Màster Universitari
Anàlisi de Dades Òmiques / Omics Data Analysis
FACULTAT DE CIÈNCIES I TECNOLOGIA
UVIC | UVIC·UCC

Master of Science in Omics Data Analysis

Master Thesis

HRAS Physical Feature Analysis: Predicting Protein Activation Through a Random Forest Classifier

by

Jorge González García

Supervisor: Jordi Villà i Freixa, Systems Biology Department, UVIC

Academic tutor: Jordi Villà i Freixa, Systems Biology Department, UVIC

Biosciences Department

University of Vic – Central University of Catalonia

10/09/2023

ABSTRACT

Over the past decade, increased computational capabilities have enabled us to address biological questions using data-driven methods, particularly where traditional techniques have been limiting. We hypothesize these computer-based methods can be used to predict enzyme activation status. To verify this claim, we have selected a benchmark protein for study, Human *HRAS*, sourcing a comprehensive set of experimentally labelled structures from available databases. Seven physical, computationally inexpensive features were extracted from these structures at the amino acid alpha carbon level and aligned to the canonical sequence to convey their metrics locally. Subsequently, three-dimensional tensors were generated with them from the set of all possible combinations of the obtained features. A random forest model was then trained on t-SNE preprocessed tensors to look for the highest performing combination of features. Our results strongly suggest that activation status in Human HRAS, and probably other proteins, is mainly codified in the electrostatic and Van der Waals forces, with solvation forces playing a lesser role. These forces, when processed through machine learning models, offer substantial predictive capability. In contrast, methods based on the physical three-dimensional position of residues, such as coordinate-based data and pairwise Root Mean Standard Deviation, were not independently effective in distinguishing activation states.

TABLE OF CONTENTS

1 - INTRODUCTION

In the last ten years we've been seeing an exponential growth of computational resources available for tackling biological problems previously thought only achievable through empirical methods. Since the publication of Alphafold by Deepmind in 2018 (Jumper et al., 2021), proving the solution to the folding problem through graph neural networks, the explosion of solutions to biological problems has escaped the confines of exclusive supercomputers and is achievable in low-end terminals. Besides, both enormous databases of molecular data and supercomputing time are readily available for marginally low costs, enabling solutions faster and more affordable than ever.

The folding problem has been one of the philosopher stones of computational biology for decades. The solution space of transforming moderately sized proteins of a few hundreds of amino acids into their final three-dimensional conformation yielded numbers in the order of 10^{500} possibilities, completely outside the capacities of our most potent supercomputers to deal with. It was thought as well that we couldn't achieve it until a sufficiently powerful quantum computer was developed, capable of simulating the complex interatomic forces that give proteins their shapes and functions. This cornerstone is yet many decades away.

It would turn out we wouldn't have to wait that long. The Critical Assessment of Structure Prediction (CASP14) competition of 2020 achieved unprecedented levels of accuracy in protein predictions and can be plausibly seen as the kickstart of the race to develop purely computational pipelines that could save much money and time.

Perhaps one of the most fundamental questions one can ponder when studying almost any protein with enzymatic capacity is whether it is active or inactive. Activation status of a protein is a complex deal, mediated by post-translational modifications, subcellular location, environmental properties, allosteric molecules, feedback mechanisms and a myriad of other circumstances. In this regard, a purely brute-force computational approach to the simulation of a protein status is an intensive endeavor, not shy of the requirements the very folding problem would have had were not for the steps taken in machine learning and neural networks.

A simpler approach to this problem could be not trying to tackle all the possibilities that can determine whether a protein is active or inactive based on atomic magnitudes, but whether some of the measurable properties of the structures are statistically associated to their activation. For example, we can hypothesize that closer, more compact structures are harder for substrates to access to, and perhaps over a large sample of similarly behaving molecules, a pattern can be detected that can be clearly associated with the catalytic capacity of the protein.

In that regard, several properties of a figurative protein come to mind that have an instant association with the location of its constituent components in relation to one another and with the medium. These are relatively easily calculated through well-known computational methods, and perhaps within that high-dimensional photography of several metrics and coordinates there is a particular pattern that can unequivocally denote or not a protein as active.

In a preliminary unpublished work, the laboratory of Jordi Villà (Lingner et al., 2010) investigated the relevance of residue level solvation free energy values for the discrimination of active and inactive structures of Ras, as showed in Figure 1.1.

They went for a probabilistic approach that did not conclude any statistically significant association, although preliminary results hint at the existence of two distinct conformational patterns separating the active and inactive structures. Other study can be tangentially associated to the goal of this one (Tahir & Hayat, 2017), although their target was on more general protein-protein interactions (PPI). Their results are encouraging in this line of research.

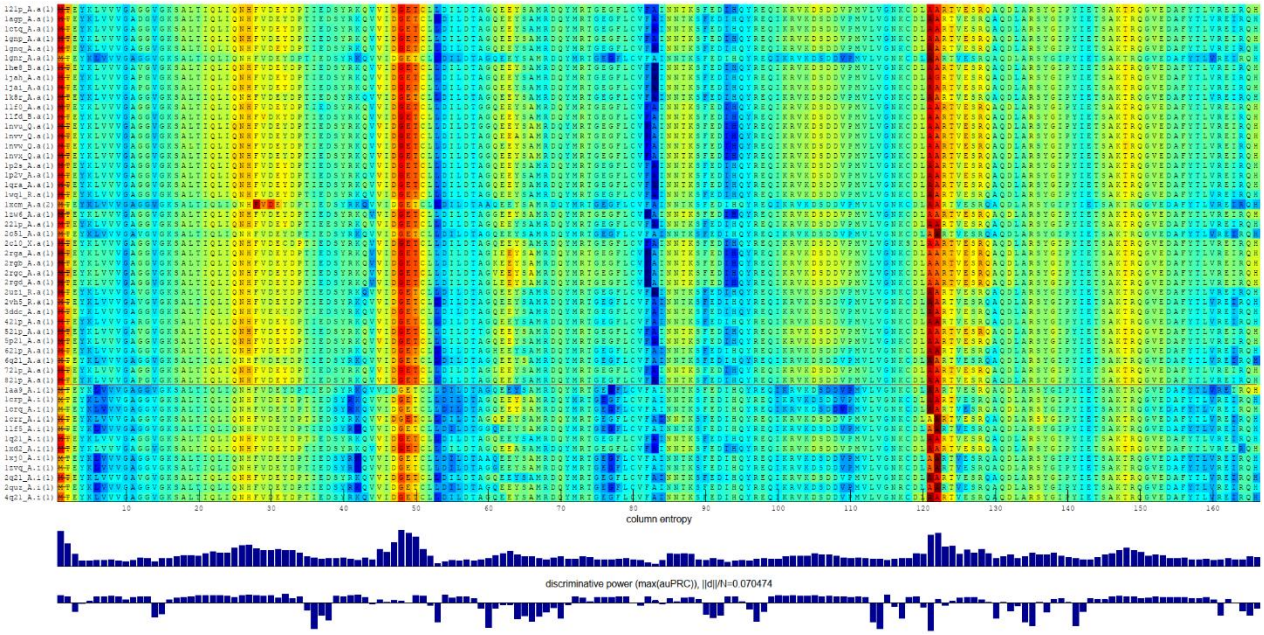

Figure 1.1 – Preliminary unpublished work by the laboratory of Jordi Villà on residue-level solvation energies for the discrimination of active versus inactive protein structures. The heatmap clearly distinguishes two groups of structures, upper (active) and lower (inactive).

Considering the aforementioned factors, prior to employing machine learning techniques, it is advisable to ascertain whether traditional statistical methods cannot achieve comparable classification outcomes. The latter methods are notably more straightforward to implement.

Considering our objective, we aim to investigate the feasibility of developing a machine-learning classifier that can interpret information from existing databases by analysing multiple physical properties of a benchmark protein. This classifier should reliably determine whether a specific protein is in its active or inactive state.

Thus, the two main goals of the project are:

- Demonstrating that a statistical method applied to the physical properties of our benchmark protein is insufficient to categorize its experimental structures into active and inactive forms.
- Select and train a machine learning model based on physical feature metrics that can successfully predict with sufficient efficacy the activation status of our labelled dataset.

More specific goals of this project are:

- Successfully generating a computational pipeline that can automatically generate a model from existing available data.
- Extract relevant physical-chemical features from existing data on our benchmark protein and study their predictive potential on activation status.

- Label the retrieved data for activation status to train a supervised machine learning model.

1.1 - HRAS

The proposed model protein for this project is human HRAS (Uniprot Accession Code P01112). HRAS is a member of the RAS family, a great community of GTPases, which means it hydrolizes guanosine triphosphate (GTP) into guanosine diphosphate (GDP). RAS proteins function as molecular relays, transmitting signals from activated receptors. They switch between two states: an inactive state when bound to GDP and an active state when bound to GTP. In the active GTP-bound state, RAS can engage with and activate a variety of downstream proteins, leading to a spectrum of cellular responses, such as cell growth, survival, differentiation, and potential transformation into cancerous cells.

The structures can be visualized in Figure 1.1 (Fetics et al., 2010). For more information on HRAS structural and functional characteristics, refer to Annex 1.

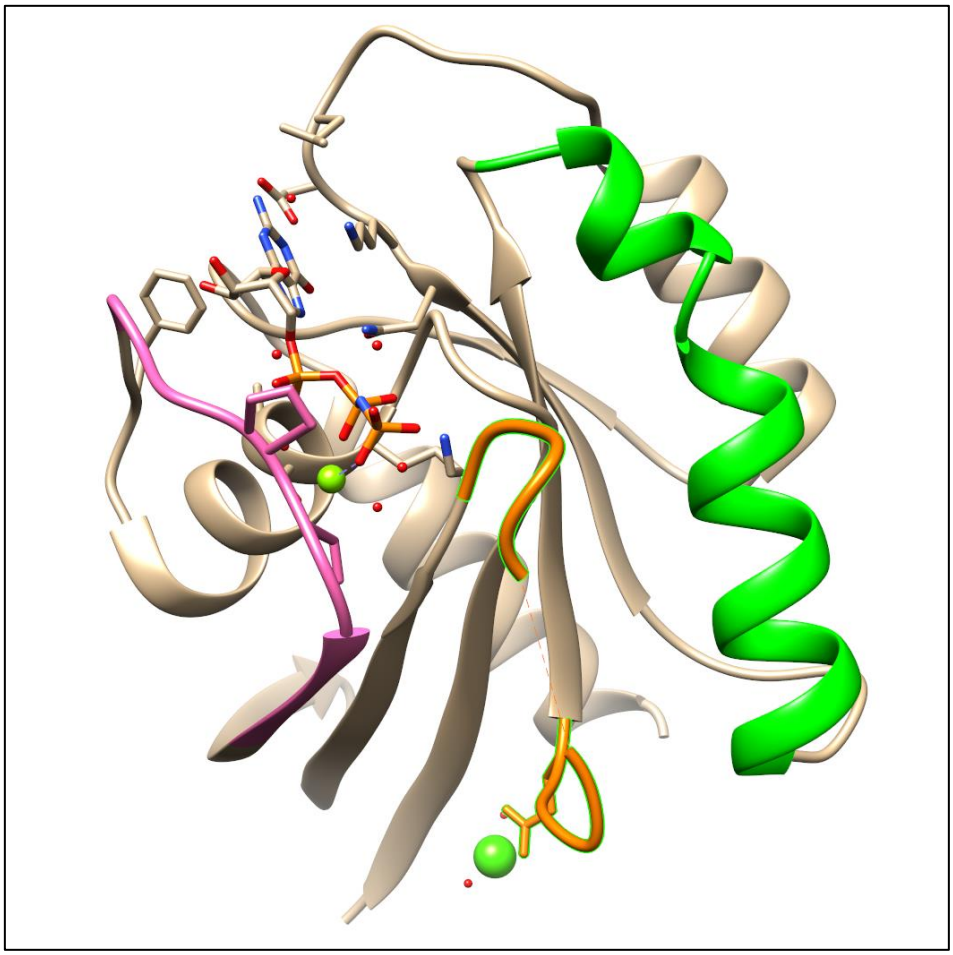

Figure 1.2 – HRAS 3K9L structure. Pink regions correspond to Switch I. Orange regions correspond to Switch II. Altogether, they are mainly responsible for the enzymatic activity of HRAS. Green regions constitute the Linker joining both Switches.

HRAS is a well-studied protein with a critically important physiological function. That is the regulation of several intracellular signaling cascades. It interacts with several effector proteins, and have a role to play

in cell proliferation, differentiation, and survival. Thus, it is a well conserved protein, because mutations that render it dysfunctional usually lead to lethality.

