

## RESEARCH ARTICLE

# Robust principal component analysis-based prediction of protein-protein interaction hot spots

Divya Sitani<sup>1,2</sup>  | Alejandro Giorgetti<sup>3,4</sup>  | Mercedes Alfonso-Prieto<sup>3,5</sup>  | Paolo Carloni<sup>1,3,6,7</sup> 

<sup>1</sup>JARA-Institute: Molecular Neuroscience and Neuroimaging, Institute for Neuroscience and Medicine INM-11/JARA-BRAIN Institute JBI-2, Forschungszentrum Jülich GmbH, Jülich, Germany

<sup>2</sup>Department of Biology, RWTH Aachen University, Aachen, Germany

<sup>3</sup>Institute for Advanced Simulations IAS-5 / Institute for Neuroscience and Medicine INM-9, Computational Biomedicine, Forschungszentrum Jülich GmbH, Jülich, Germany

<sup>4</sup>Department of Biotechnology, University of Verona, Verona, Italy

<sup>5</sup>Cécile and Oskar Vogt Institute for Brain Research, University Hospital Düsseldorf, Medical Faculty, Heinrich Heine University Düsseldorf, Düsseldorf, Germany

<sup>6</sup>Department of Physics, RWTH Aachen University, Aachen, Germany

<sup>7</sup>JARA-HPC, IAS-5/INM-9 Computational Biomedicine, Forschungszentrum Jülich GmbH, Jülich, Germany

## Correspondence

Divya Sitani and Paolo Carloni, JARA-Institute: Molecular Neuroscience and Neuroimaging, Institute for Neuroscience and Medicine INM-11/JARA-BRAIN Institute JBI-2, Forschungszentrum Jülich GmbH, 52428 Jülich, Germany.  
Email: d.sitani@fz-juelich.de (D. S.) and p.carloni@fz-juelich.de (P. C.)

## Abstract

Proteins often exert their function by binding to other cellular partners. The hot spots are key residues for protein-protein binding. Their identification may shed light on the impact of disease associated mutations on protein complexes and help design protein-protein interaction inhibitors for therapy. Unfortunately, current machine learning methods to predict hot spots, suffer from limitations caused by gross errors in the data matrices. Here, we present a novel data pre-processing pipeline that overcomes this problem by recovering a low rank matrix with reduced noise using Robust Principal Component Analysis. Application to existing databases shows the predictive power of the method.

## KEYWORDS

F1-score, feature selection, hot spot residues, imbalanced datasets, machine learning, noiseless data matrices, protein-protein interactions, robust PCA (principal component analysis)

## 1 | INTRODUCTION

Proteins rarely act alone.<sup>1,2</sup> Most often, they interact with other proteins and ligands to carry out biological processes, from metabolism to signal transduction, to cellular motion and to synaptic transmission.<sup>3,4</sup> Protein-protein interactions (PPIs) occur in specific areas of the protein surface

known as protein-protein interfaces. Although PPIs usually involve many residues on two opposite interfaces,<sup>5,6</sup> often only a few residues (the so-called hot spots) contribute significantly to the overall free energy of binding ( $\Delta G_{\text{binding}}$ ).<sup>7</sup> Mutations of these hot spots impact PPIs and entire biological pathways, leading to severe diseases including cancer and various neurological disorders.<sup>8</sup> Drug molecules that interact with these hot

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2021 The Authors. *Proteins: Structure, Function, and Bioinformatics* published by Wiley Periodicals LLC.

spots may interfere with PPIs and with the downstream pathways they mediate.<sup>9,10</sup> Thus, predicting hot spots is crucial to understand the effect of disease associated mutations on PPIs and drug discovery.<sup>11</sup>

Experimental Alanine Scanning Mutagenesis (ASM) identifies hot spots experimentally by systematically mutating each interface residue to alanine and measuring the change in  $\Delta G_{\text{binding}}$  ( $\Delta\Delta G_{\text{binding}}$ ).<sup>7,12</sup> If  $\Delta\Delta G_{\text{binding}} \geq 2.0$  kcal/mol, the residue is defined as a hot spot, otherwise as a “null spot”.<sup>7,13</sup> Hot spots tend to be highly enriched in disease causing mutations as compared to null spots.<sup>14,15</sup> These mutations significantly affect the protein structure, function and protein-protein complex thermodynamics when occurring in hot spots.<sup>14</sup> Experimental ASM is rather expensive and time consuming. Alternatively, computational approaches may also be used.<sup>16,17</sup> Some of them rely on different scoring functions (either based on energy or statistical functions)<sup>18,19</sup> or on molecular dynamic simulations.<sup>20-23</sup> These approaches have so far turned out to be partially successful, identifying accurately some, but not all hot spots in a variety of protein-protein complexes.<sup>24,25</sup> In the past years, approaches based on machine learning algorithms including Random Forests,<sup>26</sup> Support Vector Machines,<sup>27</sup> Neural Networks,<sup>28</sup> Ensemble Learning,<sup>29</sup> Bayesian networks,<sup>30,31</sup> have become quite popular to predict hot spots.<sup>32-39</sup> These methods use a data matrix that contains protein sequence- and structure-based features. Often such data matrices contain entries that can be corrupted by errors associated with experimental issues, computational issues and/or human oversight.<sup>40,41</sup> Of course, this adversely affects the predictive power of the current hot spot prediction algorithms. Therefore, it would be highly desirable to use an approach devoid of this problem. Here, we address this issue by pre-processing the data matrix using robust principal component analysis (RPCA). RPCA is a variant of the traditional PCA method and it is particularly useful for data matrices that may contain corrupted entries, such as the ones considered here. In reference [41], the authors showed that a noisy matrix  $D$  can be decomposed exactly into a low rank noiseless matrix  $A$  and a sparse matrix  $S$ , regardless of the number of corrupted or missing entries (ie, robustly). The thus-obtained low rank matrix is then the new data matrix for the identification of hot spots. We apply this method, that we call RBHS, or RPCA Based approach for PPI Hot Spot prediction, to the curated benchmark datasets HB-34 and BID-18.<sup>24</sup> RBHS turns out to outperform several state-of-the-art approaches for hot spot prediction.

## 2 | MATERIALS AND METHODS

We first describe the preparation of the dataset in section 2.1 and then the steps of our approach are explained in section 2.2. The final output of the workflow is a set of metrics that describes the predictive power of the approach used (see section 3).

