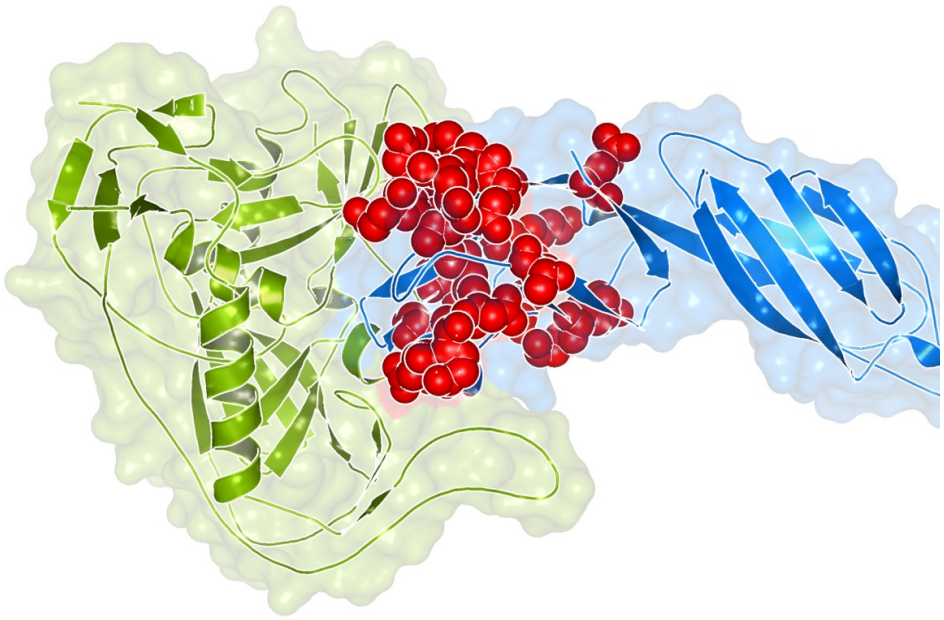


Supervised Classification Applied to Hot-Spot Detection in Protein-Protein Interfaces

Irina de Sousa Moreira
Master's Degree in Mathematical Engineering
Mathematics Department
2016

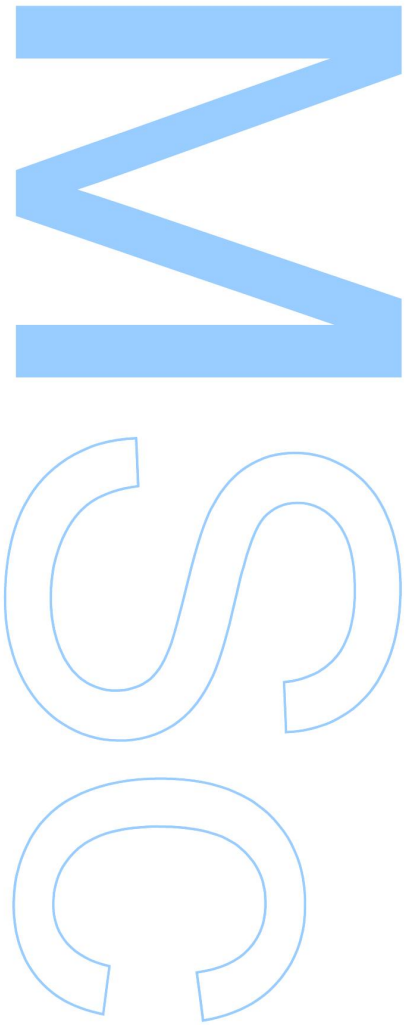
Advisor
Joaquim Pinto Costa, Assistant Professor, Faculty of Science, University of Oporto

Co-Advisor
Alexandre M. J. J. Bonvin, Full Professor, Faculty of Science, Utrecht University

Todas as correções determinadas pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, _____/_____/_____



ACKNOWLEDGEMENTS

I am grateful for all the guidance and support of my advisors, Prof. Joaquim Pinto da Costa (Oporto University) and Prof. Alexandre Bonvin (Utrech University) whose insightful comments and advices allowed me to improve my knowledge on machine-learning application to structural biology. I wish also to acknowledge all the collaboration of students and colleagues in the course of this thesis.

A special thanks to family and close friends that followed closely this adventure and never let me gave up. Their strength was crucial in various moments in the last year. And a very special thanks to my kids, Diogo and Sofia, that endure mommy trying to balance motherhood, managing her scientific group in Coimbra, being a post-doctoral fellow in Utrecht and a master student in Porto. I am sure it was not easy.

RESUMO

A identificação de complexos proteína-proteína e das suas interações é fundamental para a compreensão da organização da maquinaria celular. Devido à elevada dificuldade na obtenção de dados experimentais, novas metodologias e ferramentas computacionais estão a surgir, proporcionando alternativas fiáveis. É especialmente verdade que algoritmos de *Machine Learning* (ML) são extremamente promissores para a pesquisa de interação de proteínas através da identificação de padrões biológicos relevantes, o que levará ao aumento do nosso conhecimento do mecanismo funcional de proteínas dentro das células. Ao longo das últimas décadas, a melhoria de um grande número de técnicas computacionais levou à diminuição de custos e ao aumento de base de dados por ordens de grandeza. No entanto, a precisão ainda se encontra longe do que seria de esperar e existe espaço para melhorias.

Neste trabalho, aplicamos técnicas de ML que vão além do atual estado da arte e que nos levam a previsões precisas de *Hot-Spots* em complexos de proteína-proteína. Exploramos a capacidade de usar ML para o problema biológico em causa e comparamos diferentes classificadores e condições de pré-processamento. Com base nesta avaliação, concluímos que a aplicação do algoritmo C5.0 com super-amostragem da classe menor leva a resultados em concordância com a realidade e que apresenta uma precisão global numa base de dados independente de 0.88. Devido à relevância do tema para a comunidade científica que trabalha em biologia estrutural, criámos um web-server que se encontra disponível de forma grátis: http://milou.science.uu.nl/cgi/servicesdevel/SPOTON/spoton/

PALAVRAS-CHAVE

Bioinformática, Interações proteína-proteína, Hot-spots, Machine-learning, Classificação supervisionada, Avaliação de resultados.

ABSTRACT

The identification of protein complexes and interactions is crucial for the understanding of cellular organization and machinery. Due to the high difficulty in attaining experimental data about such an important subject, computational tools and methodologies are emerging as reliable alternatives. It is especially true that Machine-Learning (ML) algorithms hold an incredible promise for protein interaction research by identifying biological relevant patterns, which accelerates our knowledge of the functional mechanism of proteins within the cells. Over the last decades the improvement of a large number of computational techniques led to significant cost decreases and, also, increases in throughput by orders of magnitude. However, there is still room for improvement as their accuracy is still far from optimal.

In this work, we have developed and applied computer modelling techniques that went beyond the current state-of-the-art, leading to quantitative and reliable molecular-level predictions of Hot-Spots at protein-protein complexes. We explored the feasibility of using ML in the HS detection and compared different classifiers as well as different pre-processing conditions. Based on this evaluation, we concluded that applying the C5.0 algorithm with minor class up-sampling leads to accurate results. The overall accuracy in an independent test set demonstrated to be 0.88. Due to the theme's relevance to the large scientific community working on structural biology, we have assembled a freely available web-server that can be found at: http://milou.science.uu.nl/cgi/servicesdevel/SPOTON/spoton/

KEYWORDS

Bioinformatics, Protein-protein interactions, hot-spots, Machine-learning, Supervised learning, Classification, Performance metrics.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

3D	*Three-Dimensional*
AAC	*Amino Acid Composition*
ADA	*Boosted Classification Trees*
AMDAi	*Adaptive Mixture Discriminant Analysis*
ANN	*Artificial Neural Networks*
ANOSIM	*Analysis of Group Similarities*
APAAC	*Amphiphilic Pseudo Amino Acid Composition*
ASAEdb	*Alanine Scanning Energetics Database*
ASM	*Alanine Scanning Mutagenesis*
AUROC	*Area Under the Receiver Operator Curve*
avNNet	*Model Averaged Neural Network*
bagEarth	*Bagged MARS*
bagEarthGCV	*Bagged MARS using gCV Pruning*
bagFDA	*Bagged Flexible Discriminant Analysis*
bagFDAGCV	*Bagged FDA with gCV Pruning*
Bagging	*Bootstrapped Aggregation*
BBN	*Bayesian Belief Network*
BID	*Binding Interface Database*
Binda	*Binary Discriminant Analysis*
BLOSUM	*BLOcks Substitution Matrix*
BN	*Bayesian Network*
Boruta	*Random Forest with Additional Feature Selection*
C5.0Cost	*Cost-Sensitive C5.0*
CAPRI	*Critical Assessment of Predicted Interactions*
CART	*Classification and Regression Tree*
CHAID	*Chi-squared Automatic Interaction Detection*
CompASM	*Computational Alanine Scanning Mutagenesis*

CPORT	*Consensus Prediction Of interface Residues in Transient complexes*
Ctree	*Conditional Inference Tree*
Ctree2	*Conditional Inference Tree*
dwdPoly	*Distance Weighted Discriminant with Polynomial Kernel*
dwdRadial	*Distance Weighted Discrimination with Radial Basis Function Kernel*
Earth	*Multivariate Adaptive Regression Spline*
EDA	*Exploratory Data Analysis*
extraTrees	*Random Forest by Randomization*
FDA	*Flexible Discriminant Analysis*
FN	*False Negative*
FNR	*False Negative Rate*
FP	*False Positive*
FPR	*False Positive Rate*
gamboost	*Boosted generalized Additive Model*
GBM	*Gradient Boosting Machines*
GBRT	*Gradient Boosted Regression Trees*
gcvEarth	*Multivariate Adaptive regression Splines*
GLM	*Generalized Linear Model*
glmboost	*Boosted generalized Linear Model*
Had	*Heteroscedastic Discriminant Analysis*
HADDOCK	*High Ambiguity Driven protein-protein DOCKing*
Hdda	*High Dimensional Discriminant Analysis*
HS	*Hot-Spots*
ID3	*Iterative Dichotomiser 3*
J48	*C4.5-like Trees*
LARS	*Least-Angle Regression*
LASSO	*Least Absolute Shrinkage and Selection Operator*
LDA	*Linear Discriminant Analysis*
LDA2	*Linear Discriminant Analysis with discriminate functions*
LocLDA	*Localized Linear Discriminant Analysis*

LogitBoot	*Boosted Logistic Regression*
LVQ	*Learning Vector Quantization*
LWL	*Locally Weighted Learning*
MANOVA	*Multivariate Analysis of Variance*
MD	*Molecular Dynamic*
MDA	*Mixture Discriminant Analysis*
MDS	*MultiDimensional Scaling*
ML	*Machine-Learning*
MLP	*MultiLayer Perceptron*
MRPP	*Multi-response Permutation Procedures*
multinom	*Penalized Multinomial Regression*
NB	*Naïve Bayesian*
NMR	*Nuclear Magnetic Resonance*
NPV	*Negative Predictive Value*
NS	*Null-Spots*
ORFlog	*Oblique Random Forest*
ORFpls	*Oblique Random Forest*
ORFridge	*Oblique Random Forest*
ORFsvm	*Oblique Random Forest*
PAAC	*Pseudo Amino Acid Composition*
PCA	*Principal Components Analysis*
pcaNNet	*Neural networks with Feature extraction*
PCM	*ProteoChemometric Modeling*
PDA	*Penalized Discriminant Analysis*
PDB	*Protein Data Bank*
PINT	*Protein-protein Interaction Thermodynamic Database*
PLR	*Penalized Logistic Regression*
PPI	*Protein-Protein Interaction*
PPV	*Positive Predictive Value*
PSSM	*Position Scoring Matrices*

QDA	*Quadratic Discriminant Analysis*
Ranger	*Random Forest*
RBF	*Radial Basis Function*
RDA	*Regularized Discriminant Analysis*
RF	*Random Forest*
RMSD	*Root-Mean-Square-Deviation*
RRF	*Regularized Random Forest*
RRFglobal	*Regularized Random Forest*
SASA	*Solvent Accessible Surface Area*
SKEMPI	*Structural database of kinetic and Energetic of Mutant Protein Interactions*
SOM	*Self-Organizing Map*
StepLDA	*Linear Discriminant Analysis with Stepwise Feature Selection*
StepQDA	*Quadratic Discriminant Analysis with Stepwise Feature Selection*
SVM	*Support Vector Machines*
svmLinear	*Support vector Machine with Linear Kernel*
svmLinear2	*Support vector Machine with Linear Kernel*
svmLinearWeights	*Linear Support Vector Machines with Class Weights*
svmPoly	*Support vector Machine with Polynomial Kernel*
svmRadial	*Support vector Machine with Radial Basis Function Kernel*
svmradialCost	*Support vector Machine with Radial Basis Function Kernel*
svmradialSigma	*Support vector Machine with Radial Basis Function Kernel*
TN	*True Negative*
TNR	*True Negative Rate*
TP	*True Positive*
TPR	*True Positive Rate*
WSRF	*Weighted subspace Random Forest*
WT	*Wild-Type*

1 INTRODUCTION

Structural Bioinformatics is a key research area in the field of Computational Biology and it focuses on the analysis and prediction of 3D structures of nucleic acid- and protein-based machineries (1). A high-resolution structural model of such assemblies is crucial for the correct understanding of their function and mechanism (2) as protein structure, dynamics and function are interdependent (3). Various experimental techniques such as X-ray diffraction, electron microscopy and NMR are widely used to gain structural insight into biomolecules but they are simultaneously time consuming, experimentally expensive and often with inherent technical difficulties. That led to a big discrepancy between the number of published sequences and published 3D structures (Figure 1). Therefore, there is an urgent need for complementary computational procedures capable of reliably generating and identifying 3D protein-protein structures and, especially, identifying their interactions. Moreover, it is imperative to obtain accurate predictions that, integrated with experimental data, can potentially enlarge the structural understanding of the most relevant biological targets. The functional characterization of the complex cellular machinery and of the impact of mutations on protein structure is crucial for the development of new drugs (4).

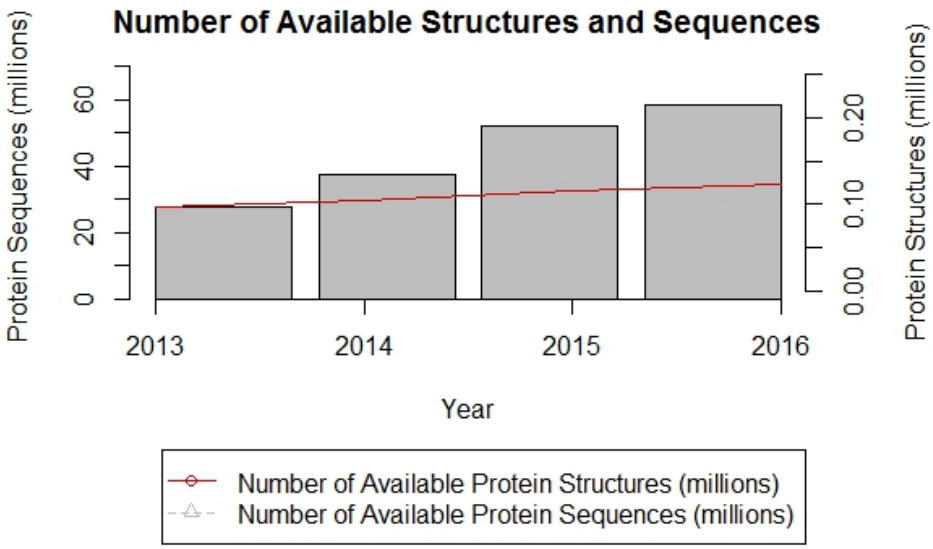


Figure 1. Number of available protein sequences (retrieved from NCBI Reference Sequence (5, 6) Database) and structures (retrieved from PDB (7)) between the years 2013 and 2016.

1.1 PROTEIN-PROTEIN INTERACTIONS

The human interactome consists of more than 400.000 PPIs, which are fundamental for a wide-range of biological pathways (8-10). Interactome-level descriptions of molecular function are becoming crucial for a detailed picture and understanding of the nature of complex traits and diseases (11). Characterizing the critical residues involved in these interactions, which can be performed by experimental or computational methods, is crucial in PPI fine tuning. Furthermore, only through gaining an atomistic-level of detail of PPIs can we develop new methods and drugs that modulate their binding (11, 12). Critical to PPI understanding has been the discovery that the driving forces of protein coupling are not evenly distributed across their surfaces: instead, a usually small set of residues contribute the most to the binding process – the so called Hot-Spots (HS).

1.2 BINDING HOT-SPOTS

Protein-protein interfaces often involve a large number of residues. However, it is generally recognized that small regions of a few residues, termed HS, are essential for maintaining the integrity of the interface. ASM is the method of choice for mapping functional epitopes and can be used to infer energy contributions of individual side-chains to protein binding. The contribution of a residue to the binding energy is measured by the binding free energy difference ($\Delta\Delta G_{binding}$) between WT and mutant complex upon mutation of a specific residue to an alanine (13). Bogan and Thorn (14) defined the residues with $\Delta\Delta G_{binding} \geq 2.0$ kcal mol-1 as HS; and the residues with $\Delta\Delta G_{binding} < 2.0$ kcal mol-1 as NS. HS, apart from providing stability to the complex, also contributes to the specificity at the binding sites. Figure 2 illustrates a protein-protein complex in which the HS and NS are highlighted by a van der Waals representation.

HS are conserved residues tightly packed at the center of the interface (15, 16) with an amino-acid composition similar to the core residues defined by Chakrabarti *et al.* (17). The amino acid composition of HS has shown not to be equally distributed, being enriched with Trp, Tyr and Arg residues (14). The number of HS was shown to increase with the increase of the interface surface area, maybe to overcome a larger configurational entropic cost [29]. At the end, functional and structural epitope were defined as comprising HS or all residues that participate in the interface, respectively [30].

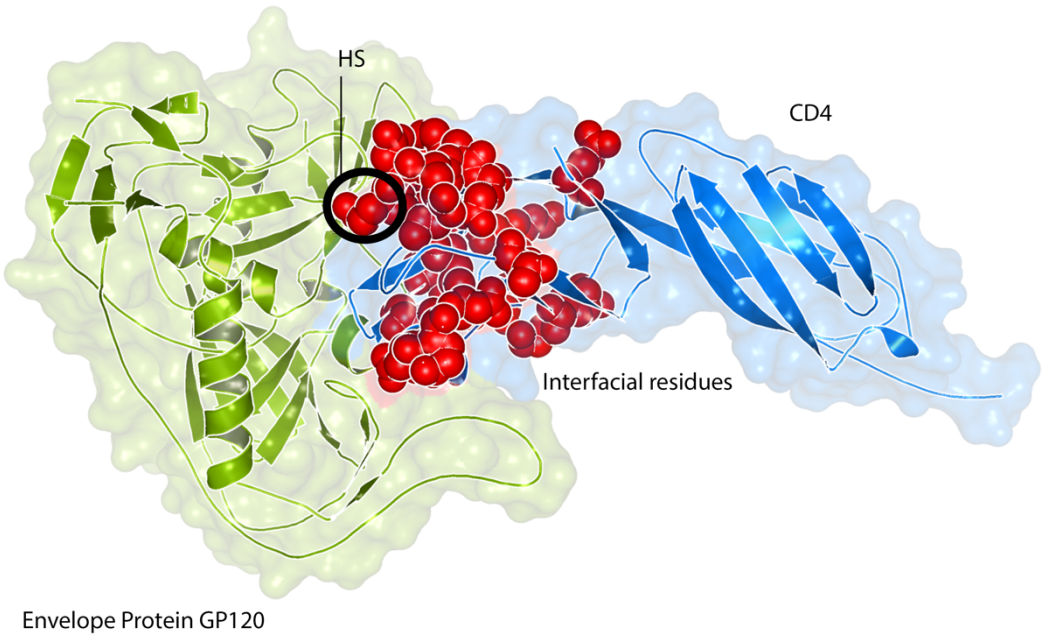


Figure 2. Structural representation of a protein-protein complex (1GC1 (18)). Interfacial residues are highlighted in a van der Walls representation.

HS tend to be found on both monomers and show a high degree of complementarity with buried charged residues forming salt bridges and hydrophobic residues fitting into the nooks on the opposite face (10, 11). Also, PPIs have shown to have a high degree of plasticity, and so, specific proteins may bind to different partners re-utilizing the same HS, although possibly with different combinations (15).

Experimental methods for identifying HS such as the mentioned ASM are based on molecular biology techniques that are accurate but complex, time-consuming and expensive. The inherent low-throughput of these techniques due to the need to express and purify each individual protein before measurement is a major bottleneck (19). Highly efficient computational methods for predicting HS can therefore provide a viable alternative to experimentation. Molecular modelling tools like MD simulations are largely used to construct and analyze protein-protein interactions models and to investigate the dynamic behavior of complex formation or inhibition (12, 20-25). However, due to the complexity and typically large size of protein-protein complexes, these methods are still computationally expensive. Recently, ML approaches trained on various features of experimentally determined HS residues have been developed in order to predict HS in new protein complexes (19, 26-37). ML techniques are especially suitable to deal with HS prediction due to their ability to infer input-output relationships without explicitly assuming a pre-determined model. They tend to work quite well even in non-linear and

noisy environments. Computational ML approaches to PPI prediction tend to fall into two broad categories:

i) sequence-based methods which use an encoding of sequence-derived features of the residues and their neighbours and then explore amino-acid identity, physicochemical properties of amino-acids, predicted solvent accessibility, PSSMs, conservation in evolution and interface propensities;

ii) structure-based methods that use an encoding of structure-based features of the target residues and neighbours such as propensities at interface and surface, interface size, geometry, chemical composition, roughness, SASA, atomic interactions, among others.

A detailed review of current ML algorithms applied to HS detection can be found at Moreira (10).

1.3 THESIS STRUCTURE

This master's dissertation consists of eight sections. In Section 1 we introduce the specific field of HS detection. Section 2 consists of a literature review on the basis of ML methods with special focus on the mathematical foundations of the tested algorithms. Section 3 is a detailed description of the methodology used to attain a reliable predictive model. In Section 4 we present the computational results and in Section 5 we make final considerations and plans for future work. After references (Section 6) you can find the Annexes Section (Section 7) with further tables of results as well as the two publications that resulted from this master's thesis (Section 8); one that will be submitted and another already published in a peer-reviewed journal:

Melo,R., Fieldhouse,R., Melo,A., Correia,J.D.G., Cordeiro,M.N.D.S., Gümüş,Z.H., Costa,J., Bonvin,A.M.J.J. and Moreira,I.S. (2016) A Machine Learning Approach for Hot-Spot Detection at Protein-Protein Interfaces. IJMS, 17, 1215

Moreira,I.S*., Koukos,P*., Melo,R., Almeida,J.G., Gomes, A., Schaarschmidt,J., Trellet,M., Gumus,Z.H., Costa,J. and Bonvin,A.M.J.J. (2016) SpotON: a web server for prediction of protein-protein binding hot-spots.

2 METHODOLOGY

Contrary to the reductionist approach aimed at understanding individual components, the new data revolution will allow the understanding of complicated interactions and pathways through the use of statistical and ML techniques (38). The volume of biological data that becomes available every day is transforming the way research is done in the field of bioinformatics. However, the gap between raw protein data and functional knowledge extraction can be attributed to the fact that experimental bench work is highly costly from a time and money point of view. Computational approaches arise as a practical and viable solution in understanding structure and function as a dual relationship (4). ML algorithms were already successfully applied in a variety of subjects such as chemogenomics approaches in virtual screening against G-Protein Coupled Receptors (39), gene expression (40, 41), proteomics mass spectrometry data (42), metabolomics (43), just to name a few.

2.1 ML BASIS

ML are general-purpose approaches defined as the automatic extraction of information from large amounts of data by efficient algorithms, in order to discover patterns and correlations and build predictive models. ML involves the creation of algorithms that improve their performance when undertaking a certain task based on its own experience. They should be fully automatic and *off-the-shelf* methods that process the available data and maximize a problem dependent performance criterion (44).

ML algorithms can be broadly classified into 3 main categories:

i) Supervised – based on training a model on data samples that have known categorical class labels associated with them.

ii) Unsupervised – which aim to discover patterns from the data without knowing their labels.

iii) Semi-supervised or active learning – based on training a model using unlabelled data (a small set of labelled data with a large amount of unlabelled data).

The large majority of ML algorithms are designed for binary classification scenarios: positive and negative classification. During training, the algorithms learn a decision boundary in the feature space that separate data points into positive and negative

classes. If there are more than two categorization classes, the problem is said to be multi-class. Not all algorithms are ready to tackle this type of problem but a general approach is to turn a multi-class problem into a binary problem in which one class is classified in opposite to all others. The typical ML workflow, explained in more detail in the following sub-sections, involves:

i) Data cleaning and pre-processing;

ii) Feature extraction;

iii) Model fitting;

iv) Evaluation.

2.1.1 Data Cleansing and Pre-Processing

Data is the key ingredient of all ML systems (45) and ideally should be a uniformly random sample from the database. Due to its relevance in algorithm performance it is important to carefully collect, label, clean and normalize data by subtracting the variable mean (μ) and divide by its standard deviation (σ) (Equation 1). To avoid inaccuracy, outliers should be detected and removed and missing values imputed (46).

$$X' = \frac{X-\mu}{\sigma} \qquad [1]$$

2.1.2 Feature Extraction

The multivariate structure of data causes problems in computation and visualization. Therefore, dimensionality reduction is crucial in ML. This dimensionality reduction is relevant by removing noise and redundancy or by combining relevant measurements into a smaller number of features that still describe the data with sufficient accuracy (42). Features strongly coupled with other features do not provide extra information and unnecessarily bias the result (44).

There are two main methods of dimension reduction: feature selection and feature extraction. In the first method, *k* features of the *d* dimensions are selected and on the second the original *d* dimensions are transformed to a new set of dimensions of which *k* are selected. Feature selection involves the selection of significant attributes for reduction of datasets by removal of redundant or irrelevant features with the aim to increase the accuracy of models and increasing the computational speed. Feature selection can be both supervised and unsupervised. As it is computationally intensive, a number of sub-optimal methods have been proposed. The most common ones for a supervised approach are:

i) Filtering that orders all features using a criterion of how useful they are, such as t-statistics, and adds one feature at a time until performance stops improving;

ii) Wrapper with forward and backward search methods that take into account that two individually poor features may together be informative. In the feature extraction, there is an attempt to find a transformation *y=f(x)* for the original vector x.

For the unsupervised that does not use the categorical class label but rather intrinsic properties of selected or extracted features we have:

i) PCA;

ii) MDS that uses a nonlinear transformation to preserve distances or dissimilarities between objects.

PCA is an orthogonal linear transformation of data to a new coordinate system and constitutes one of the most common practices in ML to reduce data size but still maintain all the useful information. It works by projecting the data into a lower dimension linear space formed by the principal components, in which the variance of the projected data is maximized. For that it computes the covariance matrix:

$$\Sigma = \frac{1}{m}\sum_{i=1}^{n}(x^i)(x^i)^T \qquad [2]$$

That is then diagonalized to calculate the eigenvectors. PCA is often implemented using the Single Value Decomposition, a more stable mathematical procedure:

$$\mathrm{X} = \mathrm{UD}V^T \qquad [3]$$

2.1.3 Model Fitting

It is important to first split the database into the training set that represent a percentage of the data and the remaining data should be used as a test set, where an independent performance analysis can be made. Also, bootstrapping or *k*-fold cross-validation should be used, which is especially important if the dataset is small. In the cross-validation method, which was used in this work, the dataset is usually divided randomly into *K* equal-sized parts. *K-1* parts are then used to train the model and the remaining one (validation set) to evaluate it. This process is repeated *K* times. Extensive tests on various datasets with different learning algorithms have shown that *K = 10* is about the right number of folds to get the best estimate of error. Moreover, here we went beyond simple cross-validation and we have followed a *k* repeat of *k*-cross-validation, in which

the cross-validation procedure is executed *k* times in order to increase performance. The typical workflow on a ML study is depicted in Figure 3.

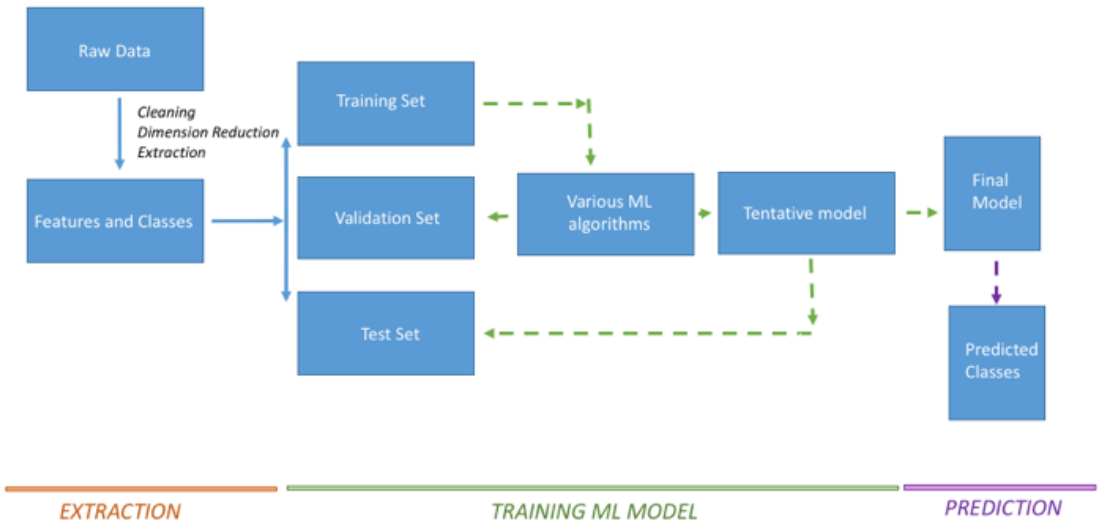


Figure 3. Workflow of a typical ML in HS detection.

The goal of model training is to find parameters *w* that minimize an objective function *L(w)*, which measures the fit between the predictions the model parameterized by *w* and the actual observations.

2.1.4 Evaluation

As already mentioned, evaluation of classification models is essential and should be performed by producing the model on the training set and testing it on an independent test set as performance estimates on the training set would be too much optimistic and heavily overfitted (47). In bioinformatics applications the more natural choices are: the Area Under the Receiver Operator Curve (AUROC), the Accuracy (equation 4), True Positive Rate (TPR/recall/sensitivity, equation 5), True Negative Rate (TNR/specificity, equation 6), Positive Predictive Value (PPV/Precision, equation 7), Negative Predictive Value (NPV, equation 8), False Discovery Rate (FDR, equation 9), False Negative Rate (FNR, equation 10) and F1-score (equation 11).

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{FN}+\text{TN}} \qquad [4]$$

$$\text{TPR} = \frac{\text{TP}}{\text{TP}+\text{FN}} \qquad [5]$$

$$\text{TNR} = \frac{\text{TN}}{\text{FP}+\text{TN}} \qquad [6]$$

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \qquad [7]$$

$$\mathrm{NPV} = \frac{\mathrm{FP}}{\mathrm{FP+TN}} \qquad [8]$$

$$\mathrm{FDR} = \frac{\mathrm{FP}}{\mathrm{FP+TP}} = 1 - \mathrm{PPV} \qquad [9]$$

$$\mathrm{FNR} = \frac{\mathrm{FN}}{\mathrm{TP+FN}} = 1 - \mathrm{TPR} \qquad [10]$$

$$\mathrm{F1-score} = \frac{2\mathrm{TP}}{2\mathrm{TP+FP+FN}} \qquad [11]$$

in the equations above, TP stands for true positive (predicted hot-spots that are actual hot-spots), FP stands for false positive (predicted hot-spots that are not actual hot-spots), FN stands for false negative (non-predicted hot-spots that are actual hot-spots), and TN stands the true negatives (correctly predicted null-spots).

2.2 ML ALGORITHMS

One of the main questions in applying ML to structural biology is finding the optimal classifier complexity for a given problem, which constitutes one of the focus of this thesis. As ML algorithms can usually be separated by similarity in terms of their function, we will briefly explain the main assumptions of the field as well as mathematical notations and formulations for some of the most common approaches.

2.2.1 Discriminant Analysis

Like clustering approaches, dimensionality reduction seeks to exploit the inherent structure in the data. Some of the most common supervised methods are: LDA and QDA. LDA and the related Fisher's linear discriminant are methods used in statistics, pattern recognition and ML to find a linear combination of features, which characterize or separate two or more classes of objects or events. The final combination may be used as a linear classifier or, more commonly, for dimensionality reduction before later classification. In more detail, assuming the density of each class is modelled as multivariate Gaussian:

$$f_k(x) = \frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}} e^{\{-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)\}} \qquad [12]$$

LDA assumes that the covariance for each predictor is the same $\Sigma_k = \Sigma \ \forall k$ (homoscedasticity assumption). In a binary classification problem with two classes *k* and *l*, it is enough to take into account the log-ratio as it gives a linear equation in x:

$$log\frac{\Pr(G=k|X=x)}{\Pr(G=l|X=x)} = log\frac{f_k(x)}{f_l(x)} + log\frac{\pi_k}{\pi_l} = log\frac{\pi_k}{\pi_l} - \frac{1}{2}(\mu_k+\mu_l)^T\Sigma^{-1}(\mu_k-\mu_l) + x^T\Sigma^{-1}(\mu_k-\mu_l) = X^T\Sigma^{-1}(\mu_k-\mu_l) + c_0 \qquad [13]$$

From the previous equation, it is derived that the linear discriminant functions are equivalent to the decision rule $G(x) = argmax_k\delta_k(x)$:

$$\delta_k(x) = x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + log\pi_k \qquad [14]$$

The parameters of the Gaussian distributions cannot be easily calculated and need to be estimated on the training set by: $\hat{\pi} = \frac{N_k}{N}$ where N_k is the number of class-k observations:

$$\widehat{\pi_k} = \sum_{g_i=k} x_i/N_k \qquad [15]$$

$$\hat{\Sigma} = \frac{\sum_{k=1}^{K}\sum_{g_i=k}(x_i-\mu_k)(x_i-\widehat{\mu_k})^T}{N-K} \qquad [16]$$

If the variables have different covariances, the method is called QDA:

$$\delta_k = -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x-\mu_k)^T\sum_k^{-1}(x-\mu_k) + log\pi_k \qquad [17]$$

in which the decision boundary is a quadratic function:

$$\{x: \delta_k(x) = \delta_l(x)\} \qquad [18]$$

2.2.2 Decision Trees

Decision trees are machine learning models that structure the knowledge used to discriminate between examples in a tree-like structure with the root at the top and the leaves at the bottom. The root splits into two or more branches that continue to split until a leaf is reached, a node that cannot be further split. These can model highly nonlinear decision boundaries. They are usually constructed top-down by choosing a variable in each step that best splits the dataset, which is usually evaluated by measuring the homogeneity of the target variable within the subsets. One of the most common metrics is the Gini impurity:

$$i(t) = 1 - \sum_{i=1}^{k} p_i^2 \qquad [19]$$

where *i* stands for the observations $i \in \{1,\ldots,k\}$ and p_i their probability. Another common approach is the calculation of entropy:

$$i(t) = -\sum_{i=1}^{k} p_i \log(p_i) \qquad [20]$$

Simple decision trees are very easy to interpret but are more prone to overfitting and to suffer from high variance (48). Their prediction sensitivity is highly influenced by the quality and complexity of input data. To overcome this pitfall, Random Forest methods were constructed using multiple randomized trees and combining their output in which the majority vote leads to the prediction of individual classes (49). An average of *N* identically distributed random variables, each with variance σ^2 has variance $\frac{1}{N}\sigma^2$. Considering the positive pairwise correlation ρ, the variance of the average is given by:

$$\rho\sigma^2 + \frac{1-\rho}{N}\sigma^2 \qquad [21]$$

As the feature's number increases, the variance of the average is restricted to the first term, and so the size of correlation of bagged trees limits the benefits. Random forest improves variance by reducing the correlation between the trees without increasing variance too much.

