UNIVERSIDAD AUTONOMA DE MADRID

FACULTAD DE MEDICINA

TRABAJO FIN DE MÁSTER

Integración y evaluación de algoritmos para microscopía electrónica

Integration and evaluation of algorithms for electron microscopy

Máster Universitario en Bioinformática y Biología Computacional

Autora: González Matatoros, Sofía.

Director: Sorzano Sánchez, Carlos Óscar.

Departamento de Anatomía, Histología y Neurociencia en Universidad Autónoma de Madrid (UAM).

Departamento de Estructura de Macromoléculas en el Centro Nacional de Biotecnología (CNB).

Director: Kieger, James Michael.

Departamento de Estructura de Macromoléculas en el Centro Nacional de Biotecnología (CNB).

Tutor: Redrejo Rodríguez, Modesto.

Departamento de Bioquímica. Universidad Autónoma de Madrid

CURSO 2023-24
FECHA: Mayo, 2024

Index

1. Abbreviations

- Cryo-EM: Cryo-electron microscopy.
- EMDB: Electron Microscopy Data Bank.
- HAP40: huntingtin-associated protein 40.
- NMR: Nuclear magnetic resonance.
- PDB: Protein Data Bank.
- RMSF: Root Mean Square Deviation.

2. Summary

The biological function of a molecule is highly related to its structure and dynamics. There are numerous techniques to study them, which include cryo-electron microscopy (cryo-EM) and molecular simulations.

Matsumoto et al. (2021) developed and trained a deep neural network called DefMap, which combines data from these two sources to predict local dynamics using data. As this tool can benefit researchers interested in studying protein dynamics, for this project, it has been incorporated into the Scipion framework, which bundles and integrates a variety of software packages for structural biology, to improve its accessibility to potential users.

The plugin created, scipion-em-defmap, includes a protocol in which DefMap is integrated along with a workflow for pre-processing and analysing the results. In addition, to facilitate the interpretation of the results, the plugin includes a visualiser with different options, allowing users to choose the one that best matches their needs. An extra protocol for adapting files has also been created, so that users can benefit from this viewer outside the main protocol.

The plugin was tested on different structures from the Human Huntingtin-HAP40 complex and SARS-CoV-2 Spike glycoprotein. The test concluded that there is a relationship between the plugin output (RMSF) with other measures of variability or uncertainty of the atomic positions, specifically B-factors and, to a lesser extent, local resolutions. Future improvements for the plugin and for the analysis of these variables have also been identified in the discussion.

3. Keywords

DefMap, Cryo-EM, Scipion, Molecular Dynamics.

4. Introduction

In order to understand the molecular mechanisms that allow different biological processes to take place, it is necessary to understand how the molecular components involved behave. Focusing on proteins, several studies have shown that their functions are associated with their three-dimensional (3D) structure and dynamic behaviour, not only from a global perspective of the protein, but also at the level of its constituent atoms (Boehr et al., 2009; Kohen, 2015; Matsumoto et al., 2023).

4.1. Experimental techniques for structural characterisation

In order to obtain atomic structures, there are several experimental techniques that allow the 3D characterisation of proteins. The most common ones are X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy and cryo-electron microscopy (cryo-EM), which each have their advantages and limitations (Table 1).

Table 1

Advantages and disadvantages of experimental techniques for structural characterisation.

Technique	Advantages	Disadvantages
X-ray crystallography	High-resolution structures	Limited dynamic information and it cannot be used with non crystallizable samples.
NMR	High-resolution results, providing information about structure and dynamics.	The output is hard to interpret and dependent on averaging of the signals.
Cryo-EM	High-resolution reconstructions, with information about structure and dynamics.	Dynamic information can be altered by physical and computational factors.

Note. The table shows information about the three most commonly used methods to study the structure of molecules.

X-ray crystallography consists of passing a beam of X-rays through a crystal of the protein under study. The beam is diffracted in several directions and it generates a pattern of intensities, which can be interpreted according to the location of the atoms in the crystal and its symmetry (Smyth & Martin, 2000). This method is often capable of generating high-resolution structures. Nevertheless, due to the methodology for the generation of the crystals, only one type of structure is usually obtained, losing most information about its dynamics. Therefore, for large and dynamic molecules, such as those with many domains that move relative to each other, it is important to complement their output with other methods (Srivastava et al., 2018; Zheng et al., 2015). An example of this would be transmembrane receptors such as viral ones (Lengyel et al., 2014) or immunoglobulin G (Yanaka et al., 2020).

In contrast, in NMR spectroscopy a magnetic field is applied to the sample, causing a change in the spin of the atomic nuclei at different frequencies. As the magnetic field is removed, the nucleus returns to equilibrium, generating an electromagnetic signal, which can be translated into an energy peak in the spectrum. Then, the experimental results will be processed with different techniques to facilitate its interpretation (Libretexts, 2023). As with the previous method, this one generates results with good resolution, but it also provides information about the conformational dynamics of the molecule, by keeping proteins in solutions with near native conditions and assembling different conformations, which can be extracted from the ensemble-averaged observables. However, it presents complications to analyse the results in case of large molecules, therefore, further improvements are still being developed, such as applying chemical transformations like selective isotope labelling, to minimise the signal of many of the atoms. In addition, in order to facilitate its interpretation, some authors combine its results with AI protein structure predictors such as AlphaFold2 and specialised AI for analysing NMR spectra (Shukla et al., 2023).

The last method to mention is the cryo-EM technique, whose outputs are the ones used in this project. The experimental procedure consists of freezing the samples in liquid nitrogen, to fix and protect them before using the electron microscope to record images of the molecule. This protection is applied to avoid damage and variations in the structure of the proteins, due to electron radiation (Murata & Wolf, 2018). One advantage of freezing over crystallising is that it allows more than one type of conformation to be recorded, as it fixes

each particle in their current structural state from the original dynamic ensemble in solution (Wang & Wang, 2017).

After obtaining the images, they are combined to reconstruct one or more average three-dimensional maps of the molecule, which are then used to build atomic models (Vilas et al., 2022). Therefore, single-particle fixation gives an advantage over NMR, since it is less dependent on general averaging and enables more direct characterisation of the structure and dynamics within particular conformational states. In this respect, there have been many advances in algorithms that analyse the conformational heterogeneity of particles. These methods use approaches such as linear and non-linear transformations or deep neural networks, sometimes in combination with structure prediction tools (Tang et al., 2023).

4.2. Resolution concepts of Cryo-EM reconstructions

As mentioned above, Cryo-EM generates reconstructions of the molecules. One of the properties to be taken into account when evaluating a reconstruction is the resolution.

According to Vilas (2019), “resolution describes the degree of detail that an optical system is able to discriminate, the higher resolution the higher quality and details can be seen in the image” (p.53). In practice, resolution is treated as a value that indicates the minimum distance at which we can distinguish two objects; therefore, high resolution corresponds to a low numeric value.

4.3. Advantages and limitations of dynamic information provided by Cryo-EM

Despite having overall high resolution, one interesting feature of these reconstructions is that the resolution varies locally over the map, with some regions having lower local resolutions. This phenomenon may have many causes, one major reason for this is the effect of structural dynamics and class averaging, since more inconsistent positions of atoms make the average of the atoms worse. In consequence, flexible regions tend to have worse resolution in the reconstruction.

Local resolution is therefore related to other measures, such as B-factors and root mean square fluctuation (RMSF) values, which have also been considered in this project. The B-factors represent the relative uncertainty of an atom's position, arising from atomic displacement due to thermal and static vibrations among other factors (Trueblood et al.,

1996), while RMSF values measure the degree to which the positions deviate from the average of a set of structures under study (Bagewadi et al., 2023).

In addition to structural dynamics, further reasons can reduce the local resolution of a region, such as preferred orientations of the molecule, damage at the air-water interface or other sample-related sources (Glaeser, 2018; Li et al., 2021). This last aspect also causes biases during computational image analysis, such as changes in the structure because of the experimental protocol or a bad recognition of the particles, among others (Sorzano et al., 2022). In this regard, computational solutions are still being developed to try to mitigate the effect of these deviations on the results.

4.4. Computational techniques for dynamic characterisation

Given the above mentioned difficulties of capturing structural dynamics from experiments, one frequent alternative is to execute molecular simulations. In these simulations, the position of each atom is calculated as a function of time, according to physical models of atomic interactions (Hollingsworth & Dror, 2018). Nevertheless, such simulations are very costly in terms of time and resources, as all the forces from non-bonded interactions have to be computed (AlRawashdeh & Barakat, 2023; Bock et al., 2023). This issue makes them not feasible for all cases, and generates limitations in the analysis of large molecules with complex assemblies. They also suffer from their own limitations, such as force field inaccuracies and insufficient sampling, due to the limitation on the timestep that often does not give enough time to explore the complete movement of the molecule.

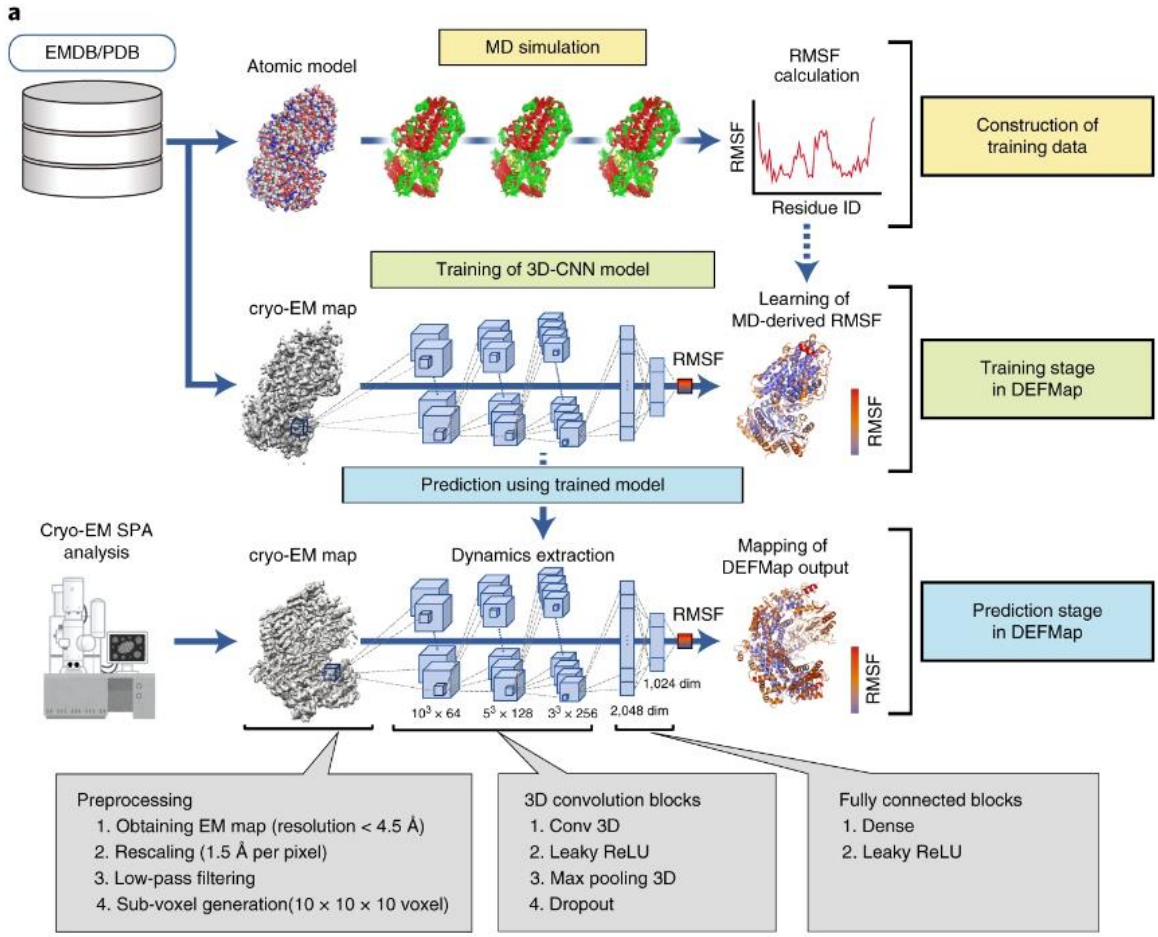
Although there are considerable efforts in overcoming these limitations (Bock et al., 2023; Hénin et al., 2022), some authors opt for combining the output from the molecular simulations with data from different sources, in order to obtain more refined models. It can be retrieved from experimental techniques, such as NMR (Doktorova et al., 2023; Zadorozhnyi et al., 2024) and cryo-EM (Costa et al., 2023; Vant et al., 2022).

Other authors combine them with information from neural network predictions (Tsai et al., 2020), in addition to the ones mentioned above (Qi et al., 2022). In this context, the work of Matsumoto et al. (2021) is particularly noteworthy. They developed and trained a deep neural network, called DefMap, to predict local dynamics using data not only from molecular simulations but from cryo-EM maps as well. As Matsumoto et al. (2021) explain in their article, for obtaining the training data, they retrieved 25 maps and atomic models from

the Electron Microscopy Data Bank (EMDB) (Turner et al., 2023b) and Protein Data Bank (PDB) and performed molecular dynamics (MD) simulations. With this data, they trained the supervised learning algorithm in such a way that it could learn the relation between local densities, from the volumes, and root mean square fluctuation (RMSF) values from simulations (see Figure 1).

Figure 1

Workflow for training DefMap Neural Network from Matsumoto et al. (2021).



Note. The first step was to perform the molecular simulations, then the network was trained with its output and volumes from the Electron Microscopy Data Bank (EMDB). Afterwards, the trained neural network was tested with experimentally obtained volumes. Image retrieved from "Extraction of protein dynamics information from cryo-EM maps using deep learning" by S. Matsumoto, S. Ishida, M. Araki, T. Kato, K. Terayama, and Y. Okuno, 2021, Nature Machine Intelligence 3(2), p. 154 (https://doi.org/10.1038/s42256-020-00290-y).

Nonetheless, they pointed out that this neural network has limitations for predicting the dynamics of structures with transmembrane regions or post-translational modifications. The reason for this lies in the fact that they could not be incorporated in the training dataset, given the difficulty of performing molecular simulations with these types of structures and issues with Cryo-EM map reconstruction around membrane mimetics and post-translational modifications.

4.5. Justification of the project

Despite its limitations, DefMap is undoubtedly a tool that can benefit researchers interested in studying protein dynamics. For this reason, incorporating it into the Scipion workflow engine for Cryo-EM and structural biology (Conesa et al., 2023) will allow users to find, install and run it more easily. Scipion is an open source framework that bundles and integrates a variety of software packages into protocols that form workflows, primarily for processing electron microscopy images, but also other functionalities such as atomic model building (Martínez et al., 2020) and molecular simulations (Del Hoyo et al., 2023). Scipion has a graphical interface, from which users can access the different protocols. Therefore, by integrating DefMap into this application, it will be more accessible to both researchers and developers.

