bioexcel
Centre of Excellence for Computational Biomolecular Research

Horizon 2020
European Union Funding
for Research & Innovation

Exaflop and Big data approaches against Covid19

Modesto Orozco 02/22/2021

What we learn in one crazy year,…

bioexcel

What we learn in one crazy year,…

https://covid.bioexcel.eu/

Develop big data tools

Biology | Viral/Host Structural Data | Therapeutics | Tools | DOIs | Contribution | About

The first open data repository of Covid-19

Targets:

3CLpro / Mpro Activity
Inhibition of PLpro Protease Activity
Host Immune Response
Inhibition of Nsp13 Helicase Activity
Blocking SARS-CoV-2 Spike Protein Binding to Human ACE2 Receptor
Inhibiting Cleavage of the SARS-CoV-2 Spike Protein
Inhibition of Formation of the Viral Fusion Core
Inhibition of Viral Polymerases

Proteins:

3CLpro ACE2 BoAT1 E protein Fc receptor Furin Helicase IL6R M protein N protein NSP1
NSP10 NSP11 NSP14 NSP15 NSP16 NSP2 NSP4 NSP6 NSP7 NSP8 NSP9 ORF10 ORF3a ORF6 ORF7a
ORF7b ORF8 PD-1 PLpro RdRP TMPRSS2 fusion core p38 spike virion

Structures:

3CLpro ACE2 BoAT1 E protein Fc receptor Furin Helicase IL6R M protein N protein NSP1
NSP10 NSP11 NSP14 NSP15 NSP16 NSP2 NSP4 NSP6 NSP7 NSP8 NSP9 ORF10 ORF3a ORF6 ORF7a
ORF7b ORF8 PD-1 PLpro RdRP TMPRSS2 fusion core p38 spike virion

Models:

3CLpro ACE2 BoAT1 E protein Fc receptor Furin Helicase IL6R M protein N protein NSP1
NSP10 NSP11 NSP14 NSP15 NSP16 NSP2 NSP4 NSP6 NSP7 NSP8 NSP9 ORF10 ORF3a ORF6 ORF7a
ORF7b ORF8 PD-1 PLpro RdRP TMPRSS2 fusion core p38 spike virion

Therapeutics:

antibody antiviral immunotherapy peptide small molecule

Simulations:

3CLpro ACE2 BoAT1 E protein Fc receptor Furin Helicase IL6R M protein N protein NSP1
NSP10 NSP11 NSP14 NSP15 NSP16 NSP2 NSP4 NSP6 NSP7 NSP8 NSP9 ORF10 ORF3a ORF6 ORF7a
ORF7b ORF8 PD-1 PLpro RdRP TMPRSS2 fusion core p38 spike virion

Links:

ANI-CAS Antiviral Archive | ANI-FDA Drugs Archive | CORD-19: COVID-19 Open Research Dataset | COV3D: A Coronavirus 3D Structure Database | Coronaviruses 101- Focus on Molecular Virology | Drug Repurposing Hub | DrugBank database | Enamine REAL Space | MolPort | Open Science Data Portal | PubChem | SWEETLEAD | Solvation Maps for COVID19-related Protein Targets | Structural Biology Task Force GitHub page | Structure models of all mature peptides in 2019-nCoV genome by C-I-TASSER | SuperDRUG | The Cambridge Structural Database | Tristan Croll ISOLDE COVID-19 models | WuXi GalaXi | zinc15 database

MolSSI bioexcel

Coronavirus nonstructural protein 7

Inhibition of viral polymerases

SARS-CoV-2 nsp7-nsp8-nsp12 RNA polymerase complex in aqueous solution

DESRES

Represented Proteins: RdRP NSP7 NSP8

Model: Files | Source Structure PDBs: 6M71 | Visualize: 3DMol.js

The C- and N-peptide termini capped with amide and acetyl groups respectively. The missing loops in the published structural models were manually built as extended peptide conformation. The missing part of Chain D was built through homology modeling using the structure of SARS-CoV-1 polymerase complex (PDB entry 6NUR). The system was neutralized and salted with NaCl, with a final concentration of 0.15 M.

Simulations:

DESRES-ANTON-10917618 10 μs simulation of SARS-CoV-2 nsp7-nsp8-nsp12 RNA polymerase complex in aqueous solution

DESRES-ANTON-10917618 10 μs simulation of SARS-CoV-2 nsp7-nsp8-nsp12 RNA polymerase complex, no water or zinc

---

SARS-CoV-2 RdRP (NSP12) in complex with NSP7 and two copies of NSP8: ISOLDE refined model

Tristan Croll

Represented Proteins: RdRP NSP7 NSP8

Model: Files | Source Structure PDBs: 6M71 | Visualize: 3DMol.js

Refinement of 6m71 that fixes multiple issues: * Corrects incorrect modeling of both zinc binding sites (originally modeled as disulfides) * Corrects 1-2 dozen rotamer adjustments and peptide flips * Models C-terminal domain of chain D (one of the NSP8s) using well-resolved chain B See a complete description of the issues remedied by this model.

Simulations:

SARS-CoV-2 RdRp complex (nsp12+2*nsp8+nsp7) + RNA template-primer + ATP model for MD simulations

Vaibhav Modi

University of Jyväskylä -- Department of Chemistry and Nanoscience Center -- Computational Biomolecular Chemistry Group

Represented Proteins: RdRP NSP7 NSP8

Model: Files | Source Structure PDBs: 6NUR 7BTF 7BV2 6YYT | Visualize: 3DMol.js

Model of the RdRp + RNA + ATP complex of the SARS-CoV-2 with non-covalently bound ATP molecule is built using homology modelling with the SARS-CoV-1 RdRp complex (PDB:6NUR) as template structure (https://doi.org/10.1038/s41467-019-10280-3). The modelled structure shows excellent fit (< 0.6 Å) to the SARS-CoV-2 RdRp complex (PDB:6M71) kindly shared with us by Gao et.al. Further, the model of RdRp complex with RNA and ATP molecule in Tri-phosphate form is modelled based on comparative fitting with previously known poliovirus and norovirus structres (with NTP molecule in hydrophobic cleft). The protein-RNA complex with Remdesivir shows excellent fit (< 0.7 Å) with the recently published RdRp complex with RNA template-primer (PDB:7BV2, 6YYT). The fitted models have been equilibriated to perform long MD simulations.

Simulations:

Gromacs 100 ns MD of SARS-CoV-2 RdRp + RNA template-primer + ATP model, All Atom model

---

SARS-CoV-2 apo-RdRp complex (nsp12+2*nsp8+nsp7) model for MD simulations

Vaibhav Modi

University of Jyväskylä -- Department of Chemistry and Nanoscience Center -- Computational Biomolecular Chemistry Group

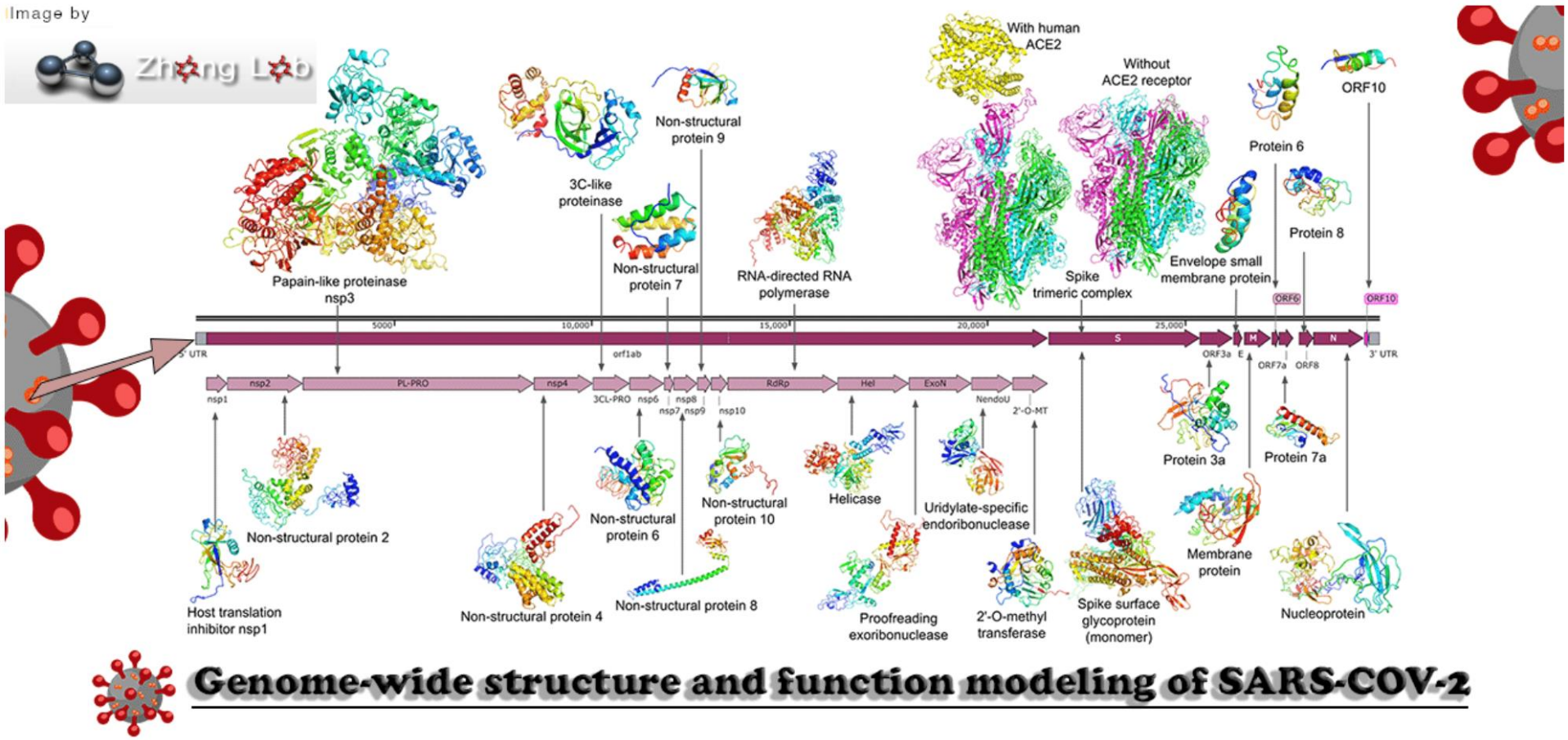
Represented Proteins: RdRP NSP7 NSP8

Model: Files | Source Structure PDBs: 6NUR 6M71 7BTF 7BV1 | Visualize: 3DMol.js

Model of the apo-protein form of RdRp complex of the SARS-CoV-2 is built using homology modelling with

Develop big data tools

https://covid.bioexcel.eu/proteins/



Genome-wide structure and function modeling of SARS-COV-2

bioexcel

BioExcel COVID-19

HOME BROWSE CONTACT REST API

accession	name	unit	analyses
MCV1900002	SARS spike receptor binding domain bound with FERR ACE2	RBD-ACE2	Fluctuation PCA Rgyr
MCV1900003	Bat-SARSr-CoV RATG13 spike receptor binding domain	RBD	
MCV1900004	Human Angiotensin-converting enzyme 2 (ACE2)	ACE2	
MCV1900005			
MCV1900006	SARS-CoV-2 spike receptor binding domain	RBD	
MCV1900007	Mutated system of Bat-SARSr-CoV RATG13 spike receptor binding domain	RBD	

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

IRB BARCELONA
INSTITUTE FOR RESEARCH IN BIOMEDICINE

Develop big data tools

BioExcel COVID-19

HOME BROWSE CONTACT REST API

OVERVIEW TRAJECTORY PCA RGYR RMSD FLUCT

Accession: MCV1900002

DATA IN THIS PAGE

Statistics

Counts

System atoms	Proteins atoms	Proteins residues	Phospholipids	Solvent molecules
136320	12119	757	0	123717

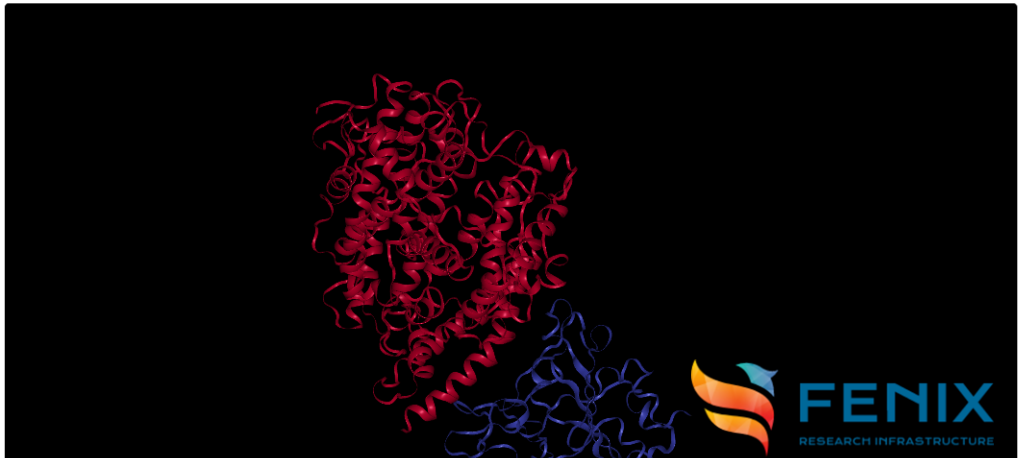
Positive ions	Negative ions
0	1

15222 nm³

Length	Timestep	Snapshots	Frequency	Force field
200 ns	2 fs	20001	10 ps	Not available

