein–protein interactions are closely associated with the development and progress of various diseases, including viral pathogenesis,[8] cancer,[9] and neurodegenerative diseases.[5,10] For example, neurological disorders such as Alzheimer's disease, Parkinson's disease, and Huntington's disease have all been linked to mutations that specifically disrupt PPIs that can prevent misfolding, leading to the effectively irreversible aggregation of proteins.[7]

The exact identification of PPIs in cellular systems remains a very difficult task. Several experimental techniques, including yeast two-hybrid (Y2H) screens,[11,12] mass spectroscopy,[13,14] and tandem affinity purification (TAP),[15,16] have been developed in recent years for detecting them. However, despite some advances, determining PPIs in labs remains technically very challenging, time-consuming, and costly. In addition, due to the complexity of the underlying processes, these experimental methods often exhibit high rates of false positives and false negatives.[17] As a result, several computational methods have been proposed to assist in predicting protein interactions more accurately and efficiently.[18–20] Such theoretical methods not only support traditional wet lab experiments but also offer a more cost-effective means to quickly identify potentially interacting protein pairs across the huge space of the entire proteome.[21] Yet, the performance of most of these techniques declines when supplemental additional biological information, such as protein structure details, protein domains, or gene neighborhood information, is not available.[21] Hence, there is an immediate need to devise new computational strategies that could efficiently predict PPIs, preferably relying only on the limited information coming mostly from protein sequence data.[22]

Because of the large volume of available biological information, machine learning methods have recently emerged as powerful tools to complement traditional experimental techniques, enabling the analysis and prediction of PPIs from amino acid sequences.[22–27] However, many advanced machine-learning models, such as deep neural networks, are black boxes, making it difficult to understand why they make specific predictions. Such methods do not provide insights into which features of the protein sequence are most relevant for these interactions. Moreover, traditional models frequently rely on simplistic representations of protein sequences, such as amino acid composition, or a very limited set of physicochemical descriptors of proteins.[22] On the other hand, despite the advancements achieved in the field of PPI prediction using machine learning, current approaches often overlook a crucial aspect—the specificity of protein–protein interactions within different biological systems. Biological processes are highly contextual, and protein interactions may vary significantly across diverse organisms and cellular environments. For example, different organisms adapted to different temperature ranges may influence the specificity, stability, and kinetics of their PPIs. Existing machine-learning methods might not fully capture the species-specific patterns and nuances of the PPI networks, limiting their ability to provide robust predictions.

Here, we present a novel computational approach that addresses this crucial gap in the abilities of PPI prediction techniques. We define protein–protein interactions as the probability for two different protein molecules to be associated with each other for significant periods during cellular functioning. Although utilizing the molecular properties of amino acid sequences to predict protein–protein interactions is a common idea in computational biology, what makes our method unique is that we used a comprehensive set of descriptors, including dipeptide composition, correlation functions, and pseudo amino acid composition. We hypothesize that interactions between different protein species correlate with specific molecular properties that may be indicative of the thermodynamics and kinetics of their binding interactions. In this approach, we extract a comprehensive set of physicochemical features of proteins using a standard bioinformatic tool.[28] Then, the concept of physicochemical similarity between protein pairs is applied to identify the correlations with protein–protein interactions. It is important to note that we use 'physical–chemical similarity' as a general term, as certain features in proteins, such as electric charges, when dissimilar can indeed support higher binding probabilities.

By incorporating species-specific features and training machine-learning models on organism-specific datasets, our method reveals the unique aspects of PPI networks in different organisms. We investigated six diverse datasets encompassing microorganisms, mammals, insects, and plants, allowing us to comprehensively capture the properties of PPI networks across different biological kingdoms. The protein–protein interaction prediction is modeled as a classification problem, applying the principles of supervised machine learning. By employing supervised machine-learning techniques, specifically logistic regression and Support Vector Machines (SVMs), we demonstrate that a selected set of physicochemical protein features can effectively predict whether proteins will interact or not. Our analysis identifies that dipeptide composition features are universal factors across all studied organisms that best correlate with the possibility of PPIs. The proposed computational method provides an enhanced approach to understanding the characteristics of proteins associated with successful interactions.

TABLE I. Summary of protein–protein interactions datasets used in our computational study. Data obtained from Ref. 32.

Species	Proteins	PPI (+/−)
Escherichia coli (EC2)	589	1167/1167
Saccharomyces cerevisiae (SC5)	454	500/500
Schizosaccharomyces pombe (SP)	904	742/742
Arabidopsis thaliana (AT)	756	541/541
Mus musculus (MM)	1088	500/500
Drosophila melanogaster (DM2)	658	321/321

II. MATERIALS AND METHODS

A. Dataset and data pre-processing

We considered protein–protein interactions in two types of living systems: (1) unicellular organisms, including the bacteria *Escherichia coli* (EC2) and two distinct species of yeast, including *Saccharomyces cerevisiae* (SC5) and *Schizosaccharomyces pombe* (SP); and (2) multicellular organisms, including *Mus musculus* (MM), *Drosophila melanogaster* (DM2), and *Arabidopsis thaliana* (AT). We utilized data collected from various databases and analyzed in previous studies.[29–31] The summary of all utilized information for different systems is presented in Table I.

The data for each organism consisted of pairs of proteins and their corresponding sequences. Each protein–protein pair in the dataset is labeled as 1 if they interact and 0 if they do not interact. In this context, protein–protein interaction refers to the probability of binding. For each organism, there was an equal number of protein–protein pairs that interact vs those that do not interact, as illustrated in Table I. This allows us to minimize the bias in the analysis of the data.

B. Generation of physicochemical descriptors for proteins

From the amino acid sequence of each protein, we extracted a comprehensive set of physicochemical descriptors using the *propy* package.[28] The features were broadly classified into different categories, including charge, residue composition features (e.g., dipeptide composition), autocorrelations, chemical composition features, and sequence order features. Proteins containing non-natural amino acids were excluded from our dataset, as the *propy* package only identifies natural amino acids, and we are also interested in finding PPIs only in real cellular systems.

For each protein, the quantitative values of the physicochemical properties have different numerical values. It is important to initially rescale all these values to fall between 0 and 1 so that every property is considered to have a similar weight. To normalize this quantity to be in the range of 0 and 1, we use the following rescaling expression:

$$\hat{z} = \frac{(z - z_{min})}{(z_{max} - z_{min})}, \tag{1}$$

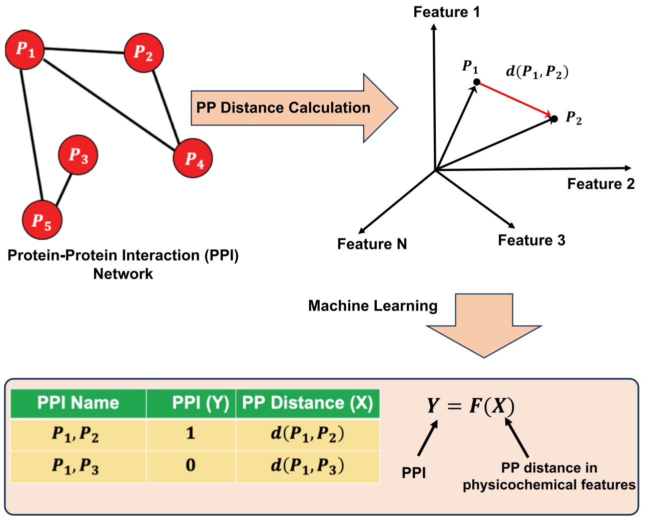
where z is the original value of the physicochemical property, z_{min} and z_{max} are the limiting values for this property for all considered proteins, and $\hat{z}$ is the normalized one that is specifically utilized in the analysis.