1.2 – Features

Weighted Contact Number

WCN has been shown to provide dynamical characterization of protein structures from the simple measurement of the distance between nodes (C-alpha, centroids, all atoms). Given a network with nodes i and j, the Weighted Contact Number is defined as the sum of the inverse of the squares of all the distances between nodes i and j.

$$WCN_i = \sum_{j \neq i} \frac{1}{d_{ij}^2}$$

When the nodes we are using are, for example, C-alpha atoms in each amino acid residue, we can evaluate a *residue*-based WCN. It provides a practical way to featurize protein conformation at the residue level. A mutation that significantly alters the WCN of a residue might disrupt the protein's stability or its capacity to undergo necessary structural shifts for its function, leading to inactivate or constituent states (Lin et al., 2008).

WCN provides insights into the local environment of an amino acid residue by quantifying the distance to surrounding residues. It is a good measure of the "compactness" of each region, which according to our hypothesis, may critically determine the function of a particular structural region.

pKa of Ionizable Groups

The pKa of an amino acid residue indicates the pH at which it is half-protonated.

Starting with the Henderson-Hasselbach equation (Bombarda & Ullmann, 2010):

$$pH = pKa + \log\left(\frac{[\mathrm{A^-}]}{[\mathrm{HA}]}\right)$$

Given the condition in which the residue is half protonated, both protonated and deprotonated forms are equal, we can arrange the equation so:

$$pH = pKa$$

Changes in protonation states can drastically affect the distribution of charges across the whole electronic domain of the protein, and thus its structure and function (Krusemark et al., 2009).

Many proteins operate at specific pH values. Trying to consider all the possible environmental conditions that can influence a protein structure, or the critical role the surrounding medium's pH can have on its activation state, is precisely the kind of computational effort this project intends to avoid. The best way then to codify the behavior of those structures under different proton concentrations is to study the pKa to know the rate of dissociation of its residues.

Many enzymes utilize acidic or basic residues as catalytic groups. In fact, only a fraction of the constituent amino acids in regular proteins account as ionizable groups that interact strongly with the surrounding medium as acids or alkalis. These "chemical hinges" can have a powerful effect on their surrounding

distribution of electric charges, and a deep impact on the final conformation of the structure, its affinity to substrate and finally its activation state (Fossat et al., 2021).

Energetic Landscape

On a wide perspective, the process of protein folding is driven by energetic considerations (Haq et al., 2010). The native state of a protein is often its lowest energy state, be it on globally or in their multiple transient states (Sorokina et al., 2022). Deviations from this state, be it because of mutations, environmental interference, or PPI, can lead to protein misfolding or new conformations altogether, metastable or persistent. These alterations can usually be associated to pathological or constituent states, and thus to the activation status.

This is critical in the case of HRAS as its function is heavily modulated by a multitude of effector proteins like GAP. Even when it's not possible or convenient to model all the possible PPI participating in the metabolic cascades HRAS regulates, the total energy of its residues will probably contain a pattern characteristic of the interaction with other molecules. Its total energy metrics can provide insights into the stability of their interaction, the potential for allosteric regulation and the probability of transient or stable interactions.

Solvent Interactions

Solvation energies reflect how the residues interact with the protein's solvent (usually water). Proteins consist of a mix of hydrophobic and hydrophilic residues. Their particular distribution of these in a protein's surface and interior plays a crucial role in its folding and function (Tomar et al., 2016).

This property in particular can provide clear insights into the degree of exposure or shielding of hydrophobic regions, indicating potential binding sites, allosteric sites or regions of structural importance.

In conjunction with the coordinates-based WCN or the electric charge based pKa already mentioned, the solvation profile can give us not only the amount of shielding of a residue, but its predisposition to occupy lower or higher energy regions. For example, a residue with a high degree of hydrophobicity can still be relatively exposed to the surrounding medium due to the particular distribution of residues in the immediate area, but under a conformational shift can greatly modify its position to one more suited to its solvation value (Hudáky et al., 2001).

Molecular Interactions and Stability

Finally, to account for all possible sources of conformational shifts in a protein, we have to move to the electronic forces determining the attraction and repulsion between residues. In this study, the metrics that have been considered to account for the electronic component of intramolecular forces are electrostatic and Van der Waals.

Electrostatic forces arise from interactions between charges groups. In proteins, these are often the side chains of amino acids with intrinsic charge like arginine, lysine, glutamate, or aspartate. For instance, an enzyme might stabilize a negatively charged transition state by positioning a positively charged residue nearby. In fact, the particular working mechanism of HRAS contains a single positive residue initializing the reaction, known as an "Arginine Finger".

Electrostatic potential provides insights into charge distribution on the protein's surface. Charge-charge interactions play a pivotal role in ligand binding, PPI and general conformation shifts (Gallicchio & Levy, 2011).

Meanwhile, Van der Waals attractive and repulsive components reflect non-covalent interactions between atoms. This ubiquitous forces are quantum-mechanical in nature and arise from fluctuations in the electronic charge density (Hermann et al., 2017). Even though the forces mediated by these components are

extremely weak, they can indicate shifts in the protein's structure, particulary in the surface (Gao et al., 2015) (Hähl et al., 2012).

1.3 - Statistical Discrimination

In order to determine if active and inactive structures can be discriminated by purely statistical mechanisms, we have used two clustering-based metrics:

Root Mean Standard Deviation (RMSD)

It is a widely used metric in homology analysis to quantify the similarity of two protein structures. It is defined as:

$$RMSD = \sqrt{\frac{1}{n}\sum_{i=1}^{n}|r_i - r'_i|^2}$$

in the two datasets being compared, where n is the number of datapoints, and r and r' are the coordinates of the i^{th} datapoint.

For all our complete training protein dataset, all samples were first superimposed to each other, and their RMSD value calculated. If the scalar divergence between all possible pairs of samples contains within itself the necessary discriminating information, we will be able to see a clustering pattern around the two categories.

Coordinate-based Clustering

As for the coordinate-based clustering, we can implement an algorithm that can be fed on the three-dimensional locations of the residues, and these processed by another clustering algorithm. If the divergence in the 3D coordinates of the atoms in all pairwise comparisons is significant enough and codifies the activation status of the proteins, we will see a clustering pattern around the two categories as well.

Classifier

Our selected method is Random Forest for the following reasons:

- Accuracy. Random Forests often provide a very high accuracy, particularly on complex datasets with many features.
- Feature Importance Discrimination. Random Forests provide insights into the importance of different features in making predictions.
- Resilience to over and underfitting. Due to the averaging or majority voting system, Random Forests are less prone to overfitting than individual decision trees or other ML-based classifiers.
- Large dataset handling. Random Forests can handle large datasets with high-level dimensionality, as well as a large number of input variables.
- Outlier management: Random Forests are resilient to outliers. They can be used to detect and account for outliers in the dataset.

2 - METHODS

2.1 - Pipeline Execution

The pipeline is explained in full detailed in Annex 2, and the full code is available in the provided Github repository. Please refer to the Resources and Code Availability Section for the link to the repository.

Data Acquisition: Four types of data were retrieved:

- Canonical sequence of human HRAS.
- BLASTp query results using the canonical sequence.
- PDB files from BLASTp output.
- Metadata from the RCSB database. (See Annex 2 for retrieval specifics)

Feature Extraction: Seven features were extracted via specific scripts:

- WCN using the WCN Standalone Library (Floor, n.d.)
- pKa using the PropKa Library (Olsson et al., 2011)
- Rosetta profiles using the Pyrosetta Library (Chaudhury et al., 2010).

The features were globally aligned (BLOSUM62) to the amino acids of the canonical sequences and stored into dataframe objects.

Data extraction: RMSD values were determined for all possible structure pairs, and alpha carbon coordinates were extracted from all selected structures. All possible combinations of feature dataframes were created using:

$$C(n,k) = \frac{n!}{k!\,(n-k)!}$$

Where n represents the total number of dataframes and k is a random number between 1 and 7. The operation yielded 127 distinct combinations.

The dataframes within these combinations were stacked along the third dimension into 3D tensor objects, normalized using the Z-score method. Null values were handled through zero imputation.

Dimensionality Reduction and Modelling: Tensors were flattened to 2D arrays and preprocessed with t-SNE of two components. Random forest models were trained on these arrays using a three-fold cross-validation and a hyperparameter grid search. The full details of the model training are available in Annex 2.

2.2 - Analysis

Hierarchical clustering of RMSD values and a 3D PCA on alpha-carbon coordinates were conducted. Each dataframe metric was used to train separate random forest models, from which feature important residues were selected. A two-sample t-test was performed between active and inactive datasets from the 15 most important of these residues. Model training outputs were also stored and analysed. For the complete list of metrics obtained, consult Annex 2.

Detailed visuals and analyses are available in the Results section and in the 'Analysis' folder on the provided Github repository. Specific values obtained from the analysis are available in Annex 3.

3 – RESULTS

3.1 - Statistical Analysis

Comprehensive statistical evaluations, including normalization, clustering, feature importance, and significance studies of the seven metrics, are accessible in the 'Analysis' subfolder of the associated Github repository.

As we can observe in Figure 3.1. Clustering study of RMSD and coordinates for our samples does not yield any aggrupation pattern corresponding to the activation states of the dataset proteins. The PCA performed on the coordinates extracted from the .PDB files can be seen in Figure 3.2 and does not reveal any clustering pattern corresponding to the activation status.

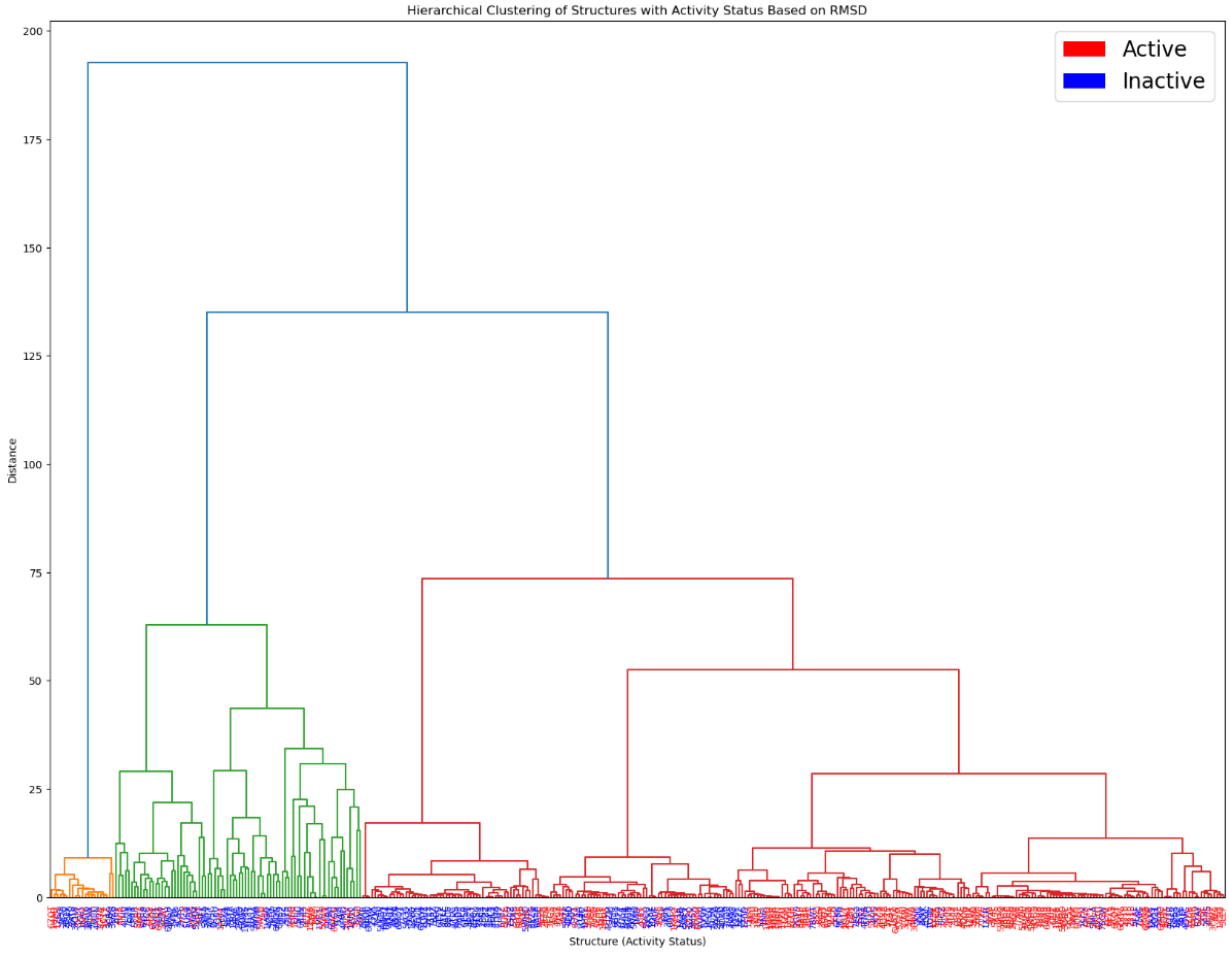

Figure 3.1 – Tree cluster visualization of RMSD values in the complete dataset. Y axis corresponds to the distance between nodes of the tree. X axis contains the different structure identifiers, marked by colour. The red names correspond to active structure identifiers, and the blue correspond to inactive structure identifiers. There is not a clearly discernible clustering pattern in the structures, and there are more than two clusters that do not correspond to the categories of active and inactive.

