Université Paris Cité

Instituts thématiques Inserm

Institut national de la santé et de la recherche médicale

SiMB Dynamique des Systèmes et Interactions des Macromolécules Biologiques

The interest of High-Performance Computing in Molecular Modelling and Structural Bioinformatics:

How molecular simulations and IA have become a key player in biology and chemistry: Some success stories.

A revolution in Biology: Omics technologies

- Genomics: Genome* sequencing, quantification of expression of genes, identification of variants
- Proteomics : Identification, quantification of proteins* within a cell, tissue, organ etc
- Metabolomics: Identification and quantification of metabolites

etc…

A HUGE AMOUNT OF DATA ORGANIZED or NOT in DATABASES

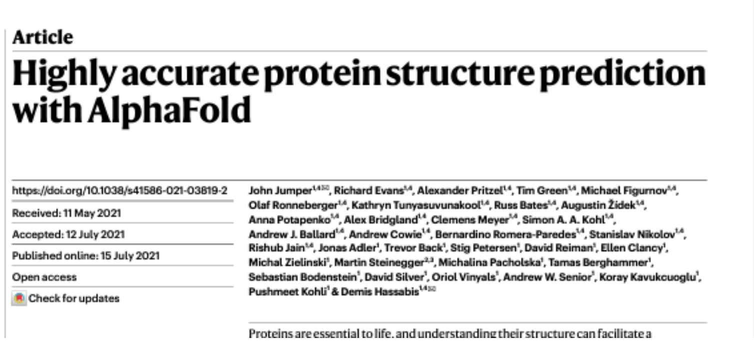
* Genome : All the genetic material of an organism, composed of DNA. A gene codes for a specific protein.
* Protein: Amino acid polymers **folded in a 3D structure** that supports the function*
* Protein Functions: Catalyze biochemical reactions, transport molecules, synthesize and repair DNA, receive and send chemical signals, respond to stimuli, provide structural support

What to do with this data?

- Machine Learning- Deep Learning Approaches to:
  - Predict the 3D structure of proteins from amino acid sequence (<=> Fonction)
  - Understand and Predict the impact of mutations in relation with a disease

  ⇨ DeepMind Successes (Nobel Prize in Chemistry (2024)), Fair Meta AI
  - Design new proteins (D. Baker, Nobel Prize in Chemistry (2024),

  To name a few…. :::



Article

Highly accurate protein structure prediction with AlphaFold

https://doi.org/10.1038/s41586-021-03819-2

Received: 11 May 2021

Accepted: 12 July 2021

Published online: 15 July 2021

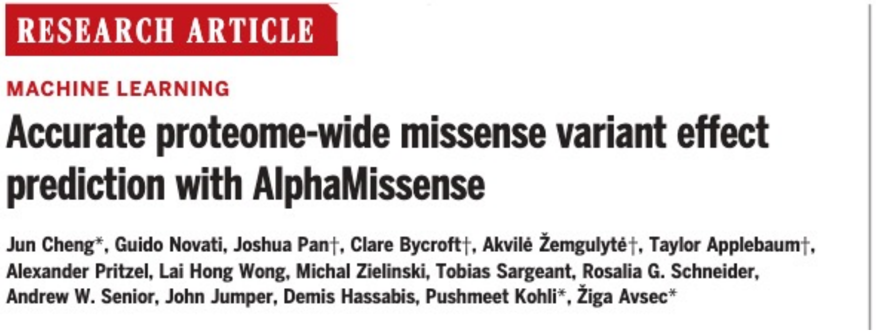
Open access

Check for updates

John Jumper[1,4✉], Richard Evans[1,4], Alexander Pritzel[1,4], Tim Green[1,4], Michael Figurnov[1,4], Olaf Ronneberger[1,4], Kathryn Tunyasuvunakool[1,4], Russ Bates[1,4], Augustin Žídek[1,4], Anna Potapenko[1,4], Alex Bridgland[1,4], Clemens Meyer[1,4], Simon A. A. Kohl[1,4], Andrew J. Ballard[1,4], Andrew Cowie[1,4], Bernardino Romera-Paredes[1,4], Stanislav Nikolov[1,4], Rishub Jain[1,4], Jonas Adler[1], Trevor Back[1], Stig Petersen[1], David Reiman[1], Ellen Clancy[1], Michal Zielinski[1], Martin Steinegger[2,3], Michalina Pacholska[1], Tamas Berghammer[1], Sebastian Bodenstein[1], David Silver[1], Oriol Vinyals[1], Andrew W. Senior[1], Koray Kavukcuoglu[1], Pushmeet Kohli[1] & Demis Hassabis[1,4✉]

Proteins are essential to life, and understanding their structure can facilitate a

Nature, 596, 583-589-(2021)



RESEARCH ARTICLE

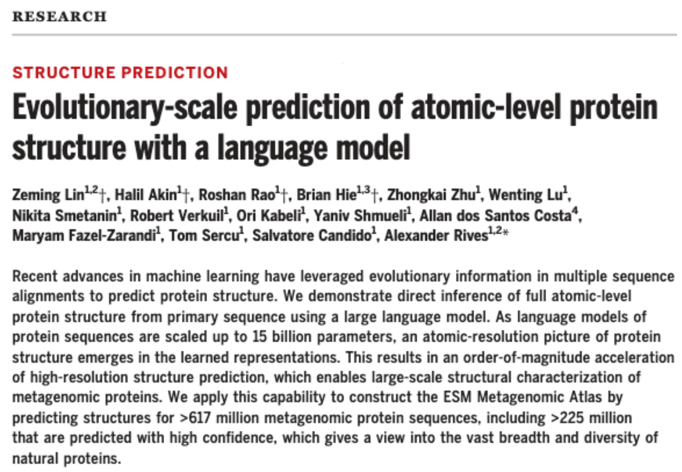
MACHINE LEARNING

Accurate proteome-wide missense variant effect prediction with AlphaMissense

Jun Cheng*, Guido Novati, Joshua Pan†, Clare Bycroft†, Akvilė Žemgulytė†, Taylor Applebaum†, Alexander Pritzel, Lai Hong Wong, Michal Zielinski, Tobias Sargeant, Rosalia G. Schneider, Andrew W. Senior, John Jumper, Demis Hassabis, Pushmeet Kohli*, Žiga Avsec*

Science, 381, 1303 (2023)

Successes but … High Computational Cost and High Memory Ressources



RESEARCH

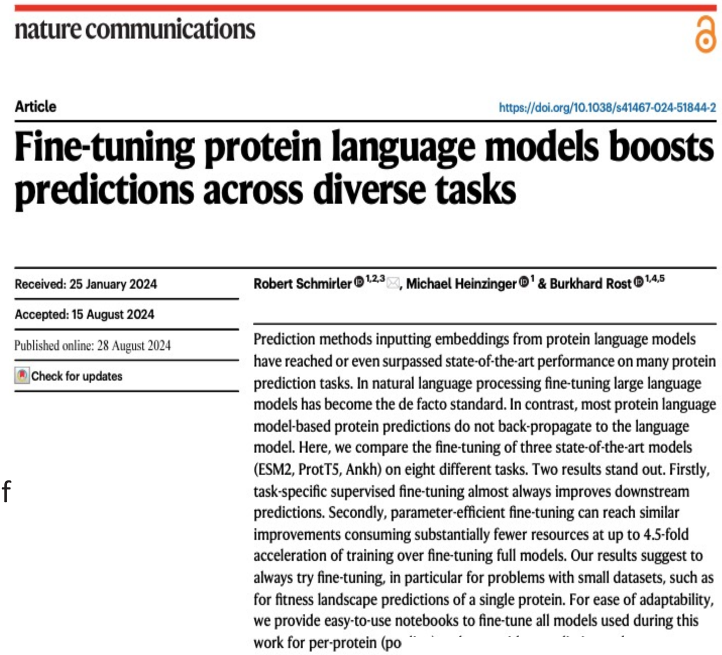
STRUCTURE PREDICTION

Evolutionary-scale prediction of atomic-level protein structure with a language model

Zeming Lin[1,2]†, Halil Akin[1]†, Roshan Rao[1]†, Brian Hie[1,3]†, Zhongkai Zhu[1], Wenting Lu[1], Nikita Smetanin[1], Robert Verkuil[1], Ori Kabeli[1], Yaniv Shmueli[1], Allan dos Santos Costa[4], Maryam Fazel-Zarandi[1], Tom Sercu[1], Salvatore Candido[1], Alexander Rives[1,2]*

Recent advances in machine learning have leveraged evolutionary information in multiple sequence alignments to predict protein structure. We demonstrate direct inference of full atomic-level protein structure from primary sequence using a large language model. As language models of protein sequences are scaled up to 15 billion parameters, an atomic-resolution picture of protein structure emerges in the learned representations. This results in an order-of-magnitude acceleration of high-resolution structure prediction, which enables large-scale structural characterization of metagenomic proteins. We apply this capability to construct the ESM Metagenomic Atlas by predicting structures for >617 million metagenomic protein sequences, including >225 million that are predicted with high confidence, which gives a view into the vast breadth and diversity of natural proteins.

"We present an evolutionary-scale structural characterization of metagenomic proteins that folds practically all sequences in MGnify90 (32), >617 million proteins. We were able to complete this characterization in 2 weeks on a heterogeneous cluster of 2000 graphics processing units (GPUs), which demonstrates scalability to far larger databases"



nature communications

Article

https://doi.org/10.1038/s41467-024-51844-2

Fine-tuning protein language models boosts predictions across diverse tasks

Received: 25 January 2024

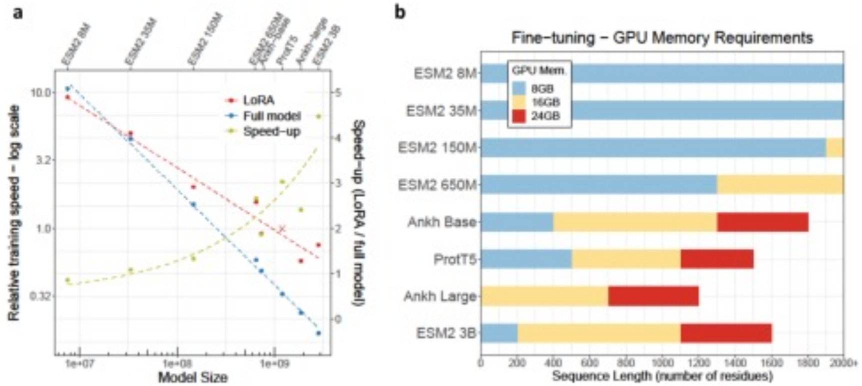
Accepted: 15 August 2024

Published online: 28 August 2024

Check for updates

Robert Schmirler[1,2,3], Michael Heinzinger[1] & Burkhard Rost[1,4,5]

Prediction methods inputting embeddings from protein language models have reached or even surpassed state-of-the-art performance on many protein prediction tasks. In natural language processing fine-tuning large language models has become the de facto standard. In contrast, most protein language model-based protein predictions do not back-propagate to the language model. Here, we compare the fine-tuning of three state-of-the-art models (ESM2, ProtT5, Ankh) on eight different tasks. Two results stand out. Firstly, task-specific supervised fine-tuning almost always improves downstream predictions. Secondly, parameter-efficient fine-tuning can reach similar improvements consuming substantially fewer resources at up to 4.5-fold acceleration of training over fine-tuning full models. Our results suggest to always try fine-tuning, in particular for problems with small datasets, such as for fitness landscape predictions of a single protein. For ease of adaptability, we provide easy-to-use notebooks to fine-tune all models used during this work for per-protein (po



Successes and Limits:

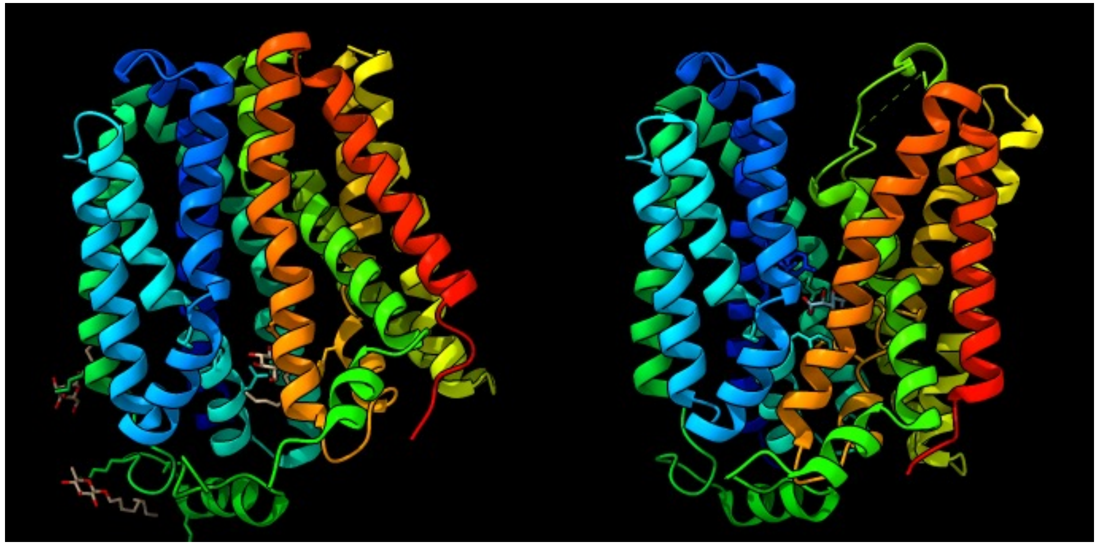
- AlphaFold’s Family: can predict the 3D structures of proteins and complexes (Protein-Protein, Protein-DNA and Protein-ligands) harbouring known folds but also novel protein folds
- AlphaFold’s Family: can’t predict neither the impact of mutation on the 3D structure nor ***alternative conformations*** resulting from the dynamics of the proteins, which is EXTREMELY IMPORTANT FOR THE BIOLOGICAL FUNCTION

(ACTUALLY, it is possible with AlphaFold by manipulating and guiding the search but with very limited success)

Limits: Case of Conformational Changes

Example : The Major Facilitator Family (MFS) Transporters

E. coli D-galactonate:proton symporter



⇨ALTERNATIVE APPROACHES TO ACCESS THIS CONFORMATIONAL LANDSCAPE:

MOLECULAR DYNAMICS SIMULATIONS

EXPLORING THE CONFORMATIONAL LANDSCAPE OF A BIOLOGICAL MACROMOLECULE

- “Ingredients” of Molecular Dynamics Simulations:
  - Based on the solution of the Newton’s Equations (Second Law)
    - The force acting on a particle* is equal to the mass*acceleration :

$$\boxed{\frac{d^2x_i}{dt^2} = \frac{F_{x_i}}{m_i}}$$

    - The force is the equal to the opposite of the gradient of the energy V of this particle interacting with the other particles:

$$\boxed{F_{x_i} = -\frac{\partial V}{\partial x_i}}$$

➥An algorithm able to integrate efficiently motion equations:

➥A model to describe the physical interactions.