C. Protein–protein interaction as a classification problem

By extracting various physicochemical features, we can mathematically represent each protein as a vector in a high-dimensional space of these properties. The overall scheme for our procedure is presented in Fig. 1. Let us consider two arbitrary proteins, A and B, for which there are N available physicochemical features. Their vector representations are $\boldsymbol{A} = [A_1, A_2, \ldots, A_N]$ and $\boldsymbol{B} = [B_1, B_2, \ldots, B_N]$, respectively. Thus, the difference between two vectors is given as another vector,

$$d(\boldsymbol{A} - \boldsymbol{B}) = \big[|A_1 - B_1|, \quad |A_2 - B_2|, \quad \ldots, \quad |A_N - B_N|\big]. \tag{2}$$

The process of identifying protein–protein interactions can be viewed as a supervised machine-learning problem. In our dataset, we assign an index y_i to each protein–protein pair. If the two proteins interact, $y_i = 1$, and if they do not interact, $y_i = 0$. The feature vector (with total n properties) $\mathbf{x_i} = \{x_{i,1}, x_{i,2}, \ldots, x_{i,n}\}$ for each protein–protein pair i describes differences between the two proteins in terms of individual features. Then, the Support Vector Machine (SVM) classification method[33] is employed for predicting protein–protein interactions from the differences in the physicochemical properties. However, using feature selection, as described below, can make the prediction model more interpretable by revealing a limited number of the most important physicochemical features and whether two proteins tend to be similar or distinct in these features.



PPI Name	PPI (Y)	PP Distance (X)
P_1, P_2	1	$d(P_1, P_2)$
P_1, P_3	0	$d(P_1, P_3)$

FIG. 1. Schematic view of the machine learning classification model for prediction of protein–protein interaction using physicochemical differences between proteins.

D. Feature selection process

The number of possible physicochemical descriptors is very large, and many of these properties strongly correlate with each other. In such a high-dimensional feature space, it is beneficial to identify a small subset of the most predictive features. This can be achieved mathematically by assigning zero weights to irrelevant or redundant features in regression and SVM methods. The LASSO (Least Absolute Shrinkage and Selection Operator) regression and support vector machine are two prevalent techniques employed for shrinkage and feature selection.[34,35] A support vector machine (SVM) approach with surface patch analysis has been successfully utilized to predict protein–protein binding sites.[36]

E. Evaluating performance of machine-learning models

In the evaluation of machine learning models, several metrics are commonly used to measure the performance of these models. Each of these metrics has its strengths and weaknesses. *Accuracy*, which is one of the most intuitive metrics, represents the proportion of correctly classified instances (both true positives and true negatives) to the total number of instances. This quantity can be evaluated via

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \tag{3}$$

Here, true positives (TP) and true negatives (TN) represent the number of correctly classified interacting protein pairs. Similarly, false positives (FP) and false negatives (FN) denote the count of incorrectly classified protein–protein interactions. Accuracy is a suitable measure when the classes in the dataset are well-balanced, meaning there are roughly an equal number of instances for each class.

Another evaluating metric is *Recall*, which measures the proportion of actual positive instances that the model correctly identified,

$$Recall = \frac{TP}{TP + FN}. \tag{4}$$

However, *Recall* only considers the positive class, and sometimes there is a need for a metric that considers both classes.

In addition to *Accuracy* and *Recall*, the model's performance can be assessed using a so-called F_1 score (also known as an F-score or F-measure).[37] It is particularly useful in situations in which the data are imbalanced,[38] although Matthew's Correlation Coefficient (MCC) has been shown to be more representative of imbalanced datasets.[39] The F_1 score is defined as

$$F_1 = \frac{2TP}{2TP + FP + FN}. \tag{5}$$

However, one limitation of the F_1 score is that it still does not take true negatives into account. In some cases, correctly identifying negatives (e.g., healthy patients in a medical test) can be just as important as identifying positives.

The final evaluating quantity is a Matthews Correlation Coefficient (MCC) that serves as a more dependable statistical measure for complex scenarios,

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}. \tag{6}$$

MCC ranges from −1 to +1, where +1 represents a perfect prediction, 0 is no better than the random prediction, and −1 indicates total disagreement between prediction and observation.

While accuracy, *Recall*, MCC, and $F1$ scores each provide valuable information about the model's performance, they generally provide an evaluation at a single threshold, often set at 0.5, missing how performance varies across all possible thresholds. To address these limitations, the Receiver Operating Characteristic (ROC) curve provides a comprehensive view of a model's performance across all thresholds, highlighting the trade-offs between true positive and false positive rates and offering a more robust evaluation. The ROC curve is a graphical plot that illustrates the diagnostic ability of a binary classifier as its discrimination threshold is varied from 0 to 1. The curve is created by plotting the true positive rate (TPR), which is the same as *Recall* [Eq. (4)], against the false positive rate (FPR) at various threshold settings. The FPR is defined as follows:

$$FPR = \frac{FP}{FP + TN}. \tag{7}$$

By plotting TPR against FPR at various threshold settings, the ROC curve shows the trade-off between the classifier's sensitivity (its ability to identify true positives) and its specificity (its ability to avoid false positives). Each point on the ROC curve represents the TPR and FPR at a specific threshold. As the threshold changes, the TPR and FPR change, creating different points on the curve.

The area under the ROC curve (AUC) is a single value that summarizes the performance of the classifier across all threshold values. The AUC represents the probability that a randomly chosen positive instance is given a higher score by the classifier than a randomly chosen negative instance. AUC values range from 0 to 1, where a value of 1 indicates a perfect classifier, 0.5 represents a random classifier, and values less than 0.5 suggest a model performing worse than a random classifier. For example, a classifier assigns a score to each protein–protein pair, indicating the likelihood of an interaction between them. If we randomly select a pair of proteins that do interact (positive instance) and one pair that does not interact (negative instance), the AUC represents the probability that the interacting pair will receive a higher score than the non-interacting pair.

III. RESULTS AND DISCUSSIONS

The results of the feature selection methods for protein–protein interaction networks in EC2, SC5, and SP organisms are shown in Figs. 2–4, respectively. The definitions of the acronyms are presented in Table S1 in the supplementary material. In these figures, negative coefficients for features indicate that the differences between two proteins' properties negatively correlate with their ability to interact, while positive coefficients suggest that those differences positively correlate with the protein–protein interaction.