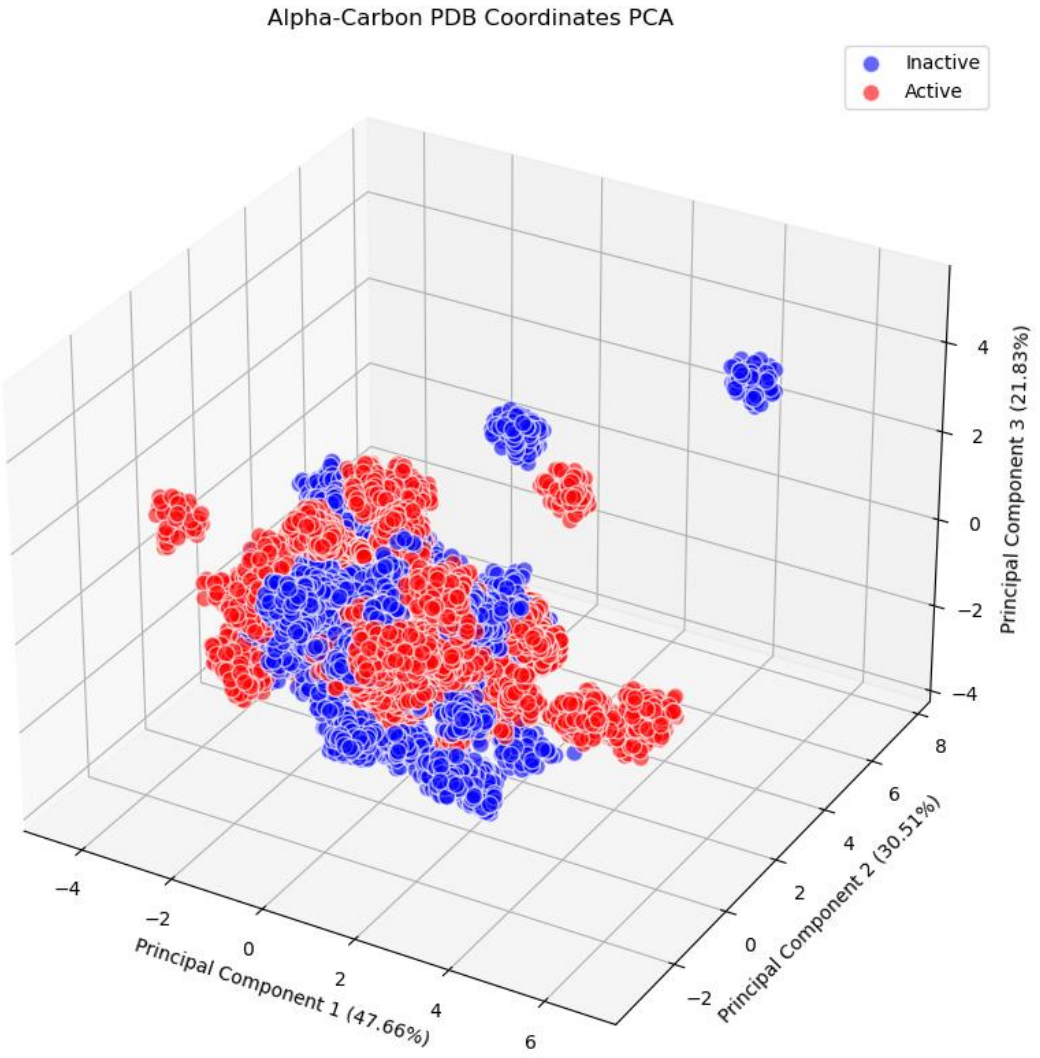

Figure 3.2 – PCA visualization of the 3D coordinates obtained from the .PDB files.Each of the axis represent the contribution of the dimensions to the PCA. There is not clustering arrangement of the datapoints in the space of the PCA into two separate, clearly discernible groups.

3.2 – Feature Importance and Discriminative Analysis

We can observe the impact each of the 15 most important amino acid residues has on random forest model generation for each of the analyzed metrics in Figure 3.3. The precise values are available in Annex 3. For these amino acids, the results of the two-sample t-test (α=0.05) reveals the following distribution in each metric:

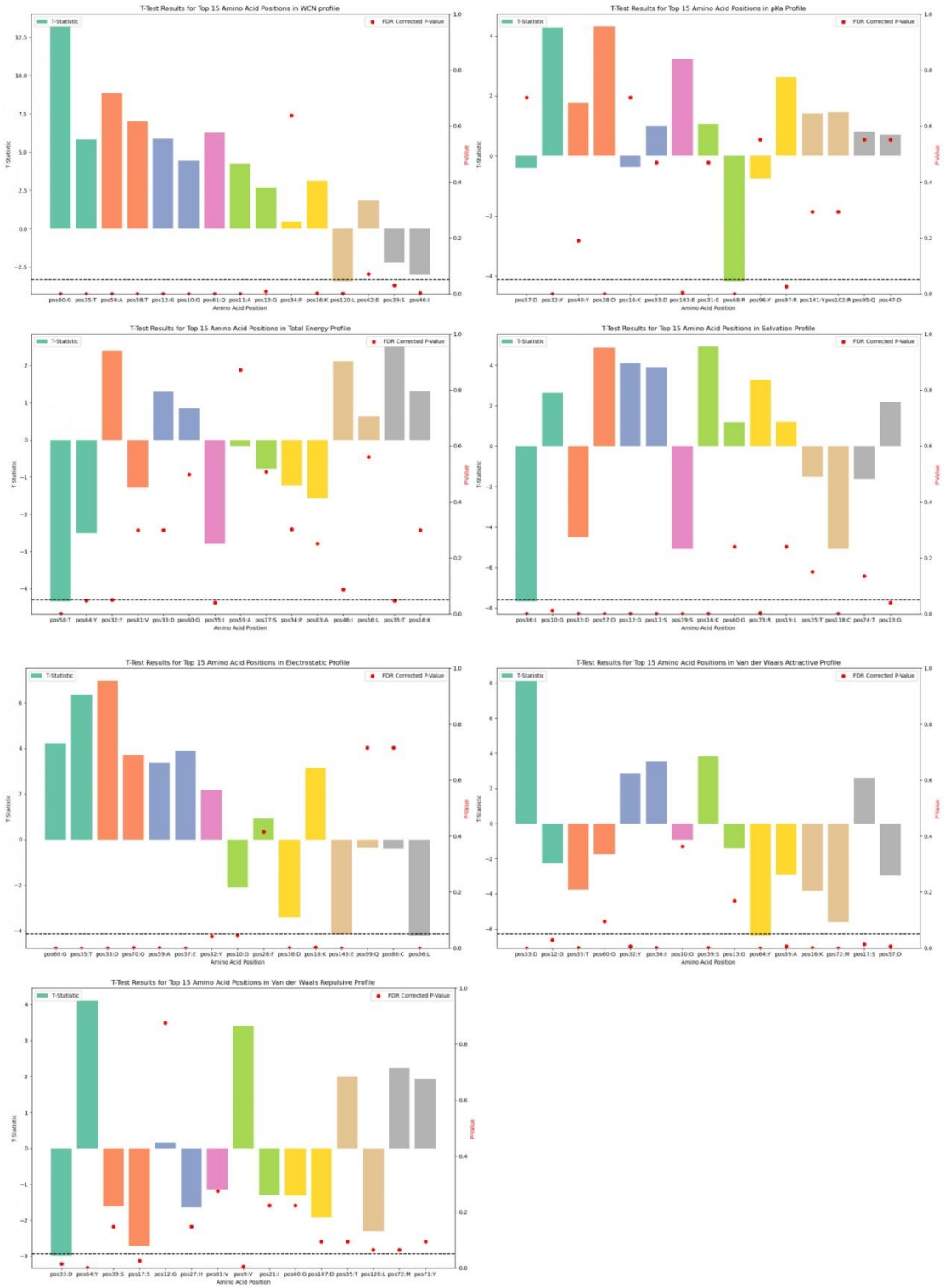

Figure 3.3 – Each of these tables represent the most important amino acids in random forest model formation for each one of the features analysed. The x-axis contains the 15 top-performing amino acids of the whole HRAS in descending order of importance. The y-axis contains the t-statistic from the comparison test between the active and the inactive datasets. The red dots represent the False Discovery Rate-corrected p-values of the test, and the dotted line represents the alpha of the analysis (0.05)

WCN Profile (Refer to Figure 3.3 and Table A3.2):

Major positions in this profile align with vital HRAS regions. The Switch I region exhibited stark disparities between active and inactive structures, with residues T35 (p=3.01E-08) and D39 (p=0.03083573) showing pronounced differences. The Switch II region brought attention to residues G60 (p=1.00E-31) and Q61 (p=3.84E-09). Additionally, T35, G60, Q61, G10 (p=2.66E-05), G12 (p=2.83E-08), and G13 (p=0.00940391) varied structurally and functionally. Comprehensive structural characteristics of HRAS can be referenced in Section 1.1 and Annex 1.

pKa Profile (Refer to Figure 3.3 and Table A3.3):

This profile highlighted significant positions in the Switch I region, specifically Y32 (p=0.000181767) and D38 (p=0.000181767). The Switch II region revealed significance at R68 (p=0.000181767). Notably, Y32 is pivotal for HRAS's functional activity. Consult Section 1.1 and Annex 1 for an in-depth exploration.

Total Energy Profile (Refer to Figure 3.3 and Table A3.4):

The variations between the datasets means for analyzed positions were rather limited in this profile. We find very little significance in T35(p=0.04730029) in Switch I, and Y64(p=0.04730029) in Switch II. Refer to Section 1.1 and Annex 1 for HRAS structural characteristics.

Solvation Profile (Refer to Figure 3.3 and Table A3.5):

This profile elucidated significant variances among key HRAS residues. Residues D33 (p=2.24E-05), I36 (p=2.24E-12), and S39 (p=2.91E-06) in the Switch I region and R73 (p=0.00190848) in the Switch II region were of interest. Residues G10 (p=0.01283681), G12 (p=0.000105356), G13 (p=0.000105356), and C118 (p=2.91E-06) showcased structural and functional variations, with C118 being of unique significance in this profile. Further insights on HRAS are available in Section 1.1 and Annex 1.

Electrostatic Force Profile (Refer to Figure 3.3 and Table A3.6):

A notable distinction in mean values for amino acids between active and inactive structures emerged. Residues Y32 (p=0.041433874), D33 (p=2.19E-10), T35 (p=4.31E-09), E37 (p=0.000299298), and D38 (p=0.00136504) in the Switch I region, and G60 (p=0.000123179) and Q70 (p=0.000485045) in the Switch II region were highlighted. Refer to Section 1.1 and Annex 1 for further structural details.

Van der Waals Attractive Force Profile (Refer to Figure 3.3 and Table A3.7):

The profile underscored several significant disparities in the mean values. Residues Y32 (p=0.007392506), D33 (p=9.81E-14), T35 (p=0.000511909), I36 (p=0.000924024), and S39 (p=0.000459318) in the Switch I region demonstrated heavy significance. In the Switch II region, Y64 (p=4.38E-09) and M72 (p=2.06E-07) were of note. For an in-depth understanding of HRAS, refer to Section 1.1 and Annex 1.

Van der Waals Repulsive Profile (Refer to Figure 3.3 and Table A3.8):

Despite the T-test analyses indicating significance in many residues, FDR Correction impacted numerous p-values. Residues D33 (p=0.015590989) in the Switch I region and Y64 (p=0.00074742) in the Switch II region remained significant post-correction. For details on HRAS structure, refer to Section 1.1 and Annex 1.

3.3 - Machine Learning Model Results

ROCAUC and Accuracy Analysis

Insights into the distribution can be found in Annex 3, Table 3.9 and Table 3.10. Figures 3.4 and 3.5 offer visual depictions of the data. It should be highlighted that the models with the best ROCAUC scores consistently showcased the best performances in other metrics, as illustrated in Figure 3.5.

