

# **SPIKE RBD-ACE2 PROTEIN-PROTEIN INTERFACE ANALYSIS**

Núria Cardona, Antonio Arcas & Ricard Garcia

---

A complex comprising the Receptor Binding Domain (RBD) of the SARS-CoV-2 Spike protein and its receptor, the Angiotensin Converting Enzyme (ACE2), was furnished to us. The primary objective of this study was to assess the relative contribution of interface residues to the interaction energy within a protein-protein complex, specifically in the context of the RBD-ACE2 interaction.

## 1. Introduction

The exponential growth of genomic data from large-scale sequencing projects has produced a significant number of protein sequence variants. Many proteins operate by forming permanent or temporary complexes with other macromolecules, and these complexes often play a central role in controlling the activity of the involved proteins. Consequently, there is a growing interest in comprehending the impact of sequence variation on the stability of such complexes.

In this study, we embark on an initial exploration of this complex landscape by assessing the relative contributions of amino acid residues that constitute protein-protein interfaces to the interaction energies between the components of these complexes. Our specific focus is on the complex formed between the **receptor-binding domain (RBD)** of the SARS-CoV-2 spike protein and its receptor, **angiotensin-converting enzyme (ACE2)**. The formation of this complex represents a critical step in viral infection, and disrupting it is a key goal of many existing vaccines aimed at preventing viral entry.

## 2. Summary

This scientific analysis delves into the structural intricacies of the Receptor Binding Domain (RBD) of the SARS-CoV-2 spike protein, specifically focusing on the complex represented by PDB structure with ID 6m0j. Our primary goal is to solve the complex interactions between sequence variants, protein-protein interfaces, and their consequential effects on the stability of these crucial complexes.

To achieve this objective, we employ a strategic approach. Initially, we identify the amino acid residues constituting the interface between the complex components, leveraging the structural information from 6m0j. Subsequently, we conduct a meticulous analysis of the stability contribution of each amino acid residue. This involves a traditional Ala-scanning experiment, where each residue is individually replaced with Ala, and the resulting changes in the complex interaction energy are evaluated.

### 3. Methodology

Our study is based on calculating the total energy of the protein-protein interaction. This energy is calculated using the following formula:

$$\Delta G^{A-B} = \Delta G_{elect}^{A-B} + \Delta G_{vdw}^{A-B} + \Delta G_{solv}^{A-B} - \Delta G_{solv}^A - \Delta G_{solv}^B$$

Where the first element corresponds to the electrostatics of the interaction, the second element are the Van Der Waals of the interaction, the third element is the solvation of the interaction and the last element is the solvation of the individual subunits.

Therefore, all our prior work is focused on strictly calculating each of the elements of the formula in order to obtain final results and draw various conclusions depending on the specific case we are studying. More precisely, the followed steps are:

#### **PREPARATION & DATA CURATION**

In initiating our energy analysis, the first step involved acquiring and refining the necessary primary data. Concretely, the structure (6m0j) was sourced from the Protein Data Bank (PDB), which is a database of structures of biological molecules (biomolecules). To enhance the raw structure, we excluded non-contributing chains and removed heteroatoms. Subsequently, a thorough quality check was conducted, followed by the addition of missing side-chains, hydrogen atoms, and atom charges. At the end a clean structure was obtained, consisting only of the residues of both chains. This allowed us to compute energies involved in the formation of the complex in a straightforward manner. Moreover, it also facilitated the inspection of the complex using the molecular visualization system PyMOL.

#### **DETERMINE INTERFACE RESIDUES**

Following data curation, we defined the protein-protein interface by establishing a list of residues on both chains with at least one atom below a specified distance. The protein-protein interface represents the interaction regions between proteins within a complex, with the distance criterion determined using PyMOL. In our case, this distance criterion is 4 Å.

In order to determine this distance cutoff, we inspected the complex structure with PyMOL. We colored both chains differently and we manually selected the residues belonging to the angiotensin-converting enzyme **ACE2** that were interacting with the **RBD** of the SARS-CoV-2 spike protein. Then, we computed the distance of all of those bonds that formed the complex and we got that the maximum bond distance was 2.9 Å. However, we set our cutoff at 4 Å to ensure that we were also checking neighboring residues that may also be interacting or helping with the interaction between both chains.