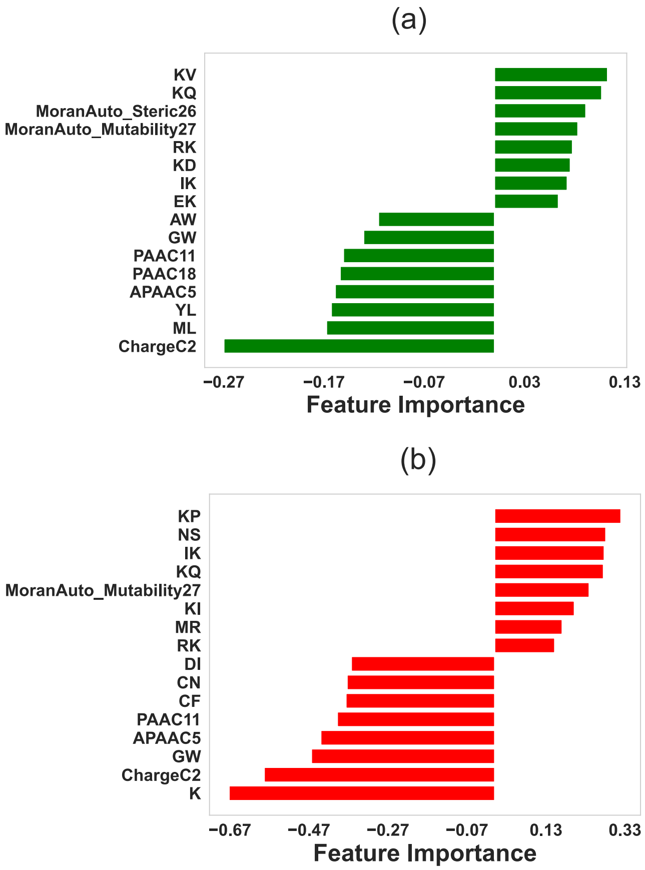


FIG. 2. Relative importance of different physicochemical properties in identifying the protein–protein interactions in the *Escherichia coli* (EC2) network using (a) the LASSO regression method and (b) the support vector machine (SVM). In computations, we utilized the following values for the hyperparameters: for LASSO, the hyperparameter was set to be λ = 0.004. For SVM, the hyperparameter C, which is calculated via the grid search optimization, is equal to C = 0.1. In both methods, the number of stratified shuffled cross-validation sets is equal to n = 18.

A. Feature selection for PPI network in *E. coli* (EC2)

Our feature selection analysis for the protein–protein interactions network in *E. coli* (EC2) has provided interesting insights. In particular, we observed that differences in dipeptide compositions between two proteins can exhibit both negative and positive correlations with protein–protein interactions (Fig. 2). The dipeptide composition here represents the fraction of each possible dipeptide (a sequence of two amino acids) within the peptide. Given that there are 20 standard amino acids, there are 20 × 20 = 400 possible dipeptides. In the dipeptide composition (DPC), a protein sequence is transformed into a fixed-length feature vector of size 400. Each element of this vector corresponds to one of the possible dipeptides and is calculated as the fraction of the total number of occurrences of that dipeptide in the sequence to the total number of all dipeptides

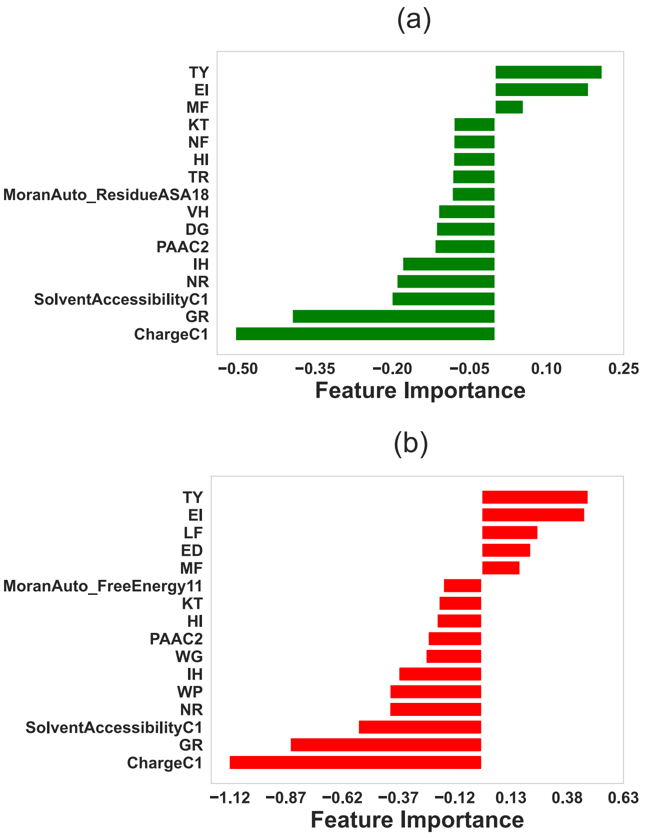


FIG. 3. Relative importance of different physicochemical features in identifying the protein–protein interactions in the *Saccharomyces cerevisiae* (SC5) network using (a) the LASSO regression method and (b) the support vector machine (SVM). In computations, we utilized the following values for the hyperparameters: for LASSO, the hyperparameter was set to be $\lambda = 0.004$. For SVM, the hyperparameter C, which is calculated via the grid search optimization, is equal to $C = 0.1$. In both methods, the number of stratified shuffled cross-validation sets is equal to $n = 15$.

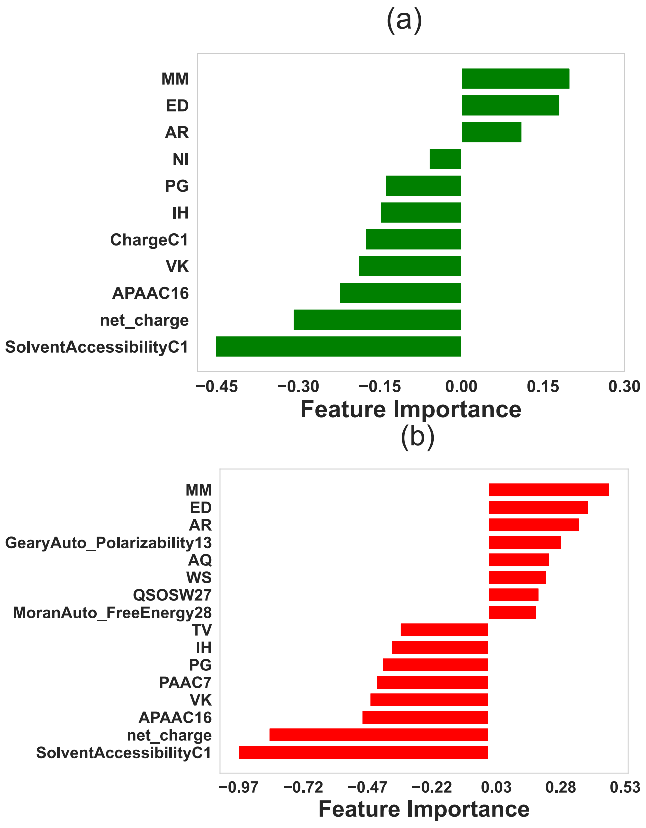


FIG. 4. Relative importance of different physicochemical features in identifying the protein–protein interactions in the *Schizosaccharomyces pombe* (SP) network using (a) the LASSO regression method and (b) the support vector machine. In computations, we utilized the following values for the hyperparameters: for LASSO, the hyperparameter was set to be $\lambda = 0.004$. For SVM, the hyperparameter C, which is calculated via the grid search optimization, is equal to $C = 0.1$. In both methods, the number of stratified shuffled cross-validation sets is equal to $n = 15$.

in the sequence. For a protein with N amino acids, the total number of dipeptides is $N - 1$, and we have

$$DPC(i) = \frac{\text{Number of occurrences of dipeptide } i}{N - 1}. \tag{8}$$

Thus, $DPC(i)$ for dipeptide i is a number between 0 and 1, which corresponds to the probability of finding the dipeptide in the given protein sequence.