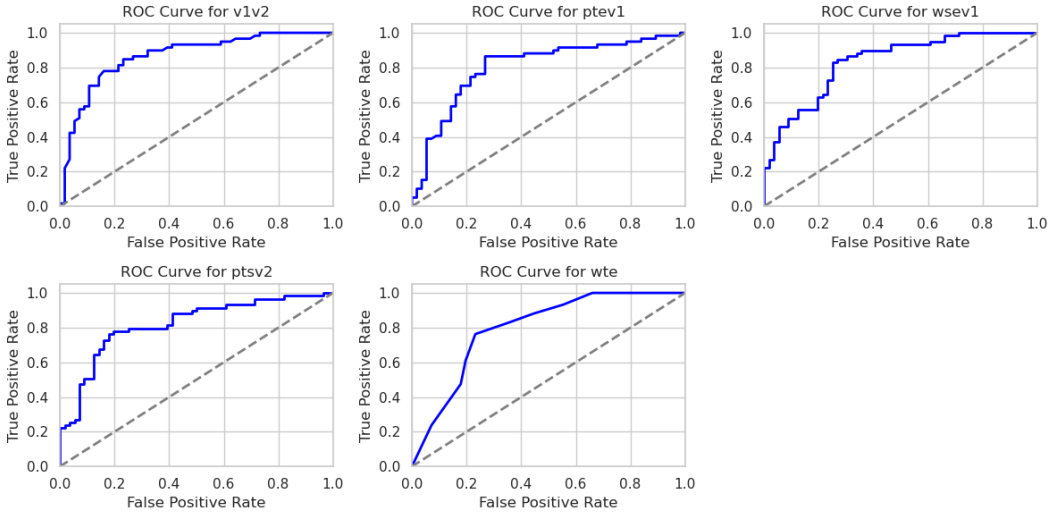

Figure 3.4 – ROC Curves of the top 5 performing models. The x-axis represents the False Positive Rate. The y-axis represents the True Positive Rate. The closer the curves are to the top-left corner of the plot, the better the classifier. The names codify which features the model was trained on: w=WCN; p=pKa; t=Total Energy; s=Solvation Profile; e=Electrostatic Profile, v1=Van der Waals Attractive Profile; v2=Van der Waals Repulsive Profile.

Figure 3.5 – Each plot represents the ranking of the 5 top-performing models for each of these metrics: Accuracy, Log Loss, Cohen's Kappa and Matthew's Correlation Coefficient. The models are ranked for each metric from the top performing (top) to the worst performing (bottom). The names codify which features the model was trained on: w=WCN; p=pKa; t=Total Energy; s=Solvation Profile; e=Electrostatic Profile, v1=Van der Waals Attractive Profile; v2=Van der Waals Repulsive Profile.

Log Loss Analysis

For a detailed assessment, consult Figure 3.5 and Table A3.11. While top-performing models display slight variations when assessed using the log loss metric, the combination of both Van der Waals components emerges as the best performer. Interestingly, weak non-covalent forces remain predominant in terms of predictive capability.

Cohen's Kappa (CK) and Matthew's Correlation Coefficient (MCC):

For visualization, see Figure 3.5. Both metrics confirm the top-scoring models. The exact values can be found in Annex A3, Table A3.12 and Table A3.13.

Feature Importance Distribution:

To understand how importance is distributed across features for each of the top-performing models in terms of accuracy, refer to Figure 3.6.

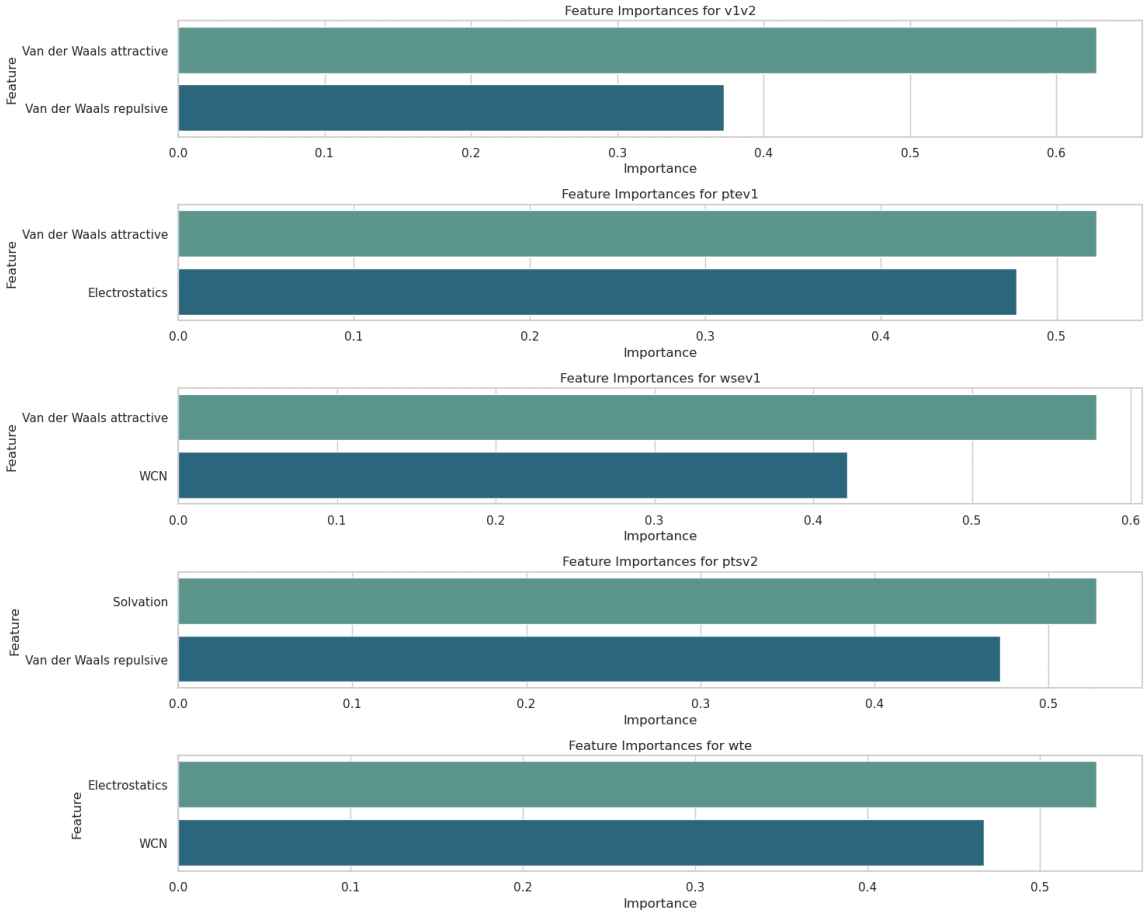

Figure 3.6- Ranking of the most important features in each of the 5 top-performing models. The models are ranked from the best performer (top) to the worst (bottom). In each, the top 2 most important features for model generation are ranked from most important (top) to second most important (bottom). The names codify which features the model was trained on: w=WCN; p=pKa; t=Total Energy; s=Solvation Profile; e=Electrostatic Profile, v1=Van der Waals Attractive Profile; v2=Van der Waals Repulsive Profile.

Confusion Matrix Analysis

A detailed examination of the Confusion Matrix for the top five performing models is presented in Figure 3.16.

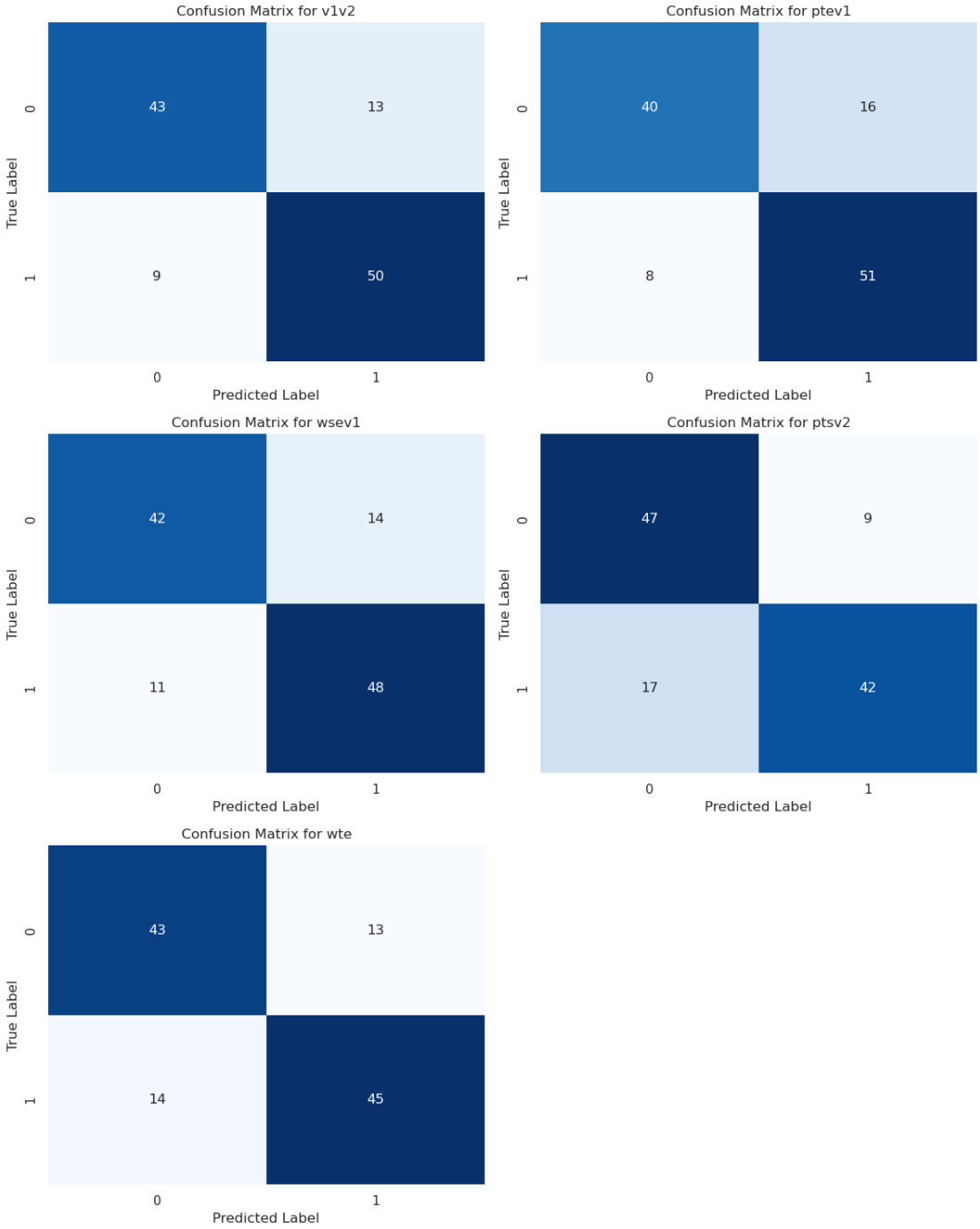

Figure 3.7 – Confusion matrices of the 5 top-performing models according to ROCAUC score. The x-axis in each plot represents the True labels. The y-axis represents the Predicted Label. The names codify which features the model was trained on: w=WCN; p=pKa; t=Total Energy; s=Solvation Profile; e=Electrostatic Profile, v1=Van der Waals Attractive Profile; v2=Van der Waals Repulsive Profile.

3.4 - Discussion

Coordinate-based analysis of the structures doesn't reveal any clear clustering pattern related to enzymatic activation. However, the battery of testing metrics performed on the models seems strongly indicative of the capacity of physical properties of a protein to codify in them its activation status. Even though analysis of the features suggests statistical significance in the means of values comparing active and inactive structures, it's more likely that machine learning and its superior pattern recognition capabilities are necessary to create a model that can make predictions based on the data.

Several metrics show very strong statistical significance when comparing the means of values in the most important positions in most metrics analyzed. Significant differences in Switch I and Switch II are to be expected given the role of these zones in the regulation of HRAS functional cycle.

Of particular interest are Y32 and T35, both highly conserved across structures and critical in function. Y32 is very important in the stabilization of HRAS as it is only accessible upon GAP binding, and can lead to intrinsic GAP independent GTPase activity in mutant isoforms of HRAS (Ilter & Sensoy, 2019), while T35 plays a role in the binding of GEFs and other downstream effectors. We also observe significant differences in G10, G12 and G13, all crucial for HRAS function. Finally, we find an interesting C118 significantly different between active and inactive datasets in the solvation profile, which might explain its role in activation prediction due to its role in GEF binding.