The plugin was tested on different structures from two molecules: Human Huntingtin-HAP40 complex and SARS-CoV-2 Spike glycoprotein. The former was chosen because it was already analysed by Matsumoto et al. (2021), therefore, it was convenient to use it to compare the new features of the plugin with those already developed by them. On the other hand, Spike glycoprotein is a molecule whose structure and dynamics have been extensively studied by Cryo-EM and molecular simulations, among other techniques, due to its role in the SARS-CoV-2 virus infection process, that triggered a global pandemic in 2020 (Abduljalil et al., 2023; Sinha et al., 2023; Zaidi & Dawoodi, 2024). Therefore, since it is so well characterised, the information available in the public databases was reliable and suitable for testing the plugin.

5. Objectives

The main objective of this project is to enhance the understanding of the molecular structure and functions of biological particles. For this purpose, the following specific goals were defined:

1. Generate a plugin that integrates the DefMap neural network approach in the Scipion framework.
2. Apply this plugin to real molecules and analyse the results.

6. Material and Methods

6.1. Design of the project

According to the specific objectives mentioned in the previous section, the first step was to create a plugin in Scipion (scipion-em-defmap) using the template recommended in the documentation. Within this plugin, two protocols and one viewer were created. The first protocol developed (defmap - prediction) was the one that implements DefMap, in which six stages can be distinguished, steps 3 to 5 being those that directly run DefMap programs:

1. Validation and handling of input file formats.
2. Preprocessing of volumes.
3. Preparation of the dataset for prediction.
4. Inference with the neural network.
5. Postprocessing of the results.
6. Analysis of the results.

Afterwards, a specific viewer has also been created to facilitate the analysis of the results. In addition, another protocol was created in case users would like to use the viewer outside DefMap. This extra protocol (defmap - analysis) converts the file formats indicated in the input to pdb and generates a PyMOL script file, similary to the analysis step of the other protocol. The code of the plugin can be retrieved in https://github.com/Sofia-GMT/scipion-em-defmap

6.2. Sources of data

The code from Matsumoto et al. (2021) was obtained from *Github* and includes example input files, the source of these and other input files used to test the plugin are shown in Table 2. The conformations with PDB ids 6vyb and 7bnn of Spike are variants of the open state of the molecule, therefore, they are expected to predict a higher flexibility than the conformation with id 6vxx, associated with a closed state.

Table 2

Sources of data for testing the plugin.

Molecule	Volumes	Atomic Structure
SARS-CoV-2 Spike glycoprotein	EMD-21457	PDB: 6vyb
	EMD-12230	PDB: 7bnn
	EMD-21452	PDB: 6vxx
Human Huntingtin-HAP40 complex	EMD-3984	PDB: 6ez8
	Matsumoto et al. (2021) GitHub	Matsumoto et al. (2021) GitHub

Note. Three conformations of Spike protein and one of Huntingtin were analysed. Most volumes and atomic structures were retrieved from the Electron Microscopy Data Bank (EMDB) and Protein Data Bank (PDB), respectively, with the exception of those in the last row.

6.3. Preprocessing of data (steps 1-3)

Before running the neural network, it is necessary to preprocess the input cryo-EM data via two types of processing.

The first one was performed by executing different protocols of an existing plugin, specifically the scipion-em-xmipp plugin for Xmipp (Střelák et al., 2021) in our case, distinguishing four functional phases that ensure the maps have appropriate characteristics for the network:

1. Resize of the sampling rate to 1.50 Å/px, to be consistent with the training dataset.
2. Filter in Fourier space to 5 Å maximum resolution, eliminating higher resolution details which are less informative for dynamics and more sensitive to noise.
3. Create and apply a mask for smoothing the shape of the volume.
4. Apply a threshold to remove contaminants.

This first preprocessing is embedded in scipion-em-defmap and is optional, when Xmipp is available, which is often the case as it is a software and plugin that users usually already have installed. However, Scipion offers similar operations with other plugins like scipion-em-eman2 or scipion-em-relion in case the user prefers to use them instead.

The second preprocessing step is mandatory and consists of executing the script "prep_dataset.py" provided by Matsumoto et al. (2021) on their GitHub, to generate the dataset for the inference in the appropriate format.

6.4. Statistical Analysis (step 6)

To analyse the results, three common measures have been used to quantify the mobility of the atomic positions: Root Mean Square Deviation (RMSF; predicted by DefMap as normalized logarithms), B factors and local resolutions. The latter were calculated using DeepRes (Ramírez-Aportela et al., 2019).

The reference values used for comparing were:

- B-factors of atomic structures obtained from Protein Data Bank
- Local resolution extracted from volumes of EMDB.

To measure the correlation of the plugin results with the references, both linear regression and Pearson's correlation coefficient with their corresponding p-value were calculated. For the linear regression, the r-squared value was also calculated, which reflects what proportion of variance is explained by the model.

The variables considered in the analysis are:

- Dependent variables: B-factors and local resolution.
- Independent variables: log RMSF values.

The null hypothesis for the different tests are:

- For the linear regression: the slope of the regression is zero, so the variables are not related.
- For the correlation: the coefficient is zero, so the variables are not related.

For the contrasts of hypotheses, the p-value 0.01 was set as the maximum threshold for accepting the null hypothesis. It is also important to note that the lowest double value that

can be determined with precision in this machine is 2.220446049250313e-16, lower values cannot be determined with precission.

In order to facilitate their interpretation of the statistical analysis, four graphs have been plotted:

- Distribution of DefMap output values.
- DefMap output values vs Residue index.
- B factors of the reference structures vs DefMap output values.
- Local resolutions of the reference volumes vs DefMap output.

The plugin allows the users to decide whether they prefer to show the DefMap output values in logarithmic scale.

6.5. Equipment

The machine used to test the plugin was carver.cnb.csic.es, which has the following characteristics:

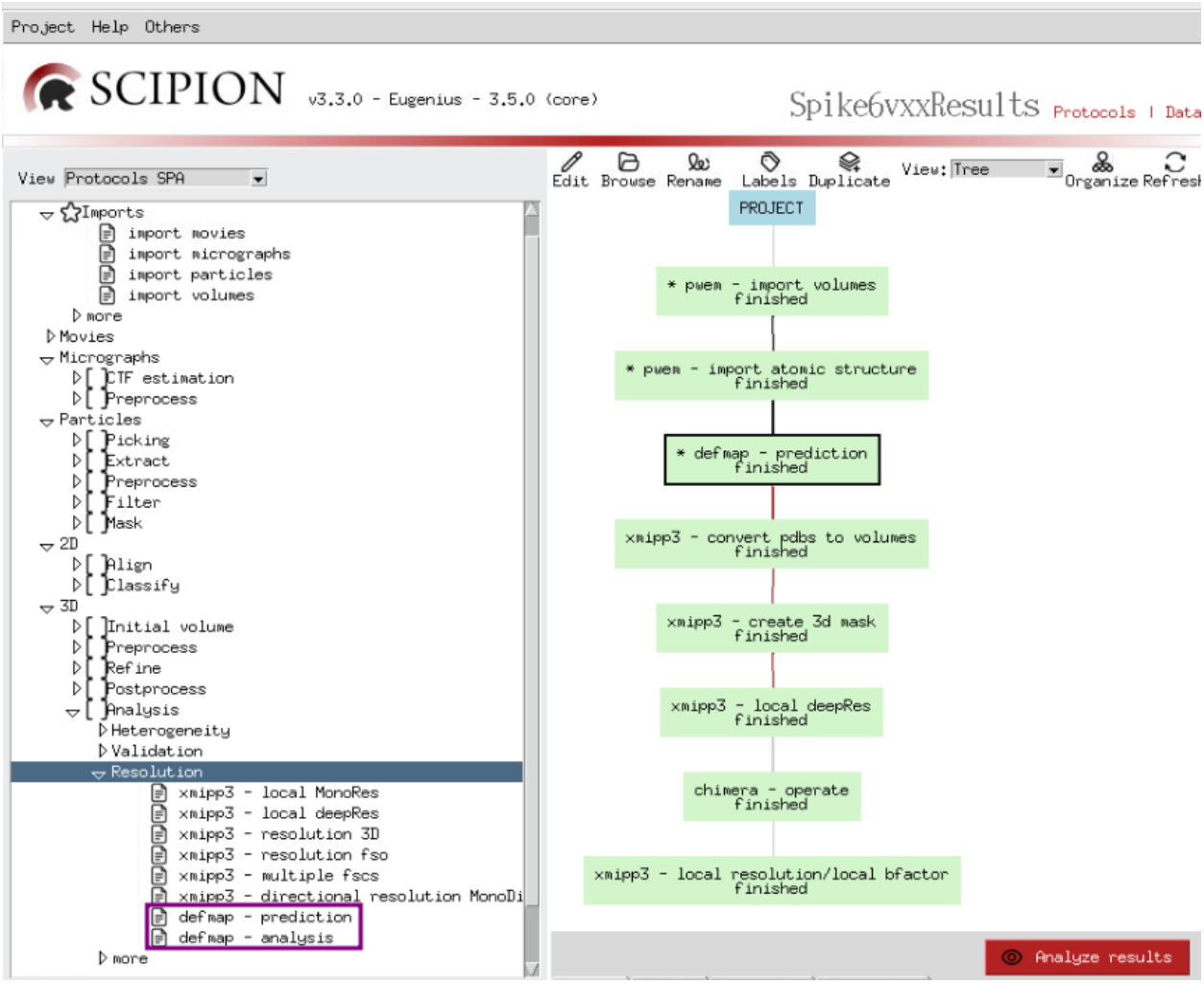
- Processor: Intel® Xeon® E5-2630 v4.
- Graphics: NVIDIA Corporation TU104GL.
- Four GPUs with 15360 MB of memory with 5060 CUDA cores. The main protocol uses one GPU for running Tensorflow in the inference step.
- Memory: 540 GB. Matsumoto et al. (2021) recommended users to have at least 96 GB.
- Machine epsilon for double precision: 2.220446049250313e-16. It is the lowest double value that can be determined with precision.

6.6. Workflow in Scipion application

As previously mentioned, Scipion has a graphical interface, from which the plugin can be used. Figure 2 illustrates the project created, inside the application, for predicting the dynamics of Spike structure 6vxx using DefMap and the DeepRes local resolution method. The rest of the molecules have followed a similar procedure. After executing this workflow, the graphs were generated by pressing the button “Analyze Results” and choosing the corresponding display option (Figure 3).

Figure 2

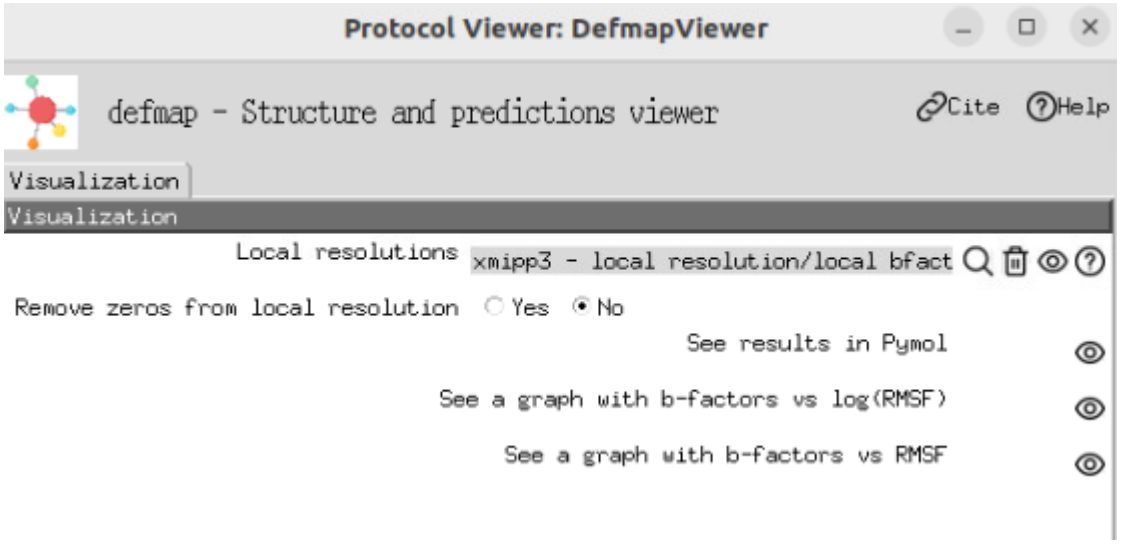
Screenshot of Scipion's project for the Spike structure 6vxx.



Note. In the left column we find the location of the two protocols generated in the plugin within a purple rectangle. The right column shows the workflow in which the “defmap - prediction” protocol is integrated. The input files were imported in the first two right rectangles and in the third one DefMap (darker outline) is executed. The following rectangles were used to extract the local resolutions. The “defmap - analysis” protocol adapts the input files generated outside the plugin, enabling them to be analysed with the viewer.

Figure 3

Viewer options panel.



Note. The results can be viewed in PyMOL or as graphs. In the first rectangle, the file with the local resolutions can optionally be specified. If it is not indicated, this graphic will not be generated and the option for removing the zeros will not be displayed.

7. Results

Five executions of the plugin have been carried out, two for the complex of huntingtin with huntingtin-associated protein 40 (HAP40) and three for the SARS-CoV-2 Spike.

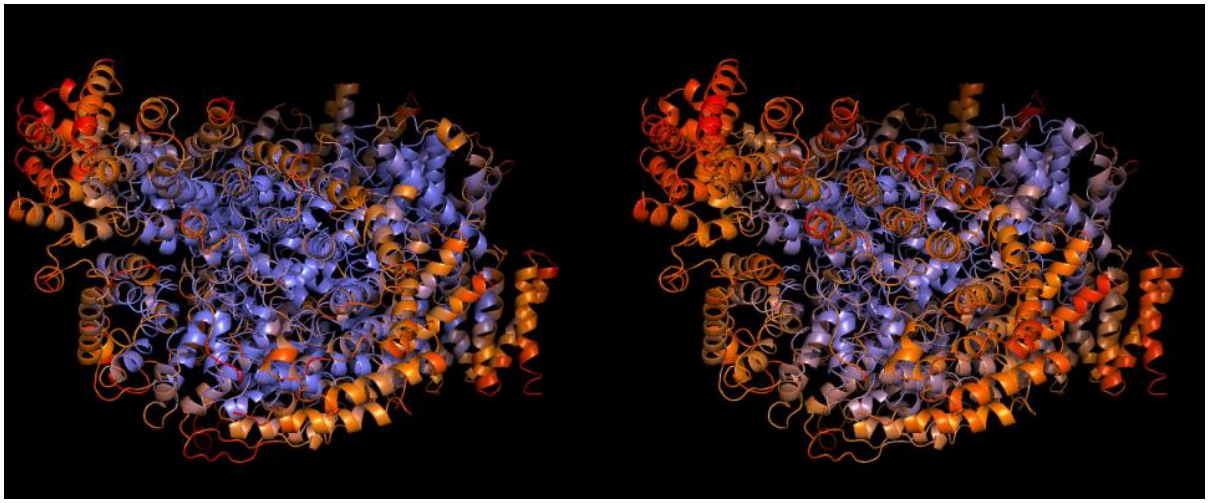
7.1. Human Huntingin-HAP40

In Figure 4, we can observe a comparison between the results generated using the files provided by Matsumoto et al. (2021) and the ones generated using public databases (PDB and EMDB). In general, both executions offer a similar prediction, which is in line with the dynamics seen in the original B-factors from the PDB. This suggests that the preprocessing provided by the plugin developed here generates reasonable predictions with DefMap, similar to the ones that performed by Matsumoto et al. (2021) using EMAN2.