2.2.3 Ensemble

These are powerful and popular techniques that are composed of multiple weaker models, which are independently trained and whose predictions are combined at the end. There are 3 types of ensemble models:

i) Bagging, also called Bootstrap aggregating: builds multiple models with equal weight of the same type from random subsamples of the training dataset. Individual classifiers are trained independently.

ii) Boosting: builds multiple models of the same type, in which the more recent learns to fix the prediction error of the previous model. The training is sequential and iterative but more prone to overfitting of the data.

iii) Voting: builds multiple models of different types and use simple metrics to combine predictions.

A boost classifier has the form:

$$H_\tau(x_i) = \sum_{k=1}^{K} f_k(x_i) \qquad [22]$$

in which f_k is the output of a weak learner with the input *x* and that returns the class of the object. The predicted class is identified by the sign and the confidence of

classification is given by the absolute value. The sum training error of the state *n* of the boost classifier is minimized by:

$$E_n = \sum_i E[H_{t-1}(x_i) + \alpha_n h(x_i)] \qquad [23]$$

$H_{t-1}(x)$ is the boosted classifier built from previous step, *E(H)* is an error function and $f_n(x) = \alpha_n h(x)$ is the weak learner considered for addition to the final classifier. AdaBoost is one of the most important ones, with solid theoretical foundation, accurate prediction, great simplicity and successful application (50).

2.2.4 SVM

SVM were first developed by Vapnik and coworkers (51) and base their prediction in the concept of linear separability between classes. They are some of most accurate and robust methods (50). These algorithms aim to minimize both the complexity of the classifier and the number of misclassifications on the training set, the so called structural risk minimization (51). For a linearly separable dataset, a linear classification function corresponds to a separating hyperplane *f(x)* that passes through the middle of the two classes. However, as there are many linear hyperplanes, SVM tries to find the best function by maximizing the margin between the two classes, which confers it the best generalization ability. SVMs are defined by the criteria used to define the optimal linear classifier based on the concept of separation margin maximization, by the identification of the so-called support vectors, the minimal set of training instances that are necessary to define the optimal linear classifier. SVM classifiers attempt to maximize the following function with respect to b and $\vec{w}$:

$$L_p = \frac{1}{2}\|\vec{w}\| - \sum_{i=1}^{t} \alpha_i \gamma_i (\vec{w}.x_i + b) + \sum_{i=1}^{t} \alpha_i \qquad [24]$$

where *t* is the number of training examples, and α_i , $i \in \{1, \ldots, t\}$, are non-negative numbers such that the derivatives *of* L_P with respect to α_i are zero. α_i are the Lagrange multipliers and L_P is called the Primal Lagrangian. In this equation, the vectors $\vec{w}$ and constant *b* define the hyperplane. To minimize the Lagrangian, we take derivatives of *w* and *b* and set them to 0:

$$\frac{\partial L_p}{\partial w} = 0 \Longrightarrow w = \sum_{i=1}^{t} \alpha_i \gamma_i x_i \qquad [25]$$

$$\frac{\partial L_p}{\partial b} = 0 \Longrightarrow \sum_{i=1}^{t} \alpha_i \gamma_i = 0 \qquad [26]$$

This Primal Lagrange function can be substituted by the Dual, which is easier to solve numerically:

$$L_D = -\frac{1}{2}\sum_{i=1}^{t}\sum_{i'=1}^{t}\alpha_i\alpha_{i'}\gamma_i\gamma_{i'}\langle h(x_i), h(x_{i'})\rangle + \sum_{i=1}^{t}\alpha_i \qquad [27]$$

$h(x_i)$ are transformed feature vectors involved through inner products and for which only the kernel function knowledge is required to compute the inner products in the transformed space:

$$K(x, x') = \langle h(x_i), h(x_{i'})\rangle \qquad [28]$$

So, in most real-world problems, which involve data distribution that are not linearly separable, it is common to use kernels (kernel-trick) to transform the original set of variables into a higher order non-linear space. The kernel function *k* is substituted into the dual of the Lagrangian, allowing the determination of a maximum margin hyperplane in the transformed space. There are four typical families of functions:

i) $\text{Linear} \quad K(x, x') = x \cdot x'$ [29]

ii) $\text{Polynomial} \quad K(x, x') = (\gamma x \cdot x' + c)^d$ [30]

iii) $\text{Sigmoid} \quad K(x, x') = \tanh(\gamma x \cdot x' + c)$ [31]

iv) $\text{RBF} \quad K(x, x') = e^{-\gamma\|x - x'\|^2}$ [32]

The first three are global ones and RBF is a local kernel. The choice of SVM kernel is dependent on empirical and experimental analysis as no well-established method was yet designed for this selection.

2.2.5 Neural Networks

ANNs are statistical ML models inspired by the workings of the brain (52) and are composed of a collection of computational elements (neurons) that are interconnected. ANNs have several advantages as the ability to perform multiple training steps, detecting all possible interactions and requiring less formal statistical training. The common MLP architecture combines layers of perceptron-like processing elements (neurons) connected by weighted connections (synapses) (53). The neurons are grouped into layers with only full synaptic connection between successive layers. The layers that receives the signal are the input layers and all others are hidden layers that propagate the signal until the output layer. The depth corresponds to the number of hidden layers and the width is related to the maximum number of neurons in one of its layers. The weights are free parameters that capture the representation of the model and are learned

from samples. Each neuron is a variant of a linear classifier, but the inclusion of multiple neurons and layers result in the construction of sophisticated nonlinear classifiers that allow for their application to complex problems. Mathematically each input is characterized by a real number x_i, where $j \in \{1,\ldots,k\}$. The mapping of information from input to output is modeled by:

$$y_j = \sigma(w_o + \sum_{j=1}^{k} w_j x_j) \qquad [33]$$

where w_i is the weight assigned to each input line and w_o plays the role of a threshold value. The activation function σ enables the use of different gradient techniques for learning algorithms and can be given by:

i) Logistic sigmoid[2] $\sigma(z) = \frac{1}{1+e^{-z}}$ [34]

ii) Hyperbolic tangent $\sigma(z) = \tanh(z)$; [35]

iii) Rectified linear $\sigma(z) = \max\{0, z\}$: [36]

The weights can be attained using the sum-of-squared error (equation 37) or cross-entropy/deviance (equation 38):

$$R(w) = \sum_{j=1}^{K} \sum_{i=1}^{N} (y_{ij} - f_j(x_i))^2 \qquad [37]$$

$$R(w) = -\sum_{i=1}^{N} \sum_{j=1}^{K} y_{ij} log f_j(x_i) \qquad [38]$$

Usually, we do not want to achieve the global minimizer *R(w)* as it would lead to an overfitted solution and instead regularization should be introduced either directly through a penalty term or indirectly by early stopping. *R(w)* is then minimized by gradient descent, by back-propagation. A major recent advance in ML is the automatization by learning a suitable representation of the data with deep artificial neural networks. A deep neural network takes the raw data at the lowest (input) layer and transforms them into increasingly abstract feature representations by successively combining outputs from the preceding layer in a data-driven manner, encapsulating highly complicated functions. The potential of deep learning in high-throughput biology is clear: in principle, it affords better exploitation of increasingly large and highly-dimensional data sets by training complex networks with multiple layers that capture their internal structure. The learned networks discover high-level features, improve performance over traditional models, increase interpretability and provide additional understanding about the structure of the biological data.

2.2.6 Naïve Bayes

The probabilistic approach to modeling uses probability theory to express all forms of uncertainty (45). NB is based on Bayes' theorem, which provides a mathematical framework for describing the probability of an event that might be the result of two or more causes. NB is easy to construct, robust and performs quite well, even outperforming more sophisticated alternatives. NB assumes that given a class $G = j$, the features x_k are independent:

$$f_j(X) = \prod_{k=1}^{p} f_{jk}(X_k) \qquad [39]$$

Using the logit-transform we get:

$$log\frac{\Pr(G=l|X)}{\Pr(G=J|X)} = \alpha_l + \sum_{k=1}^{p} g_{lk}(X_k) \qquad [40]$$

2.2.7 Instance-Based

Instance based algorithms are also called winner-take-all and memory-based learning approaches and typically compare training and test data by some similarity measure. The most popular one, which was also used in this study, is *k*-NN.

K-NN followed in this work looks into a group of k objects in the training set that are close to the object and assigns a label based on a predominance of a particular class in its neighborhood (50). *K*-NN is a simple and intuitive ML algorithm where an observation is classified according to the majority vote.

$$\text{Majority voting: } y' = \text{argmax}_v \sum_{(x_i, y_i) \in D_z} I(v = y_i) \qquad [41]$$

where v is a class label, y_i is the class label for the i^{th} nearest neighbors, and I is an indicator function that returns the value 1 if its argument is true and 0 otherwise. The *k*-NN is sensitive to the local structure of the data and therefore can be used for calculating properties with strong locality, as is the case of protein function.

2.2.8 Regression-Based

These classifiers involve a more probabilistic view of classification and aim to attain the posterior probabilities:

$$\Pr(\omega_k|X = x) = \frac{\pi_k f_k(x)}{\sum_{i=1}^{k} \pi_i f_i(x)} \qquad [42]$$

The model has a linear form:

$$log\frac{\Pr(\omega_{k-1}|X=x)}{\Pr(\omega_K|X=x)} = \beta_{(k-1)0} + \beta_{k-1}^T \mathrm{x} \qquad [43]$$

Maximum likelihood and the Newton-Raphson algorithms are used to fit this linear model. Usually, these methods make no assumption about distribution of classes in the feature space, are quickly trained, have good accuracy, resistant to overfitting and can interpret model coefficients as indicators of feature importance.

The choice of ML algorithm while studying a particular problem should be made in light of its characteristics, deep familiarity with the theoretical foundations of the field, data source and prediction performance (48). ML is an active area of research in computer science with the increasing availability of big data collections of all sorts prompting interest in the development of novel tools for data mining. It is expected that continuous improvement of software infrastructure will make ML applicable to a growing range of biological problems. Silicon Valley companies understand the value of ML in the biology world and have been investing millions of dollars to address the usage of scalable ML tools and their application to this field. As an example, Facebook founder, Mark Zuckerberg, and his wife, Priscilla Chan, have recently announced a 3 billion dollars' contribution to the creation of a network of researchers of different fields of knowledge with the intent of preventing, curing and managing disease. It's foreseeable that a significant part of this investment will be guided toward new computational methods and techniques that ensure such an outcome, as is the case of computational biology methods.

It is thus now the time to develop and apply new techniques to transform the current state-of-the-art and possibly leading to the reliable molecular-level prediction of HS at protein-protein interfaces.

3 METHODS

3.1 HS DETECTION METHOD

3.1.1 Dataset Construction

We constructed a database of complexes by combining information from the ASEdb (54), the BID (55), SKEMPI (56) and PINT (57) databases. Combined they provide both experimental $\Delta\Delta G_{binding}$ values for interfacial residues and tridimensional (3D) X-ray structure information. The protein sequences were filtered to ensure a maximum of 35% sequence identity for at least one protein in each interface. Crystal structures were retrieved from the PDB (7) and all water molecules, ions and other small ligands were removed. Our final dataset consists of 545 mutations from 53 different complexes.

3.1.2 Sequence/Structural Features

From a structural point of view, we compiled 12 previously used different SASA descriptors for all interfacial residues: i) $_{comp}SASA_i$ the solvent accessible surface area of residue i in the complex form; ii) $_{mon}SASA_i$ the residue SASA in the monomer form; iii) $\Delta SASA_i$, the SASA difference upon complexation (equation 44); iv) $_{rel}SASA_i$ the ratio between ΔSASA for each residue and the $_{mon}SASA_i$ value for the same residue (equation 45). Four additional features ($_{comp/res}SASA_i$, $_{mon/res}SASA_i$, $_{\Delta/res}SASA_i$ and $_{rel/res}SASA_i$), defined by equations 46 to 49, were determined by applying amino-acid standardization and dividing the previous features by the average protein $_{res}SASA_r$ values as determined by Miller and colleagues (58, 59), with r being the respective residue type. Four other amino-acid standardized features were calculated by replacing the values determined by Miller by our own protein averages $_{ave}SASA_r$ for each amino-acid type in its respective protein: $_{comp/ave}SASA_i$, $_{mon/ave}SASA_i$, $_{\Delta/ave}SASA_i$ and $_{rel/ave}SASA_i$ defined in equations 50 to 53.

$$\Delta SASA_i = \left| {}_{comp}SASA_i - {}_{mon}SASA_i \right| \qquad [44]$$

$$ {}_{rel}SASA_i = \frac{\Delta SASA_i}{{}_{mon}SASA_i} \qquad [45]$$

$$ {}_{comp/res}SASA_i = \frac{{}_{comp}SASA_i}{{}_{res}SASA_r} \qquad [46]$$

$$_{\mathrm{mon/res}}\mathrm{SASA_i} = \frac{_{\mathrm{mon}}\mathrm{SASA_i}}{_{\mathrm{res}}\mathrm{SASA_r}} \quad [47]$$

$$_{\Delta/\mathrm{res}}\mathrm{SASA_i} = \frac{\Delta\mathrm{SASA_i}}{_{\mathrm{res}}\mathrm{SASA_r}} \quad [48]$$

$$_{\mathrm{rel/res}}\mathrm{SASA_i} = \frac{\mathrm{relSASA_i}}{_{\mathrm{res}}\mathrm{SASA_r}} \quad [49]$$

$$_{\mathrm{comp/ave}}\mathrm{SASA_i} = \frac{_{\mathrm{comp}}\mathrm{SASA_i}}{_{\mathrm{ave}}\mathrm{SASA_r}} \quad [50]$$

$$_{\mathrm{mon/ave}}\mathrm{SASA_i} = \frac{_{\mathrm{mon}}\mathrm{SASA_i}}{_{\mathrm{ave}}\mathrm{SASA_r}} \quad [51]$$

$$_{\Delta/\mathrm{ave}}\mathrm{SASA_i} = \frac{\Delta\mathrm{SASA_i}}{_{\mathrm{ave}}\mathrm{SASA_r}} \quad [52]$$

$$_{\mathrm{rel/ave}}\mathrm{SASA_i} = \frac{\mathrm{relSASA_i}}{_{\mathrm{ave}}\mathrm{SASA_r}} \quad [53]$$

We further introduced two features directly related to the size of the interface: the total number of interfacial residues and the $\Delta SASA_{total}$ (sum of the $\Delta SASA_i$ of all residues at the protein-protein binding interfaces). Twenty other features were added by splitting the total number of interface residues into the 20 amino-acid types. Four contact features were also calculated: i) and ii) the number of protein-protein contacts within 2.5 Å and 4.0 Å distance cut-offs, respectively; iii) the number of intermolecular hydrogen bonds and iv) the number of intermolecular hydrophobic interactions. In-house scripts using the VMD molecular package (60) were used for all these calculations. In total, we used 38 structural features in our study.

The evolutionary sequence conservation information was introduced, upon using CONSURF server (61, 62), that calculates a conservation score for each amino-acid at an interfacial position for a complex, based on known sequences in different organisms. We also computed PSSM using BLAST (63, 64) as well as the weighted observed percentages, introducing them as 40 new features for all interfacial residues. Positive values in this matrix appear for substitutions more frequent than expected by random chance and negative values indicate that the substitution is not frequent. So, a total of 41 evolutionary sequence-related features were added to the structural features, resulting in 79 features total. These features were used in a previous version of the study, published during the course of this master's thesis (28). In the meanwhile, we have extended the sequence related features to include 850 ones extracted from the PROTR

(65) module from the R package: i) the ACC of protein, the fraction of each amino acid type within the protein; ii) PAAC (66) adds up to the standard 20 amino acid definition, providing information about patterns; iii) amphiphilic PAAC (67), a set of the twenty original amino acids, plus descriptors regarding the hydrophobicity/hydrophilicity of the sequences that have often displayed positive effects regarding protein-protein interaction prediction algorithms; iv) BLOSUM which provides evolutionary features in the form of a scoring matrix upon sequence alignment taking into account amino acid substitution at a 62% level of similarity; v) Protein Fingerprinting, a process that allows for the identification and differentiation of proteins by unique characteristics, sometimes despite sequence similarity and is generated from both the AAindex and by PCA; vi) PCM (68) derived from PCA of 2D and 3D descriptors, that allows for a perspective regarding protein dynamics and interaction with ligands. Due to the large increase in available data on the human genome a much deeper characterization and understanding of sequences is now possible. Therefore, we have integrated it in the context of structural understanding of proteins leading to a better description of PPIs.

We totalize a final of 929 features for which all results will be presented in section 4.1. These features were calculated for 545 observations, each one corresponding to an amino acid residue classified as HS or NS. We have written all the feature calculation code in Python and will make it available to all researchers in the area on GitHub.

3.1.3 Machine-Learning Techniques

In this study, we used the Classification and Regression Training (Caret) Package (69) from the R software, which provides a unified interface with a large number of built-in classifiers, in order to train a HS predictor. We randomly split this dataset (for details see Annex Table SI-1) into a training set consisting of 70% of data (382 mutations/observations) and an independent test set (163 mutations/observations - 30%). This is a standard division scheme demonstrated to give a good result. One of the main concerns when applying classification to the detection of HS is the natural imbalance of the data. As expected, the number of HS is lower than the number of NS at a protein-protein interface, as indicated by the presence of 185 HS and 360 NS in the main dataset. In ML classification methods, the disparity of frequencies of the observed classes may have a very negative impact on the models performance. To overcome this problem, we have tried two different subsampling techniques for the training set: down-sampling and up-sampling. In the first, there is a random sub-setting of all classes at the training set with their class frequency matching the least prevalence class (HS), whereas in the up-sampling the opposite is happening with random sampling (with the

replacement) of the minority class (HS) to reach the same size as the majority class (NS). The 54 algorithms tested were: Boruta, C5.0, C5.0Rules, C5.0Tree, J48, LogitBoost, ORFlog, ORFpls, ORFridge, ORFsvm, RRF, RRFglobal, ada, adaboost, amdai, avNNet, bagEarth, bagEarthGCV, bagFDA, bagFDAGCV, ctree, ctree2, dwdPoly, dwdRadial , evtree, fda, gamboost, gbm, glm, glmboost, hdda, knn, lda, lda2, loclda, multinom, nb , parRF , pda , plr , qda, ranger, rda, rf, stepLDA, stepQDA, svmLinear, svmLinear2, svmPoly, svmRadial, svmRadialCost, svmRadialSigma, svmRadialWeights and wsrf.

All the classification models were tested using 10-fold cross validation repeated 10 times in order to avoid overfitting and to obtain the model's generalization error. This means that the training set was split randomly into ten isolated parts, using nine of the ten parts to train the model and taking the remaining fold of data to test the final performance of the model. This process was repeated ten times. Two different sets were tested in which:

i) the variables were normalized;

ii) the variables were normalized and then subjected to PCA.

Both techniques are described in more detail in sections 2.1.1 and 2.1.2. The validity and performance of the various methods was determined by measuring the AUROC, the Accuracy, TPR, TNR, PPV, NPV, FPR, FNR and F1-score described in section 2.1.4 over our dataset. The calculations for the various algorithms were written in R and performed in parallel for the various conditions.

The 54 algorithms were analysed on different attributes for which a binary value was given (1/0 if present/absent). These were subjected to hierarchical clustering that returned a distance matrix using the Jaccard similarity coefficient as a metric and the Ward aggregation scheme. The different clusters attained were compared by three different techniques to discriminate among the various groups: two nonparametric procedures (MRPP and ANOSIM) and the parametric MANOVA. In MRPP, the *delta* (the weighted mean within-group distance) for *g* groups was calculated based on the average distance matrix calculated in each group ($\overline{d_i}$):

$$delta = \delta = \sum_{i=1}^{g} \frac{n_i}{N} \overline{d}_i \qquad [\,54\,]$$

Here, n_i is the number of items in group I and N is the total number of items. ANOSIM is also a nonparametric procedure that is based on the calculation of dissimilarity matrixes and their ranking. It calculates the test statistics *R* (an index of relative within-group dissimilarity) upon calculating the mean among- and within-group rank dissimilarities:

$$R = \frac{\overline{r_A} - \overline{r_W}}{N(N-1)} \times 4 \qquad [55]$$

In which N(N-1)/2 is the number of sample pairs. Then, for both procedures, the probability of a *delta/R* value is calculated through Monte Carlo permutations that involve randomly assigning sample observations to groups.

We have also used one-way MANOVA, a parametric test to check if the groups differ from each other significantly in one or more characteristics. The two hypotheses tested were:

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_L \; vs \;\; H_1: \mu_r \neq \mu_s \; for \; one \; pair \; r, s.$$

MANOVA calculates the two matrices of between- and within-scatter:

$$H = k \sum_{i=1}^{L} (\overline{x_{i\cdot}} - \bar{x}..)(\overline{x_{i\cdot}} - \bar{x}..)^T \qquad [56]$$

$$E = k \sum_{i=1}^{L} \sum_{j=1}^{K} (x_{ij} - \overline{x_i}.)(x_{ij} - \overline{x_i}.)^T \qquad [57]$$

Considering that A = $H \times E^{-1}$, four different statistics were calculated based on the eigenvalues λ_p of the A matrix:

i) Samuel Stanley Wilks $\lambda_{Wilks} = det(I + A)^{-1}$ [58]

ii) Pillai M S. Barlett trace $\lambda_{Pillai} = tr((I + A)^{-1})$ [59]

iii) Laeley-Hotelling $trace \; \lambda_{LH} = tr(A)$ [60]

iv) Roy's greatest root $\lambda_{Roy} = max_p(\lambda_p)$ [61]

3.1.4 Comparison with other HS Detection Software

We compared our results with some of the common methods in the literature: ROBETTA (70), KFC2-A and KFC2-B (30) and CPORT (71).

4 RESULTS

4.1 HS DETECTION METHOD

4.1.1 Exploratory Data Analysis

The accuracy of ML depends largely on the quality of the feature sets and the experimental data available to train the model. A few databases contain information about a handful of experimentally determined HS, and a non-redundant representative dataset can be construct with a vast coverage of all relevant type of interactions. However, this data, as the majority of data in biology, is still atypical for ML, too sparse and incomplete, too biased and too noisy (72). Moreover, the field is marked by imbalanced data, which turns the selection of proper performance measures and algorithms even more important.

Our final dataset includes 545 amino acids from 53 complexes (140 HS and 405 NS). More clearly, the final number of observations are 545 with 140 of them belonging to the positive class and 405 to the negative one. For these observations, we began our work by calculating 79 features that were extended to 929 used in this thesis. We calculated the percentage of the different type of amino acids within HS and NS sets:

i) NS set - SER: 7.4; GLY: 1.5; PRO: 2.0; VAL: 3.2; LEU: 2.7; ILE: 5.2; MET: 1.0; CYS: 0.7; PHE: 4.7; TYR: 5.9; TRP: 4.9; HIS: 4.4; LYS 8.9; ARG: 10.6; GLN: 5.4; ASN: 6.2; GLU: 9.9; ASP: 7.2; THR: 8.1;

ii) HS set - SER: 2.1; GLY: 2.9; PRO: 2.9; VAL: 3.6; LEU: 7.1; ILE: 4.3; MET: 0.0; CYS: 0.0; PHE: 6.4; TYR: 20.0; TRP: 5.7; HIS: 2.1; LYS 7.1; ARG: 6.4; GLN: 2.1; ASN: 5.0; GLU: 7.1; ASP: 10.7; THR: 4.3.

For both sets, there is a natural expected tendency for a higher percentage of large hydrophobic or charged residues at the interfaces, in particular TYR. Although different patterns could influence the training of a robust classifier, we have previously successfully constructed models that were bias-free for all different amino acids (27).

As in any statistical study, we began by performing an EDA to investigate the database and summarize its main characteristics. Various libraries (R-packages) were used to implement it: *PerformanceAnalytics, ggplot2, reshape2, FactoMineR, factoextra, corrplot.* This step was of particular importance to maximize the insight into the data. In

particular, we plotted comparative boxplots and histograms of the distribution of all the numeric variables within our dataset and calculated simple statistics such as mean, standard deviation, just to name a few. Correlation r^2 and *p-values* for the Pearson correlation test were also calculated using the R package *corrplot* and the *Performance Analytics* packages. Only 4 SASA-based descriptors (Equations 46 to 49) demonstrated to have high correlation with $_{comp}SASA_i$, $_{mon}SASA_i$, $\Delta SASA_i$, and $_{rel}SASA_i$, as they are simple standardization of the same metric. However, as they proved to be relevant in our previous studies (26-28) we decided to keep them in our dataset.

The features used in this work have different scales (i.e. the range of the raw data varies significantly), and therefore we have performed feature normalization or data standardization of the predictor variables at the training set by centering the data, i.e. subtracting the mean and normalizing it by dividing by the standard deviation. The same protocol was followed for the test set taking into consideration the use of the training mean and standard deviation to ensure a good estimation of the model quality and generalization power. As we have a high-dimensional dataset, we have also applied PCA to reduce the dimensionality of the data. As explained in detail at the 2.1.2 Section, PCA works by establishing an orthogonal transformation of the data to convert a set of possible correlated variables into a set of linearly uncorrelated ones, the so-called principal components. In particular, on the preprocessing function of Caret, SVD is used on covariance matrixes. PCA was shown to be an acceptable trade-off between computational time, data variance and model performance (73). We plotted the variances explained by the first 49 principal components (Figure 4), the ones that account for a cumulative percentage variance $\frac{\sum_{i=1}^{d}\lambda_i}{\sum_i \lambda_i} \geq 95\%$, and that will be considered in this study. As this is a case in which the number of observations (n) is lower than the number of features (p), the number of principal components with non-zero variance attained (49) could not exceed n-1.

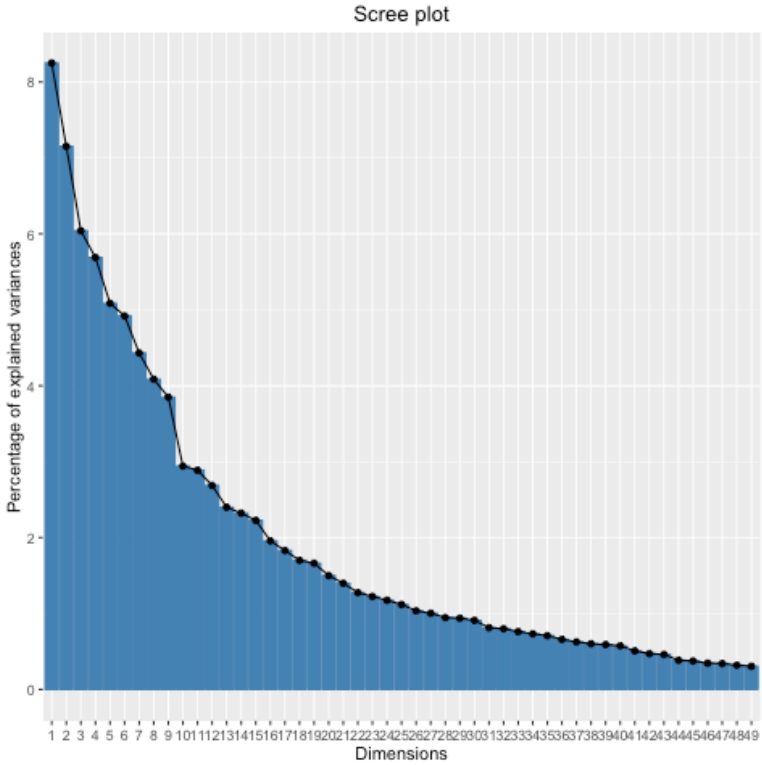


Figure 4. Plot of percentage of explained variance versus dimension considered.

Different conditions were thus established:

i) Scaled - dataset generated upon normalization of variables;

ii) Scaled_Up - dataset generated upon normalization of variables and up-sampling of the minor class (HS);

iii) Scaled_Down - dataset generated upon normalization of variables and down-sampling of the major class (NS);

iv) PCA - dataset generated upon normalization of variables and PCA;

v) PCA_Up - dataset generated upon normalization and PCA of variables and up-sampling of the minor class (HS);

vi) PCA_Down - dataset generated upon normalization and PCA of variables and down-sampling of the major class (NS).

Various statistical metrics (described in detail in Section 2.1.4) were adopted to evaluate the performance of the algorithms tested. Figure 5 illustrates the final workflow followed in this study.

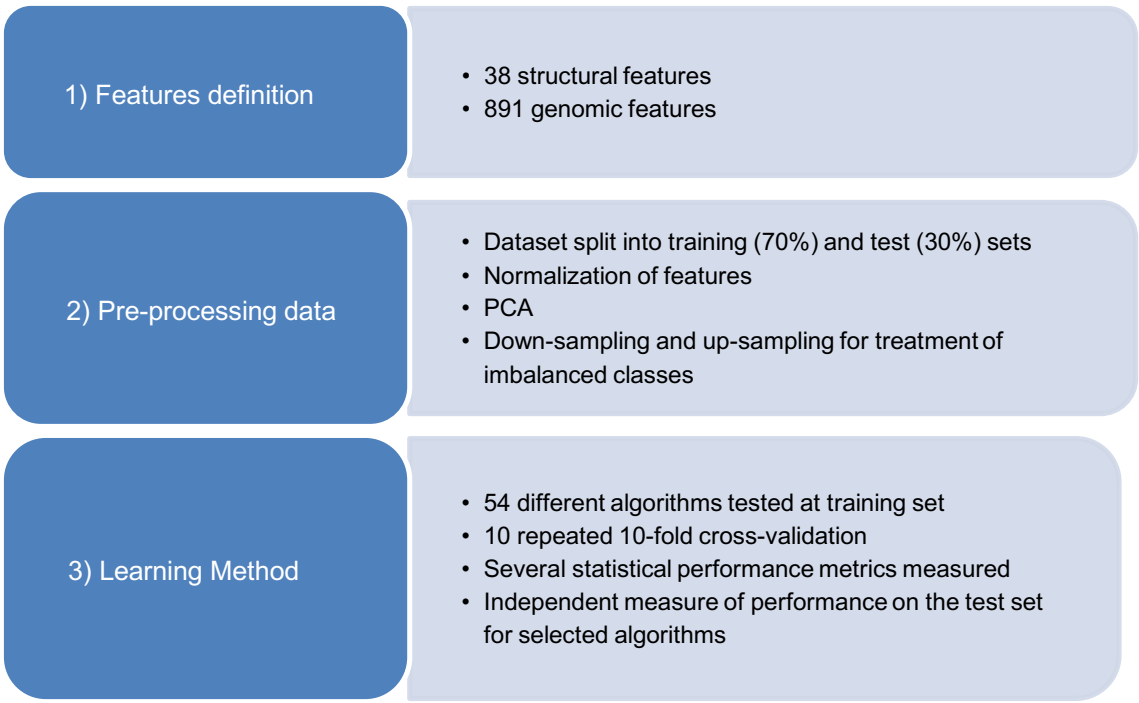


Figure 5. The flowchart of the current work.

4.1.2 Clustering of ML algorithms

54 algorithms were tested. For a better performance comparison, and due to the difficulty in categorizing ML approaches in a simple way, we began by characterizing them in agreement with Caret's tags (69): Accepts Case Weights, Bagging, Bayesian Model, Binary Predictors Only, Boosting, Categorical Predictors Only, Cost Sensitive Learning, Discriminant Analysis, Distance Weighted Discrimination, Ensemble Model, Feature Extraction, Feature Extraction Models, Feature Selection Wrapper, Gaussian Process, Generalized Additive Model, Generalized Linear Model, Generalized Linear Models, Handle Missing Predictor Data, Implicit Feature Selection, Kernel Method, L1 Regularization, L1 Regularization Models, L2 Regularization, L2 Regularization Models, Linear Classifier, Linear Classifier Models, Linear Regression, Linear Regression Models, Logic Regression, Logistic Regression, Mixture Model, Model Tree ,Multivariate Adaptive Regression Splines, Neural Network, Oblique Tree, Ordinal Outcomes, Partial Least Squares, Polynomial Model, Prototype Models, Quantile Regression, Radial Basis Function, Random Forest, Regularization, Relevance Vector Machines, Ridge Regression, Robust Methods, Robust Model, ROC Curves, Rule-Based Model, Self-Organizing Maps, String Kernel, Support Vector Machines, Text Mining, Tree-Based Model and Two Class Only. For all tags, a binary attribute was given with a value of 1 (if present) or 0 (if not present).