As for the Random Forest model, we find accuracies up to 80%. Interestingly, top performers seem to consistently have three features in common corresponding to weak molecular interactions (Electrostatic and Van der Waals), down to the best 10 performers.

In the testing metrics of the models, we can very clearly observe a predominance of the weak forces as the most valuable metrics for the prediction of the activation status of proteins, followed by the solvation profile. While the winning model is the one trained on the exclusively Van der Waals forces tensor, we can observe at least one occurrence of the electrostatics or Van der Waals features in any of the best performers, across all metrics. We can also observe that the only occurrence of one of these forces not being the most impactful in model generation is when Van der Waals repulsive is in conjunction with the solvation profile, hinting to the possibility that solvation might be more influential in prediction than the Van der Waals repulsive force.

That exactly the winning model would be the combination of the weakest forces is simultaneously surprising and a good indicator that they codify the most information about the conformational configuration of the structures leading to activated status.

It's important to note that these weak non-covalent forces not only consistently lead to better prediction outcomes (ROCAUC, Accuracy, MCC, CK), but to greater resilience to stochastic artifacts of detection as well (Log loss).

Our primary hypothesis for this phenomenon is that the Van der Waals and electrostatic forces predominantly, followed by solvation to a lesser extent, influence higher-level folding more than other analysis metrics combined. Furthermore, they may be more specifically associated with the events leading to activation.These forces may have a more predominant role in tertiary and quaternary structure determination. Simply said, the models could be capturing subtle configuration shifts indicative of the adaptation of the protein to allosteric ligands or the substrate itself.

While total energies or pKa distributions can have a deeper impact on the complete configuration of the protein from its linear state, determining arrangement the location of each residue deep within or in the surface, these shifts are simply too broad to accurately represent the minimal conformational changes happening at the active site, in the switch regions or other allosteric sites where activators or the substrate join in.

For example, while residue protonation plays a crucial role in protein folding and stability, this metric applied to the alpha carbon of the amino acids might not capture the nuanced changes associated with the very

specific interactions that lead to activation or deactivation. In the medium, the intricate interplay between electrolytes and the ionization of polar groups collectively contribute to general enzyme function (Ou et al., 2016). However, this likely doesn't provide discernible patterns concerning ligand binding and active site topology.

Moreover, both Van der Waals and electrostatic forces are sensitive enough to the local environment. They can quantify with great precision small interactions and movements in space; enough precision to be detected even through the masking effect other greater and more drastic forces exert on the configuration of the protein and thus in the model weights.

One of the most potent sources of evidence for this hypothesis is the relative importance of each feature. In all weak forces analysis we find hard statistical significance on amino acids constituent or adjacent to function-critical regions of HRAS.

That these forces act so predominantly as markers of activation can be attributed to two main factors:

- Some authors have observed that while electrostatic and Van der Waals forces lead only to weak intermolecular shifts, their combined action is greatly intensified (Persson et al., 2009). In this study it is shown that both interactions can work in a cooperative way in order to optimize specific biochemical mechanisms.

- Van der Waals and other non-covalent weak forces mainly influence surface processes, whereas other stronger forces, like solvation energies determine at a more fundamental level the structure and function of the protein. This, in turn, means that while greater forces acting at longer distances determine the folding phase solution, lighter forces operating outside the more compact, energy-dense interior are almost solely responsible for interactions with regulator molecules and the substrate itself. Particularly in the case of electrostatic interactions, evidence exists that the influence of surface charges on ion binding to proteins may be more common than generally supposed and could have important consequences for protein function (Linse et al., 1988).

Several works have shown that active site electrostatic preorganization can be linked to the instability of residues vital for catalysis. A notable investigation (Bonet et al., 2005) examined RAS/GAP complexes, identifying protein-protein interaction (PPI) "hotspots". These are specific regions on the HRAS exterior that undergo conformational alterations, positioning residues in an energetically advantageous alignment for interaction with residues on partner protein surfaces. Such findings underscore the significance of electrostatic and weak forces in modulating allosteric regulation and suggest their potential as predictors for protein activation status.

4 - CONCLUSION

It can be concluded that classical statistical analyses of the physical properties in enzyme structures fall short in predicting activation status on their own. However, these structures exhibit patterns linked to their activation status. Leveraging machine learning with feature analysis harnesses these patterns, enabling the creation of predictive classifiers. Notably, weak surface forces appear especially informative, capturing nuances like allosteric regulation and subtle conformational shifts at active sites.

Utilizing existing databases, a classifier can be developed to predict catalytic activation status, leveraging data that can be inexpensively processed from available sources. While this computational approach offers a valuable layer of insight, enabling significant advancements in the understanding of proteins as molecular switches and facilitating the identification of pathological mutations leading to aberrant activity, it does not eliminate the need for experimental methods. Rather, it complements and streamlines the research process by providing an additional layer of information, making experimental endeavors more targeted and efficient.

By integrating the innovations of Alphafold with contemporary machine learning strategies, the field is advancing towards enhanced precision in protein dynamics simulations. These computational advancements offer a refined understanding of protein behavior, for example in the context of protein mutations. It's evident that the confluence of technology and biology is shaping a promising trajectory for protein research.

A1 - HRAS STRUCTURAL CHARACTERISTICS

There are three primary RAS genes which produce four closely related RAS proteins: NRAS, HRAS, and two splice variants of KRAS, KRAS4B and KRAS4A. While these proteins can interact with the same downstream partners due to having identical effector binding domains, their activities are not entirely redundant. Variations in their posttranslational modifications mean they follow different pathways within the cell and position themselves in specific regions of cell membranes. This positioning may grant them access to varied effector groups, leading to different signaling results. For example, research shows that the cancer-causing forms of HRAS are more efficient than NRAS or KRAS in altering fibroblast cells, but NRAS excels in transforming blood-forming cells. Supporting this, gene studies reveal unique roles: eliminating NRAS or HRAS individually or in tandem in mice yields mostly normal outcomes, but a KRAS deficiency is fatal during embryonic development.

Approximately 30% of human cancers, encompassing both solid tumors and blood-related cancers, have mutations linked to RAS genes. Intriguingly, specific RAS gene variants are more commonly associated with certain organ-specific cancers. For instance, KRAS mutations appear in almost 90% of pancreatic cancers. In blood cell-related cancers, NRAS mutations are more prevalent than those in KRAS, while HRAS mutations are uncommon. The reasons for these varied mutation rates among RAS gene variants in blood cell-related cancers remain unclear (Parikh et al., 2007).

The HRAS structure consists of two main parts: a highly conserved G-domain, which is similar across different isoforms and is essential for its GTPase function, and a highly variable region that determines its positioning on the cell membrane.

HRAS oscillates between two conformational states: the GTP-bound active form and the GDP-bound quiescent form. The meticulous modulation of these biophysical transitions is fundamental to ensuring homeostatic cell signaling and is orchestrated by a set of regulatory proteins with high specificity.

Central to this regulatory machinery are the Guanine Nucleotide Exchange Factors (GEFs), particularly the SOS (Son of Sevenless) proteins. These factors, through their Ras Exchange Motif(REM) and CDC25 domains, engage with HRAS in its GDP-associated form. The interaction causes steric hindrance, inducing a conformational alteration in the Switch I and Switch II regions of HRAS. This destabilizes the Mg^{2+} ion coordination, facilitating the egress of GDP. Subsequently, given the high intracellular concentration of GTP relative to GDP, GTP competitively binds to the nucleotide-binding pocket of HRAS, transmuting it into its active conformation.

Conversely, GTPase-activating Proteins (GAPs), including the p120GAP and neurofibromin (NF1) isoforms, serve to attenuate HRAS signaling. By interacting with the Switch I and Switch II regions of GTP-bound HRAS, GAPs allosterically augment the GTP hydrolysis rate, a process that inherently is relatively slow. The provided catalytic asparagine residue from GAPs accelerates the nucleophilic attack on the gamma-phosphorus of GTP by a water molecule, engendering the rapid formation of GDP and inorganic phosphate (Pi). This shifts HRAS back to its quiescent state, truncating the signaling cascade (Herrero et al., 2020).

Anomalies or aberrations in these regulatory processes can precipitate pathophysiological conditions, including oncogenesis, given the pivotal role of HRAS in multiple signaling pathways (Glennon et al., 2000).

Its particular domain distribution is as follows (Downward, 2003; Gremer et al., 2010; Herrero et al., 2020; Ilter & Sensoy, 2019; Rhett et al., 2020):

G-domain (GTPase domain):

- Comprises the majority of the HRAS protein (Residues 1 – 166).
- It is responsible for binding to GTP, GDP or analogous.
- Contains motifs essential for GPT binding and hydrolysis.
- Allows HRAS to switch between GTP or GDP bound states (Active or inactive)

Switch I and Switch II regions:

- Flexible regions within the G-domain
- Switch I goes from residues 30 – 40, while Switch II goes from residues 60-76.
- Undergo conformational changes upon GTP or GDP binding.
- These conformational changes determine the interactions with other effectors and regulators. For example, when HRAS is GTP-bound (active), Switches I and II adopt a conformation that allows the union of effector protein RAF.

Hypervariable region (HVR):

- Located at the terminal C of the protein. It's the part of the protein outside the G-domain.
- It's the location that most differentiates the RAS isoforms (HRAS, KRAS, NRAS).
- Contains important post-translational modifications, such as farnesylation, which determines the membrane location of the protein.

C-terminal CAAX motif: The "C" stands for cysteine, "A" for aliphatic amino acids and "X" for any amino acid. This motif is the site for the forementioned lipid modification that binds HRAS to the cell membrane, located at the terminal carbon of the protein.

Linker region: Connects the G-domain to the HVR region between residues 86 and 104, although this can vary based on the whether HRAS is GTP-bound or not.

At the amino acid level, specific residues in HRAS have a vital role for its function. The most important are (Downward, 2003; Rhett et al., 2020):

Glycine-10 (G10):

- This residue is a part of the P-loop (phosphate-binding loop) and is involved in binding the phosphate groups of GDP/GTP.

Glycine-12 and Glycine-13 (G12 and G13):

- These residues are situated in the phosphate-binding loop (often referred to as the P-loop) of the G domain of HRAS. The P-loop plays a vital role in binding the phosphates of GDP and GTP.
- Mutations on these residues can render HRAS constitutively active, leading to oncogenic signaling. These are common mutated residues in HRAS-associated tumoral processes.

Tryptophan-28 (W28):

- This residue plays a role in effector binding, helping in interaction with downstream partners.

Valine-29 (V29):

- Positioned close to the nucleotide binding pocket, this residue can affect nucleotide affinity and potentially influence the GDP/GTP-bound states of HRAS.

Tyrosine-32 (Y32):

- Located in the Switch I region.
- It is important for the stabilization of the activate conformation of GTP-bound HRAS (active).

Threonine-35 (T35):

- Located in the Switch I region.

- Important for effector interaction and the binding of downstream effectors like RAF.

Glutamate-37 (E37):

- A 2010 study found that duplication of E37 in the Switch I region impairs effector/GAP binding, and it is associated with increased GTP to GDP dissociation (Gremer et al., 2010).

Glycine-60 and Glutamine-61 and Aspartate-119 (G60 ,Q61, D119):

- Located in the Switch II region.
- They play critical roles in GTP hydrolysis.
- Q61, in particular, is critical for catalysing GTP hydrolysis by stabilizing the transition state. Mutations at Q61 can hinder GTP hydrolysis, leading to HRAS persisting in its active form.
- D119 plays a role in GTP hydrolysis, working in conjunction with Q61 to stabilize the transition state.
- Mutations in them impair the intrinsic GTPase activity, and are also visible in neoplastic processes.

Lysine-117 (K117):

- This lysine is involved in binding the γ-phosphate of GTP and plays a role in HRAS's intrinsic GTPase activity. Mutations here could affect the GTPase activity of the protein.

Cysteine-118 (C118):

    - Located away from the nucleotide binding site.
- It is important for the interaction with regulatory proteins like GEFs.