### 2.1 | Preparation of the data for the workflow: Datasets and Features

We reproduced the data matrix for our workflow similar to ref. [24]. We used the HB-34 dataset<sup>24</sup> to construct the training data matrix,

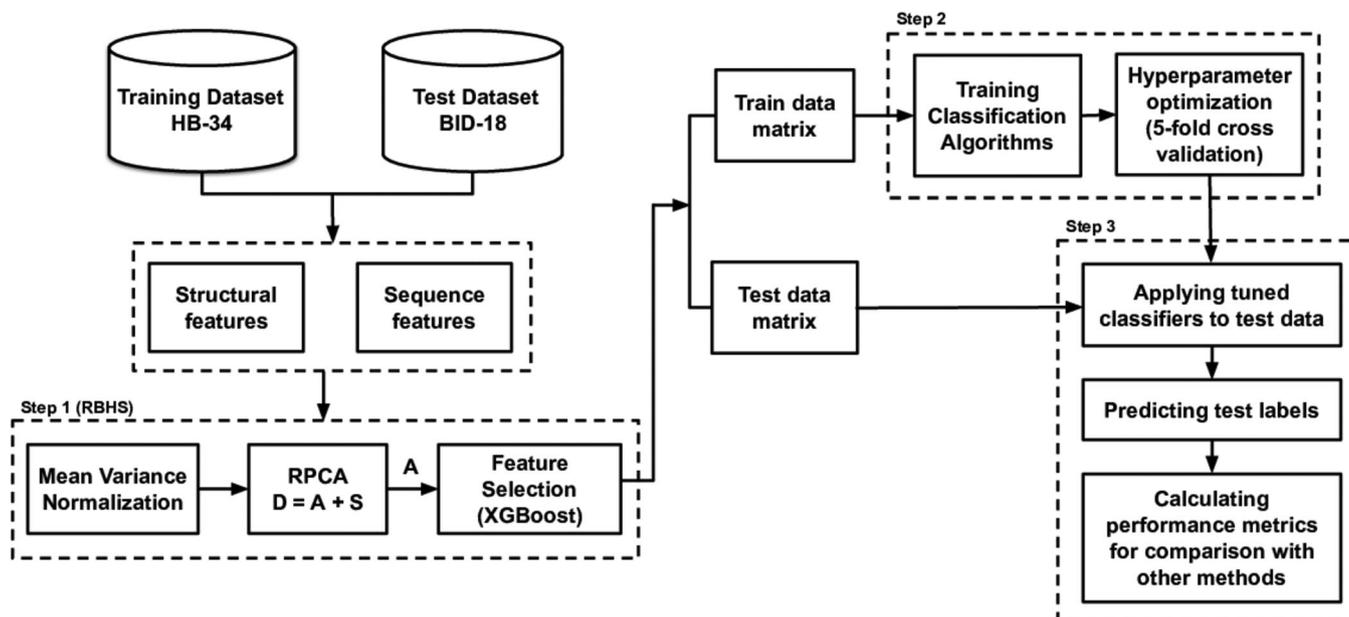
whereas the BID-18 dataset was used to construct the test data matrix.<sup>24</sup> HB-34 consists of 34 protein complexes with 313 characterized interface residues, out of which 133 are hot spots. BID-18 includes 18 protein complexes and 126 interface residues, out of which 39 are hot spots. The authors of the benchmark HB-34 dataset<sup>24</sup> extracted the alanine-mutation data from four databases: Alanine scanning energetics (ASEdb),<sup>42</sup> SKEMPI database,<sup>43</sup> Ab+ data,<sup>44</sup> and Alexov\_sDB.<sup>45</sup> Then, they excluded the complexes present in the BID dataset<sup>46</sup> and removed the redundant proteins, obtaining a benchmark of 34 protein complexes (HB-34). Afterwards, the authors generated an independent test dataset, BID-18, by selecting complexes from the BID database that are non-homologous to those in the training dataset (HB-34). Therefore, the authors of reference [24] stated that the HB-34 and BID-18 datasets are completely independent. However, we have further analyzed the two datasets using the CD-HIT-2D webserver,<sup>47</sup> which searches for protein sequences that are similar within the following, stringent criteria: the sequence identity should be larger than 40% and the coverage larger than 20% of the whole sequence. No protein with these characteristics was identified, with the exception of the coagulation factor VIIA. However, this protein forms a complex with the soluble tissue factor (PDB code 1DAN) in the BID-18 dataset, and with the peptide exosite inhibitor E-76 (PDB code 1DVA) in the HB-34 dataset. The two protein partners are not evolutionarily related (sequence identity lower than 20% and sequence coverage below 20%). Therefore, we do not expect our results to be affected by the presence of the coagulation factor VIIA protein in common between the two datasets.

Next, we constructed the data matrix  $D$  for the HB-34 and BID-18 datasets by computing 58 features for all the residues in HB-34 and BID-18. These include several structural features and sequence-based features, as it is common in most machine learning-based methods for hot spot prediction. The structural features are: (i) six physicochemical features obtained from the AAindex database<sup>48</sup>; (ii) five solvent accessible area features,<sup>49</sup> computed using Dictionary of Protein Secondary Structure<sup>50</sup>; and (iii) seven solvent exposure features, computed using hsexpo.<sup>51</sup> The sequence features include: (a) twenty position-specific score matrix (PSSM) profiles, calculated using PSI-BLAST<sup>52</sup>; and (b) twenty block substitution matrix based features, computed using Blosum62.<sup>53</sup> Consequently, there is a total of 58 features. Hence, the training data matrix obtained is of size 58 features  $\times$  313 interface residues and the testing data matrix is of size 58  $\times$  126.

### 2.2 | Workflow

After calculating the features and obtaining the training and test data matrices, the workflow (Figure 1) involves data pre-processing (Step 1 of the workflow), training and validating suitable machine learning models (Step 2 of the workflow), and applying the models on test data and predicting the output labels (Step 3 of the workflow). The steps of the workflow are described below:

1. RBHS: This is our novel pre-processing pipeline for recovering a data matrix with reduced noise from a noisy data matrix and then



**FIGURE 1** Workflow illustrating the steps of the RBHS approach

performing feature selection on the reduced noise matrix. The RBHS pipeline for data pre-processing consists of the following steps:

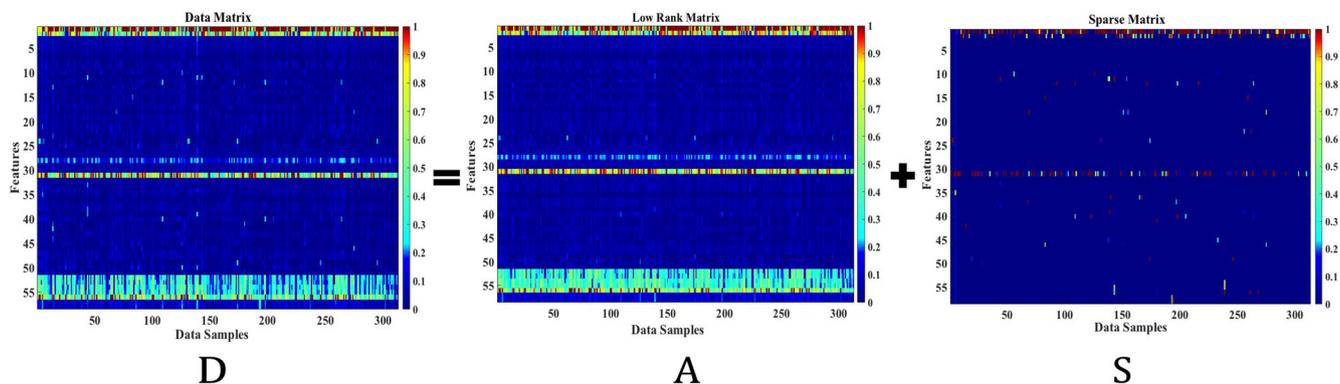
- We normalize the data matrices  $D$  built from the training dataset (HB-34) and from the testing dataset (BID-18), respectively. Normalization is commonly done when the values of different features in the data matrix have different scales; this is the case here. We tested various normalization techniques, as in reference [54]. We used mean variance normalization on our data matrices because it gave the best results among the different techniques tested. The final results using this approach are reported in Table 2.
- We apply RPCA to both the normalized data matrices  $D$  and obtain the corresponding matrices  $A$  that contain reduced noise. In particular, RPCA splits  $D$ , which may have corrupted entries, into a low rank matrix  $A$  and a sparse matrix  $S$  ( $D = A + S$ ).<sup>41</sup> Considering a matrix of size  $m \times n$  (where  $m$  is the number of rows and  $n$  is the number of columns) with  $m < n$ , then the matrix is full rank when all  $m$  rows are linearly independent<sup>†</sup> and its rank is  $m$ . Instead, if  $m > n$ , the matrix is full rank when all  $n$  columns are linearly independent and its rank is  $n$ . The matrix that does not have full rank is a low rank matrix,<sup>55</sup> whereas the sparse matrix is a matrix in which most elements are zero,<sup>56</sup> except for those that correspond to the putative corrupted entries. Among several approaches available to solve RPCA,<sup>57</sup> we use the Principal Component Pursuit method described in.<sup>41</sup> We calculate  $A$  and  $S$  by solving the following optimization problem:(1)

$$\begin{aligned} \min_{A,S} & \|A\|_* + \lambda \|S\|_1, \\ \text{subject to} & D = A + S \end{aligned} \quad (1)$$

Here,  $\|\cdot\|_*$  is the nuclear norm (i.e., the sum of the singular values of a matrix) of matrix  $A$ ,  $\|\cdot\|_1$  is the  $L_1$ -norm (i.e., the sum of the

absolute values of the entries) of matrix  $S$  and  $\lambda$  is a regularization parameter. The value of  $\lambda$  was experimentally determined to obtain the best performance values (metrics are specified in Step 3 of the workflow). The details to solve Equation 1 can be found in.<sup>58</sup> As can be seen in Figure 2, the original matrix  $D$  is corrupted by noise (shown as random, spike-like elements in  $D$ ). This noise is inherent to the sequence- and structure-based data of the complexes included in the matrix. The matrix  $A$  recovered from  $D$  using RPCA exhibits reduced noise. Moreover, the matrix  $S$  is sparse. Therefore,  $S$  can be discarded and  $A$  can be used as the new data matrix for both training and test sets.

- Perform feature selection on matrices  $A$ , to obtain reduced matrices  $A'$ . The importance of each of the features in the training and test matrices  $A$  was calculated using the SciKit module in reference [59] and the Extreme Gradient Boosting (XGB) algorithm.<sup>60</sup> Then, we select those features in both  $A$  matrices, whose feature importance is above an empirically determined threshold. To determine this threshold, we observe whether the performance of the XGB classifier (in terms of accuracy) increases or decreases with the number of selected features. Then, the value of feature importance at which we observe maximum value of accuracy is set as the threshold for feature selection. The plot of accuracy values versus the threshold values can be seen in Figure 3. Based on the plot, we conclude that the optimal value of the threshold is 0.008. Nonetheless, the accuracy-threshold curve is rather rough and thus other threshold values may also lead to accurate results. As shown in the Supplementary Information (Table S1), modifying the threshold by a small extent does not impact dramatically the results. Indeed, when using as threshold 0.009 instead of 0.008, the F1-score decreases by only 0.02 with RBHS +XGB and by 0.05 or less with RBHS combined with other classifiers (see Table S1). This is also the case when using cutoffs

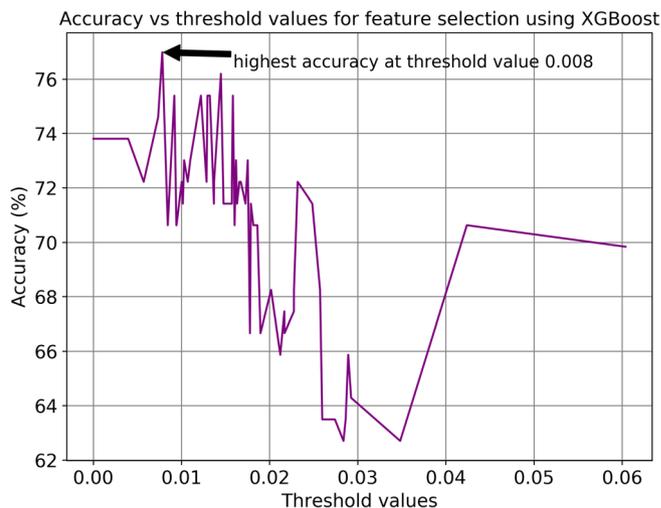


**FIGURE 2** Robust principal component analysis (RPCA) applied to the data matrix  $D$  of the training set HB-34.  $D$  contains entries corrupted by noise that appear as random, spike-like elements in the matrix,  $A$  is the matrix with reduced noise obtained from  $D$  and  $S$  is the sparse matrix

below 0.016, which have the next best-ranked accuracies (see Fig. 3). The F1-score of RBHS+XGB turns out to decrease by 0.08 or less (Table S2). Thus, choosing thresholds with similar accuracies does not affect largely the results. Instead, choosing cutoff values larger than 0.016, which have lower accuracies (see Fig. 3), decreases significantly the performance of the RBHS+XGB method (Table S2). Altogether, our results suggest that, although several thresholds might be chosen, feature selection is an important step in our workflow to improve hot spot prediction. Feature selection identifies the effective feature subspace for building our prediction models, obtaining two new data matrices  $A'$ . The reduced matrix  $A'$  consists of 51 features, instead of the original 58.