All methods were then subjected to hierarchical clustering that returned a distance matrix using the Jaccard similarity coefficient as a metric and the Ward aggregation scheme. The dendrogram is depicted in Figure 6, which allows us to distinguish 5 main clusters:

i) Cluster I (mainly random forest and bagging based): bagEarth, bagEarthGDV, bagFDA, bagFDAGCV, RRF, RRFglobal, wsrf, ranger, parRF, rf;

ii) Cluster II (mainly tree-based and random forest models): LogistBoost, ada, adaboost, C5.0, gbm, fda, C5.0Rules, C5.0Tree, J48, evtree, ctree, ctree2;

iii) Cluster III (random forest to neural models): Orfridge, ORFsvm, PRFlog, PRFpls, multinom, plr, glmboost, glm, nb, knn, avNNet, Boruta

iv) Cluster IV (SVM models): svmLinear, svmLinear2, svmPoly, svmRadial, svmRadialCost, svmRadialSigma, and svfmRadialWeights;

v) Cluster V (mainly linear pr quadratic models): stepLDA, loclda, lda2, hdda, kda, dwdPoly, amdai, rda, stepQDA, pda, qda

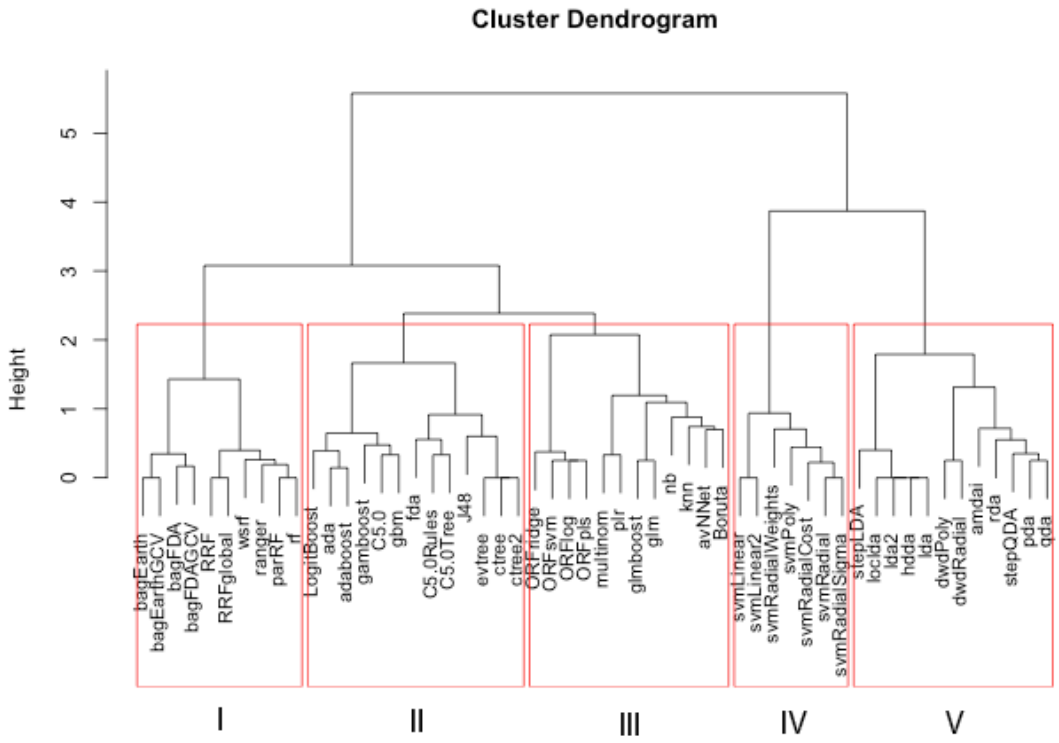


Figure 6. Cluster Dendrogram of the ML algorithms tested in this work.

4.1.3 ML algorithms Performance Discrimination

We present extensive statistical measures, covering all possible aspects of the assessment proposed so far, for the six conditions at Annexes Tables SI-2 to SI-7. For some approaches, the algorithms used did not converged and are not listed in the Annexes section. Figure 7 and Figure 8 illustrate the mean values and box-plot distributions of the sum of AUROC, TPR and TNR metrics for all six pre-processing conditions studied. From the Scaled conditions, it seems that C4 algorithms performed differently from the remaining ones with lower mean and wider distributions. For the PCA conditions it seems that C3 members present on average higher mean.

We have then performed various statistical analysis to access the real discrimination power between the 5 attained clusters: C1, C2, C3, C4 and C5. In particular, we used MRPP, ANOSIM and MANOVA for all 6 pre-processing conditions, and the p-values attained are listed in Table 1.

Table 1. P-Values for the statistical methods used to discriminate between groups. 1- All 8 metrics were used; 2- AUROC, TPR, TNR metrics were used.

PRE-PROCESSING	METHODS / P-VALUE				
	MRPP 1	MRPP 2	ANOSIM 1	ANOSIM 2	MANOVA
Scaled	0.089	0.069	0.081	0.093	0.069
Scaled_Up	0.198	0.094	0.186	0.158	0.276
Scaled_Down	0.039	0.023	0.033	0.059	0.022
PCA	0.039	0.021	0.058	0.029	0.002
PCA_Up	0.042	0.018	0.083	0.091	0.001
PCA_Down	0.047	0.015	0.040	0.028	0.001

MRPP and ANOSIM are nonparametric procedures for testing the hypothesis of no difference between the 5 groups based on permutation test of among- and within- group dissimilarities. With the exception of Scaled-Up pre-processing condition, it can be concluded for this test that at a significant level of 0.10, the 5 clusters differ significantly in terms of the measured performance metrics.

MANOVA is a parametric test that has some assumptions: multivariate normality of the data, multivariate homoscedasticity, no multicollinearity, and that there are no multivariate outliers. As all algorithms are organized already by similarity, they are not independent and these assumptions are not fulfilled by our data. However, fortunately MANOVA is usually robust to violations of theses assumptions, which nevertheless are

hard to prove on a multivariate perspective, and we can have confidence on the attained results. The same conclusion retrieved from the non-parametric procedures was achieved by application of MANOVA.

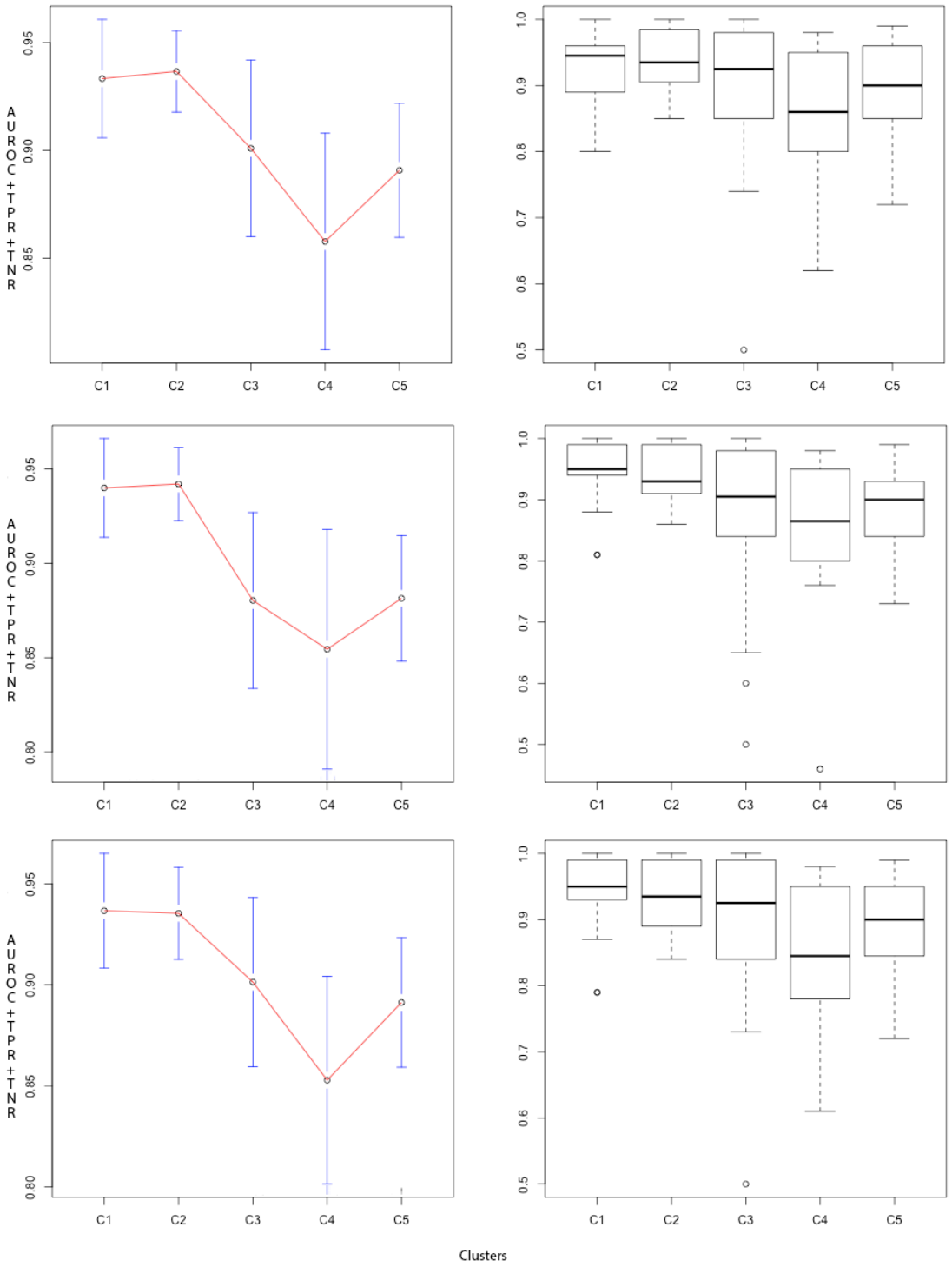


Figure 7. Mean of AUROC, TPR and TNR metrics for the Scaled, Scaled-Up and Scaled-down pre-processing conditions on the left panel. Right panels are the box-plots of the same metrics over the 5 clusters.

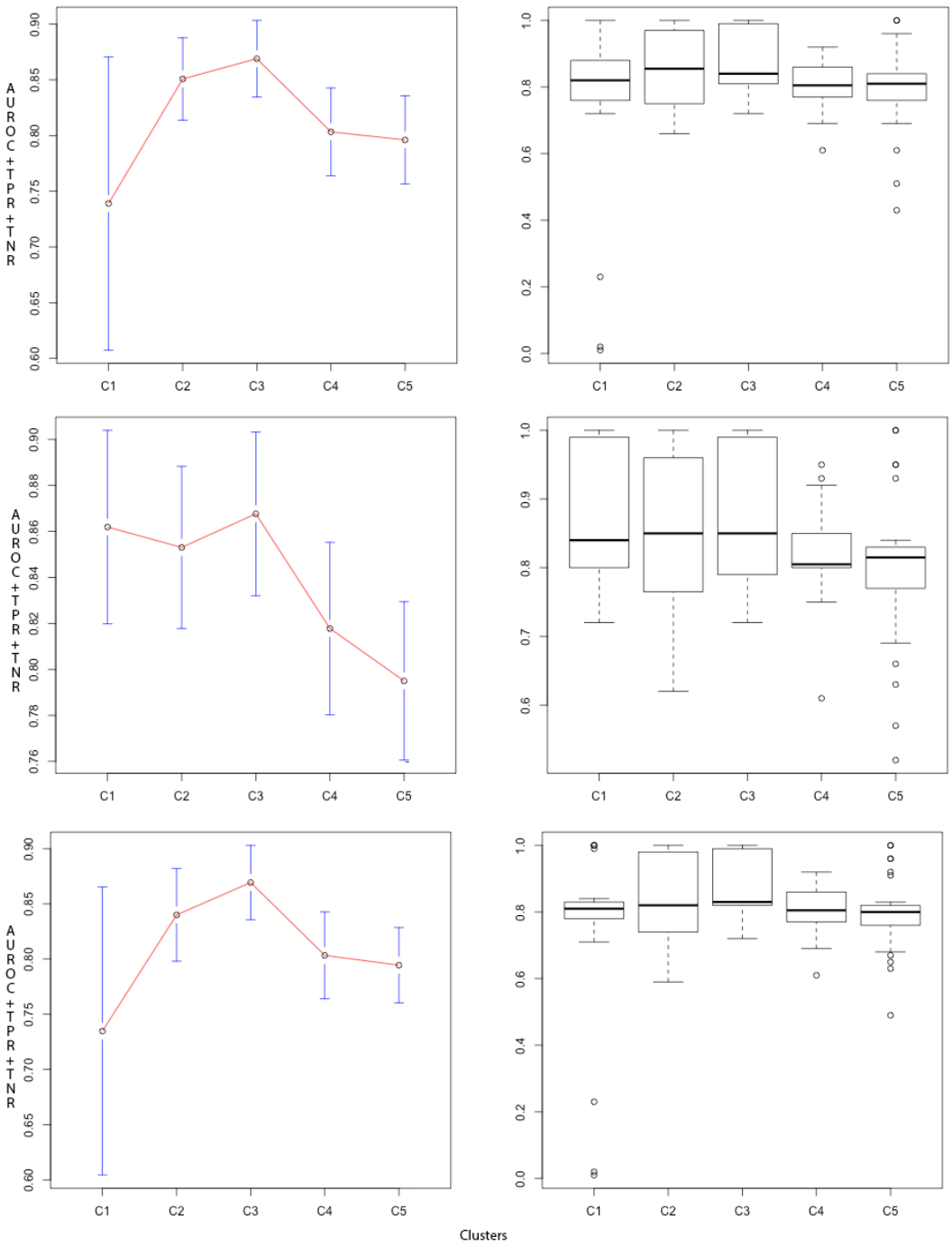


Figure 8. Mean of AUROC, TPR and TNR metrics for the PCA, PCA-Up and PCA-Down pre-processing conditions on the left panel. Right panels are the box-plots of the same metrics over the 5 clusters.

4.1.4 ML algorithms Performance Comparison

The results for the training set for the 5 best algorithms (as clustered in Figure 6) and for each of the 6 conditions studied are listed in Table 2.

Table 2. Statistical metrics attained for 5 algorithms with top performance for each of the studied conditions at the training set.

PRE-PROCESSING	METRICS	ALGORITHMS				
		Cluster I	**Cluster II**	**Cluster III**	**Cluster IV**	**Cluster V**
PCA		*bagEarthGCV*	*adaboost*	*ORFlog*	*svmPoly*	*dwdPoly*
	AUROC	0.83	0.83	0.84	0.81	0.81
	Accuracy	0.81	1.00	1.00	0.89	0.87
	TPR	0.79	1.00	1.00	0.86	0.82
	TNR	0.83	1.00	1.00	0.92	0.91
	PPV	0.81	1.00	1.00	0.91	0.89
	NPV	0.81	1.00	1.00	0.88	0.85
	FDR	0.19	0.00	0.00	0.09	0.11
	F1-score	0.80	1.00	1.00	0.89	0.86
PCA_Up		*parRF*	*adaboost*	*ORFlog*	*svmLinear*	*lda2*
	AUROC	0.84	0.85	0.86	0.83	0.83
	Accuracy	1.00	1.00	1.00	0.82	0.80
	Sensitivity	1.00	1.00	1.00	0.80	0.77
	Specificity	1.00	1.00	1.00	0.85	0.83
	PPV	1.00	1.00	1.00	0.84	0.82
	NPV	1.00	1.00	1.00	0.81	0.78
	FPR	0.00	0.00	0.00	0.16	0.18
	F1-score	1.00	1.00	1.00	0.82	0.79
PCA_Down		*parRF*	*adaboost*	*ORFridge*	*svmPoly*	*lda2*
	AUROC	0.82	0.82	0.83	0.81	0.81
	Accuracy	1.00	1.00	0.99	0.89	0.80
	Sensitivity	1.00	1.00	0.99	0.86	0.77
	Specificity	1.00	1.00	1.00	0.92	0.82
	PPV	1.00	1.00	1.00	0.91	0.81
	NPV	1.00	1.00	0.99	0.88	0.78
	FPR	0.00	0.00	0.00	0.09	0.19
	F1-score	1.00	1.00	0.99	0.89	0.79
Scaled		*bagEarthGCV*	*gbm*	*glmboost*	*svmLinear*	*dwdPoly*
	AUROC	0.96	0.94	0.92	0.91	0.91
	fAccuracy	0.96	1.00	0.90	0.97	0.99
	Sensitivity	0.96	1.00	0.91	0.98	0.98

	Specificity	0.96	1.00	0.9	0.97	0.99
	PPV	0.96	1.00	0.89	0.97	0.99
	NPV	0.96	1.00	0.91	0.98	0.98
	FPR	0.04	0.00	0.11	0.03	0.01
	F1-score	0.96	1.00	0.90	0.97	0.99
Scaled_Up		*bagEarthGCV*	*C5.0*	*glmboost*	*svmLinear*	*lda*
	AUROC	0.96	0.93	0.92	0.92	0.90
	Accuracy	0.96	0.99	0.90	0.98	0.94
	Sensitivity	0.96	0.98	0.90	0.98	0.93
	Specificity	0.96	0.99	0.90	0.97	0.94
	PPV	0.96	0.99	0.90	0.97	0.94
	NPV	0.96	0.98	0.90	0.98	0.93
	FPR	0.04	0.01	0.10	0.03	0.06
	F1-score	0.96	0.99	0.90	0.98	0.94
Scaled_Down		*bagEarthGCV*	*Gbm*	*glmboost*	*svmLinear*	*dwdPoly*
	AUROC	0.95	0.94	0.92	0.91	0.90
	Accuracy	0.95	1.00	0.90	0.98	0.99
	Sensitivity	0.96	1.00	0.91	0.98	0.99
	Specificity	0.95	1.00	0.90	0.97	0.99
	PPV	0.95	1.00	0.90	0.97	0.99
	NPV	0.96	1.00	0.91	0.98	0.99
	FPR	0.05	0.00	0.10	0.03	0.01
	F1-score	0.95	1.00	0.90	0.98	0.99

The statistical measures presented, and commonly used in the ML field, have inherent problems. Accuracy can provide deceivingly high numbers for unbalanced data, both AUROC and Accuracy put the emphasis on performance in areas not of interest for researchers and F1-score relies on unknown class priors (72). However, AUROC is one of the most widely used measures as shows the trade-off between the fraction of true positive and false positives as a function of a threshold on the output of the classifier.

So, in this work, we used AUROC as the main statistical measure to rank the performance of the classifiers. In case of draw between different classifiers, we used the TPR as second choice since, from a biological point of view, the correct classification of HS is more important than the one of the NS. Nevertheless, TNR was taken into account when necessary as these 3 metrics have shown the best discriminative power. Table 2 shows that various ML techniques perform quite well. The best classifiers from the six different pre-processing conditions are:

i) Cluster I: bagEarthGCV, parRF;

ii) Cluster II: adaboost, gbm, C5.0;

iii) Cluster III: ORFlog, ORFridge, glmboost;

iv) Cluster IV: svmLinear, svmPoly;

v) Cluster V: dwdPoly, lda2, lda.

The AUROC values vary between 0.76 and 0.83 on the 3 PCA-based cases and between 0.80 and 0.88 on the remaining 3 of the best 5 classifiers. In this case, it seems that PCA leads to a decrease of the performance due to some loss of important information.

The performance of a classifier on the training set from which it was constructed gives a poor estimate of its accuracy in new cases as the training error is likely to be lower than the actual generalization error. Overfitting on algorithms without regularization terms (such as decision trees and random forests) is harder to address on the training set. The overfitting problem is even bigger for these biological datasets as the number of observations at the training set (382) is roughly 40% of the number of tested features (929) and the large discrepancy between observations and number of features could lead to incorrect prediction on a new dataset. This is known as the *"Hughes effect"* or curse of dimensionality and it appears when the number of predictors (p) is much higher than the number of available training examples (n). This problem is however quite common. For example, in the Kaggle competitions there are situations with less than 300 data points in the training set and around 28.000 dimensions. In these p >>n situations, one of the major problems is the inclusion of irrelevant/noise attributes as a set of them can become the truly relevant ones due to random fluctuations, not contributing to the reduction of classification error. Introduction of more data can lead to sparseness of the training data and therefore the accurate estimation of the classifier's parameters (e.g. decision boundaries) becomes more difficult. Also, sparseness is not uniformly distributed over the search space. There is no fixed rule that defines how many features can be used in a classification problem as it depends of the amount of training data available, the complexity of decision boundaries and the type of classifier used. To overcome this problem, we have used PCA to reduce the dimensionality space and a built-in feature selection method available in 50% of all tested algorithms (ada, adaboost, bagEarth, bagEarthGCV, bagFDA, bagFDAGCV, C5.0, C5.0Rules, C5.0Tree, ctree, ctree2, evtree, fda, gamboost, gbm, J48, LogitBoost, ORFlog, ORFpls, ORFridge, ORFsvm, parRF, ranger, rf, RRF, RRFglobal and wsrf). Built-in feature selection can be more efficient than algorithms where search routine for the right predictors is external to the model, and typically couples the predictor search algorithm with the parameter estimation and are usually optimized with a single objective function. Also, the true

predictive accuracy of the classifier was estimated on a separate test set corresponding to 30% of the main dataset. Table 3 summarizes the performance on the independent test set for the best classifiers shown in Table 3.

Table 3. Statistical metrics attained for 5 algorithms with top performance for each of the studied conditions at the independent test set.

PRE-PROCESSING	METRICS	ALGORITHMS				
		Cluster I	**Cluster II**	**Cluster III**	**Cluster IV**	**Cluster V**
PCA		*bagEarthGCV*	*adaboost*	*ORFlog*	*svmPoly*	*dwdPoly*
	AUROC	0.78	0.78	0.78	0.78	0.76
	Accuracy	0.78	0.78	0.78	0.78	0.76
	Sensitivity	0.78	0.75	0.74	0.71	0.70
	Specificity	0.78	0.81	0.82	0.84	0.81
	PPV	0.77	0.78	0.79	0.81	0.77
	NPV	0.79	0.78	0.77	0.76	0.74
	FDR	0.22	0.22	0.21	0.19	0.23
	F1-score	0.75	0.77	0.77	0.76	0.73
PCA_Up		*parRF*	*adaboost*	*ORFlog*	*svmLinear*	*lda2*
	AUROC	0.77	0.78	0.78	0.83	0.80
	Accuracy	0.76	0.78	0.78	0.83	0.80
	Sensitivity	0.81	0.75	0.77	0.84	0.81
	Specificity	0.72	0.81	0.80	0.82	0.80
	PPV	0.73	0.78	0.78	0.81	0.78
	NPV	0.80	0.78	0.79	0.85	0.81
	FPR	0.27	0.22	0.22	0.19	0.22
	F1-score	0.77	0.77	0.77	0.83	0.79
PCA_Down		*parRF*	*adaboost*	*ORFridge*	*svmPoly*	*lda2*
	AUROC	0.74	0.78	0.76	0.78	0.79
	Accuracy	0.74	0.78	0.76	0.78	0.79
	Sensitivity	0.75	0.75	0.73	0.71	0.77
	Specificity	0.73	0.81	0.80	0.84	0.82
	PPV	0.73	0.78	0.77	0.81	0.80
	NPV	0.76	0.78	0.76	0.76	0.79
	FPR	0.27	0.22	0.23	0.19	0.2
	F1-score	0.74	0.77	0.75	0.76	0.78
Scaled		*bagEartGCV*	*gbm*	*glmboost*	*svmLinear*	*dwdPoly*
	AUROC	0.84	0.87	0.83	0.81	0.83
	Accuracy	0.84	0.86	0.82	0.81	0.82
	Sensitivity	0.87	0.91	0.87	0.84	0.83
	Specificity	0.82	0.82	0.78	0.78	0.82
	PPV	0.82	0.82	0.79	0.78	0.81

	NPV	0.87	0.91	0.87	0.84	0.84
	FPR	0.18	0.18	0.21	0.22	0.19
	F1-score	0.84	0.86	0.83	0.81	0.82
Scaled_Up		*bagEarthGCV*	*C5.0*	*glmboost*	*svmLinear*	*Lda*
	AUROC	0.84	0.88	0.82	0.81	0.80
	Accuracy	0.84	0.88	0.82	0.81	0.80
	Sensitivity	0.87	0.91	0.86	0.84	0.79
	Specificity	0.82	0.84	0.78	0.77	0.81
	PPV	0.82	0.84	0.79	0.84	0.79
	NPV	0.87	0.91	0.86	0.84	0.81
	FPR	0.18	0.16	0.21	0.23	0.21
	F1-score	0.84	0.88	0.82	0.81	0.79
Scaled_Down		*bagEarthGCV*	*gbm*	*glmboost*	*svmLinear*	*dwdPoly*
	AUROC	0.85	0.86	0.82	0.76	0.80
	Accuracy	0.84	0.86	0.82	0.76	0.80
	Sensitivity	0.88	0.88	0.83	0.74	0.78
	Specificity	0.81	0.84	0.82	0.78	0.82
	PPV	0.81	0.84	0.81	0.76	0.80
	NPV	0.88	0.89	0.84	0.76	0.80
	FPR	0.19	0.16	0.19	0.24	0.20
	F1-score	0.84	0.86	0.82	0.75	0.79

From Table 3 it is clear than even the more overfitted methods still perform really well on an independent test set. The AUROC values at the test set still range between 0.76 and 0.83 for the PCA pre-processing cases and between 0.76 and 0.88 for the remaining, which are clearly high.

From all methods, C5.0, trained on the normalized up-scaling set, had the highest performance metrics on the independent test set. It was thus chosen as a final model. C5.0 is significantly faster than its precedent C4.5, more efficient, uses smaller decision trees, has support for boosting, introduces different weights and also allows the winnowing of the attributes that could potentially decrease performance. C5.0 can produce two kinds of models: a decision tree or a rule set. A decision tree follows the explanation of Section 2.2.2, and exactly one prediction is possible. In contrast, a rule set is a set of rules that makes predictions for individual observations. These are derived from decisions trees but in a more simplified way. The crucial difference is that for a rule set, more than one rule may apply for a particular observation or no rules at all may apply. In the first situation, the observation will be classified according to a combination of the weights for all the applied rules. If no rules apply, a default prediction is assigned to the observation. Figure 9 (Panel A) illustrates the AUROC values for the various tuning parameters tested for the C5.0 algorithm: rules or tree-based and with or without

winnowing. Although all of them are quite high, the best model consists on rule-based algorithm with 20 boosting interactions and without winnowing of features.

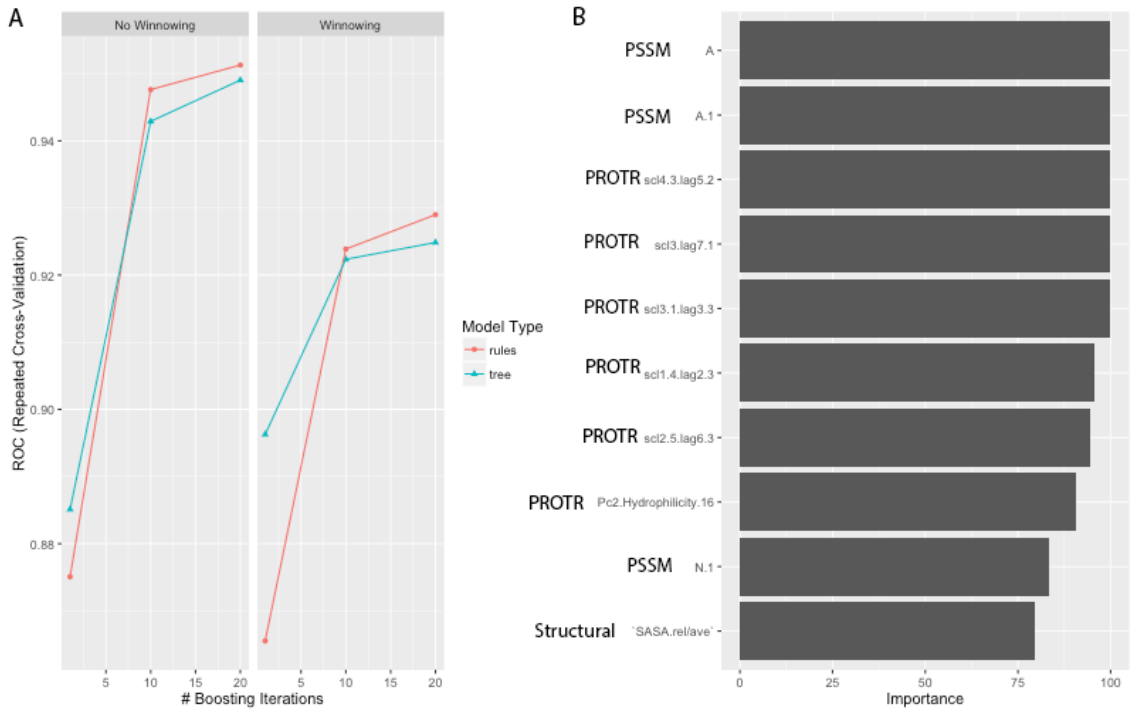


Figure 9. A: ROC plot for the best C5.0 classifier: B: Top 10 features used by the chosen C5.0 algorithm.

In our analysis of this classifier (Figure 9 – panel B), we observed that the key features are sequence-related ones: 3 PSSM values, 6 PROTR values introduced more recently and a structural one that seems consisted in all our applications of the method (related to $_{rel}SASA_i$ or one of its standardizations) (26-28).

To validate the accuracy of the best predictor, we performed the HS predictions with other methods reported in literature such as ROBETTA [19] , KFC2-A [20], KFC2-B [20] and CPORT (not specialized in HS prediction but instead a protein-protein interface predictor) [21] on the same training and test sets. Comparison among these ML methods (Table 4) demonstrates that our new method achieves the best performance with F1-scores/AUROC values of 0.88/0.88 on the test set, compared to our previous approach 0.73/0.78 and 0.39/0.62, 0.56/0.66, 0.42/0.67 and 0.43/0.54 for ROBETTA, KFC2-A, KFC2-B and CPORT, respectively.

Table 4. Comparison of the statistical metrics attained for the best predictor in this work and some of the most common ones in literature

METRICS	ALGORITHMS													
	C5.0/UP-SCALING		c-forest/ up-scaling classes		SBHD2		Robetta		KFC2-A		KF2-B		CPORT	
	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST
AUROC	0.93	0.88	0.85	0.78	0.74	0.69	0.62	0.62	0.72	0.66	0.60	0.67	0.54	0.54
Accuracy	0.99	0.88	0.93	0.80	0.70	0.71	0.66	0.66	0.76	0.71	0.70	0.73	0.49	0.49
Sensitivity	0.98	0.91	0.93	0.76	0.70	0.70	0.38	0.29	0.57	0.53	0.26	0.28	0.55	0.54
Specificity	0.99	0.84	0.93	0.82	0.70	0.71	0.85	0.88	0.85	0.81	0.93	0.96	0.45	0.47
PPV	0.99	0.84	0.93	0.70	0.55	0.56	0.61	0.60	0.67	0.59	0.65	0.80	0.34	0.35
NPV	0.98	0.91	0.93	0.86	0.82	0.82	0.68	0.67	0.79	0.77	0.71	0.72	0.66	0.66
F1-score	0.99	0.88	0.93	0.73	0.62	0.62	0.47	0.39	0.62	0.56	0.37	0.42	0.42	0.42

Figure 10 clearly shows that SVM is the most common algorithm applied in the field. As we observed in this work the reason is clear as they tend to perform really well. However, other ML algorithms are shown to be has good as SVM or even better, and should be applied in the future to structural bioinformatics studies. In particular, during this thesis we concluded that C5.0 seems especially indicated to HS detection.

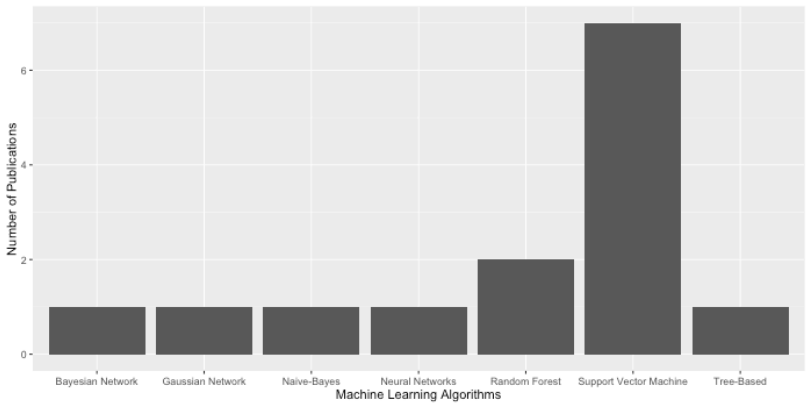


Figure 10. ML-based algorithms for HS detection based on the ones reviewed by Moreira *et al*. (10) as well as our other 2 recent approaches (27, 28).

4.2 SPOT-ON: WEB SERVER FOR HS PREDICTION

In the current era of shared information, it is crucial that all methodologies, algorithms and scripts are free-available and easy to use for any researcher interested in the subject. Thus, we implemented our accurate predictor in a user-friendly web-server that serves a wide community of non-experts in ML.

4.2.1 Input

A screenshot of the submission page can be seen in Figure 11. The interface requires the user to upload a 3D structure of the protein-protein complex in the Protein Data Bank (PDB) format (9) and a CONSURF (10, 11) conservation scores file for it. The conservation scores can be easily calculated at http://consurf.tau.ac.il/2016/. The user should also specify the chain identifiers of the two monomers. The choice of the chains that constitute monomer A or B is completely arbitrary. Instructions for all the input are available in the Help section in addition to popups in the submission page. The first step every SpotOn user needs to complete is to register with an email address of their choice, which is used to authenticate him/her during job submission. Although the server is freely available, registration is required since the user email is used for various notifications about the progress of the job. Upon successful job submission the user receives an email with the URL address where the output of the run will appear as soon as the analysis is complete. An additional email notification containing the URL of the results page is sent upon completion, informing the user of the success or failure of the run.

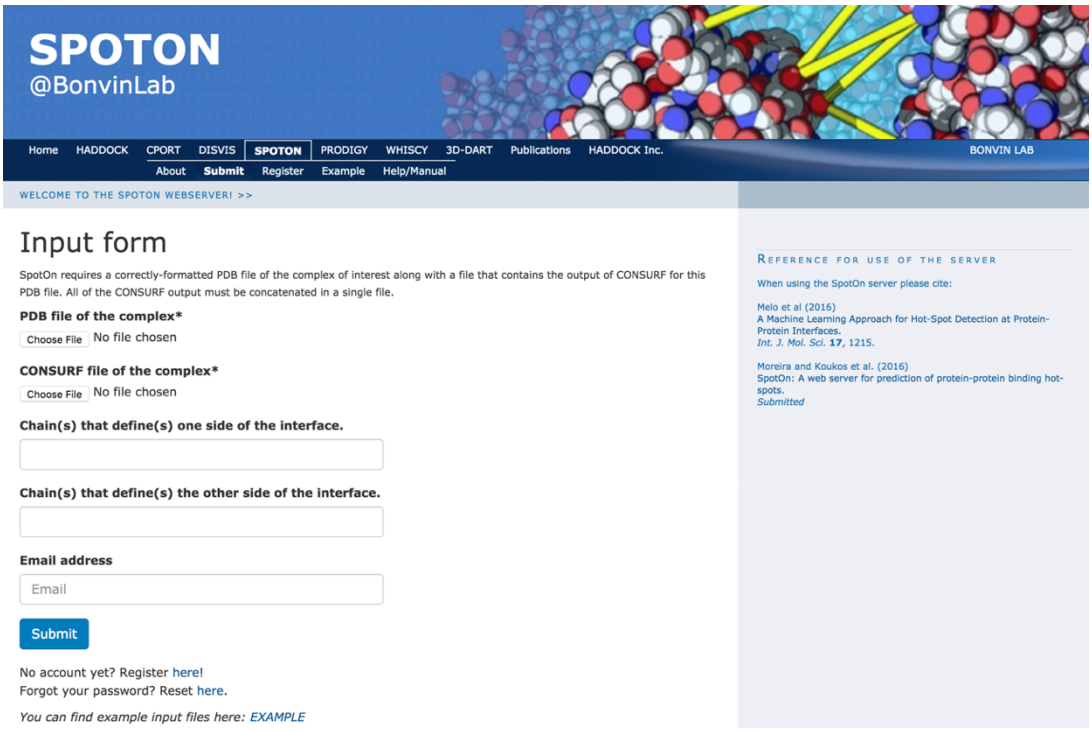


Figure 11. Screenshot of the SpotOn server submission page.