Temperature	Water type	Ensemble	Pressure coupling	Membrane
310 K	Not available	NPT	Isotropic	No

Adding value to the trajectories



FENIX RESEARCH INFRASTRUCTURE

https://bioexcel-cv19.bsc.es

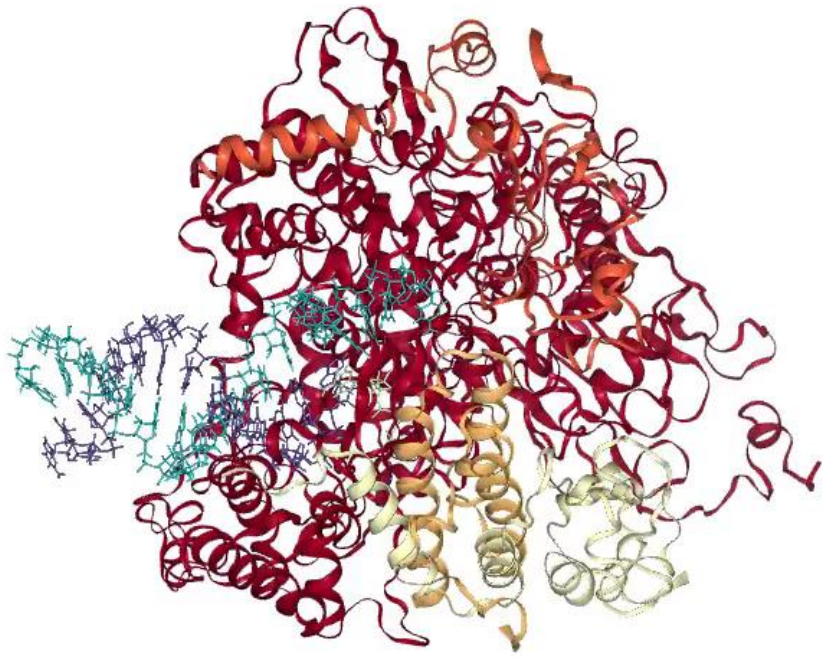
JÜLICH Forschungszentrum

Human Brain Project

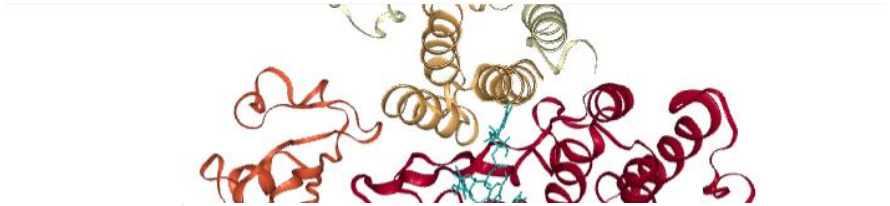
MCV1900213 - Trajectory

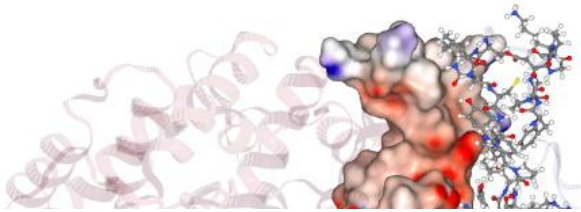
DATA IN THIS PAGE

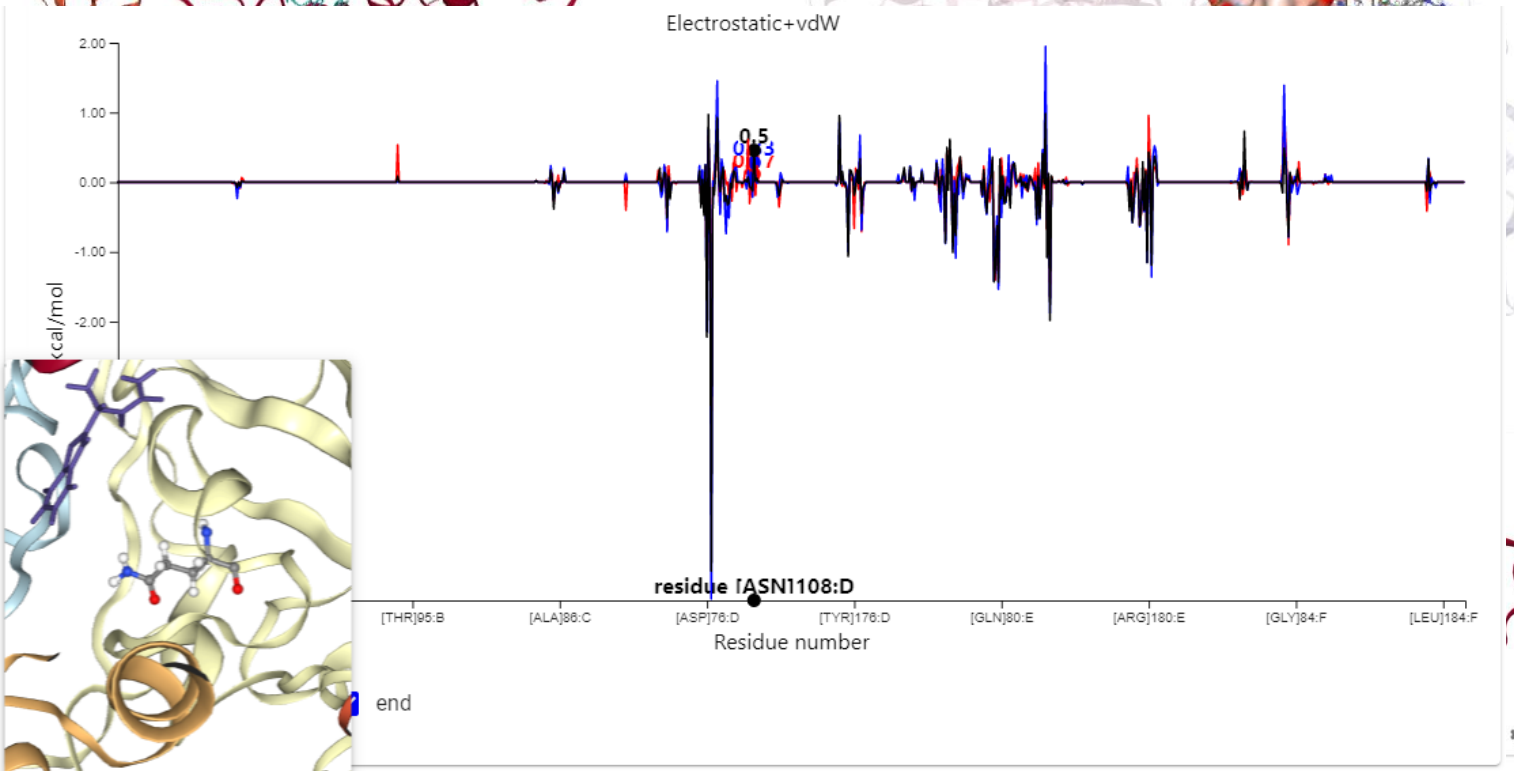
Domains: Overall | Polymerase | Cofactor 1 | Cofactor 2 | Cofactor 3 | Complex | RTP | RNA



▶ Trajectory metadata



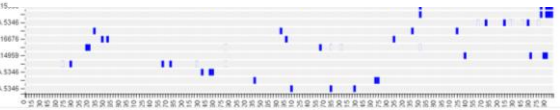


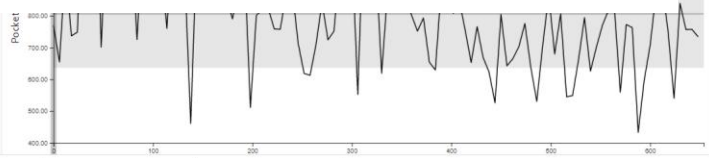

Electrostatic+vdW
2.00
1.00
0.00
-1.00
-2.00
kcal/mol
0.5
residue [ASN1108:D
[THR]95:B
[ALA]86:C
[ASP]76:D
[TYR]176:D
[GLN]80:E
[ARG]180:E
[GLY]84:F
[LEU]184:F
Residue number
end

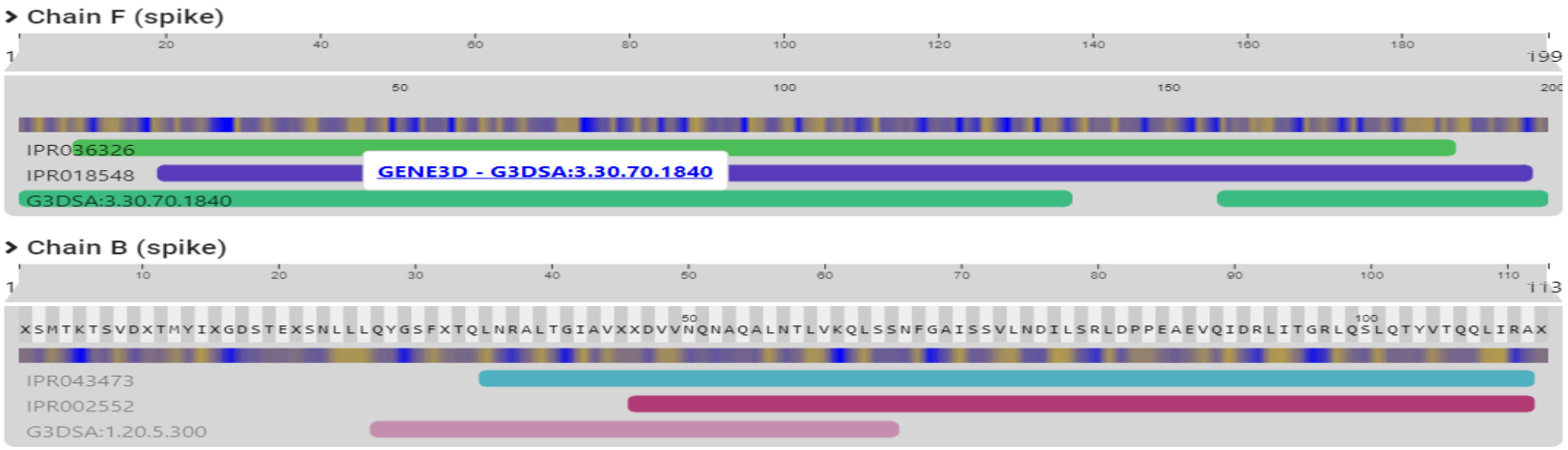
MCV1900213 - Hydroge
Interaction:
Polymera
Polymera
Polymera
Cofactor
Cofactor
Complex
Complex
Hydrogen bonds

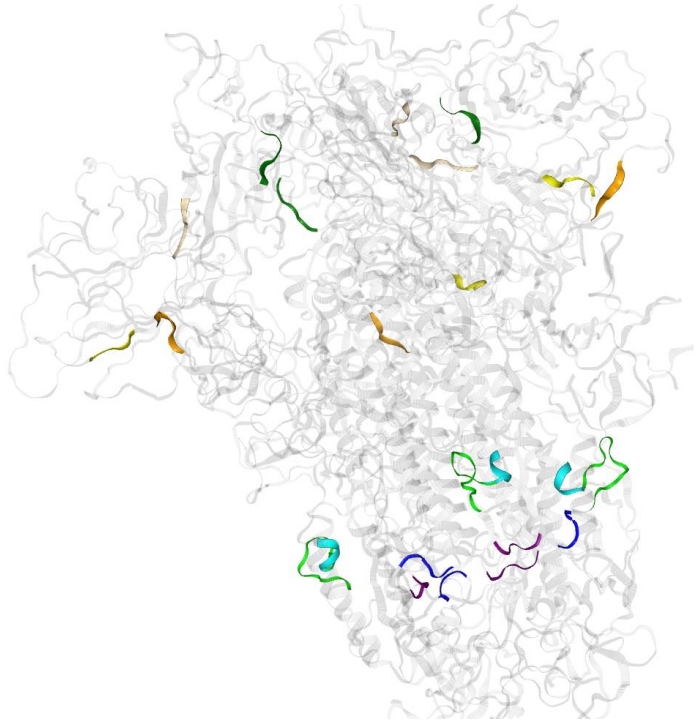
Barcelona
Supercomputing
Center
Centro Nacional de Supercomputación