2. Training and validation of classification algorithms on the training data set matrix  $A'$ . We train popular classifiers like Support Vector Machines (SVM),<sup>27</sup> Gradient Boosting Machines (GBM),<sup>61</sup> Extreme Gradient Boosting (XGB)<sup>60</sup> and Random Forests<sup>62</sup> on the training data matrix  $A'$ . Next, we use 5-fold cross validation for hyper-parameter tuning of the trained classifiers, using SciKitlearn.<sup>63,64</sup> In such validation, the training data is divided into five subsets. One of the five subsets is used as the validation set and the other four are put together to form a training set; this method is repeated five times. Hence, every data point is in the validation set exactly once, and in the training set four times. During the 5-fold cross validation the F1-score is used as the scoring parameter to assess the total effectiveness of our model. The F1-score calculated each time is then averaged over all five iterations. Besides hyper-parameter tuning, cross validation helps to reduce overfitting on the training set, because the dataset is split into multiple folds/subsets and the algorithm is trained on different folds each time. In this way, the model becomes more generalizable.

3. Applying the validated classifiers to the test dataset matrix  $A'$ . We apply the validated models on  $A'$  (Figure 1) to calculate labels for the residues in the test set. If the label = 1, the residue is classified as hot spot, if label = 0, it is a null spot. The computationally predicted hot and null spot labels are compared with the experimentally known labels and the following performance metrics are calculated:



**FIGURE 3** Accuracy vs threshold values plot for feature selection using Extreme Gradient Boosting (XGB). The highest value of accuracy of the XGB classifier is at the threshold value 0.008. All features with feature importance less than 0.008 are thus discarded from the data matrix [Color figure can be viewed at wileyonlinelibrary.com]

$$\text{Recall} = \frac{TP}{TP + FN}, \quad (2)$$

$$\text{Specificity} = \frac{TN}{TN + FP}, \quad (3)$$

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}, \quad (4)$$

$$\text{Precision} = \frac{TP}{TP + FP}, \quad (5)$$

$$\text{F1-Score} = \frac{2TP}{2TP + FP + FN}, \quad (6)$$

$$\begin{aligned} &\text{Matthew's Correlation Coefficient (MCC)} \\ &= \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FN)}}, \end{aligned} \quad (7)$$

Here, TP, FP, TN, and FN represent the number of true positives (predicted hot spot residues are indeed known experimentally to be

**TABLE 1** Performance comparison of various methods on the training dataset HB-34. These values are computed in Step 2 of our workflow in Figure 1

Method	Recall	Specificity	Accuracy	Precision	F1-Score	MCC
Original Data+SVM	0.59	0.80	0.71	0.69	0.63	0.40
PCA + SVM	0.36	0.89	0.67	0.72	0.48	0.30
RBHS+SVM	0.67	0.74	0.71	0.66	0.66	0.41
Original Data+GBM	0.65	0.78	0.72	0.69	0.66	0.43
PCA + GBM	0.58	0.71	0.65	0.60	0.59	0.29
RBHS+GBM	0.67	0.73	0.71	0.65	0.66	0.40
Original Data+XGB	0.57	0.74	0.67	0.62	0.59	0.32
PCA + XGB	0.61	0.71	0.67	0.61	0.60	0.31
RBHS+XGB	0.56	0.76	0.68	0.64	0.59	0.33
Original Data+RF	0.51	0.79	0.67	0.65	0.57	0.31
PCA + RF	0.53	0.74	0.67	0.57	0.52	0.24
RBHS+RF	0.56	0.78	0.68	0.65	0.60	0.34

**TABLE 2** Performance of different methods on the testing dataset BID-18. These values are computed in Step 3 of our workflow in Figure 1

Method	Recall	Specificity	Accuracy	Precision	F1-Score	MCC
Original Data+SVM	0.79	0.66	0.70	0.51	0.62	0.42
PCA + SVM	0.59	0.67	0.71	0.44	0.51	0.24
RBHS+SVM	0.80	0.69	0.72	0.53	0.64	0.45
Original Data+GBM	0.54	0.66	0.62	0.41	0.47	0.18
PCA + GBM	0.54	0.64	0.61	0.40	0.46	0.17
RBHS+GBM	0.69	0.76	0.74	0.56	0.62	0.43
Original Data+XGB	0.54	0.78	0.71	0.53	0.53	0.32
PCA + XGB	0.59	0.67	0.64	0.44	0.51	0.24
RBHS+XGB	0.72	0.79	0.77	0.61	0.66	0.49
Original Data+RF	0.54	0.80	0.72	0.55	0.54	0.34
PCA + RF	0.56	0.76	0.70	0.51	0.54	0.31
RBHS+RF	0.67	0.78	0.75	0.58	0.62	0.43

so), false positives (predicted hot spot residues are experimentally null spots), true negatives (predicted null spots are indeed so) and false negatives (predicted null spots are actual hot spots).

### 3 | RESULTS

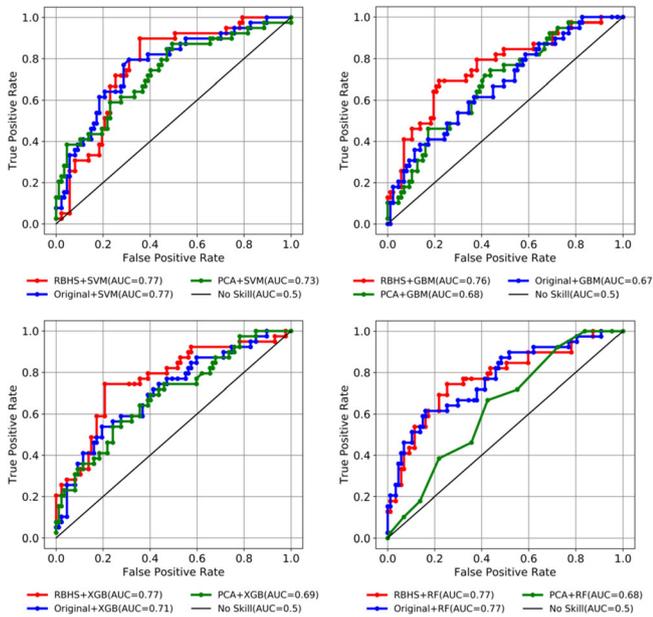
We compare the performance of the classification algorithms specified in the workflow<sup>†</sup> to all the three training data matrices: the original data matrix  $D$ , the matrix obtained after applying Principal Component Analysis (PCA)<sup>‡65</sup> on  $D$  and the reduced matrix  $A'$  (calculated in Section 2.2) obtained after performing RBHS on  $D$ .