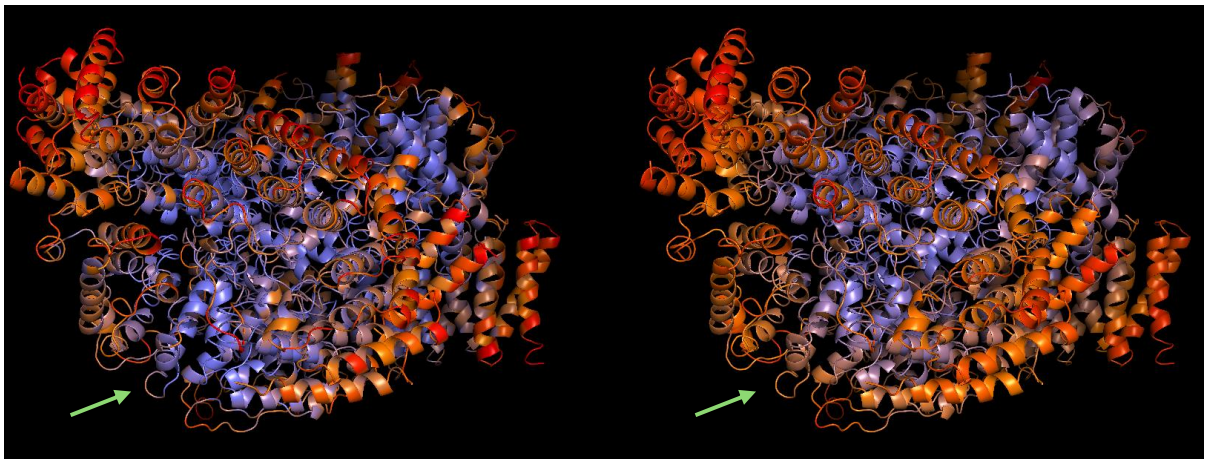
Figure 4

Visualisation in PyMOL of predicted Huntingtin-HAP40 dynamics against the reference structure B-factors.

(A)



(B)



Note. Atoms have been coloured using the command “spectrum b, slate_orange_red” in PyMOL. This command also had the arguments (minimum=1, maximum=2) to colour the prediction. The regions in red reflect more mobility than the ones in slate blue.

(A) The prediction (left) using as input the volume and the atomic structure (right) provided by Matsumoto et al. (2021).

(B) As in (A) but using input files from EMDB and PDB. The prediction (left) is quite similar to the reference (right), although in the latter the basal region has traces with higher mobility in orange (green arrows).

In Figures A1 to A4 of the Annexes, we can observe the statistics calculated for the runs with the input from Matsumoto et al. (2021) and from the public databases. In both executions it is observed that most of the log(RMSF) values are between -1 Å and 1, although there are some peaks until 3.

Considering that the p-value reflects the probability of observing the results assuming the null hypothesis; as it is lower than 0.01, we can consider that there is a significant correlation between B-factors and the values of log(RMSF), in both executions, with r-squared values between 60% and 70%.

On the other hand, the contrast of hypothesis for the local resolutions in the execution with input files from Matsumoto et al. (2021) accepts the null hypothesis, with a p-value of 0.04, while on the other it is rejected, but with a p-value of 3.87e-3. In all cases, the r-squared value in these comparisons with local resolutions are extremely low, not reaching 1%. This is consistent with the graphs since there are many points that are relatively far away from the regression line.

7.2. Spike

In Figure 5, we can compare the predictions generated using different conformations of the Spike glycoprotein. Overall, the three runs offer a similar prediction, the main differences are on the periphery, although in all cases it is predicted a higher mobility there than in the rest of the structure. However, both open structures (Figures 5A and 5B) predict higher mobility in the RBD and NTD domains (S1) in contrast to the closed conformation (Figure 5C) as expected.

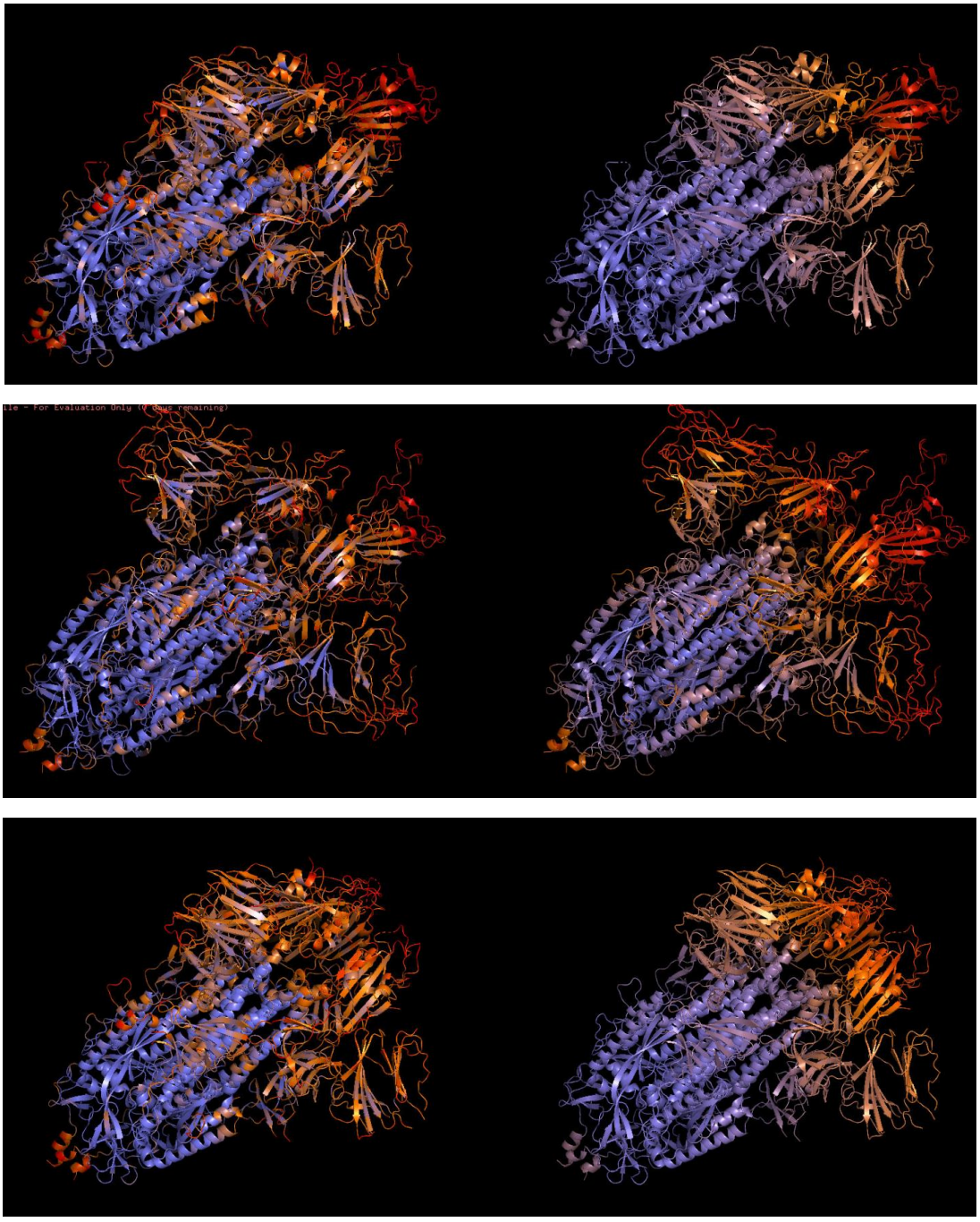
Figure 5

Visualisation in PyMOL of predicted Spike against the reference structure.

(A)

(B)

(C)



Note. Atoms have been coloured using the command “spectrum b, slate_orange_red”. The regions in red reflect more mobility than the ones in slate blue. The predictions are on the left while the reference with B-factors is on the right. Both the predictions and the references show a higher mobility in the S1 (top right) and S2 (bottom left) domains, although the predictions show more movement in the S2 domain in comparison to the reference. The conformations 6vyb (A) and 7bnn (B) correspond to an open state of Spike, while the conformation 6vxx (C) is associated to a closed state.

The graphs and statistics calculated using the three different structures of Spike can be observed from Figures A4 to A12 of the Annexes. In the three of them, most log(RMSF) values are between -1 and 1, although many residues close to the C-terminal have values between -2 and 0. However, extreme C-terminal residues have higher values.

When checking the statistics for the comparison with the B-factors, it is found that the p-values are less than 0.01, and that the regression models explain between 60% and 70% of the variability of the data. The graphs from the structure 7bnn from the open conformation show that the atoms from chain B have higher log(RMSF) values than those predicted in the regression model, while in the graphs from 6vxx the origin of the outliers is distributed between the three chains.

On the other hand, when analysing the local resolution against log(RMSF) values, the regression models explain less than 5% of the data, even if a statistically significant positive correlation is calculated. Consistently with the graphs, a high dispersion with respect to the regression line is noted.

8. Discussion

The comparison of the B-factors and the local resolutions against the output of the scipion-em-defmap plugin (RMSF in logarithmic scale), has shown that in four of the five cases both relationships are proportional and significant. This is consistent with what was explained in the introduction, since the three variables measure the degree of variability or uncertainty of the atomic positions in different ways. The only execution, where the relationship has not been accepted, has a p-value greater than 0.01 but less than 0.05, thus the support for the null hypothesis is very low.

In all cases, both statistics and graphs show that the relationship with the B-factors is much stronger than with the local resolutions. In the case of B-factors, the models explained between 50% and 70% of the variability, being generally more explanatory in the Huntingtin-HAP40 complex executions than in the Spike executions.

On the contrary, the models in the Spike executions were a bit more explanatory in the local resolutions, but it was less than 5% and the slope of the regression was close to 0. Consequently, it can be concluded that there might be a relationship between the RMSF

values and local resolutions, but that they are insufficient to serve as a unique predictor of each other depending on the molecule.

Furthermore, the graphs show that different chains from distinct structures can fit a logarithmic model in a different way. In the cases of the Huntingtin-HAP40 complex, and the chains A and C of Spike, the trend of the points is roughly similar. However, for the open conformation 6vyb, the distribution of the points in chain B clearly follows a different logarithmic function, with an initial increase in the vertical axis much more pronounced than the rest of the chains, as the B-factors are higher than the predictions in that area. Therefore, a more in-depth study for each chain in the different conformations would be recommendable, and it would be convenient to analyse this relationship with more molecules in future studies, in order to identify a more suitable model for predicting B-factors as a function of RMSF, and vice versa. For example, it would be desirable to study further cases with asymmetry of conformation and dynamics across states, similar to Spike.

Observing the PyMOL representations in Figures 4 an 5, it is clear that, despite the intensity differences in the intermediate regions, predictions and references show a high degree of similarity. Considering that representations are coloured along a spectrum based on the B-factor column of the PDB files, where the DefMap log RMSF values are also stored, this is consistent with the detected relationship between the DefMap log RMSFs and the B-factors. However, there are still some differences, generally taking warmer colors in the plugin predictions. These differences can be better understood thanks to the graphs and statistics.

Moreover, comparative analysis was also carried out between predictions generated using the pre-processed volumes of Matsumoto et al. (2021) and those in which the volumes were pre-processed with the Xmipp plugin. The graphs and the statics show that both preprocessing methods offer quite similar results; therefore, the preprocessing integrated within the plugin it is a good option for the majority of Scipion’s users that have Xmipp and not EMAN2.

In addition, when looking at the graphs with the local resolutions, there are some outliers at 0 Å. This was the reason why there is an option in the viewer to prevent them from being displayed and included in the statistics. A possible cause could be the application of a too tight mask in the protocol for obtaining the local resolutions.

Additionally, during the construction and development of the plugin, some difficulties were encountered.

Firstly, indicating a threshold in the command for the creation of the dataset produced an error in the post-processing step, when relating it with the atomic structure. As an alternative, the threshold was more convenient to apply prior to the dataset creation command. In addition, an issue was created in their Github repository.

Secondly, one of the concerns reported in the follow-up study by Matsumoto et al. (2023) was corroborated. They indicated that the inference step was notably longer, when predicting the dynamics of molecules that were not included in the test dataset. Considering that the Huntington-HAP40 complex was included and Spike was not, both executions of the former took around 10 minutes, while the executions from the second one took between 30 and 50 minutes. Nevertheless, it is still much faster than running a molecular simulation (on the order of weeks).

Another concern from Matsumoto et al. (2023) was that DefMap was not thought to be used with molecules with transmembrane regions or complex post-translational modifications, like Spike, due to the computational difficulty of executing molecular simulations with them. In addition to extending the learning dataset, as they pointed out, it would also be recommendable, in future versions of the plugin, to give users the option to train the neural network with their own data, instead of using the already trained network.

Overall, despite the limitations mentioned above, DefMap's results are promising, making it a useful tool for Scipion users. Furthermore, its integration in the plugin, with the pre-processing workflow with Xmipp and with the statistical analyses, will facilitate its accessibility and the interpretation of its results.

9. Conclusion

In conclusion, this project could be summarised in the following points:

- The scipion-em-defmap plugin has been incorporated into the Scipion framework, as a tool for predicting molecular dynamics.
- The plugin integrates the DefMap neural network with a pre-processing and analysis workflow.
- Analyses show a relationship between the plugin output (RMSF) with B-factors and, to a lesser extent, with local resolutions.

For future versions of the plugin, it would be desirable to allow users to train the neural network with their own data. Additionally, further studies on the relationship of RMSFs with B-factors and local resolutions would also be helpful in order to obtain more significant results.