## **COMPUTING INTERACTION ENERGIES**

Once the interface residue list was established, we computed the interaction energy among chains. This involved calculating the difference between the total energy of each chain in the bound state (complex) and the unbound state (isolated chains in solution). The calculation was performed twice: first considering all residues and then focusing solely on those within the interface region, yielding insightful conclusions.

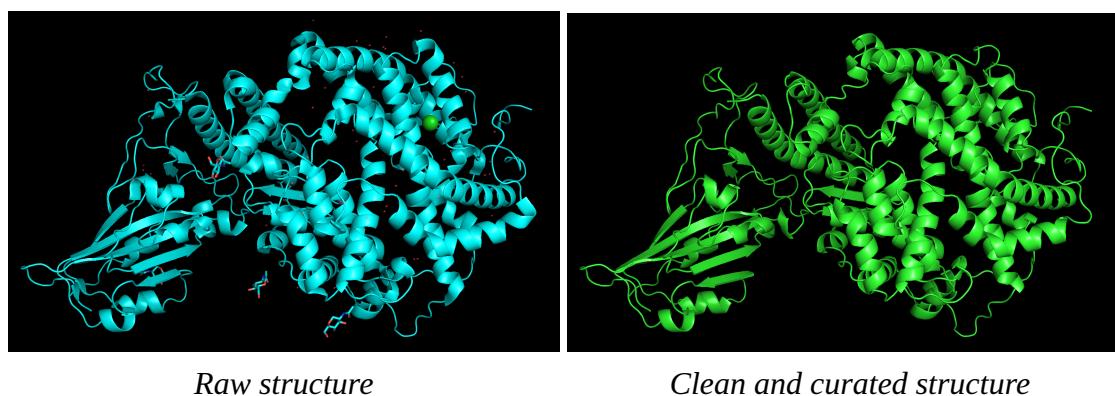
## **ALANINE-SCANNING**

To evaluate the significance of each interface residue, we employed Alanine-Scanning, a method involving the substitution of each residue with alanine (Ala). This technique, leveraging alanine's structurally simple and smaller side chain, allowed us to observe the impact of substitutions on the protein's function or structure, facilitating the identification of key residues crucial for interface stability.

## **4. Results**

### **PREPARATION & DATA CURATION**

At the left side and colored in blue the raw structure is represented. On the other side, in green, is the structure obtained after running the section of data curation and preparation.



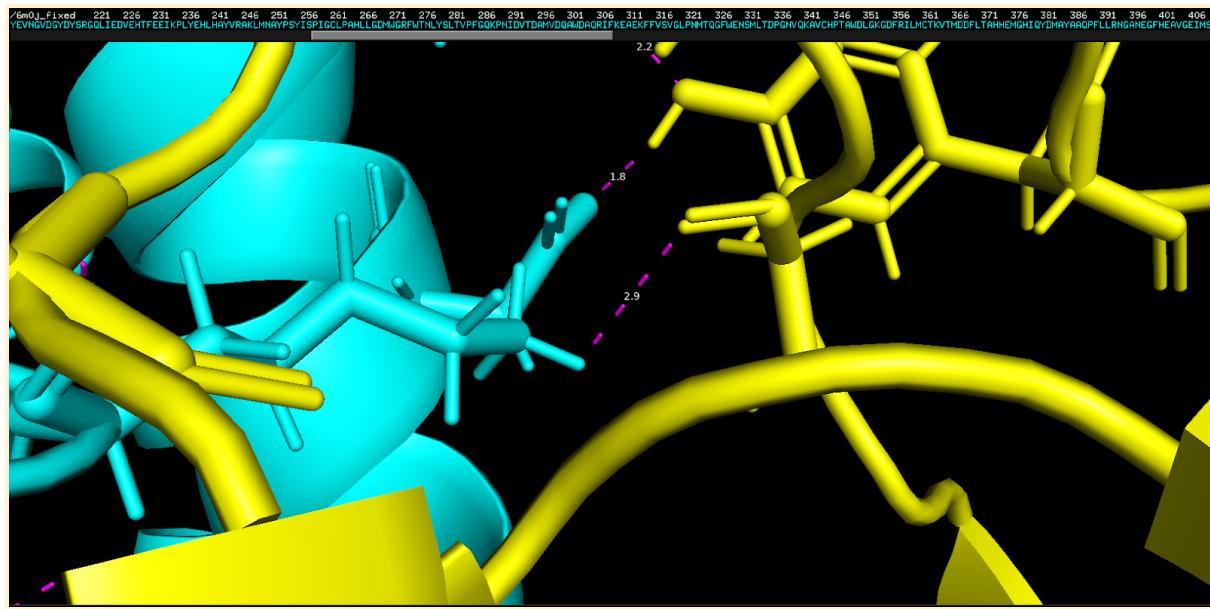
Mainly all differences can be spotted at first sight. After curation, the complex is essentially the same except for the fact that we have:

- removed all heteroatoms and surrounding molecules such as water and other ligands.
- checked and allocated residues with alternative coordinates.
- removed unusual contacts involving amide atoms.
- restored two missing backbone atoms.
- added missing hydrogen atoms.