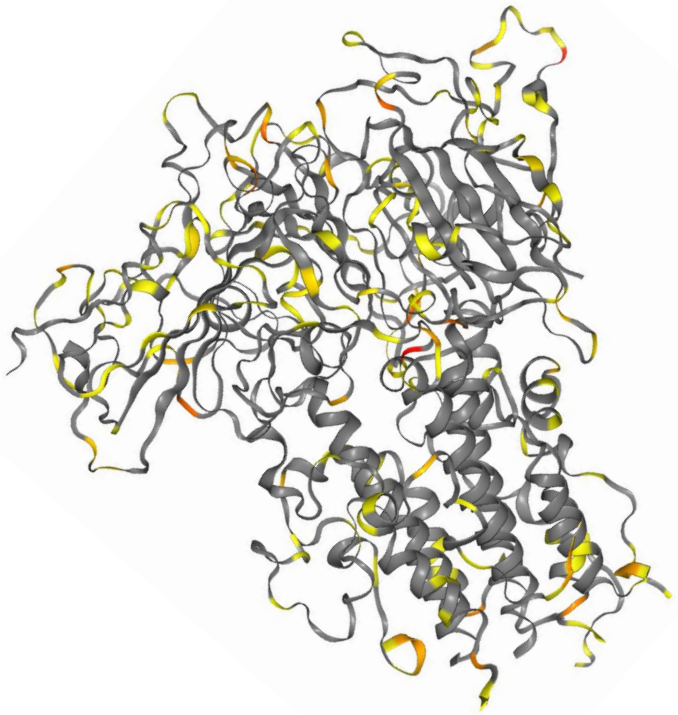
IRB
BARCELONA




Pocket


Chain F (spike)
IPR036326
IPR018548
GENE3D - G3DSA:3.30.70.1840
G3DSA:3.30.70.1840
Chain B (spike)
IPR043473
IPR002552
G3DSA:1.20.5.300

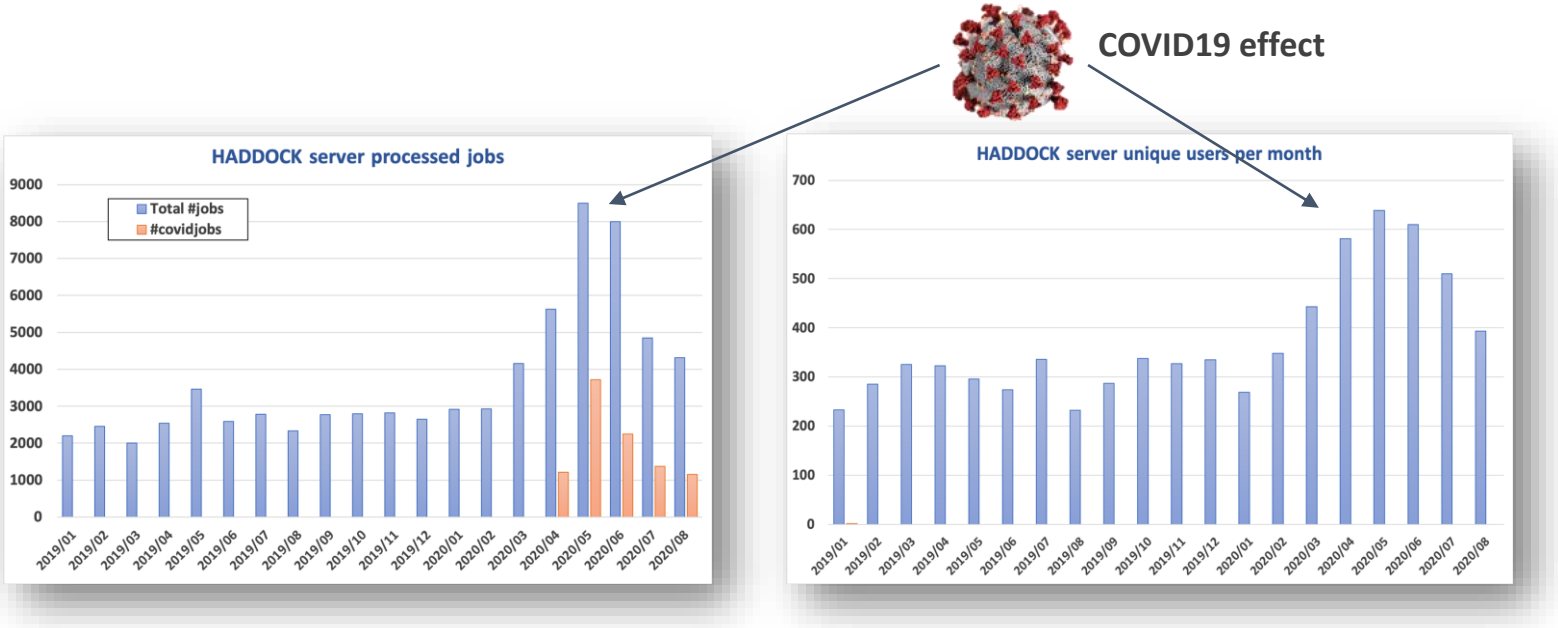




Universiteit Utrecht

HADDOCK: Meeting the increased demand

The HADDOCK workflow machinery was modified to improve its efficiency and meet the increased demand (allows to run more processes in // - relevant toward exascale).



By now more than 10500 COVID-related runs!!

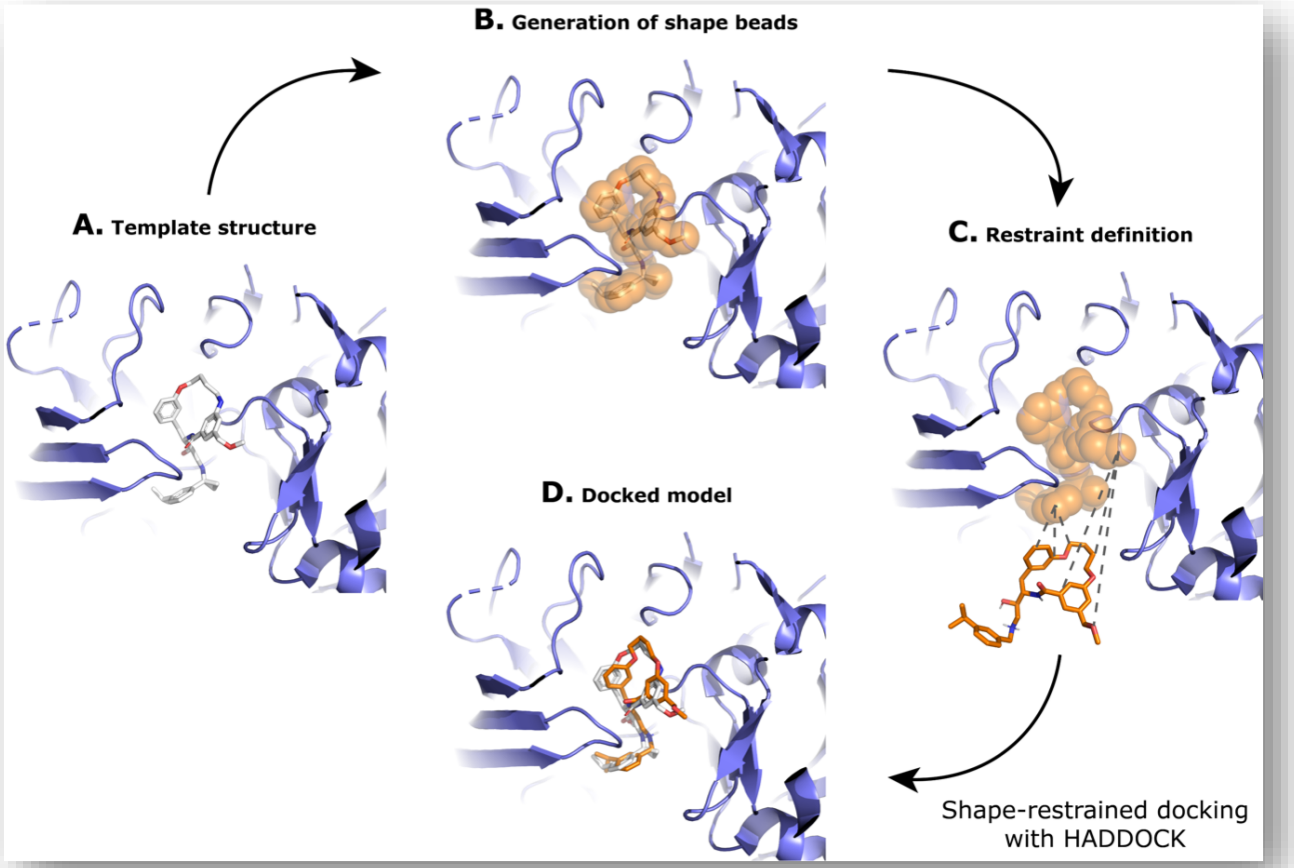
bioexcel

Develop HPC tools

Template, shape-driven HADDOCKing

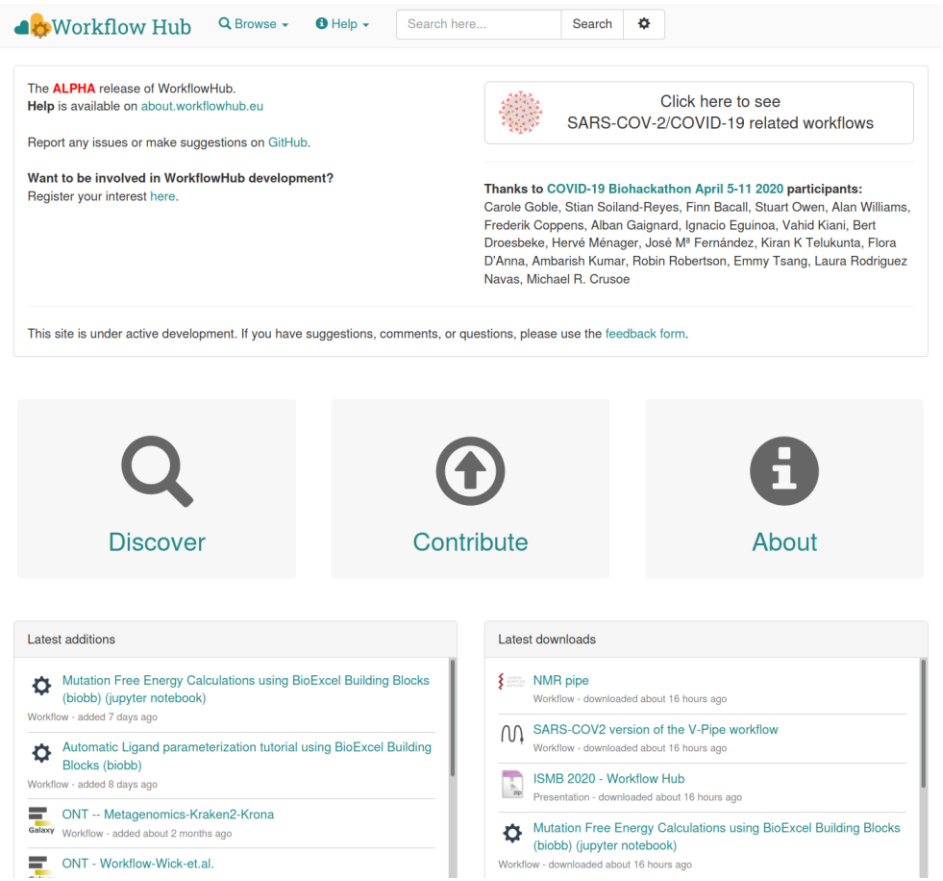
Universiteit Utrecht

- Identify template structures
- Transform template compound atoms to dummy atoms
- Dock using restraints from the dummy shape atoms to the conformers without pre-selecting conformers



bioexcel

Workflow Hub



https://covid19.workflowhub.eu

s://elixir-europe.org/news/hacking-pandemic

Development of the Workflow Hub for workflows **fast-tracked** for COVID-19

- COVID-19 Virtual **BioHackathon April 2020**
- Community hackathons during 2020
- A pan-project collaboration
- Expanded beyond COVID

>25 **public COVID-19 workflows** identified, curated and registered

- Galaxy, Nextflow, CWL, Snakemake
- Listed on ELIXIR COVID-19 data portal

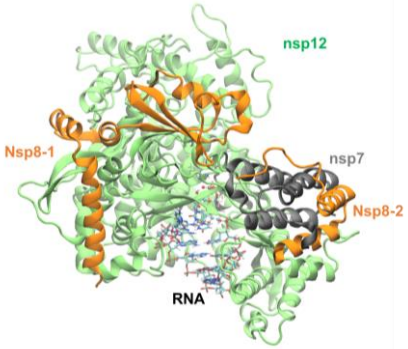
A catalyst for **community collaboration**