* Particle: atom, residue, group of residues => different resolution scales

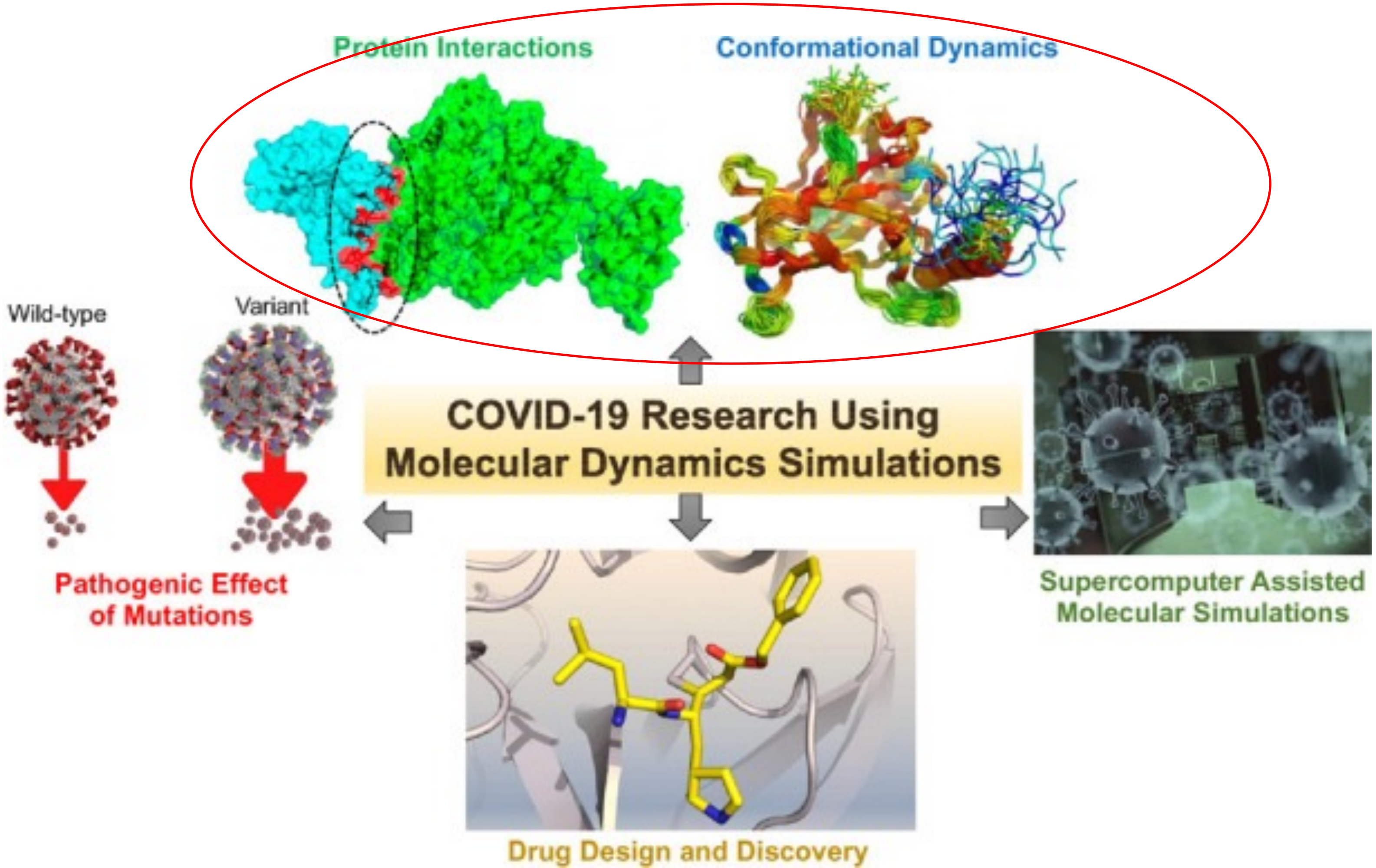
A few examples of applications:

“PRACE support to mitigate impact of COVID-19 pandemic”

- Biomolecular research to understand the mechanisms of the virus infection
- Bioinformatics research to understand mutations, evolution, etc.
- Bio-simulations to develop therapeutics and/or vaccines
- Epidemiologic analysis to understand and forecast the spread of the disease
- Other analyses to understand and mitigate the impact of the pandemic

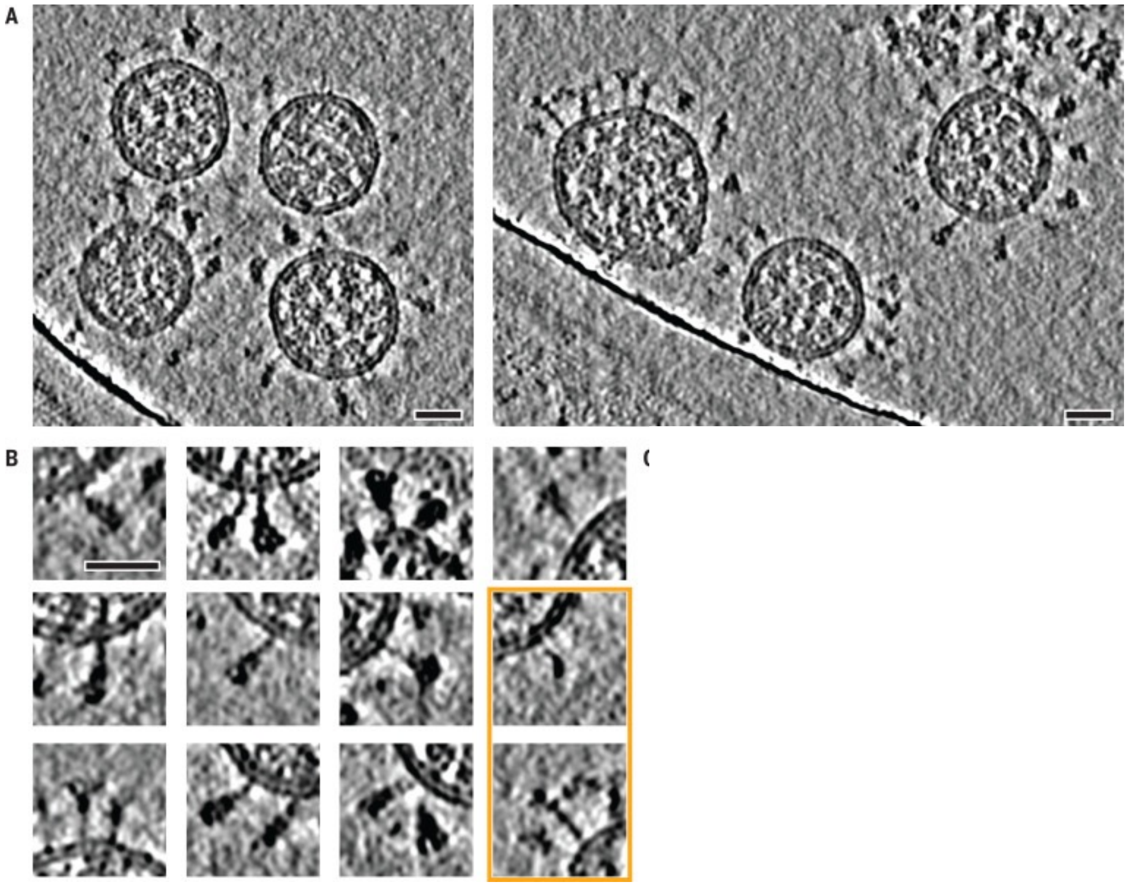
Covid19 Spike2 Protein & Theoretical approaches

Accelerating COVID-19 Research Using Molecular Dynamics Simulation, Aditya K. Padhi, Soumya Lipsa Rath, and Timir Tripathi, The Journal of Physical Chemistry B 2021 125 (32), 9078-9091

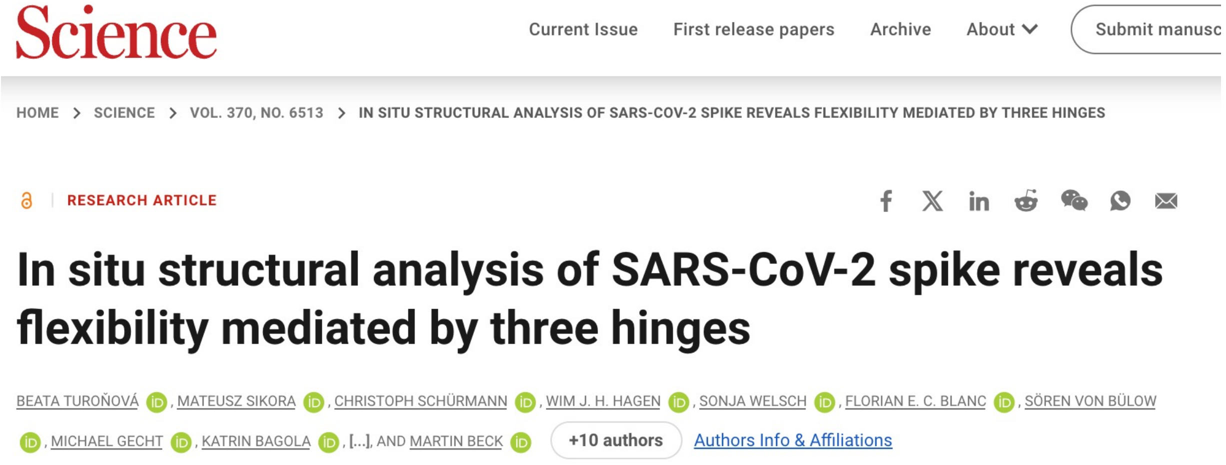


An Emblematic Case: Covid19 Spike2 Protein

Cryo–electron tomography of SARS-CoV-2 virions.



In situ structural analysis of SARS-CoV-2 spike reveals **flexibility** mediated by three hinges



Science Current Issue First release papers Archive About Submit manusc

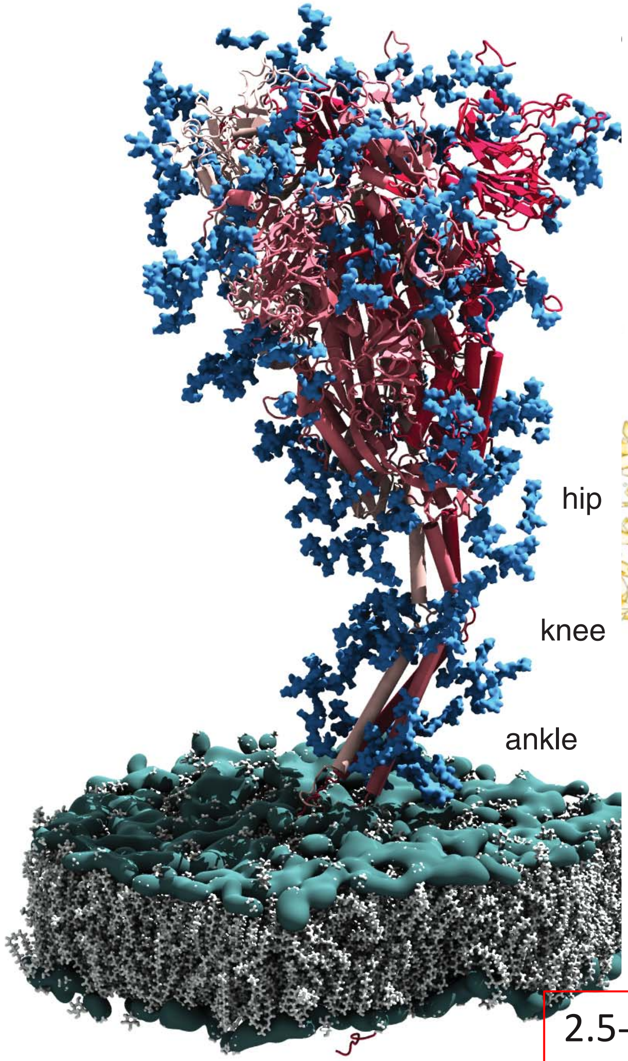
HOME > SCIENCE > VOL. 370, NO. 6513 > IN SITU STRUCTURAL ANALYSIS OF SARS-COV-2 SPIKE REVEALS FLEXIBILITY MEDIATED BY THREE HINGES