10. Bibliography

Abduljalil, J. M., Elghareib, A. M., Samir, A., Ezat, A. A., & Elfiky, A. A. (2023). How helpful were molecular dynamics simulations in shaping our understanding of SARS-CoV-2 spike protein dynamics? *International Journal Of Biological Macromolecules*, *242*, 125153. https://doi.org/10.1016/j.ijbiomac.2023.125153

AlRawashdeh, S., & Barakat, K. (2023). Applications of Molecular Dynamics Simulations in Drug Discovery. En *Methods in molecular biology* (pp. 127-141). https://doi.org/10.1007/978-1-0716-3441-7_7

Bagewadi, Z. K., Khan, T. M. Y., Gangadharappa, B., Kamalapurkar, A., Shamsudeen, S. M., & Yaraguppi, D. A. (2023). Molecular dynamics and simulation analysis against superoxide dismutase (SOD) target of Micrococcus luteus with secondary metabolites from Bacillus licheniformis recognized by genome mining approach. *Saudi journal of biological sciences*, *30*(9), 103753. https://doi.org/10.1016/j.sjbs.2023.103753

Bock, L. V., Gabrielli, S., Kolář, M., & Grubmüller, H. (2023). Simulation of Complex Biomolecular Systems: The Ribosome Challenge. *Annual Review Of Biophysics*, *52*(1), 361-390. https://doi.org/10.1146/annurev-biophys-111622-091147

Boehr, D. D., Nussinov, R., & Wright, P. E. (2009). The role of dynamic conformational ensembles in biomolecular recognition. *Nature Chemical Biology*, *5*(11), 789-796. https://doi.org/10.1038/nchembio.232

Conesa, P., Fonseca, Y. C., Jiménez de la Morena, J., Sharov, G., de la Rosa-Trevín, J. M., Cuervo, A., … Sorzano, C. O. S. (2023). Scipion3: A workflow engine for cryo-electron microscopy image processing and structural biology. *Biological Imaging*, *3*, e13. https://doi.org/10.1017/S2633903X23000132

Costa, M. G. S., Gür, M., Krieger, J., & Bahar, İ. (2023). Computational biophysics meets cryo-EM revolution in the search for the functional dynamics of biomolecular systems. *Wiley Interdisciplinary Reviews. Computational Molecular Science*, *14*(1). https://doi.org/10.1002/wcms.1689

Del Hoyo, D., Salinas, M., Lomas, A., Ulzurrun, E., Campillo, N. E., & Sorzano, C. (2023). Scipion-Chem: An Open Platform for Virtual Drug Screening. *Journal Of Chemical Information And Modeling*, *63*(24), 7873-7885. https://doi.org/10.1021/acs.jcim.3c01085

Doktorova, M., Khelashvili, G., Ashkar, R., & Brown, M. F. (2023). Molecular simulations and NMR reveal how lipid fluctuations affect membrane mechanics. *Biophysical Journal*, *122*(6), 984-1002. https://doi.org/10.1016/j.bpj.2022.12.007

Glaeser, R. M. (2018). Proteins, interfaces, and cryo-EM grids. *Current Opinion In Colloid & Interface Science*, *34*, 1-8. https://doi.org/10.1016/j.cocis.2017.12.009

Hénin, J., Lelièvre, T., Shirts, M. R., Valsson, Ó., & Delemotte, L. (2022). Enhanced Sampling Methods for Molecular Dynamics Simulations [Article v1.0]. *Living Journal Of Computational Molecular Science*, *4*(1). https://doi.org/10.33011/livecoms.4.1.1583

Hollingsworth, S. A., & Dror, R. O. (2018). Molecular Dynamics Simulation for All. *Neuron*, *99*(6), 1129–1143. https://doi.org/10.1016/j.neuron.2018.08.011

Kohen, A. (2014). Role of Dynamics in Enzyme Catalysis: Substantial versus Semantic Controversies. *Accounts Of Chemical Research*, *48*(2), 466-473. https://doi.org/10.1021/ar500322s

Lengyel, J., Hnath, E., Storms, M., & Wohlfarth, T. (2014). Towards an integrative structural biology approach: combining Cryo-TEM, X-ray crystallography, and NMR. *Journal Of Structural And Functional Genomics*, *15*(3), 117-124. https://doi.org/10.1007/s10969-014-9179-9

Li, B., Zhu, D., Shi, H., & Zhang, X. (2021). Effect of charge on protein preferred orientation at the air–water interface in cryo-electron microscopy. *Journal Of Structural Biology*, *213*(4), 107783. https://doi.org/10.1016/j.jsb.2021.107783

Libretexts. (2023, 30 enero). *Introduction to NMR*. Chemistry LibreTexts. https://chem.libretexts.org/Bookshelves/Physical_and_Theoretical_Chemistry_Textbook_Maps/Supplemental_Modules_(Physical_and_Theoretical_Chemistry)/Spectroscopy/Magnetic_Resonance_Spectroscopies/Nuclear_Magnetic_Resonance/Nuclear_Magnetic_Resonance_II

Martínez, M., Jiménez-Moreno, A., Maluenda, D., Ramírez-Aportela, E., Melero, R., Cuervo, A., Conesa, P., Del Caño, L., Fonseca, Y. C., Sánchez-García, R. J., Střelák, D., Conesa, J. J., Fernández-Giménez, E., De Isidro, F., Sorzano, C., Carazo, J. M., & Marabini, R. (2020). Integration of Cryo-EM Model Building Software in Scipion. *Journal Of Chemical Information And Modeling*, *60*(5), 2533-2540. https://doi.org/10.1021/acs.jcim.9b01032

Matsumoto, S., Ishida, S., Araki, M., Kato, T., Terayama, K., & Okuno, Y. (2021). Extraction of protein dynamics information from cryo-EM maps using deep learning. *Nature Machine Intelligence*, *3*(2), 153-160. https://doi.org/10.1038/s42256-020-00290-y

Matsumoto, S., Ishida, S., Terayama, K., & Okuno, Y. (2023). Quantitative analysis of protein dynamics using a deep learning technique combined with experimental cryo-EM density data and MD simulations. *Biophysics and physicobiology*, *20*(2), e200022. https://doi.org/10.2142/biophysico.bppb-v20.0022

Murata, K., & Wolf, M. (2018). Cryo-electron microscopy for structural analysis of dynamic biological macromolecules. *Biochimica Et Biophysica Acta. G, General Subjects/Biochimica Et Biophysica Acta. General Subjects (Online)*, *1862*(2), 324-334. https://doi.org/10.1016/j.bbagen.2017.07.020

Qi, G., Vrettas, M. D., Biancaniello, C., Sanz-Hernández, M., Cafolla, C. T., Morgan, J. W. R., Wang, Y., De Simone, A., & Wales, D. J. (2022). Enhancing Biomolecular Simulations with Hybrid Potentials Incorporating NMR Data. *Journal Of Chemical Theory And Computation*, *18*(12), 7733-7750. https://doi.org/10.1021/acs.jctc.2c00657

Ramírez-Aportela, E., Mota, J., Conesa, P., Carazo, J. M., & Sorzano, C. (2019). DeepRes: a new deep-learning- and aspect-based local resolution method for electron-microscopy maps. *IUCrJ*, *6*(6), 1054-1063. https://doi.org/10.1107/s2052252519011692

Shukla, V. K., Heller, G. T., & Hansen, D. F. (2023). Biomolecular NMR spectroscopy in the era of artificial intelligence. *Structure*, *31*(11), 1360-1374. https://doi.org/10.1016/j.str.2023.09.011

Sinha, A., Sangeet, S., & Roy, S. (2023). Evolution of Sequence and Structure of SARS-CoV-2 Spike Protein: A Dynamic Perspective. *ACS Omega*, *8*(26), 23283-23304. https://doi.org/10.1021/acsomega.3c00944

Smyth, M. S., & Martin, J. H. (2000). x ray crystallography. *Molecular pathology: MP*, *53*(1), 8–14. https://doi.org/10.1136/mp.53.1.8

Sorzano, C. O. S., Jiménez-Moreno, A., Maluenda, D., Martínez, M., Ramírez-Aportela, E., Krieger, J., Melero, R., Cuervo, A., Conesa, J., Filipovic, J., Conesa, P., Del Caño, L., Fonseca, Y. C., Jiménez-de la Morena, J., Losana, P., Sánchez-García, R., Strelak, D., Fernández-Giménez, E., de Isidro-Gómez, F. P., Herreros, D., … Carazo, J. M. (2022). On bias, variance, overfitting, gold standard and consensus in single-particle analysis by cryo-electron microscopy. *Acta crystallographica. Section D, Structural biology*, *78*(4), 410–423. https://doi.org/10.1107/S2059798322001978

Srivastava, A., Nagai, T., Srivastava, A., Miyashita, O., & Tama, F. (2018). Role of Computational Methods in Going beyond X-ray Crystallography to Explore Protein Structure and Dynamics. *International Journal Of Molecular Sciences*, *19*(11), 3401. https://doi.org/10.3390/ijms19113401

Střelák, D., Jiménez-Moreno, A., Vilas, J. L., Ramírez-Aportela, E., Sánchez-García, R. J., Maluenda, D., Vargas, J., Herreros, D., Fernández-Giménez, E., De Isidro-Gómez, F. P., Horáček, J., Myška, D., Horáček, M., Conesa, P., Fonseca-Reyna, Y. C., Jiménez, J., Martínez, M., Harastani, M., Jonić, S., . . . Sorzano, C. Ó. S. (2021). Advances in Xmipp for Cryo–Electron Microscopy: From Xmipp to Scipion. *Molecules/Molecules Online/Molecules Annual*, *26*(20), 6224. https://doi.org/10.3390/molecules26206224

Tang, W., Zhong, E. D., Hanson, S. M., Thiede, E. H., & Cossio, P. (2023). Conformational heterogeneity and probability distributions from single-particle cryo-electron microscopy. *Current Opinion In Structural Biology*, *81*, 102626. https://doi.org/10.1016/j.sbi.2023.102626

Trueblood, K. N., Bürgi, H., Burzlaff, H., Dunitz, J. D., Gramaccioli, C. M., Schulz, H., Shmueli, U., & Abrahams, S. C. (1996). Atomic Dispacement Parameter Nomenclature. Report of a Subcommittee on Atomic Displacement Parameter Nomenclature. *Acta Crystallographica. Section A, Foundations Of Crystallography/Acta Crystallographica. Section A*, *52*(5), 770-781. https://doi.org/10.1107/s0108767396005697

Tsai, S. T., Kuo, E. J., & Tiwary, P. (2020). Learning molecular dynamics with simple language model built upon long short-term memory neural network. *Nature Communications*, *11*(1). https://doi.org/10.1038/s41467-020-18959-8

Turner, J., Abbott, S., Da Fonseca, N. J., Carrijo, L., Duraisamy, A. K., Salih, O., Wang, Z., Kleywegt, G. J., Morris, K. L., Patwardhan, A., Burley, S., Crichlow, G., Feng, Z., Flatt, J. W., Ghosh, S., Hudson, B. P., Lawson, C. L., Liang, Y., Peisach, E., . . . Ma, X. (2023b). EMDB—The Electron Microscopy Data Bank. *Nucleic Acids Research*, *52*(D1), D456-D465. https://doi.org/10.1093/nar/gkad1019

Vant, J., Sarkar, D., Nguyen, J., Baker, A., Vermaas, J. V., & Singharoy, A. (2022). Exploring cryo-electron microscopy with molecular dynamics. *Biochemical Society Transactions*, *50*(1), 569-581. https://doi.org/10.1042/bst20210485

Vilas, J. L., Carazo, J., & Sorzano, C. (2022). Emerging Themes in CryoEM─Single Particle Analysis Image Processing. *Chemical Reviews*, *122*(17), 13915-13951. https://doi.org/10.1021/acs.chemrev.1c00850

Vilas, J.L. (2019). Local quality assessment of cryo-EM reconstructions and its applications [Doctoral dissertation, Universidad Autónoma de Madrid]. UAM Institutional repository. https://repositorio.uam.es/handle/10486/688556

Wang, H. W., & Wang, J. W. (2016). How cryo-electron microscopy and X-ray crystallography complement each other. *Protein science: a publication of the Protein Society*, *26*(1), 32–39. https://doi.org/10.1002/pro.3022

Yanaka, S., Yogo, R., & Kato, K. (2020). Biophysical characterization of dynamic structures of immunoglobulin G. *Biophysical Reviews*, *12*(3), 637-645. https://doi.org/10.1007/s12551-020-00698-1

Zadorozhnyi, R., Gronenborn, A. M., & Polenova, T. (2024). Integrative approaches for characterizing protein dynamics: NMR, CryoEM, and computer simulations. *Current opinion in structural biology*, *84*, 102736. https://doi.org/10.1016/j.sbi.2023.102736

Zaidi, A. K., & Dawoodi, S. (2024). Structural biology of SARS-CoV-2. *Progress in molecular biology and translational science*, *202*, 31-23. https://doi.org/10.1016/bs.pmbts.2023.11.001

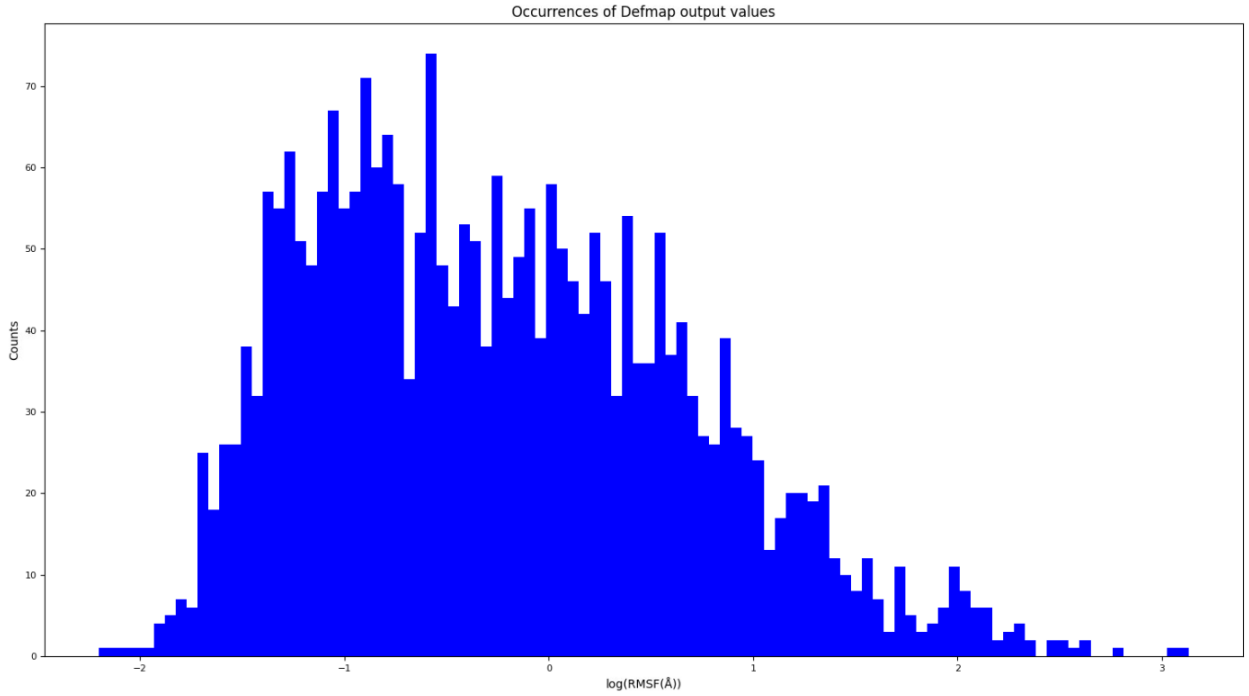
Zheng, H., Handing, K., Zimmerman, M., Shabalin, I., Almo, S. C., & Minor, W. (2015). X-ray crystallography over the past decade for novel drug discovery – where are we heading next? *Expert Opinion On Drug Discovery*, *10*(9), 975-989. https://doi.org/10.1517/17460441.2015.1061991