As a result of all those changes, we got the final structure in green. This process can be reproduced with other proteins.

## DETERMINE INTERFACE RESIDUES

As previously mentioned, a visual inspection of the curated structure in PyMOL allowed to determine that a distance of 2.9 Å included all contact residues. Yet, this value was rounded to 4 Å in order to include other potentially relevant residues.



After loading the clean structure and processing it (i.e. applying a function that takes as parameter the structure and the 4 Å cutoff distance), the interface residues are returned in a dictionary with two keys (one for each chain). As can be seen in the image below, the associated value to each key is a set containing information about the interface residues located in that chain (i.e. details about the residues forming the interface of chain A (ACE2) and E (RBD), respectively).

```
In [44]: interface
Out[44]: {'A': {<Residue GLN het= resseq=24 icode= >,
 <Residue THR het= resseq=27 icode= >,
 <Residue PHE het= resseq=28 icode= >,
 <Residue ASP het= resseq=30 icode= >,
 <Residue LYS het= resseq=31 icode= >,
 <Residue HIE het= resseq=34 icode= >,
 <Residue GLU het= resseq=35 icode= >,
 <Residue GLU het= resseq=37 icode= >,
 <Residue ASP het= resseq=38 icode= >,
 <Residue TYR het= resseq=41 icode= >,
 <Residue GLN het= resseq=42 icode= >,
 <Residue LEU het= resseq=79 icode= >,
 <Residue MET het= resseq=82 icode= >,
 <Residue TYR het= resseq=83 icode= >,
 <Residue ASN het= resseq=330 icode= >,
 <Residue LYS het= resseq=353 icode= >,
 <Residue GLY het= resseq=354 icode= >,
 <Residue ASP het= resseq=355 icode= >,
 <Residue ARG het= resseq=357 icode= >,
 <Residue ARG het= resseq=393 icode= >},
 'E': {<Residue LYS het= resseq=417 icode= >,
 <Residue GLY het= resseq=446 icode= >,
 <Residue TYR het= resseq=449 icode= >,
 <Residue TYR het= resseq=453 icode= >,
 <Residue LEU het= resseq=455 icode= >,
 <Residue PHE het= resseq=456 icode= >,
 <Residue ALA het= resseq=475 icode= >,
 <Residue PHE het= resseq=486 icode= >,
 <Residue ASN het= resseq=487 icode= >,
 <Residue TYR het= resseq=489 icode= >,
 <Residue GLN het= resseq=493 icode= >,
 <Residue GLY het= resseq=496 icode= >,
 <Residue GLN het= resseq=498 icode= >,
 <Residue THR het= resseq=500 icode= >,
 <Residue ASN het= resseq=501 icode= >,
 <Residue GLY het= resseq=502 icode= >,
 <Residue TYR het= resseq=505 icode= >}}
```

It's worth mentioning that the clean structure is loaded using Biopython, a python package meant for biological analysis in which structures are an object containing one or several models. In our case, only one model was provided. Inside this single model, the different chains (as well as their associated residues) can be found.

### **COMPUTING INTERACTION ENERGIES**

This is the output generated after calculating the energies involved in the complex for the interaction residues only:

	<b>Energy</b>	<b>Value</b>
<b>1</b>	Total Elec Int.	<b>8.7271</b>
<b>2</b>	Total Vdw Int.	<b>-139.4030</b>
<b>3</b>	Total Solv AB	<b>0.6690</b>
<b>4</b>	Total Solv A	<b>-25.9476</b>
<b>5</b>	Total Solv E	<b>24.2248</b>
<b>6</b>	DGintAB-A-B	<b>-128.2841</b>

And this is the calculation of energies involved in the complex of all residues:

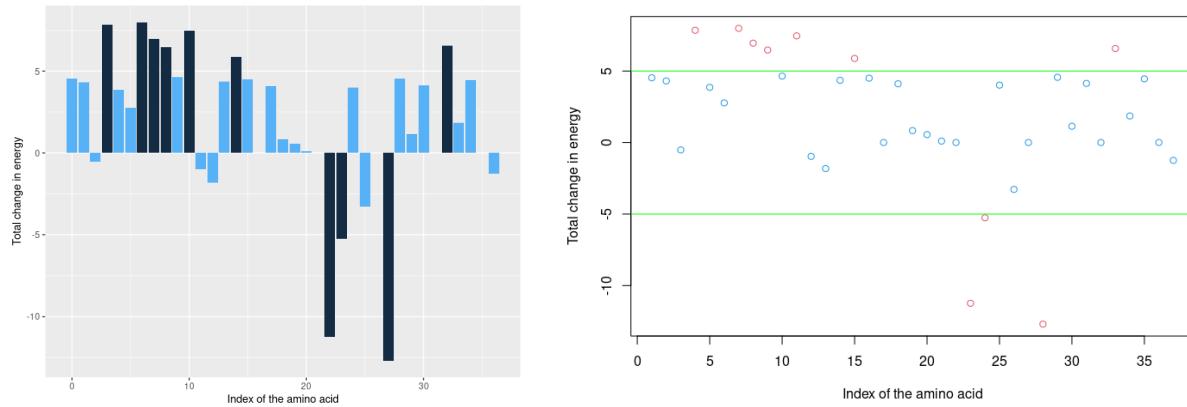
	<b>Energy</b>	<b>Value</b>
<b>1</b>	Total Elec Int.	<b>48.7068</b>
<b>2</b>	Total Vdw Int.	<b>-175.9310</b>
<b>3</b>	Total Solv AB	<b>-518.4848</b>
<b>4</b>	Total Solv A	<b>-423.5458</b>
<b>5</b>	Total Solv E	<b>-97.7981</b>
<b>6</b>	DGintAB-A-B	<b>-124.3650</b>

### **ALANINE-SCANNING**

Finally, the first ten lines of the output generated in the alanine scanning section are the following energies:

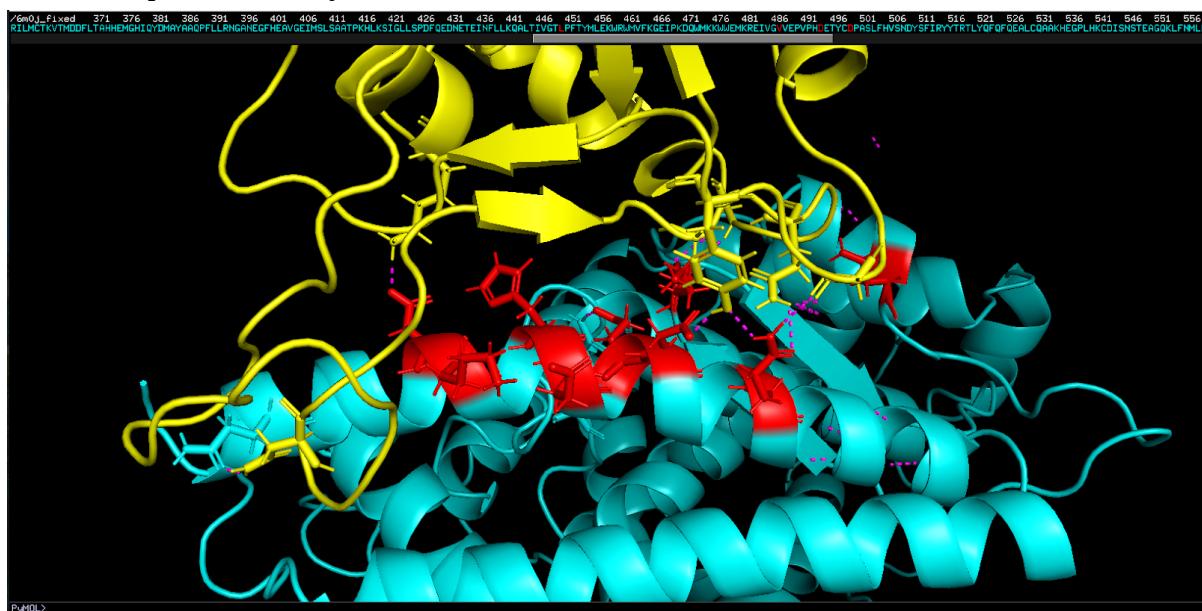
<b>1</b>	GLN	A24	-0.4788	0.8239	1.6975	2.4856	4.5282
<b>2</b>	THR	A27	0.9489	2.8536	0.7679	-0.2643	4.3062
<b>3</b>	PHE	A28	0.0596	1.3121	-0.4867	-1.4038	-0.5189
<b>4</b>	ASP	A30	0.6242	2.2177	1.6563	3.3476	7.8458
<b>5</b>	LYS	A31	-3.2739	4.4258	0.2075	2.4986	3.8580
<b>6</b>	HIE	A34	0.1352	-2.0487	1.6662	3.0157	2.7683
<b>7</b>	GLU	A35	2.0584	1.3245	1.9700	2.6278	7.9807
<b>8</b>	GLU	A37	2.7909	1.9790	0.9107	1.2671	6.9476
<b>9</b>	ASP	A38	1.6613	1.5816	0.8174	2.4061	6.4663
<b>10</b>	TYR	A41	0.2241	6.0650	0.0000	-1.6475	4.6416