Next, we analyze the performance of the classifiers in terms of the metrics<sup>§</sup> described in step 3 of the workflow in section 2.2. To identify the best metric for this analysis, we notice that the two datasets have imbalanced classes (in that they are biased towards null spots): indeed, as mentioned in the Methods Section, HB-34 has 133 hotspots and 180 null spots and BID-18 has 39 hot spots and 87 null spots. For learning from imbalanced datasets, one needs to

improve recall (Equation (2)), which provides information about a classifier's performance with respect to false negatives, without hurting the precision metric (Equation (5)), which deals with false positives. Unfortunately, trying to reach this goal can often be challenging, since when increasing the true positives for the minority class (in our case the class of hot spots), the number of false positives can also increase, reducing precision.<sup>66</sup> The problem is greatly alleviated by using a metric that combines the trade-offs of both precision and recall.<sup>66</sup> This is their harmonic mean, i.e., the F1-score. It reaches its highest value at 1 (perfect precision and recall) and the lowest at 0. Thus, the F1-score was used as the scoring parameter for tuning the hyperparameters and also to assess the performance of our approach.

Table 1 shows the performance of the classifiers (specified in section 2.2) upon implementing 5-fold cross validation on the training dataset HB-34. Using RBHS with SVM and GBM classifiers and Original Data with GBM, yields the highest F1-score (0.66) while PCA+SVM and PCA+RF exhibit the lowest value (0.48 and 0.52, respectively).

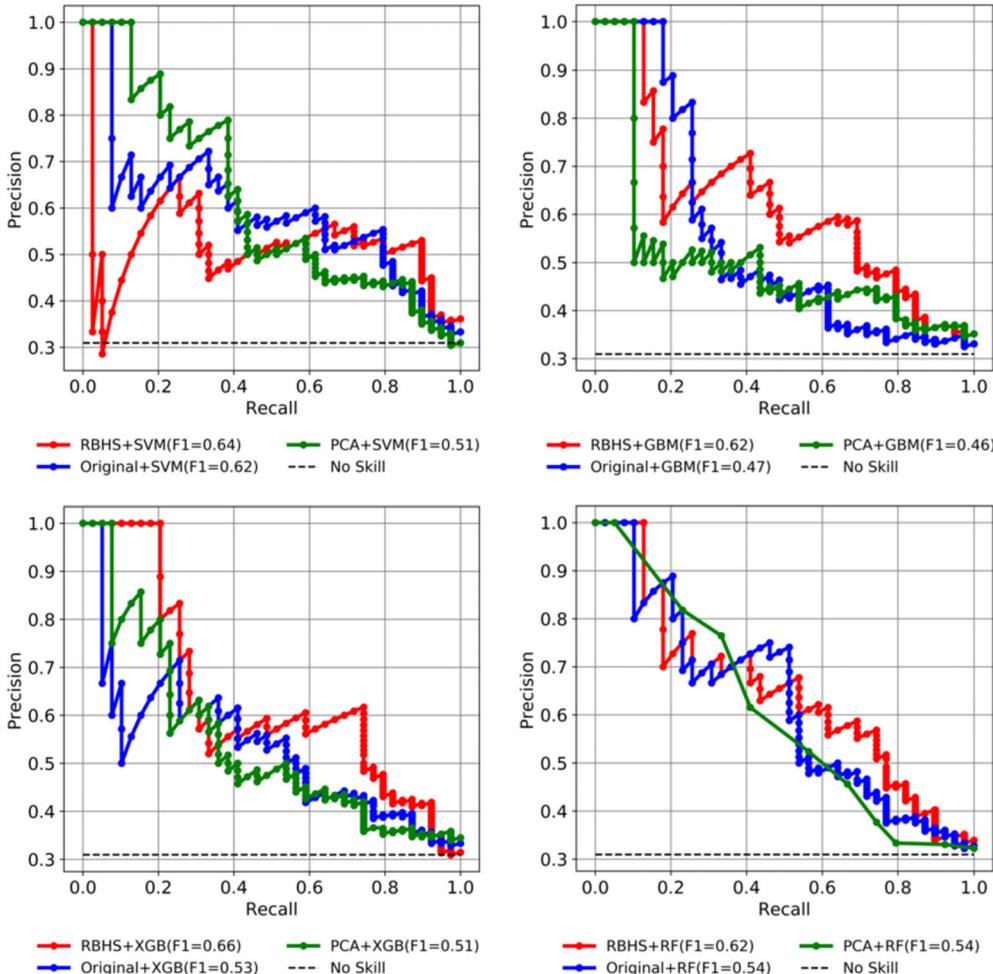
Next, the trained models were applied on the testing data BID-18 to predict class labels for test data (Table 2). The best F1-score value on the test set is obtained by RBHS+XGB (0.66) and RBHS+SVM



**FIGURE 4** ROC (Receiver Operating Characteristic) Curves to compare the performance of all the methods on the independent test set along with the AUC (Area under the curve) values for each method [Color figure can be viewed at wileyonlinelibrary.com]

(0.64), while the lowest value by Original Data+GBM (0.47) and PCA+GBM (0.46). Thus, RBHS performs better than PCA and Original data, regardless of the classifier being used. We now compare the results of training with those of testing. There is a significant increase in F1 scores for RBHS+XGB and RBHS+RF during testing. Instead, for RBHS+SVM, the F1 score does not change and it slightly decreases for RBHS+GBM, as shown in Table 2. Based on these observations, we conclude that there is no overfitting on the training data when using RBHS. In contrast, the F1-scores of classifiers applied to the original data during testing (Table 2) are overall lower than those of the training set (Table 1). Most likely, this is caused by the overfitting of the classifiers on the original training data. The same is observed in case of PCA, except in case of PCA+SVM, where there is a slight increase of 0.03 during testing.