It is known that dipeptides play a critical role in protein stability and function.[40–42] The impact of dipeptides on protein–protein interactions (PPIs) can be attributed to several factors. First, the specific arrangement of dipeptides can influence the structural conformations of proteins, affecting their interactions.[43] Second, dipeptide compositions may contain critical binding sites that facilitate or hinder PPI.[44,45] Third, the presence of charged amino acids in dipeptides can lead to electrostatic interactions that modulate PPI, especially between positively and negatively charged amino acids.[46] Fourth, differences in hydrophobicity within dipeptides can also influence interactions, particularly hydrophobic interactions.[47] Finally, the significance of dipeptide compositions on PPI may be context-dependent, varying based on organism biology, cellular environment, or the specific protein network under consideration. Since dipeptide compositions influence the stability of proteins, the differences in dipeptide compositions can also play a role in forming stable protein–protein interactions.

Our analysis also shows that other selected features, such as differences in Amphiphilic Pseudo Amino Acid Composition (APAAC) and Pseudo Amino Acid Composition (PAAC), as predicted by both LASSO and SVM feature selection methods, negatively correlate with protein–protein interactions. These differences may lead to structural incompatibility, altering the distribution of hydrophobic and hydrophilic residues along protein sequences and affecting binding site accessibility.[48] APAAC and PAAC variations might also correspond to hydrophobic–hydrophilic interactions and electrostatic repulsion, reducing the likelihood of stable binding.[49,50]

Moreover, the impact of APAAC and PAAC on PPIs can be context-specific, depending on the organism's biology and cellular environment. The cumulative effect of these factors can hinder the formation of stable protein complexes and weaken the interactions between proteins, leading to a negative impact on PPI.

Another feature that positively correlates with PPIs in EC2 (see Fig. 2) is differences in Moran's autocorrelation of the mutability and steric properties of the amino acids at certain distances. For example, *MoranAuto_Mutability27* refers to Moran's autocorrelation function of mutability for amino acids that are 27 positions apart in a protein sequence. The mutability of residues in proteins corresponds to the likelihood or rate at which the amino acid residues in a protein sequence change over time due to genetic mutations. This can be influenced by various factors, such as the structural and functional constraints on the protein as well as the physicochemical properties of the amino acids themselves. The Moran autocorrelation function, which is similar to Pearson's correlation between the mutability of residue i and residue $i + d$, is defined as[51]

$$M(d) = \frac{\frac{1}{N-d}\sum_{i=1}^{N-d}(Z_i - \bar{P})(Z_{i+d} - \bar{P})}{\frac{1}{N}\sum_{i=1}^{N}(Z_i - \bar{P})^2}, \qquad (9)$$

where $Z_i = \frac{P_i - \mu_P}{\sigma_P}$ is defined as the corresponding z-score of the property P_i. In addition, the mean $\mu_P = \frac{1}{20}\sum_{j=1}^{20} P_j$ and the variance $\sigma_P = \left(\frac{1}{20}\sum_{j=1}^{20}(P_j - \mu_P)^2\right)^{1/2}$ of a physicochemical property P are averaged over the 20 types of amino acids. In addition, $\bar{P} = \frac{1}{N}\sum_{i=1}^{N} Z_i$ is the average of the z-scores of the amino acid properties in the protein sequence. A positive Moran's value indicates that amino acids that are d positions apart in the protein sequence tend to have similar mutability values. While a negative, Moran's value indicates that amino acids that are d positions apart in the protein sequence tend to have dissimilar mutability values.

Alternatively, we can define the Geary autocorrelation function,[52]

$$G(d) = \frac{\frac{1}{2(N-d)}\sum_{i=1}^{N-d}(Z_i - Z_{i+d})^2}{\frac{1}{N-1}\sum_{i=1}^{N}(Z_i - \bar{P})^2}. \qquad (10)$$

In calculating Moran and Geary autocorrelation, employing z-scores means that amino acids with properties near the average value have minimal influence on the autocorrelation function. This approach essentially removes the ordinary variations of average properties and highlights the roles of unique or exceptional properties. Such insights are especially valuable for exploring the structure and function of proteins. These autocorrelation functions enable us to measure the spatial distribution of physicochemical properties along the protein sequence, taking into account both local and distant interactions. As these methods require centering the property P values by subtracting the mean, the resulting autocorrelation values can range from positive to negative. Our feature selection methods suggest that large differences between two proteins in terms of the distribution of mutability in the sequences correlate with protein–protein interactions. It could mean that proteins with similar patterns of mutability at a distance of 27 amino acids are more likely to interact with each other. This could potentially be related to how the proteins fold and fit together, as similar patterns of mutability might lead to complementary structural features that facilitate the interaction. Alternatively, it could be related to functional similarities between the interacting proteins, such that they are subject to similar evolutionary pressures that affect their mutability in a coordinated way. Further analysis and validation would be needed to fully understand the underlying mechanisms behind this association.

B. Feature selection for PPIs in *S. cerevisiae* (SC5) and *S. pombe* (SP)

For SP and SC5 systems, our analysis of the protein–protein interaction networks using both LASSO and SVM methods again predicts that differences in dipeptide compositions exhibit the strongest correlation with protein–protein interactions; see Figs. 3 and 4. Thus, the role of dipeptide composition in PPIs is not context-specific, indicating that it might be a universal phenomenon valid across all organisms. To test this idea, we applied our method to three different multicellular organisms (see Figs. S1, S2, and S3 in the supplementary material), and it was found that differences in dipeptide compositions also strongly correlate with the protein–protein interactions for multicellular organisms, supporting the hypothesis of the universality of dipeptide compositions as a predictor of PPIs.

Moreover, our computational approach predicts that differences in solvent accessibility between two proteins negatively correlate with PPIs. Solvent accessibility measures how accessible the individual amino acids are to the solvent molecules (typically water) in the protein's environment.[53] Differences in solvent accessibility between two proteins can have diverse implications. Steric hindrance may arise when exposed regions of one protein obstruct the buried regions of the other, hindering effective interaction. Distinct hydrophobic and hydrophilic regions influenced by solvent accessibility may impact the affinity of hydrophobic interactions. Surface complementarity might play a role here; proteins with complementary solvent-accessible surfaces are more likely to form stable interactions. Electrostatic interactions can also be influenced by charged residue exposure, leading to attractive or repulsive forces. In addition, solvent accessibility may influence conformational changes, affecting the propensity for structural alterations upon interaction. Overall, these factors collectively contribute to the potential impact of solvent accessibility on protein–protein interactions.