4.2.2 Output and Representation of the Results

The main outputs of the server are the two tables that list the residues classified as HS and NS. Figure 12 illustrates the output for an example case (PDBid: 1Z7X (21)) and contains the list of residues predicted as HS. Any column can be used to sort the table.

This table along with the NS table are also made available as CSV files in the archive of the run that the user can download. The information contained in those two tables is also visualized in the form of a line plot (e.g. Figure 13) which provides pertinent information when the user hovers the cursor over it (chain identifier, name and index of each residue). This enables the user to quickly identify the residues that have been identified as HS.

HOT-SPOT TABLE

This table contains a list of the residues which have been classified as HotSpots by the algorithm.

Residue Index	Residue Name	Residue Chain	HotSpot Probability
409	LEU	W	0.7
41	LYS	X	0.678
111	GLU	X	0.678
39	ARG	X	0.666
436	ILE	W	0.662
67	ASN	X	0.654
91	ARG	X	0.645
407	ASN	W	0.634
38	GLY	W	0.625
11	GLN	X	0.605

Figure 12. Example table of residues identified as Hot Spots along with their probabilities for the complex with PDBid 1Z7X (21). Only the top 10 Hot-Spots are shown.

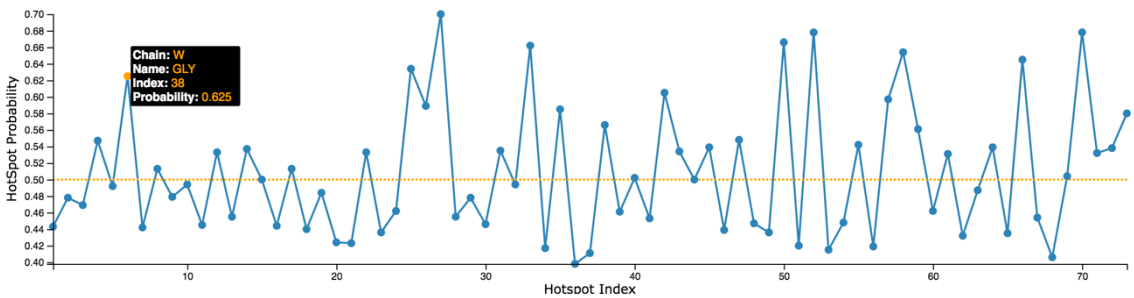


Figure 13. Probability chart of an interface residue being a Hot Spot. Residues above the orange line at 0.50 are predicted as HS and those below as NS. Such a chart is presented to users on the results page.

Finally, the result page provides a direct visualization of the identified HS within the interface of the complex in the form of pre-generated, publication quality views of the complex (Figure 14), that are outputs of the Chimera software (22).

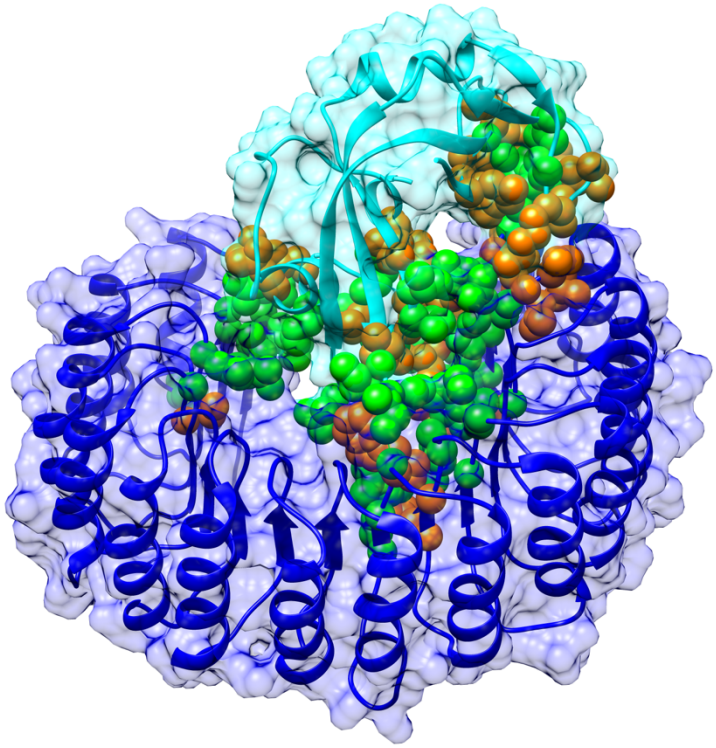


Figure 14. Graphical output example of SpotOn server showing a view of the complex between ribonuclease inhibitor (blue ribbons) and ribonuclease (cyan ribbons), respectively, with a transparent surface representation (PDBid 1Z7X (21)).

For each run, all generated results are provided as two gzipped archives, which the user can download from the provided links. The first contains all the graphical outputs of the program: Chimera images, a static version of the plot described above as well as similar plots that display the probability of a residue being a HS for the entire molecule, broken down by chain identifier. The second archive contains all the text outputs: the CSV file that details all the features (refer to the method paper for details (8)) for the interfacial residues, and the CSV files of the two tables of the results page.

4.2.3 Implementation

The SpotOn server runs alongside the other servers of our group on a local Linux cluster. The backend is implemented in Python and R, but also makes use of external programs, including VMD (18), BLAST (19, 20) and Chimera (22) during the analysis. It makes use of the Flask micro-framework for web development and, in addition to the standard languages of the web (HTML, CSS, JS), utilizes the charting library D3.js (22) for the interactive plots in the results page. All scripts are available on Github (http:github.com/haddocking). Documentation is kept up-to-date and support is offered

via spoton.csbserver@gmail.com and the BioExcel support forum (http://ask.bioexcel.eu). Calculations submitted by users are anonymous and output data to separate directories with randomly generated 12-character key names. Results are kept on the server for 2 weeks. The server workflow is illustrated in Figure 15. If any errors occur at any point of the pipeline illustrated in this figure the analysis will be terminated and an email will be sent to the user prompting them to review the output of the program. Submissions from users are processed in parallel with a maximum number of 15 jobs running simultaneously. Every user is limited to 3 concurrent runs. Typical runtimes for a prediction range between 30 and 90 minutes.

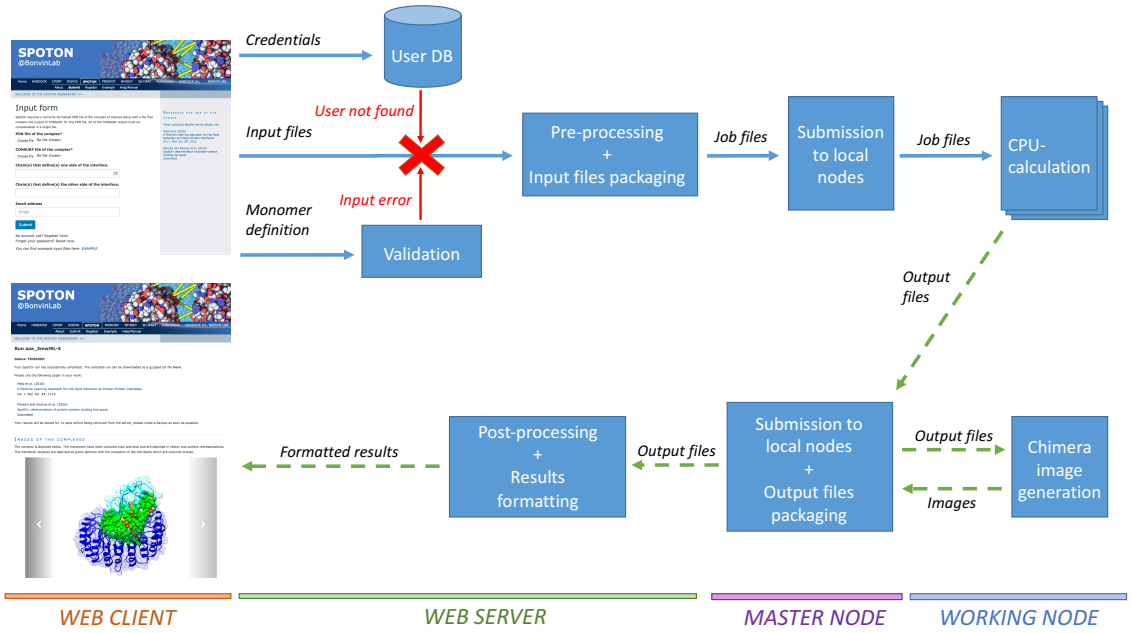


Figure 15. Workflow chart of the entire SpotOn pipeline.

Each box corresponds to a step in the pipeline and the horizontal bars at the bottom of the image indicate the environment in which this step takes place. At the very beginning, the user is required to upload the PDB file and the Consurf output for the same molecule, in addition to defining the two monomers of the interface. After the credentials of the user have been checked and the input data validated, the web server creates the run directory with all the necessary files. Should the data be badly formatted or the user not recognized as a registered user of SpotOn a helpful message will be displayed on screen indicating the problem. The master node of the Linux cluster where SpotOn is hosted monitors the directory where the run folders are located and if the global maximum number of SpotOn jobs or the number of jobs the particular user has submitted hasn't exceeded the limits defined in the Implementation paragraph, the analysis is submitted to the queue. Depending on the load of the system at the time of submission, the analysis might start

immediately or with a small delay. The user is notified as soon as the job starts running. The actual run takes place in one of the working nodes of the cluster and as soon as it is finished, the master node submits another job for the generation of the chimera images based on the results of the analysis. At the same time the result archives are generated on the master node and the user is notified of the job completion via email. With the exception of the chimera images, the rest of the elements of the page are generated by the client in real time.

5 CONCLUSIONS AND FUTURE WORK

Most biological processes within the cell involve the coupling of proteins to form stable complexes able to interact with other proteins or complexes, and activating various cellular pathways. It is fundamental to attain a faithful picture of all interactions made by these complexes to be able to understand high level cellular organization.

In recent years ML has been proven to be crucial to capture protein function from a vast majority of biomolecular data resources and has become widely used in a variety of areas due to its reduced application time and high performance. Over the past years a few algorithms have been applied for the specific problem in this study: the detection of HS at PPIs (19, 26-37). However, dataset selection and treatment as well as performance estimation proven to be major challenges in the application of ML to the field. To advance these application it was necessary to compare the performance of various algorithms and different data extraction techniques and propose a more general methodology. Some classifiers (linear discriminant analysis or generalized linear models) come from statistics, others come from data mining (tree-based) and some are connectionist approaches (such as neural networks), and all can behave differently when applied to different databases. So, the look for the best classifier for this particular subject is crucial, as the No-Free-Lunch Theorem from Wolpert (74) states: "*The best classifier may not be the same for all the datasets".*

We evaluated 54 classifiers arising from different families and compared their performance in 6 different pre-processing sets. These classifiers were subjected to hierarchical clustering and grouped in 5 different clusters. We have compared the algorithms' performance in each cluster and chosen the best of each for a global comparison. The classifiers tested were implement in the caret package, which uses an automatically way of parameterizing them. Caret's in-build function allows parameter tuning and selects the values that maximize the AUROC according to the validation selected (in this case a 10 repeat 10-fold cross-validation). So, every single one of them was tuned to be best possible choice of parameters. Various gave meaningful AUROC values in the range of 0.78 to 0.88, which were especially high if all features were considered (without PCA feature reduction). From a broad perspective the attained accuracy increase is clearly visible when compared to other reported methods. At the end, we chose a C5.0 rule-base algorithm with 20 boosting interactions and no winnowing of features that gave an AUROC value of 0.88 in the independent test set.

The values in the independent test were also very high compared to the ones currently reported in the literature, and surpassing all the other methods tested in this study, including the one achieved at the beginning of the work (Table 4). One important aspect that seemed to improve the results compared to our previous approaches (such as SBHD (26)) was the use of in-build R techniques to balance the training data: up-scaling of the data led to a substantial improvement of the F1-score and to a decrease of the FPR to about 0.19 on the independent test set. Also, the use of more sequence-related features improve the AUROC value from 0.78 (28) to 0.88 in the latest model. In this particular classifier, the first 9 features with higher importance were all sequence-based and one structural that had already been used in previous versions of our algorithm. In conclusion, we were thus able to train an accurate and robust predictor using C5.0 learning method, and up-sampling of the minor class (HS) for dataset balance. These new methods can now be widely applied to the detection of HS in protein-protein interfaces by use of the web-server that we have developed: SpotON.

SpotOn is an easy to use, publicly accessible web server that enables accurate Hot-Spot identification for protein-protein complexes, with minimal input requirements. The method behind it is robust and is arguably the most accurate to date. A successful run will present the user with meaningful results displayed in a user-friendly interactive format that should be equally useful to experts in the field of computational structural biology as well as less computationally trained researchers. SpotOn is part of a family of widely-used web portals operated by the Utrecht group (71, 75, 76) in the general area of biomolecular interaction. As such it is part of services for which we aim at ensuring a high reliability and availability. The ML algorithm behind the webserver is still the one that we recently published (28) but will be updated with the new model developed during the remaining part of this master's thesis.

The work presented here serves as proof of concept about the importance of application of ML in the Bioinformatics field. The use of these techniques to other relevant biological problems such as the construction of 3D structures of protein-protein complexes will allow to go beyond the understanding of the function of individual proteins to the understanding of group proteins and various other iterators, and ultimately to the understanding of the biological pathways. We have now open up this window of opportunity and knowledge, and we intent to explore it in future works.

6 REFERENCES

1. Dorn,M., Silva,M.B.E., Buriol,L.S. and Lamb,L.C. (2014) Three-dimensional protein structure prediction: Methods and computational strategies. *Computational Biology and Chemistry*, **53**, 251–276.

2. Xu,X., Yan,C., Wohlhueter,R. and Ivanov,I. (2015) Integrative Modeling of Macromolecular Assemblies from Low to Near-Atomic Resolution. *CSBJ*, **13**, 492–503.

3. Chakravarty,D., Janin,J., Robert,C.H. and Chakrabarti,P. (2015) Changes in protein structure at the interface accompanying complex formation. *IUCrJ (2015). M2, 643-652 [doi:10.1107/S2052252515015250]*, 10.1107/S2052252515015250.

4. jamil (2013) A Review of Protein Function Prediction Under Machine Learning Perspective.

5. Pruitt,K.D. (2004) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **33**, D501–D504.

6. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–5.

7. Berman,H.M. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

8. Petta,I., Lievens,S., Libert,C., Tavernier,J. and De Bosscher,K. (2015) Modulation of Protein–Protein Interactions for the Development of Novel Therapeutics. *Mol Ther*, **24**, 707–718.

9. Clackson,T. and Wells,J.A. (1995) A hot spot of binding energy in a hormone-receptor interface. *Science*, **267**, 383–386.

10. Moreira,I.S. (2015) The Role of Water Occlusion for the Definition of a Protein Binding Hot-Spot. *Curr Top Med Chem*, **15**, 2068–2079.

11. Moreira,I.S., Fernandes,P.A. and Ramos,M.J. (2007) Hot spots--a review of the protein-protein interface determinant amino-acid residues. *Proteins*, **68**, 803–812.

12. Ramos,R.M. and Moreira,I.S. (2013) Computational Alanine Scanning Mutagenesis-An Improved Methodological Approach for Protein-DNA Complexes. *J Chem Theory Comput*, **9**, 4243–4256.

13. Cunningham,B.C. and Wells,J.A. (1989) High-resolution epitope mapping of hGH-receptor interactions by alanine-scanning mutagenesis. *Science*, **244**, 1081–1085.

14. Bogan,A.A. and Thorn,K.S. (1998) Anatomy of hot spots in protein interfaces. *J. Mol. Biol.*, **280**, 1–9.

15. Keskin,O., Ma,B. and Nussinov,R. (2005) Hot Regions in Protein–Protein Interactions: The Organization and Contribution of Structurally Conserved Hot Spot

Residues. *J. Mol. Biol.*, **345**, 1281–1294.

16. Guharoy,M. and Chakrabarti,P. (2010) Conserved residue clusters at protein-protein interfaces and their use in binding site identification. *BMC Bioinformatics 2009 10:1*, **11**, 286.

17. Chakrabarti,P. and Janin,J. (2002) Dissecting protein–protein recognition sites. *Proteins*, **47**, 334–343.

18. Kwong,P.D., Wyatt,R., Robinson,J., Sweet,R.W., Sodroski,J. and Hendrickson,W.A. (1998) Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody. *Nature*, **393**, 648–659.

19. Brender,J.R. and Zhang,Y. (2015) Predicting the Effect of Mutations on Protein-Protein Binding Interactions through Structure-Based Interface Profiles. *PLoS Comput Biol*, **11**, e1004494–25.

20. Moreira,I.S., Fernandes,P.A. and Ramos,M.J. (2006) Unraveling the importance of protein-protein interaction: application of a computational alanine-scanning mutagenesis to the study of the IgG1 streptococcal protein G (C2 fragment) complex. *J Phys Chem B*, **110**, 10962–10969.

21. Massova,I. and Kollman,P.A. (1999) Computational Alanine Scanning To Probe Protein−Protein Interactions: A Novel Approach To Evaluate Binding Free Energies. *Journal of the American Chemical Society*, **121**, 8133–8143.

22. Moreira,I.S., Ramos,R.M., Martins,J.M., Fernandes,P.A. and Ramos,M.J. (2014) Are hot-spots occluded from water? *J. Biomol. Struct. Dyn.*, **32**, 186–197.

23. Moreira,I.S., Fernandes,P.A. and Ramos,M.J. (2006) Detailed microscopic study of the full zipA:FtsZ interface. *Proteins*, **63**, 811–821.

24. Moreira,I.S., Fernandes,P.A. and Ramos,M.J. (2007) Hot spot occlusion from bulk water: a comprehensive study of the complex between the lysozyme HEL and the antibody FVD1.3. *J Phys Chem B*, **111**, 2697–2706.

25. Moreira,I.S., Martins,J.M., Ramos,R.M., Fernandes,P.A. and Ramos,M.J. (2013) Understanding the importance of the aromatic amino-acid residues as hot-spots. *Biochim. Biophys. Acta*, **1834**, 404–414.

26. Martins,J.M., Ramos,R.M., Pimenta,A.C. and Moreira,I.S. (2014) Solvent-accessible surface area: How well can be applied to hot-spot detection? *Proteins*, **82**, 479–490.

27. Munteanu,C.R., Pimenta,A.C., Fernandez-Lozano,C., Melo,A., Cordeiro,M.N.D.S. and Moreira,I.S. (2015) Solvent accessible surface area-based hot-spot detection methods for protein-protein and protein-nucleic acid interfaces. *J Chem Inf Model*, **55**, 1077–1086.

28. Melo,R., Fieldhouse,R., Melo,A., Correia,J.D.G., Cordeiro,M.N.D.S., Gümüş,Z.H., Costa,J., Bonvin,A.M.J.J. and Moreira,I.S. (2016) A Machine Learning Approach for Hot-Spot Detection at Protein-Protein Interfaces. *IJMS*, **17**, 1215.

29. Darnell,S.J., Page,D. and Mitchell,J.C. (2007) An automated decision-tree approach to predicting protein interaction hot spots. *Proteins*, **68**, 813–823.

30. Zhu,X. and Mitchell,J.C. (2011) KFC2: A knowledge-based hot spot prediction method based on interface solvation, atomic density, and plasticity features. *Proteins*, **79**, 2671–2683.

31. Wang,L., Zhang,W., Gao,Q. and Xiong,C. (2014) Prediction of hot spots in protein interfaces using extreme learning machines with the information of spatial neighbour residues. *IET Syst Biol*, **8**, 184–190.

32. Cho,K.-I., Kim,D. and Lee,D. (2009) A feature-based approach to modeling protein-protein interaction hot spots. *Nucleic Acids Res.*, **37**, 2672–2687.

33. Chen,P., Li,J., Wong,L., Kuwahara,H., Huang,J.Z. and Gao,X. (2013) Accurate prediction of hot spot residues through physicochemical characteristics of amino acid sequences. *Proteins*, **81**, 1351–1362.

34. Tuncbag,N., Keskin,O. and Gursoy,A. (2010) HotPoint: hot spot prediction server for protein interfaces. *Nucleic Acids Res.*, **38**, W402–6.

35. Mora,J.S., Assi,S.A. and Fernandez-Fuentes,N. (2010) Presaging Critical Residues in Protein interfaces- Web Server (PCRPi- W): A Web Server to Chart Hot Spots in Protein Interfaces. *PLoS ONE*, **5**, e12352.

36. Guo,F., Li,S.C., Wei,Z., Zhu,D., Shen,C. and Wang,L. (2015) Structural neighboring property for identifying protein-protein binding sites. *BMC Systems Biology*, **9**, S3.

37. Peri,C., Morra,G. and Colombo,G. (2016) Surface energetics and protein- protein interactions: analysis and mechanistic implications. *Nature Publishing Group*, 10.1038/srep24035.

38. Petrey,D. and Honig,B. (2014) Structural Bioinformatics of the Interactome. *Annu. Rev. Biophys.*, **43**, 193–210.

39. Vass,M., Kooistra,A.J., Ritschel,T., Leurs,R., de Esch,I.J. and de Graaf,C. (2016) ScienceDirect Molecular interaction fingerprint approaches for GPCR drug discovery. *Current Opinion in Pharmacology*, **30**, 59–68.

40. Karlić,R., Chung,H.-R., Lasserre,J., Vlahovicek,K. and Vingron,M. (2010) Histone modification levels are predictive for gene expression. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 2926–2931.

41. Libbrecht,M.W. and Noble,W.S. (2015) Machine learning applications in genetics and genomics. *Nat. Rev. Genet.*, **16**, 321–332.

42. Swan,A.L., Mobasheri,A., Allaway,D., Liddell,S. and Bacardit,J. (2013) Application of machine learning to proteomics data: classification and biomarker identification in postgenomics biology. *OMICS*, **17**, 595–610.

43. Kell,D.B. (2005) Metabolomics, machine learning and modelling: towards an understanding of the language of cells. *Biochem. Soc. Trans.*, **33**, 520–524.

44. Baştanlar,Y. and Özuysal,M. (2013) Introduction to Machine Learning. In Yousef,M., Allmer,J. (eds), *miRNomics: MicroRNA Biology and Computational Analysis*, Methods in Molecular Biology. Humana Press, Totowa, NJ, Vol. 1107, pp. 105–128.

45. Ghahramani,Z. (2015) Probabilistic machine learning and artificial intelligence.

Nature, **521**, 452–459.

46. Donders,A.R.T., van der Heijden,G.J.M.G., Stijnen,T. and Moons,K.G.M. (2006) Review: a gentle introduction to imputation of missing values. *J Clin Epidemiol*, **59**, 1087–1091.

47. de Ridder,D., de Ridder,J. and Reinders,M.J.T. (2013) Pattern recognition in bioinformatics. *Briefings in Bioinformatics*, **14**, 633–647.

48. Lavecchia,A. (2015) Machine-learning approaches in drug discovery: methods and applications. *Drug Discovery Today*, **20**, 318–331.

49. Kittler,J., Hatef,M. and Duin,R. (1998) On combining classifiers. *IEEE transactions*

50. Wu,X., Kumar,V., Ross Quinlan,J., Ghosh,J., Yang,Q., Motoda,H., McLachlan,G.J., Ng,A., Liu,B., Yu,P.S., *et al.* (2007) Top 10 algorithms in data mining. *Knowl Inf Syst*, **14**, 1–37.

51. Vapnik,V. (2013) The nature of statistical learning theory.

52. ROSENBLATT,F. (1958) The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol Rev*, **65**, 386–408.

53. Manning,T., Sleator,R.D. and Walsh,P. (2014) Biologically inspired intelligent decision making. *Bioengineered*, **5**, 80–95.

54. Thorn,K.S. and Bogan,A.A. (2001) ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. *Bioinformatics*, **17**, 284–285.

55. Fischer,T.B., Arunachalam,K.V., Bailey,D., Mangual,V., Bakhru,S., Russo,R., Huang,D., Paczkowski,M., Lalchandani,V., Ramachandra,C., *et al.* (2003) The binding interface database (BID): a compilation of amino acid hot spots in protein interfaces. *Bioinformatics*, **19**, 1453–1454.

56. Moal,I.H. and Férnandez-Recio,J. (2012) SKEMPI: a Structural Kinetic and Energetic database of Mutant Protein Interactions and its use in empirical models. *Bioinformatics*, **28**, 2600–2607.

57. Kumar,M.D.S. and Gromiha,M.M. (2006) PINT: Protein-protein Interactions Thermodynamic Database. *Nucleic Acids Res.*, **34**, D195–8.

58. Miller,S., Janin,J., Lesk,A.M. and Chothia,C. (1987) Interior and surface of monomeric proteins. *J. Mol. Biol.*, **196**, 641–656.

59. Miller,S., Lesk,A.M., Janin,J. and Chothia,C. (1987) The accessible surface area and stability of oligomeric proteins. *Nature*, **328**, 834–836.

60. Humphrey,W., Dalke,A. and Schulten,K. (1996) VMD: Visual molecular dynamics. *Journal of Molecular Graphics*, **14**, 33–38.

61. Ashkenazy,H., Abadi,S., Martz,E., Chay,O., Mayrose,I., Pupko,T. and Ben-Tal,N. (2016) ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Res.*, **44**, W344–50.

62. Glaser,F., Pupko,T., Paz,I., Bell,R.E., Bechor-Shental,D., Martz,E. and Ben-Tal,N.

(2003) ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics*, **19**, 163–164.

63. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics 2009 10:1*, **10**, 421.

64. Altschul,S. (1990) Basic Local Alignment Search Tool. *J. Mol. Biol.*, **215**, 403–410.

65. Xiao,N., Cao,D.-S., Zhu,M.-F. and Xu,Q.-S. (2015) protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences. *Bioinformatics*, **31**, 1857–1859.

66. Du,P., Gu,S. and Jiao,Y. (2014) PseAAC-General: Fast Building Various Modes of General Form of Chou's Pseudo-Amino Acid Composition for Large-Scale Protein Datasets. *IJMS*, **15**, 3495–3506.

67. Chou,K.-C. (2005) Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics*, **21**, 10–19.

68. van Westen,G.J.P., Wegner,J.K., IJzerman,A.P., van Vlijmen,H.W.T. and Bender,A. (2011) Proteochemometric modeling as a tool to design selective compounds and for extrapolating to novel targets. *Med. Chem. Commun.*, **2**, 16–30.

69. Kuhn,M. (2008) Building Predictive Models in R Using the caretPackage. *Journal of Statistical Software*, **28**, 1–26.

70. Kim,D.E., Chivian,D. and Baker,D. (2004) Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Res.*, **32**, W526–31.

71. de Vries,S.J. and Bonvin,A.M.J.J. (2011) CPORT: a consensus interface predictor and its performance in prediction-driven docking with HADDOCK. *PLoS ONE*, **6**, e17695.

72. Rost,B., Radivojac,P. and Bromberg,Y. (2016) Protein function in precision medicine: deep understanding with machine learning. *FEBS Lett*, **590**, 2327–2341.

73. Shlens,J. (2014) A Tutorial on Principal Component Analysis.

74. Wolpert,D.H. (1996) The Lack of A Priori Distinctions Between Learning Algorithms. *Neural Computation*, **8**, 1341–1390.

75. van Dijk,M. and Bonvin,A.M.J.J. (2009) 3D-DART: a DNA structure modelling server. *Nucleic Acids Res.*, **37**, W235–9.

76. van Zundert,G.C.P., Rodrigues,J.P.G.L.M., Trellet,M., Schmitz,C., Kastritis,P.L., Karaca,E., Melquiond,A.S.J., van Dijk,M., de Vries,S.J. and Bonvin,A.M.J.J. (2016) The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *J. Mol. Biol.*, **428**, 720–725.

7 ANNEXES

Table SI - 1. Table SI 1: $\Delta\Delta G_{binding}$ experimental values/HS-NS classification for the residues at our dataset.

COMPLEX	CPX_PDBID	REFERENCE	MUTATION		
			RESIDUE	CHAIN	ΔΔG
Ribonuclease Inhibitor/ Angiogenin	1A4Y	[1]	TRP	261	0.10
			TRP	263	1.20
			SER	289	0.00
			TRP	318	1.50
			LYS	320	-0.30
			GLU	344	0.20
			TRP	375	1.00
			GLU	401	0.90
			TYR	434	3.30
			ASP	435	3.50
			TYR	437	0.80
			ARG	457	-0.20
			ILE	459	0.70
			ARG	5	2.30
			HIS	8	0.90
			GLN	12	0.30
			HIS	13	-0.30
			ARG	31	0.20
			ARG	32	0.90
			ASN	68	0.20
			HIS	84	0.20
			TRP	89	0.20
			GLU	108	-0.30
			HIS	114	0.65
Tissue Factor/Fab(5G9)	1AHW	[2]	TYR	156	4.00
			THR	167	0.00
			THR	170	1.00
			LEU	176	1.00
			ASP	178	-0.50
			THR	197	1.30
			VAL	198	-0.30
Barnase/barnstar	1BRS	[3]	LYS	27	5.40
			ARG	59	5.20
			GLU	60	-0.20
			GLU	73	2.80
			ARG	87	5.50
			HIS	102	6.00
			TYR	29	3.40
			ASP	35	4.50

			ASP	39	7.70
			THR	42	1.80
			GLU	76	1.30
E. Coli colicin E9 dnase domain/ cognate immunity protein IM9	1BXI	[4]	CYS	23	0.92
			ASN	24	0.14
			THR	27	0.73
			SER	28	0.17
			SER	29	0.96
			GLU	30	1.14
			LEU	33	3.42
			VAL	34	2.58
			VAL	37	1.66
			THR	38	0.90
			GLU	41	2.08
			SER	48	0.01
			GLY	49	1.49
			SER	50	2.19
			ASP	51	5.92
			TYR	55	4.63
			PRO	56	1.24
Bovine alpha-chymotrypsin/BPTI	1CBW	[5]	THR	11	0.20
			LYS	15	2.00
			ARG	17	0.50
			ILE	19	0.10
			VAL	34	0.00
			ARG	39	0.20
Factor VIIA/Tissue factor	1DAN	[6]	LYS	15	-0.40
			THR	17	0.10
			ASN	18	0.20
			LYS	20	2.60
			THR	21	-0.20
			ILE	22	0.70
			GLU	24	0.70
			LYS	41	-0.04
			SER	42	-0.05
			ASP	44	0.70
			LYS	46	0.25
			SER	47	0.05
			LYS	48	0.40
			PHE	50	0.40
			ASP	58	2.18
			LYS	68	-0.10
IgG1-kappa D1.3 Fv/E5.2 Fv	1DVF	[7]	HIS	30	1.70
			TYR	32	2.00
			TYR	49	1.70
			TYR	50	0.70
			TRP	92	0.30
			SER	93	1.20
			THR	30	0.90

			TYR	32	1.80
			TRP	52	4.20
			ASP	54	4.30
			ASN	56	1.20
			ASP	58	1.60
			GLU	98	4.20
			ARG	99	1.90
alpha-thrombin/thrombomodulin	1DX5	[8]	ILE	24	NS
			LYS	235	NS
			PHE	34	2.60
			LYS	36	NS
			PRO	37	NS
			GLN	38	NS
			GLU	39	NS
			LEU	65	NS
			ARG	67	3.4
			THR	74	NS
			ARG	75	NS
			TYR	76	3.00
			GLU	80	HS
			LYS	81	NS
			ILE	82	2.6
			MET	84	0.3
			LYS	110	0.00
HIV gp120/CD4	1GC1	[9]	SER	23	0.29
			GLN	25	0.03
			HIS	27	0.28
			LYS	29	0.59
			ASN	32	0.18
			GLN	33	0.10
			LYS	35	0.32
			GLN	40	-0.41
			SER	42	0.00
			LEU	44	1.04
			THR	45	-0.15
			ASN	52	0.70
			ARG	59	1.16
			SER	60	-0.09
			ASP	63	-0.32
			GLN	64	0.44
			GLU	85	1.31
Subtype N9 neuraminidase/Antibody NC10	1NMB	[10]	ASP	56	2.80
			TYR	99	2.13
			THR	93	0.30
IgG1-kappa D1.3 Fv/HEW lysozyme	1VFB	[11]	HIS	30	0.80
			TYR	32	1.30
			TYR	49	0.80
			TYR	50	0.40
			THR	53	-0.23

			TRP	92	2.70
			SER	93	0.30
			THR	30	0.10
			TYR	32	0.50
			TRP	52	0.40
			ARG	99	0.10
			ASP	100	3.10
			TYR	101	4.00
			ASP	18	0.30
			ASN	19	0.30
			TYR	23	0.40
			SER	24	0.80
			LYS	116	0.70
			THR	118	0.80
			ASP	119	1.00
			VAL	120	0.90
			GLN	121	2.90
			ILE	124	1.20
			ARG	125	1.80
			LEU	129	0.20
HyHEL-10/HEW Lysozyme	3HFM	[12]	SER	31	0.20
			ASP	32	2.00
			TYR	33	6.00
			TYR	50	7.50
			TYR	53	3.29
			TYR	58	1.70
			TYR	20	5.00
			ARG	21	1.00
			TRP	63	0.30
			ARG	73	-0.20
			LEU	75	1.25
			THR	89	0.00
			ASN	93	0.60
			LYS	96	7.00
			LYS	97	6.00
			SER	100	0.25
			ASP	101	1.02
			HIS	15	-0.50
			ASN	31	5.25
			ASN	32	5.20
			TYR	50	4.60
			GLN	53	1.00
			TYR	96	2.80
Protein A/Z/IgG1 MO61 Fc	1FC2	[13]	ASN	147	0.60
			ILE	150	2.20
			LYS	154	1.20
Ribonuclease A/Ribonuclease inhibitor	1DFJ	[14]	GLU	202	1.00
			TRP	257	1.30
			TRP	259	2.20

			GLU	283	1.30
			SER	285	0.80
			TRP	314	1.00
			LYS	316	1.30
			GLU	340	1.60
			GLU	397	1.30
			TYR	430	5.90
			ASP	431	3.60
			TYR	433	2.60
			ARG	453	0.80
			GLU	202	1.00
			TRP	257	1.30
Integrin alpha2 I domain/collagen	1DZI	[15]	ASN	154	NS
			TYR	157	NS
			GLN	215	HS
			ASP	219	NS
			LEU	220	NS
			THR	221	HS
			GLU	256	NS
			HIS	258	NS
BMP-2/BMP receptor IA extracellular domains	1ES7	[16]	PHE	49	NS
			PRO	50	NS
			VAL	26	NS
			TRP	31	HS
NIDOGEN-1/PERLECAN IG3	1GL4	[17]	ARG	403	NS
			ASP	427	HS
			HIS	429	HS
			TYR	431	HS
			TYR	440	NS
			GLU	616	HS
			ARG	620	HS
MazE (antidote)/ MazF (toxin)	1UB4	[18]	PHE	453	NS
			LEU	455	HS
			LEU	458	HS
IGG1 FC/ streptococcal protein G	1FCC	[19]	THR	25	0.24
			GLU	27	>4.90
			LYS	28	1.30
			LYS	31	3.50
			ASN	35	NS
			ASP	40	0.30
			GLU	42	0.40
			TRP	43	3.80
Oligomerization domain of P53	3SAK	[20]	GLU	8	NS
			PHE	10	HS
			THR	11	NS
			LEU	12	HS
			GLN	13	NS
			ILE	14	HS
			ARG	15	NS