11. Annexes

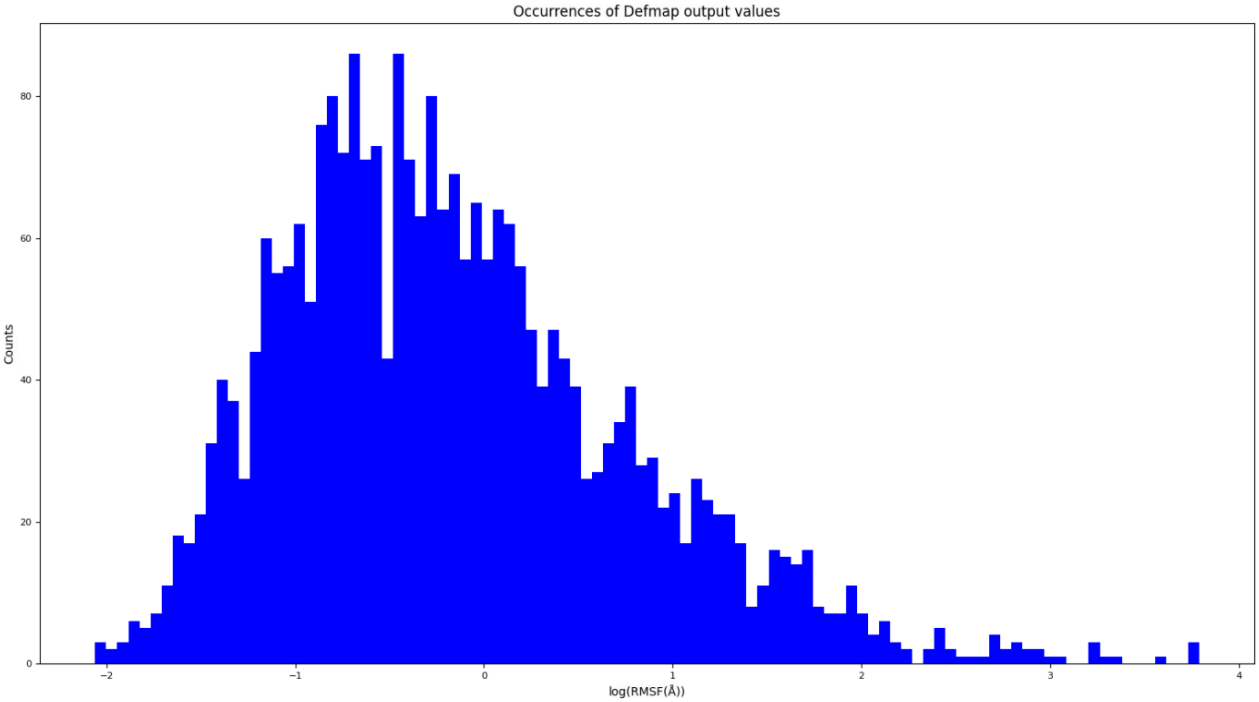
Figure A1

Representation of the occurrences of RMSF values in logarithmic scale from the prediction for the Human Huntingin-HAP40 complex.

(A)



(B)

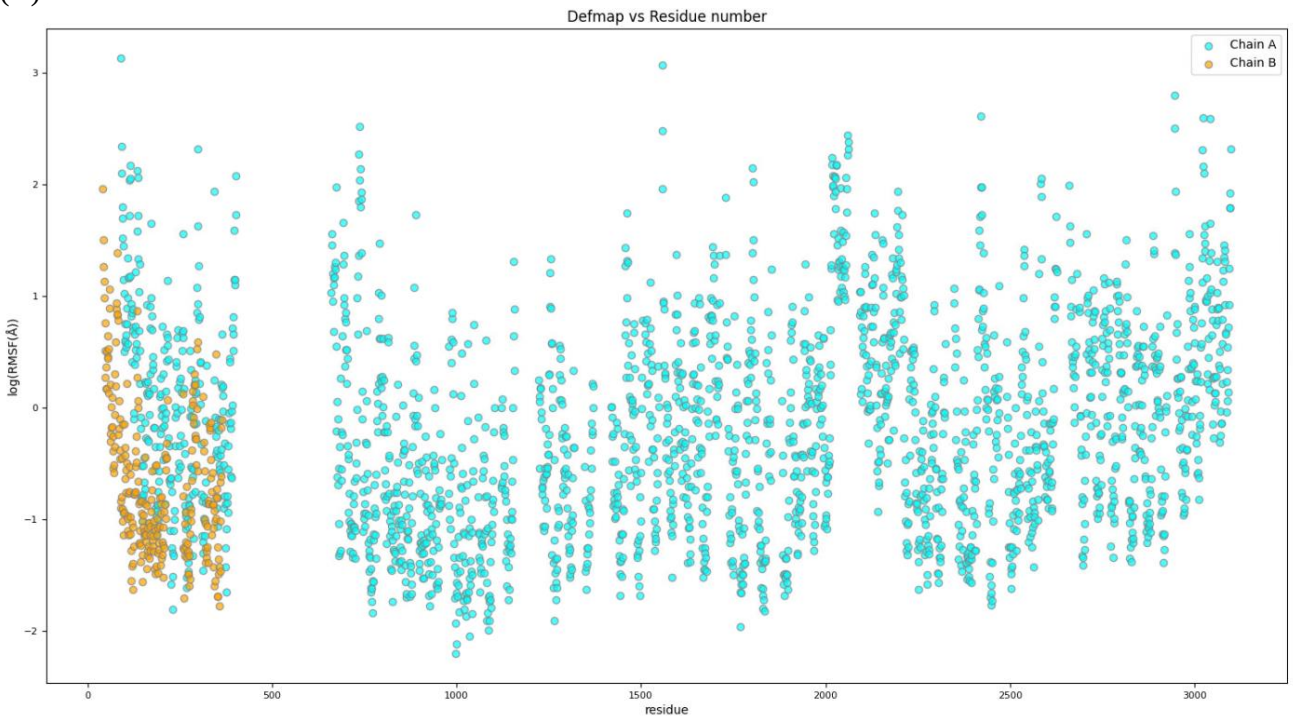


Note. In (A), the prediction was generated using the input from Matsumoto et al. (2021), while in (B) the input was retrieved from EMDB and PDB databases and pre-processed with Xmipp.

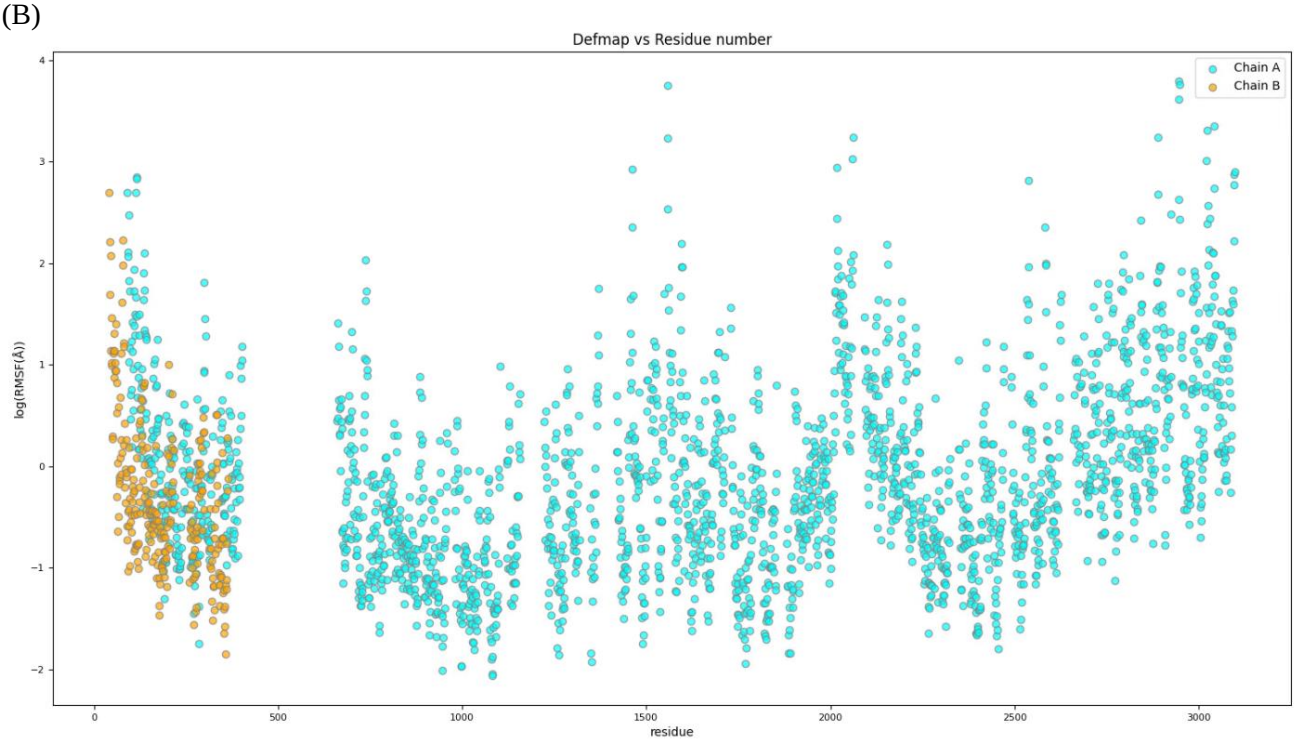
Figure A2

Representation of the RMSF values in logarithmic scale from the prediction against the residue number, for the Human Huntingin-HAP40 complex.

(A)



(B)

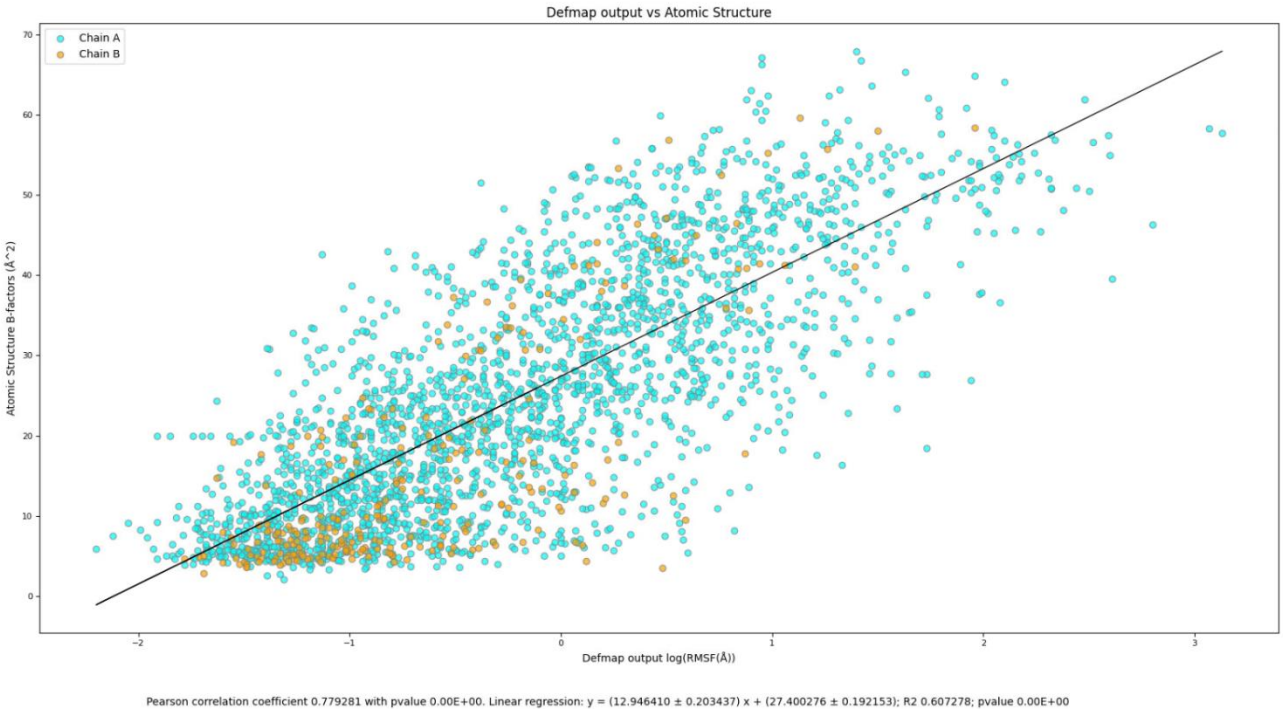


Note. In (A), the prediction was generated using the input from Matsumoto et al. (2021), while in (B) the input was retrieved from EMDB and PDB databases and pre-processed by Xmipp.

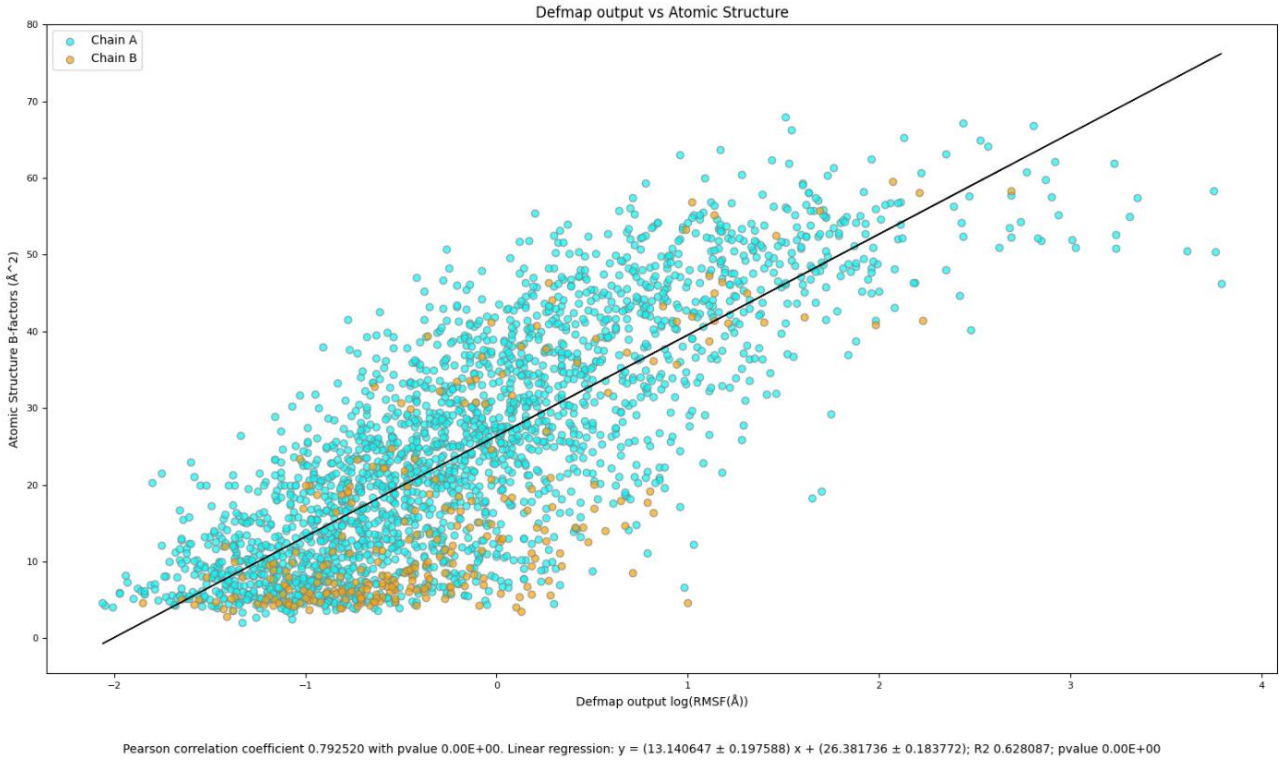
Figure A3

B-factors of the reference structure against RMSF values in logarithmic scale from the prediction, for the Human Huntingin-HAP40 complex.