- Workflow creators
- Workflow infrastructure developers

Improved **standards** and **best practices**

- Bioschemas, RO-Crate, CWL, nf-core

Develop HPC tools



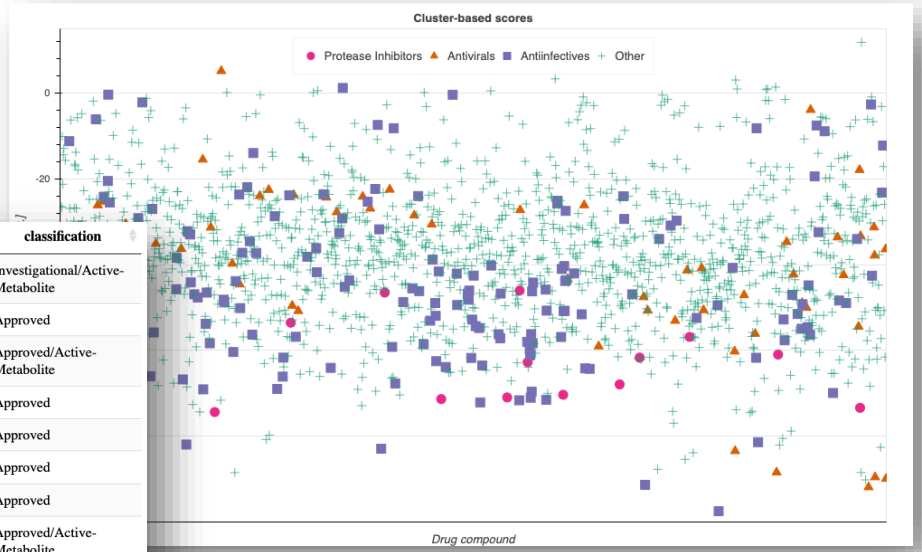
RdRNA polymerase

Modeling studies

Binding site-driven docking protocol

target		atc	score	rank	category	classification
remTP	Remdesivir-triphosphate	NA	-94.866	1	Antivirals	Investigational/Active-Metabolite
DB00923	Ceforanide	J01DC11	-94.804	2	Antiinfectives	Approved
CID23725128	GS-461203	J05AP08,J05AP51,J05AP55,J05AP56	-94.721	3	Antivirals	Approved/Active-Metabolite
DB09050	Ceftolozane	J01DI54	-94.271	4	Antiinfectives	Approved
DB06717	Fosaprepitant	A04AD12	-94.241	5	Other	Approved
DB06590	Ceftaroline-fosamil	J01DI02	-92.098	6	Antiinfectives	Approved
DB11574	Elbasvir	J05AP54	-91.283	7	Antivirals	Approved
CID5271809	favipiravir-RTP	J05AX27	-90.675	8	Antivirals	Approved/Active-Metabolite
DB01632	5-O-phosphono-alpha-D-ribofuranosyl-diphosphate	NA	-90.028	9	NA	Approved
DB09335	Alatrofloxacin	J01MA13	-88.158	10	Antiinfectives	Approved

Showing 1 to 10 of 2,026 entries Previous 1 2 3 4 5 … 203 Next

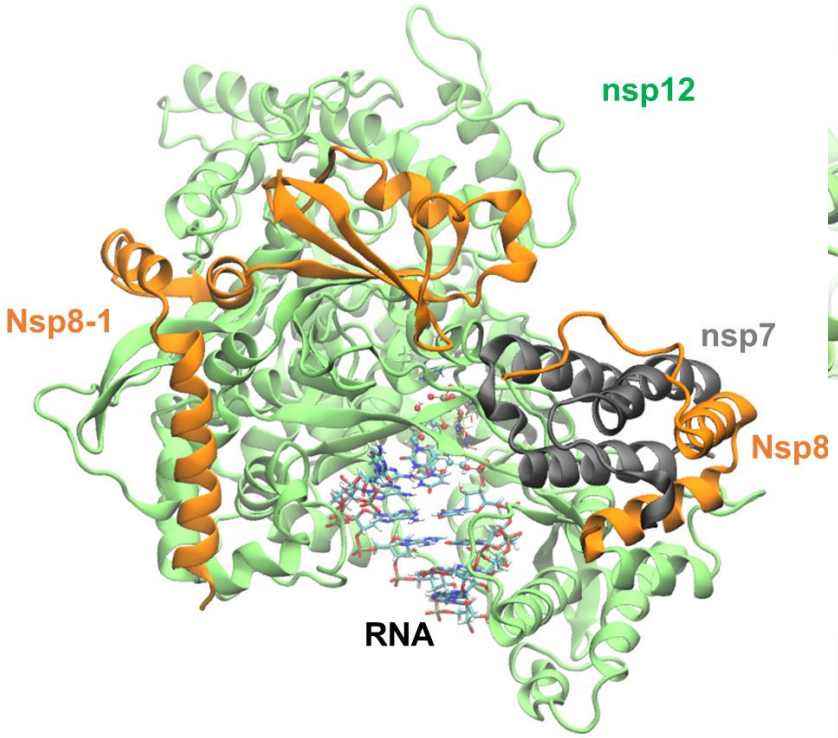


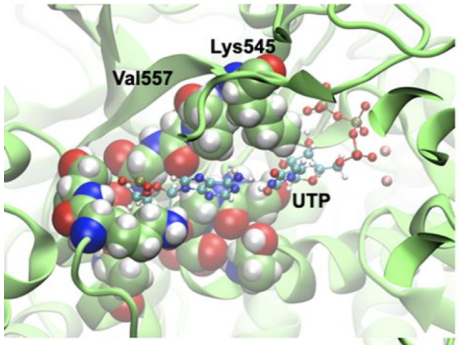
bonvinlab.org/covid

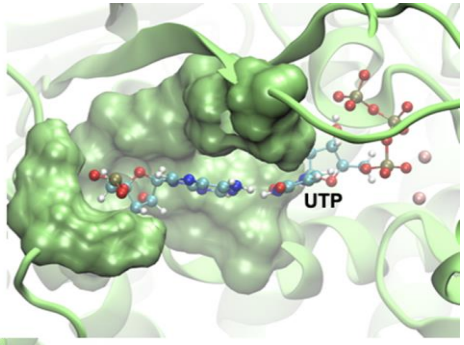
Universiteit Utrecht

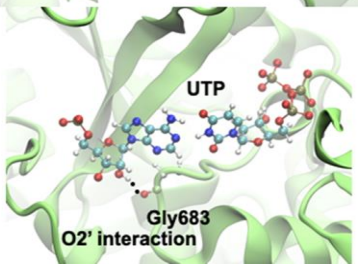
bioexcel

RdRNA Polymerase. The core of Covid-19 replication







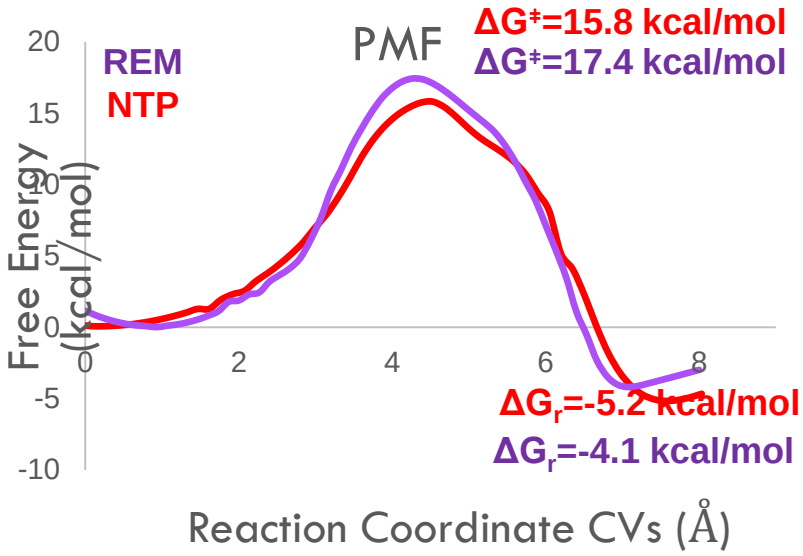


Complex of NSP7 (83 Aa) & NSP8 (198 Aa) : Primase. NSP12: RNA dependent RNA polymerase