As shown in Figure 2, matrix  $D$  contains a significant amount of noise, and thus the PCA algorithm generates a noisy representation of  $D$ . Classifiers utilizing representations based on PCA and on the original matrix  $D$  tend to overfit on the noisy training data matrix and perform poorly on the test data, as also observed in reference [41]. In contrast, our approach uses the reduced noise matrix  $A'$ , that has been obtained from the noisy matrix  $D$  using RBHS (Section 2.2). Hence, the model does not overfit on the training data and works well during testing.



**FIGURE 5** Precision-Recall Curves of different methods applied on the independent test set. The F1-Score values for each method are also reported [Color figure can be viewed at wileyonlinelibrary.com]

**TABLE 3** Comparison of our approach (RBHS) when used with XGB classifier, with other state of the art methods for hot spot prediction. For each metric, the top scoring method is highlighted in blue, the second one in green and the third one in yellow

Method	Recall	Specificity	Accuracy	Precision	F1-Score	MCC
PredHS-SVM	0.79	0.93	0.83	0.59	0.68	0.57
HEP	0.60	0.76	0.79	0.84	0.70	0.56
RBHS+XGB	0.72	0.79	0.77	0.61	0.66	0.49
KFC2a	0.55	0.73	0.73	0.74	0.63	0.44
KFC2b	0.64	0.87	0.77	0.55	0.60	0.44
MINERVA	0.65	0.90	0.76	0.44	0.52	0.38
APIS	0.57	0.76	0.75	0.72	0.64	0.45
Robetta	0.52	0.88	0.72	0.33	0.41	0.25
FOLDEF	0.48	0.88	0.69	0.26	0.34	0.17
PCRPI	0.51	0.75	0.69	0.39	0.44	0.25
KFC	0.48	0.85	0.69	0.31	0.38	0.19

Next, we calculate the receiver operating characteristic (ROC) curve along with the Area Under the ROC curve (AUROC) (Figure 4) in order to measure discrimination (i.e., the ability of the algorithm to correctly identify those residues which are hot spots from those which are not). We use also the Precision-Recall curves, as they are more informative in case of imbalanced datasets<sup>67</sup> (Figure 5 and Table 2).

Table 2 along with the ROC curves (Figure 4) and the Precision-Recall curves (Figure 5) allow us to identify the best classifier to be used with RBHS. We can see from Table 2 and the precision-recall curves that RBHS+SVM (0.8) performs best for recall and RBHS+XGB (0.61) performs best for precision. The latter method, when applied to the test set BID-18, is also the best for F1-score (0.66), prediction accuracy (0.77) and MCC (0.49). It performs second best for specificity (0.79), after Original Data+RF (0.80). Hence, RBHS, along with XGB classifier, shows a reliable performance in identifying hot spots.

Although the results presented here are based on the training matrix constructed by computing 58 features from the state-of-the-art HB-34 dataset, we performed additional tests using other training matrices, generated by artificially introducing Gaussian noise (with zero mean and either 1 or 10 standard deviation). This noise is randomly added to an increasing number of locations in the data matrix, resulting in a percentage of corrupted entries ranging between 1 and 50% (see Supplementary Information, Table S3). These tests showed that RBHS+XGB is still capable to predict hotspots for up to 20% corrupted entries when using a standard deviation of the added Gaussian noise of 10 (see Supplementary Information, Table S3).

Finally, we compare the performance of RBHS+XGB on the independent test set BID-18 with other popular hot spot prediction algorithms, including HEP,<sup>68</sup> PredHS-SVM,<sup>36</sup> KFC2a and KFC2b,<sup>35</sup> PCRPI,<sup>44</sup> MINERVA,<sup>69</sup> APIS,<sup>34</sup> KFC,<sup>70</sup> Robetta,<sup>19</sup> and FOLDEF<sup>18</sup> (Table 3). The performance metric values on the BID-18 dataset for the aforementioned methods is taken from.<sup>24</sup> Our methodology turns out to be in the top three after HEP<sup>68</sup> and PredHS-SVM,<sup>36</sup> when considering the F1-score. Moreover, it also performs second best in case of Recall values and third best in terms of Accuracy and MCC values.

As mentioned before, F1-Score is the best indicator for the predictive power of the methods with imbalance datasets,<sup>66</sup> like in our case. Hence, the results in Table 3 establish unambiguously the predictive power of the RBHS method presented here.

## 4 | DISCUSSION

Protein-protein interactions (PPI) databases, as other biological data repositories, often contain noise caused by gross errors associated with computational issues and/or human oversight. This affects significantly the quality of predictions made based on these data. Therefore, it is very important to identify methodologies which drastically reduce the level of noise. So far, the existing machine learning-based hot spot predictors have not addressed this important issue: their predictions are based on data matrices that are inherently noisy. Here, we present a novel, machine learning-based pre-processing technique for hot spot prediction, called Robust Principal Component Analysis-based Prediction of PPI Hot spots (RBHS). This recovers a matrix with reduced noise from a corrupted data matrix by using RPCA.<sup>41</sup> RBHS is then used with different classifiers, including Support Vector Machines (SVM), Gradient Boosting Machines (GBM), Extreme Gradient Boosting (XGB) and Random Forests (RF) (Table 1 and Table 2). We were able to show that RBHS, when combined with the Extreme Gradient Boosting (XGB) classifier, is a rather reliable approach: indeed, the data matrices obtained from the HB-34<sup>24</sup> and BID-18<sup>24</sup> PPI databases contained reduced noise. The approach applied on the independent test set BID-18 identified as many as 77% of the known hot spots of the complexes investigated,<sup>46</sup> as shown in Table 2. Arguably, our approach emerges as the method of choice for hot spot prediction in the frequent cases where there is a highly noisy data matrix, because our preprocessing pipeline includes RPCA. The authors of reference [41] have demonstrated mathematically that RPCA is a method of choice for recovering a low rank matrix from a highly noisy data matrix, regardless of the number of corrupt or missing entries.

## 5 | CONCLUSION

Machine learning is a powerful tool for the analysis of hot-spots in protein-protein complexes.<sup>32-39</sup> This analysis is very important from a medical and pharmacological perspective, because hot spot residues may undergo mutations in a variety of diseases and can be used to design PPI inhibitors. Here, we have presented a machine learning method that increases the reliability of the predictions, by reducing

the noise that is frequently present in the data matrices. Further improvements of the predictive power of the approach may include the application of semi-supervised learning based approaches to leverage the large amount of unlabeled data available.

## ACKNOWLEDGEMENTS

We would like to thank Dr. Florian Lemmerich and Prof. Dr. Markus Strohmaier from RWTH Aachen University (Germany) and Prof. Ira Assent from Aarhus University (Denmark) for very useful suggestions. Open access funding enabled and organized by Projekt DEAL.