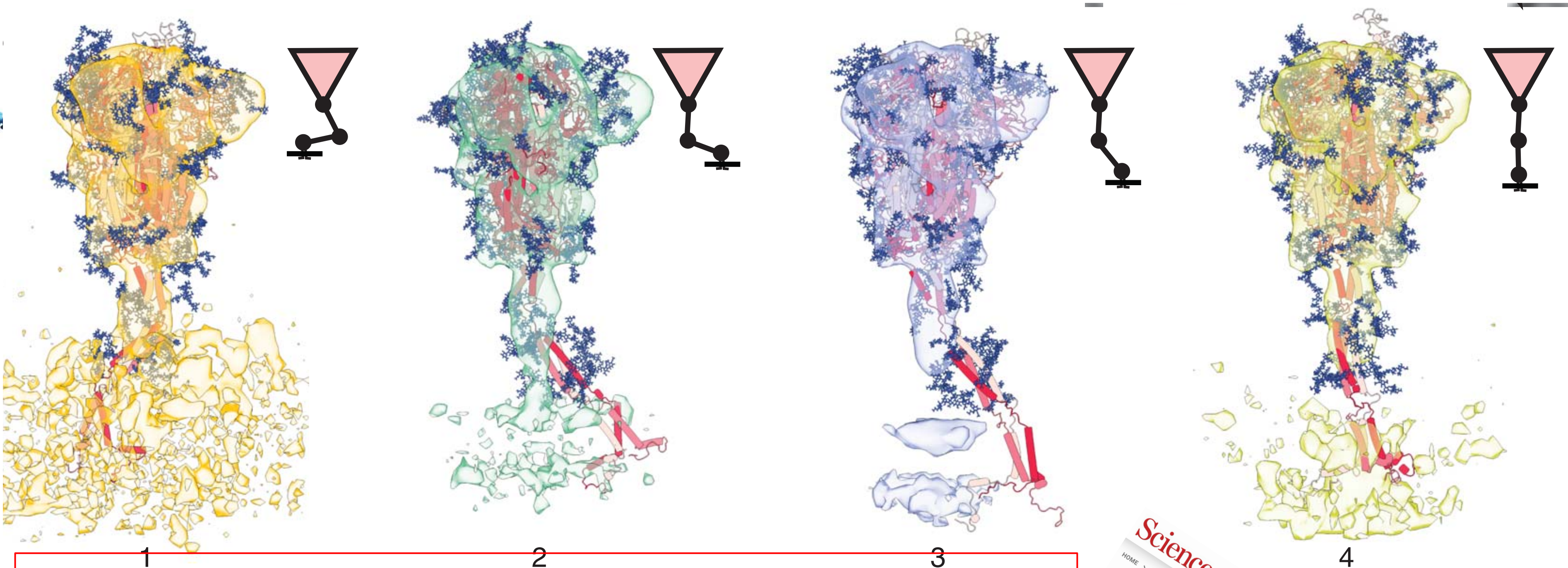
RESEARCH ARTICLE

In situ structural analysis of SARS-CoV-2 spike reveals flexibility mediated by three hinges

BEATA TUROŇOVÁ, MATEUSZ SIKORA, CHRISTOPH SCHÜRMANN, WIM J. H. HAGEN, SONJA WELSCH, FLORIAN E. C. BLANC, SÖREN VON BÜLOW, MICHAEL GECHT, KATRIN BAGOLA, [...], AND MARTIN BECK +10 authors Authors Info & Affiliations

Molecular dynamics simulations coupled to experiments.





Fit of snapshots of MD simulations into the classes obtained for different distances of the head from the membrane (1 to 4), calculated from different tomograms. Shorter distances are concomitant with a stronger bending of the hinges and a lateral displacement of the stalk. (Fig. 4 from *Science* 2020, 370(6513): 203–208.)

2.5-μs-long all-atom MD simulation of a 4.1 million atom system containing four glycosylated S proteins anchored into a patch of viral membrane and embedded in aqueous solvent (Fig. 3 from *Science* 2020, 370(6513): 203–208.

Science
HOME > SCIENCE > VOL. 370, NO. 6513 > IN SITU STRUCTURAL ANALYSIS OF SARS-COV-2 SPIKE REVEALS FLEXIBILITY MEDIATED BY THREE HINGES
Current Issue
First release papers
Archive
About
Submit manusc

RESEARCH ARTICLE

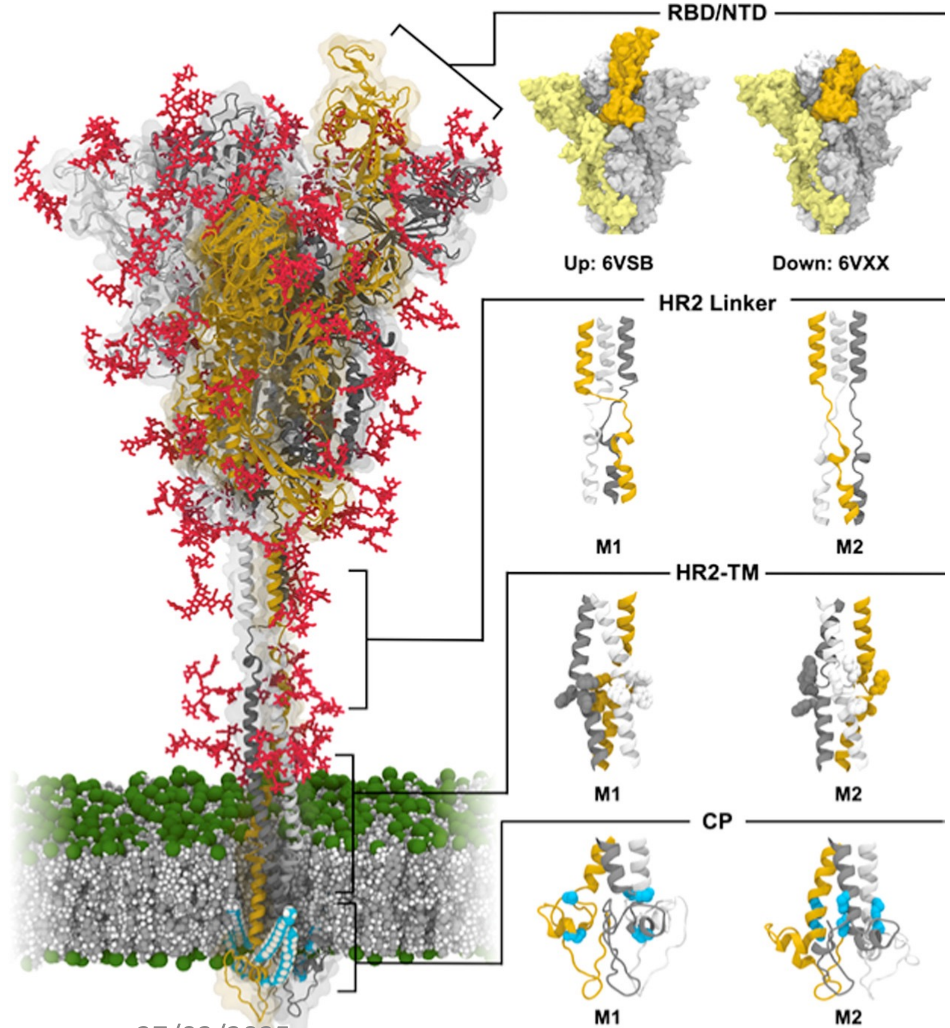
In situ structural analysis of SARS-CoV-2 spike reveals flexibility mediated by three hinges

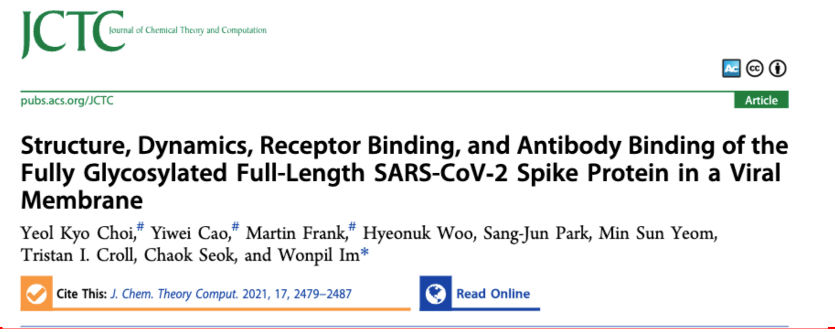
BEATA TUROŇOVÁ, MICHAEL GECHT, MATEUSZ SIKORA, CHRISTOPH SCHÜRMANN, KATRIN BAGOLA, [...], AND MARTIN BECK, WIM J. H. HAGEN, SONJA WELSCH, FLORIAN E. C. BLANC, SÖREN VON BÜLOW

+10 authors

Authors Info & Affiliations

Fully Glycosylated Full-Length SARS-CoV-2 Spike Protein in a Viral Membrane





JCTC Journal of Chemical Theory and Computation

pubs.acs.org/JCTC

Article

Structure, Dynamics, Receptor Binding, and Antibody Binding of the Fully Glycosylated Full-Length SARS-CoV-2 Spike Protein in a Viral Membrane

Yeol Kyo Choi,# Yiwei Cao,# Martin Frank,# Hyeonuk Woo, Sang-Jun Park, Min Sun Yeom, Tristan I. Croll, Chaok Seok, and Wonpil Im*

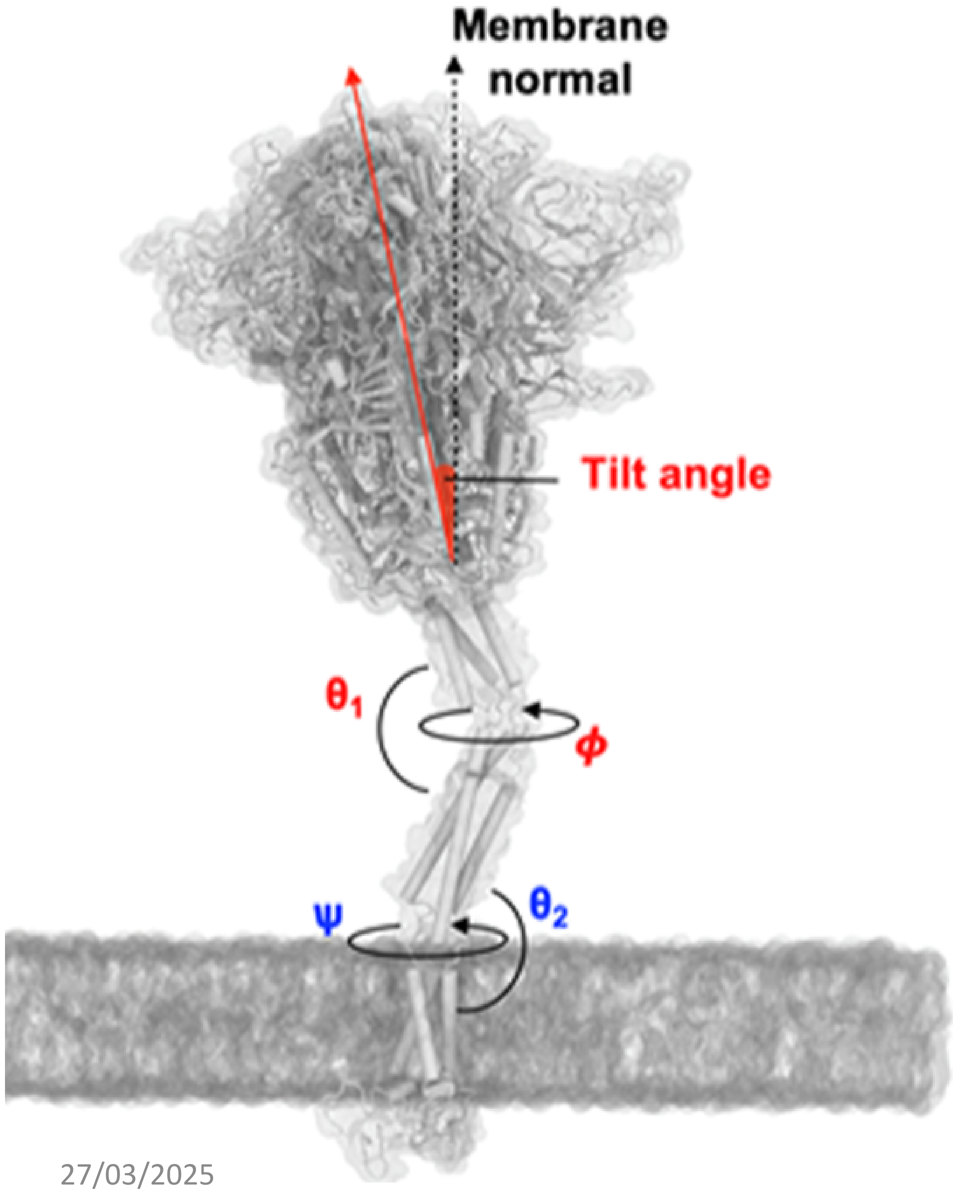
Cite This: *J. Chem. Theory Comput.* 2021, 17, 2479−2487