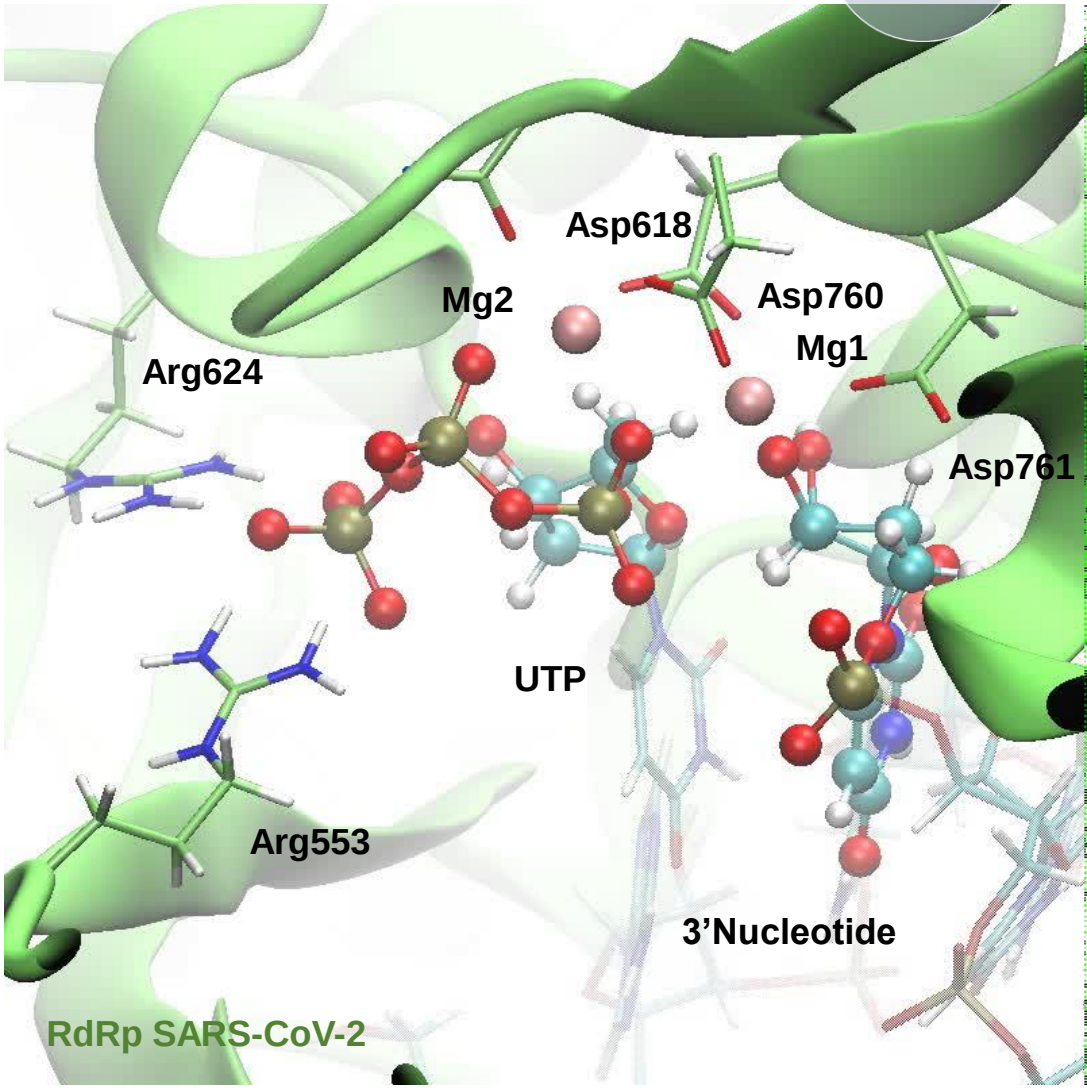
Modeling studies

Polimerization mechanism of RdRp SARS-CoV-2

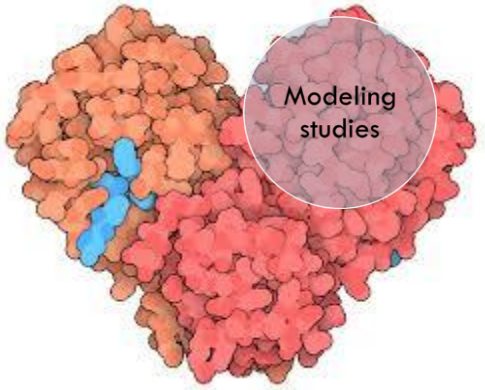
B3LYP/6-311++G**:DFTB3/MM



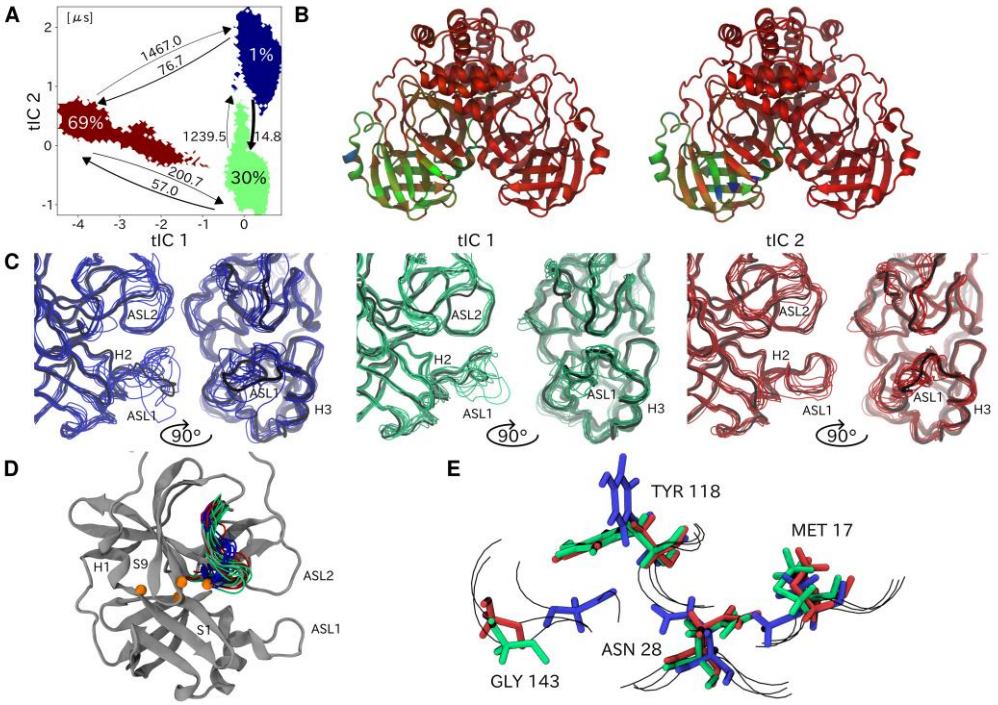
Very efficient enzyme!

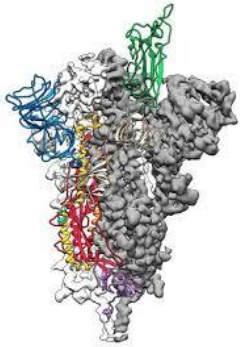


Massive-scale simulations of SARS-CoV-2 proteins



- Markov State Modeling of 100μs of simulations of **main protease (Mpro)** revealed alternative loop conformations that produce two distinct states of the active site
  (Cathrine Bergh, KTH)
- Affects binding in docking studies, improves correlation with SARS-CoV-1 binding assays
- Initiated new collaborations with CINECA and pharma sector (Dompé) that led to EXSCALATE4CoV project

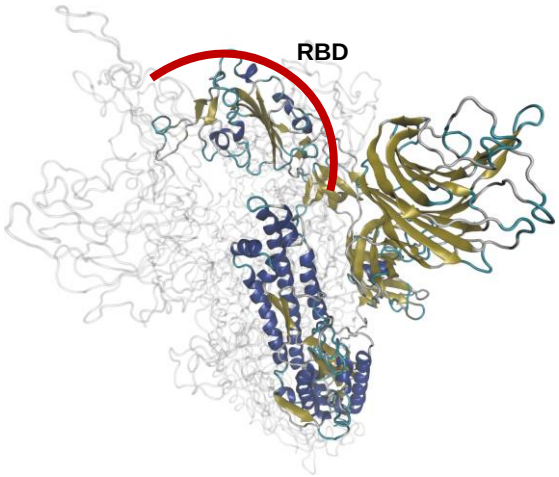




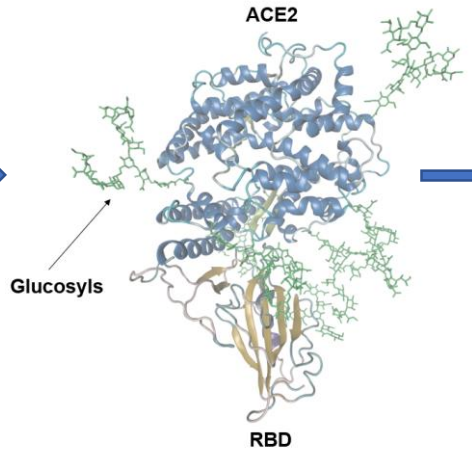
Modeling studies

SARS-CoV-2 (COVID-19) Spike protein trimer

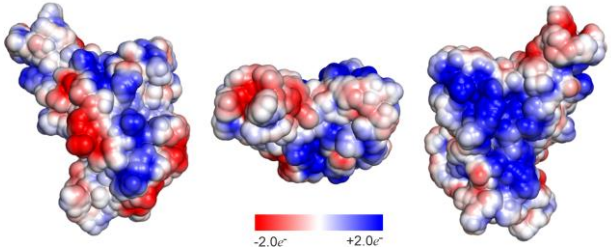
S-protein trimer



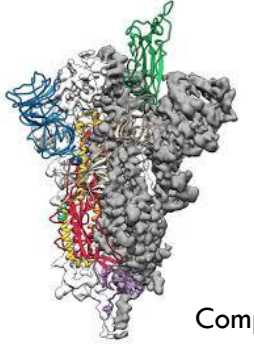
RBD complex with human ACE2



Electrostatic potential of RBD



- Responsible for recognition and initial interaction with human cells
- Structure is a trimer with ~1200 amino acids in each individual monomer
- Part of S-protein responsible for binding is called Receptor Binding Domain (RBD)
- Target for the binding found to be human Angiotensin Converting Enzyme 2 (ACE2).
- Target for COVID-19 therapy, through blocking binding towards ACE2

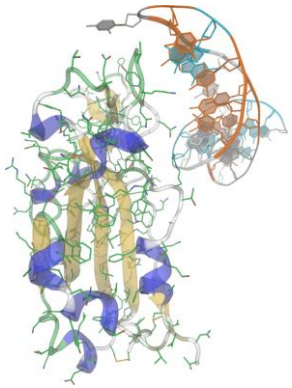


Modeling studies

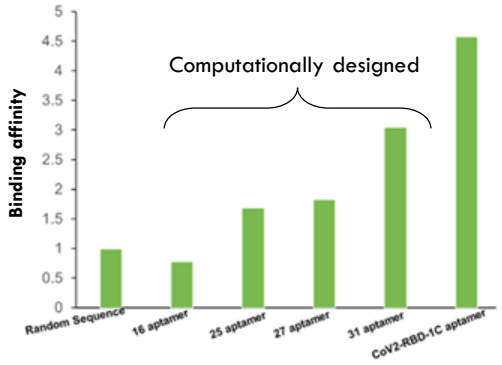
Aptamers design for selective binding

Computationally designed oligo-DNA based aptamer

Binding mode with SARS-CoV-2 RBD



Experimental binding check



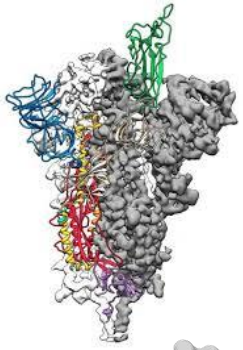
- Synthetic ligands designed to specifically bind with the chosen target.
  - Oligomers of nucleic acids (RNA, DNA)
  - Oligopeptides
  - Hybrid substances which combines nucleic acids with peptides and hydrocarbons.
- Modelling of aptamers allows systematic:
  - Structure prediction
  - Binding sites with estimation of binding affinity
  - Iterative design of the new aptamers using combination of simulations, machine learning techniques and experiments
- The best designed aptamer 31 shows binding affinity close to the much bigger aptamer CoV2-RBD-1C made by *in vitro* evolutional protocol SELEX

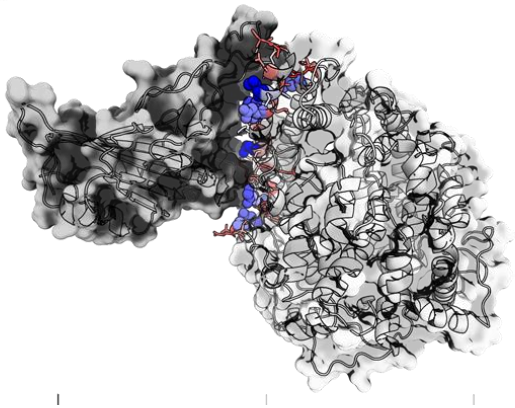
JYVÄSKYLÄN YLIOPISTO
UNIVERSITY OF JYVÄSKYLÄ

bioexcel

Massive exploration of human variability Towards antibody design

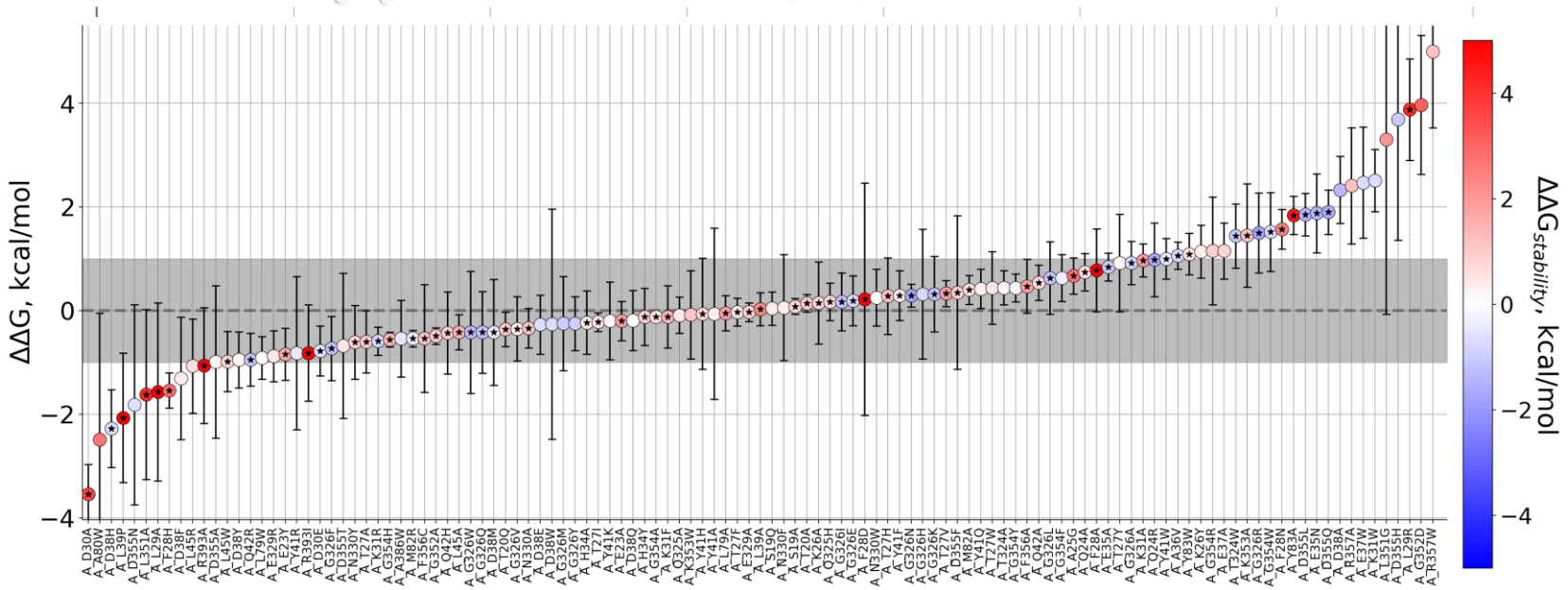
Modeling studies





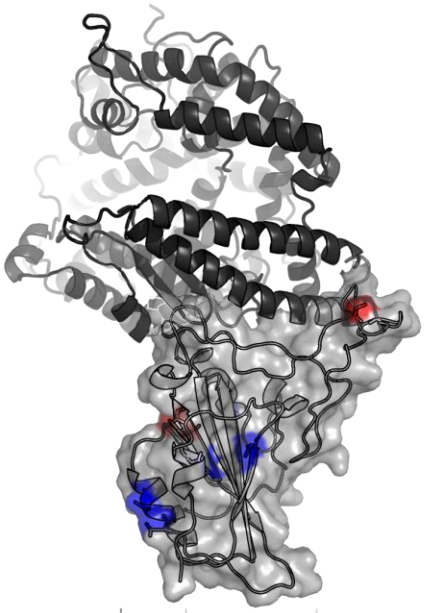
ACE2 residues in the interface with SARS-Cov2 RBD were probed in a site saturation mutational scan by means of Rosetta's flex_ddg protocol.

A number of mutations were identified to increase the binding affinity (negative free energy differences in the figure).