We also observed that for the EC2 network, there are considerable differences between the predictions of SVM and LASSO. To further analyze this observation, we performed Principal Component Analysis (PCA) for all networks (see Figs. S4 and S5 in the supplementary material) and realized that only for the EC2 network can PCA relatively well separate interacting from non-interacting pairs [see Fig. S4(a) in the supplementary material]. The data distribution in the EC2 network, characterized by non-uniformity and higher variance, might have favored the SVM approach. SVM's capability to construct a hyperplane that maximizes the margin between classes would be particularly beneficial in such a complex dataset. On the other hand, when PCA shows that the data points are not distinctly separable, it suggests that neither LASSO nor SVM can easily distinguish the classes without potentially complex transformations. This scenario could lead to both methods performing similarly.

C. Prediction of protein–protein interactions using selected features

After extracting the most important physicochemical properties of each PPI network, our objective is to utilize those features to accurately predict protein–protein interactions. The performance metrics used for comparison include *Accuracy*, *Recall*, Matthews Correlation Coefficient (*MCC*), *F*1 score, and ROC curves, as described above. We employed the SVM method for classifying interacting vs non-interacting protein pairs. Results from unicellular organisms are presented in Table II. One can see that selected features from the SVM method generally lead to slightly higher metrics, suggesting a better description of correlations. The corresponding *ROC* curves are presented in Figs. S6 and S7 in the supplementary material.

It is important to note that although our dataset contains an equal number of interacting and non-interacting pairs, the complete PPI data for all organisms is highly imbalanced.[54] For N proteins in an organism, the total number of possible protein–protein pairs is $\sim \frac{N^2}{2}$. This indicates that in real biological systems, there are far fewer interacting pairs than non-interacting ones, a discrepancy that poses significant challenges in correctly predicting the interactions. Despite these challenges, our *F*1 score closely aligns with our recall metrics (Table II), suggesting that the model maintains a balanced performance regarding false positives (*FP*) and false negatives (*FN*). The near-equal rates of *FP* and *FN* imply that our approach neither excessively predicts protein–protein interactions where none exist nor fails to identify true interactions. Moreover, the values of AUC are generally high, indicating that the SVM classifier is very effective in distinguishing between interacting and non-interacting protein–protein pairs.

The prediction of the SVM model for multicellular organisms is shown in Table S3 in the supplementary material. One can see that our machine-learning models yield better prediction metrics for unicellular organisms compared to multicellular organisms such as AT, MM, and DM2. Regardless of the limited protein–protein interaction data in this study, the relatively poor performance of multicellular organisms can be attributed to two important factors. First, multicellular organisms are inherently more complex than unicellular ones. This complexity increases the difficulty of accurately predicting protein–protein interactions (PPIs), as there are more players to consider. Furthermore, multicellular organisms also exhibit more complex genetic and epigenetic landscapes, which might affect protein production and interactions differently depending on the tissue or developmental stage. Second, in multicellular organisms, different cells can express different subsets of proteins, leading to diverse interaction networks that are highly context-dependent. This variability can complicate the prediction of interactions across all cell types.

TABLE II. Results of feature selection for protein–protein interactions in *Escherichia coli* (EC2), *Saccharomyces cerevisiae* (SC5), and *Schizosaccharomyces pombe* (SP) networks. Comparison of accuracy, recall, Matthews's correlation coefficient (*MCC*), *F1* score, and AUC for the trained baseline models (SVM) using selected features from SVM and LASSO. Each metric reflects the average value among 15 test cross-fold validation sets. A standard splitting of 80/20 (training/test) was applied for each fold.

Network	Feature selection	Accuracy	Recall	*MCC*	*F1* score	AUC
EC2	LASSO	0.8	0.8	0.61	0.77	0.84
	SVM	0.81	0.81	0.64	0.79	0.87
SC5	LASSO	0.72	0.72	0.44	0.74	0.81
	SVM	0.75	0.75	0.50	0.76	0.82
SP	LASSO	0.68	0.67	0.34	0.67	0.72
	SVM	0.69	0.69	0.38	0.69	0.76

D. Illustrative example 1: Two-component PhoB–PhoR system in *E. coli*

To illustrate our computational approach, let us apply it to identifying the protein–protein interactions in a two-component system in *E. coli* bacteria. We specifically focus on interactions of histidine kinase PhoR proteins with transcriptional response regulators PhoB proteins.[3,4] The PhoB–PhoR system in *E. coli* functions to detect low phosphate levels in the environment. When the amount of phosphate species in the medium is low, the PhoR proteins activate the PhoB proteins. The activated (phosphorylated) PhoB proteins then activate genes that help the bacteria absorb more phosphate molecules and use them more efficiently. This system ensures that *E. coli* gets enough phosphate molecules, a vital nutrient, even when they are limited in their surroundings.

Our objective is to show that differences in certain physicochemical features between a PhoR and a PhoB correlate with their abilities to interact. We chose an arbitrary response regulator, NarL, for which it is known that it does not interact with PhoR. The response regulator NarL is part of the NarL-NarX/NarQ two-component system. While both protein systems (PhoB–PhoR and NarL-NarX/NarQ) are two-component regulatory systems in *E. coli*, they are tuned to detect and respond to different environmental signals and, thus, they have distinct regulatory outcomes. In Fig. 5, we compared three proteins in terms of the contributions of four dipeptide compositions: VV, KK, IK, and EE. One can see that PhoR and PhoB contain different compositions of the corresponding dipeptides, while PhoR and NarL are similar. This suggests that strong differences in the dipeptide compositions correlate with the abilities of these proteins to interact.

The interactions between histidine kinases (HK) and response regulators (RR) in two-component systems (TCS) are governed by specific protein–protein properties that can be significantly affected by dipeptide and amino acid composition differences. The following molecular picture might be proposed: The specific amino acid sequences and dipeptides at the interaction interfaces of HK and RR determine their cognate pairings, ensuring that a particular HK interacts with its intended RR species. Changes in these sequences could disrupt this specificity. The efficiency of phosphoryl transfer between the conserved histidine of HK and the conserved aspartate of RR can be influenced by the surrounding amino acids. For instance, any changes in the nearby residues that might hinder the approach of RR to HK could also affect this transfer.

Furthermore, the strength or affinity of the interaction between HK and RR proteins can be controlled by the nature of the amino acids and dipeptides at the binding interface. Hydrophobic, ionic, and hydrogen bond interactions contribute to this binding, and changes in these residues can either enhance or diminish the affinity.

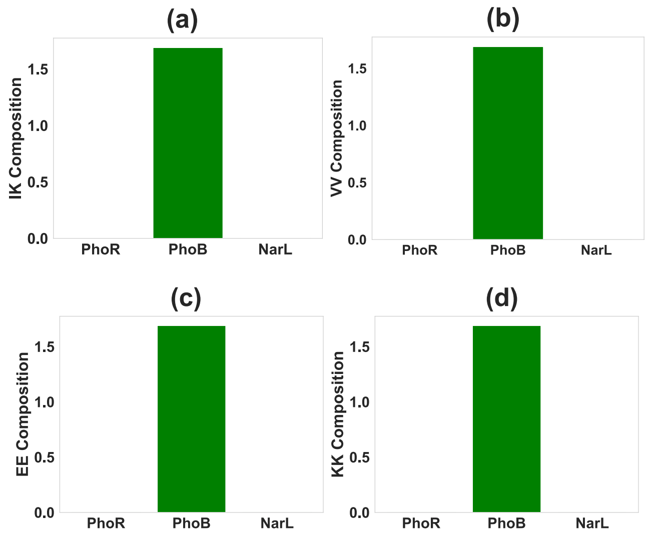


FIG. 5. Comparison of charged dipeptide compositions of histidine kinase PhoR with two response regulators, PhoB (interacting) and NarL (non-interacting). (a) Dipeptide IK composition, (b) dipeptide VV composition, (c) dipeptide EE composition, and (d) dipeptide KK composition.