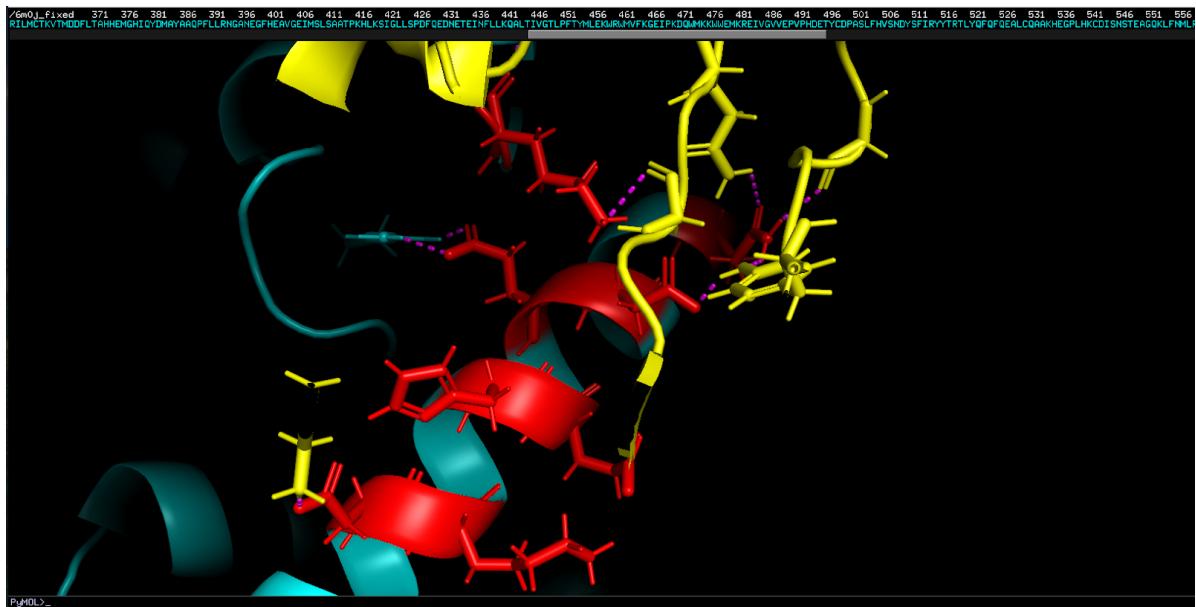
Using the last column of the previous table (which corresponds to the total change in energy observed when exchanging each interface residue by an alanine), two different graphical representations have been generated:



In the plot at the left the bars in dark blue represent residues whose associated total change in energy in the alanine scanning (in absolute value) is higher than 5. To the contrary, bars in light blue correspond to residues with change in energy smaller than 5 (also in absolute value), meaning that they are less important for the interaction that the ones in dark blue. A similar view is seen in the graph at the right, where instead of bars representing each amino acid's energy change there are dots. In this case, the 2 horizontal lines represent the up and low threshold energies, so that the residues outside the middle region are considered to be the most important ones.

