Mapping 2D-flexibility and large conformational transitions from HS-AFM with parsimonious data-driven models

Ian Addison-Smith[1,2], Willy Menacho[1], Celica Krigul[1] and Horacio V. Guzman[1*]

[1]Institut de Ciència de Materials de Barcelona, CSIC, 08193 Barcelona, Spain
[2]Universidad de Chile, Beauchef 851, Santiago, Chile

Abstract

AFM has emerged as a transformative experimental platform capable of directly probing biophysical properties such as elasticity with unprecedented spatial and force sensitivity. This opens crucial avenues for **scientific discovery of "real" (i.e. ambient conditions) dynamical properties of biomacromolecules**. In particular for those biological ensembles that exhibit multiple domains, which are the great majority that can be imaged with HS-AFM. However, although HS-AFM experimental groups increase, **the interpretability of their images/videos is still mostly done by eye**, and/or **there are a few "fitting" tools** that **are not considering that the proteinaceous structure deforms upon adsorption, and also may change elastic properties.**

The **problem we tackle in this hackathon is** to develop ***a lightweight and self-contained ML tool*** that helps dissect, interpret and identify both **2D-flexibility and large conformational transitions** of biomacromolecules. In addition, HS-AFM could require external control and/or deep knowledge of the system, which we attempt to substitute with a conceptual CG model that reflects interfacial dynamics with surfaces of different hydropathy. The difference between our simulations and existing fitting tools, is that in our case the effects of the surface are included, and the defined observables are reduced to 2D. Hence, tackling a 2D-flexibility of domains makes it more understandable and interpretable, as well.

Hackathon targets. We have several specific questions for this 3 days of intensive work, namely:

1. Iterative reconstruction of HS-AFM videos using the dynamic mode decomposition (DMD) framework[1,2,3]. Are we able to create an algorithm for this?
   This is important to keep the model lightweight but accurate enough to reproduce gyration radii in 2D (image plane).
2. Are we capable of identifying most flexible regions (domains) of the biomacromolecules, by performing a detailed mapping of the DMD amplitudes and frequencies?
3. Are we capable of identifying most rigid regions (domains) of the biomacromolecules, by performing a detailed mapping of the DMD amplitudes and frequencies?
4. Are we capable of improving resolution gradually, by detecting sub-domains beyond the visible domains of the biomacromolecule, by combining DMD, PCA and TICA methods?
5. Are we able to predict regions (domains) of the biomacromolecule where large conformational transitions occur, so that we could identify molecular substates (microstates)?

Methodology:

In this work, we address the dynamic characterization of biomacromolecules through four complementary approaches: DMD, PCA, TICA, molecular dynamics (MD) simulations, among other diverse pythonic tools. From the HS-AFM measurements of the SARS-CoV-2 spike protein[4], we collect multipage TIFF files that were processed using the Python imaging library to extract individual image frames. Subsequently, these frames were reshaped into a data matrix wherein each column corresponds to a flattened image. The DMD framework was then employed on this matrix, again using an iterative number of modes, to capture the dominant spatial and temporal features in the experimental data. The reconstructed snapshots from the DMD model were visually compared to the original AFM images at selected frames for one spike variants: WT (Figure 1 a-b) in the downward conformation. Furthermore, the nonzero pixel intensities—interpreted as a macromolecular distribution—were used to compute the center of mass and the radius of gyration for each frame. The temporal evolution of these structural metrics and RMSD calculations relative to a reference frame provided a

quantitative measure of the model's performance. As a proof-of-concept for interpreting and mapping the experimental results into molecular resolution with MD simulations, we performed a short MD simulation of the spike protein in the downwards conformation, following the protocols used in ref.[5]
Moreover, experiments of the spike protein show that the 3 RBDs are rotating based on the imaged center of the STEM (As shown in Figure 1 b and c).