HK and RR undergo conformational changes during their interaction. Amino acid or dipeptide composition differences can impact the protein's ability to undergo necessary conformational changes, which can in turn affect their interaction dynamics. Over evolutionary timescales, if one protein (either HK or RR) changes its amino acid composition and modifies its interaction potential, the interacting partner might co-evolve to accommodate or compensate for this change, maintaining the same interaction as before the change. This might be the main reason for strong correlations between the differences in the amino-acid and dipeptide compositions and the abilities of proteins to interact.

E. Illustrative example 2: Barnase–barstar complex in *B. amyloliquefaciens*

As another example, we consider the interaction of barnase and barstar proteins, which are, respectively, an extracellular ribonuclease and its intracellular inhibitor, both produced by the bacterium *Bacillus amyloliquefaciens*.[55] Although barnase functions extracellularly to degrade RNA from other sources, any intracellular activity of barnase can degrade the cell's own RNA. This RNA degradation, in turn, can disrupt various cellular processes, including, but not limited to, protein synthesis. Barstar protein, by inhibiting barnase, effectively prevents this damaging ribonuclease activity. This inhibitory mechanism allows the organism to utilize the ribonuclease as an effective defense tool externally without compromising its cellular physiology. The interaction between barnase and barstar proteins is extremely strong, with a dissociation equilibrium constant of $K_d = 10^{-14}$M.

Multiple studies, utilizing double mutant cycle analysis, have been conducted to determine how residues of barnase and barstar contribute to their binding interaction.[56,57] It is known that the binding hotspots of the barnase–barstar complex are located between residues 29 and 42 of the barstar molecule. Analyzing the binding energy of wild-type (WT) barnase with two mutants of barstar yields two different results. When the amino acid tyrosine at position 29 is replaced by phenylalanine (Y29F), the WT barnase protein interacting with Y29F mutated barstar yields a dissociation constant (K_d) value of 0.008×10^{-12}M. Conversely, when tyrosine is replaced by alanine (Y29A), the interaction results in a significantly higher K_d of 3.5×10^{-12}M. In Fig. 6, we compared WT barnase and barstar species with two mutants of barstar in terms of the composition of two dipeptides, EF and FY. As one can see, the frequency of these dipeptides in WT barnase and barstar is zero, while introducing a mutation at location 29 of barstar results in the formation of both EF and FY. Since the other mutant of barstar Y29A weekly interacts with the WT barstar protein, it can be argued that this difference

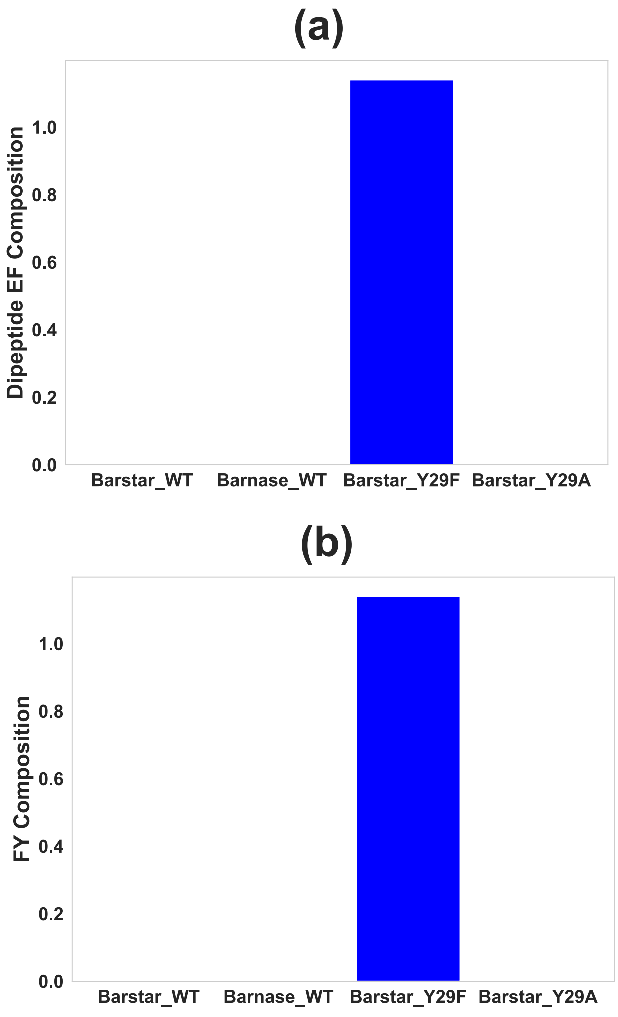


FIG. 6. Comparison of dipeptide compositions in WT barnase and two mutants of barstar: (a) dipeptide EF composition and (b) dipeptide FY composition. WT barnase interacts with the barstar mutant Y29F, exhibiting a dissociation constant (K_d) of 0.008×10^{-12}M. Conversely, the interaction between WT barnase and the barstar mutant Y29A shows a significantly higher K_d of 3.5×10^{-12}M. The data were obtained from Ref. 57.

in dipeptide composition contributes to a significant decrease in the dissociation constant ($K_d = 0.008 \times 10^{-12}$M) between the WT barnase and the Y29F mutant of barstar.

IV. DISCUSSION

In this study, we introduced a novel computational method for assessing the probability of protein–protein interactions in any biological system. It is based on the idea that certain physicochemical properties correlate with these interactions. We performed our analysis on six different datasets belonging to various domains of life: three types of microorganisms (*Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, and *Escherichia coli*), mammal species (Mus musculus), insects (*Drosophila melanogaster*), and plants (Arabidopsis thaliana). Despite their differences in complexity, all these organisms rely on PPI networks to perform essential biological functions. These organisms have well-characterized genomes and proteomes, enabling the study of their protein–protein interactions. Examining their PPI networks allows us to gain more microscopic insights into the cellular processes and functions specific to each organism.