A2 - COMPUTATIONAL PIPELINE

A-2.1 - Overall Function

The pipeline is designed to run from end-to-end on a single command. In order to generate the machine learning model, it first obtains from the internet the canonical sequence associated to the Uniprot accession code, as well as a large collection of protein structures via BLASTp query under the specified identity thresholds. The retrieved structures allow for the download of the whole .PDB files associated with them, which are programmatically trimmed to obtain the chain corresponding to the BLASTp hit. Subsequentally, several nested functions obtain relevant metadata from the structures and store them in a Pandas dataframe structure. The chains are then converted to new .PDB files, their sequences aligned to the canonical sequence previously stored, so each relevant metric is associated to its value in the alpha carbons of the experimental structures and aligned to their corresponding position in the canonical sequence. The data extracted for all these proteins at the alpha carbon level is:

Weighted Contact Number (WCN)

PropKa Profiles (PKa)

Pyrosetta profiles for:

- Total Energy
- Solvation
- Electrostatic Potential
- Van der Waals Attractive Component (1)
- Van der Waals Repulsive Component (2)

The script extracts several .CSV objects through the execution for statistical analysis and quality control, as well as logging all activity throughout the execution. The first goal of the script is the generation of seven dataframes, one for each of the metrics analyzed, whose rows correspond to the analyzed experimental structures, while the columns correspond to the position of the aminoacids in the canonical sequence (pe: Pos 1: M, Pos 2: T)

These dataframes contain the values of each metric associated to each of the corresponding alfa carbons in the alignment. Therefore, different numerical parameters for the alignment yield different substitutions and gaps, modifying the results. The gaps are represented through NaN values.

These dataframes are then inputted to the second script (Tensor#.py), whose first task is the normalization and preparation of the raw metric data. The script performs a Z-score normalization of all values and performs a 0-input operation on all NaN values of the dataframes in preparation for the Random Forest analysis.

Then it proceeds to group all the normalized dataframes in all the possible permutations, stacking them across the third dimension to obtain the set of all possible 3D tensors where the first dimension contains the aminoacid positions, the second contains the experimental structures and the third a Pandas array of the normalized metrics for each amino acid. The operation results in the creation of 127 distinct tensors, going from 1 metric (original dataframes) to 7 metrics (the complete stacking of all metrics), identified by a unique code that distinguish the analysis contained in them.

The third script (RF_model.py) has two functions. First, it compresses the 3D tensors into 2D structures using a feature importance statistical analysis tool (t-SNE), before initializing a Random Forest model and training it with the obtained structure.

The second is the training of the models, one for each tensor in the set. Hyperparameter selection is done programmatically using the Grid Search algorithm across a number of options for several distinct hyperparameters. The model is then trained using cross-validation with three folds on each of the 127

permutations of normalized metrics dataframes. The results are then stored for subsequent analysis and the execution of the pipeline stops.

A2.2 - Overall Structure

The pipeline consists of a string of three Python scripts with sequential instructions for retrieval, preprocessing and analysis of .PDB files obtained from the PDB database in order to train the most effective binary classifier through Random Forest modelling.

The scripts are orchestrated through the use of Nextflow, and callable from a single bash command specifying the analysis parameters. The required arguments for the execution of the pipeline are:

- --uniprot_code (Required)
- --active_ligands (Required)
- --inactive_ligands (Required)
- --query_coverage_threshold (Default: 50)
- --identity_threshold (Default: 30)
- --gap_open_penalty (Default: -0.2)
- --gap_extend_penalty (Default: -0.2)
- --seed (Default: 1)

It logs their own execution in runtime, on top of NextFlow responding to halting states and having its own log control during the execution. It will retry restarting three times after encountering a non-exceptionable problem before halting the process.

It also generates their own folders for storing of the analysis files. NextFlow ensures the assignment of execution permissions for the subsequent directories, subdirectories, and stored files, for which it is convenient to assign the script itself permissions under the sudoer group. The pipeline generates such folders in the same directory where it is located at the time of execution, generating absolute paths in the same directory address as the location of the NextFlow file, and given no permission problems, expands the directory structure downwards until the pipeline is completed and the execution halted.

The pseudo-random number used for the reproducibility of the analysis is specified in the seed parameter supplied during the script call, and it is shared among all the scripts until the end of the execution.

All the files generate several .PKL (Pickle) files along the execution. Pickle files are binary storage files that store the result of particularly heavy or critical processes that would otherwise take significantly more time to process again and again throughout the execution of the pipeline. These are saved to the “pickles” folder after the analysis. Inside this one there is the “tensors” subdirectory, which stores the generated tensors from the permutations of all metrics dataframes.

The retrieval of the .PDB files is done via an intermediary file type (.ENT), and stored in the “ent_files” folder after the analysis. These contain all the retrieved files from the BLASTp query, but many of them remain without use after the filtering process.

The resulting filtering yields the .PDB files trimmed to the chain corresponding to the hits in the BLASTp query, and are stored in the “full_pdb_files” folder.

The metadata analysis includes a function for the generation of FASTA files from the selected chain .PDB files in order to perform phylogenetic tree analysis on them. These are stored in the “chain_fasta_files” folder after analysis.

PropKa analysis generates its own files, necessary for the extraction of the pKa data. These files are stored in the “propk_files” folder.

Exception to this are the “pyrosetta_pdb_files” and “wcn_full_pdb_files” folders, which contain the same .PDB files analyzed, but whose B-factor value is substituted at each alpha-carbon with the associated metric from the corresponding analysis, allowing the visualization of the given values through tools like UCSF Chimera.

The “result_dataframes” folder contains the several dataframes extracted along the analysis in .CSV format. These files are reported in Table 2.1.

Table A2.1 - .CSV files generated during the execution of the pipeline. The Analysis column contains the type of data contained in the files. The Raw file column contains the names of the files with non-normalized data. The Normalized file column contains the names with the normalized data.

Analysis	Raw File	Normalized File
Alignment	alignment_dataframe.csv	-
Metadata	metadata_dataframe.csv	-
WCN	wcn_dataframe.csv	w_dataframe_normalized.csv
PropKa	propka_dataframe.csv	p_dataframe_normalized.csv
Total Energy	pyrosetta_total_energy_dataframe.csv	t_dataframe_normalized.csv
Solvation	pyrosetta_solvation_dataframe.csv	s_dataframe_normalized.csv
Electrostatics	pyrosetta_electrostatics_dataframe.csv	pyrosetta_electrostatics_dataframe.csv
Van Der Waals attractive	pyrosetta_vanderwaals1_dataframe.csv	pyrosetta_vanderwaals1_dataframe.csv
Van Der Waals repulsive	pyrosetta_vanderwaals2_dataframe.csv	v2_dataframe_normalized.csv

From the alignment_dataframe.csv, all structure ids in subsequent dataframes follow the same ordering of structures along the second dimension. That is, alphabetically.

Finally, the “models” folder contains the “RF” subdirectory, where all the models corresponding to all the possible permutations of metrics dataframes are stored after training. Each of these subdirectories contain two files: The trained model and another file with extensive data regarding the output and selected parameters of said model / combination, both in pickle format.

A2.3 - Data Extraction

The first script (Aligner#.py) is the largest, amounting to 1447 lines of code. Its purpose is the retrieval, transformation, and data extraction of the .PDB files. It revolves around the class Datahub, that contains all the necessary data throughout the analysis, as well as storing the input parameters for the execution of the script. The script per se requires the input via bash of the Uniprot Accession code (in this case P01112 corresponding to HRAS), as well as gap open and extension penalties for global (Needleman-Wunsch) alignment (default BLOSUM62). It also requires the desired thresholds for identity search and query

coverage, as well as two comma separated strings of the ligands that unequivocally distinguish the given protein as active or inactive for labeling purposes.

Retrieval and Preprocessing

Three types of queries are performed:

- A HTTP request to Uniprot, querying the Uniprot accession code (P01112 HRAS) in search of the canonical FASTA file containing the raw aminoacid sequence.
- A BLASTp query against PDB, looking for hits on the previously obtained canonical sequence. The default Expect value is 0.01, with a hitlist size of 10000 samples and a word seed of 3. All the other parameters are left as default.
- Download of all the selected PDB files after filtering them.

The result of the FASTA query is stored and parsed, removing the header, and turning it into a Seq object for ease of use. This Seq object is used to perform the BLASTp search. The parameters of the search are stored inside a pickle file, and parsed to obtain a dictionary whose keys are the returned experimental structure ids, and whose keys are the chain identifiers of the HSP (High-Scoring Segment Pairs) of the BLAST search. Structures with repeated chains are deleted from the dictionary, and the selected .PDB files are downloaded via the Biopython library.

The structures are then trimmed to contain only the chain containing the HSP from the BLASTp query, and the new .PDB files are stored for later analysis. Their sequences are extracted with Biopython based on query coverage and identity criteria, and will serve as the template for alignment in posterior feature extraction functions.

Two metadata files for the analysis are also generated: A pairwise global alignment of all the structures against the canonical sequence and a large metadata dataframe. See metadata_dataframe in the Analysis folder of the provided Github repository.

This data is obtained simultaneously from the RCSB database using the structure identification codes and from the proper .PDB files. The script will also generate FASTA files from the sequences of the selected chains using a simplified header (> [Structure ID]) for the purpose of performing a phylogenetic tree.

The structures can be classified under the column “Read Activity Status” as “Active”, “Inactive” and “Unknown”, based on the ligand list supplied at script call. Those structures identified as “Unknown” don’t contain enough information on their metadata to be labeled as either active or inactive and are purged from the list of available structures for analysis.

In the case of HRAS P01112, the script will look for following ligands to determine activity:

Active:

- Guanosine Triphosphate (GTP)
- Guanosine-5'-[(β,γ)-imido]triphosphate (GNP), a non-hydrolyzable analog of GTP that blocks HRAS in its active state.

Inactive:

- Guanosine Diphosphate (GDP)
- 5'-Guanosine-diphosphate-monothiophosphate (GDPβS or GSP), a non-hydrolyzable analog of GDP that blocks HRAS in its inactive state.

The retrieval script recovered 554 structures, of which 435 were programmatically selected at the chosen filtering parameters based on bound ligands.

Weighted Contact Number

The extraction is performed through the WCN module by Martin Floor (Floor, M.). It contains a method to directly obtain the WCN values of all alpha carbons from a given .PDB structure object. We first proceeded to feed the structures to the module and store these values as Numpy ndarrays. We performed the alignment stored in the alignment dataframe and associate the values corresponding to the substitutions or deletions in our query (structure) sequences.

Then, we proceed to save the aligned WCN arrays to a dataframe. The structure will be the same as in the previously generated alignment dataframe. Refer to table of dataframes in Methods Section II.

Subsequentially, we generate new .PDB files, substituting the alpha carbons B-factor value for the newly obtained WCN value for visualization with UCSF Chimera.

PropKa

PropKa predicts the pKa values of ionizable groups in proteins based on the 3D structure of the protein (Søndergaard et al., 2011) (Olsson et al., 2011). It can handle standard amino acids and a few non-standard ones as well as ligands.

We iterated over all the selected .PDB files and generate with them Molecular Container objects, using the default options and parameters of the module once the structure files are loaded. The Molecule Container objects are then read and their pKa profiles calculated before being written to a new set of files.

These files once stored have to be parsed to obtain the desired information. The data obtained this way is imputed on the alignment and stored in a dataframe. Refer to Table A2.1.

Pyrosetta

Pyrosetta is a well-known library, derived from the Rosetta project, whose main goal is the computational generation of molecular dynamics simulations for the purposes of molecule analysis (Chaudhury et al., 2010). It contains powerful features and a wide set of proprietary profiles to quantify the properties of proteins, their amino acids and their subsequent components.

We followed a common pipeline for the five metrics obtained from the selected .PDB files. It starts by generating a score function (“ref2015” is used for being the most up-to-date scoring system available to Pyrosetta). A general energy profile is calculated for each file, and then 5 profiles are extracted from them:

- **Total Energy**
- **Van der Waals Attractive**
- **Van der Waals Repulsive**
- **Solvation**
- **Electrostatic**

These profiles are stored in dictionaries, aligned to the canonical sequence, and exported to their respective dataframes. Refer to Table A2.1.

A2.4 - Statistical Control

Apart from the analysis, 2 more analysis were performed on the selected files for statistical analysis: RMSD profiles and 3D coordinates of the alpha carbons for PCA study.

RMSD

We generated all possible pairs (combinations) of .PDB files and stores them as tuples. Then, it will attempt to obtain the Root Square Mean Deviation (RMSD) profile between them.

We then used the Superimposer module from Biopython to perform this calculation. It obtained the alpha carbon identifiers from the .PDB files and performed a pairwise alignment between them. Once aligned, the RMSD value was calculated between them. The structure pairs and their RMSD value were stored in a pickle file for later analysis.

Spatial Coordinates

We performed pairwise global alignment between the canonical sequence and each of the selected .PDB files structures. Then, extracted the coordinates from the alpha carbons and stored them in a list of tuples associated to the structure IDs in a dictionary. That dictionary was stored in a pickle file and a 3D PCA was performed on the coordinates. Refer to the Analysis folder of provided Github repository for more information.