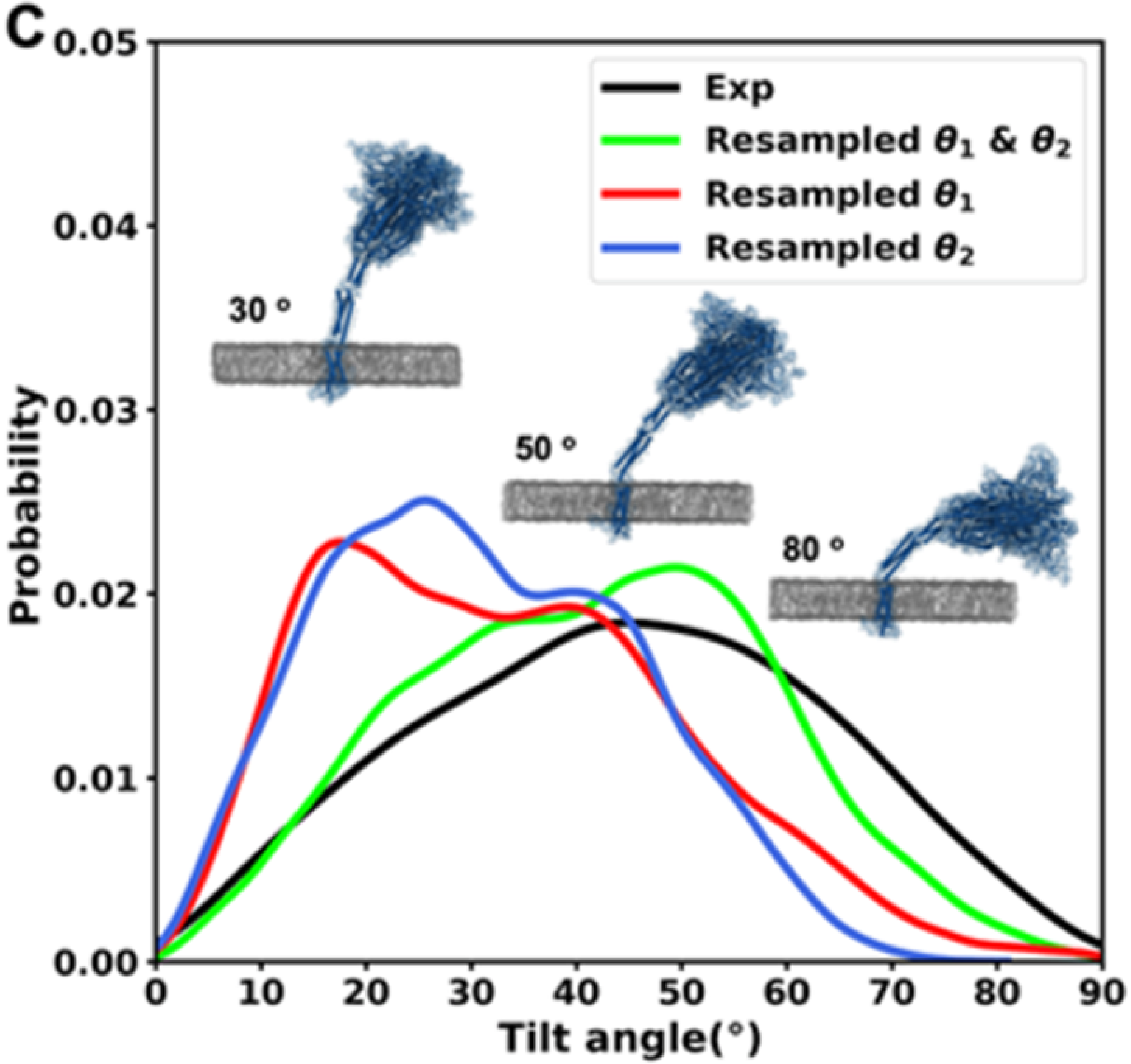
Read Online

Three monomers composed of 2 subunits S1 (Responsible of Receptor Binding) & S2 (membrane fusion) separated by a cleavage site.

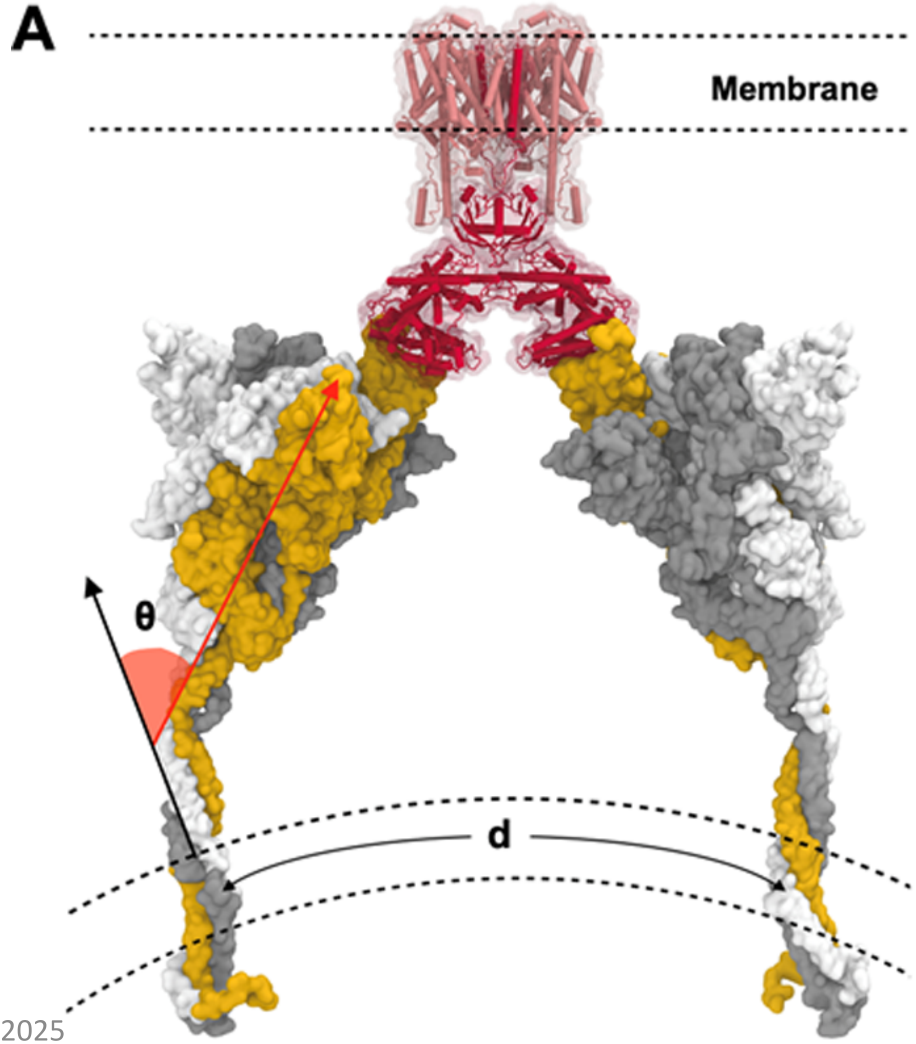
S1: Signal Peptide, two models (Up & Down) for Receptor Binding Domain(RBD)/Nter Domain (NTD) ,
S2: Fusion Peptide, HR2 (Heptad repeat) linker, HR2-TM (Transmembrane Region) , and Cytoplasmic (CP)

All-atom MD simulations of the fully glycosylated full-length S protein in a viral bilayer, multiple µs-long trajectories: RBD in open and closed states, Different models of S stalk (16 models), Glycosylated and non-glycosylated S head-only systems.


Membrane normal
Tilt angle
θ1
ϕ
ψ
θ2

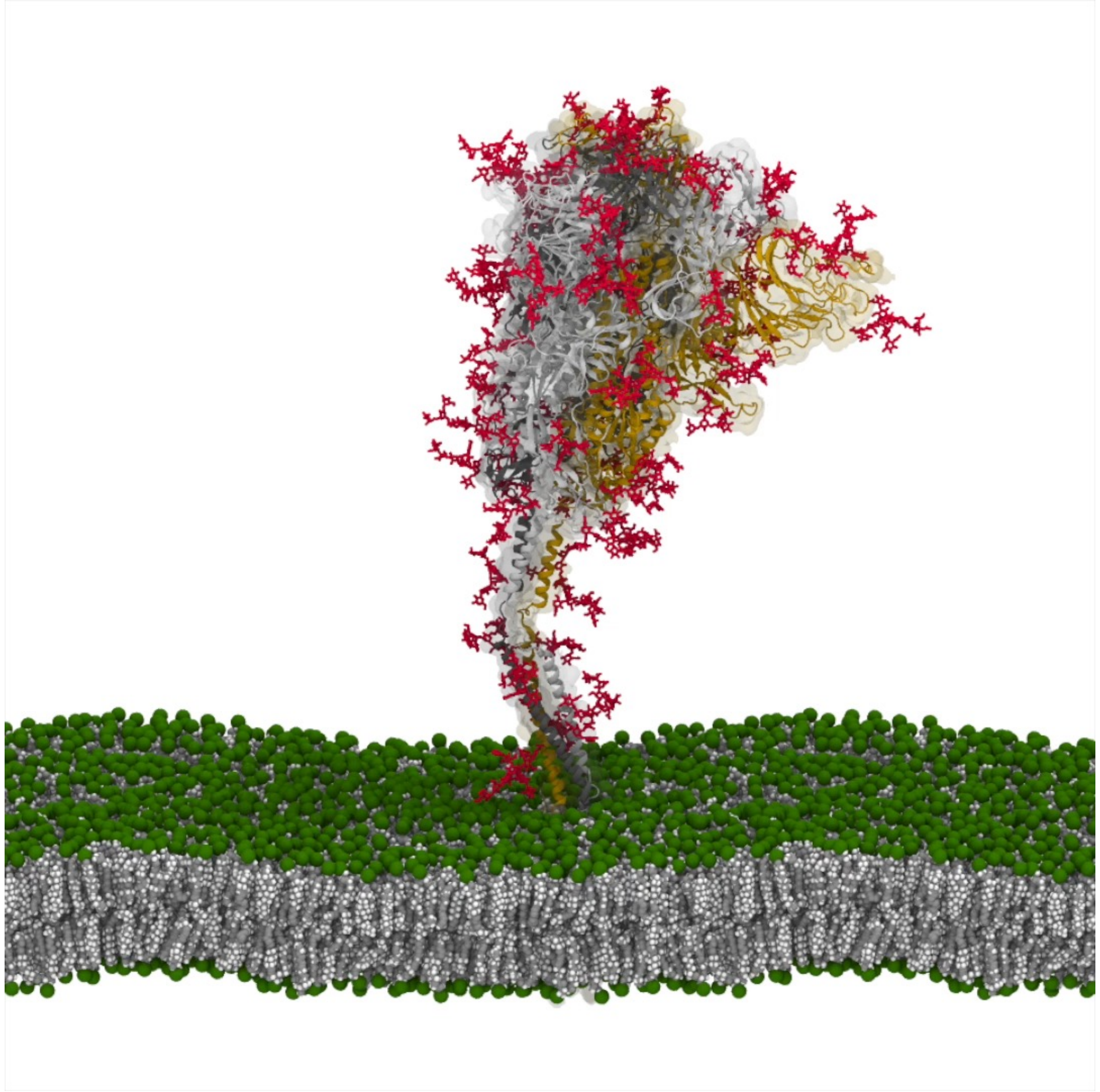

C
Probability
Tilt angle(°)
0.00
0.01
0.02
0.03
0.04
0.05
0
10
20
30
40
50
60
70
80
90
Exp
Resampled θ1 & θ2
Resampled θ1
Resampled θ2
30 °
50 °
80 °