We utilized two feature selection methods, LASSO and SVM, to select the most important set of physicochemical descriptors, which have a positive or negative correlation with the protein–protein interactions. These methods reveal that, for all organisms, the differences between two proteins in terms of dipeptide compositions are critically important for identifying PPIs. This is a universal feature that seems to work for all the organisms that we investigated. Furthermore, our feature selection methods suggest that there are other physicochemical features specific to each organism that contribute to correlations with protein–protein interactions. These types of features, however, are context-dependent.[58] They might be specific to the organism's biology, the cellular environment, or the specific protein network being considered. Different organisms or cell types might have distinct requirements for protein interactions, leading to different preferences for certain physicochemical properties.[59,60]

The impact of correlations between dipeptide compositions and PPIs can be attributed to several sources. First, it could be related to the structural conformations of proteins. Dipeptides are short sequences of two amino acids, and their specific arrangement can influence the overall secondary and tertiary structure of proteins.[61] The three-dimensional structure of proteins is important in determining how they interact with other proteins.[62] Differences in dipeptide compositions might also lead to variations in protein folding,[63,64] which, in turn, can affect their ability to interact with other proteins. Second, dipeptide compositions may contain specific amino acid pairs that serve as critical binding sites for PPIs. These binding sites can mediate physical interactions between proteins and are essential for the formation of protein complexes. Variations in dipeptide compositions can alter the presence or accessibility of these binding sites, influencing the potential for PPIs. Third, amino acids in dipeptide compositions can have different charges, such as positively charged (e.g., lysine), negatively charged (e.g., aspartic acid), or neutral (e.g., alanine). These charged amino acids can engage in electrostatic interactions with other proteins, either promoting or inhibiting their interactions. Dipeptides with specific combinations of charged amino acids may create favorable or unfavorable electrostatic environments for PPI. Finally, some dipeptide compositions may contain hydrophobic amino acids, which tend to cluster together in the protein's core,[65,66] while others may have hydrophilic amino acids exposed on the protein's surface.[67,68] Differences in dipeptide compositions can lead to variations in hydrophobic and hydrophilic regions, influencing protein–protein interactions, especially those that involve hydrophobic interactions, and dipeptide composition can be targeted to affect protein–protein interactions.[69]

It was also argued that our theoretical approach provides a new tool for investigating the mechanisms of biological processes. Understanding the physicochemical properties of the interface formed by protein–protein association might help to clarify the mechanisms of the formation of protein interaction networks on the one hand and to design molecules that can engage with a given interface and thereby control protein function on the other hand. For example, synthetic molecules that resemble the chemical structure of proteins, called peptidomimetics, can be used to inhibit protein–protein interactions associated with diseases.[70] This means that PPIs might be excellent targets for drug development.[71] By considering the specific physicochemical features of each PPI network, our computational approach can capture the network-specific patterns and relationships that govern protein–protein interactions in different biological contexts. This allows for a more accurate and context-specific prediction of protein–protein interactions, enhancing our understanding of how these interactions contribute to cellular processes and functions in each organism.

It is important to note that protein–protein interactions are highly complex and multifaceted processes, involving various molecular forces and structural features. Machine-learning models can help identify patterns and correlations in large datasets, but they may not capture the full microscopic intricacies of protein–protein interactions. As with any predictive model, it is essential to interpret the results cautiously and complement them with experimental validations and further analysis to gain a deeper understanding of the underlying biology. In addition, considering other physicochemical properties and features in combination with solvent accessibility can lead to a more comprehensive understanding of protein–protein interactions. It will be interesting to investigate how the differences in tri-peptide compositions (the frequency of a given set of three adjacent amino acids in a sequence) are correlated with protein–protein interactions. Our method can be applied to PPI systems in humans, including virus–host systems[72] and cancer.[9,73]

SUPPLEMENTARY MATERIAL

See the supplementary material for feature selection in determining the protein–protein interactions in multicellular organisms, description of selected features for all organisms, model evaluation for prediction of protein–protein interactions in multicellular organisms, principal component analysis results, and Receiver Operating Characteristic (ROC) curves.

ACKNOWLEDGMENTS

The work was supported by the Welch Foundation (C-1559), the NIH (Grant Nos. R01GM148537 and R01HL157714-04), the

NSF (Grant No. CHE-2246878), and the Center for Theoretical Biological Physics sponsored by the NSF (Grant No. PHY-2019745).

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Hamid Teimouri: Conceptualization (equal); Data curation (equal); Formal analysis (equal); Investigation (equal); Validation (equal); Visualization (equal); Writing – original draft (equal). **Angela Medvedeva**: Data curation (equal); Formal analysis (equal); Investigation (equal); Validation (equal); Writing – review & editing (equal). **Anatoly B. Kolomeisky**: Conceptualization (equal); Funding acquisition (equal); Project administration (equal); Resources (equal); Supervision (equal); Writing – review & editing (equal).

DATA AVAILABILITY

The data that support the findings of this study are available within the article and its supplementary material.