A2.5 - Data Preparation

The next script (Tensor#.py) is a lighter script with two main functions.

First, we performed Z-score normalization and 0 imputation on all null values, and verified the shape of the seven resulting dataframes from the prior script.

Then, we generated all possible combinations without repetition of these dataframes, and stacked them along a new dimension so spatial data on the position of each metric is retained.

Afterwards, we imported all the result dataframes (WCN, Propka, Pyrosetta Total Energy, Pyrosetta Solvation, Pyrosetta Electrostatic, Pyrosetta Van der Waals Attractive, Pyrosetta Van der Waals Repulsive), as well as the alignment dataframe and the metadata dataframe.

The 7 result dataframes were stored in a dictionary whose keys codify the type of analysis they contain. These would later be used to identify the tensors and the models trained on them. The identifier names can be checked in table A3.1 of Annex 3.

A2.6 - Normalization and Verification

In order to preserve the same order of analysis, protein structure IDs were extracted from the index of the alignment dataframe. Its structures were contrasted with those in the other result dataframes to verify the shape is exactly the same in all instances. That was a required step to the posterior creation of the tensors.

Z-score normalization was performed first, not accounting for null values and thus preserving the mean and variance of the statistic from artifacts. The Z-score normalization was performed using an instance of the StandardScaler object from sklearn.

After normalization, 0 imputation was performed on all null values using an instance of the SimpleImputer object from sklearn. The resulting data was then reconverted into dataframes and saved to new .CSV files. Refer to Refer to Table A2.1 for the specific files generated.

A2.7 - Tensor Generation

The script generated and stored as a pickle file a structure identifier. This is a dictionary that unequivocally identifies each structure ID with its index number in the reference alignment dataframe. This would be used later during model creation for labeling purposes.

Then, using the combinations module from itertools, the script looped over all possible combinations without repetition of dataframes to generate all possible tensors.

For identification, the single full combination of all metrics was given a distinct name: “Complete”.

Finally, the tensors were converted to Pytorch format, which enables several powerful operations to be effected on them.

The resulting operation yielded 127 distinct tensors codified by the specific combination of features they contain. Each tensor is a three-dimensional object whose first dimension represents the positions and amino acids of the canonical sequence from the study protein (HRAS P01112), the second dimension represents the selected structure IDs and the third dimension is an array of Z-score normalized metrics corresponding to the particular combination the tensor contains.

These tensors were saved as pickle files.

Model Generation

The data so prepared, tensors were loaded and processed for model training and evaluation. The third script (RF_model.py) takes care of two roles.

First, it flattened the tensors into arrays and applied t-distributed Stochastic Neighbor Embedding (t-SNE) in preparation for the Random Forest model training.

Then, a grid of hyperparameters is defined for the training of the models. The model is trained trying on all possible combinations of hyperparameters using cross-validation.

The resulting models, their training and test metrics and the winning selection of parameters are then stored for later statistical analysis.

Tensor Preprocessing

Random Forest models cannot process 3D objects. Therefore, first, we processed each tensor flattening it using a Pytorch method into an array across the first dimension, so all the data is preserved.

t-SNE of two components was applied to the flattened tensors using the TSNE module from sklearn.

The resulting objects were then stored for model training.

A2.8 - Model Training

A Random Forest model was initialized for each of the t-SNE transformed tensors and trained on them.

The labels were extracted from the metadata dataframe, and the ordering was kept by the structure identifier, both obtained previously.

The algorithm employed a test size of 30%. That is, 30% of the samples were hidden from the training algorithm for testing purposes while the remaining 70% was used to train the model.

The training was performed through the GridSearchCV function from sklearn. using all possible combinations of hyperparameters featuring in the hyperparameter grid.

The grid of hyperparameters is defined like this:

- n_estimators ([10, 50, 100, 200, 500]): Number of trees in the Random Forest
- max_depth ([None, 10, 20, 30, 40, 50]): Maximum depth of the trees.
- min_samples_split ([2, 5, 10, 20]): Minimum number of samples required to split an internal node.
- min_samples_leaf ([1, 2, 4, 8, 16]): Minimum number of samples required to be a leaf node.
- max_features (['sqrt', 'log2', None]): Number of features to consider when looking for the best split, in reference to the total number of features.
- criterion (['gini', 'entropy']): Quality of a split, Gini impurity or entropy information gain.

The algorithm employed cross-validation of 3 folds. That is, it subdivided the training dataset into three subsets, training sequentially the model on each of those segments and validating the model on the remaining subsets. This helps prevent overfitting and provides an estimate of the model´s performance on unseen data.

The training process employed n_jobs = -1. That means it used all available resources to perform the training.

The total for the 3 folds and all possible parameters is 10800 models trained per tensor.

The resulting model was stored as a pickle file. Several metrics were extracted from the models training and performed for later analysis, and also stored into pickle files along the models. These are:

- Accuracy
- Classification Report (To calculate F1)
- Confusion Matrix
- ROC Curve
- ROCAUC
- Precision Recall Curve
- Log Loss
- Cohen’s Kappa
- Matthews Correlation Coefficient
- Brier Score
- Feature Importances
- Best Parameter Selection

For the complete statistical analysis of the top performers, please refer to the Analysis folder in the provided Github repository.

A3 - RESULTS TABLES

Table A3.1 – Reference table for the codes denoting the different combinations of features present in the model-training tensors. Each character or combination of characters in the left side correspond to a feature. The complete identifier code gives the composition of the tensors: w=WCN; p=pKa; t=Total Energy; s=Solvation Profile; e=Electrostatic Profile, v1=Van der Waals Attractive Profile; v2=Van der Waals Repulsive Profile.

FEATURE IDENTIFIER REFERENCE	
w	WCN
p	PropKa
t	Total Energy
s	Solvation Energy
e	Electrostatics
v1	Van der Waals Attractive
v2	Van der Waals Repulsive

Table A3.2 - Ranking of the 15 most important amino acid residues for the WCN profile according to the Random Forest Analysis, from most important (top) to least important (bottom). The Amino Acid Position column contains the position and the type of amino acid present therein. The T-statistic contains the value of t for the two-sample T-test performed between active and inactive samples of the dataset. The FDR Corrected P-Value column contains the p-value of such statistic, corrected for False Discovery Rate.

Amino Acid Position	T-Statistic	FDR Corrected P-Value
pos60:G	13.1762587	1.00E-31
pos35:T	5.82726183	3.01E-08
pos59:A	8.85657624	2.42E-16
pos58:T	7.00705125	5.59E-11
pos12:G	5.87148669	2.83E-08
pos10:G	4.42854065	2.66E-05
pos61:Q	6.26302539	3.84E-09
pos11:A	4.23332139	5.42E-05
pos13:G	2.68718453	0.00940391
pos34:P	0.46992142	0.63868143
pos16:K	3.14051319	0.00272836
pos120:L	-3.42486209	0.0011373
pos62:E	1.83318234	0.07238532
pos39:S	-2.22418878	0.03083573
pos46:I	-2.98495416	0.00411841

Table A3.3 - Ranking of the 15 most important amino acid residues for the pKa profile according to the Random Forest Analysis, from most important (top) to least important (bottom). The Amino Acid Position column contains the position and the type of amino acid present therein. The T-statistic contains the value of t for the two-sample T-test performed between active and inactive samples of the dataset. The FDR Corrected P-Value column contains the p-value of such statistic, corrected for False Discovery Rate.

Amino Acid Position	T-Statistic	FDR Corrected P-Value
pos57:D	-4.03E-01	0.701697053
pos32:Y	4.27E+00	0.000181767
pos40:Y	1.78E+00	0.190870153
pos38:D	4.31E+00	0.000181767
pos16:K	-3.83E-01	0.701697053
pos33:D	1.01E+00	0.469113438
pos143:E	3.22E+00	0.005165992
pos31:E	1.06E+00	0.469113438
pos68:R	-4.179209152	0.000181767
pos96:Y	-0.756416403	0.552238577
pos97:R	2.623447708	0.027167931
pos141:Y	1.4198444	0.293390622
pos102:R	1.464023741	0.293390622
pos95:Q	0.810572545	0.552238577
pos47:D	0.709247829	0.552238577

Table A3.4 - Ranking of the 15 most important amino acid residues for the Total Energy profile according to the Random Forest Analysis, from most important (top) to least important (bottom). The Amino Acid Position column contains the position and the type of amino acid present therein. The T-statistic contains the value of t for the two-sample T-test performed between active and inactive samples of the dataset. The FDR Corrected P-Value column contains the p-value of such statistic, corrected for False Discovery Rate.

Amino Acid Position	T-Statistic	FDR Corrected P-Value
pos58:T	-4.337606816	0.00027738
pos64:Y	-2.508162595	0.04730029
pos32:Y	2.40319861	0.05019641
pos81:V	-1.284273555	0.2997461
pos33:D	1.292770271	0.2997461
pos60:G	0.844657592	0.49854364
pos55:I	-2.799424964	0.04035924
pos59:A	-0.162384663	0.87108949
pos17:S	-0.772355982	0.50813651
pos34:P	-1.224188894	0.30223828

pos83:A	-1.568979127	0.25175847
pos46:I	2.119191294	0.0868102
pos56:L	0.639154637	0.56047333
pos35:T	2.506439832	0.04730029
pos16:K	1.306121713	0.2997461

Table A3.5 - Ranking of the 15 most important amino acid residues for the Solvation profile according to the Random Forest Analysis, from most important (top) to least important (bottom). The Amino Acid Position column contains the position and the type of amino acid present therein. The T-statistic contains the value of t for the two-sample T-test performed between active and inactive samples of the dataset. The FDR Corrected P-Value column contains the p-value of such statistic, corrected for False Discovery Rate.

Amino Acid Position	T-Statistic	FDR Corrected P-Value
pos36:I	-7.66708813	2.24E-12
pos10:G	2.643019777	0.01283681
pos33:D	-4.502481276	2.24E-05
pos57:D	4.866433293	5.01E-06
pos12:G	4.106845424	0.000105356
pos17:S	3.904239015	0.00020977
pos39:S	-5.093304819	2.91E-06
pos16:K	4.920105799	4.85E-06
pos60:G	1.177807041	0.239612581
pos73:R	3.277225434	0.00190848
pos19:L	1.204955081	0.239612581
pos35:T	-1.515775107	0.150472839
pos118:C	-5.083982003	2.91E-06
pos74:T	-1.610326266	0.135198951
pos13:G	2.175013003	0.041245666

Table A3.6 - Ranking of the 15 most important amino acid residues for the Electrostatic profile according to the Random Forest Analysis, from most important (top) to least important (bottom). The Amino Acid Position column contains the position and the type of amino acid present therein. The T-statistic contains the value of t for the two-sample T-test performed between active and inactive samples of the dataset. The FDR Corrected P-Value column contains the p-value of such statistic, corrected for False Discovery Rate.

Amino Acid Position	T-Statistic	FDR Corrected P-Value
pos60:G	4.220586903	0.000123179
pos35:T	6.361850912	4.31E-09

pos33:D	6.964539952	2.19E-10
pos70:Q	3.723446753	0.000485045
pos59:A	3.363153474	0.001415214
pos37:E	3.887159104	0.000299298
pos32:Y	2.173189012	0.041433874
pos10:G	-2.107620594	0.04465011
pos28:F	0.916893767	0.415132358
pos38:D	-3.406682315	0.00136504
pos16:K	3.141063653	0.002723366
pos143:E	-4.150155479	0.000123179
pos99:Q	-0.363568164	0.716383216
pos80:C	-0.394586502	0.716383216
pos56:L	-4.202160699	0.000123179

Table A3.7 - Ranking of the 15 most important amino acid residues for the Attractive Van der Waals profile according to the Random Forest Analysis, from most important (top) to least important (bottom). The Amino Acid Position column contains the position and the type of amino acid present therein. The T-statistic contains the value of t for the two-sample T-test performed between active and inactive samples of the dataset. The FDR Corrected P-Value column contains the p-value of such statistic, corrected for False Discovery Rate.

Amino Acid Position	T-Statistic	FDR Corrected P-Value
pos33:D	8.121724256	9.81E-14
pos12:G	-2.274192649	0.029391001
pos35:T	-3.74960071	0.000511909
pos60:G	-1.74E+00	0.096233231
pos32:Y	2.83E+00	0.007392506
pos36:I	3.55155159	0.000924024
pos10:G	-0.910838445	0.362959432
pos39:S	3.826063233	0.000459318
pos13:G	-1.411764664	0.170185422
pos64:Y	-6.359388365	4.38E-09
pos59:A	-2.887402309	0.006845101
pos16:K	-3.824618265	0.000459318
pos72:M	-5.600229605	2.06E-07
pos17:S	2.597354407	0.01330966

pos57:D	-2.966499547	0.00600565

Table A3.8 - Ranking of the 15 most important amino acid residues for the Repulsive Van der Waals profile according to the Random Forest Analysis, from most important (top) to least important (bottom). The Amino Acid Position column contains the position and the type of amino acid present therein. The T-statistic contains the value of t for the two-sample T-test performed between active and inactive samples of the dataset. The FDR Corrected P-Value column contains the p-value of such statistic, corrected for False Discovery Rate.