(A)



(B)

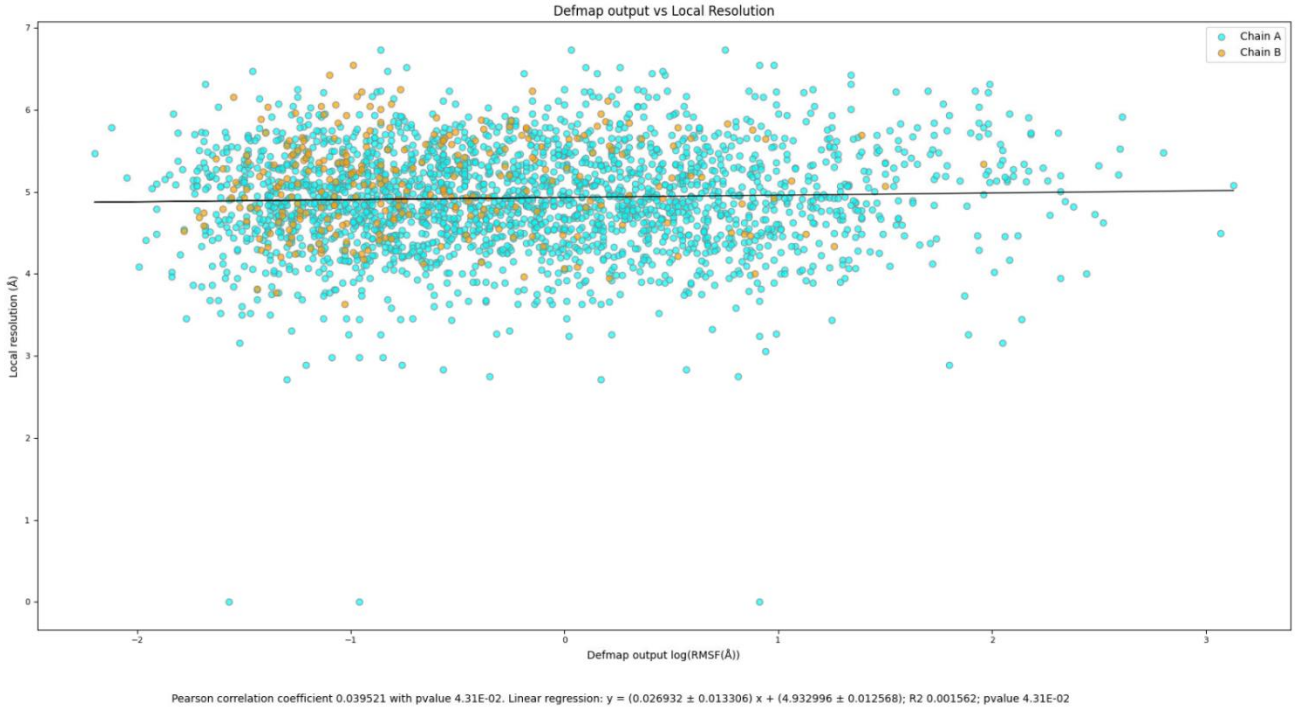


Note. P-values equal to zero means that its real value is lower than the epsilon machine value. In (A), the prediction was generated using the input from Matsumoto et al. (2021). Pearson's coefficient is 0.77, r-squared is 0.61 and p-values lower than 0.01. In (B) the input was retrieved from EMDB and PDB databases and pre-processed with Xmipp. Pearson's coefficient is 0.79, r-squared is 0.63, with a p-values lower than 0.01.

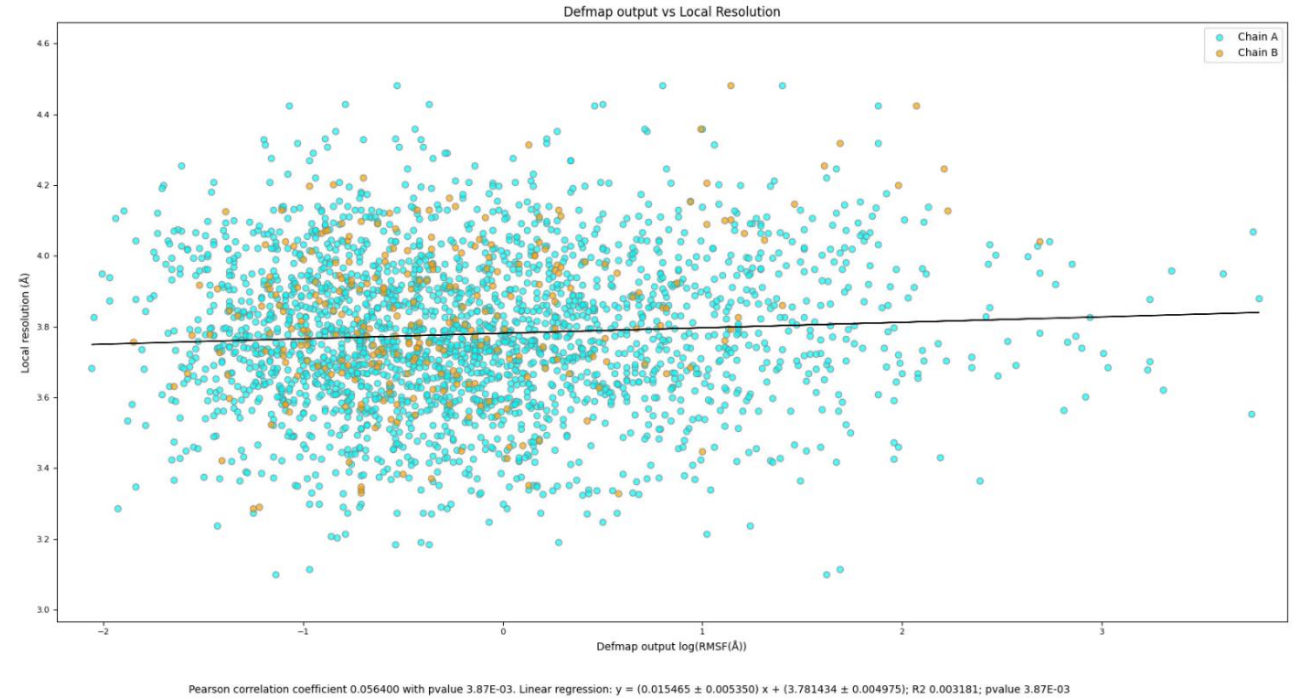
Figure A4

Local resolution of the reference structure against RMSF values in logarithmic scale from the prediction, for the Human Huntingin-HAP40 complex.

(A)



(B)

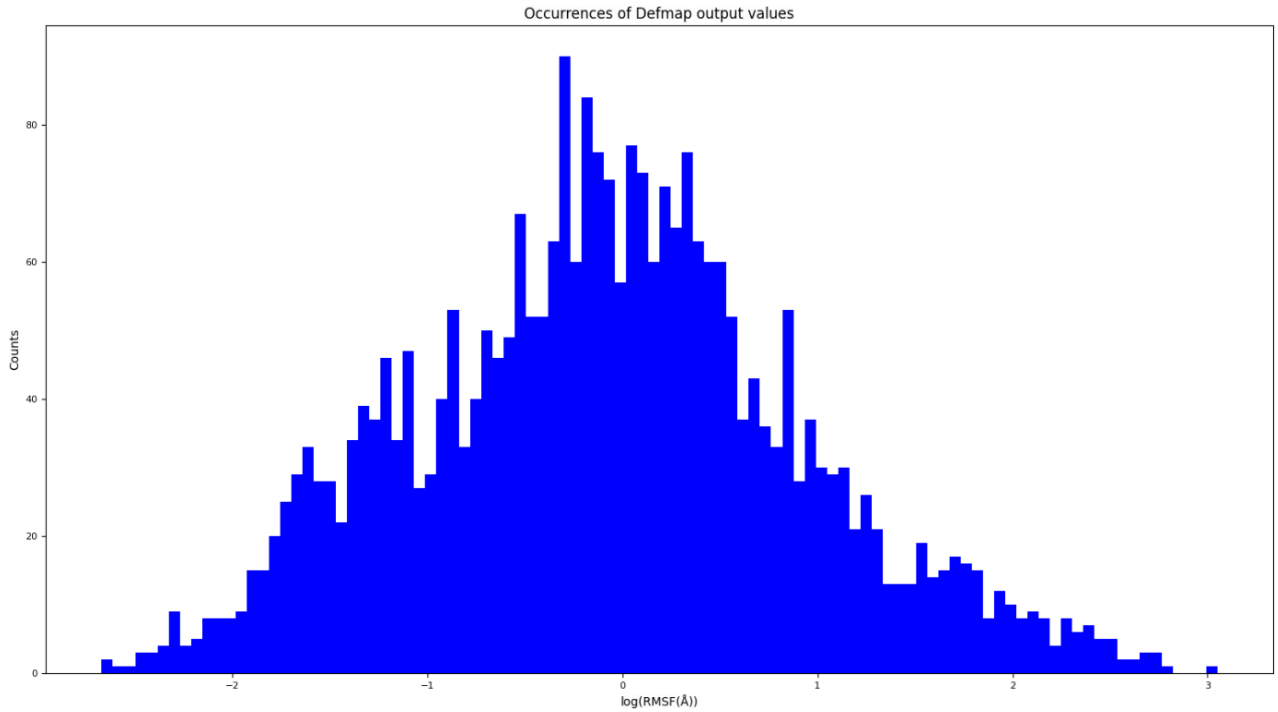


Note. In (A), the prediction was generated using the input from Matsumoto et al. (2021). Pearson's coefficient is 0.04 and r-squared is 0.0016 with p-values higher than 0.01 and lower than 0.05. In (B) the input was retrieved from EMDB and PDB databases and pre-processed with Xmipp. Pearson's coefficient is 0.056 and r-squared is 0.003 with p-values lower than 0.01.

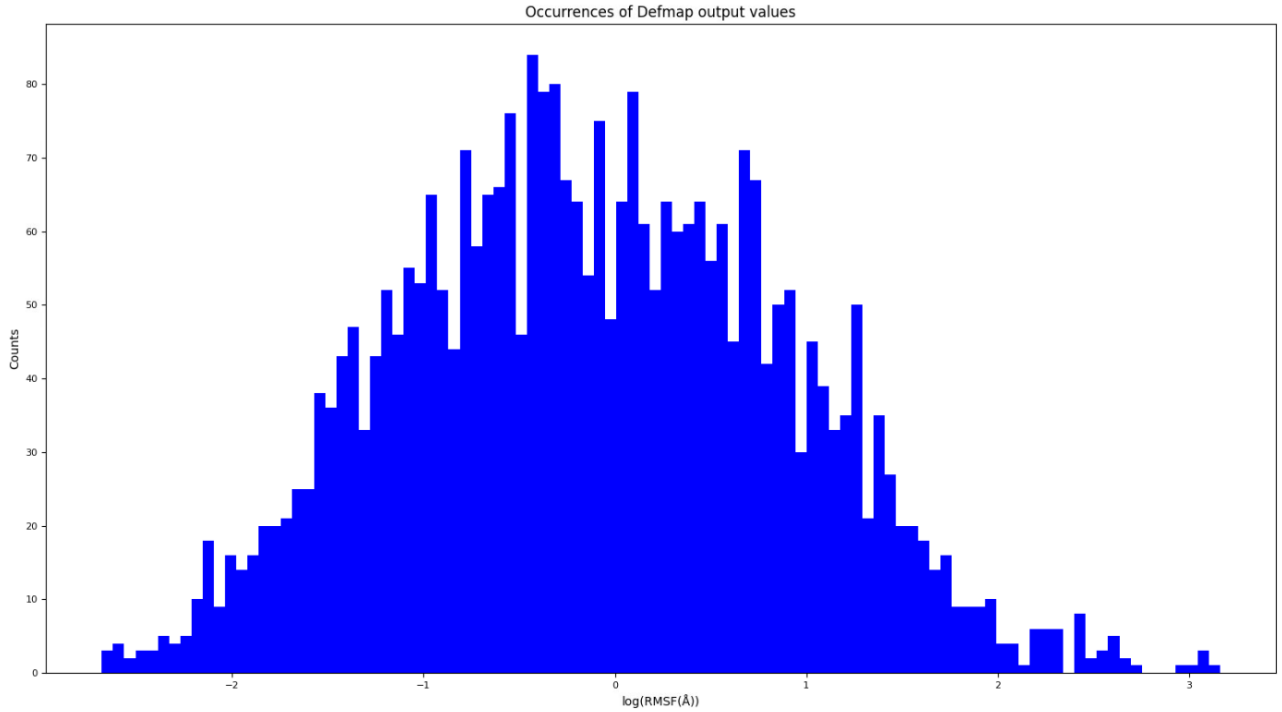
Figure A5

Representation of the occurrences of RMSF values in logarithmic scale from the prediction for the open conformation of Spike.