REFERENCES

[1] S. Jones and J. M. Thornton, Proc. Natl. Acad. Sci. U. S. A. **93**, 13 (1996).
[2] H. F. Lodish, *Molecular Cell Biology* (Macmillan, 2008).
[3] M. G. Lamarche, B. L. Wanner, S. Crepin, and J. Harel, FEMS Microbiol. Rev. **32**, 461 (2008).
[4] K. S. Choudhary, J. A. Kleinmanns, K. Decker, A. V. Sastry, Y. Gao, R. Szubin, Y. Seif, and B. O. Palsson, Msystems **5**, e00980 (2020).
[5] H. Lu, Q. Zhou, J. He, Z. Jiang, C. Peng, R. Tong, and J. Shi, Signal Transduction Targeted Ther. **5**, 213 (2020).
[6] B. T. MacDonald, K. Tamai, and X. He, Dev. Cell **17**, 9 (2009).
[7] G. Calabrese, C. Molzahn, and T. Mayor, J. Biol. Chem. **298**, 102062 (2022).
[8] A. Chakraborty, S. Mitra, M. Bhattacharjee, D. De, and A. J. Pal, Med. Novel Technol. Devices **18**, 100228 (2023).
[9] G. Kar, A. Gursoy, and O. Keskin, PLoS Comput. Biol. **5**, e1000601 (2009).
[10] T. L. Nero, C. J. Morton, J. K. Holien, J. Wielens, and M. W. Parker, Nat. Rev. Cancer **14**, 248 (2014).
[11] S. Fields and O.-K. Song, Nature **340**, 245 (1989).
[12] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki, Proc. Natl. Acad. Sci. U. S. A. **98**, 4569 (2001).
[13] Y. Ho, A. Gruhler, A. Heilbut, G. D. Bader, L. Moore, S.-L. Adams, A. Millar, P. Taylor, K. Bennett, K. Boutilier *et al.*, Nature **415**, 180 (2002).
[14] M. Mann, R. C. Hendrickson, and A. Pandey, Annu. Rev. Biochem. **70**, 437 (2001).
[15] X. Xu, Y. Song, Y. Li, J. Chang, L. An, and L. An, Protein Expression Purif. **72**, 149 (2010).
[16] G. Rigaut, A. Shevchenko, B. Rutz, M. Wilm, M. Mann, and B. Séraphin, Nat. Biotechnol. **17**, 1030 (1999).
[17] C. Von Mering, R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields, and P. Bork, Nature **417**, 399 (2002).
[18] M. D. Dyer, T. Murali, and B. W. Sobral, Bioinformatics **23**, i159 (2007).
[19] M. Singhal and H. Resat, BMC Bioinf. **8**, 199 (2007).
[20] C. Chen, Q. Zhang, B. Yu, Z. Yu, P. J. Lawrence, Q. Ma, and Y. Zhang, Comput. Biol. Med. **123**, 103899 (2020).
[21] O. Keskin, N. Tuncbag, and A. Gursoy, Chem. Rev. **116**, 4884 (2016).
[22] J. Shen, J. Zhang, X. Luo, W. Zhu, K. Yu, K. Chen, Y. Li, and H. Jiang, Proc. Natl. Acad. Sci. U. S. A. **104**, 4337 (2007).
[23] Y. Guo, L. Yu, Z. Wen, and M. Li, Nucleic Acids Res. **36**, 3025 (2008).
[24] Y.-A. Huang, Z.-H. You, X. Chen, K. Chan, and X. Luo, BMC Bioinf. **17**, 184 (2016).
[25] Y. Murakami and K. Mizuguchi, Biophys. Rev. **14**, 1393 (2022).
[26] J. Bernett, D. B. Blumenthal, and M. List, Briefings Bioinf. **25**, bbae076 (2024).
[27] J. Wu, B. Liu, J. Zhang, Z. Wang, and J. Li, BMC Bioinf. **24**, 473 (2023).
[28] D.-S. Cao, Q.-S. Xu, and Y.-Z. Liang, Bioinformatics **29**, 960 (2013).
[29] T. K. Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, S. Ramabadran, R. Chaerkady, A. Pandey *et al.*, "Human protein reference database–2009 update," Nucleic Acids Res. **37**, D767 (2009).
[30] M. Oesterheld, P. Pagel, A. Ruepp, H.-W. Mewes *et al.*, Nucleic Acids Res. **34**, D436–D441 (2006).
[31] L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie, and D. Eisenberg, Nucleic Acids Res. **32**, D449 (2004).
[32] K.-H. Chen, T.-F. Wang, and Y.-J. Hu, BMC Bioinf. **20**, 1 (2019).
[33] W. S. Noble, Nat. Biotechnol. **24**, 1565 (2006).
[34] R. Tibshirani, J. Royal Stat. Soc. Ser. B: Stat. Methodol. **58**, 267 (1996).
[35] H. Teimouri, A. Medvedeva, and A. B. Kolomeisky, J. Chem. Inf. Model. **63**, 1723 (2023).
[36] J. R. Bradford and D. R. Westhead, Bioinformatics **21**, 1487 (2005).
[37] C. Goutte and E. Gaussier, *European Conference on Information Retrieval* (Springer, 2005), pp. 345–359.
[38] Q. Zou, S. Xie, Z. Lin, M. Wu, and Y. Ju, Big Data Res. **5**, 2 (2016).
[39] D. Chicco and G. Jurman, BMC Genomics **21**, 6 (2020).
[40] Y. Ding, Y. Cai, G. Zhang, and W. Xu, FEBS Lett. **569**, 284 (2004).
[41] K. Guruprasad, B. B. Reddy, and M. W. Pandit, Protein Eng., Des. Sel. **4**, 155 (1990).
[42] H.-M. Zhang, J.-F. Li, M.-C. Wu, H.-L. Shi, and C.-D. Tang, Ann. Microbiol. **63**, 307 (2013).
[43] M. Ghadimi, K. Khalifeh, and E. Heshmati, Amino Acids **49**, 1641 (2017).
[44] M. Tang, L. Wu, X. Yu, Z. Chu, S. Jin, and J. Liu, Front. Genet. **12**, 784863 (2021).
[45] M. Tang, L. Wu, X. Yu, Z. Chu, S. Jin, and J. Liu, eLife **12**, e82819 (2023).
[46] N.-Z. Xie, Q.-S. Du, J.-X. Li, and R.-B. Huang, PLoS One **10**, e0137113 (2015).
[47] A. Mishra and R. Sankararamakrishnan, Front. Mol. Biosci. **8**, 706002 (2021).
[48] Y. Liu, J. Li, X. Li, S. C. Li, L. Wong, and S. T. C. Wong, Front. Aging Neurosci. **13**, 699024 (2021).
[49] M. Radhakrishna, J. Grimaldi, G. Belfort, and S. K. Kumar, Langmuir **29**, 8922 (2013).
[50] C. N. Pace, H. Fu, K. L. Fryar, J. Landua, S. R. Trevino, B. A. Shirley, M. M. Hendricks, S. Iimura, K. Gajiwala, J. M. Scholtz, and G. R. Grimsley, J. Mol. Biol. **408**, 514 (2011).
[51] Z.-R. Li, H. H. Lin, L. Han, L. Jiang, X. Chen, and Y. Z. Chen, Nucleic Acids Res. **34**, W32 (2006).
[52] S. A. Ong, H. H. Lin, Y. Z. Chen, Z. R. Li, and Z. Cao, BMC Bioinf. **8**, 1 (2007).
[53] G. D. Rose, A. R. Geselowitz, G. J. Lesser, R. H. Lee, and M. H. Zehfus, Science **229**, 834 (1985).
[54] C.-Y. Yu, L.-C. Chou, and D. T.-H. Chang, BMC Bioinf. **11**, S3 (2010).
[55] R. W. Hartley, Trends Biochem. Sci. **14**, 450 (1989).
[56] A. Horovitz, Folding Des. **1**, R121 (1996).
[57] G. Schreiber and A. R. Fersht, J. Mol. Biol. **248**, 478 (1995).
[58] P. Zhou, Q. Miao, F. Yan, Z. Li, Q. Jiang, L. Wen, and Y. Meng, Mol. Omics **15**, 280 (2019).
[59] S. Mika and B. Rost, PLoS Comput. Biol. **2**, e79 (2006).
[60] S. Sultana, M. Abdullah, J. Li, M. Hochstrasser, and A. H. Kachroo, Genetics **225**, iyad117 (2023).
[61] T. Liu, Y. Qin, Y. Wang, and C. Wang, Int. J. Mol. Sci. **17**, 15 (2015).
[62] X. Liu, Y. Luo, P. Li, S. Song, and J. Peng, PLoS Comput. Biol. **17**, e1009284 (2021).
[63] K. Patil and U. Chouhan, Curr. Bioinf. **14**, 688 (2019).
[64] J. X. Guo and N. N. Rao, Adv. Mater. Res. **378–379**, 157 (2011).

[65] E. Van Dijk, A. Hoogeveen, and S. Abeln, PLoS Comput. Biol. **11**, e1004277 (2015).

[66] M. Banach, P. Fabian, K. Stapor, L. Konieczny, and I. Roterman, Biomolecules **10**, 767 (2020).

[67] F. C. Almeida, K. Sanches, R. Pinheiro-Aguiar, V. S. Almeida, and I. P. Caruso, Front. Mol. Biosci. **8**, 706002 (2021).

[68] C. Strub, C. Alies, A. Lougarre, C. Ladurantie, J. Czaplicki, and D. Fournier, BMC Biochem. **5**, 9 (2004).

[69] L. Li, L. Xie, R. Zheng, and R. Sun, Front. Chem. **9**, 739791 (2021).

[70] P. Gupta, S. Srivastava, and P. Kumar, Curr. Protein Pept. Sci. **20**, 329 (2019).

[71] L. Alzyoud, R. A. Bryce, M. Al Sorkhy, N. Atatreh, and M. A. Ghattas, Sci. Rep. **12**, 7975 (2022).

[72] A. F. Brito and J. W. Pinney, Front. Microbiol. **8**, 1557 (2017).

[73] X. Zhou, B. Park, D. Choi, and K. Han, BMC Genomics **19**, 568 (2018).