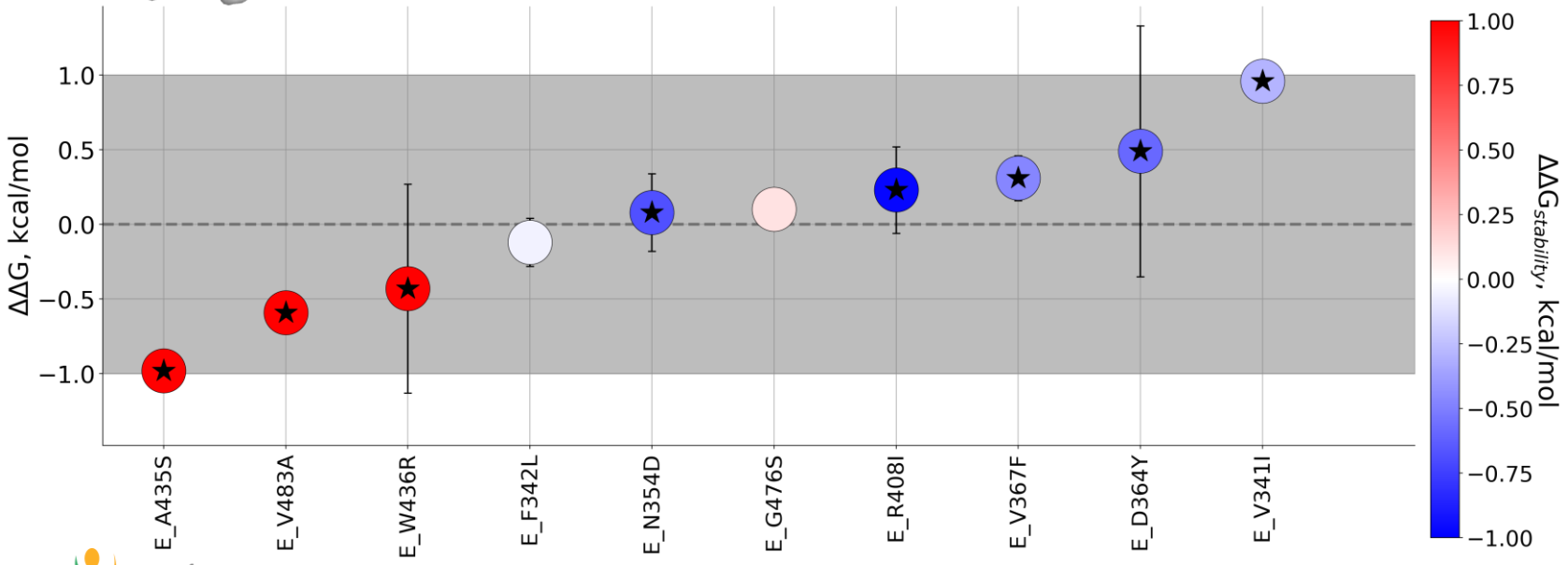
SARS-Cov2 RBD mutations

Modeling studies



Alchemical calculations predict that the probed mutations have only moderate effect (+- 1 kcal/mol) on the complex binding affinity in contrast to the previously reported predictions (10.1101/2020.03.15.991844).

The mutations also have moderate effect within the range of +- 2 kcal/mol on the stability of the RBD apo state.



bioexcel

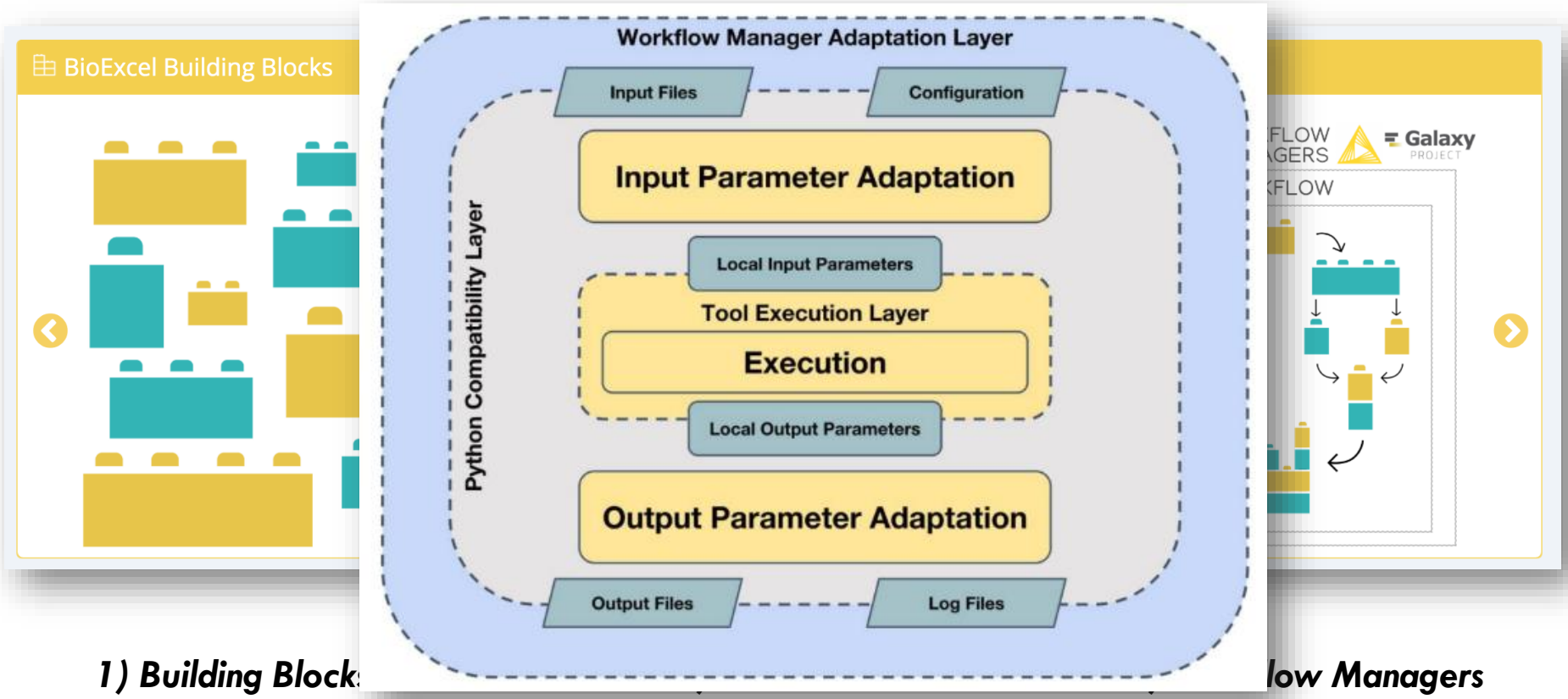
MAX-PLANCK-GESELLSCHAFT

BioExcel Building Blocks software library

Modeling studies

Develop HPC tools

HPC



1) Building Block

low Managers

http://mmb.irbbarcelona.org/biobb/

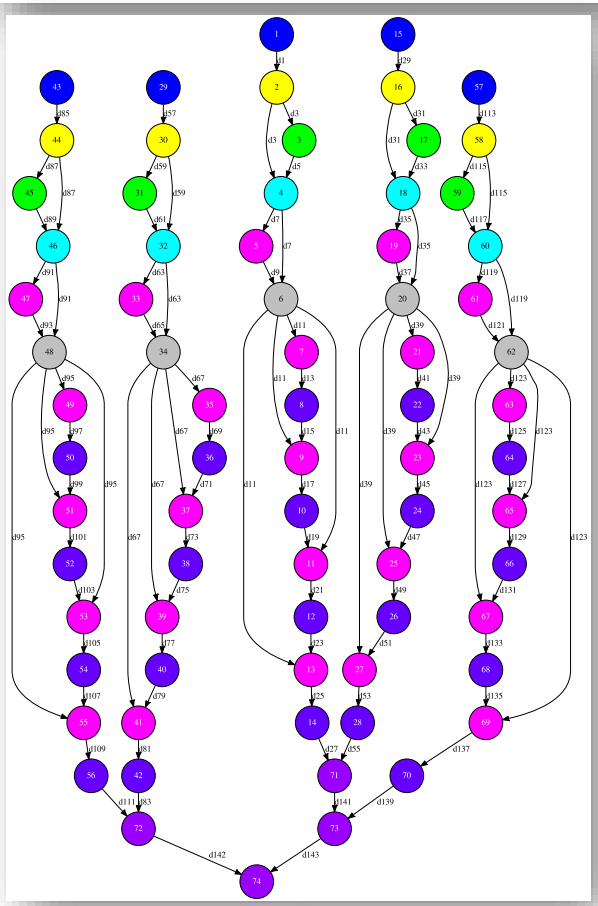
BioExcel Building Blocks: CLI - Mutations

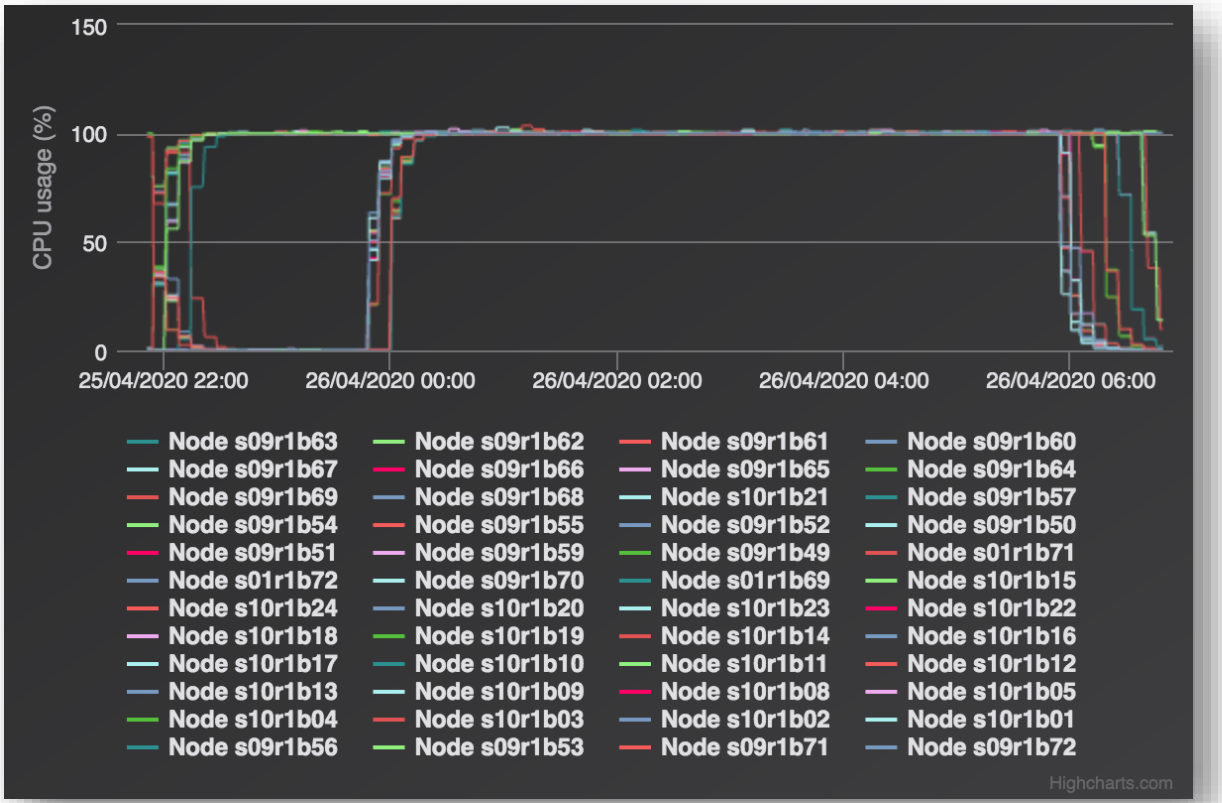
```python
# Loading the biobb configuration reader
conf = settings.ConfReader(sys.argv[1])

global_log, _ = fu.get_logs(path=conf.get_working_dir_path())
global_prop = conf.get_prop_dic(global_log=global_log)
global_paths = conf.get_paths_dic()

global_log.info("step1_pdb: Download the initial Structure")
Pdb(**global_paths["step1_pdb"], properties=global_prop["step1_pdb"]).launch()

global_log.info("step2_fixsidechain: Modeling the missing heavy atoms in the structure side chains")
FixSideChain(**global_paths["step2_fixsidechain"], properties=global_prop["step2_fixsidechain"]).launch()

for mutation in conf.properties['mutations_list']:

    mut_paths = conf.get_paths_dic(mutation)
    mut_prop = conf.get_prop_dic(mutation, global_log=global_log)

    mut_paths['step3_mutate']['input_pdb_path'] = global_paths['step2_fixsidechain']['output_pdb_path']

    global_log.info("step3_mutate: Modeling a particular residue mutation")
    Mutate(**mut_paths["step3_mutate"], properties=mut_prop["step3_mutate"]).launch()

    global_log.info("step4_pdb2gmx: Generate the topology")
    Pdb2gmx(**mut_paths["step4_pdb2gmx"], properties=mut_prop["step4_pdb2gmx"]).launch()
```

```yaml
mutations_list: ["A:Arg5Ala", "A:Arg5Gly", "A:Arg5Lys"]

step1_pdb:
  paths:
    output_pdb_path: structure.pdb
  properties:
    pdb_code: 1aki
```