			ARG	17	NS
			PHE	20	HS
			PHE	23	HS
			LEU	26	HS
			ASN	27	NS
			LEU	30	HS
			ASP	34	NS
Factor VIIA/Tissue factor	1FAK	[21]	ASN	37	NS
			LYS	41	NS
			SER	42	NS
			ASP	44	NS
			TYR	94	NS
			LYS	15	-0.40
			THR	17	0.10
			ASN	18	0.20
			LYS	20	2.60
			ILE	22	1.70
			GLU	24	NS
			SER	47	0.10
			LYS	48	0.40
			PHE	50	0.40
			ASP	58	2.50
			GLU	128	0.10
			LEU	133	0.10
			ARG	135	0.50
			PHE	140	1.30
			THR	203	0.10
			VAL	207	NS
Subtilisin BPN' precursor/chymotrypsin inhibitor 2	1TM1	[22]	THR	58	2.64
			MET	59	1.02
			GLU	60	2.98
			TYR	61	2.57
			ARG	62	1.25
			ARG	65	3.40
			ARG	67	2.99
			VAL	70	0.02
Interleukin-4/Interleukin-4 receptor alpha chain	1IAR	[23]	ILE	5	0.22
			THR	6	1.17
			GLN	8	-0.10
			ILE	11	-0.22
			THR	13	0.07
			ASN	15	0.97
			SER	16	-0.03
			GLU	19	-0.18
			LYS	77	-0.32
			GLN	78	0.15
			ARG	81	0.12
			PHR	82	0.48
			LYS	84	-0.90

			ARG	85	0.34
			ARG	88	0.42
			ASN	89	3.74
			TRP	91	1.55
14.3.D T cell antigen receptor/Staphylococcal enterotoxin C3	1JCK	[24]	THR	20	1.65
			TYR	26	1.77
			ASN	60	1.64
			TYR	90	2.89
			VAL	91	2.22
			LYS	103	0.67
			PHE	176	2.13
Growth factor receptor-bound protein 2/Vav proto-oncogene	1GCQ	[25]	PRO	595	0.76
			PRO	608	1.31
			PRO	609	0.12
			PRO	657	0.08
Cyclophilin A/HIV-1 capsid	1AK4	[26]	PRO	485	2.44
			VAL	486	2.35
			HIS	487	2.36
			GLY	489	3.43
			PRO	490	3.52
			ILE	491	1.60
			PRO	493	2.04
ATF-urokinase receptor	2I9B	[[27]	ARG	137	-0.29
			LYS	139	0.67
			ARG	142	0.36
			HIS	143	0.66
			ARG	145	0.41
Lyzozyme C/inhibitor	1UUZ	[28]	CYS	64	0.65
Mlc/ EIICB	3BP8	[29]	PHE	136	0.71
IMME2/ E9 DNASE	2WPT	[30]	GLU	30	1.73
			VAL	37	3.79
			GLU	41	4.48
			SER	50	2.42
			PRO	56	2.92
			ARG	54	0.87
			ASN	72	0.70
			SER	74	-0.13
			ASN	75	1.25
			SER	77	-0.46
			SER	78	-0.09
			SER	84	-0.07
			PHE	86	1.05
			THR	87	0.38
			GLN	92	0.38
			LYS	97	0.65
			VAL	98	0.26
Cytochrome C peroxidase/Cytochrome C	2PCC	[31]	ASP	34	-0.89
			VAL	197	2.09
			GLU	290	6.18

			LYS	87	0.90
JEL42 FAB/HPR	2JEL	[[32]	THR	62	0.00
			GLU	68	0.41
			GLU	70	2.72
			HIS	76	-0.41
			GLU	83	0.00
Nuclease A/inhibitor	2O3B	[33]	GLU	24	5.45
			GLN	74	3.22
			TRP	76	4.06
Profilin/beta-Actin	2BTF	[34]	PHE	59	4.27
			LYS	125	0.00
UCHL3/UbVME	1XD3	[35]	LYS	6	1.64
			LEU	8	2.10
			GLU	24	1.59
			LYS	27	0.46
			ASP	39	1.34
			ILE	44	2.47
			GLU	51	-0.24
			ASP	52	-0.06
			ASP	58	-0.41
TSG101(UEV)/ ubiquitin	1S1Q	[36]	VAL	43	0.67
			PHE	44	0.20
			ASN	45	1.23
			ASP	46	0.96
			TRP	75	0.27
			PHE	88	0.77
RALGDS/ RAS	1LFD	[37]	ARG	20	1.13
			LYS	32	1.32
			LYS	48	0.26
			ASP	51	-0.58
			LYS	52	1.17
			ASP	56	-0.28
			GLU	57	-0.25
TGF-BETA3/ TBR-2	1KTZ	[38]	ARG	25	1.48
			ARG	94	2.87
			LEU	27	2.26
			PHE	30	3.41
			ASP	32	1.96
			ASN	47	0.72
			SER	49	0.78
			ILE	50	2.33
			THR	51	1.95
			SER	52	0.66
			ILE	53	1.81
			GLU	55	1.66
			VAL	62	1.09
			GLU	75	1.52
			VAL	77	0.86
			HIS	79	0.74

			PHE	110	1.37
			MET	112	1.31
			ASP	118	1.26
			GLU	119	1.93
			ILE	125	0.98
AML1/CBF-BETA	1H9D	[39]	ARG	3	1.16
			VAL	4	1.40
			GLY	61	2.07
			GLN	67	1.36
			LEU	103	0.94
			ASN	104	2.29
Chemotaxis protein Chey/Chea	1FFW	[40]	GLU	171	0.71
			GLU	178	0.64
			HIS	181	0.03
			ASP	202	-0.07
			ASP	207	0.10
			CYS	213	0.20
			PHE	214	3.63
			ILE	216	0.43
MT-SP1/ S4 FAB	3NPS	[41]	GLN	38	0.03
			ILE	41	0.64
			ARG	87	-0.15
			PHE	94	1.59
			ASN	95	0.25
			ASP	96	1.50
			PHE	97	0.46
			THR	98	0.72
			HIS	143	1.87
			GLN	145	0.29
			TYR	146	1.77
			THR	150	0.17
			GLU	169	0.61
			GLN	177	-0.06
			GLN	175	0.74
			ASP	217	1.46
			ARG	222	-0.08
			LYS	224	-0.10
Beta-trypsin/BPTI	2FTL	[42]	GLY	12	4.37
			LYS	15	10.36
			ILE	18	5.00
			GLY	36	2.01
RNASE 1/RNASE inhibitor	1Z7X	[43]	GLU	206	1.01
			TRP	261	1.33
			TRP	263	2.20
			GLU	287	1.32
			SER	289	0.81
			TRP	318	0.99
			LYS	320	1.32
			GLU	344	1.56

			TRP	375	1.66
			GLU	401	1.30
			TYR	434	5.93
			ASP	435	3.65
			TYR	437	2.61
			ARG	457	0.84
			ILE	459	0.34
Human leukocyte elastase/OMTKY3	1PPF	[44]	LYS	13	0.75
			PRO	14	-0.12
			THR	17	3.18
			LEU	18	1.01
			GLU	19	1.20
			TYR	20	3.20
			ARG	21	0.21
			GLY	32	0.26
			ASN	36	-1.64
Proteinase B/OMTKY3	3SGB	[45]	LYS	13	-2.56
			PRO	14	-0.19
			THR	17	3.40
			LEU	18	2.96
			GLU	19	1.02
			TYR	20	1.94
			ARG	21	0.05
			GLY	32	1.29
			ASN	36	0.33
Efb-C / C3d	2GOX	[46]	ARG	131	2.25
			ASN	138	1.57
Interstitial collagenase/Metalloprotei nase inhibitor 1	2J0T	[47]	VAL	4	0.00
			SER	68	2.11
			THR	2	4.29
			MET	66	1.64
Bone morphogenetic protein 2/ Crossveinless 2	3BK3	[48]	LEU	1	0.00
			ILE	2	1.04
			ILE	18	0.49
			ILE	21	1.31
			ILE	27	1.26
Membrane-type serine protease 1/BPTI	1EAW	[49]	GLN	38	-0.52
			ILE	41	-0.82
			ILE	60	-0.19
			ASP	60A	-0.17
			ASP	60B	1.50
			ARG	60C	0.59
			PHE	60E	-0.43
			ARG	60F	0.23
			TYR	60G	-0.08
			ARG	87	-0.15
			PHE	94	0.73
			ASN	95	0.31
			ASP	96	0.65

			PHE	97	0.89
			THR	98	0.25
			HIS	143	-0.01
			GLN	145	0.31
			TYR	146	0.50
			THR	150	0.09
			LEU	153	0.50
			GLU	169	0.70
			GLN	174	0.56
			GLN	175	-0.13
			ASP	217	2.23
			GLN	221A	0.14
			ARG	222	-0.09
			LYS	224	0.48
Membrane-type serine protease 1/E2 Fab	3BN9	[50]	GLN	38	-0.42
			ILE	41	0.00
			ILE	60	0.84
			ASP	60a	0.42
			ASP	60b	0.31
			ARG	60c	-0.04
			PHE	60e	-0.04
			ARG	60f	-0.07
			TYR	60g	0.02
			ARG	87	-0.16
			PHE	94	0.64
			ASN	95	0.77
			THR	98	1.13
			HIS	143	0.09
			GLN	145	0.13
			TYR	146	1.08
			THR	150	0.29
			LEU	153	0.34
			GLU	169	0.37
			GLN	174	-0.03
			GLN	175	2.51
			ASP	217	0.57
			GLN	221a	0.71
			ARG	222	-0.09
			LYS	224	0.78
HyHEL-63 Fab/HEW Lysozyme	1DQJ	[51]	TYR	20	3.29
			ARG	21	1.21
			LYS	97	3.52
			ASP	101	1.45
			TRP	62	0.76
			TRP	63	1.35
			LEU	75	1.45
			THR	89	0.84
			ASN	93	0.65
			LYS	96	6.16

			LYS	97	3.52
			SER	100	0.78
			ASP	101	1.30
			ASN	31	2.01
			ASN	32	4.09
			TYR	50	2.68
			SER	91	1.43
			TYR	96	1.14
			ASP	32	2.01
			TYR	33	5.52
			TYR	50	6.89
			TYR	53	1.18
			TRP	98	4.93
SHV-1 beta-lactamase/BLIP	2G2U	[52]	GLU	31	0.65
			SER	35	-0.95
			PHE	36	2.76
			SER	39	-0.96
			HIS	41	1.72
			GLY	48	-0.43
			TYR	50	-2.07
			TYR	51	-0.63
			TYR	53	2.30
			SER	71	-0.51
			GLU	73	-1.98
			LYS	74	-0.22
			TRP	112	0.96
			SER	113	-0.61
			GLY	141	-0.41
			PHE	142	0.28
			TYR	143	-1.85
			ARG	144	-0.34
			HIS	148	1.12
			TRP	150	1.78
			ARG	160	0.67
			TRP	162	0.53
			SER	12	1.90
			THR	10	2.05
			ILE	13	3.51
Bovine alpha-chymotrypsin/Turkey ovomucoid third domain	1CHO	[53]	GLY	32	-0.77
			THR	17	4.32
			LEU	18	4.93

Table SI - 2. Statistical measures of ML algorithms applied to HS detection.

Scaled		METRICS								
Cluster I	***Algorithms***	**AUROC**	**Accuracy**	**TPR**	**TNR**	**PPV**	**NPV**	**FDR**	**FNR**	**F1-score**
	bagEarth	0.86	0.94	0.94	0.95	0.94	0.94	0.06	0.06	0.94
	bagEarthGCV	0.96	0.96	0.96	0.96	0.96	0.96	0.04	0.04	0.96
	bagFDA	0.88	0.90	0.91	0.89	0.89	0.91	0.11	0.09	0.90
	bagFDAGCV	0.95	0.94	0.94	0.94	0.94	0.94	0.06	0.06	0.94
	parRF	0.80	1.00	1.00	1.00	1.00	1.00	0.00	0.00	1.00
	wsrf	0.87	0.99	1.00	0.99	0.99	1.00	0.01	0.00	0.99
Cluster II										
	C5.0	0.88	0.97	0.94	1.00	1.00	0.95	0.00	0.06	0.97
	C5.0Rules	0.89	0.97	0.96	0.99	0.99	0.96	0.01	0.04	0.97
	C5.0Tree	0.91	0.97	0.94	1.00	1.00	0.95	0.00	0.06	0.97
	ctree	0.93	0.90	0.9	0.91	0.90	0.90	0.10	0.10	0.90
	evtree	0.85	0.92	0.91	0.94	0.93	0.91	0.07	0.09	0.92
	fda	0.88	0.90	0.91	0.90	0.89	0.91	0.11	0.09	0.90
	gbm	0.94	1.00	1.00	1.00	1.00	1.00	0.00	0.00	1.00
	LogitBoost	0.93	0.99	0.99	0.98	0.98	0.99	0.02	0.01	0.99
Cluster III										
	avNNet	0.84	0.97	0.97	0.97	0.97	0.97	0.03	0.03	0.97
	glm	0.76	1.00	1.00	1.00	1.00	1.00	0.00	0.00	1.00
	glmboost	0.92	0.90	0.91	0.90	0.89	0.91	0.11	0.09	0.90
	multinom	0.79	0.99	0.98	0.99	0.99	0.98	0.01	0.02	0.99
	nb	0.50	0.80	0.74	0.85	0.82	0.78	0.18	0.26	0.78

	ORFlog	0.81	0.87	0.81	0.93	0.91	0.84	0.09	0.19	0.86
	ORFpls	0.85	0.99	0.98	0.99	0.99	0.98	0.01	0.02	0.99
	ORFridge	0.85	0.98	0.98	0.99	0.99	0.98	0.01	0.02	0.98
	ORFsvm	0.88	0.99	0.99	0.99	0.99	0.99	0.01	0.01	0.99
	plr	0.90	0.98	0.98	0.98	0.98	0.98	0.02	0.02	0.98
Cluster IV										
	svmLinear	0.91	0.97	0.98	0.97	0.97	0.98	0.03	0.02	0.97
	svmLinear2	0.91	0.96	0.97	0.95	0.95	0.97	0.05	0.03	0.96
	svmPoly	0.81	0.95	0.95	0.95	0.95	0.95	0.05	0.05	0.95
	svmRadial	0.80	0.78	0.73	0.84	0.8	0.77	0.2	0.27	0.77
	svmRadialCost	0.80	0.78	0.73	0.84	0.8	0.77	0.2	0.27	0.77
	svmRadialWeights	0.80	0.75	0.88	0.62	0.68	0.85	0.32	0.12	0.77
Cluster V										
	dwdPoly	0.91	0.99	0.98	0.99	0.99	0.98	0.01	0.02	0.99
	dwdRadial	0.88	0.99	0.99	0.99	0.99	0.99	0.01	0.01	0.99
	hdda	0.80	0.79	0.72	0.85	0.81	0.77	0.19	0.28	0.77
	lda	0.90	0.93	0.91	0.96	0.95	0.92	0.05	0.09	0.93
	lda2	0.90	0.93	0.91	0.96	0.95	0.92	0.05	0.09	0.93
	pda	0.90	0.93	0.91	0.96	0.95	0.92	0.05	0.09	0.93
	stepLDA	0.85	0.82	0.85	0.79	0.79	0.85	0.21	0.15	0.82
	stepQDA	0.83	0.82	0.87	0.77	0.78	0.86	0.22	0.13	0.82

Table SI - 3. Statistical measures of ML algorithms applied to HS detection upon up-sampling of minor class.

ScaledUp		METRICS								
Cluster I	***Algorithms***	**AUROC**	**Accuracy**	**TPR**	**TNR**	**PPV**	**NPV**	**FDR**	**FNR**	**F1-score**
	bagEarth	0.88	0.96	0.94	0.97	0.97	0.94	0.94	0.06	0.96
	bagEarthGCV	0.96	0.96	0.96	0.96	0.96	0.96	0.04	0.04	0.96
	bagFDA	0.88	0.94	0.94	0.94	0.94	0.94	0.06	0.06	0.94
	bagFDAGCV	0.95	0.94	0.94	0.94	0.94	0.94	0.06	0.06	0.94
	parRF	0.81	1.00	1.00	1.00	1.00	1.00	0.00	0.00	1.00
	Ranger	0.81	1.00	1.00	0.99	0.99	1.00	0.01	0.00	1.00
	Wsrf	0.88	1.00	1.00	0.99	0.99	1.00	0.01	0.00	1.00
Cluster II										
	C5.0	0.91	1.00	1.00	1.00	1.00	1.00	0.00	0.00	1.00
	C5.0Rules	0.90	0.99	0.98	0.99	0.99	0.98	0.01	0.02	0.99
	C5.0Tree	0.93	0.99	0.98	0.99	0.99	0.98	0.01	0.02	0.99
	Ctree	0.93	0.90	0.90	0.91	0.91	0.90	0.09	0.10	0.90
	Evtree	0.86	0.92	0.93	0.92	0.92	0.93	0.08	0.07	0.92
	Fda	0.88	0.88	0.91	0.86	0.86	0.90	0.09	0.09	0.88
	Gbm	0.95	1.00	1.00	1.00	1.00	1.00	0.00	0.00	1.00
	LogitBoost	0.93	0.98	0.99	0.96	0.97	0.99	0.03	0.01	0.98
Cluster III										
	avNNet	0.84	0.97	0.97	0.96	0.96	0.97	0.04	0.03	0.97
	Glm	0.78	0.63	0.65	0.60	0.62	0.63	0.38	0.35	0.64
	glmboost	0.92	0.90	0.90	0.90	0.90	0.90	0.10	0.10	0.90
	multinom	0.80	0.98	0.99	0.98	0.98	0.99	0.02	0.01	0.98

	Nb	0.50	0.80	0.76	0.84	0.82	0.78	0.18	0.24	0.79
	ORFlog	0.83	0.88	0.84	0.92	0.91	0.85	0.09	0.16	0.87
	ORFpls	0.86	0.99	0.97	1.00	1.00	0.97	0.00	0.03	0.99
	ORFridge	0.86	0.99	0.98	1.00	1.00	0.98	0.00	0.02	0.99
	ORFsvm	0.89	0.99	1.00	0.99	0.99	1.00	0.01	0.00	0.99
	Plr	0.91	0.98	0.99	0.98	0.98	0.99	0.02	0.01	0.98
Cluster IV										
	svmLinear	0.92	0.98	0.98	0.97	0.97	0.98	0.03	0.02	0.98
	svmLinear2	0.92	0.96	0.96	0.95	0.95	0.96	0.05	0.04	0.96
	svmPoly	0.81	0.95	0.95	0.95	0.95	0.95	0.05	0.05	0.95
	svmRadial	0.80	0.78	0.76	0.81	0.8	0.77	0.2	0.24	0.78
	svmRadialCost	0.80	0.79	0.77	0.8	0.79	0.78	0.21	0.23	0.78
	svmRadialWeights	0.80	0.72	0.97	0.46	0.64	0.95	0.36	0.03	0.77
Cluster V										
	dwdRadial	0.89	0.99	0.99	0.99	0.99	0.99	0.01	0.01	0.99
	Hdda	0.78	0.79	0.73	0.85	0.83	0.76	0.17	0.27	0.78
	Lda	0.90	0.94	0.93	0.94	0.94	0.93	0.06	0.07	0.94
	lda2	0.90	0.94	0.93	0.94	0.94	0.93	0.06	0.07	0.94
	Pda	0.90	0.94	0.93	0.94	0.94	0.93	0.06	0.07	0.94
	stepLDA	0.84	0.82	0.86	0.79	0.80	0.85	0.20	0.14	0.83
	stepQDA	0.83	0.82	0.90	0.75	0.78	0.87	0.11	0.11	0.83

Table SI - 4. Statistical measures of ML algorithms applied to HS detection and down-sampling of major class.

ScaledDown		METRICS								
Cluster I	***Algorithms***	**AUROC**	**Accuracy**	**TPR**	**TNR**	**PPV**	**NPV**	**FDR**	**FNR**	**F1-score**
	bagEarth	0.88	0.95	0.94	0.96	0.96	0.95	0.06	0.06	0.95
	bagEarthGCV	0.95	0.95	0.96	0.95	0.95	0.96	0.05	0.04	0.95
	bagFDA	0.88	0.94	0.95	0.94	0.94	0.95	0.06	0.05	0.95
	bagFDAGCV	0.95	0.94	0.95	0.93	0.93	0.95	0.07	0.05	0.94
	parRF	0.79	1.00	1.00	1.00	1.00	1.00	0.00	0.00	1.00
	ranger	0.79	0.99	1.00	0.99	0.99	1.00	0.01	0.00	0.99
	wsrf	0.87	1.00	1.00	0.99	0.99	1.00	0.01	0.00	1.00
Cluster II										
	C5.0	0.88	1.00	1.00	1.00	1.00	1.00	0.00	0.00	1.00
	C5.0Rules	0.88	0.98	0.99	0.98	0.98	0.99	0.02	0.01	0.98
	C5.0Tree	0.90	0.99	0.98	0.99	0.99	0.98	0.01	0.02	0.99
	ctree	0.90	0.89	0.85	0.94	0.93	0.86	0.07	0.15	0.89
	evtree	0.84	0.88	0.86	0.91	0.90	0.86	0.10	0.14	0.88
	fda	0.88	0.91	0.92	0.91	0.91	0.92	0.09	0.08	0.92
	gbm	0.94	1.00	1.00	1.00	1.00	1.00	0.00	0.00	1.00
	LogitBoost	0.93	0.98	0.98	0.99	0.99	0.98	0.01	0.02	0.98
Cluster III										
	avNNet	0.84	0.97	0.97	0.98	0.98	0.97	0.02	0.03	0.97
	glm	0.76	1.00	1.00	1.00	1.00	1.00	0.00	0.00	1.00
	glmboost	0.92	0.90	0.91	0.90	0.90	0.91	0.10	0.09	0.90
	multinom	0.79	0.99	0.99	0.99	0.99	0.99	0.01	0.01	0.99

	nb	0.50	0.80	0.73	0.87	0.85	0.77	0.15	0.27	0.79
	ORFlog	0.80	0.87	0.81	0.93	0.92	0.83	0.08	0.19	0.86
	ORFpls	0.84	0.98	0.97	0.99	0.99	0.97	0.01	0.03	0.98
	ORFridge	0.83	0.99	0.98	1.00	1.00	0.98	0.00	0.02	0.99
	ORFsvm	0.88	0.99	1.00	0.99	0.99	1.00	0.01	0.00	0.99
	plr	0.89	0.99	0.99	0.99	0.99	0.99	0.01	0.01	0.99
Cluster IV										
	svmLinear	0.91	0.98	0.98	0.97	0.97	0.98	0.03	0.02	0.98
	svmLinear2	0.91	0.96	0.97	0.96	0.96	0.97	0.04	0.03	0.96
	svmPoly	0.79	0.95	0.95	0.94	0.95	0.95	0.05	0.05	0.95
	svmRadial	0.79	0.77	0.78	0.76	0.77	0.78	0.23	0.22	0.78
	svmRadialCost	0.79	0.77	0.78	0.77	0.77	0.78	0.23	0.22	0.77
	svmRadialWeights	0.79	0.75	0.90	0.61	0.70	0.85	0.30	0.10	0.78
Cluster V										
	dwdPoly	0.90	0.99	0.99	0.99	0.99	0.99	0.01	0.01	0.99
	dwdRadial	0.88	0.99	0.99	0.99	0.99	0.99	0.01	0.01	0.99
	hdda	0.81	0.78	0.73	0.84	0.82	0.76	0.18	0.27	0.77
	lda	0.90	0.94	0.92	0.95	0.95	0.92	0.05	0.08	0.94
	lda2	0.89	0.94	0.92	0.95	0.95	0.92	0.05	0.08	0.94
	pda	0.90	0.93	0.92	0.95	0.95	0.92	0.05	0.08	0.93
	stepLDA	0.85	0.83	0.86	0.79	0.80	0.85	0.20	0.14	0.83
	stepQDA	0.84	0.81	0.91	0.72	0.76	0.89	0.24	0.09	0.83

Table SI - 5. Statistical measures of ML algorithms applied to HS detection upon PCA.

PCA		METRICS								
Cluster I	***Algorithms***	**AUROC**	**Accuracy**	**TPR**	**TNR**	**PPV**	**NPV**	**FDR**	**FNR**	**F1-score**
	bagEarth	0.72	0.82	0.76	0.87	0.85	0.79	0.15	0.24	0.80
	bagEarthGCV	0.83	0.81	0.79	0.83	0.81	0.81	0.19	0.21	0.80
	bagFDA	0.72	0.82	0.76	0.88	0.85	0.79	0.15	0.24	0.80
	bagFDAGCV	0.82	0.82	0.77	0.88	0.85	0.80	0.15	0.23	0.81
	parRF	0.82	1.00	1.00	1.00	1.00	1.00	0.00	0.00	1.00
	ranger	0.23	0.01	0.02	0.01	0.02	0.01	0.98	0.98	0.02
	wsrf	0.82	0.99	1.00	0.99	0.99	1.00	0.01	0.00	0.99
Cluster II										
	ada	0.77	0.95	0.92	0.97	0.97	0.93	0.03	0.08	0.94
	adaboost	0.83	1.00	1.00	1.00	1.00	1.00	0.00	0.00	1.00
	C5.0	0.72	1.00	1.00	1.00	1.00	1.00	0.00	0.00	1.00
	C5.0Rules	0.72	0.90	0.90	0.91	0.91	0.90	0.09	0.10	0.90
	C5.0Tree	0.73	0.97	0.97	0.97	0.97	0.97	0.03	0.03	0.97
	ctree	0.72	0.81	0.77	0.85	0.83	0.80	0.17	0.23	0.80
	evtree	0.72	0.83	0.77	0.88	0.86	0.80	0.14	0.23	0.81
	fda	0.71	0.79	0.78	0.80	0.79	0.80	0.21	0.22	0.79
	gamboost	0.80	0.80	0.73	0.87	0.84	0.78	0.16	0.27	0.78
	gbm	0.78	0.99	0.98	0.99	0.99	0.98	0.01	0.02	0.99
	J48	0.66	0.98	0.98	0.97	0.97	0.98	0.03	0.02	0.98
	LogitBoost	0.72	0.87	0.86	0.88	0.87	0.87	0.13	0.14	0.86

Cluster III										
	avNNet	0.79	0.94	0.91	0.97	0.96	0.92	0.04	0.09	0.94
	glm	0.74	0.84	0.82	0.86	0.85	0.84	0.15	0.18	0.83
	glmboost	0.78	0.77	0.76	0.79	0.77	0.78	0.23	0.24	0.76
	multinom	0.82	0.84	0.82	0.86	0.85	0.84	0.15	0.18	0.83
	nb	0.72	0.82	0.74	0.90	0.88	0.79	0.12	0.26	0.80
	ORFlog	0.84	1.00	1.00	1.00	1.00	1.00	0.00	0.00	1.00
	ORFpls	0.83	0.99	0.99	1.00	1.00	0.99	0.00	0.01	0.99
	ORFridge	0.84	1.00	0.99	1.00	1.00	0.99	0.00	0.01	1.00
	ORFsvm	0.83	0.99	0.99	0.99	0.99	0.99	0.01	0.01	0.99
	plr	0.82	0.84	0.81	0.86	0.84	0.83	0.16	0.19	0.83
Cluster IV										
	svmLinear	0.81	0.82	0.75	0.90	0.87	0.79	0.13	0.25	0.80
	svmLinear2	0.80	0.82	0.77	0.86	0.84	0.80	0.16	0.23	0.80
	svmPoly	0.81	0.89	0.86	0.92	0.91	0.88	0.09	0.14	0.89
	svmRadial	0.80	0.77	0.70	0.84	0.80	0.75	0.20	0.30	0.74
	svmRadialCost	0.80	0.77	0.69	0.84	0.80	0.74	0.20	0.31	0.74
	svmRadialWeights	0.80	0.75	0.90	0.61	0.68	0.86	0.32	0.10	0.78
Cluster V										
	amdai	0.81	0.80	0.76	0.84	0.82	0.79	0.21	0.24	0.79
	dwdPoly	0.81	0.87	0.82	0.91	0.89	0.85	0.11	0.18	0.86
	dwdRadial	0.80	1.00	1.00	1.00	1.00	1.00	0.00	0.00	1.00
	hdda	0.79	0.77	0.72	0.82	0.79	0.76	0.21	0.28	0.75
	lda	0.81	0.80	0.76	0.84	0.82	0.79	0.18	0.24	0.79
	lda2	0.81	0.80	0.76	0.84	0.82	0.79	0.18	0.24	0.79

	loclda	0.69	0.94	0.92	0.96	0.96	0.93	0.04	0.08	0.94
	pda	0.81	0.80	0.76	0.84	0.82	0.79	0.18	0.24	0.79
	qda	0.76	0.79	0.61	0.95	0.92	0.73	0.08	0.39	0.74
	rda	0.77	0.80	0.76	0.84	0.82	0.79	0.18	0.24	0.79
	stepLDA	0.70	0.68	0.51	0.84	0.74	0.65	0.26	0.49	0.60f
	stepQDA	0.71	0.67	0.43	0.90	0.79	0.63	0.21	0.57	0.55

Table SI - 6. Statistical measures of ML algorithms applied to HS detection upon PCA and up-sampling of minor class.

PCAUp		METRICS								
Cluster I	***Algorithms***	**AUROC**	**Accuracy**	**TPR**	**TNR**	**PPV**	**NPV**	**FDR**	**FNR**	**F1-score**
	bagEarth	0.72	0.81	0.80	0.82	0.82	0.81	0.18	0.20	0.81
	bagEarthGCV	0.83	0.82	0.80	0.84	0.83	0.81	0.17	0.20	0.82
	bagFDA	0.72	0.81	0.80	0.81	0.81	0.80	0.19	0.20	0.81
	bagFDAGCV	0.83	0.82	0.80	0.84	0.83	0.81	0.17	0.20	0.81
	parRF	0.84	1.00	1.00	1.00	1.00	1.00	0.00	0.00	1.00
	Ranger	0.84	0.99	0.99	0.99	0.99	0.99	0.01	0.01	0.99
	Wsrf	0.84	0.99	1.00	0.99	0.99	1.00	0.01	0.00	0.99
Cluster II										
	Ada	0.78	0.96	0.94	0.97	0.97	0.94	0.03	0.06	0.96
	adaboost	0.85	1.00	1.00	1.00	1.00	1.00	0.00	0.00	1.00
	C5.0	0.73	1.00	1.00	1.00	1.00	1.00	0.00	0.00	1.00
	C5.0Rules	0.74	0.87	0.93	0.82	0.84	0.92	0.16	0.07	0.88
	C5.0Tree	0.75	0.91	0.95	0.87	0.88	0.95	0.12	0.05	0.91
	Ctree	0.77	0.87	0.86	0.88	0.87	0.86	0.13	0.14	0.87
	Evtree	0.72	0.83	0.81	0.85	0.84	0.82	0.16	0.19	0.83
	Fda	0.62	0.77	0.75	0.80	0.79	0.76	0.21	0.25	0.77
	gamboost	0.80	0.78	0.76	0.81	0.80	0.77	0.20	0.24	0.78
	Gbm	0.81	0.98	0.98	0.98	0.98	0.98	0.02	0.02	0.98
	J48	0.70	0.97	0.97	0.98	0.98	0.97	0.02	0.03	0.97
	LogitBoost	0.75	0.89	0.93	0.85	0.86	0.93	0.14	0.07	0.89

Cluster III										
	avNNet	0.80	0.96	0.94	0.97	0.97	0.94	0.03	0.06	0.96
	Glm	0.73	0.77	0.76	0.78	0.77	0.76	0.23	0.24	0.77
	glmboost	0.79	0.78	0.78	0.77	0.77	0.78	0.23	0.22	0.78
	multinom	0.83	0.83	0.81	0.85	0.84	0.82	0.16	0.19	0.83
	Nb	0.72	0.82	0.74	0.89	0.87	0.78	0.13	0.26	0.80
	ORFlog	0.86	1.00	1.00	1.00	1.00	1.00	0.00	0.00	1.00
	ORFpls	0.85	1.00	0.99	1.00	1.00	0.99	0.00	0.01	1.00
	ORFridge	0.85	1.00	0.99	1.00	1.00	0.99	0.00	0.01	1.00
	ORFsvm	0.85	0.99	0.99	1.00	1.00	0.99	0.00	0.01	0.99
	Plr	0.83	0.83	0.81	0.85	0.84	0.82	0.16	0.19	0.83
Cluster IV										
	svmLinear	0.83	0.82	0.80	0.85	0.84	0.81	0.16	0.20	0.82
	svmLinear2	0.82	0.82	0.78	0.87	0.85	0.80	0.15	0.22	0.81
	svmPoly	0.82	0.94	0.95	0.93	0.93	0.95	0.07	0.05	0.94
	svmRadial	0.81	0.78	0.75	0.80	0.79	0.76	0.21	0.25	0.77
	svmRadialCost	0.80	0.79	0.78	0.80	0.80	0.78	0.20	0.22	0.79
	svmRadialWeights	0.80	0.77	0.92	0.61	0.70	0.89	0.30	0.08	0.80
Cluster V										
	Amdai	0.82	0.8	0.77	0.83	0.82	0.78	0.18	0.23	0.79
	dwdPoly	0.82	0.81	0.78	0.84	0.83	0.79	0.17	0.22	0.80
	dwdRadial	0.82	1.00	1.00	1.00	1.00	1.00	0.00	0.00	1.00
	Hdda	0.79	0.74	0.76	0.72	0.73	0.75	0.27	0.24	0.75
	Lda	0.82	0.80	0.77	0.83	0.82	0.78	0.18	0.23	0.79
	lda2	0.83	0.80	0.77	0.83	0.82	0.78	0.18	0.23	0.79

	Loclda	0.72	0.94	0.93	0.95	0.95	0.93	0.05	0.07	0.94
	Pda	0.82	0.80	0.77	0.83	0.82	0.78	0.18	0.23	0.79
	Qda	0.78	0.79	0.63	0.95	0.92	0.72	0.08	0.37	0.75
	Rda	0.77	0.80	0.77	0.83	0.82	0.78	0.18	0.23	0.79
	stepLDA	0.66	0.67	0.52	0.82	0.74	0.63	0.26	0.48	0.61
	stepQDA	0.69	0.69	0.57	0.81	0.76	0.66	0.24	0.43	0.65

Table SI - 7. Statistical measures of ML algorithms applied to HS detection upon PCA and down-sampling of major class.