(A)



(B)

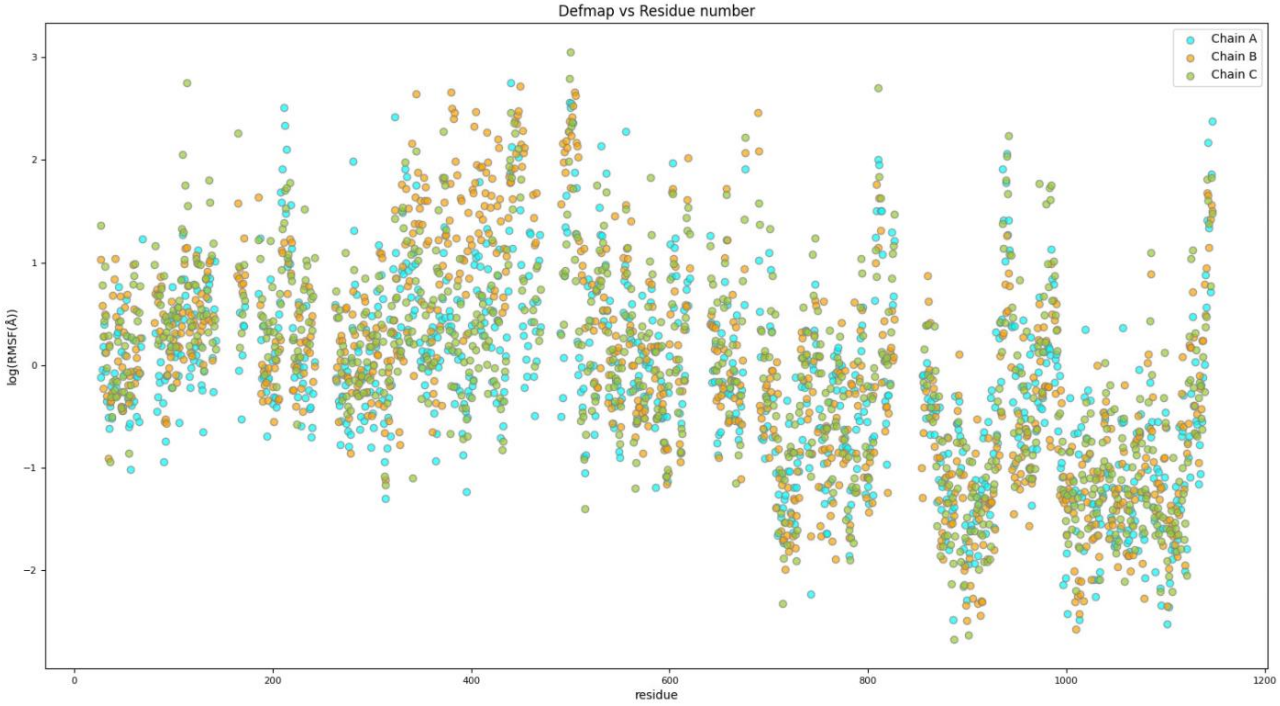


Note. (A) represents the values for 6vyb structure while (B) shows the values for 7bnn structure. Both are unimodal distributions, with the main peak between -1 and 1 in the horizontal axis.

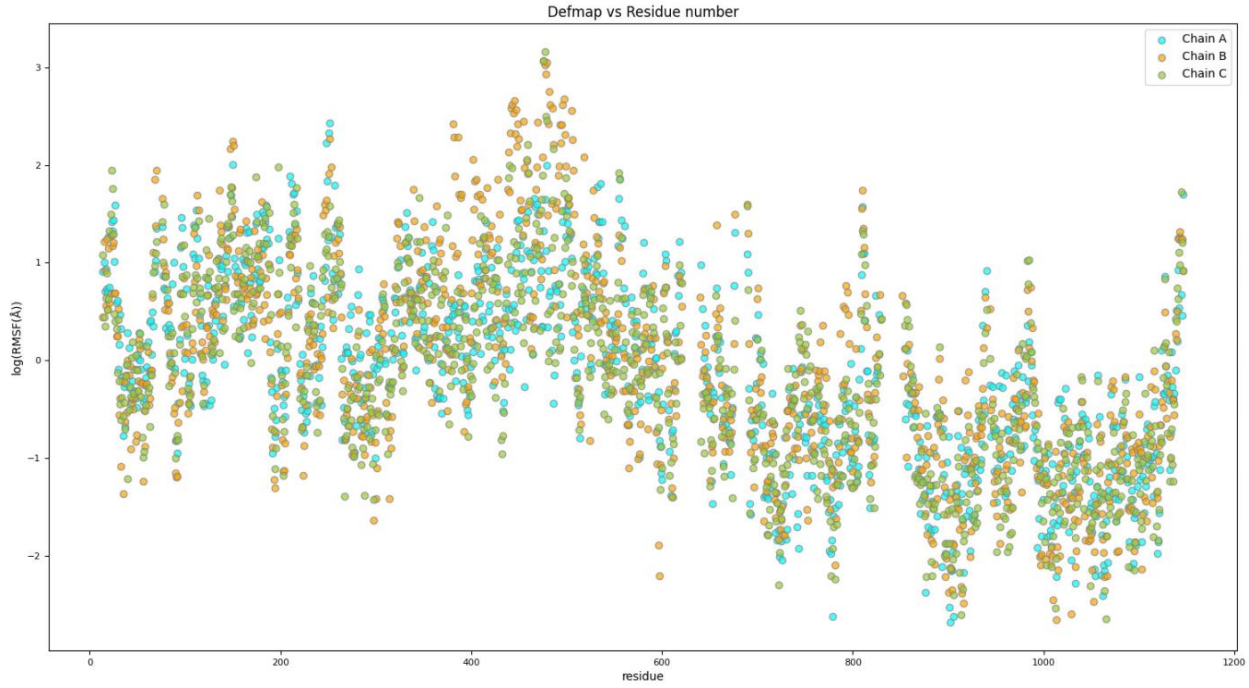
Figure A6

Representation of the RMSF values in logarithmic scale from the prediction against the residue number, for open conformation of Spike.

(A)



(B)

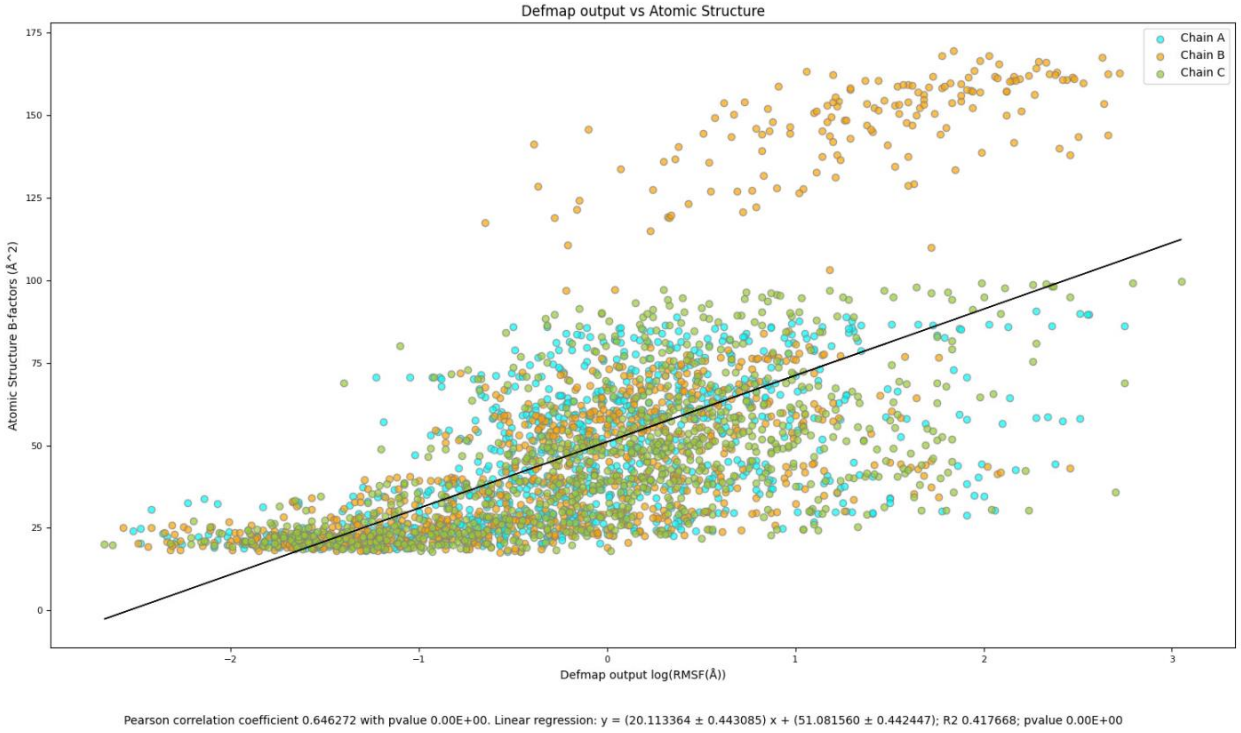


Note. (A) represents the values for 6vyb structure while (B) shows the values for 7bnn structure. Most residues close to the C-terminal have more atoms with low values.

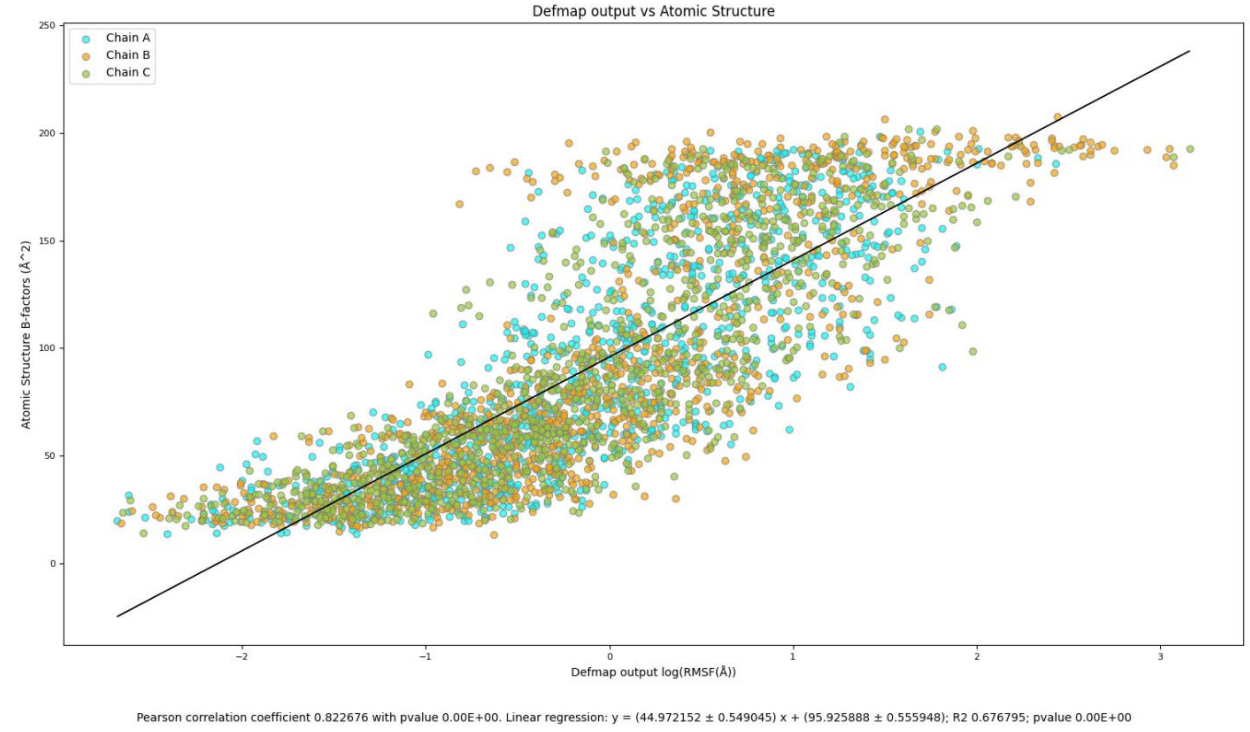
Figure A7

B-factors of the reference structure against RMSF values in logarithmic scale from the prediction, for the open conformation of Spike.

(A)



(B)

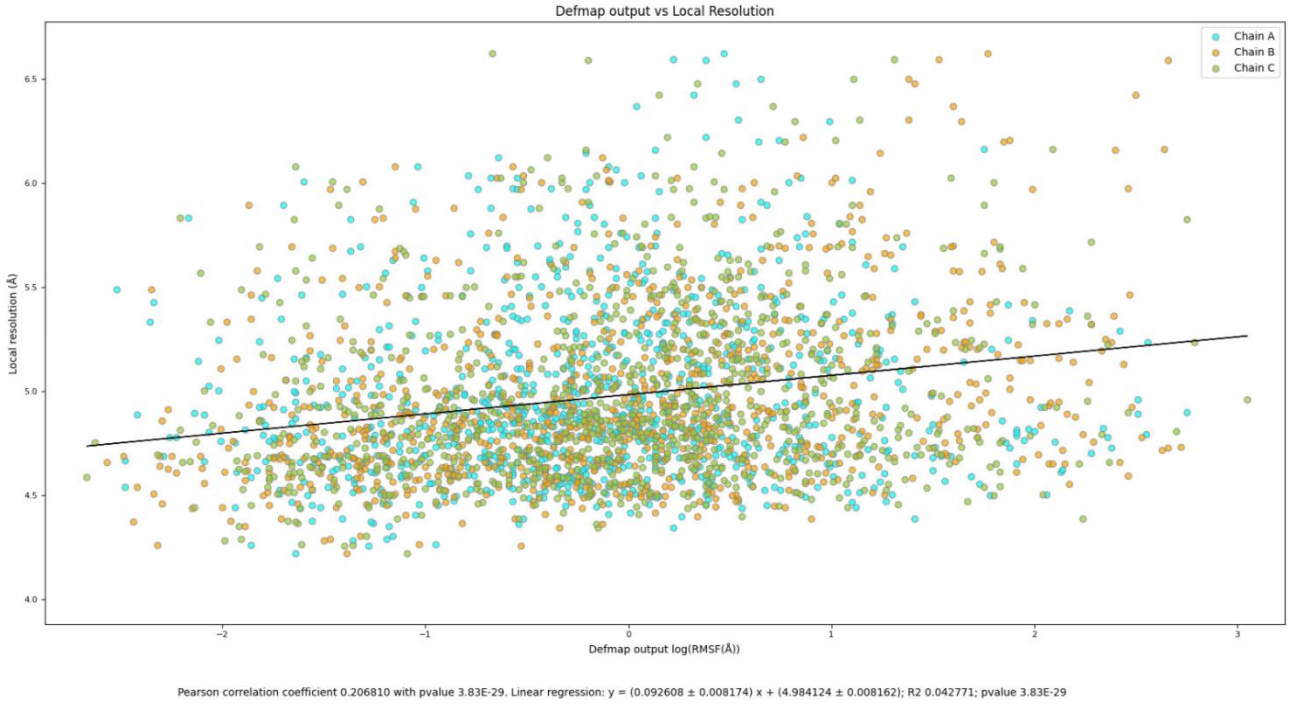


Note. P-values equal to zero means that its real value is lower than the epsilon machine value. (A) represents the values for 6vyb structure. Pearson's coefficient is 0.65, r-squared is 0.42, with a p-values lower than 0.01. (B) shows the values for 7bnn structure. Pearson's coefficient is 0.82, r-squared is 0.68, with a p-values lower than 0.01.