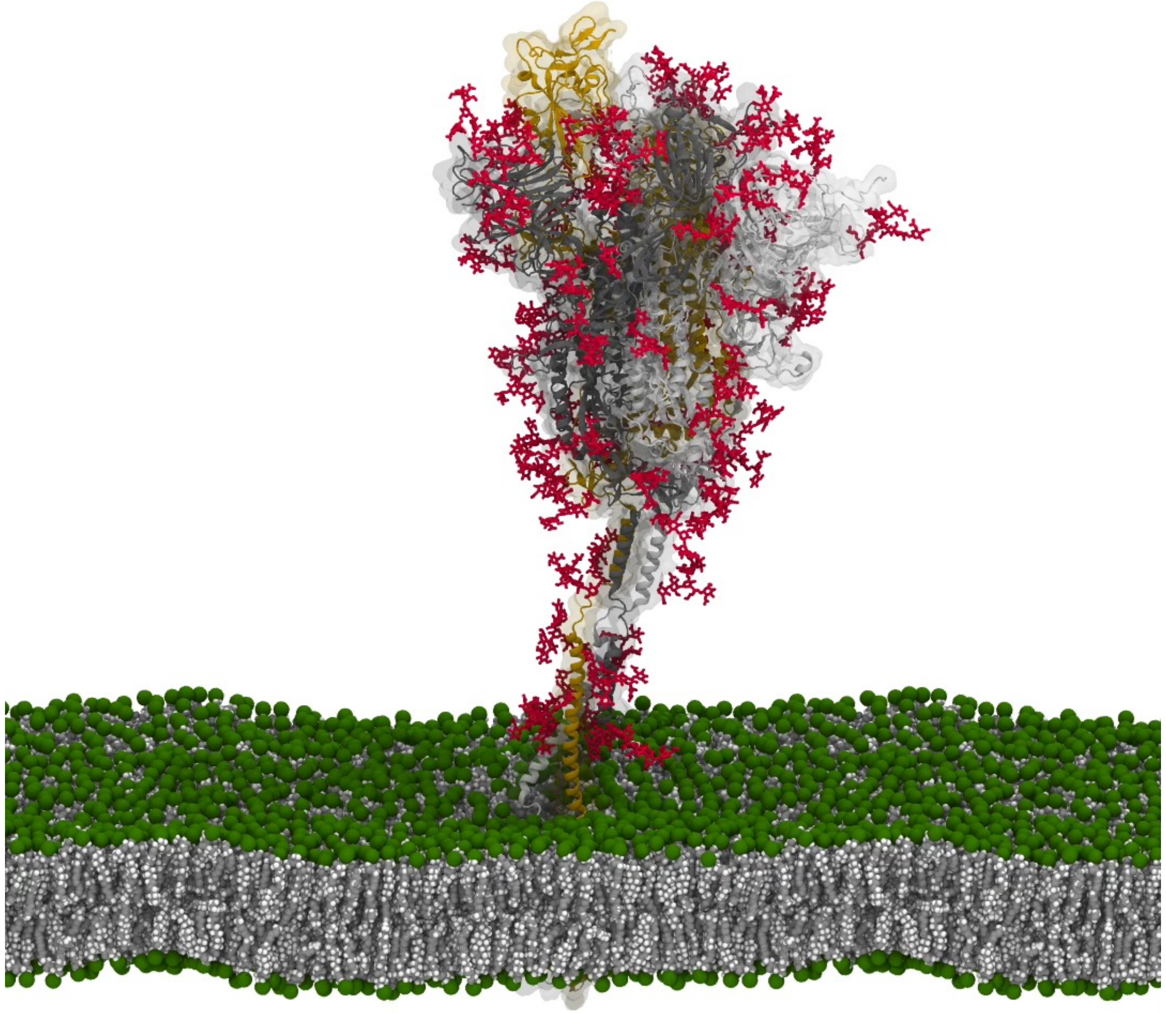
A putative model of Spike with ACE2 receptor



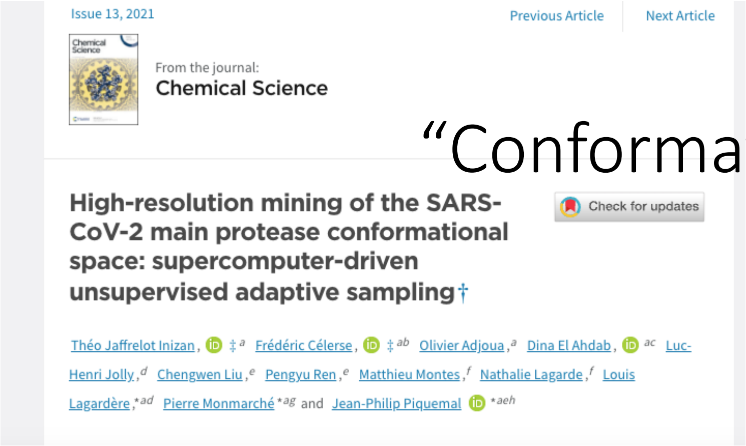
Important Results

- Glycan Impact:
  - (some) on RBD and NTD Motions => S Trimer Stability
  - Shields for immune evasion
  - Contribution to antibody binding.





A Granted PRACE Project:
“Conformational spaces of SARS-CoV-2 drug targets”
J.P Piquemal, Sorbonne University



Issue 13, 2021

Previous Article | Next Article

From the journal:
Chemical Science

High-resolution mining of the SARS-CoV-2 main protease conformational space: supercomputer-driven unsupervised adaptive sampling†

Check for updates

Théo Jaffrelot Inizan, ‡[a] Frédéric Célerse, ‡[ab] Olivier Adjoua,[a] Dina El Ahdab, [ac] Luc-Henri Jolly,[d] Chengwen Liu,[e] Pengyu Ren,[e] Matthieu Montes,[f] Nathalie Lagarde,[f] Louis Lagardère,*[ad] Pierre Monmarché*[ag] and Jean-Philip Piquemal *[aeh]

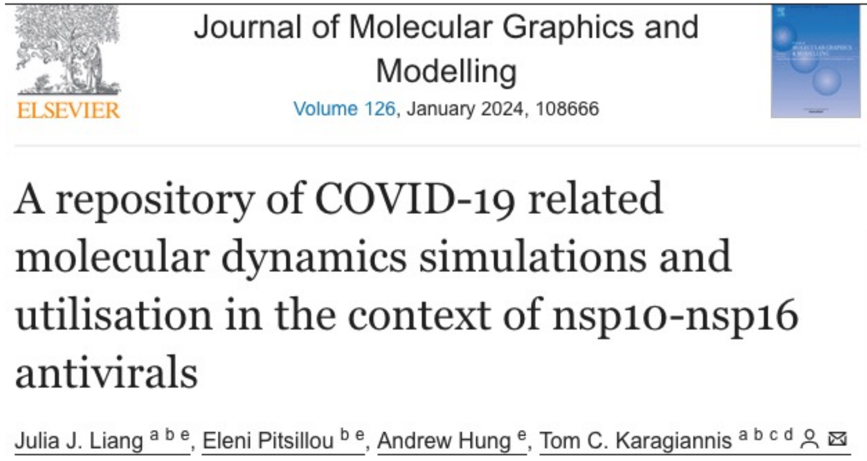
Main features:

- The use of a polarizable force field, which is supposed to overcome current force field limitations
- A density-driven unsupervised adaptive sampling method that **exploits pre–exascale machine and 100 GPUs**

Computer Resources: HPE Jean Zay Supercomputer (IDRIS, GENCI, France): 15.14 µs in two weeks.

Main Results

- Efficient Sampling
- Role of water molecules
- Validation of some results with experimental data
- Identification of a new druggable pocket.

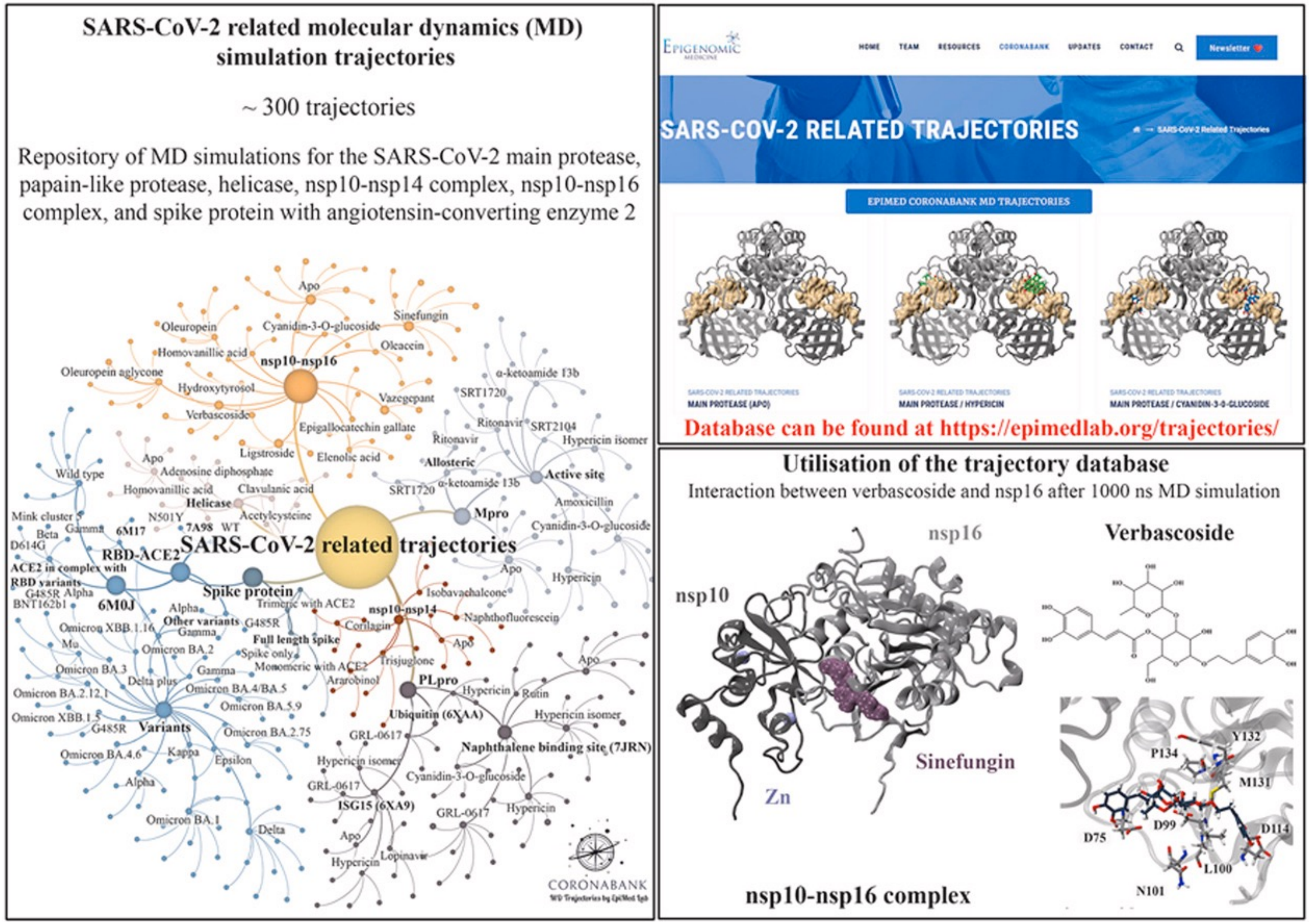


Journal of Molecular Graphics and Modelling
Volume 126, January 2024, 108666

A repository of COVID-19 related molecular dynamics simulations and utilisation in the context of nsp10-nsp16 antivirals

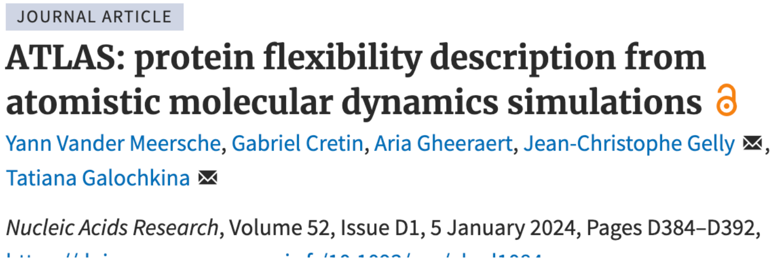
Julia J. Liang [a b e], Eleni Pitsillou [b e], Andrew Hung [e], Tom C. Karagiannis [a b c d]

300 Classical MD. : High performance computing services



SARS-CoV-2 related molecular dynamics (MD) simulation trajectories
~ 300 trajectories
Repository of MD simulations for the SARS-CoV-2 main protease, papain-like protease, helicase, nsp10-nsp14 complex, nsp10-nsp16 complex, and spike protein with angiotensin-converting enzyme 2
SARS-CoV-2 related trajectories
nsp10-nsp16
Mpro
Spike protein
RBD-ACE2
Helicase
nsp10-nsp14
PLpro
Variants
Apo
Oleuropein
Cyanidin-3-O-glucoside
Sinefungin
Homovanillic acid
Oleacein
Oleuropein aglycone
Hydroxytyrosol
Verbascoside
Vazegepant
Epigallocatechin gallate
Ligstroside
Elenolic acid
Allosteric
Active site
α-ketoamide 13b
Ritonavir
Hypericin
Ubiquitin (6XAA)
Naphthalene binding site (7JRN)
ISG15 (6XA9)
GRL-0617
Omicron BA.1
Delta
Alpha
CORONABANK
EPIGENOMIC MEDICINE
HOME TEAM RESOURCES CORONABANK UPDATES CONTACT Newsletter
SARS-COV-2 RELATED TRAJECTORIES
EPIMED CORONABANK MD TRAJECTORIES
MAIN PROTEASE (APO)
MAIN PROTEASE / HYPERICIN
MAIN PROTEASE / CYANIDIN-3-O-GLUCOSIDE
Database can be found at https://epimedlab.org/trajectories/
Utilisation of the trajectory database
Interaction between verbascoside and nsp16 after 1000 ns MD simulation
nsp16
nsp10
Zn
Sinefungin
nsp10-nsp16 complex
Verbascoside
Y132
P134
M131
D75
D99
D114
L100
N101

Beyond Covid :