PCADown		METRICS								
Cluster I	***Algorithms***	**AUROC**	**Accuracy**	**TPR**	**TNR**	**PPV**	**NPV**	**FDR**	**FNR**	**F1-score**
	bagEarth	0.72	0.82	0.81	0.83	0.83	0.81	0.17	0.19	0.82
	bagEarthGCV	0.81	0.81	0.8	0.82	0.81	0.80	0.19	0.20	0.80
	bagFDA	0.71	0.81	0.78	0.84	0.83	0.79	0.17	0.22	0.80
	bagFDAGCV	0.81	0.81	0.79	0.83	0.83	0.8	0.17	0.21	0.81
	parRF	0.82	1.00	1.00	1.00	1.00	1.00	0.00	0.00	1.00
	ranger	0.23	0.01	0.02	0.01	0.02	0.01	0.98	0.98	0.01
	wsrf	0.81	0.99	1.00	0.99	0.99	1.00	0.01	0.00	0.99
Cluster II	***Cluster II***									
	ada	0.74	0.97	0.96	0.98	0.98	0.96	0.02	0.04	0.97
	adaboost	0.82	1.00	1.00	1.00	1.00	1.00	0.00	0.00	1.00
	C5.0	0.70	1.00	1.00	1.00	1.00	1.00	0.00	0.00	1.00
	C5.0Rules	0.70	0.94	0.91	0.98	0.98	0.91	0.02	0.09	0.94
	C5.0Tree	0.71	0.99	0.99	0.98	0.98	0.99	0.02	0.01	0.99
	ctree	0.74	0.82	0.82	0.82	0.82	0.82	0.18	0.18	0.82
	evtree	0.68	0.73	0.77	0.69	0.71	0.75	0.29	0.23	0.74
	fda	0.59	0.77	0.74	0.80	0.78	0.75	0.22	0.26	0.76
	gamboost	0.78	0.78	0.76	0.81	0.8	0.77	0.20	0.24	0.78
	gbm	0.77	0.98	0.98	0.98	0.98	0.98	0.02	0.02	0.98
	J48	0.69	0.93	0.88	0.99	0.99	0.89	0.01	0.12	0.93
	LogitBoost	0.71	0.88	0.82	0.95	0.94	0.84	0.06	0.18	0.88

Cluster III	***Cluster III***									
	avNNet	0.78	0.94	0.94	0.94	0.94	0.94	0.06	0.06	0.94
	glm	0.82	0.84	0.83	0.84	0.84	0.84	0.16	0.17	0.84
	glmboost	0.78	0.77	0.76	0.79	0.77	0.78	0.23	0.24	0.76
	multinom	0.82	0.84	0.82	0.86	0.86	0.83	0.14	0.18	0.84
	nb	0.72	0.82	0.74	0.90	0.88	0.79	0.12	0.26	0.80
	ORFlog	0.82	1.00	1.00	1.00	1.00	1.00	0.00	0.00	1.00
	ORFpls	0.82	0.99	0.99	1.00	1.00	0.99	0.00	0.01	0.99
	ORFridge	0.83	0.99	0.99	1.00	1.00	0.99	0.00	0.01	0.99
	ORFsvm	0.82	0.99	0.99	0.99	0.99	0.99	0.01	0.01	0.99
	plr	0.82	0.84	0.81	0.86	0.84	0.83	0.16	0.19	0.82
Cluster IV	***Cluster IV***									
	svmLinear	0.81	0.82	0.75	0.90	0.87	0.79	0.13	0.25	0.80
	svmLinear2	0.80	0.82	0.77	0.86	0.84	0.80	0.16	0.23	0.80
	svmPoly	0.81	0.89	0.86	0.92	0.91	0.88	0.09	0.14	0.89
	svmRadial	0.80	0.77	0.70	0.84	0.80	0.75	0.20	0.30	0.74
	svmRadialCost	0.80	0.77	0.69	0.84	0.80	0.74	0.20	0.31	0.74
	svmRadialWeights	0.80	0.75	0.90	0.61	0.68	0.86	0.32	0.10	0.78
Cluster V	***Cluster V***									
	amdai	0.80	0.80	0.77	0.82	0.81	0.78	0.19	0.23	0.79
	dwdPoly	0.80	0.87	0.83	0.91	0.90	0.84	0.10	0.17	0.86
	dwdRadial	0.78	1.00	1.00	1.00	1.00	1.00	0.00	0.00	1.00
	hdda	0.79	0.78	0.75	0.8	0.79	0.76	0.21	0.25	0.77
	lda	0.80	0.8	0.77	0.82	0.81	0.78	0.19	0.23	0.79
	lda2	0.81	0.8	0.77	0.82	0.81	0.78	0.19	0.23	0.79

	loclda	0.75	0.94	0.92	0.96	0.96	0.92	0.04	0.08	0.94
	pda	0.80	0.80	0.77	0.82	0.81	0.78	0.19	0.23	0.79
	qda	0.76	0.80	0.63	0.96	0.94	0.72	0.06	0.37	0.75
	rda	0.76	0.80	0.77	0.82	0.81	0.78	0.19	0.23	0.79
	stepLDA	0.67	0.66	0.49	0.82	0.74	0.62	0.26	0.51	0.59
	stepQDA	0.68	0.69	0.65	0.73	0.71	0.68	0.29	0.35	0.68

7.1 SUPPLEMENTARY INFORMATION REFERENCES

1. Papageorgiou, A.C.; Shapiro, R.; Acharya, K.R. Molecular recognition of human angiogenin by placental ribonuclease inhibitor - an x-ray crystallographic study at 2.0 angstrom resolution. Embo Journal 1997, 16, 5162-5177.

2. Huang, M.; Syed, R.; Stura, E.A.; Stone, M.J.; Stefanko, R.S.; Ruf, W.; Edgington, T.S.; Wilson, I.A. The mechanism of an inhibitory antibody on tf-initiated blood coagulation revealed by the crystal structures of human tissue factor, fab 5g9 and tf·5g9 complex1. Journal of Molecular Biology 1998, 275, 873-894.

3. Buckle, A.M.; Schreiber, G.; Fersht, A.R. Protein-protein recognition: Crystal structural analysis of a barnase-barstar complex at 2.0-.Ang. Resolution. Biochemistry 1994, 33, 8878-8889.

4. Kuhlmann, U.C. Crystal structure of the e.Coli colicin e9 dnase domain with its cognate immunity protein im9. 1998.

5. Scheidig, A.J.; Hynes, T.R.; Pelletier, L.A.; Wells, J.A.; Kossiakoff, A.A. Crystal structures of bovine chymotrypsin and trypsin complexed to the inhibitor domain of alzheimer's amyloid beta-protein precursor (appi) and basic pancreatic trypsin inhibitor (bpti): Engineering of inhibitors with altered specificities. Protein Science : A Publication of the Protein Society 1997, 6, 1806-1824.

6. Banner, D.W.; D'Arcy, A.; Chène, C.; Winkler, F.K.; Guha, A.; Konigsberg, W.H.; Nemerson, Y.; Kirchhofer, D. The crystal structure of the complex of blood coagulation factor viia with soluble tissue factor. Nature 1996, 380, 41-46.

7. Braden, B.C.; Fields, B.A.; Ysern, X.; Dall'Acqua, W.; Goldbaum, F.A.; Poljak, R.J.; Mariuzza, R.A. Crystal structure of an fv–fv idiotope–anti-idiotope complex at 1.9 å resolution. Journal of Molecular Biology 1996, 264, 137-151.

8. Fuentes-Prior, P.; Iwanaga, Y.; Huber, R.; Pagila, R.; Rumennik, G.; Seto, M.; Morser, J.; Light, D.R.; Bode, W. Structural basis for the anticoagulant activity of the thrombin-thrombomodulin complex. Nature 2000, 404, 518-525.

9. Kwong, P.D.; Wyatt, R.; Robinson, J.; Sweet, R.W.; Sodroski, J.; Hendrickson, W.A. Structure of an hiv gp120 envelope glycoprotein in complex with the cd4 receptor and a neutralizing human antibody. Nature 1998, 393, 648-659.

10. Malby, R.L.; Tulip, W.R.; Harley, V.R.; McKimm-Breschkin, J.L.; Laver, W.G.; Webster, R.G.; Colman, P.M. The structure of a complex between the nc10 antibody and influenza virus neuraminidase and comparison with the overlapping binding site of the nc41 antibody. Structure 1994, 2, 733-746.

11. Bhat, T.N.; Bentley, G.A.; Boulot, G.; Greene, M.I.; Tello, D.; Dall'Acqua, W.; Souchon, H.; Schwarz, F.P.; Mariuzza, R.A.; Poljak, R.J. Bound water molecules and conformational stabilization help mediate an antigen-antibody association. Proceedings of the National Academy of Sciences of the United States of America 1994, 91, 1089-1093.

12. Padlan, E.A.; Silverton, E.W.; Sheriff, S.; Cohen, G.H.; Smithgill, S.J.; Davies, D.R. Structure of an antibody antigen complex - crystal-structure of the hyhel-10 fab-lysozyme complex. Proceedings of the National Academy of Sciences of the United States of America 1989, 86, 5938-5942.

13. Deisenhofer, J. Crystallographic refinement and atomic models of a human fc fragment and its complex with fragment-b of protein-a from staphylococcus-aureus at 2.9-a and 2.8-a resolution. Biochemistry 1981, 20, 2361-2370.

14. Kobe, B.; Deisenhofer, J. A structural basis of the interactions between leucine-rich repeats and protein ligands. Nature 1995, 374, 183-186.

15. Emsley, J.; Knight, C.G.; Farndale, R.W.; Barnes, M.J.; Liddington, R.C. Structural basis of collagen recognition by integrin alpha 2 beta 1. Cell 2000, 101, 47-56.

16. Kirsch, T.; Sebald, W.; Dreyer, M.K. Crystal structure of the bmp-2-bria ectodomain complex. Nature Structural Biology 2000, 7, 492-496.

17. Kvansakul, M.; Hopf, M.; Ries, A.; Timpl, R.; Hohenester, E. Structural basis for the high-affinity interaction of nidogen-1 with immunoglobulin-like domain 3 of perlecan. Embo Journal 2001, 20, 5342-5346.

18. Kamada, K.; Hanaoka, F.; Burley, S.K. Crystal structure of the maze/mazf complex: Molecular bases of antidote-toxin recognition. Molecular Cell 2003, 11, 875-884.

19. Sauererikssson, A.E.; Kleywegt, G.J.; Uhl, M.; Jones, T.A. Crystal-structure of the c2 fragment of streptococcal protein-g in complex with the fc domain of human-igg. Structure 1995, 3, 265-278.

20. Kuszewski, J.; Gronenborn, A.M.; Clore, G.M. Improving the packing and accuracy of nmr structures with a pseudopotential for the radius of gyration. Journal of the American Chemical Society 1999, 121, 2337-2338.

21. Zhang, E.; St Charles, R.; Tulinsky, A. Structure of extracellular tissue factor complexed with factor viia inhibited with a bpti mutant. Journal of Molecular Biology 1999, 285, 2089-2104.

22. Radisky, E.S.; Kwan, G.; Lu, C.J.K.; Koshland, D.E. Binding, proteolytic, and crystallographic analyses of mutations at the protease-inhibitor interface of the subtilisin bpn '/chymotrypsin inhibitor 2 complex. Biochemistry 2004, 43, 13648-13656.

23. Hage, T.; Sebald, W.; Reinemer, P. Crystal structure of the interleukin-4/receptor alpha chain complex reveals a mosaic binding interface. Cell 1999, 97, 271-281.

24. Fields, B.A.; Malchiodi, E.L.; Li, H.M.; Ysern, X.; Stauffacher, C.V.; Schlievert, P.M.; Karjalainen, K.; Mariuzza, R.A. Crystal structure of a t-cell receptor beta-chain complexed with a superantigen. Nature 1996, 384, 188-192.

25. Nishida, M.; Nagata, K.; Hachimori, Y.; Horiuchi, M.; Ogura, K.; Mandiyan, V.; Schlessinger, J.; Inagaki, F. Novel recognition mode between vav and grb2 sh3 domains. Embo Journal 2001, 20, 2995-3007.

26. Gamble, T.R.; Vajdos, F.F.; Yoo, S.H.; Worthylake, D.K.; Houseweart, M.; Sundquist, W.I.; Hill, C.P. Crystal structure of human cyclophilin a bound to the amino-terminal domain of hiv-1 capsid. Cell 1996, 87, 1285-1294.

27. Barinka, C.; Parry, G.; Callahan, J.; Shaw, D.E.; Kuo, A.; Bdeir, K.; Cines, D.B.; Mazar, A.; Lubkowski, J. Structural basis of interaction between urokinase-type plasminogen activator and its receptor. Journal of Molecular Biology 2006, 363, 482-495.

28. Abergel, C.; Monchois, V.; Byrne, D.; Chenivesse, S.; Lembo, F.; Lazzaroni, J.-C.; Claverie, J.-M. Structure and evolution of the ivy protein family, unexpected lysozyme inhibitors in gram-negative bacteria. Proceedings of the National Academy of Sciences of the United States of America 2007, 104, 6394-6399.

29. Nam, T.-W.; Il Jung, H.; An, Y.J.; Park, Y.-H.; Lee, S.H.; Seok, Y.-J.; Cha, S.-S. Analyses of mlc-iibglc interaction and a plausible molecular mechanism of mlc inactivation by membrane sequestration. Proceedings of the National Academy of Sciences of the United States of America 2008, 105, 3751-3756.

30. Meenan, N.A.G.; Sharma, A.; Fleishman, S.J.; MacDonald, C.J.; Morel, B.; Boetzel, R.; Moore, G.R.; Baker, D.; Kleanthous, C. The structural and energetic basis for high selectivity in a high-affinity protein-protein interaction. Proceedings of the National Academy of Sciences of the United States of America 2010, 107, 10080-10085.

31. Pelletier, H.; Kraut, J. Crystal-structure of a complex between electron-transfer partners, cytochrome-c peroxidase and cytochrome-c. Science 1992, 258, 1748-1755.

32. Prasad, L.; Waygood, E.B.; Lee, J.S.; Delbaere, L.T.J. The 2.5 angstrom resolution structure of the jei42 fab fragment hpr complex. Journal of Molecular Biology 1998, 280, 829-845.

33. Ghosh, M.; Meiss, G.; Pingoud, A.M.; London, R.E.; Pedersen, L.C. The nuclease a-inhibitor complex is characterized by a novel metal ion bridge. Journal of Biological Chemistry 2007, 282, 5682-5690.

34. Schutt, C.E.; Myslik, J.C.; Rozycki, M.D.; Goonesekere, N.C.W.; Lindberg, U. The structure of crystalline profilin beta-actin. Nature 1993, 365, 810-816.

35. Misaghi, S.; Galardy, P.J.; Meester, W.J.N.; Ovaa, H.; Ploegh, H.L.; Gaudet, R. Structure of the ubiquitin hydrolase uch-l3 complexed with a suicide substrate. Journal of Biological Chemistry 2005, 280, 1512-1520.

36. Sundquist, W.I.; Schubert, H.L.; Kelly, B.N.; Hill, G.C.; Holton, J.M.; Hill, C.P. Ubiquitin recognition by the human tsg101 protein. Molecular Cell 2004, 13, 783-789.

37. Huang, L.; Hofer, F.; Martin, G.S.; Kim, S.H. Structural basis for the interaction of ras with raigds. Nature Structural Biology 1998, 5, 422-426.

38. Hart, P.J.; Deep, S.; Taylor, A.B.; Shu, Z.Y.; Hinck, C.S.; Hinck, A.P. Crystal structure of the human t beta r2 ectodomain-tgf-beta 3 complex. Nature Structural Biology 2002, 9, 203-208.

39. Bravo, J.; Li, Z.; Speck, N.A.; Warren, A.J. The leukemia-associated aml1 (runx1)-cbf[beta] complex functions as a DNA-induced molecular clamp. Nat Struct Mol Biol 2001, 8, 371-378.

40. Gouet, P.; Chinardet, N.; Welch, M.; Guillet, V.; Cabantous, S.; Birck, C.; Mourey, L.; Samama, J.P. Further insights into the mechanism of function of the response

regulator chey from crystallographic studies of the chey-chea(124-257) complex. Acta Crystallographica Section D-Biological Crystallography 2001, 57, 44-51.

41. Schneider, E.L.; Lee, M.S.; Baharuddin, A.; Goetz, D.H.; Farady, C.J.; Ward, M.; Wang, C.-I.; Craik, C.S. A reverse binding motif that contributes to specific protease inhibition by antibodies. Journal of Molecular Biology 2012, 415, 699-715.

42. Hanson, W.M.; Domek, G.J.; Horvath, M.P.; Goldenberg, D.P. Rigidification of a flexible protease inhibitor variant upon binding to trypsin. Journal of Molecular Biology 2007, 366, 230-243.

43. Johnson, R.J.; McCoy, J.G.; Bingman, C.A.; Phillips, G.N., Jr.; Raines, R.T. Inhibition of human pancreatic ribonuclease by the human ribonuclease inhibitor protein. Journal of Molecular Biology 2007, 368, 434-449.

44. Bode, W.; Wei, A.Z.; Huber, R.; Meyer, E.; Travis, J.; Neumann, S. X-ray crystal-structure of the complex of human-leukocyte elastase (pmn elastase) and the 3rd domain of the turkey ovomucoid inhibitor. Embo Journal 1986, 5, 2453-2458.

45. Read, R.J.; Fujinaga, M.; Sielecki, A.R.; James, M.N.G. Structure of the complex of streptomyces-griseus protease-b and the 3rd domain of the turkey ovomucoid inhibitor at 1.8-a resolution. Biochemistry 1983, 22, 4420-4433.

46. Hammel, M.; Sfyroera, G.; Ricklin, D.; Magotti, P.; Lambris, J.D.; Geisbrecht, B.V. A structural basis for complement inhibition by staphylococcus aureus. Nat Immunol 2007, 8, 430-437.

47. Iyer, S.; Wei, S.; Brew, K.; Acharya, K.R. Crystal structure of the catalytic domain of matrix metalloproteinase-1 in complex with the inhibitory domain of tissue inhibitor of metalloproteinase-1. Journal of Biological Chemistry 2007, 282, 364-371.

48. Zhang, J.-l.; Qiu, L.-y.; Kotzsch, A.; Weidauer, S.; Patterson, L.; Hammerschmidt, M.; Sebald, W.; Mueller, T.D. Crystal structure analysis reveals how the chordin family member crossveinless 2 blocks bmp-2 receptor binding. Developmental Cell 2008, 14, 739-750.

49. Friedrich, R.; Fuentes-Prior, P.; Ong, E.; Coombs, G.; Hunter, M.; Oehler, R.; Pierson, D.; Gonzalez, R.; Huber, R.; Bode, W., et al. Catalytic domain structures of mt-sp1/matriptase, a matrix-degrading transmembrane serine proteinase. Journal of Biological Chemistry 2002, 277, 2160-2168.

50. Farady, C.J.; Egea, P.F.; Schneider, E.L.; Darragh, M.R.; Craik, C.S. Structure of an fab-protease complex reveals a highly specific non-canonical mechanism of inhibition. Journal of Molecular Biology 2008, 380, 351-360.

51. Li, Y.L.; Li, H.M.; Smith-Gill, S.J.; Mariuzza, R.A. Three-dimensional structures of the free and antigen-bound fab from monoclonal antilysozyme antibody hyhel-63. Biochemistry 2000, 39, 6296-6309.

52. Reynolds, K.A.; Thomson, J.M.; Corbett, K.D.; Bethel, C.R.; Berger, J.M.; Kirsch, J.F.; Bonomo, R.A.; Handel, T.M. Structural and computational characterization of the shv-1 beta-lactamase-beta- lactamase inhibitor protein interface. Journal of Biological Chemistry 2006, 281, 26745-26753.

53. Fujinaga, M.; Sielecki, A.R.; Read, R.J.; Ardelt, W.; Laskowski, M.; James, M.N.G. Crystal and molecular-structures of the complex of alpha-chymotrypsin with its inhibitor turkey ovomucoid 3rd domain at 1.8 a resolution. Journal of Molecular Biology 1987, 195, 397-418.

8 PAPERS

8.1 PAPER NUMBER 1

Melo,R., Fieldhouse,R., Melo,A., Correia,J.D.G., Cordeiro,M.N.D.S., Gümüş,Z.H., Costa,J., Bonvin,A.M.J.J. and Moreira,I.S. (2016) A Machine Learning Approach for Hot-Spot Detection at Protein-Protein Interfaces. IJMS, 17, 1215.

International Journal of
Molecular Sciences

MDPI

Article

A Machine Learning Approach for Hot-Spot Detection at Protein-Protein Interfaces

Rita Melo [1,2], Robert Fieldhouse [3], André Melo [4], João D. G. Correia [1], Maria Natália D. S. Cordeiro [4], Zeynep H. Gümüş [3], Joaquim Costa [5], Alexandre M. J. J. Bonvin [6] and Irina S. Moreira [2,6,*]

1 Centro de Ciências e Tecnologias Nucleares, Instituto Superior Técnico, Universidade de Lisboa, Estrada Nacional 10 (ao km 139,7), 2695-066 Bobadela LRS, Portugal; ritamelo@ctn.ist.utl.pt (R.M.); jgalamba@ctn.tecnico.ulisboa.pt (J.D.G.C.)

2 CNC—Center for Neuroscience and Cell Biology; Rua Larga, Faculdade de Medicina, Polo I, 1ºandar, Universidade de Coimbra, 3004-504 Coimbra, Portugal

3 Department of Genetics and Genomics and Icahn Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA; robert.fieldhouse@mssm.edu (R.F.); zeynep.gumus@gmail.com (Z.H.G.)

4 REQUIMTE (Rede de Química e Tecnologia), Faculdade de Ciências da Universidade do Porto, Departamento de Química e Bioquímica, Rua do Campo Alegre, 4169-007 Porto, Portugal; asmelo@fc.up.pt (A.M.); ncordeir@fc.up.pt (M.N.D.S.C.)

5 CMUP/FCUP, Centro de Matemática da Universidade do Porto, Faculdade de Ciências, Rua do Campo Alegre, 4169-007 Porto, Portugal; Jpcosta@fc.up.pt

6 Bijvoet Center for Biomolecular Research, Faculty of Science—Chemistry, Utrecht University, Utrecht 3584CH, The Netherlands; a.m.j.j.bonvin@uu.nl

* Correspondence: irina.moreira@cnc.uc.pt; Tel.: +351-239-820-190

Academic Editor: Humberto González-Díaz
Received: 24 May 2016; Accepted: 18 July 2016; Published: 27 July 2016

Abstract: Understanding protein-protein interactions is a key challenge in biochemistry. In this work, we describe a more accurate methodology to predict Hot-Spots (HS) in protein-protein interfaces from their native complex structure compared to previous published Machine Learning (ML) techniques. Our model is trained on a large number of complexes and on a significantly larger number of different structural- and evolutionary sequence-based features. In particular, we added interface size, type of interaction between residues at the interface of the complex, number of different types of residues at the interface and the Position-Specific Scoring Matrix (PSSM), for a total of 79 features. We used twenty-seven algorithms from a simple linear-based function to support-vector machine models with different cost functions. The best model was achieved by the use of the conditional inference random forest (c-forest) algorithm with a dataset pre-processed by the normalization of features and with up-sampling of the minor class. The method has an overall accuracy of 0.80, an *F1*-score of 0.73, a sensitivity of 0.76 and a specificity of 0.82 for the independent test set.

Keywords: protein-protein interfaces; hot-spots; machine learning; Solvent Accessible Surface Area (SASA); evolutionary sequence conservation

1. Introduction

Among all of the cellular components of living systems, proteins are the most abundant and the most functionally versatile. The specific interactions formed by these macromolecules are vital in a wide-range of biological pathways [1]. Protein-protein interactions involved in both transient and long-lasting networks of specific complexes play important roles in many biological processes [2–4]. Characterizing the critical residues involved in these interactions by both experimental and computational methods is therefore crucial to a proper understanding of living systems.

Furthermore, only by gaining a complete understanding at atomistic detail can new methods be developed to modulate their binding [5,6].

Protein-protein interfaces often involve a large number of residues. However, it is generally recognized that small regions of a few residues, termed "Hot-Spots (HS)", are essential for maintaining the integrity of the interface. The development of techniques to identify and characterize protein-based interfaces has become widespread. Experimental Alanine Scanning Mutagenesis (ASM) continues to be a valuable technique for both detecting and analyzing protein-binding interfaces. The contribution of a residue to the binding energy is measured by the binding free energy difference ($\Delta\Delta G_{\text{binding}}$) between the wild-type (WT) and mutant complex upon mutation of a specific residue to alanine [7]. Bogan and Thorn [8] defined the residues with $\Delta\Delta G_{\text{binding}} \geqslant 2.0$ kcal$\cdot$ mol^{-1} as HS; and the residues with $\Delta\Delta G_{\text{binding}} < 2.0$ kcal$\cdot$ mol^{-1} as Null-Spots (NS). Experimental methods for identifying HS are based on molecular biology techniques that are accurate, but still complex, time-consuming and expensive [9]. Highly efficient computational methods for predicting HS can provide a viable alternative to experiments. Molecular Dynamics (MD) simulations can be used to predict changes in the binding strength of protein complexes by calculating the free energy difference from an initial to a final state [10,11]. However, due to the complexity and typical large size of protein-protein complexes, these methods are still computationally expensive. Recently, machine learning approaches trained on various features of experimentally-determined HS residues have been developed in order to predict HS in new protein complexes [6,12–14].

In previous work, we have investigated feature-based methods combining Solvent Accessible Surface Area (SASA) descriptors calculated from static structures and MD ensembles and trained predictors using a Support Vector Machine (SVM) algorithm [15]. However, we only applied these to a small number of complexes, and the prediction performance was hampered by a high number of false positives. More recently, we added an extra feature (residue evolutionary sequence conservation) on a significantly larger dataset. In that study, we explored additional Machine Learning (ML) techniques, which led us to develop a more accurate and time-efficient HS detection methodology. This resulted in new HS predictor models for both protein-protein and protein-nucleic acid interactions, and we implemented the best performing models into two web tools [14].

In this study, we significantly expand both the number of studied protein-protein complexes and the number of 3D complex structure-based features used for prediction, including: interface size, the type of interaction between residues at the interface of the complex and the number of different types of residues at the interface. To the evolutionary sequence-based features, we added the Position-Specific Scoring Matrix (PSSM), for a total of 79 features. We have further tested a total of 27 algorithms from a simple linear-based function to support-vector machine models with different cost functions. The best predictor, based on a conditional inference random forest (c-forest) algorithm, achieves an overall performance characterized with an *F*1-score of 0.73, an accuracy of 0.80, a sensitivity of 0.76 and a specificity of 0.82. To the best of our knowledge, these values are higher than all other available prediction techniques.

2. Results

In the current study, we have used the Classification And Regression Training (Caret) Package [16] from the R software [17], which provides a unified interface with a large number of built-in classifiers, in order to train an HS predictor. The dataset used for this purpose includes 545 amino acids from 53 complexes (140 HS and 405 NS). We calculated the percentage of the different types of amino acids within the NS set (Ser: 7.4; Gly: 1.5; Pro: 2.0; Val: 3.2; Leu: 2.7; Ile: 5.2; Met: 1.0; Cys: 0.7; Phe: 4.7; Tyr: 5.9; Trp: 4.9; His: 4.4; Lys 8.9; Arg: 10.6; Gln: 5.4; Asn: 6.2; Glu: 9.9; Asp: 7.2; Thr: 8.1) and within the HS set (Ser: 2.1; Gly: 2.9; Pro: 2.9; Val: 3.6; Leu: 7.1; Ile: 4.3; Met: 0.0; Cys: 0.0; Phe: 6.4; Tyr: 20.0; Trp: 5.7; His: 2.1; Lys 7.1; Arg: 6.4; Gln: 2.1; Asn: 5.0; Glu: 7.1; Asp: 10.7; Thr: 4.3). For both sets, there is a natural expected tendency for a higher percentage of large hydrophobic or charged residues at the interfaces, in particular Tyr. Although different patterns could influence the training of a robust classifier, we have previously successfully constructed models that were bias-free

for all different amino acids [14]. We randomly split this dataset (see for details Supplementary Information Table S1) into a training set consisting of 70% of data (382 mutations) and an independent test set (163 mutations, 30%). This is a standard division scheme demonstrated to give a good result. All 27 classification models (listed in the Methods Section) were tested using 10-fold cross-validation repeated 10 times in order to avoid overfitting and to obtain the model's generalization error. This means that the training set was split randomly into ten isolated parts, using nine of the ten parts to train the model and taking the remaining fold of data to test the final performance of the model. This process was repeated ten times. The performance of the five best algorithms for each tested condition was independently evaluated on the test set to ensure an unbiased assessment of the accuracy of the final model.

The 79 features used in this work have different scales (i.e., the range of the raw data varies significantly), and therefore, we have performed feature normalization or data standardization of the predictor variables at the training set by centering the data, i.e., subtracting the mean and normalizing it by dividing by the standard deviation. The same protocol was followed for the test set taking into account the use of the training mean and standard deviation to ensure a good estimation of the model quality and generalization power. As we have a high-dimensional dataset (79 features), we have also applied Principal Components Analysis (PCA) to reduce the dimensionality of the data. PCA works by establishing an orthogonal transformation of the data to convert a set of possible correlated variables into a set of linearly-uncorrelated ones, the so-called principal components.

One of the main concerns when applying classification to the detection of HS is the natural imbalance of the data. As expected, the number of HS is lower than the number of NS at a protein-protein interface, as indicated by the presence of 185 HS and 360 NS in the main dataset. In ML classification methods, the disparity of the frequencies of the observed classes may have a very negative impact on the models' performance. To overcome this problem, we have tried two different subsampling techniques for the training set: down-sampling and up-sampling. In the first, there is a random sub-setting of all classes at the training set with their class frequency matching the least prevalence class (HS), whereas in the up-sampling, the opposite is happening with random sampling (with the replacement) of the minority class (HS) to reach the same size as the majority class (NS). Different conditions were thus established: (i) Scaled; (ii) Scaled Up; (iii) Scaled Down; (iv) PCA; (v) PCA Down; and (vi) PCA Up. Various statistical metrics (described in detail in the Methods Section) were adopted to evaluate the performance of the algorithms tested: Area Under the Receiver Operator Curve (AUROC), accuracy, True Positive Rate (TPR), True Negative Rate (TNR), Positive Predictive Value (PPV), False Positive Rate (FPR), False Negative Rate (FNR) and *F*1-score. Figure 1 illustrates the workflow followed in this study.

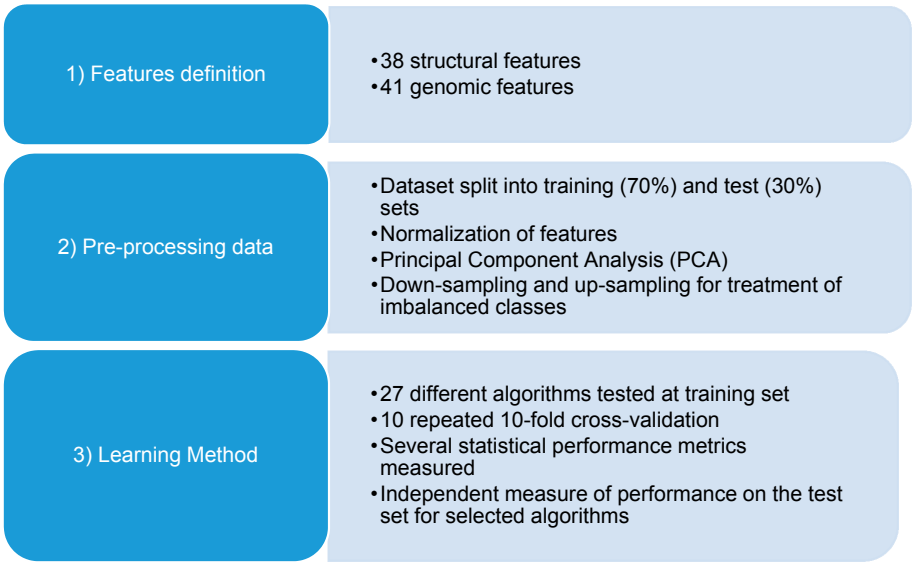


Figure 1. The flowchart of the current work.

The results for the training set for the best five algorithms for each of the six conditions studied are listed in Table 1. All statistical metrics obtained for the complete set of algorithms can be found in Supplementary Information Table S2, in which a more straightforward comparison by type of method can be made. The best classifiers seem to be almost constant in all six different pre-processing conditions, including one neuronal network (avNNET: model averaged Neural Network) and two tree-based methods (C5.0 Tree, C5.0 Rules). The fourth and fifth classifiers vary from nnet (neuronal network), to c-forest, GBM (stochastic gradient boosting machine) and svmRadialSigma (support vector machines with the Radial basis function kernel). The up-sampling of the HS class seems to improve the classifier performance presenting AUROC values higher than 0.80 in the majority of the cases.

Table 1. Statistical metrics attained for five algorithms with top performance for each of the studied conditions for the training set.