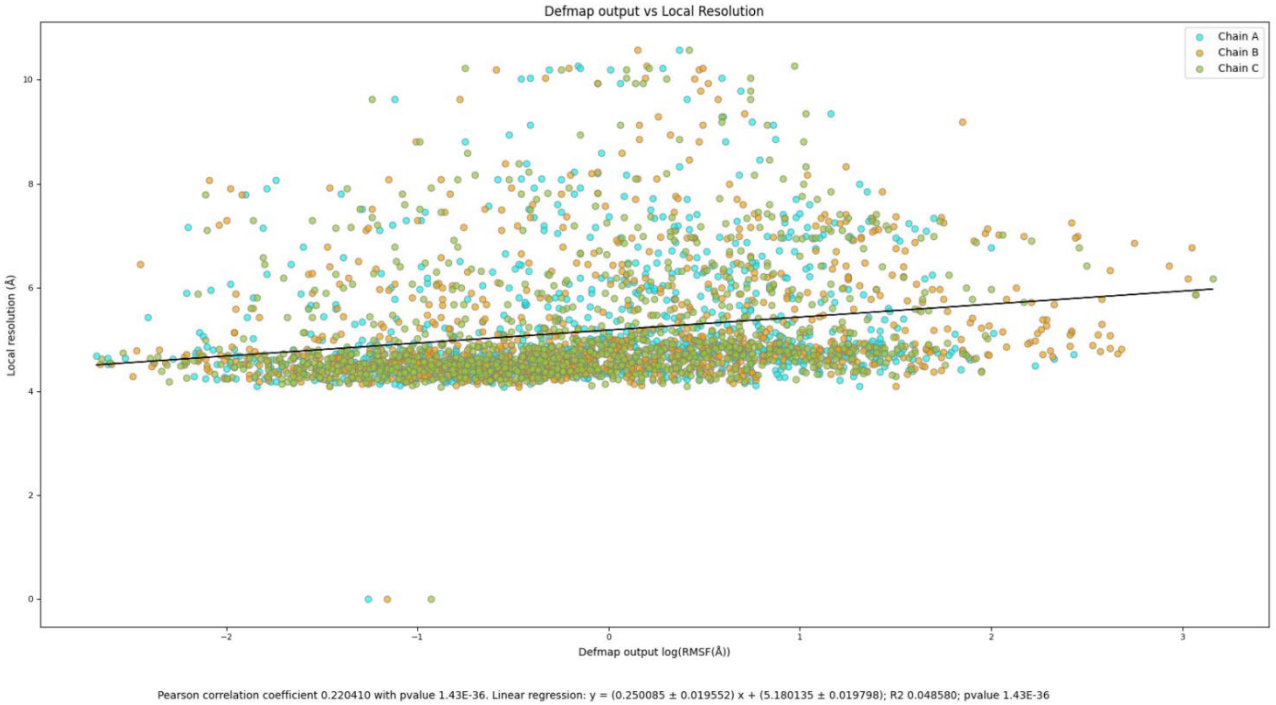
Figure A8

Local resolution of the reference structure against RMSF values in logarithmic scale from the prediction, for the open conformation of Spike.

(A)



(B)



Note. (A) represents the values for 6vyb structure. Pearson's coefficient is 0.21 and r-squared is 0.04 with p-values lower than 0.01. (B) shows the values for 7bnn structure. Pearson's coefficient is 0.22 and r-squared is 0.05 with p-values lower than 0.01.

Figure A9

Representation of the occurrences of the values from the prediction for the closed conformation of Spike.

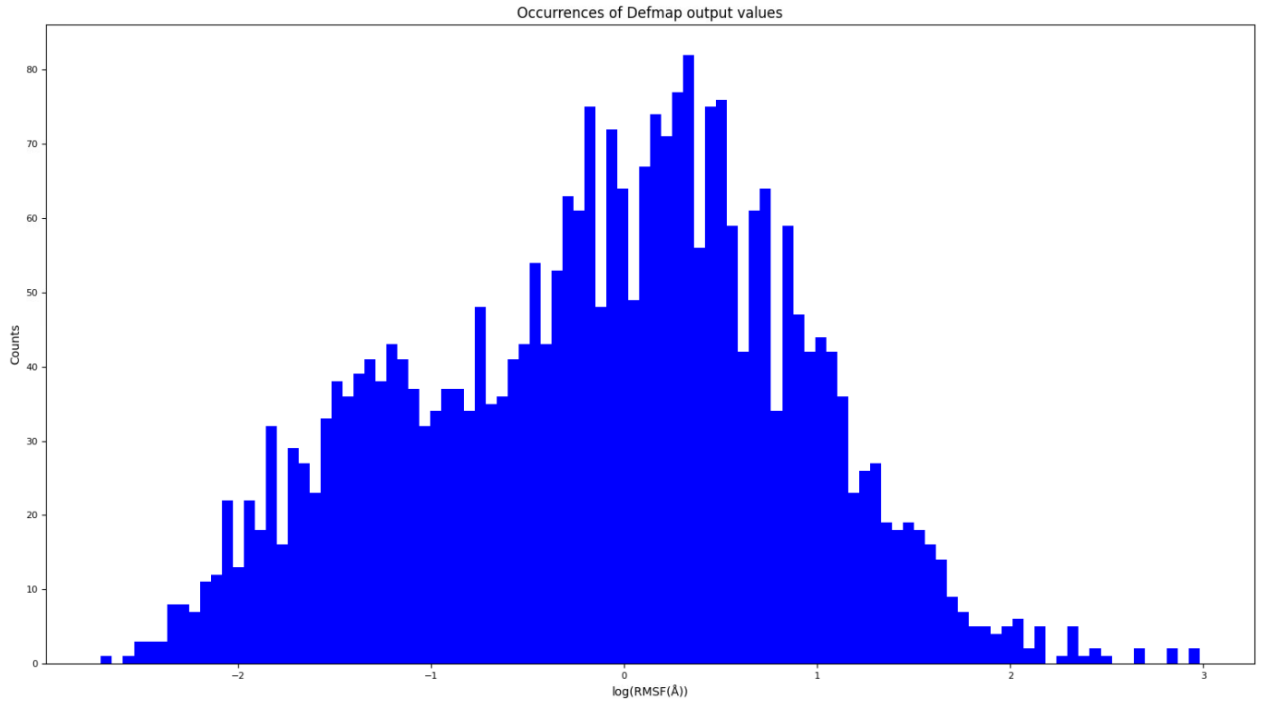
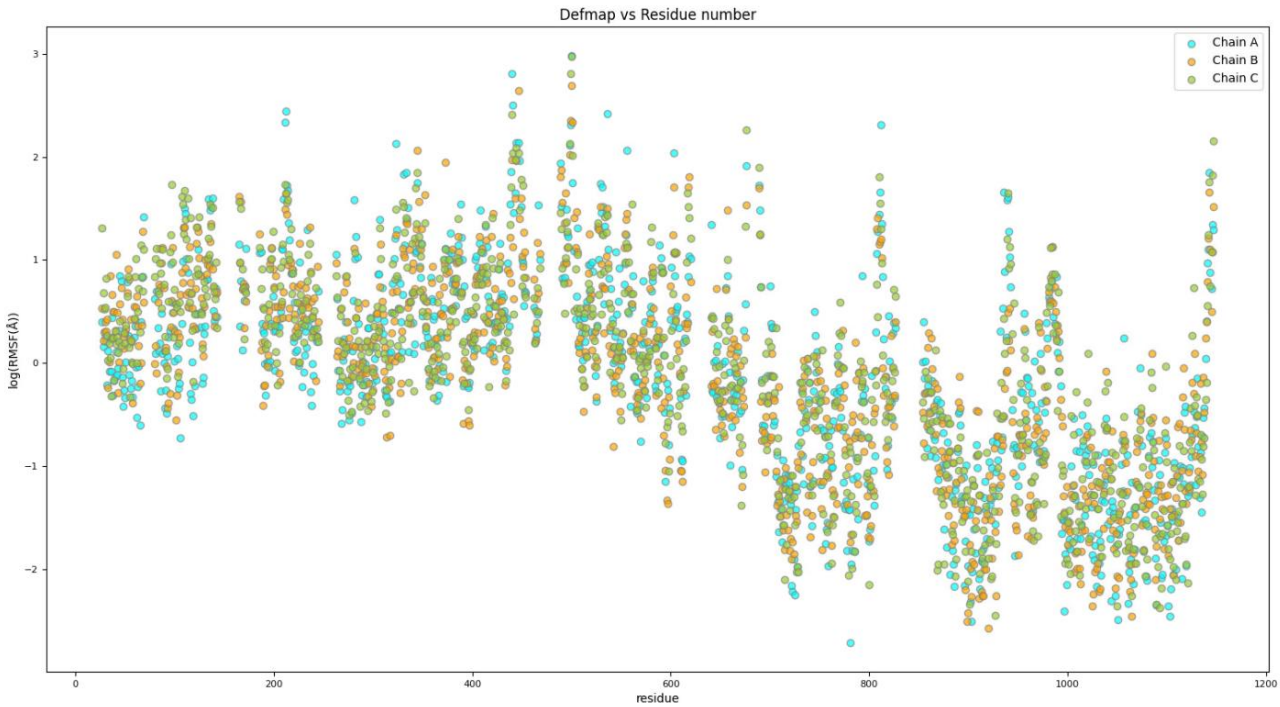


Figure A10

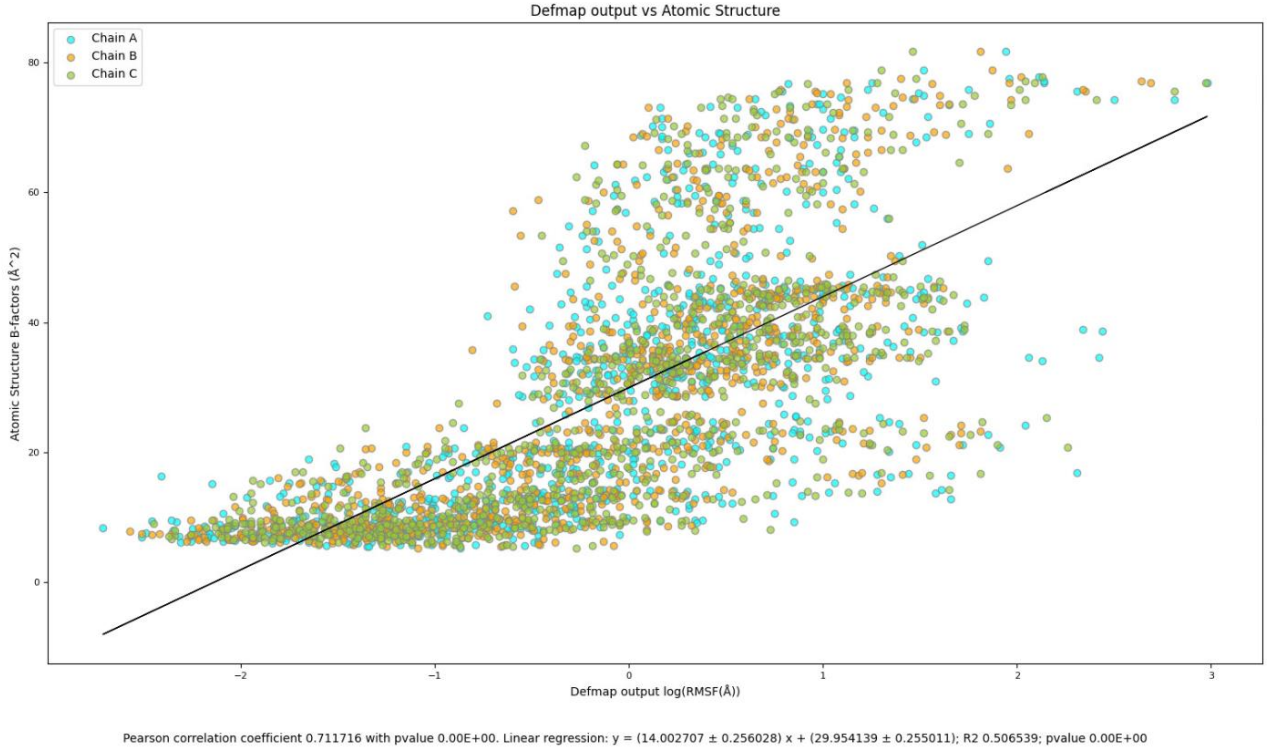
Representation of values from the prediction for closed conformation of Spike against the residue number.



Note. Residues close to the C-terminal have more atoms with low RMSF values.

Figure A11

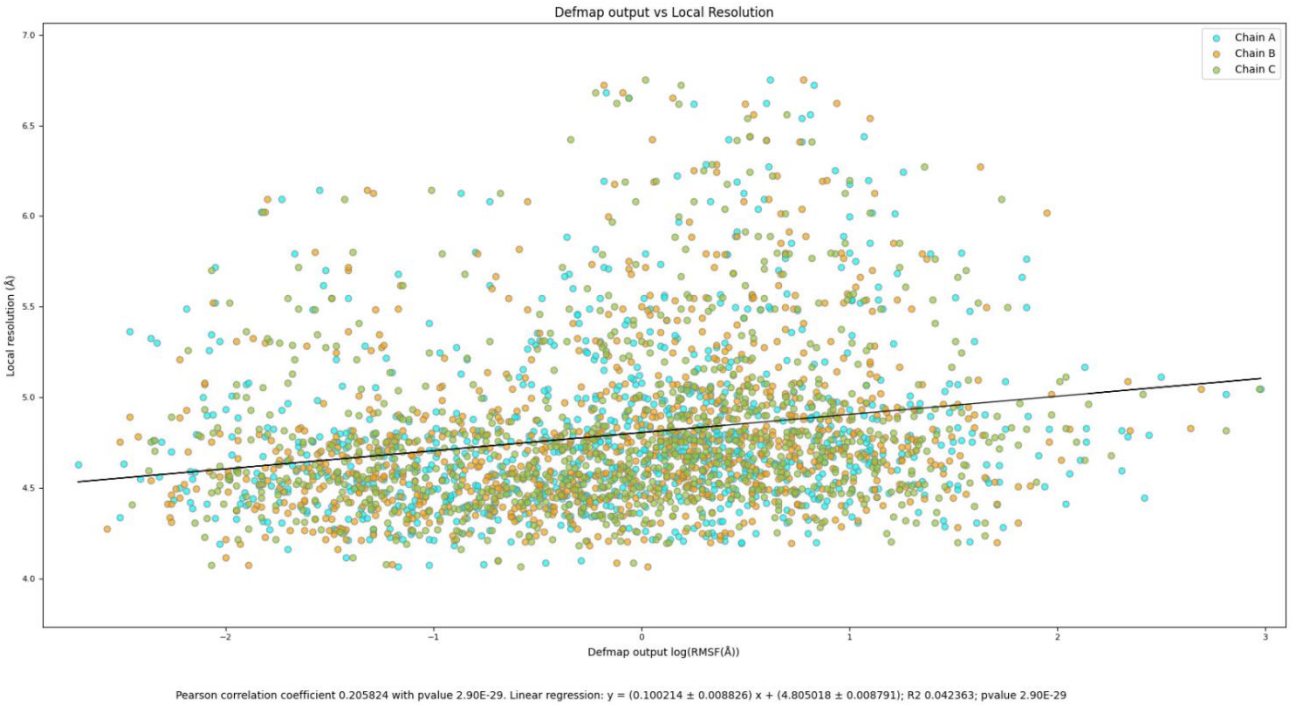
B-factors of the reference structure against values from the prediction for the open conformation of Spike.



Note. P-values equal to zero means that its real value is lower than the epsilon machine value. Pearson's coefficient is 0.71, r-squared is 0.51, with a p-values lower than 0.01.

Figure A12

Local resolution of the reference structure against values from the prediction for the closed conformation of Spike.



Note. Pearson's coefficient is 0.21 and r-squared is 0.04 with p-values lower than 0.01.