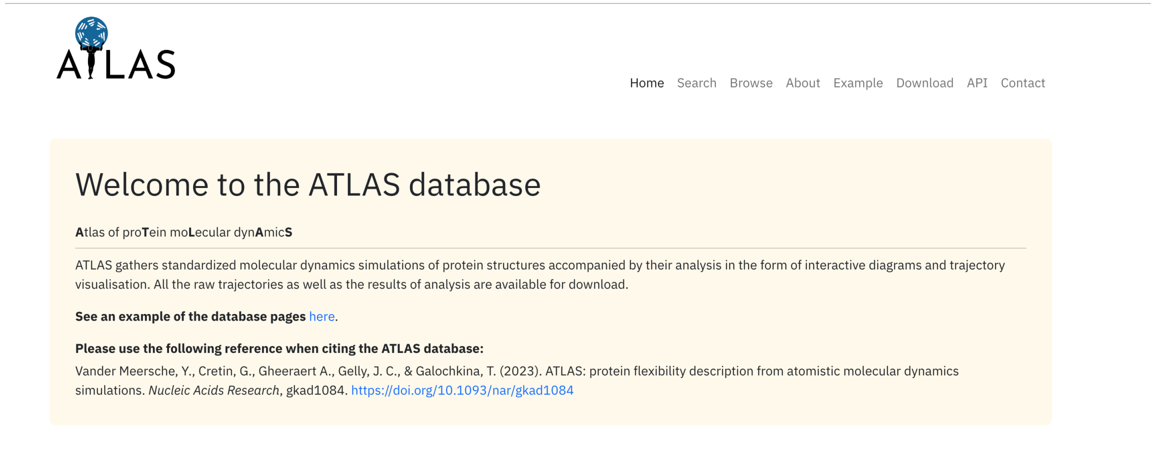


JOURNAL ARTICLE

ATLAS: protein flexibility description from atomistic molecular dynamics simulations

Yann Vander Meersche, Gabriel Cretin, Aria Gheeraert, Jean-Christophe Gelly ✉, Tatiana Galochkina ✉

Nucleic Acids Research, Volume 52, Issue D1, 5 January 2024, Pages D384–D392,



Welcome to the ATLAS database

Atlas of pro**T**ein mo**L**ecular dyn**A**mic**S**

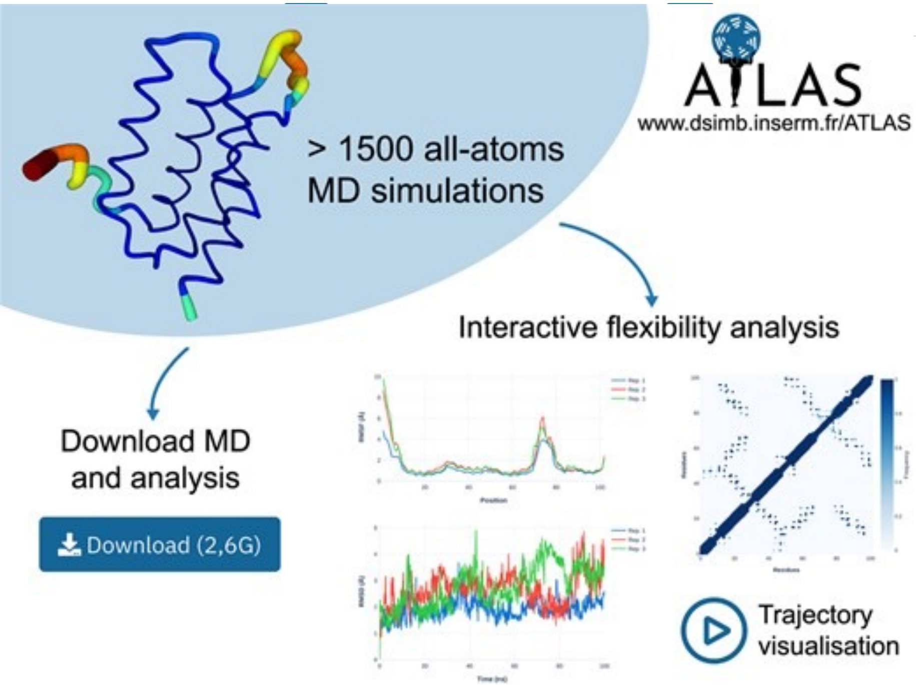
ATLAS gathers standardized molecular dynamics simulations of protein structures accompanied by their analysis in the form of interactive diagrams and trajectory visualisation. All the raw trajectories as well as the results of analysis are available for download.

See an example of the database pages here.

Please use the following reference when citing the ATLAS database:

Vander Meersche, Y., Cretin, G., Gheeraert A., Gelly, J. C., & Galochkina, T. (2023). ATLAS: protein flexibility description from atomistic molecular dynamics simulations. *Nucleic Acids Research*, gkad1084. https://doi.org/10.1093/nar/gkad1084

The user can c



ATLAS
www.dsimb.inserm.fr/ATLAS
> 1500 all-atoms MD simulations
Interactive flexibility analysis
Download MD and analysis
Download (2,6G)
Trajectory visualisation

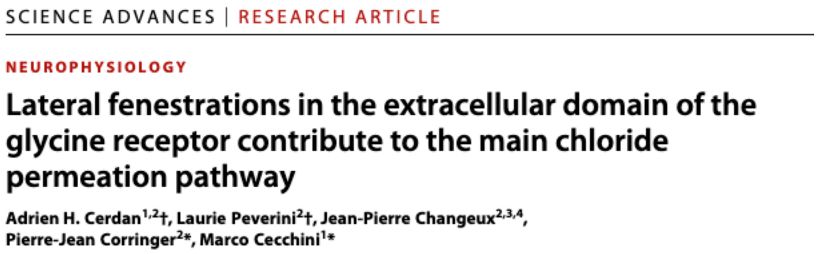
ATLAS: A database collecting protein MD dynamics simulations:

~2000 proteins, 100 ns x 3 replicates

10 M Hours-Cpu GENCI Juliot-Curie's Irene Rome supercomputer (TGCC/CEA), utilising dual-processor compute nodes running at 2.6 GHz with 64 cores per processor.

A few other important biological questions:

Identification of the ion pathway throug the Glycine Receptor



SCIENCE ADVANCES | RESEARCH ARTICLE

NEUROPHYSIOLOGY

Lateral fenestrations in the extracellular domain of the glycine receptor contribute to the main chloride permeation pathway

Adrien H. Cerdan[1,2]†, Laurie Peverini[2]†, Jean-Pierre Changeux[2,3,4], Pierre-Jean Corringer[2]*, Marco Cecchini[1]*

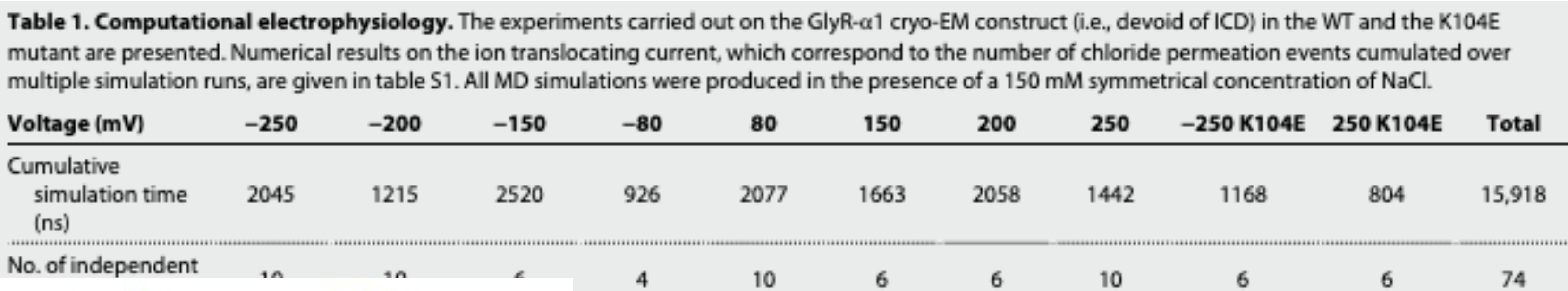
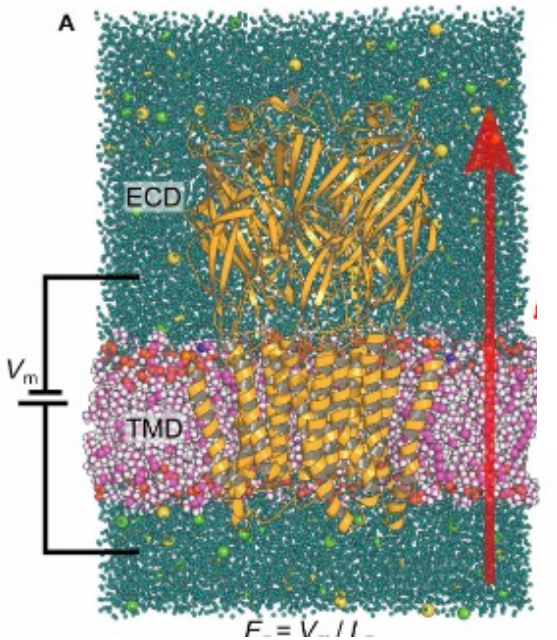
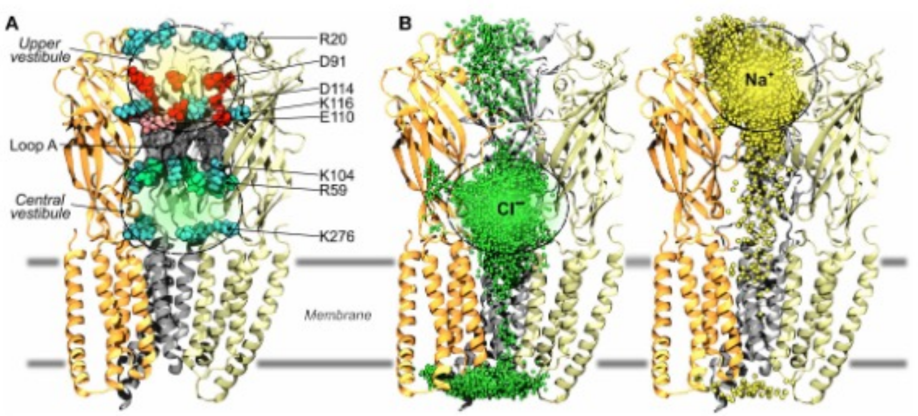


Table 1. Computational electrophysiology. The experiments carried out on the GlyR-α1 cryo-EM construct (i.e., devoid of ICD) in the WT and the K104E mutant are presented. Numerical results on the ion translocating current, which correspond to the number of chloride permeation events cumulated over multiple simulation runs, are given in table S1. All MD simulations were produced in the presence of a 150 mM symmetrical concentration of NaCl.

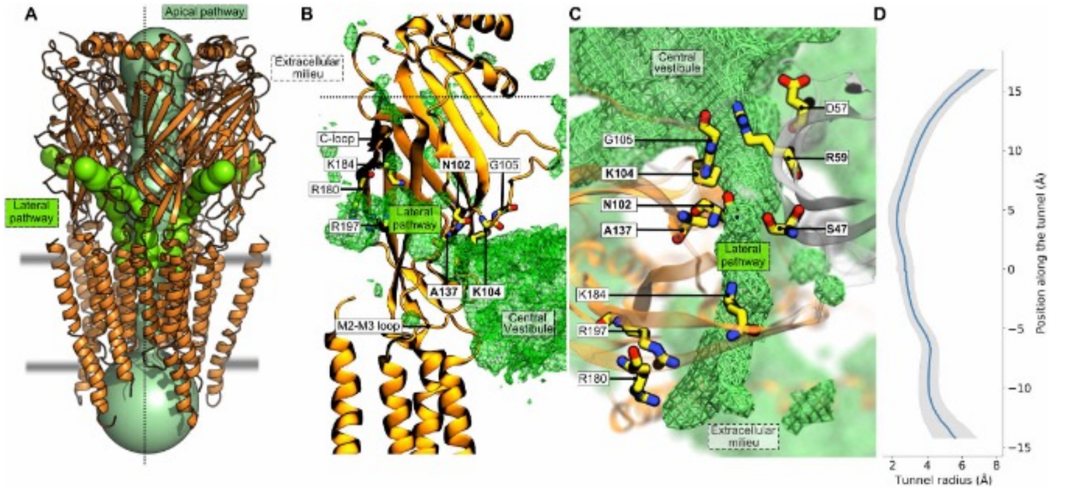
Voltage (mV)	−250	−200	−150	−80	80	150	200	250	−250 K104E	250 K104E	Total
Cumulative simulation time (ns)	2045	1215	2520	926	2077	1663	2058	1442	1168	804	15,918
No. of independent	10	10	6	4	10	6	6	10	6	6	74