Pre-Processing	Metrics	Algorithms				
Scaled		Nnet	avNNET	C5.0 Tree	C5.0 Rules	svmRadialSigma
	AUROC	0.52	0.65	0.77	0.72	0.78
	Accuracy	0.92	0.94	0.96	0.92	0.91
	Sensitivity	0.92	0.88	0.88	0.85	0.80
	Specificity	0.91	0.98	1.00	0.96	0.97
	PPV	0.86	0.95	0.99	0.92	0.93
	NPV	0.95	0.94	0.94	0.92	0.89
	FPR	0.09	0.02	0.00	0.04	0.03
	*F*1-score	0.89	0.92	0.93	0.89	0.86
Scaled_Down		c-Forest	avNNET	C5.0Tree	C5.0Rules	GBM
	AUROC	0.79	0.70	0.73	0.71	0.80
	Accuracy	0.91	0.95	0.96	0.90	1.00
	Sensitivity	0.93	0.96	0.96	0.89	0.99
	Specificity	0.90	0.93	0.95	0.91	1.00
	PPV	0.90	0.93	0.95	0.9	1.00
	NPV	0.92	0.96	0.96	0.89	0.99
	FPR	0.1	0.07	0.05	0.09	0
	F1-score	0.91	0.95	0.96	0.9	1.00
Scaled_Up		c-Forest	avNNET	C5.0Tree	C5.0Rules	GBM
	AUROC	0.85	0.75	0.85	0.82	0.84
	Accuracy	0.93	0.94	0.98	0.95	0.98
	Sensitivity	0.93	0.96	0.99	0.96	0.97
	Specificity	0.93	0.92	0.97	0.94	0.99
	PPV	0.93	0.92	0.97	0.94	0.99
	NPV	0.93	0.96	0.99	0.95	0.97
	FPR	0.07	0.08	0.03	0.06	0.01
	F1-score	0.93	0.94	0.98	0.95	0.98
PCA		nnet	avNNET	C5.0Tree	C5.0Rules	svmRadialSigma
	AUROC	0.69	0.75	0.61	0.59	0.76
	Accuracy	1.00	0.99	0.98	0.92	0.91
	Sensitivity	1.00	0.97	0.98	0.91	0.76
	Specificity	1.00	1.00	0.98	0.93	0.99
	PPV	1.00	0.99	0.96	0.89	0.97
	NPV	1.00	0.98	0.99	0.95	0.88
	FPR	0	0	0.02	0.07	0.01
	F1-score	1.00	0.98	0.97	0.90	0.85
PCA_Down		nnet	avNNET	C5.0Tree	C5.0Rules	svmRadialSigma
	AUROC	0.70	0.78	0.67	0.67	0.75
	Accuracy	0.87	0.91	0.97	0.91	0.91
	Sensitivity	0.88	0.88	0.96	0.96	0.88
	Specificity	0.87	0.93	0.99	0.87	0.93
	PPV	0.87	0.92	0.99	0.88	0.93
	NPV	0.88	0.89	0.96	0.95	0.89
	FPR	0.13	0.07	0.01	0.13	0.07
	F1-score	0.87	0.90	0.97	0.92	0.91
PCA_Up		nnet	avNNET	C5.0Tree	C5.0Rules	svmRadialSigma
	AUROC	0.75	0.82	0.80	0.78	0.80
	Accuracy	0.95	0.98	0.98	0.96	0.94
	Sensitivity	0.94	0.97	0.99	0.96	0.92
	Specificity	0.96	0.99	0.98	0.96	0.95
	PPV	0.96	0.99	0.98	0.96	0.95
	NPV	0.94	0.97	0.99	0.96	0.92
	FPR	0.04	0.01	0.02	0.04	0.05
	F1-score	0.95	0.98	0.98	0.96	0.94

avNNET: model averaged Neural Network; C5.0 Rules (single C5.0 Ruleset); C5.0 Tree (single C5.0 Tree); c-forest (conditional inference random forest); GBM (stochastic gradient boosting machine); nnet (neuronal network); svmRadialSigma (support vector machines with the Radial basis function kernel); Positive Predictive Value (PPV); Negative Predictive Value (NPV); False Positive Rate (FPR).

The performance of a classifier on the training set from which it was constructed gives a poor estimate of its accuracy in new cases. Furthermore, overfitting on algorithms without regularization

terms (such as decision trees and neural networks) is harder to address on the training set. Therefore, the true predictive accuracy of the classifier was estimated on a separate test set corresponding to 30% of the main dataset. Table 2 summarizes the performance on the independent test set for the best classifiers shown in Table 1.

Table 2. Statistical metrics attained for 5 algorithms with the top performance for each of the studied conditions for the independent test set.

Pre-Processing	Metrics	Algorithms				
		Nnet	avNNET	C5.0 Tree	C5.0 Rules	svmRadialSigma
	AUROC	0.71	0.68	0.68	0.72	0.70
	Accuracy	0.74	0.71	0.71	0.74	0.73
	Sensitivity	0.57	0.57	0.5	0.60	0.55
Scaled	Specificity	0.83	0.79	0.83	0.82	0.83
	PPV	0.65	0.6	0.62	0.65	0.64
	NPV	0.78	0.77	0.75	0.79	0.77
	FPR	0.43	0.43	0.4	0.4	0.45
	F1-score	0.61	0.58	0.55	0.62	0.59
		c-forest	avNNET	C5.0 Tree	C5.0 Rules	GBM
	AUROC	0.75	0.68	0.63	0.71	0.73
	Accuracy	0.76	0.69	0.64	0.72	0.75
	Sensitivity	0.79	0.71	0.67	0.76	0.74
Scaled_Down	Specificity	0.74	0.69	0.62	0.70	0.75
	PPV	0.63	0.55	0.49	0.59	0.62
	NPV	0.87	0.81	0.77	0.84	0.84
	FPR	0.21	0.29	0.33	0.24	0.26
	F1-score	0.7	0.62	0.57	0.66	0.68
		c-forest	AvNNET	C5.0 Tree	C5.0 Rules	GBM
	AUROC	0.78	0.73	0.65	0.70	0.80
	Accuracy	0.80	0.75	0.69	0.73	0.82
	Sensitivity	0.76	0.66	0.48	0.59	0.76
Scaled_Up	Specificity	0.82	0.80	0.80	0.81	0.85
	PPV	0.70	0.64	0.57	0.63	0.73
	NPV	0.86	0.81	0.74	0.78	0.86
	FPR	0.24	0.34	0.52	0.41	0.24
	F1-score	0.73	0.65	0.52	0.61	0.75
		Nnet	avNNET	C5.0 Tree	C5.0 Rules	svmRadialSigma
	AUROC	0.65	0.73	0.68	0.71	0.71
	Accuracy	0.67	0.75	0.7	0.74	0.74
	Sensitivity	0.60	0.60	0.66	0.67	0.52
PCA	Specificity	0.71	0.84	0.72	0.77	0.86
	PPV	0.54	0.67	0.57	0.62	0.67
	NPV	0.77	0.79	0.79	0.81	0.76
	FPR	0.4	0.4	0.34	0.33	0.48
	F1-score	0.57	0.64	0.61	0.64	0.58
		Nnet	avNNET	C5.0 Tree	C5.0 Rules	svmRadialSigma
	AUROC	0.70	0.68	0.59	0.61	0.69
	Accuracy	0.71	0.69	0.61	0.63	0.70
	Sensitivity	0.76	0.71	0.55	0.60	0.72
PCA_Down	Specificity	0.68	0.69	0.64	0.64	0.69
	PPV	0.56	0.55	0.46	0.48	0.56
	NPV	0.84	0.81	0.72	0.74	0.82
	FPR	0.24	0.29	0.45	0.4	0.28
	F1-score	0.65	0.62	0.50	0.53	0.63
		Nnet	avNNET	C5.0 Tree	C5.0 Rules	svmRadialSigma
	AUROC	0.67	0.75	0.56	0.61	0.69
	Accuracy	0.7	0.77	0.59	0.63	0.71
	Sensitivity	0.59	0.64	0.48	0.55	0.64
PCA_Up	Specificity	0.76	0.84	0.65	0.68	0.75
	PPV	0.58	0.69	0.43	0.48	0.59
	NPV	0.77	0.81	0.69	0.73	0.79
	FPR	0.41	0.36	0.52	0.45	0.36
	F1-score	0.58	0.66	0.46	0.52	0.61

avNNet: model averaged Neural Network; C5.0 Rules (single C5.0 Ruleset); C5.0 Tree (single C5.0 Tree); c-forest (conditional inference random forest); GBM (stochastic gradient boosting machine); nnet (neuronal network); svmRadialSigma (support vector machines with the Radial basis function kernel).

From all of methods, c-forest, trained on the normalized up-scaling set, had the highest performance metrics on both training and test sets. It was therefore chosen as a final model. In our analysis of this classifier (Figure 2), we observed that the key features are structural ones: specifically, $relSASA_i$, $\Delta SASA_i$, the number of contacts established by the interfacial residues at 4 Å and the number of LEU, VAL and HIS residues at the interface. All of these features were calculated using built-in functions of the VMD package [18] and in-house scripts.

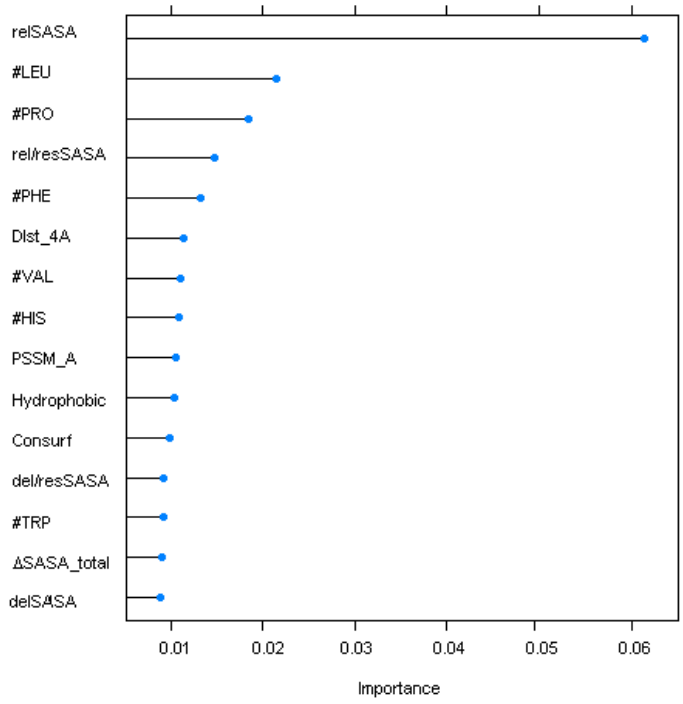


Figure 2. Top 15 variables for the c-forest method. SASA, Solvent Accessible Surface Area; #, Number of residues

To validate the accuracy of the best predictor, we performed the HS predictions with other methods reported in the literature, such as Robetta [19], KFC2-A (Knowledge-based FADE and Contacts) [20], KFC2-B [20] and CPORT (Consensus Prediction Of interface Residues in Transient complexes)(not specialized in HS prediction, but instead, a protein-protein interface predictor) [21] on the same training and test sets. The comparison among these ML methods (Table 3) demonstrates that our new method achieves the best performance with *F*1-scores/AUROC values of 0.73/0.78 on the test set against 0.39/0.62, 0.56/0.66, 0.42/0.67 and 0.43/0.54 for Robetta, KFC2-A, KFC2-B and CPORT, respectively.

Table 3. Comparison of the statistical metrics attained for the best predictor in this work and some of the most common ones in the literature.

Perfomance	Algorithms											
	c-Forest/ Up-Scaling Classes		SBHD2		Robetta		KFC2-A		KFC2-B		CPORT	
	Training	Test	Training	Test	Training	Test	Training	Test	Training	Test	Training	Test
AUROC	0.85	0.78	0.74	0.69	0.62	0.62	0.72	0.66	0.60	0.67	0.54	0.54
Accuracy	0.93	0.80	0.70	0.71	0.66	0.66	0.76	0.71	0.70	0.73	0.49	0.49
Sensitivity	0.93	0.76	0.70	0.70	0.38	0.29	0.57	0.53	0.26	0.28	0.55	0.54
Specificity	0.93	0.82	0.70	0.71	0.85	0.88	0.85	0.81	0.93	0.96	0.45	0.47
PPV	0.93	0.70	0.55	0.56	0.61	0.60	0.67	0.59	0.65	0.80	0.34	0.35
NPV	0.93	0.86	0.82	0.82	0.68	0.67	0.79	0.77	0.71	0.72	0.66	0.66
F1-score	0.93	0.73	0.62	0.62	0.47	0.39	0.62	0.56	0.37	0.42	0.42	0.42

3. Discussion

Machine learning is an area of artificial intelligence that is data driven with a focus on the development of computational techniques for making inferences or predictions. It has become widely used in a variety of areas due to its reduced application time and high performance. Over the past few years, a few algorithms have been applied for the specific problem in this study: the detection of hot-spots at protein-protein interfaces [13–15,22–35].

Here, neural networks and tree-based methods were highlighted as some of the high performance classifiers. Neural networks are inspired by biological nervous systems transmitting the information by a vast network of interconnecting processing elements (neurons). Decision trees organize the knowledge extracted from a hierarchy by using simple tests over the features of the training set. Both have been shown in the past to be promising ML algorithms in the bioinformatics field. Random forests were also shown to be able to predict the impact of each variable in high dimensional problems even in the presence of complex interactions [36]. In particular, c-forest [36], an implementation of the random forest and bagging ensemble method that uses conditional inference trees as base learners, achieved the top performance (Table 2) with a high F1-score of 0.93 on the training set using a 10 repeated 10-fold cross-validation. The values in the independent test (F1 score 0f 0.73) were also very high compared to the ones currently reported in the literature and surpassing all of the other methods tested in this study (Table 3; SBHD (Sasa-Based Hot-spot Detection) 0.61, Robetta 0.39, KFC2-A 0.56, KFC2-B 0.42 and CPORT 0.42). One important aspect that seemed to improve the results compared to our previous approaches (SBHD) was the use of in-built R techniques to balance the training data: up-scaling of the data led to a substantial improvement of the F1-score and to a decrease of the FPR to about 0.19 on the independent test set. In this particular classifier, the first seven features with higher importance were all structure-based: two already used in previous versions of our algorithm ($\Delta SASA_i$ and $relSASA_i$, check Material and Methods) and five new ones (the number of residues at a 4 Å distance and the number of LEU, VAL, HIS and PRO residues at the interface). The PSSM value for the TYR residues, one of the most common residues as HS, was the first genomic-based feature to be ranked as important.

4. Material and Methods

4.1. Dataset Construction

We constructed a database of complexes by combining information from the Alanine Scanning Energetics database (ASEdb) [37], the Binding Interface Database (BID) [38] and the SKEMPI (Structural database of Kinetics and Energetics of Mutant Protein Interactions) [39] and PINT (Protein-protein Interactions Thermodynamic Database) [40] databases, which provide both experimental $\Delta\Delta G_{binding}$ values for interfacial residues and tridimensional (3D) X-ray structure information. The protein sequences were filtered to ensure a maximum of 35% sequence identity for at least one protein in each interface. Crystal structures were retrieved from the Protein Data Bank (PDB) [41], and all water molecules, ions and other small ligands were removed. Our final dataset consists of 545 mutations from 53 different complexes.

4.2. Sequence/Structural Features

From a structural point of view, we compiled 12 previously-used different SASA descriptors for all interfacial residues [14,15]: (i) $_{comp}SASA_i$, the solvent accessible surface area of residue *i* in the complex form; (ii) $_{mon}SASA_i$, the residue SASA in the monomer form; (iii) $\Delta SASA_i$, the SASA difference upon complexation (Equation (1)); (iv) $relSASA_i$, the ratio between ΔSASA for each residue and the $_{mon}SASA_i$ value for the same residue (Equation (2)). A further four features ($_{comp/res}SASA_i$, $_{mon/res}SASA_i$, $_{\Delta/res}SASA_i$ and $_{rel/res}SASA_i$), defined by Equations (3)–(6), were determined applying amino acid standardization by dividing the previous features by the average protein $_{res}SASA_r$ values as determined by Miller and colleagues [42,43], with r being the respective residue type. Four additional, amino-acid standardized features were calculated by replacing the values determined by Miller by our own protein averages $_{ave}SASA_r$ for each amino acid type in its respective protein: $_{comp/ave}SASA_i$, $_{mon/ave}SASA_i$, $_{\Delta/ave}SASA_i$ and $_{rel/ave}SASA_i$, defined in Equations (7)–(10).

$$\Delta SASA_i = \left| {}_{comp}SASA_i - {}_{mon}SASA_i \right| \quad (1)$$

$${}_{rel}SASA_i = \frac{\Delta SASA_i}{{}_{mon}SASA_i} \quad (2)$$

$$_{comp/res}SASA_i = \frac{_{comp}SASA_i}{_{res}SASA_r} \quad (3)$$

$$_{mon/res}SASA_i = \frac{_{mon}SASA_i}{_{res}SASA_r} \quad (4)$$

$$_{\Delta/res}SASA_i = \frac{\Delta SASA_i}{_{res}SASA_r} \quad (5)$$

$$_{rel/res}SASA_i = \frac{relSASA_i}{_{res}SASA_r} \quad (6)$$

$$_{comp/ave}SASA_i = \frac{_{comp}SASA_i}{_{ave}SASA_r} \quad (7)$$

$$_{mon/ave}SASA_i = \frac{_{mon}SASA_i}{_{ave}SASA_r} \quad (8)$$

$$_{\Delta/ave}SASA_i = \frac{\Delta SASA_i}{_{ave}SASA_r} \quad (9)$$

$$_{rel/ave}SASA_i = \frac{relSASA_i}{_{ave}SASA_r} \quad (10)$$

As the SASA features described in Equations (3)–(10) are rather small, the results presented here were multiplied by a factor of 10^3.

We further introduced two features directly related to the size of the interface: the total number of interfacial residues and the $\Delta SASA_{total}$ (sum of the $\Delta SASA_i$ of all residues at the protein-protein binding interfaces). Twenty other features were added by splitting the total number of interface residues into the 20 amino acid types. Four contact features were also calculated: (i) the number of protein-protein contacts within 2.5 Å and (ii) 4.0 Å distance cut-offs, respectively; (iii) the number of intermolecular hydrogen bonds; and (iv) the number of intermolecular hydrophobic interactions. In-house scripts using the VMD molecular package [18] were used for all of these calculations. We used in total 38 structural features in our study.

To utilize evolutionary sequence conservation information, we used the ConSurf server [44] that calculates a conservation score for each amino acid at an interfacial position for a complex, based on known sequences in different organisms. We also computed, PSSM using BLAST [45,46], as well as the weighted observed percentages, introducing them as 40 new features for all interfacial residues. Positive values in this matrix appear for substitutions more frequent than expected by random chance, and negative values indicate that the substitution is not frequent. Therefore, a total of 41 evolutionary sequence-related features were added to the structural features, resulting in 79 features in total for this study.

4.3. Machine Learning Techniques

We first pre-processed the dataset by eliminating missing values or NZV (Near Zero Variance) features. Next, as mentioned in the Results section, we normalized the dataset and performed PCA. The algorithms tested were: avNNet (model averaged Neural Network); bagEarth (bagged MARS (multivariate adaptive regression splines)); bagEarthGCV Bagged MARS using gCV pruning; bagFDA (bagged Flexible Discriminant Analysis); C5.0Rules (single C5.0 Ruleset); C5.0Tree (single C5.0 Tree); c-forest (conditional inference random forest); ctree (conditional inference tree); ctree2 (conditional inference tree); earth (multivariate adaptive regression spline); fda (flexible discriminant analysis); gaussprLinear (Gaussian process); GBM (stochastic gradient boosting machine); gcvEarth (multivariate adaptive regression splines); hdda (high dimensional discriminant analysis); knn (k-nearest neighbors); lda (linear discriminant analysis); lda2 (linear discriminant analysis); multinom (penalized multinomial regression); nnet (neuronal networks); nb (naive Bayes); pda2 (penalized discriminant analysis); svmLinear (Support Vector Machines with Linear Kernel); svmLinear2 (Support Vector Machines with Linear Kernel); svmPoly (Support Vector Machines with Polynomial Kernel); svmRadial

(support vector machines with the Radial basis function kernel); svmRadialCost (support vector machines with the Radial basis function kernel); svmRadialSigma (support vector machines with the Radial basis function kernel); svmRadialWeights (support vector machines with class Weights).

The validity and performance of the various methods was determined by measuring the Area Under the Receiver Operator Curve (AUROC), the accuracy (Equation (11)), True Positive Rate (TPR/recall/sensitivity, Equation (12)), True Negative Rate (TNR/specificity, Equation (13)), Positive Predictive Value (PPV/Precision, Equation (14)), Negative Predictive Value (NPV) (Equation (15)), False Positive Rate (FPR/fall-out, Equation (16)), False Negative Rate (FNR, Equation (17)) and F1-score (Equation (18)) over our dataset.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \tag{11}$$

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{12}$$

$$\text{TNR} = \frac{\text{TN}}{\text{FP} + \text{TN}} \tag{13}$$

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{14}$$

$$\text{NPV} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{15}$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} = 1 - \text{TNR} \tag{16}$$

$$\text{FNR} = \frac{\text{FN}}{\text{TP} + \text{FN}} = 1 - \text{TPR} \tag{17}$$

$$\text{F1 score} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \tag{18}$$

In the equations above, TP stands for True Positive (predicted hot-spots that are actual hot-spots), FP stands for False Positive (predicted hot-spots that are not actual hot-spots), FN stands for False Negative (non-predicted hot-spots that are actual hot-spots) and TN stands the True Negatives (correctly-predicted null-spots).

4.4. Comparison with Other Software

We compared our results with some of the common methods in the literature: Robetta [19], KFC2-A [20] and KFC2-B [20] and CPORT [21].

5. Conclusions

In conclusion, we were thus able to train an accurate and robust predictor using c-forest, a random forest ensemble learning method, and up-sampling of the minor class (HS) for dataset balance. This new method can now be widely applied to the detection of HS in protein-protein interfaces. The code is available upon request, will be implemented as a web-server in the near future and made available for the scientific community at the HADDOCK GitHub repository (http:github.com/haddocking).

Supplementary Materials: Supplementary materials can be found at http://www.mdpi.com/1422-0067/17/8/1215/s1.

Acknowledgments: Rita Melo acknowledges support from the Fundação para a Ciência e a Tecnologia (FCT—SFRH/BPD/97650/2013). The Centre for Nuclear Sciences and Technologies (C^2TN) of Instituto Superior Técnico (IST) authors gratefully acknowledge the FCT support through the UID/Multi/04349/2013 project. Irina S. Moreira acknowledges support by the FCT Investigator program—IF/00578/2014 (co-financed by European Social Fund and Programa Operacional Potencial Humano) and by a Marie Skłodowska-Curie Individual Fellowship MSCA-IF-2015 (MEMBRANEPROT 659826). Irina S. Moreira also acknowledges FEDER

(Programa Operacional Factores de Competitividade—COMPETE 2020) and FCT–project UID/NEU/04539/2013. Zeynep H. Gümüş acknowledges support from the Center for Basic and Translational Research on Disorders of the Digestive System, Rockefeller University, through the generosity of the Leona M. and Harry B. Helmsley Charitable Trust and from the start-up funds of the Icahn School of Medicine at Mount Sinai. This work was supported in part through the computational resources and staff expertise provided by Scientific Computing at the Icahn School of Medicine at Mount Sinai.

Author Contributions: Rita Melo, Robert Fieldhouse and Irina S. Moreira performed the experiments. André Melo, João D. G. Correia, Maria Natália N. D. S. Cordeiro, Zeynep H. Gumus, Joaquim Costa, Alexandre M. J. J. Bonvin and Irina S. Moreira conceived of and designed the experiments. All authors analyzed the data and wrote the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Sudarshan, S.; Kodathala, S.B.; Mahadik, A.C.; Mehta, I.; Beck, B.W. Protein-protein interface detection using the energy centrality relationship (ECR) characteristic of proteins. *PLoS ONE* **2014**, *9*, e97115. [CrossRef] [PubMed]
2. Phizicky, E.M.; Fields, S. Protein-protein interactions: Methods for detection and analysis. *Microbiol. Rev.* **1995**, *59*, 94–123. [PubMed]
3. Clackson, T.; Wells, J.A. A hot spot of binding energy in a hormone-receptor interface. *Science* **1995**, *267*, 383–386. [CrossRef] [PubMed]
4. Uetz, P.; Giot, L.; Cagney, G.; Mansfield, T.A.; Judson, R.S.; Knight, J.R.; Lockshon, D.; Narayan, V.; Srinivasan, M.; Pochart, P.; et al. A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae. *Nature* **2000**, *403*, 623–627. [PubMed]
5. Cho, H.; Wu, M.; Bilgin, B.; Walton, S.P.; Chan, C. Latest developments in experimental and computational approaches to characterize protein–lipid interactions. *Proteomics* **2012**, *12*, 3273–3285. [CrossRef] [PubMed]
6. Moreira, I.S. The role of water occulsion for the definition of a protein binding hot-spot. *Curr. Top. Med. Chem.* **2015**, *15*, 2068–2079. [CrossRef] [PubMed]
7. Cunningham, B.; Wells, J. High-resolution epitope mapping of hgh-receptor interactions by alanine-scanning mutagenesis. *Science* **1989**, *244*, 1081–1085. [CrossRef] [PubMed]
8. Bogan, A.A.; Thorn, K.S. Anatomy of hot spots in protein interfaces 1. *J. Mol. Biol.* **1998**, *280*, 1–9. [CrossRef] [PubMed]
9. Wan, H.; Li, Y.; Fan, Y.; Meng, F.; Chen, C.; Zhou, Q. A site-directed mutagenesis method particularly useful for creating otherwise difficult-to-make mutants and alanine scanning. *Anal. Biochem.* **2012**, *420*, 163–170. [CrossRef] [PubMed]
10. Massova, I.; Kollman, P.A. Computational alanine scanning to probe protein-protein interactions: A novel approach to evaluate binding free energies. *J. Am. Chem. Soc.* **1999**, *121*, 8133–8143. [CrossRef]
11. Moreira, I.S.; Fernandes, P.A.; Ramos, M.J. Computational alanine scanning mutagenesis—An improved methodological approach. *J. Comput. Chem.* **2007**, *28*, 644–654. [CrossRef] [PubMed]
12. Bromberg, Y.; Rost, B. Comprehensive in silico mutagenesis highlights functionally important residues in proteins. *Bioinformatics* **2008**, *24*, i207–i212. [CrossRef] [PubMed]
13. Darnell, S.J.; Page, D.; Mitchell, J.C. An automated decision-tree approach to predicting protein interaction hot spots. *Proteins: Struct. Funct. Bioinform.* **2007**, *68*, 813–823. [CrossRef] [PubMed]
14. Munteanu, C.R.; Pimenta, A.C.; Fernandez-Lozano, C.; Melo, A.; Cordeiro, M.N.D.S.; Moreira, I.S. Solvent accessible surface area-based hot-spot detection methods for protein–protein and protein–nucleic acid interfaces. *J. Chem. Inform. Model.* **2015**, *55*, 1077–1086. [CrossRef] [PubMed]
15. Martins, J.M.; Ramos, R.M.; Pimenta, A.C.; Moreira, I.S. Solvent-accessible surface area: How well can be applied to hot-spot detection? *Proteins: Struct. Funct. Bioinform.* **2014**, *82*, 479–490. [CrossRef] [PubMed]
16. Caret: Classification and Regression Training. Available online: https://cran.r-project.org/web/packages/caret/index.html (accessed on 25 July 2016).
17. R Development Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2010.
18. Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual molecular dynamics. *J. Mol. Graph.* **1996**, *14*, 33–38. [CrossRef]

19. Kim, D.E.; Chivian, D.; Baker, D. Protein structure prediction and analysis using the robetta server. *Nucleic Acids Res.* **2004**, *32*, W526–W531. [CrossRef] [PubMed]
20. Zhu, X.; Mitchell, J. KFC2: A knowledge-based hot spot prediction method based on interface solvation, atomic density and plasticity features. *Proteins* **2011**, *79*, 2671–2683. [CrossRef] [PubMed]
21. De Vries, S.J.; Bonvin, A.M.J.J. Cport: A consensus interface predictor and its performance in prediction-driven docking with haddock. *PLoS ONE* **2011**, *6*, e17695. [CrossRef] [PubMed]
22. Oshima, H.; Yasuda, S.; Yoshidome, T.; Ikeguchi, M.; Kinoshita, M. Crucial importance of the water-entropy effect in predicting hot spots in protein-protein complexes. *Phys. Chem. Chem. Phys.* **2011**, *13*, 16236–16246. [CrossRef] [PubMed]
23. Liu, Q.; Hoi, S.; Kwoh, C.; Wong, L.; Li, J. Integrating water exclusion theory into betacontacts to predict binding free energy changes and binding hot spots. *BMC Bioinform.* **2014**, *15*, 57. [CrossRef] [PubMed]
24. Guharoy, M.; Chakrabarti, P. Empirical estimation of the energetic contribution of individual interface residues in structures of protein–protein complexes. *J. Comput. Aided Mol. Des.* **2009**, *23*, 645–654. [CrossRef] [PubMed]
25. Guharoy, M.; Pal, A.; Dasgupta, M.; Chakrabarti, P. Price (protein interface conservation and energetics): A server for the analysis of protein-protein interfaces. *J. Struct. Funct. Genom.* **2011**, *12*, 33–41. [CrossRef] [PubMed]
26. Chen, H.; Zhou, H.-X. Prediction of interface residues in protein-protein complexes by a consensus neural network method: Test against NMR data. *Proteins* **2005**, *61*, 21–35. [CrossRef] [PubMed]
27. Chen, P.; Li, J.; Wong, L.; Kuwahara, H.; Huang, J.; Gao, X. Accurate prediction of hot spot residues through physicochemical characteristics of amino acid sequences. *Proteins: Struct. Funct. Bioinform.* **2013**, *81*, 1351–1362. [CrossRef] [PubMed]
28. Darnell, S.J.; LeGault, L.; Mitchell, J.C. KFC server: Interactive forecasting of protein interaction hot spots. *Nucleic Acids Res.* **2008**, *36*, W265–W269. [CrossRef] [PubMed]
29. Deng, L.; Guan, J.; Wei, X.; Yi, Y.; Zhou, S. Boosting prediction performance of protein-protein interaction hot spots by using structural neighborhood properties. *Res. Comput. Mol. Biol. Lecture Notes Comput. Sci.* **2013**, *7821*, 333–344.
30. Cho, K.; Kim, D.; Lee, D. A feature-based approach to modeling protein–protein interaction hot spots. *Nucleic Acids Res.* **2009**, *37*, 2672–2687. [CrossRef] [PubMed]
31. Segura Mora, J.; Assi, S.A.; Fernandez-Fuentes, N. Presaging critical residues in protein interfaces: A web server to chart hot spots in protein interfaces. *PLoS ONE* **2010**, *5*, e12352. [CrossRef] [PubMed]
32. Xia, J.; Zhao, X.; Song, J.; Huang, D. Apis: Accurate prediction of hot spots in protein interfaces by combining protrusion index with solvent accessibility. *BMC Bioinform.* **2010**, *11*, 174. [CrossRef] [PubMed]
33. Wang, L.; Liu, Z.; Zhang, X.; Chen, L. Prediction of hot spots in protein interfaces using a random forest model with hybrid features. *Protein Eng. Des. Sel.* **2012**, *25*, 119–126. [CrossRef] [PubMed]
34. Xu, B.; Wei, X.; Deng, L.; Guan, J.; Zhou, S. A semi-supervised boosting svm for predicting hot spots at protein-protein interfaces. *BMC Syst. Biol.* **2012**, *6*. [CrossRef] [PubMed]
35. Ozbek, P.; Soner, S.; Haliloglu, T. Hot spots in a network of functional sites. *PLoS ONE* **2013**, *8*, e74320. [CrossRef] [PubMed]
36. Strobl, C.; Malley, J.; Tutz, G. An introduction to recursive partitioning: Rationale, application and characteristics of classification and regression trees, bagging and random forests. *Psychol. Methods* **2009**, *14*, 323–348. [CrossRef] [PubMed]
37. Thorn, K.S.; Bogan, A.A. ASEdb: A database of alanine mutations and their effects on the free energy of binding in protein interactions. *Bioinformatics* **2001**, *17*, 284–285. [CrossRef] [PubMed]
38. Fischer, T.B.; Arunachalam, K.V.; Bailey, D.; Mangual, V.; Bakhru, S.; Russo, R.; Huang, D.; Paczkowski, M.; Lalchandani, V.; Ramachandra, C.; et al. The binding interface database (BID): A compilation of amino acid hot spots in protein interfaces. *Bioinformatics* **2003**, *19*, 1453–1454. [CrossRef] [PubMed]
39. Moal, I.H.; Fernández-Recio, J. Skempi: A structural kinetic and energetic database of mutant protein interactions and its use in empirical models. *Bioinformatics* **2012**, *28*, 2600–2607. [CrossRef] [PubMed]
40. Kumar, M.D.S.; Gromiha, M.M. Pint: Protein–protein interactions thermodynamic database. *Nucleic Acids Res.* **2006**, *34*, D195–D198. [CrossRef] [PubMed]

41. Bernstein, F.C.; Koetzle, T.F.; Williams, G.J.; Meyer, E.F.; Brice, M.D.; Rodgers, J.R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. The protein data bank. A computer-based archival file for macromolecular structures. *Eur. J. Biochem.* **1977**, *80*, 319–324. [CrossRef] [PubMed]
42. Miller, S.; Janin, J.; Lesk, A.M.; Chothia, C. Interior and surface of monomeric proteins. *J. Mol. Biol.* **1987**, *196*, 641–656. [CrossRef]
43. Miller, S.; Lesk, A.M.; Janin, J.; Chothia, C. The accessible surface area and stability of oligomeric proteins. *Nature* **1987**, *328*, 834–836. [CrossRef] [PubMed]
44. Ashkenazy, H.; Erez, E.; Martz, E.; Pupko, T.; Ben-Tal, N. Consurf 2010: Calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res.* **2010**, *38*, W529–W533. [CrossRef] [PubMed]
45. Altschul, S.; Gish, W.; Miller, W.; Myers, E.; Lipman, D. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410. [CrossRef]
46. Camacho, C.; Coulouris, G.; Avagyan, V.; Ma, N.; Papadopoulos, J.; Bealer, K.; Madden, T.L. Blast+: Architecture and applications. *BMC Bioinform.* **2009**, *10*, 1–9. [CrossRef] [PubMed]
47. Papageorgiou, A.C.; Shapiro, R.; Acharya, K.R. Molecular recognition of human angiogenin by placental ribonuclease inhibitor—An x-ray crystallographic study at 2.0 angstrom resolution. *Embo J.* **1997**, *16*, 5162–5177. [CrossRef] [PubMed]
48. Huang, M.; Syed, R.; Stura, E.A.; Stone, M.J.; Stefanko, R.S.; Ruf, W.; Edgington, T.S.; Wilson, I.A. The mechanism of an inhibitory antibody on TF-initiated blood coagulation revealed by the crystal structures of human tissue factor, fab 5g9 and tf·5g9 complex1. *J. Mol. Biol.* **1998**, *275*, 873–894. [CrossRef] [PubMed]
49. Buckle, A.M.; Schreiber, G.; Fersht, A.R. Protein-protein recognition: Crystal structural analysis of a barnase-barstar complex at 2.0-.Ang. Resolution. *Biochemistry* **1994**, *33*, 8878–8889. [CrossRef] [PubMed]
50. Crystal structure of the *E. Coli* colicin E9 dnase domain with its cognate immunity protein im9. Available online: http://www.rcsb.org/pdb/explore.do?structureId=1bxi (accessed on 26 July 2016).
51. Scheidig, A.J.; Hynes, T.R.; Pelletier, L.A.; Wells, J.A.; Kossiakoff, A.A. Crystal structures of bovine chymotrypsin and trypsin complexed to the inhibitor domain of alzheimer's amyloid beta-protein precursor (APPI) and basic pancreatic trypsin inhibitor (BPTI): Engineering of inhibitors with altered specificities. *Protein Sci.: Publ. Protein Soc.* **1997**, *6*, 1806–1824. [CrossRef] [PubMed]
52. Banner, D.W.; D'Arcy, A.; Chène, C.; Winkler, F.K.; Guha, A.; Konigsberg, W.H.; Nemerson, Y.; Kirchhofer, D. The crystal structure of the complex of blood coagulation factor viia with soluble tissue factor. *Nature* **1996**, *380*, 41–46. [CrossRef] [PubMed]
53. Braden, B.C.; Fields, B.A.; Ysern, X.; Dall'Acqua, W.; Goldbaum, F.A.; Poljak, R.J.; Mariuzza, R.A. Crystal structure of an fv–fv idiotope–anti-idiotope complex at 1.9 å resolution. *J. Mol. Biol.* **1996**, *264*, 137–151. [CrossRef] [PubMed]
54. Fuentes-Prior, P.; Iwanaga, Y.; Huber, R.; Pagila, R.; Rumennik, G.; Seto, M.; Morser, J.; Light, D.R.; Bode, W. Structural basis for the anticoagulant activity of the thrombin-thrombomodulin complex. *Nature* **2000**, *404*, 518–525. [CrossRef] [PubMed]
55. Kwong, P.D.; Wyatt, R.; Robinson, J.; Sweet, R.W.; Sodroski, J.; Hendrickson, W.A. Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody. *Nature* **1998**, *393*, 648–659. [PubMed]
56. Malby, R.L.; Tulip, W.R.; Harley, V.R.; McKimm-Breschkin, J.L.; Laver, W.G.; Webster, R.G.; Colman, P.M. The structure of a complex between the NC10 antibody and influenza virus neuraminidase and comparison with the overlapping binding site of the NC41 antibody. *Structure* **1994**, *2*, 733–746. [CrossRef]
57. Bhat, T.N.; Bentley, G.A.; Boulot, G.; Greene, M.I.; Tello, D.; Dall'Acqua, W.; Souchon, H.; Schwarz, F.P.; Mariuzza, R.A.; Poljak, R.J. Bound water molecules and conformational stabilization help mediate an antigen-antibody association. *Proc. Natl. Acad. Sci. USA* **1994**, *91*, 1089–1093. [CrossRef] [PubMed]
58. Padlan, E.A.; Silverton, E.W.; Sheriff, S.; Cohen, G.H.; Smithgill, S.J.; Davies, D.R. Structure of an antibody antigen complex: Crystal-structure of the HyHEL-10 Fab-lysozyme complex. *Proc. Natl. Acad. Sci. USA* **1989**, *86*, 5938–5942. [CrossRef] [PubMed]
59. Deisenhofer, J. Crystallographic refinement and atomic models of a human Fc fragment and its complex with fragment-B of protein-A from staphylococcus-aureus at 2.9- and 2.8-ANG resolution. *Biochemistry* **1981**, *20*, 2361–2370. [CrossRef] [PubMed]