PyCOMPSs automatic parallelization

```
print 'step2: mmbuniprot -- Get mutations'
mmbuniprot = uniprot.MmbVariants(prop['pdb_code'])
mutations = mmbuniprot.fetch_variants()

for mut in mutations:
    mut_path = cdir(wd, mut)

    print 'step3: scw -- Model mutation'
    scw_path = cdir(mut_path, 'step3_scw')
    scw_pdb = opj(scw_path, prop['mutated_pdb'])
    scw = scwrl.Scwrl4(mmbpdb_pdb, scw_pdb, mut, scwrl_path=scwrl_path)
    scw_pdb2 = scw.launchPyCOMPSs()
```





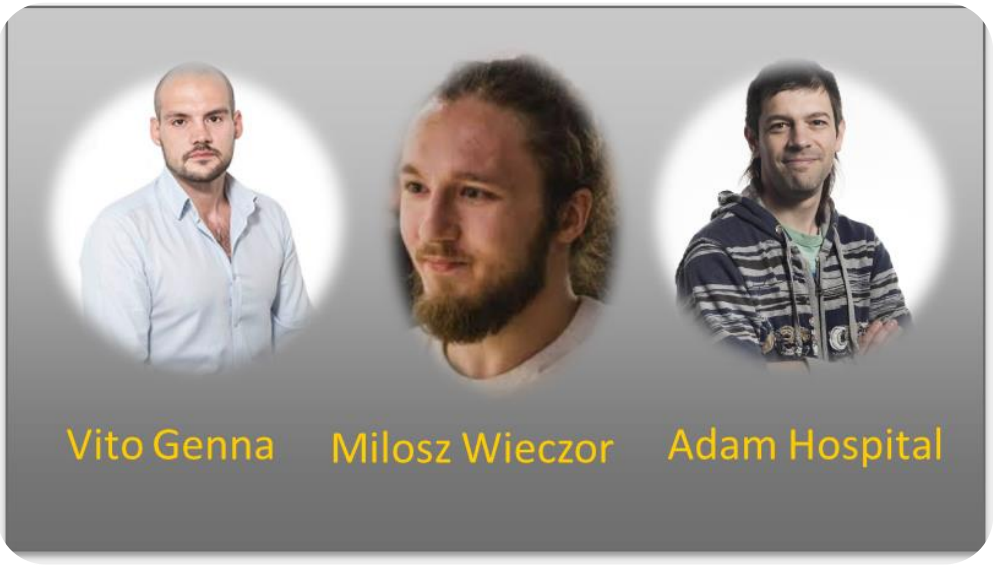
https://github.com/bioexcel/pymdsetup

https://doi.org/10.1177/1094342015594678

Modeling studies

Develop HPC tools

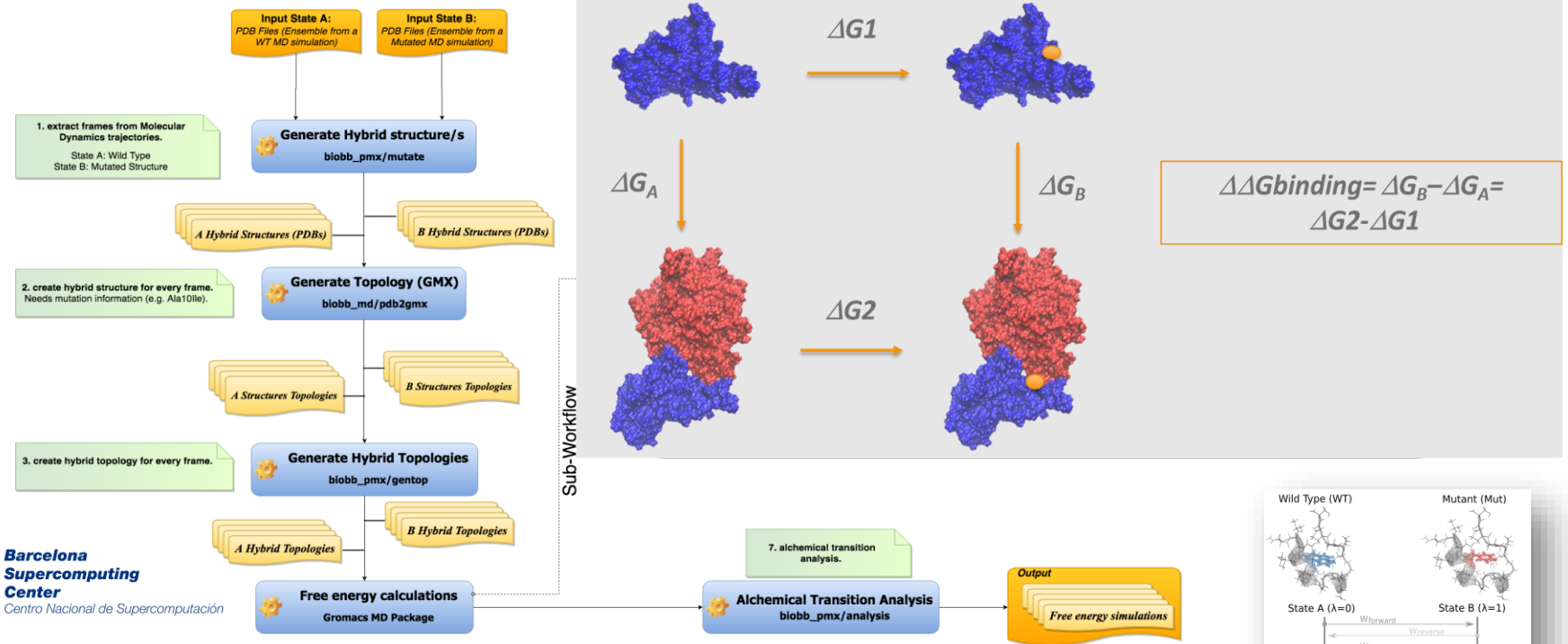
The COVID-19 bbb workflow team



PMX BBB workflows

Modeling studies

Develop HPC tools

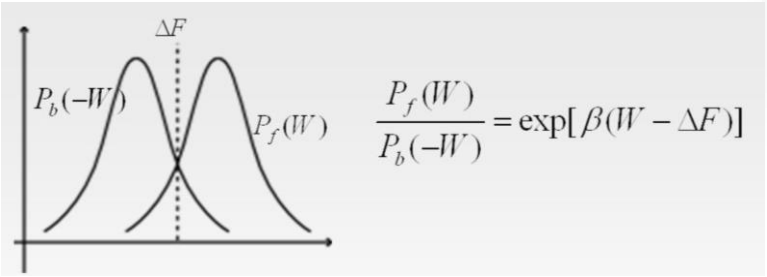


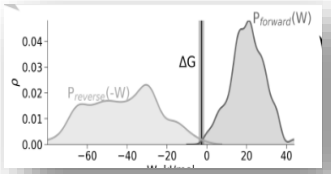
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