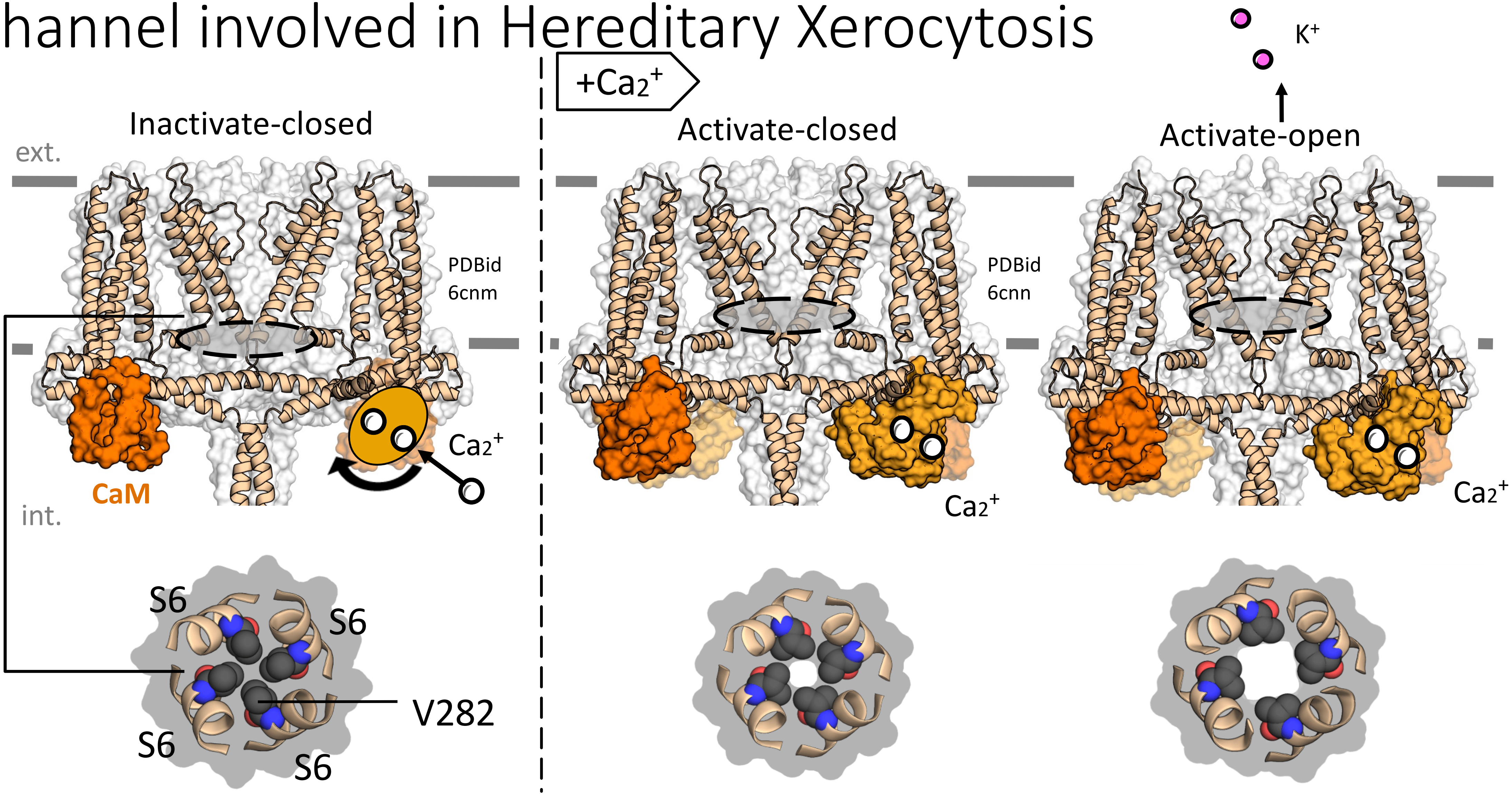


Identification of a central vestibular cavity in the ECD of GlyR that concentrates chloride at the entrance of the ion-transmembrane pore

. Lateral fenestrations connect the extracellular milieu with the central vestibule for chloride translocation in GlyR



An hemolotic disease and the role of KCNN4: a potassium channel involved in Hereditary Xerocytosis

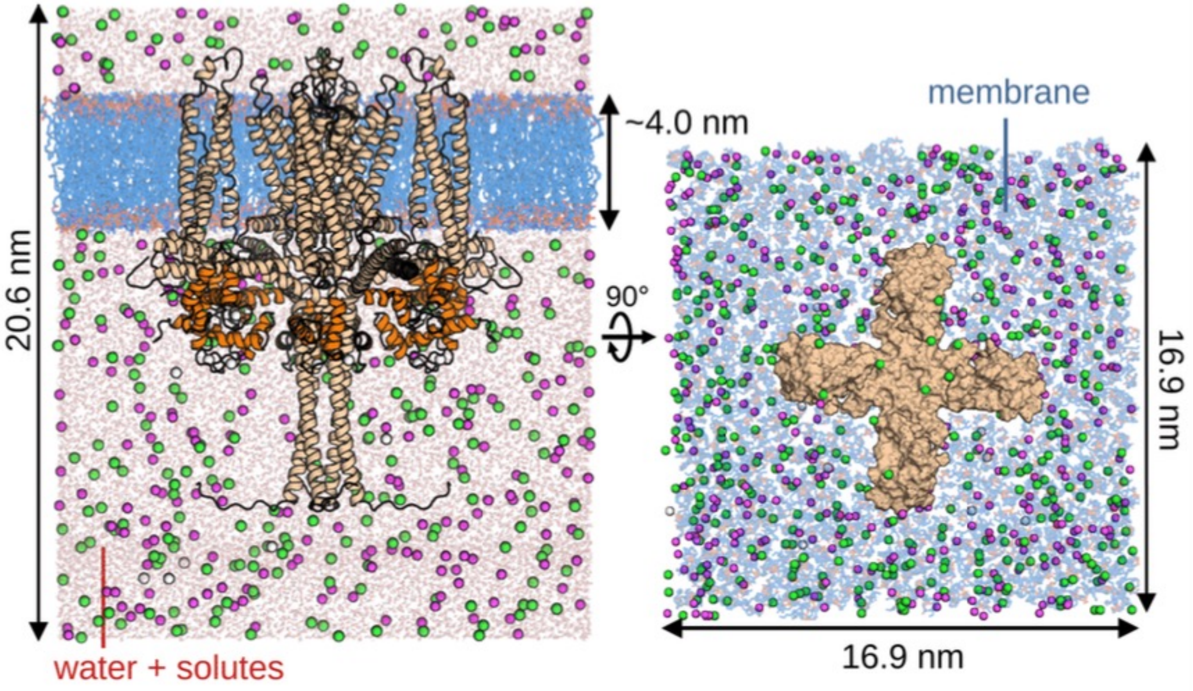


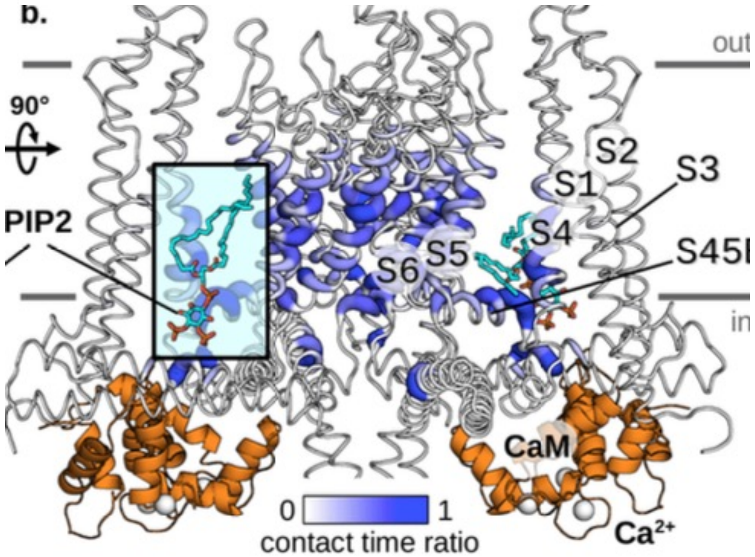
Jedélé S et al , JCIM 2025

Question: What is the dynamics of the channel in the different states?

System	State	membrane	Time	rep.
I.	Pre-activate closed	POPC	2μs	2
I.RBC		RBC	2μs	1
I.PIP2		POPC + PIP2 (bound)	2μs	2
II.	Activate-closed	POPC	2μs	2
II.RBC		RBC	2μs	1
II.PIP2		POPC + PIP2 (bound)	2μs	2
III.	Activate-open	POPC	2μs	2
III.RBC		RBC	2μs	1

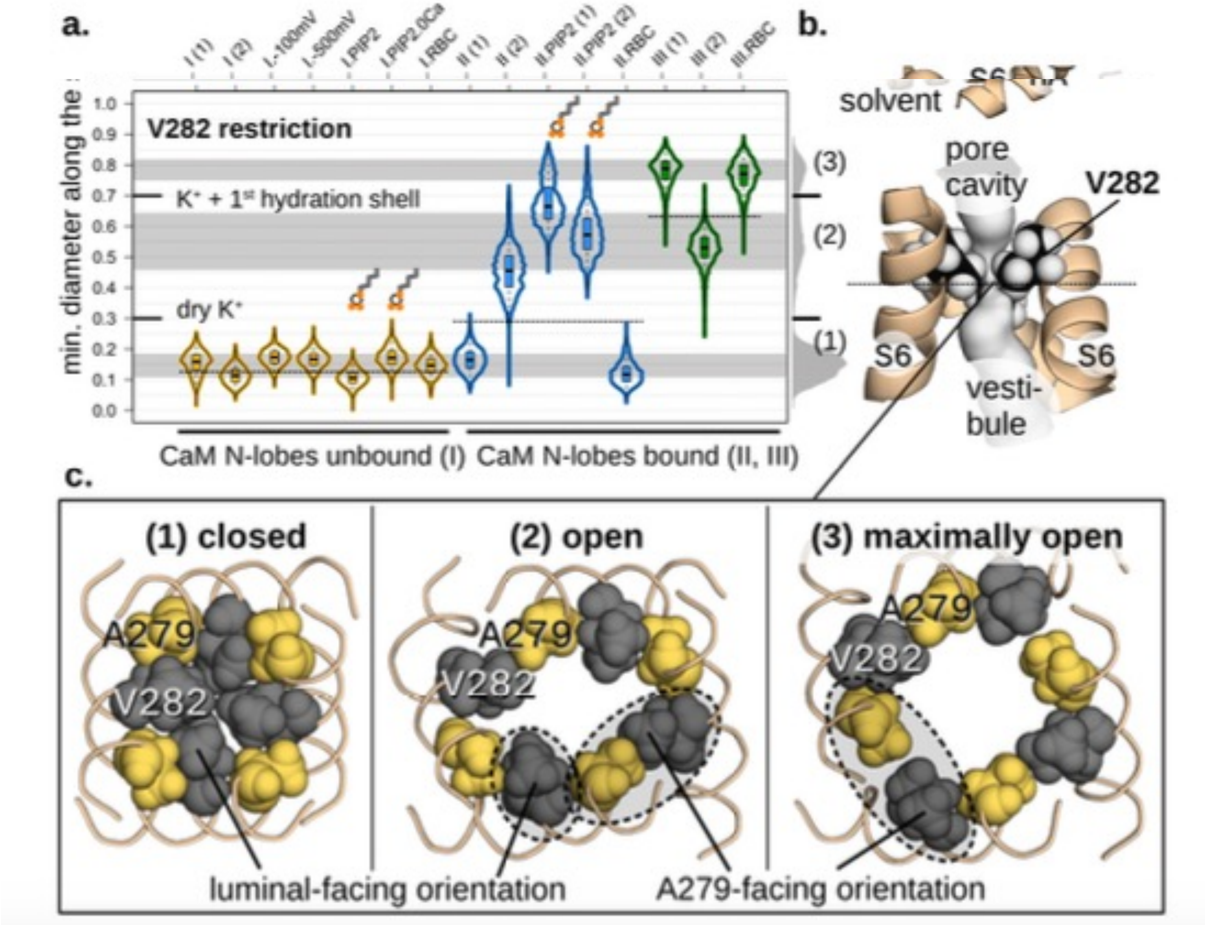
> 7. Millions CPU core Genci + 10 000 GPU



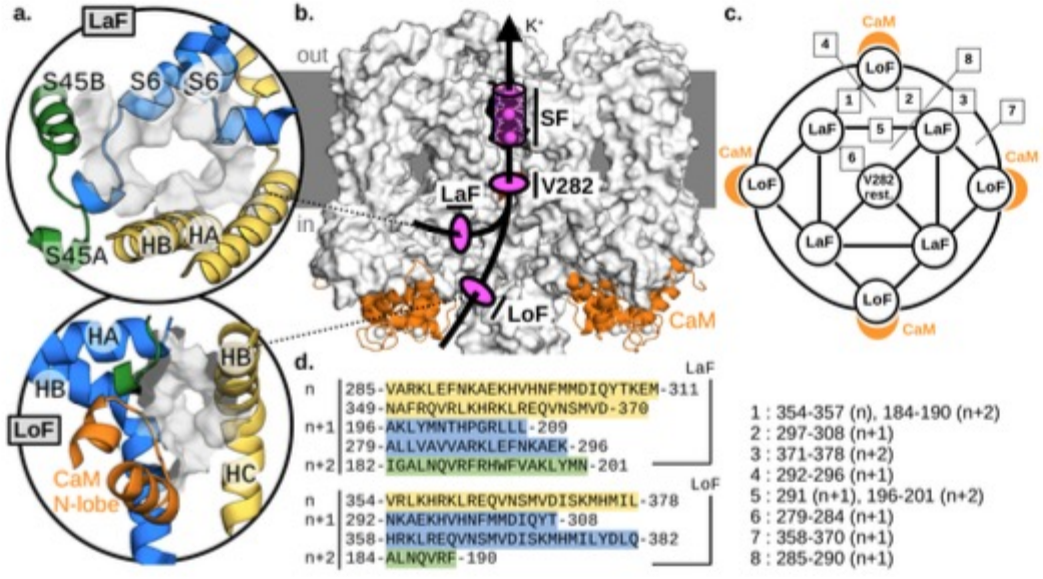


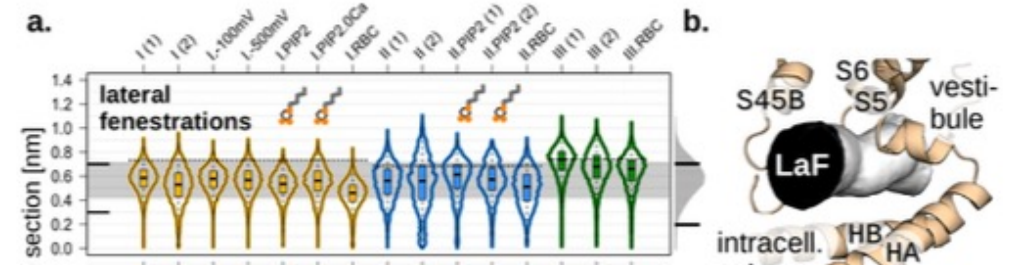
Main results

- Opening of the channel in the presence of PIP2

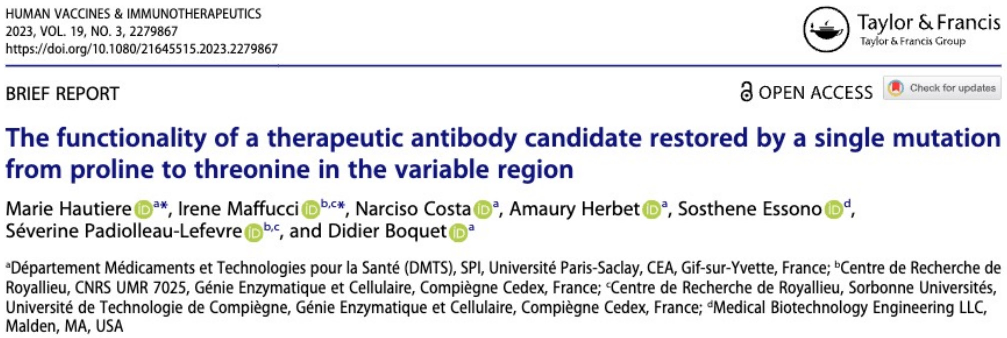


- Identification of Lateral Fenestrations





Cancer & Antibody Design



HUMAN VACCINES & IMMUNOTHERAPEUTICS
2023, VOL. 19, NO. 3, 2279867
https://doi.org/10.1080/21645515.2023.2279867