Next, the amino acids which were above threshold (i.e. the residues colored in dark blue in the graph at the left and colored in red in the graph at the right) were colored in red in the structure represented in PyMOL:

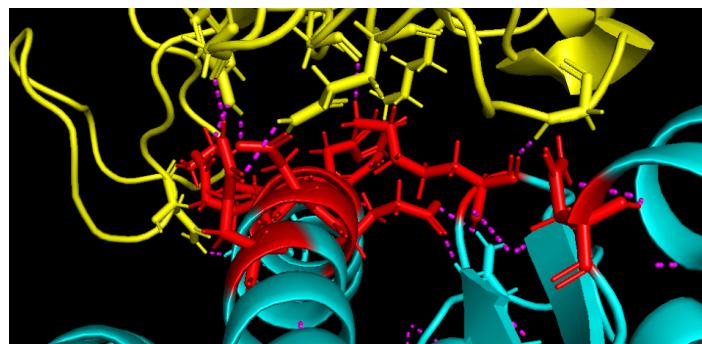




## 5. Conclusions

With the results obtained from the output of our workflow we can extract the following conclusions:

- The total energy considering all the residues and the total energy considering only the interface residues are very similar. This indicates that the interaction residues are the ones who contribute and explain the interaction energy of the complex (these residues are key). The other residues are not energetically relevant for the formation of the complex.



*Interaction residues (colored in red)*

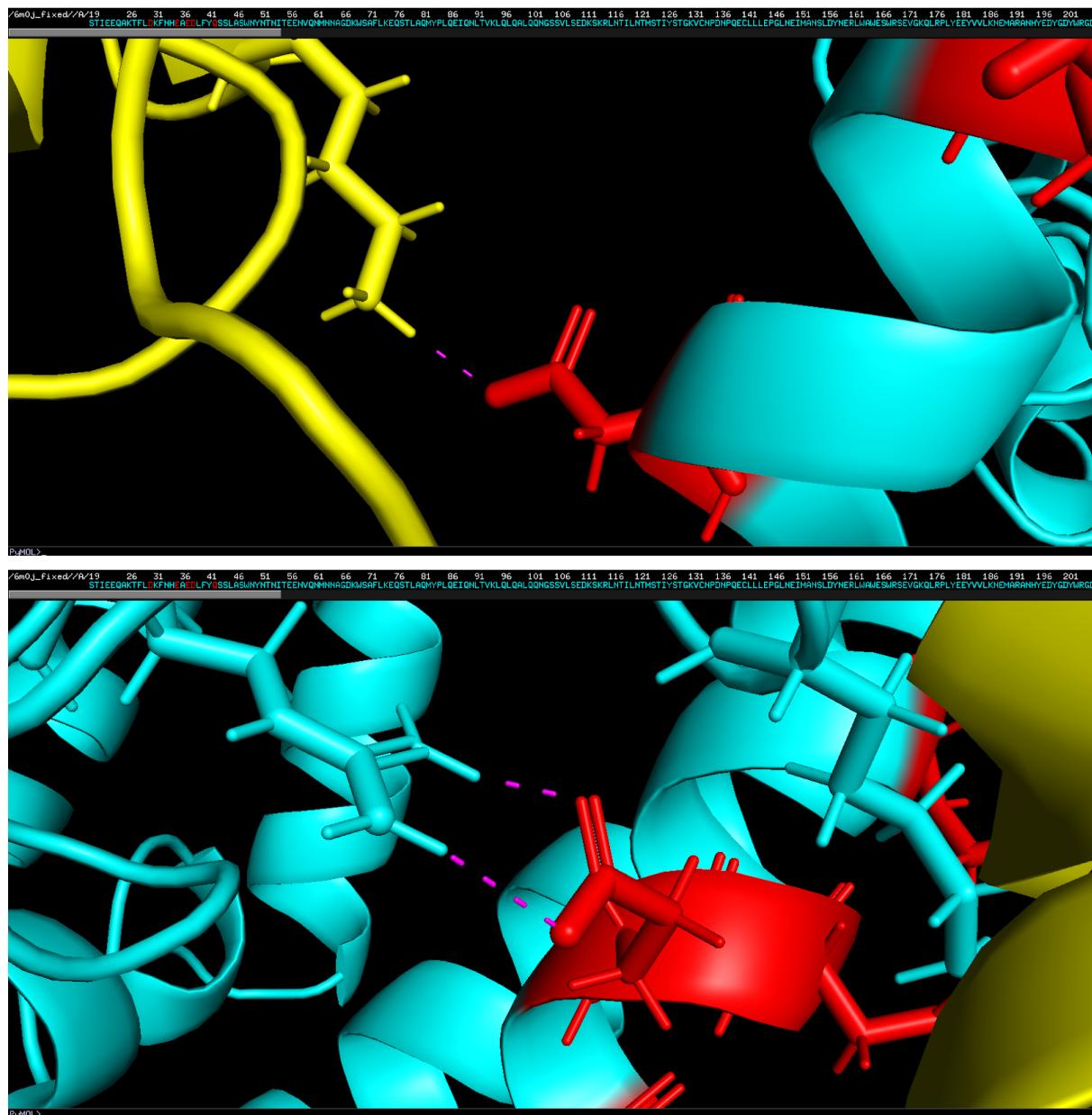
- ALA scanning: there are some residues that after being replaced by ALA, the free energy of the complex changes considerably (see the table below). Consequently, it can be concluded that these key residues have a great impact on the formation of the complex (in energetical and interaction terms). That is, if one of them would be altered, this would have a remarkable impact on the energy of the complex. If we wanted to inhibit or control this complex, we should focus on these amino acids.