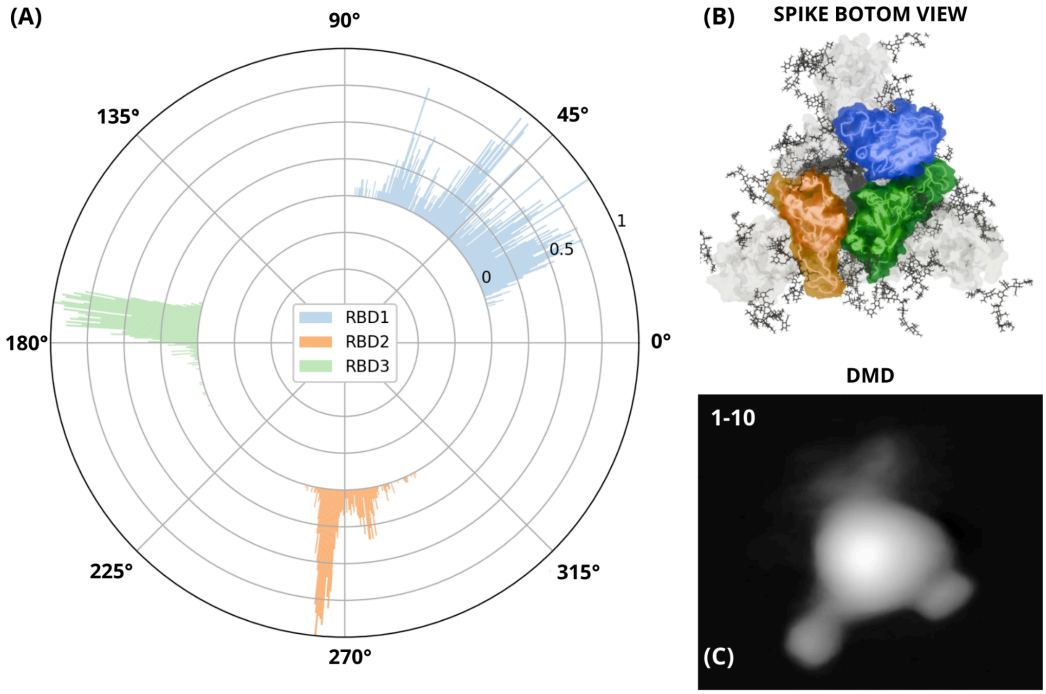

Figure 1. DMD applied to HS-AFM images of the SARS-CoV-2 spike protein. WT variant in downward conformation (A) Polar plots showing the 2D-flexibility of the 3 RBDs comprising the spike protein and (B) Bottom view of the CryoEM structure for the spike protein highlighting the RBDs (C) DMD reconstruction of the WT spike in downwards conformation (snapshots of the video, using the first 10 modes).

Our deliverables are:

1. Through TICA analysis we are now able to **identify the conformations available during the HS-AFM imaging time.**
2. Through PCA analysis we are now able to identify the variance of the domains conforming to the biomacromolecule and determine an average signal for **defining the Center-of-mass of the molecule.**
3. Through DMD analysis, we are able to **identify most flexible regions (domains)** of the **biomacromolecules, most rigid ones**, as well.
4. Through DMD and different modes analysis we are also able **to visualize subdomains** of one of the imaged RBDs.
5. Finally, we can detect microstates of the RBDs, such as open, closed and closed, among others.

References:

1. P. J. Schmid, Annual Review of Fluid Mechanics 54, 225–254 (2022).
2. N. Demo, M. Tezzele, G. Rozza, The Journal of Open Source Software 3, 530 (2018).
3. S. M. Ichinaga, F. Andreuzzi, N. Demo, et al., arXiv (2024).
4. R. Zhu, D. Canena, M. Sikora, et al., Nature Communications 13 (2022)
5. A. M. Bosch, H. V. Guzman, R. Perez, Journal of Chem. Inf. & Modeling 64 (15), 5977–5990 (2024).