Taylor & Francis
Taylor & Francis Group

BRIEF REPORT

OPEN ACCESS Check for updates

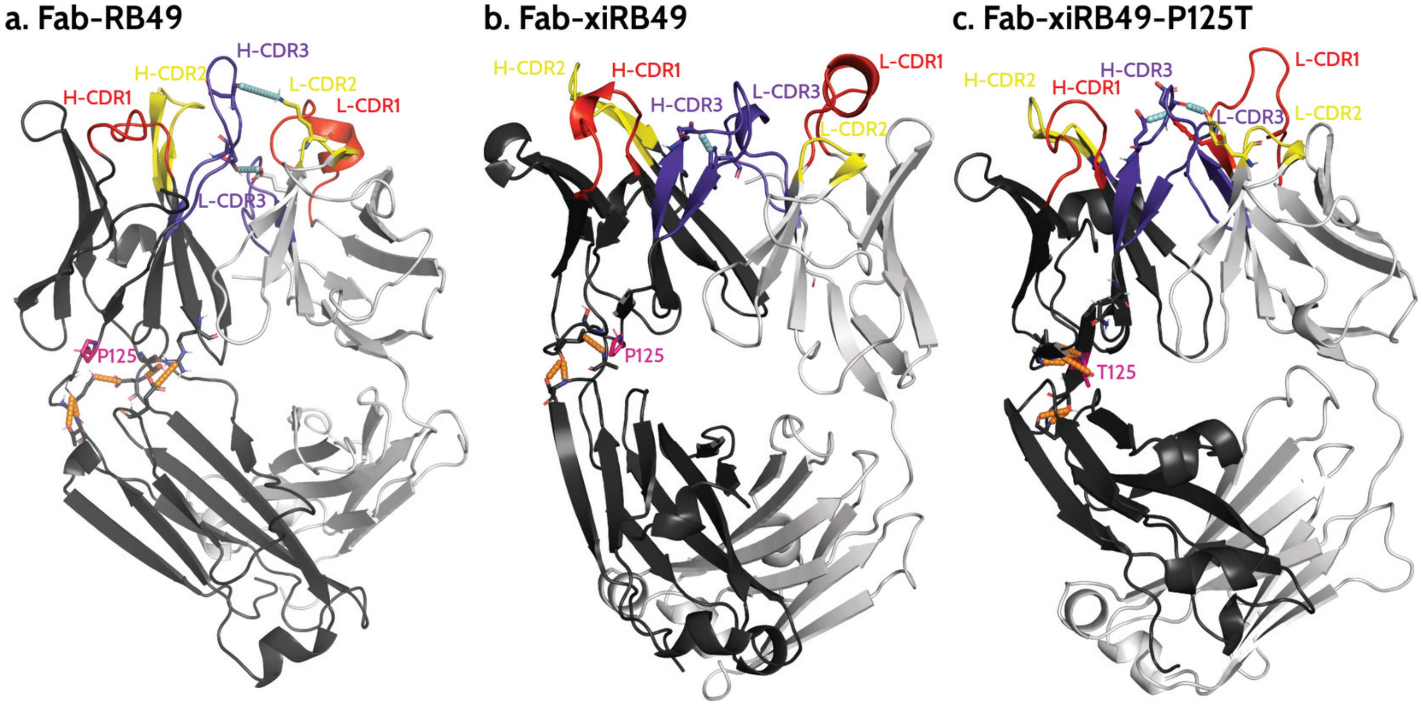
The functionality of a therapeutic antibody candidate restored by a single mutation from proline to threonine in the variable region

Marie Hautiere[a*], Irene Maffucci[b,c*], Narciso Costa[a], Amaury Herbet[a], Sosthene Essono[d], Séverine Padiolleau-Lefevre[b,c], and Didier Boquet[a]

[a]Département Médicaments et Technologies pour la Santé (DMTS), SPI, Université Paris-Saclay, CEA, Gif-sur-Yvette, France; [b]Centre de Recherche de Royallieu, CNRS UMR 7025, Génie Enzymatique et Cellulaire, Compiègne Cedex, France; [c]Centre de Recherche de Royallieu, Sorbonne Universités, Université de Technologie de Compiègne, Génie Enzymatique et Cellulaire, Compiègne Cedex, France; [d]Medical Biotechnology Engineering LLC, Malden, MA, USA

RB49 is an antibody targeting the endothelin B receptor, a GPCR molecule that plays a role in tumour cancel progression. Modificatiion (chimerization) is required to become a human therapeutic agent but this may alter the efficiency.

By combining experiments, molecular modelling and molecular dynamics simulations (μs simulations), the authors identified the key role of a Proline residue in the loss of recognition. Mutation to Thr restores the function.



Representative structures of the most populated cluster of (a) Fab-RB49, (b) Fab-xiRB49, and (c) Fab-xiRB49-P125T. The heavy chain and the light chain are represented in dark and light gray, respectively. CDR1, CDR2, and CDR3 are colored red, yellow, and purple, respectively. The residue in position 125 (either proline or threonine) is represented as ball and sticks and colored magenta. The hydrogen bond network within the region between the heavy chain variable and constant regions is indicated as dotted orange lines, while the interactions involving the H- and L-CDRs mentioned in the manuscript are indicated as dotted cyan lines. The indicated hydrogen bonds come from the analysis of the 3 simulations for each system.

Summary : Methods and Systems studied in the community

- Methods:
  - Molecular Dynamics Simulations (Classical and Enhanced Sampling),
  - Docking
  - now Large Scale 3D structure Prediction with AlphaFold, (AlphaFold)
- Systems :
  - Complexes and assemblies (protein/protein, peptide/protein, nucleic acid/protein)
  - Soluble & Membrane Proteins, Lipids, carbohydrates
  - Protein/Drug interaction (Drug Design)
- Force Fields: Classical or Polarizable, All-atom or Coarse Grained, QM/MM
- Free Energy calculations & now evaluation of Kinetics Constants;

Summary: Example of High Computational Needs

A summary of GENCI Committee “Dynamique moléculaire appliquée à la biologie »

- 49 applications assessed (calls A15 and A16) by 17 experts
- 160 Mh CPU allocated
- 9 Mh GPU allocated
- mainly academic laboratories but also start-ups (subject to publication)

Example of High Computational Needs in constant evolution

- A summary of GENCI Committee “**Dynamique moléculaire appliquée à la biologie »**

A17	JZV100	JZCSL	JZA100	JZH100
	0.72Mh	3.7Mh	0.50Mh	0.31Mh
	AdGenoa	AdMI250x	AdMi300	
	17.8Mh	0.39Mh	0.063Mh	
	JCSKL	JCRome	JCV100	
	8.1Mh	14.3Mh	0.18Mh	
A16	JZV100	JZCSL	JZA100	JZH100
	0.65Mh	0.85Mh	0.34Mh	0.05Mh
	AdGenoa	AdMI250x		
	29.3Mh	1.75Mh		
	JCSKL	JCRome	JCV100	
	rien	46.5Mh	0.41Mh	
A15	JZV100	JZCSL	JZA100	
	2.1Mh	16.6Mh	0.41Mh	
	AdGenoa	AdMI250x		
	2.6Mh	0.327Mh		
	JCSKL	JCRome	JCV100	
	4.9Mh	36.6Mh	0.65Mh	

A14	JZV100	JZCSL	JZA100	
	1.8Mh	11.1Mh	0.88Mh	
	AdGenoa	AdMI200		
	2.3Mh	0.31Mh		
	JCSKL	JCV100	JCRome	
	rien	0.25Mh	67.0Mh	
A13	JZV100	JZCSL	JZA100	
	3.24Mh	21.0Mh	1.21Mh	
		AdMI200		
		0.07Mh		
	JCSKL	JCKNL	JCRome	JCV100
	6.05Mh	3.0Mh	49.3Mh	0.35Mh

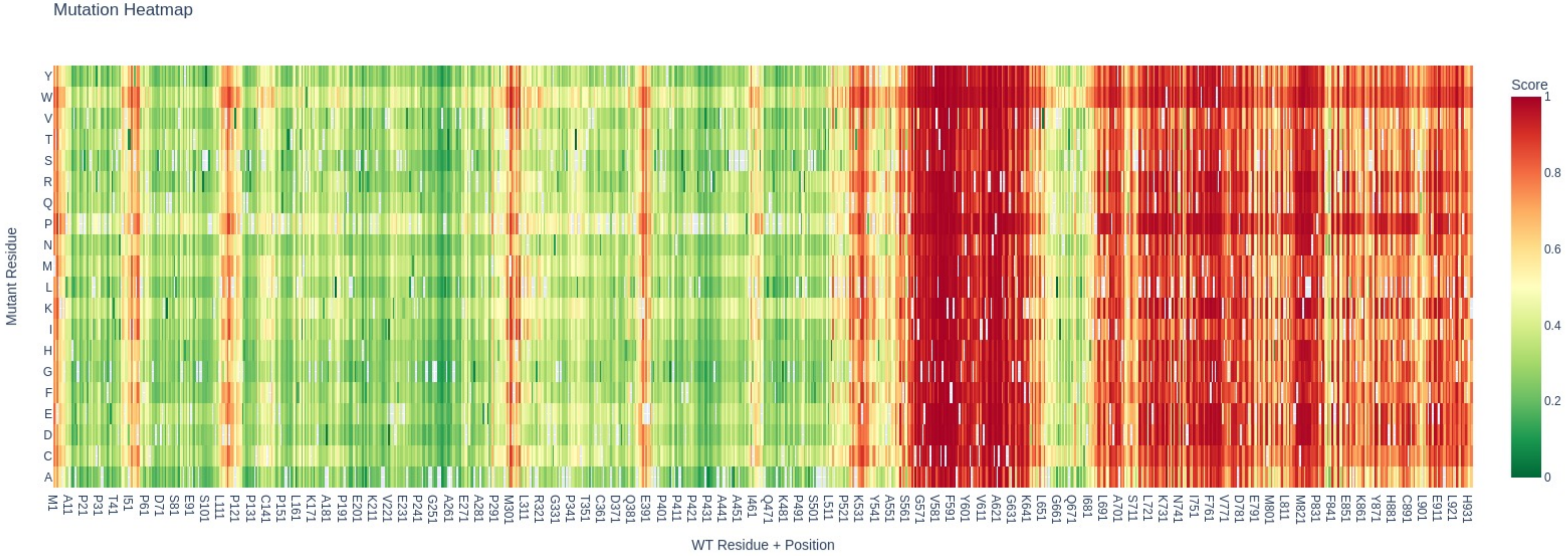
Deep Learning, Structural Bioinformatics and Medical Applications:

- Example 1:
  - Aim: to speed up current applications in structural bioinformatics, i.e. homologous protein searches, secondary structure prediction, cell localisation prediction, prediction of different levels of protein structure (fold, superfamily, family), etc.
    - Strategy: Development of an auto-encoder to reduce the dimensionality of internal protein representations (embeddings) derived from the best protein language models (PLMs) The reduction in dimensionality must be achieved while maintaining the maximum possible information from the original embeddings.
  - => 25 000 GPU hours type A100 with 80 Gb Memory (ÒParallelized)

Deep Learning, Structural Bioinformatics and Medical Applications:

- Example 2:
  - Aim: to predict pathogenicity of mutations
    - Challenge: Protein of ~ 500 residues => 19x500 => For 10 000 proteins ~ 100 million variants
    - Strategy : Use three Protein Language Models (PLM), to generate the variant embeddings

    => 100 million embeddings per PLM.
  - **A100 GPU: [3000-7000] hours depending on the PLM => ~13 000 hours GPU**
  - **H100 GPUs: 7 000 H**



THANK YOU