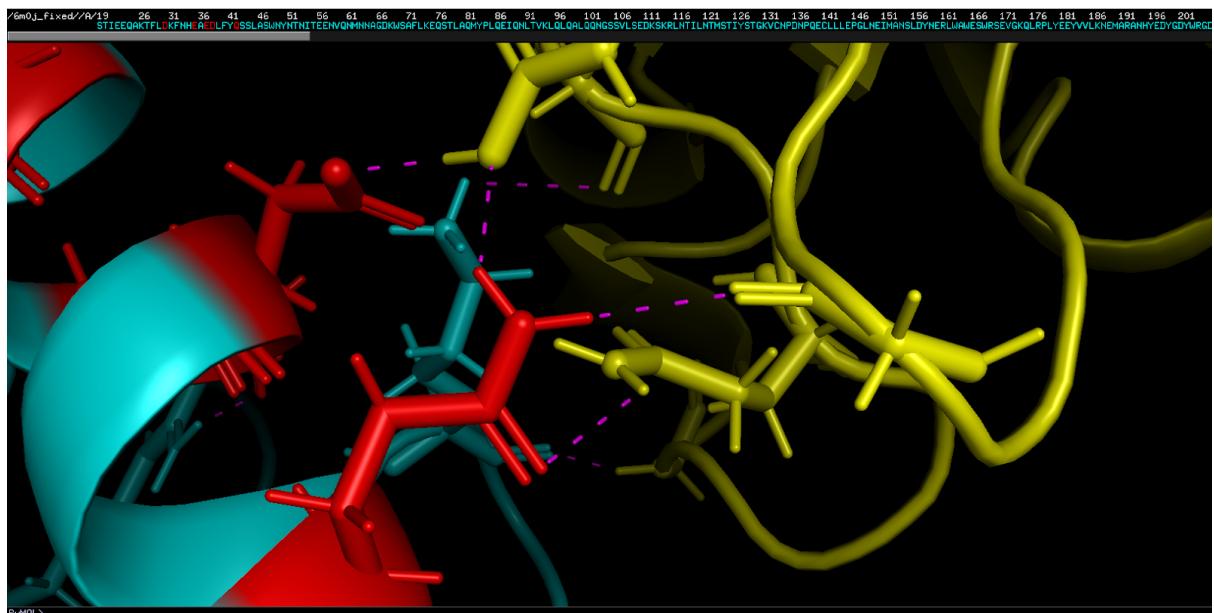
RESIDUE	POSITION	CHANGE IN ENERGY	
ASP	A30	7.8458	→ POLAR (negatively charged)
GLU	A35	7.9807	→ POLAR (negatively charged)
GLU	A37	6.9476	→ POLAR (negatively charged)
ASP	A38	6.4663	→ POLAR (negatively charged)
GLN	A42	7.4560	→ POLAR (neutral)
ASN	A330	5.8774	→ POLAR (neutral)
TYR	E449	-11.2373	→ POLAR (neutral)
TYR	E453	-5.2617	→ POLAR (neutral)
PHE	E486	-12.6984	→ NON-POLAR (neutral)
GLN	E498	6.5714	→ POLAR (neutral)

*Residues with higher energy change in the Ala scanning*

It turns out that all these key residuals (except for PHE E486) are polar, either with a negative charge or neutral.

Here are showcased some of the key residues, with positions A30, A37 and A42 respectively:





The dashed lines represent the polar bonds between residues.