Amino Acid Position	T-Statistic	FDR Corrected P-Value
pos33:D	-2.97E+00	0.015590989
pos64:Y	4.103616816	0.00074742
pos39:S	-1.60736086	0.148374494
pos17:S	-2.711490937	0.026261311
pos12:G	0.157134416	0.87522259
pos27:H	-1.650463369	0.148374494
pos81:V	-1.135892502	0.275056414
pos9:V	3.405340781	0.00548625
pos21:I	-1.302372837	0.223361343
pos60:G	-1.31E+00	0.223361343
pos107:D	-1.911909019	0.09440269
pos35:T	2.005315949	0.09440269
pos120:L	-2.31E+00	0.06412307
pos72:M	2.235492235	0.064917068
pos71:Y	1.9286393	0.09440269

Table A3.9 – Ranking of the 5 best-performing models according to ROCAUC Score, from best performing (top) to worst performing (bottom). The names codify which features the model was trained on: w=WCN; p=pKa; t=Total Energy; s=Solvation Profile; e=Electrostatic Profile, v1=Van der Waals Attractive Profile; v2=Van der Waals Repulsive Profile.

Top 5 Feature Combinations by ROCAUC	
Feature Combination	**ROCAUC**
v1v2	0.8654
wsev1	0.8359
ptsv2	0.8211

psv1	0.8067
ptev1	0.8043

Table A3.10 – Ranking of the 5 best-performing models according to Accuracy, from best performing (top) to worst performing (bottom). The names codify which features the model was trained on: w=WCN; p=pKa; t=Total Energy; s=Solvation Profile; e=Electrostatic Profile, v1=Van der Waals Attractive Profile; v2=Van der Waals Repulsive Profile.

Top 5 Feature Combinations by Accuracy	
Feature Combination	**Accuracy**
v1v2	0.8086
ptev1	0.7913
wsev1	0.7826
ptsv2	0.7739
psv1	0.7652

Table A3.11 – Ranking of the 5 best-performing models according to Log Loss, from best performing (top) to worst performing (bottom). The names codify which features the model was trained on: w=WCN; p=pKa; t=Total Energy; s=Solvation Profile; e=Electrostatic Profile, v1=Van der Waals Attractive Profile; v2=Van der Waals Repulsive Profile.

Top 5 Feature Combinations by Log Loss	
Feature Combination	**Log Loss**
v1v2	0.4744
wsev1	0.4836
ptsv2	0.5313
psv1	0.5392
ptev1	0.5661

Table A3.12 – Ranking of the 5 best-performing models according to Cohen's Kappa, from best performing (top) to worst performing (bottom). The names codify which features the model was trained on: w=WCN; p=pKa; t=Total Energy; s=Solvation Profile; e=Electrostatic Profile, v1=Van der Waals Attractive Profile; v2=Van der Waals Repulsive Profile.

Top 5 Feature Combinations by Cohen's Kappa	
Feature Combination	**Cohen's Kappa**
v1v2	0.6164
ptev1	0.5808

wsev1	0.5643
ptsv2	0.5492
psv1	0.5295

Table A3.13 – Ranking of the 5 best-performing models according to Matthew's Correlation Coefficient, from best performing (top) to worst performing (bottom). The names codify which features the model was trained on: w=WCN; p=pK_a; t=Total Energy; s=Solvation Profile; e=Electrostatic Profile, v1=Van der Waals Attractive Profile; v2=Van der Waals Repulsive Profile.

Top 5 Feature Combinations by Matthew's Correlation Coefficient	
Feature Combination	**Matthew's Correlation Coefficient**
v1v2	**0.6179**
ptev1	**0.5866**
wsev1	**0.5651**
ptsv2	**0.5545**
psv1	**0.5302**

RESOURCES AND CODE AVAILABILITY

For the full computational pipeline code, further information, analysis and resources, see the Github repository at JorgeAndOmics/HRAS-Activation-Classifier (github.com).

STRUCTURE REFERENCES

For the complete list of PDB structures used in this project, please refer to the Analysis folder in the provided Github repository. In the metadata dataframe are available the structure identifiers, their title and their abstract, the later when accessible.

REFERENCES

Bombarda, E., & Ullmann, G. M. (2010). pH-Dependent p K_a Values in Proteins—A Theoretical Analysis of Protonation Energies with Practical Consequences for Enzymatic Reactions. *The Journal of Physical Chemistry B*, *114*(5), 1994–2003. https://doi.org/10.1021/jp908926w

Bonet, J., Caltabiano, G., Khan, A. K., Johnston, M. A., Corbí, C., Gómez, À., Rovira, X., Teyra, J., & Villà-Freixa, J. (2005). The role of residue stability in transient protein-protein interactions involved in enzymatic phosphate hydrolysis. A computational study. *Proteins: Structure, Function, and Bioinformatics*, *63*(1), 65–77. https://doi.org/10.1002/prot.20791

Chaudhury, S., Lyskov, S., & Gray, J. J. (2010). PyRosetta: A script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics*, *26*(5), 689–691. https://doi.org/10.1093/bioinformatics/btq007

Downward, J. (2003). Targeting RAS signalling pathways in cancer therapy. *Nature Reviews Cancer*, *3*(1), 11–22. https://doi.org/10.1038/nrc969

Fetics, S., Young, M., Buhrman, G., & Mattos, C. (2010). *Allosteric modulation of H-Ras GTPase*.

Floor, M. (n.d.). *WCN*. GitHub. Retrieved 6 September 2023, from https://github.com/Martin-Floor/WCN

Fossat, M. J., Posey, A. E., & Pappu, R. V. (2021). Quantifying charge state heterogeneity for proteins with multiple ionizable residues. *Biophysical Journal*, *120*(24), 5438–5453. https://doi.org/10.1016/j.bpj.2021.11.2886

Gallicchio, E., & Levy, R. M. (2011). Advances in all atom sampling methods for modeling protein–ligand binding affinities. *Current Opinion in Structural Biology*, *21*(2), 161–166. https://doi.org/10.1016/j.sbi.2011.01.010

Gao, H.-Y., Wagner, H., Held, P. A., Du, S., Gao, H.-J., Studer, A., & Fuchs, H. (2015). In-plane Van der Waals interactions of molecular self-assembly monolayer. *Applied Physics Letters*, *106*(8), 081606. https://doi.org/10.1063/1.4907777

Gremer, L., De Luca, A., Merbitz-Zahradnik, T., Dallapiccola, B., Morlot, S., Tartaglia, M., Kutsche, K., Ahmadian, M. R., & Rosenberger, G. (2010). Duplication of Glu37 in the switch I region of HRAS impairs effector/GAP binding and underlies Costello syndrome by promoting enhanced growth factor-dependent

MAPK and AKT activation. *Human Molecular Genetics*, *19*(5), 790–802. https://doi.org/10.1093/hmg/ddp548

Hähl, H., Evers, F., Grandthyll, S., Paulus, M., Sternemann, C., Loskill, P., Lessel, M., Hüsecken, A. K., Brenner, T., Tolan, M., & Jacobs, K. (2012). Subsurface Influence on the Structure of Protein Adsorbates as Revealed by in Situ X-ray Reflectivity. *Langmuir*, *28*(20), 7747–7756. https://doi.org/10.1021/la300850g

Haq, S. R., Jürgens, M. C., Chi, C. N., Koh, C.-S., Elfström, L., Selmer, M., Gianni, S., & Jemth, P. (2010). The Plastic Energy Landscape of Protein Folding. *Journal of Biological Chemistry*, *285*(23), 18051–18059. https://doi.org/10.1074/jbc.M110.110833

Hermann, J., DiStasio, R. A., & Tkatchenko, A. (2017). First-Principles Models for van der Waals Interactions in Molecules and Materials: Concepts, Theory, and Applications. *Chemical Reviews*, *117*(6), 4714–4758. https://doi.org/10.1021/acs.chemrev.6b00446

Hudáky, P., Jákli, I., Császár, A. G., & Perczel, A. (2001). Peptide models XXXI. Conformational properties of hydrophobic residues shaping the core of proteins. An *ab initio* study of N-formyl- L -valinamide and N-formyl- L -phenylalaninamide: Hydrophobic Residues Shaping the Core of Proteins. *Journal of Computational Chemistry*, *22*(7), 732–751. https://doi.org/10.1002/jcc.1040

Ilter, M., & Sensoy, O. (2019). Catalytically Competent Non-transforming H-RASG12P Mutant Provides Insight into Molecular Switch Function and GAP-independent GTPase Activity of RAS. *Scientific Reports*, *9*(1), 10967. https://doi.org/10.1038/s41598-019-47481-1

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Krusemark, C. J., Frey, B. L., Belshaw, P. J., & Smith, L. M. (2009). Modifying the charge state distribution of proteins in electrospray ionization mass spectrometry by chemical derivatization. *Journal of the American Society for Mass Spectrometry*, *20*(9), 1617–1625. https://doi.org/10.1016/j.jasms.2009.04.017

Lin, C.-P., Huang, S.-W., Lai, Y.-L., Yen, S.-C., Shih, C.-H., Lu, C.-H., Huang, C.-C., & Hwang, J.-K. (2008). Deriving protein dynamical properties from weighted protein contact number: Derivation of Protein Dynamical Properties. *Proteins: Structure, Function, and Bioinformatics*, *72*(3), 929–935. https://doi.org/10.1002/prot.21983

Lingner, T., Alcántara, R., Villà i Freixa, J., & Notredame, C. (2010). Evaluating dynamics conservation in the RAS family. *Unpublished*.

Linse, S., Brodin, P., Johansson, C., Thulin, E., Grundström, T., & Forsén, S. (1988). The role of protein surface charges in ion binding. *Nature*, *335*(6191), 651–652. https://doi.org/10.1038/335651a0

Olsson, M. H. M., Søndergaard, C. R., Rostkowski, M., & Jensen, J. H. (2011). PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical p K_a Predictions. *Journal of Chemical Theory and Computation*, *7*(2), 525–537. https://doi.org/10.1021/ct100578z

Ou, G., He, B., & Halling, P. (2016). Ionization basis for activation of enzymes soluble in ionic liquids. *Biochimica et Biophysica Acta (BBA) - General Subjects*, *1860*(7), 1404–1408. https://doi.org/10.1016/j.bbagen.2016.04.004

Persson, B. A., Jönsson, B., & Lund, M. (2009). Enhanced Protein Steering: Cooperative Electrostatic and van der Waals Forces in Antigen−Antibody Complexes. *The Journal of Physical Chemistry B*, *113*(30), 10459–10464. https://doi.org/10.1021/jp904541g

Rhett, J. M., Khan, I., & O'Bryan, J. P. (2020). Biology, pathology, and therapeutic targeting of RAS. In *Advances in Cancer Research* (Vol. 148, pp. 69–146). Elsevier. https://doi.org/10.1016/bs.acr.2020.05.002

Søndergaard, C. R., Olsson, M. H. M., Rostkowski, M., & Jensen, J. H. (2011). Improved Treatment of Ligands and Coupling Effects in Empirical Calculation and Rationalization of p K_a Values. *Journal of Chemical Theory and Computation*, *7*(7), 2284–2295. https://doi.org/10.1021/ct200133y

Sorokina, I., Mushegian, A. R., & Koonin, E. V. (2022). Is Protein Folding a Thermodynamically Unfavorable, Active, Energy-Dependent Process? *International Journal of Molecular Sciences*, *23*(1), 521. https://doi.org/10.3390/ijms23010521

Tahir, M., & Hayat, M. (2017). Machine learning based identification of protein–protein interactions using derived features of physiochemical properties and evolutionary profiles. *Artificial Intelligence in Medicine*, *78*, 61–71. https://doi.org/10.1016/j.artmed.2017.06.006

Tomar, D. S., Weber, V., Pettitt, B. M., & Asthagiri, D. (2016). Importance of Hydrophilic Hydration and Intramolecular Interactions in the Thermodynamics of Helix–Coil Transition and Helix–Helix Assembly in a Deca-Alanine Peptide. *The Journal of Physical Chemistry B*, *120*(1), 69–76. https://doi.org/10.1021/acs.jpcb.5b09881