## CONFLICT OF INTERESTS

The authors declare that they have no conflict of interest.

## PEER REVIEW

The peer review history for this article is available at <https://publons.com/publon/10.1002/prot.26047>.

## DATA AVAILABILITY STATEMENT

Data sharing not applicable to this article as no new datasets were generated during the current study.

## ORCID

Divya Sitani  <https://orcid.org/0000-0001-5138-6108>

Alejandro Giorgetti  <https://orcid.org/0000-0001-8738-6150>

Mercedes Alfonso-Prieto  <https://orcid.org/0000-0003-4509-4517>

Paolo Carloni  <https://orcid.org/0000-0002-9010-0149>

## ENDNOTES

\* A set of vectors is called linearly independent if no vector in the set can be expressed as a linear combination of the other vectors in the set.

† These include Support Vector Machines (SVM), Gradient Boosting Machines (GBM), Extreme Gradient Boosting (XGB) and Random Forests (RF).

‡ For PCA representation, the principal components explaining 95% variance were chosen, after analyzing the results.

§ These include: Recall, Specificity, accuracy, precision, Mathew's Correlation Constant (MCC) and the F1-score.

## REFERENCES

- Lesk A. *Introduction to Protein Science: Architecture, Function, and Genomics*. Oxford, England: Oxford university press; 2010.
- Alberts B, Miake-Lye R. Unscrambling the puzzle of biological machines: the importance of the details. *Cell*. 1992;68(3):415-420.
- Stites WE. Protein-protein interactions: interface structure, binding thermodynamics, and mutational analysis. *Chem Rev*. 1997;97(5):1233-1250.
- Janin J. Elusive affinities. *Proteins: structure, function. Bioinformatics*. 1995;21(1):30-39.
- Jones S, Thornton JM. Principles of protein-protein interactions. *Proc Natl Acad Sci*. 1996;93(1):13-20.
- Janin J, Chothia C. The structure of protein-protein recognition sites. *J Biol Chem*. 1990;265(27):16027-16030.
- Bogan AA, Thorn KS. Anatomy of hot spots in protein interfaces. *J Mol Biol*. 1998;280(1):1-9.
- Kann MG. Protein interactions and disease: computational approaches to uncover the etiology of diseases. *Brief Bioinform*. 2007;8(5):333-346.
- Petta I, Lievens S, Libert C, Tavernier J, Bosscher KD. Modulation of protein-protein interactions for the development of novel therapeutics. *Mol Ther*. 2016;24(4):707-718.
- Scott DE, Bayly AR, Abell C, Skidmore J. Small molecules, big targets: drug discovery faces the protein-protein interaction challenge. *Nat Rev Drug Discov*. 2016;15(8):533-550.
- Murakami Y, Tripathi LP, Prathipati P, Mizuguchi K. Network analysis and in silico prediction of protein-protein interactions with applications in drug discovery. *Curr Opin Struct Biol*. 2017;44:134-142.
- Cunningham BC, Wells JA. High-resolution epitope mapping of hGH-receptor interactions by alanine-scanning mutagenesis. *Science*. 1989;244(4908):1081-1085.
- Moreira IS, Fernandes PA, Ramos MJ. Hot spots—a review of the protein-protein interface determinant amino acid residues. *Proteins: Structure, Function Bioinformatics*. 2007;68(4):803-812.
- David A, Sternberg MJ. The contribution of missense mutations in Core and rim residues of protein-protein interfaces to human disease. *J Mol Biol*. 2015;427(17):2886-2898.
- David A, Razali R, Wass MN, Sternberg MJ. Protein-protein interaction sites are hot spots for disease-associated nonsynonymous SNPs. *Hum Mutat*. 2012;33(2):359-363.
- Ibarra AA, Bartlett GJ, Hegedüs Z, et al. Predicting and experimentally validating hot-spot residues at protein-protein interfaces. *ACS Chem Biol*. 2019;14(10):2252-2263.
- Keskin O, Tuncbag N, Gursoy A. Predicting protein-protein interactions from the molecular to the proteome level. *Chem Rev*. 2016;116(8):4884-4909.
- Guerois R, Nielsen JE, Serrano L. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J Mol Biol*. 2002;320(2):369-387.
- Kortemme T, Baker D. A simple physical model for binding energy hot spots in protein-protein complexes. *Proc Natl Acad Sci*. 2002;99(22):14116-14121.
- Massova I, Kollman PA. Computational alanine scanning to probe protein-protein interactions: a novel approach to evaluate binding free energies. *J Am Chem Soc*. 1999;121(36):8133-8143.
- Huo S, Massova I, Kollman PA. Computational alanine scanning of the 1:1 human growth hormone-receptor complex. *J Comput Chem*. 2002;23(1):15-27.
- Grosdidier S, Fernández-Recio J. Identification of hot-spot residues in protein-protein interactions by computational docking. *BMC Bioinformatics*. 2008;9(1):447.
- Brenke R, Kozakov D, Chuang GY, et al. Fragment-based identification of druggable 'hot spots' of proteins using Fourier domain correlation techniques. *Bioinformatics*. 2009;25(5):621-627.
- Liu S, Liu C, Deng L. Machine learning approaches for protein-protein interaction hot spot prediction: Progress and comparative assessment. *Molecules*. 2018;23(10):2535.
- Melo R, Fieldhouse R, Melo A, et al. A machine learning approach for hot-spot detection at protein-protein interfaces. *Int J Mol Sci*. 2016;17(8):1215.
- Quinlan JR. Induction of decision trees. *Machine Learning*. 1986;1(1):81-106.
- Cortes C, Vapnik V. *Support-Vector Networks Machine Learning (Pp. 237-297), Vol. 20*. Boston, MA: Kluwer Academic Publisher; 1995.
- Yao X. Evolving artificial neural networks. *Proc IEEE*. 1999;87(9):1423-1447.
- Dietterich TG. Ensemble methods in machine learning. *Multiple Classifier Systems*. Berlin, Heidelberg: Springer Berlin Heidelberg; 2000:1-15.
- Andersen SK. Book review: Judea pearl, probabilistic reasoning in intelligent systems: networks of plausible inference. *Artificial Intelligence*. 1991;48(1):117-124.
- Jordan MI. *Learning in Graphical Models*. Vol 89. New York, NY: Springer Science & Business Media; 1998.