60. Kobe, B.; Deisenhofer, J. A structural basis of the interactions between leucine-rich repeats and protein ligands. *Nature* **1995**, *374*, 183–186. [CrossRef] [PubMed]
61. Emsley, J.; Knight, C.G.; Farndale, R.W.; Barnes, M.J.; Liddington, R.C. Structural basis of collagen recognition by integrin α2β 1. *Cell* **2000**, *101*, 47–56. [CrossRef]
62. Kirsch, T.; Sebald, W.; Dreyer, M.K. Crystal structure of the BMP-2-BRIA ectodomain complex. *Nat. Struct. Biol.* **2000**, *7*, 492–496. [PubMed]
63. Kvansakul, M.; Hopf, M.; Ries, A.; Timpl, R.; Hohenester, E. Structural basis for the high-affinity interaction of nidogen-1 with immunoglobulin-like domain 3 of perlecan. *Embo J.* **2001**, *20*, 5342–5346. [CrossRef] [PubMed]
64. Kamada, K.; Hanaoka, F.; Burley, S.K. Crystal structure of the maze/mazf complex: Molecular bases of antidote-toxin recognition. *Mol. Cell* **2003**, *11*, 875–884. [CrossRef]
65. Sauereriksson, A.E.; Kleywegt, G.J.; Uhl, M.; Jones, T.A. Crystal-structure of the C2 fragment of streptococcal protein-G in complex with the Fc domain of human-IgG. *Structure* **1995**, *3*, 265–278. [CrossRef]
66. Kuszewski, J.; Gronenborn, A.M.; Clore, G.M. Improving the packing and accuracy of nmr structures with a pseudopotential for the radius of gyration. *J. Am. Chem. Soc.* **1999**, *121*, 2337–2338. [CrossRef]
67. Zhang, E.; St Charles, R.; Tulinsky, A. Structure of extracellular tissue factor complexed with factor VIIa inhibited with a BPTI mutant. *J. Mol. Biol.* **1999**, *285*, 2089–2104. [CrossRef] [PubMed]
68. Radisky, E.S.; Kwan, G.; Lu, C.J.K.; Koshland, D.E. Binding, proteolytic, and crystallographic analyses of mutations at the protease-inhibitor interface of the subtilisin BPN′/chymotrypsin inhibitor 2 complex. *Biochemistry* **2004**, *43*, 13648–13656. [CrossRef] [PubMed]
69. Hage, T.; Sebald, W.; Reinemer, P. Crystal structure of the interleukin-4/receptor alpha chain complex reveals a mosaic binding interface. *Cell* **1999**, *97*, 271–281. [CrossRef]
70. Fields, B.A.; Malchiodi, E.L.; Li, H.M.; Ysern, X.; Stauffacher, C.V.; Schlievert, P.M.; Karjalainen, K.; Mariuzza, R.A. Crystal structure of a t-cell receptor β-chain complexed with a superantigen. *Nature* **1996**, *384*, 188–192. [CrossRef] [PubMed]
71. Nishida, M.; Nagata, K.; Hachimori, Y.; Horiuchi, M.; Ogura, K.; Mandiyan, V.; Schlessinger, J.; Inagaki, F. Novel recognition mode between vav and grb2 sh3 domains. *Embo J.* **2001**, *20*, 2995–3007. [CrossRef] [PubMed]
72. Gamble, T.R.; Vajdos, F.F.; Yoo, S.H.; Worthylake, D.K.; Houseweart, M.; Sundquist, W.I.; Hill, C.P. Crystal structure of human cyclophilin a bound to the amino-terminal domain of HIV-1 capsid. *Cell* **1996**, *87*, 1285–1294. [CrossRef]
73. Barinka, C.; Parry, G.; Callahan, J.; Shaw, D.E.; Kuo, A.; Bdeir, K.; Cines, D.B.; Mazar, A.; Lubkowski, J. Structural basis of interaction between urokinase-type plasminogen activator and its receptor. *J. Mol. Biol.* **2006**, *363*, 482–495. [CrossRef] [PubMed]
74. Abergel, C.; Monchois, V.; Byrne, D.; Chenivesse, S.; Lembo, F.; Lazzaroni, J.-C.; Claverie, J.-M. Structure and evolution of the ivy protein family, unexpected lysozyme inhibitors in gram-negative bacteria. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 6394–6399. [CrossRef] [PubMed]
75. Nam, T.-W.; Il Jung, H.; An, Y.J.; Park, Y.-H.; Lee, S.H.; Seok, Y.-J.; Cha, S.-S. Analyses of MLc-IIBGLc interaction and a plausible molecular mechanism of Mlc inactivation by membrane sequestration. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 3751–3756. [CrossRef] [PubMed]
76. Meenan, N.A.G.; Sharma, A.; Fleishman, S.J.; MacDonald, C.J.; Morel, B.; Boetzel, R.; Moore, G.R.; Baker, D.; Kleanthous, C. The structural and energetic basis for high selectivity in a high-affinity protein-protein interaction. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 10080–10085. [CrossRef] [PubMed]
77. Pelletier, H.; Kraut, J. Crystal-structure of a complex between electron-transfer partners, cytochrome-c peroxidase and cytochrome-c. *Science* **1992**, *258*, 1748–1755. [CrossRef] [PubMed]
78. Prasad, L.; Waygood, E.B.; Lee, J.S.; Delbaere, L.T.J. The 2.5 angstrom resolution structure of the jei42 fab fragment hpr complex. *J. Mol. Biol.* **1998**, *280*, 829–845. [CrossRef] [PubMed]
79. Ghosh, M.; Meiss, G.; Pingoud, A.M.; London, R.E.; Pedersen, L.C. The nuclease a-inhibitor complex is characterized by a novel metal ion bridge. *J. Biol. Chem.* **2007**, *282*, 5682–5690. [CrossRef] [PubMed]
80. Schutt, C.E.; Myslik, J.C.; Rozycki, M.D.; Goonesekere, N.C.W.; Lindberg, U. The structure of crystalline profilin beta-actin. *Nature* **1993**, *365*, 810–816. [CrossRef] [PubMed]

81. Misaghi, S.; Galardy, P.J.; Meester, W.J.N.; Ovaa, H.; Ploegh, H.L.; Gaudet, R. Structure of the ubiquitin hydrolase uch-l3 complexed with a suicide substrate. *J. Biol. Chem.* **2005**, *280*, 1512–1520. [CrossRef] [PubMed]
82. Sundquist, W.I.; Schubert, H.L.; Kelly, B.N.; Hill, G.C.; Holton, J.M.; Hill, C.P. Ubiquitin recognition by the human tsg101 protein. *Mol. Cell* **2004**, *13*, 783–789. [CrossRef]
83. Huang, L.; Hofer, F.; Martin, G.S.; Kim, S.H. Structural basis for the interaction of ras with raigds. *Nat. Struct. Biol.* **1998**, *5*, 422–426. [CrossRef] [PubMed]
84. Hart, P.J.; Deep, S.; Taylor, A.B.; Shu, Z.Y.; Hinck, C.S.; Hinck, A.P. Crystal structure of the human TβR2 ectodomain-TGF-β3 complex. *Nat. Struct. Biol.* **2002**, *9*, 203–208. [CrossRef] [PubMed]
85. Bravo, J.; Li, Z.; Speck, N.A.; Warren, A.J. The leukemia-associated AML1 (Runx1)-CBFβ complex functions as a DNA-induced molecular clamp. *Nat. Struct. Mol. Biol.* **2001**, *8*, 371–378. [CrossRef] [PubMed]
86. Gouet, P.; Chinardet, N.; Welch, M.; Guillet, V.; Cabantous, S.; Birck, C.; Mourey, L.; Samama, J.P. Further insights into the mechanism of function of the response regulator chey from crystallographic studies of the chey-chea(124–257) complex. *Acta Crystallogr. Sect. D-Biol. Crystallogr.* **2001**, *57*, 44–51. [CrossRef]
87. Schneider, E.L.; Lee, M.S.; Baharuddin, A.; Goetz, D.H.; Farady, C.J.; Ward, M.; Wang, C.-I.; Craik, C.S. A reverse binding motif that contributes to specific protease inhibition by antibodies. *J. Mol. Biol.* **2012**, *415*, 699–715. [CrossRef] [PubMed]
88. Hanson, W.M.; Domek, G.J.; Horvath, M.P.; Goldenberg, D.P. Rigidification of a flexible protease inhibitor variant upon binding to trypsin. *J. Mol. Biol.* **2007**, *366*, 230–243. [CrossRef] [PubMed]
89. Johnson, R.J.; McCoy, J.G.; Bingman, C.A.; Phillips, G.N., Jr.; Raines, R.T. Inhibition of human pancreatic ribonuclease by the human ribonuclease inhibitor protein. *J. Mol. Biol.* **2007**, *368*, 434–449. [CrossRef] [PubMed]
90. Bode, W.; Wei, A.Z.; Huber, R.; Meyer, E.; Travis, J.; Neumann, S. X-ray crystal-structure of the complex of human-leukocyte elastase (pmn elastase) and the 3rd domain of the turkey ovomucoid inhibitor. *Embo J.* **986**, *5*, 2453–2458.
91. Read, R.J.; Fujinaga, M.; Sielecki, A.R.; James, M.N.G. Structure of the complex of streptomyces-griseus protease-b and the 3rd domain of the turkey ovomucoid inhibitor at 1.8-a resolution. *Biochemistry* **1983**, *22*, 4420–4433. [CrossRef] [PubMed]
92. Hammel, M.; Sfyroera, G.; Ricklin, D.; Magotti, P.; Lambris, J.D.; Geisbrecht, B.V. A structural basis for complement inhibition by staphylococcus aureus. *Nat. Immunol.* **2007**, *8*, 430–437. [CrossRef] [PubMed]
93. Iyer, S.; Wei, S.; Brew, K.; Acharya, K.R. Crystal structure of the catalytic domain of matrix metalloproteinase-1 in complex with the inhibitory domain of tissue inhibitor of metalloproteinase-1. *J. Biol. Chem.* **2007**, *282*, 364–371. [CrossRef] [PubMed]
94. Zhang, J.-l.; Qiu, L.-y.; Kotzsch, A.; Weidauer, S.; Patterson, L.; Hammerschmidt, M.; Sebald, W.; Mueller, T.D. Crystal structure analysis reveals how the chordin family member crossveinless 2 blocks BMP-2 receptor binding. *Dev. Cell* **2008**, *14*, 739–750. [CrossRef] [PubMed]
95. Friedrich, R.; Fuentes-Prior, P.; Ong, E.; Coombs, G.; Hunter, M.; Oehler, R.; Pierson, D.; Gonzalez, R.; Huber, R.; Bode, W.; et al. Catalytic domain structures of MT-SP1/matriptase, a matrix-degrading transmembrane serine proteinase. *J. Biol. Chem.* **2002**, *277*, 2160–2168. [CrossRef] [PubMed]
96. Farady, C.J.; Egea, P.F.; Schneider, E.L.; Darragh, M.R.; Craik, C.S. Structure of an Fab-protease complex reveals a highly specific non-canonical mechanism of inhibition. *J. Mol. Biol.* **2008**, *380*, 351–360. [CrossRef] [PubMed]
97. Li, Y.L.; Li, H.M.; Smith-Gill, S.J.; Mariuzza, R.A. Three-dimensional structures of the free and antigen-bound Fab from monoclonal antilysozyme antibody hyhel-63. *Biochemistry* **2000**, *39*, 6296–6309. [CrossRef] [PubMed]
98. Reynolds, K.A.; Thomson, J.M.; Corbett, K.D.; Bethel, C.R.; Berger, J.M.; Kirsch, J.F.; Bonomo, R.A.; Handel, T.M. Structural and computational characterization of the SHV-1 β-lactamase-β lactamase inhibitor protein interface. *J. Biol. Chem.* **2006**, *281*, 26745–26753. [CrossRef] [PubMed]
99. Fujinaga, M.; Sielecki, A.R.; Read, R.J.; Ardelt, W.; Laskowski, M.; James, M.N.G. Crystal and molecular-structures of the complex of α-chymotrypsin with its inhibitor turkey ovomucoid 3rd domain at 1.8 a resolution. *J. Mol. Biol.* **1987**, *195*, 397–418. [CrossRef]

8.2 PAPER NUMBER 2

Moreira,I.S*., Koukos,P*., Melo,R., Almeida,J.G., Gomes, A., Schaarschmidt,J., Trellet,M., Gumus,Z.H., Costa,J. and Bonvin,A.M.J.J. (2016) SpotON: a web server for prediction of protein-protein binding hot-spots

SpotON: a web-server for prediction of protein-protein binding Hot-Spots

Irina S. Moreira[1,2*#], Panos Koukos[2#], Rita Melo[1,3], Jose G. Almeida[1], Antonio Gomes[1], Jorg Schaarschmidt[2], Mikael Trellet[2], Zeynep H. Gümüş[4], Joaquim Costa[5], Alexandre M.J.J. Bonvin[2]

[1] CNC - Center for Neuroscience and Cell Biology; Rua Larga, FMUC, Polo I, 1ºandar, Universidade de Coimbra, 3004-517, Coimbra, Portugal.

[2] Bijvoet Center for Biomolecular Research, Faculty of Science - Chemistry, Utrecht University, Utrecht, 3584CH, the Netherlands

[3] Centro de Ciências e Tecnologias Nucleares, Instituto Superior Técnico, Universidade de Lisboa, Estrada Nacional 10 (ao km 139,7), 2695-066 Bobadela LRS, Portugal

[4] Department of Genetics and Genomics and Icahn Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY, USA

[5] CMUP/FCUP, Centro de Matemática da Universidade do Porto, Faculdade de Ciências, Rua do Campo Alegre, 4169-007 Porto, Portugal

Joint first authors
*irina.moreira@cnc.uc.pt , a.m.j.j.bonvin@uu.nl

Abstract- We present SpotOn, a web server that implements a robust algorithm to identify and classify interfacial residues as Hot-Spots (HS) and Null-Spots (NS) with a demonstrated accuracy of 0.80 on an independent test set. The predictor was developed using a random forest ensemble learning algorithm with up-sampling of the minor class and was trained on a large number of complexes and on a high number of different structural- and evolutionary sequence-based features. The SpotOn web interface, which required as input a protein structure in PDB format and a CONSUR files, is freely available at: http://milou.science.uu.nl/cgi/servicesdevel/SPOTON/spoton/.

.

I. INTRODUCTION

The human interactome consists of more than 400.000 protein-protein interactions (PPIs), which are fundamental for a wide-range of biological pathways (1-3). Interactome-level descriptions of molecular function are becoming crucial for a detailed picture and understanding of the nature of complex traits and diseases (4). Characterizing the critical residues involved in these interactions, which can be performed by experimental or computational methods, is therefore crucial for fine tuning PPIs. Furthermore, only through gaining an atomistic-level detail of PPIs can we develop new methods and drugs that modulate their binding (4, 5). Critical for the understanding of PPIs has been the discovery that the driving forces for protein coupling are not evenly distributed across their surfaces: Instead, typically a small set of residues contribute to binding the most, which are – the so called Hot-Spots (HS). These have been defined as the residues which, upon alanine mutation, generate a binding free energy difference ($\Delta\Delta G_{binding}$) ≥2.0 kcal/mol. Oppositely, Null-spots (NS) corresponds to the residue with a $\Delta\Delta G_{binding}$ lower than 2.0 kcal/mol when mutated to alanine (4).

Experimental methods for identifying HS are based on molecular biology techniques that are accurate but still complex, time-consuming and expensive. The necessity of expressing and purifying each individual protein before measurement lead to low-throughput of these techniques, which is is a major bottleneck in HS identification (6). Highly efficient computational methods for HS prediction can therefore provide a viable alternative to experiment. Statistical and Machine-Learning-based (ML) methods are now highly attractive approaches for computational biology as they can be utilized in large scales at relatively low computational costs (7, 8). For the last few years we have been developing new tools and methodologies to accurate predict HS. The initial database used by Martins et al. (9) to train their first predictor

consisted of 15 complexes with a total of 248 interfacial residues and was subsequently extended (8, 10). Our current database includes 53 non-redundant protein complexes with alanine scanning mutagenesis data, genetic conservation scores and three dimensional (3D) crystallographic structures, for a total of 545 mutations. It was derived from the Alanine Scanning Energetics database (11), the Binding Interface Database (12) and the PINT (13) and SKEMPI (14) databases.

Initially, we took into account only 12 solvent accessible surface area (SASA)-related features (9), considering mainly the monomer and complex SASA values and comparing them between with each other and with standard SASA values for each amino acid according to Miller et al. (15). The different SASA-related features submitted to a Support Vector Machine (SVM) algorithm demonstrated the importance of occlusion of HS to the solvent. The following step (10) consisted on gathering evolutionary conservation scores from CONSURF (16, 17) for individual amino acids and using them along the already established SASA relations, as well as a higher number of Machine Learning algorithms. Lastly (8), besides the already mentioned features, a considerable number of additional features were included: two regarding the size of the interface (total number of interfacial residues and total difference of monomer and complex SASA values), four regarding the contact (number of protein-protein contacts within 2.5 Å and 5.0 Å, the number of intramolecular hydrogen bonds and the number of intermolecular hydrophobic interactions, calculated using VMD (18)), 20 related to the residue's percentage at PPIs and 40 regarding the protein sequence (PSSM scores for each amino acid, calculated using BLAST (19, 20), as well as their weighted percentages), amounting to a total of 79 features. From the several ML algorithms analysed, which consisted of variations of, among others, SVMs, neural networks, random forests, multinomial regressions and naïve Bayes, the top performing ML algorithm was found to be c-forest, a random forest implementation with a bagging ensemble which features conditional inference trees as base learners. This was assessed through its F1 score (which can be seen in equation 1, with TP as true positives, FP as false positives and FN as false negatives) using a 10 repeated 10-fold cross-validation.

$$\text{F1 score} = \frac{2TP}{2TP+FP+FN} \qquad (1)$$

The method showed a F1-score 0.73 larger than those reported in the literature so far. The predictor is now implemented in a new and user-friendly web-server, "SpotOn" (hot SPOTs ON protein complexes), that is freely available at: http://milou.science.uu.nl/cgi/servicesdevel/SPOTON/spoton/

II. DESCRIPTION OF THE WEB SERVER

Input

A screenshot of the submission page can be seen in Figure 1. The interface requires the user to upload a 3D structure of the protein-protein complex in the Protein Data Bank (PDB) format (9) and a CONSURF (10, 11) conservation scores file for it. The conservation scores can be easily calculated at http://consurf.tau.ac.il/2016/. The user should also specify the chain identifiers of the two monomers. The choice of the chains that constitute monomer A or B is completely arbitrary. Instructions for all the input are available in the Help section in addition to popups in the submission page.

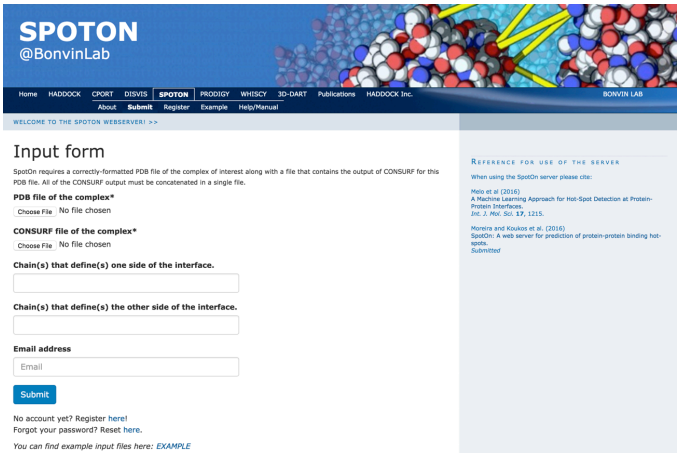


Figure 1: Screenshot of the SpotOn server submission page.

Output and representation of the results

The first step every SpotOn user needs to complete is to register with an email address of their choice, which is used to authenticate them during job submission. Although the server is freely available, registration is required since the user email is used for various notifications about the progress of the job. Upon successful job submission the user receives an email with the URL address where the output of the run will appear as soon as the analysis is complete. An additional email notification containing the URL of the results page is sent upon completion, informing the user of the success or failure of the run.

The main outputs of the server are the two tables that list the residues classified as HS and NS. Figure 2 illustrates the output for an example case (PDBid: 1Z7X (21)) and contains the list of residues predicted as HS. Any column can be used to sort the table. This table along with the NS table are also made available as CSV files in the archive of the run that the user can download. The information contained in those two tables is also visualized in the form of a line plot (e.g. Figure 3) which provides pertinent information when the user hovers the cursor over it (chain identifier, name and index of each residue). This enables the user to quickly identify the residues that have been identified as HS.

HOT-SPOT TABLE

This table contains a list of the residues which have been classified as HotSpots by the algorithm.

Residue Index	Residue Name	Residue Chain	HotSpot Probability
409	LEU	W	0.7
41	LYS	X	0.678
111	GLU	X	0.678
39	ARG	X	0.666
436	ILE	W	0.662
67	ASN	X	0.654
91	ARG	X	0.645
407	ASN	W	0.634
38	GLY	W	0.625
11	GLN	X	0.605

Figure 2: Example table of residues identified as Hot Spots along with their probabilities for the complex with PDBid 1Z7X (21). Only the top 10 Hot-Spots are shown.

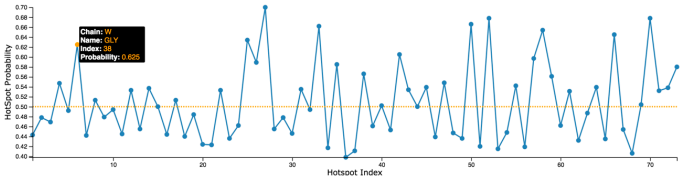


Figure 3: Probability chart of an interface residue being a Hot Spot. Residues above the orange line at 0.50 are predicted as HS and those below as NS. Such a chart is presented to users on the results page. Hovering over a point in this plot will reveal additional information about the residue as shown in the top left of the image.

Finally, the result page provides a direct visualization of the identified HS within the interface of the complex in the form of pre-generated, publication quality views of the complex (Figure 4), that are outputs of the Chimera software (22).

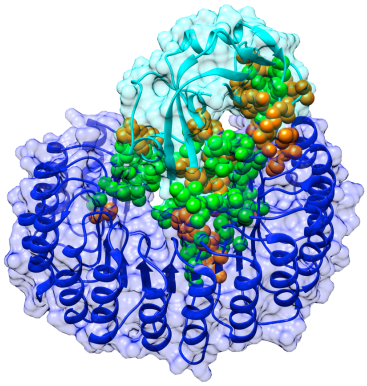


Figure 4: Graphical output example of SpotOn server showing a view of the complex (PDBid 1Z7X (21)) between ribonuclease inhibitor (blue ribbons) and ribonuclease (cyan ribbons), respectively, with a transparent surface representation. Spheres represent interface residues and the HS are in orange.

For each run, all generated results are provided as two gzipped archives, which the user can download from the provided links. The first contains all the graphical outputs of the program: the Chimera images, a static version of the plot described above as well as similar plots that display the probability of a residue being a HS for the entire molecule, broken down by chain identifier. The second archive contains all the text outputs: the CSV file that details all the features (refer to the method paper for details (8)) for the interfacial residues, and the CSV files of the two tables of the results page.

Implementation

The SpotOn server runs alongside the other servers of our group (available at http://milou.science.uu.nl) on a local Linux cluster. The backend is implemented in Python and R, but also makes use of external programs, including VMD (18) BLAST (19, 20) and Chimera (22) during the analysis. It makes use of the Flask microframework for web development and, in addition to the standard languages of the web (HTML, CSS, JS), utilizes the charting library D3.js (22) for the interactive plots in the results page. All scripts are available on Github (http:github.com/haddocking). Documentation is kept up-to-date and support is offered via spoton.csbserver@gmail.com and the BioExcel support forum (http://ask.bioexcel.eu). Calculations submitted by users are anonymous runs on separate directories with randomly generated 12-character key names. Results are kept on the server for 2 weeks. The server workflow is illustrated in Figure 5. If any errors occur at any point of the pipeline illustrated in this figure the analysis will be terminated and an email will be sent to the user prompting them to review the output of the program. Submissions from users are processed in parallel with a maximum number of 15 jobs running simultaneously. Every user is limited to 3 concurrent runs. Typical runtimes for a prediction range between 30 and 90 minutes.

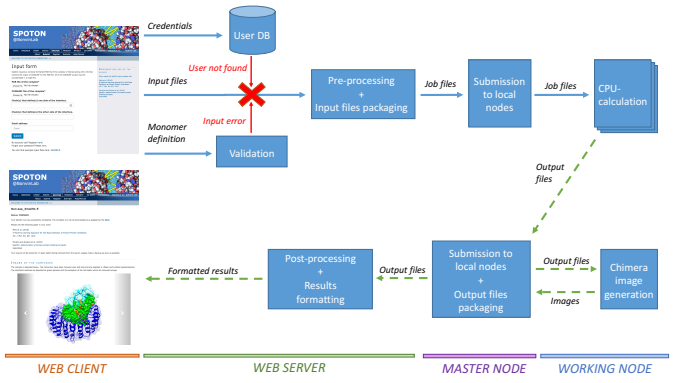


Figure 4: Workflow chart of the entire SpotOn pipeline. Each box corresponds to a step in the pipeline and the horizontal bars at the bottom of the image indicate the environment in which this step takes place. At the very beginning, the user is required to upload the PDB file and the Consurf output for the same molecule in addition to defining the two monomers of the interface. After the credentials of the user have been checked and the input data validated, the web server will generate the run directory with all the necessary files. Should the data be badly formatted or the user not recognized as a registered user of SpotOn a helpful message will be displayed on screen indicating the exact problem. The master node of the Linux cluster where SpotOn is hosted monitors the directory where the run folders are located and if the global maximum number of SpotOn jobs or the number of jobs the particular user has submitted has not exceeded the limits defined in the Implementation paragraph, the analysis is submitted to the queue. Depending on the load of the system at the time of submission, the analysis might start running immediately or with a small delay. The user is notified as soon as the job starts running. The actual run takes place in one of the working nodes of the cluster and as soon as it is finished, the master node submits another job for the generation of the chimera images based on the results of the analysis. At the same time the result archives are generated on the master node and the user is notified of the job completion via email. With the exception of the chimera images, the rest of the elements of the page are generated by the client in real time.

3. CONCLUSIONS AND FUTURE DEVELOPMENT

SpotOn is an easy to use, publicly accessible web server that enables accurate Hot-Spot identification for protein-protein complexes, with minimal input requirements. The method behind it is robust and is arguably the most accurate to date. A successful run will present the user with meaningful results displayed in a user-friendly interactive formats that should be equally useful to experts in the field of computational structural biology as well as less computationally trained researchers.

SpotOn is part of a family of widely-used web portals operated by the Utrecht group in the general area of biomolecular interaction. As such it is part of services for which we aim at ensuring a high reliability and availability. The ML algorithm behind the webserver will be updated as new, more accurate models will be developed.

FUNDING

Irina S. Moreira acknowledges support by the FCT Investigator programme - IF/00578/2014 (co-financed by

European Social Fund and Programa Operacional Potencial Humano), a Marie Skłodowska-Curie Individual Fellowship MSCA-IF-2015 [MEMBRANEPROT 659826], the FEDER (Programa Operacional Factores de Competitividade - COMPETE 2020) and FCT–project: UID/NEU/04539/2013. Rita Melo acknowledges support from the Fundação para a Ciência e a Tecnologia (FCT - SFRH/BPD/97650/2013 and UID/Multi/04349/2013 project). Zeynep H. Gümüş acknowledges support from the Center for Basic and Translational Research on Disorders of the Digestive System, Rockefeller University, through the generosity of the Leona M. and Harry B. Helmsley Charitable Trust and from the start-up funds of Icahn School of Medicine at Mount Sinai. The development of SpotOn was supported by grants from the Netherlands Organization for Scientific Research (NWO) (TOP-PUNT grant no. 718.015.001) and by the European H2020 e-Infrastructure grants West-Life grant no. 675858 and BioExcel grant no. 675728.

REFERENCES

1. Petta,I., Lievens,S., Libert,C., Tavernier,J. and De Bosscher,K. (2015) Modulation of Protein–Protein Interactions for the Development of Novel Therapeutics. Mol Ther, 24, 707–718.

2. Clackson,T. and Wells,J.A. (1995) A hot spot of binding energy in a hormone-receptor interface. Science, 267, 383–386.

3. Moreira,I.S. (2015) The Role of Water Occlusion for the Definition of a Protein Binding Hot-Spot. Curr Top Med Chem, 15, 2068–2079.

4. Moreira,I.S., Fernandes,P.A. and Ramos,M.J. (2007) Hot spots--a review of the protein-protein interface determinant amino-acid residues. Proteins, 68, 803–812.

5. Ramos,R.M. and Moreira,I.S. (2013) Computational Alanine Scanning Mutagenesis-An Improved Methodological Approach for Protein-DNA Complexes. J Chem Theory Comput, 9, 4243–4256.

6. Brender,J.R. and Zhang,Y. (2015) Predicting the Effect of Mutations on Protein-Protein Binding Interactions through Structure-Based Interface Profiles. PLoS Comput Biol, 11, e1004494–25.

7. Xue,L.C., Dobbs,D., Bonvin,A.M.J.J. and Honavar,V. (2015) Computational prediction of protein interfaces: A review of data driven methods. FEBS Lett, 589, 3516–3526.

8. Melo,R., Fieldhouse,R., Melo,A., Correia,J.D.G., Cordeiro,M.N.D.S., Gümüş,Z.H., Costa,J., Bonvin,A.M.J.J. and Moreira,I.S. (2016) A Machine Learning Approach for Hot-Spot Detection at Protein-Protein Interfaces. IJMS, 17, 1215.

9. Martins,J.M., Ramos,R.M., Pimenta,A.C. and Moreira,I.S. (2014) Solvent-accessible surface area: How well can be applied to hot-spot detection? Proteins, 82, 479–490.

10. Munteanu,C.R., Pimenta,A.C., Fernandez-Lozano,C., Melo,A., Cordeiro,M.N.D.S. and Moreira,I.S. (2015) Solvent accessible surface area-based hot-spot detection methods for protein-protein and protein-nucleic acid interfaces. J Chem Inf Model, 55, 1077–1086.

11. Thorn,K.S. and Bogan,A.A. (2001) ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. Bioinformatics, 17, 284–285.

12. Fischer,T.B., Arunachalam,K.V., Bailey,D., Mangual,V., Bakhru,S., Russo,R., Huang,D., Paczkowski,M., Lalchandani,V., Ramachandra,C., et al. (2003) The binding interface database (BID): a compilation of amino acid hot spots in protein interfaces. Bioinformatics, 19, 1453–1454.

13. Kumar,M.D.S. and Gromiha,M.M. (2006) PINT: Protein-protein Interactions Thermodynamic Database. Nucleic Acids Res., 34, D195–8.

14. Moal,I.H. and Férnandez-Recio,J. (2012) SKEMPI: a Structural Kinetic and Energetic database of Mutant Protein Interactions and its use in empirical models. Bioinformatics, 28, 2600–2607.

15. Miller,S., Lesk,A.M., Janin,J. and Chothia,C. (1987) The accessible surface area and stability of oligomeric proteins. Nature, 328, 834–836.

16. Glaser,F., Pupko,T., Paz,I., Bell,R.E., Bechor-Shental,D., Martz,E. and Ben-Tal,N. (2003) ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. Bioinformatics, 19, 163–164.

17. Ashkenazy,H., Abadi,S., Martz,E., Chay,O., Mayrose,I., Pupko,T. and Ben-Tal,N. (2016) ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. Nucleic Acids Res., 44, W344–50.

18. Humphrey,W., Dalke,A. and Schulten,K. (1996) VMD: Visual molecular dynamics. Journal of Molecular Graphics, 14, 33–38.

19. Altschul,S. (1990) Basic Local Alignment Search Tool. J. Mol. Biol., 215, 403–410.

20. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST+: architecture and applications. BMC Bioinformatics 2009 10:1, 10, 421.

21. Johnson,R.J., McCoy,J.G., Bingman,C.A., Phillips,G.N., Raines,R.T. Inhibition of Human Pancreatic Ribonuclease by the Human Ribonuclease Inhibitor Protein. (2007) J. Mol. Biol., 368, 434-449.

22. Pettersen,E.F., Goddard,T.D., Huang,C.C., Couch,G.S., Greenblatt,D.M., Meng,E.C. and Ferrin,T.E. (2004) UCSF Chimera?A visualization system for exploratory research and analysis. J Comput Chem, 25, 1605–1612.