32. Wang L, Liu ZP, Zhang XS, Chen L. Prediction of hot spots in protein interfaces using a random forest model with hybrid features. *Protein Eng des Sel*. 2012;25(3):119-126.
33. Lise S, Buchan D, Pontil M, Jones DT. Predictions of hot spot residues at protein-protein interfaces using support vector machines. *PLoS One*. 2011;6(2):e16774.
34. Xia JF, Zhao XM, Song J, Huang DS. APIS: accurate prediction of hot spots in protein interfaces by combining protrusion index with solvent accessibility. *BMC Bioinformatics*. 2010;11(1):174.
35. Zhu X, Mitchell JC. KFC2: a knowledge-based hot spot prediction method based on interface solvation, atomic density, and plasticity features. *Proteins: Structure, Function, and Bioinformatics*. 2011;79(9):2671-2683.
36. Deng L, Guan J, Wei X, Yi Y, Zhang QC, Zhou S. Boosting prediction performance of protein-protein interaction hot spots by using structural neighborhood properties. *J Comput Biol*. 2013;20(11):878-891.
37. Ofra Y, Rost B. Protein-protein interaction hotspots carved into sequences. *PLoS Comput Biol*. 2007;3(7):e119.
38. Moreira IS, Koukos PI, Melo R, et al. SpotOn: high accuracy identification of protein-protein interface hot-spots. *Sci Rep*. 2017;7(1):1-11.
39. Assi SA, Tanaka T, Rabbitts TH, Fernandez-Fuentes N. PCRPI: presaging critical residues in protein interfaces, a new computational tool to chart hot spots in protein interfaces. *Nucleic Acids Res*. 2010;38(6):e86-e86.
40. Analysis of Errors <http://faculty.sites.uci.edu/chem1/files/2013/11/RDGerranal.pdf>.
41. Candès EJ, Li X, Ma Y, Wright J. Robust principal component analysis? *Journal of the ACM (JACM)*. 2011;58(3):1-37.
42. Thorn KS, Bogan AA. ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. *Bioinformatics*. 2001;17(3):284-285.
43. Moal IH, Fernández-Recio J. SKEMPI: a structural kinetic and energetic database of mutant protein interactions and its use in empirical models. *Bioinformatics*. 2012;28(20):2600-2607.
44. Assi SA, Tanaka T, Rabbitts TH, PCRPI F-FN. Presaging critical residues in protein interfaces, a new computational tool to chart hot spots in protein interfaces. *Nucleic Acids Res*. 2009;38(6):e86-e86.
45. Petukh M, Li M, Alexov E. Predicting binding free energy change caused by point mutations with knowledge-modified MM/PBSA method. *PLoS Comput Biol*. 2015;11(7):e1004276.
46. Fischer T, Arunachalam K, Bailey D, et al. The binding interface database (BID): a compilation of amino acid hot spots in protein interfaces. *Bioinformatics*. 2003;19(11):1453-1454.
47. Huang Y, Niu B, Gao Y, Fu L, Li W. CD-HIT suite: a web server for clustering and comparing biological sequences. *Bioinformatics*. 2010;26(5):680-682.
48. Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, Kanehisa M. AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res*. 2007;36(suppl\_1):D202-D205.
49. Rost B, Sander C. Conservation and prediction of solvent accessibility in protein families. *Proteins: structure, function*. *Bioinformatics*. 1994;20(3):216-226.
50. Joosten RP, Te Beek TA, Krieger E, et al. A series of PDB related databases for everyday needs. *Nucleic Acids Res*. 2010;39(suppl\_1):D411-D419.
51. Hamelryck T. An amino acid has two sides: a new 2D measure provides a different view of solvent exposure. *Proteins: Structure, Function, and Bioinformatics*. 2005;59(1):38-48.
52. Altschul SF, Madden TL, Schäffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 1997;25(17):3389-3402.
53. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci*. 1992;89(22):10915-10919.
54. Preprocessing data <https://scikit-learn.org/stable/modules/preprocessing.html>.
55. Strang G. *Linear Algebra and its Applications*. Thomson, Brooks/Cole: Belmont, CA; 2006 <http://www.amazon.com/Linear-Algebra-Its-Applications-Edition/dp/0030105676>.
56. Duff IS, Erisman AM, Reid JK. *Direct Methods for Sparse Matrices*. USA: Oxford University Press, Inc.; 1986.
57. Bouwmans T, Zahzah EH. Robust PCA via principal component pursuit: A review for a comparative evaluation in video surveillance. *Computer Vision and Image Understanding*. 2014;122:22-34.
58. Aravkin A, Becker S, Cevher V, Olsen P. A variational approach to stable principal component pursuit. *Conference on Uncertainty in Artificial Intelligence (UAI)*. Arlington, VA: AUAI Press; 2014.
59. Feature selection [https://scikit-learn.org/stable/modules/feature\\_selection.html](https://scikit-learn.org/stable/modules/feature_selection.html).
60. Chen T, Guestrin C. Xgboost: a scalable tree boosting system. *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*. 2016;785-794. <https://doi.org/abs/10.1145/2939672.2939785>.
61. Friedman JH. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*. 2001;29:1189-1232.
62. Breiman L. Random forests. *Machine Learning*. 2001;45(1):5-32.
63. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. *J Machine Learning Res*. 2011;12:2825-2830.
64. Cross-validation: evaluating estimator performance: [https://scikit-learn.org/stable/modules/cross\\_validation.html](https://scikit-learn.org/stable/modules/cross_validation.html).
65. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY: Springer Science & Business Media; 2009.
66. Chawla NV. Data mining for imbalanced datasets: an overview. *Data Mining and Knowledge Discovery Handbook*. New York, NY: Springer; 2009:875-886.
67. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One*. 2015;10(3):e0118432.
68. Xia J, Yue Z, Di Y, Zhu X, Zheng CH. Predicting hot spots in protein interfaces based on protrusion index, pseudo hydrophobicity and electron-ion interaction pseudopotential features. *Oncotarget*. 2016;7(14):18065-18075.
69. Ki C, Kim D, Lee D. A feature-based approach to modeling protein-protein interaction hot spots. *Nucleic Acids Res*. 2009;37(8):2672-2687.
70. Darnell SJ, Page D, Mitchell JC. An automated decision-tree approach to predicting protein interaction hot spots. *Proteins: Structure, Function, and Bioinformatics*. 2007;68(4):813-823.

## SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Sitani D, Giorgetti A, Alfonso-Prieto M, Carloni P. Robust principal component analysis-based prediction of protein-protein interaction hot spots. *Proteins*. 2021;1-9. <https://doi.org/10.1002/prot.26047>