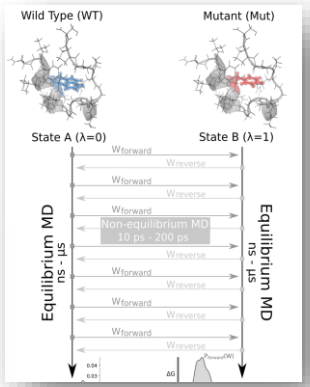
MAX-PLANCK-GESELLSCHAFT

IRB BARCELONA
INSTITUTE FOR RESEARCH IN BIOMEDICINE

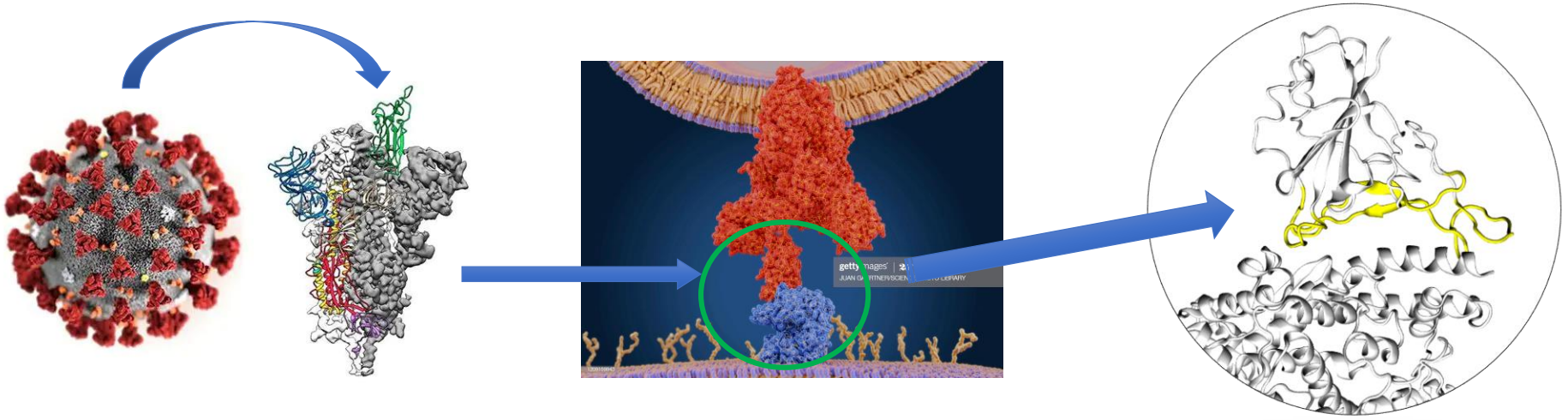
Crooks theorem

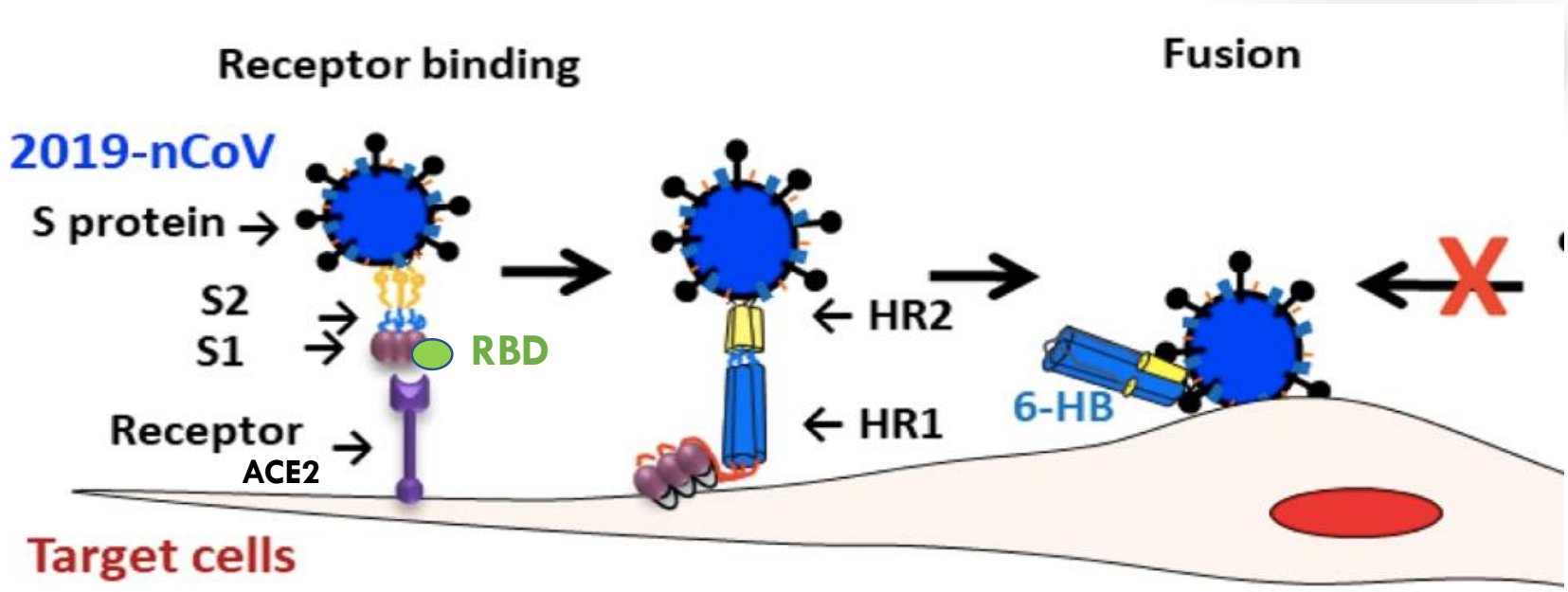


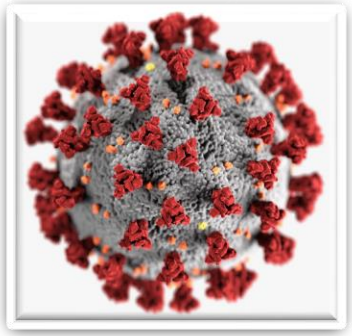




COVID-19: infective process







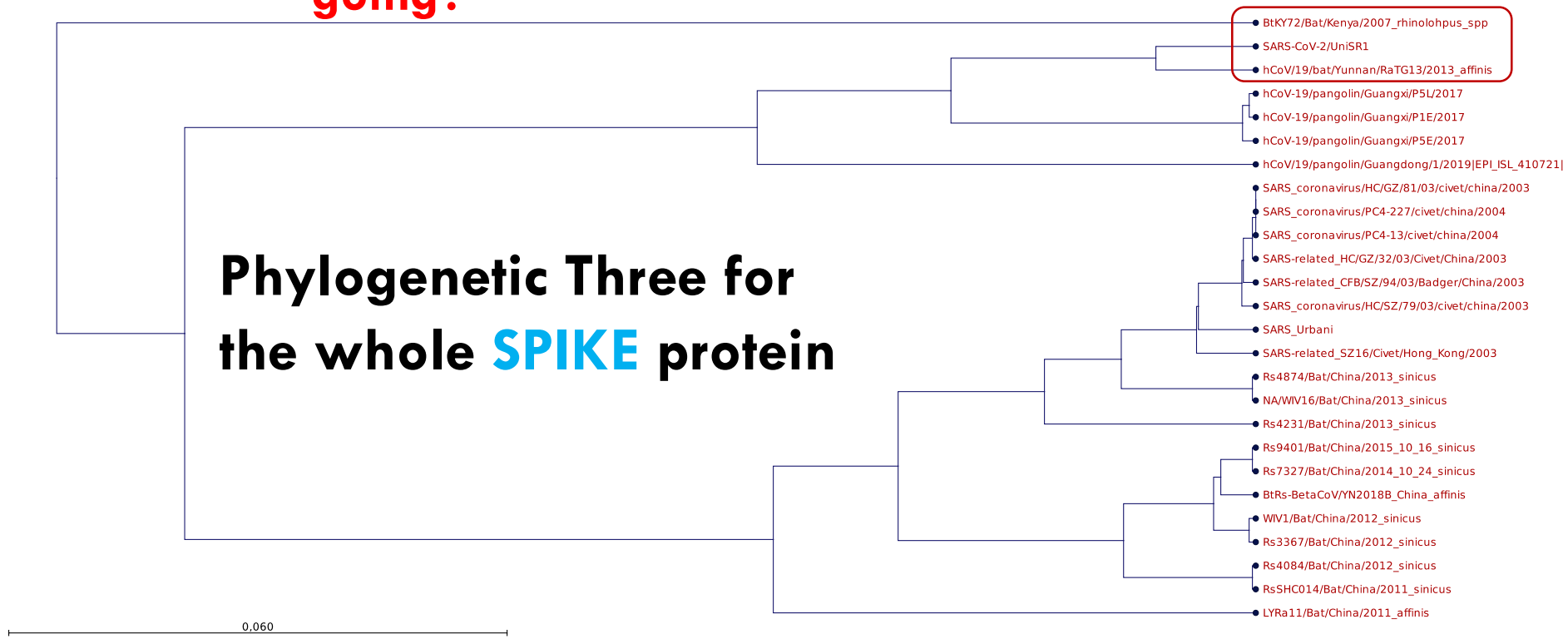
The trajectory of the virus before reaching humans

- ICEI/Fenix PRACE Access

PRACE-ICEI Call for Proposals Application Form

Project name	BioExcel biomolecular simulation workflows-3
Research field	Life Sciences, Biomolecular simulation, Structural Bioinformatics, Molecular Dynamics, Free energy, Docking

Where does SARS-Cov2 come from and where is it going?

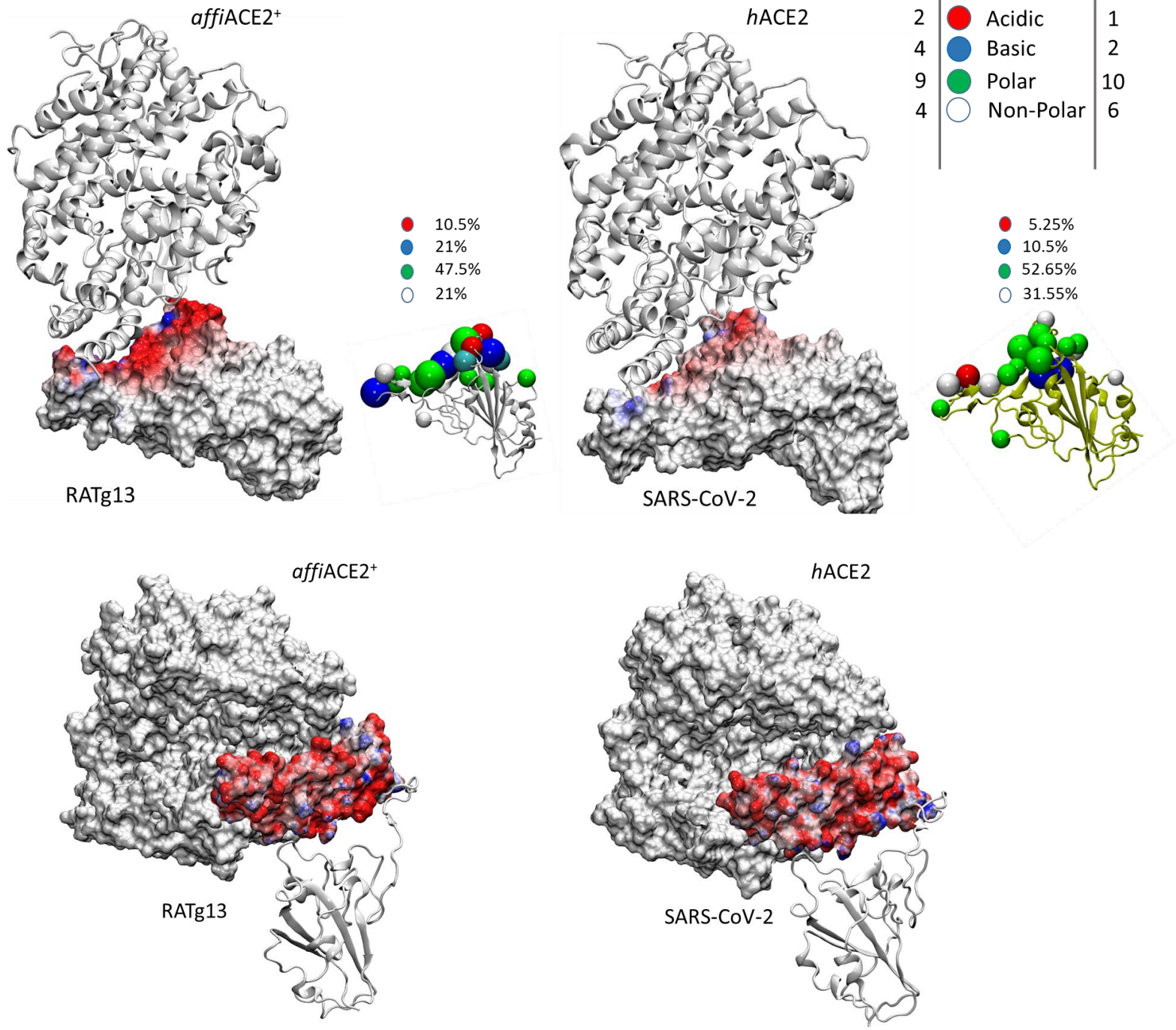


Phylogenetic Three for the whole SPIKE protein

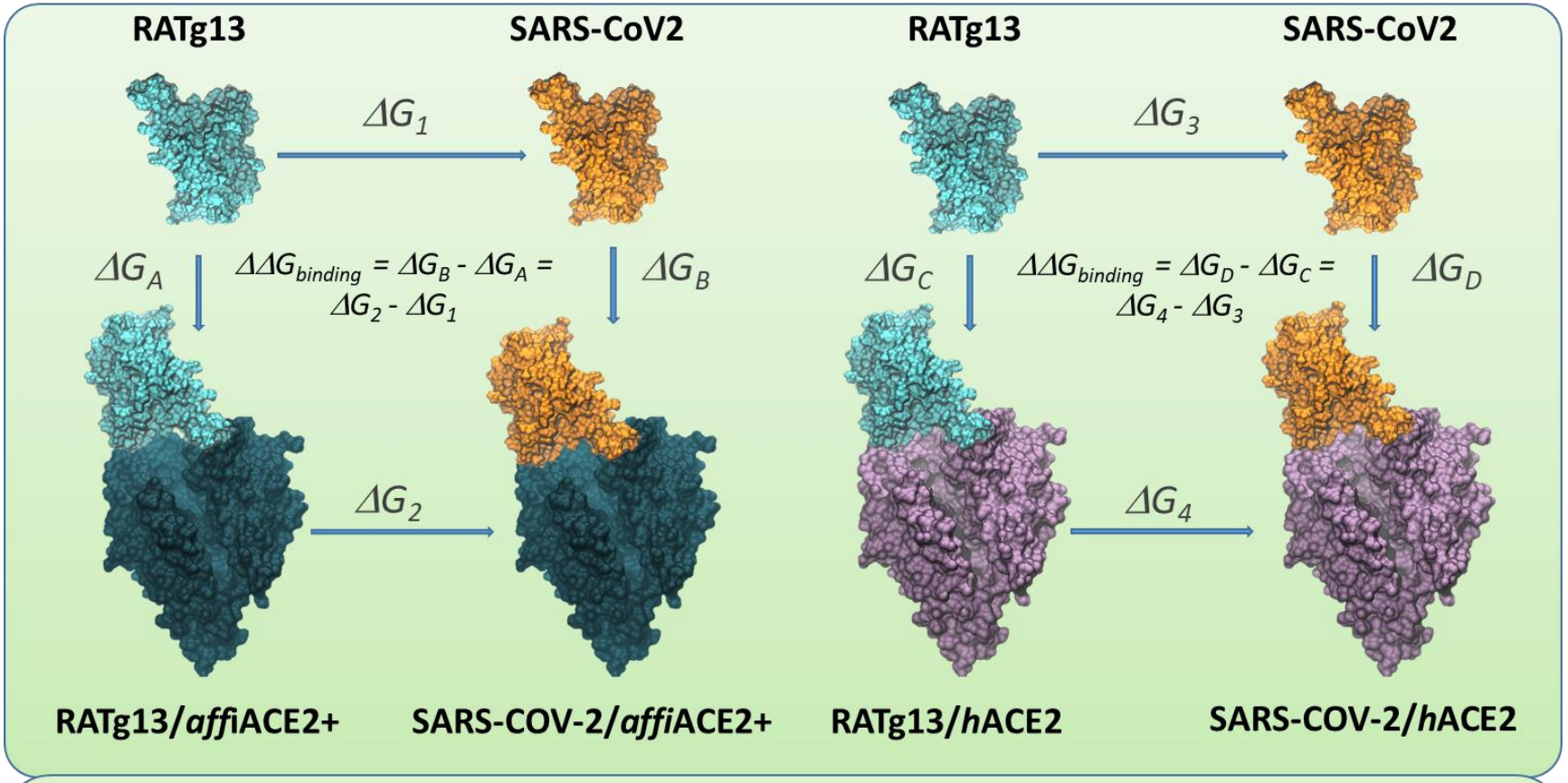
- BtKY72/Bat/Kenya/2007_rhinolohpus_spp
- SARS-CoV-2/UniSR1
- hCoV/19/bat/Yunnan/RaTG13/2013_affinis
- hCoV-19/pangolin/Guangxi/P5L/2017

RatG13 a virus from bat is the closest analog to SARS_Cov2

But RatG13 and SARS_Cov2 RBD are completely different


affiACE2+
hACE2
2
4
9
4
Acidic
Basic
Polar
Non-Polar
1
2
10
6
10.5%
21%
47.5%
21%
5.25%
10.5%
52.65%
31.55%
RATg13
SARS-CoV-2
affiACE2+
hACE2
RATg13
SARS-CoV-2

No zoonotic barrier!

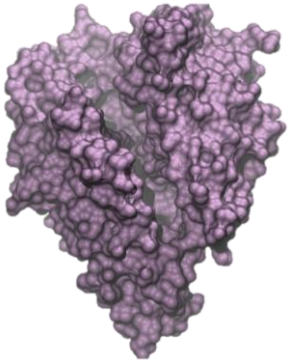


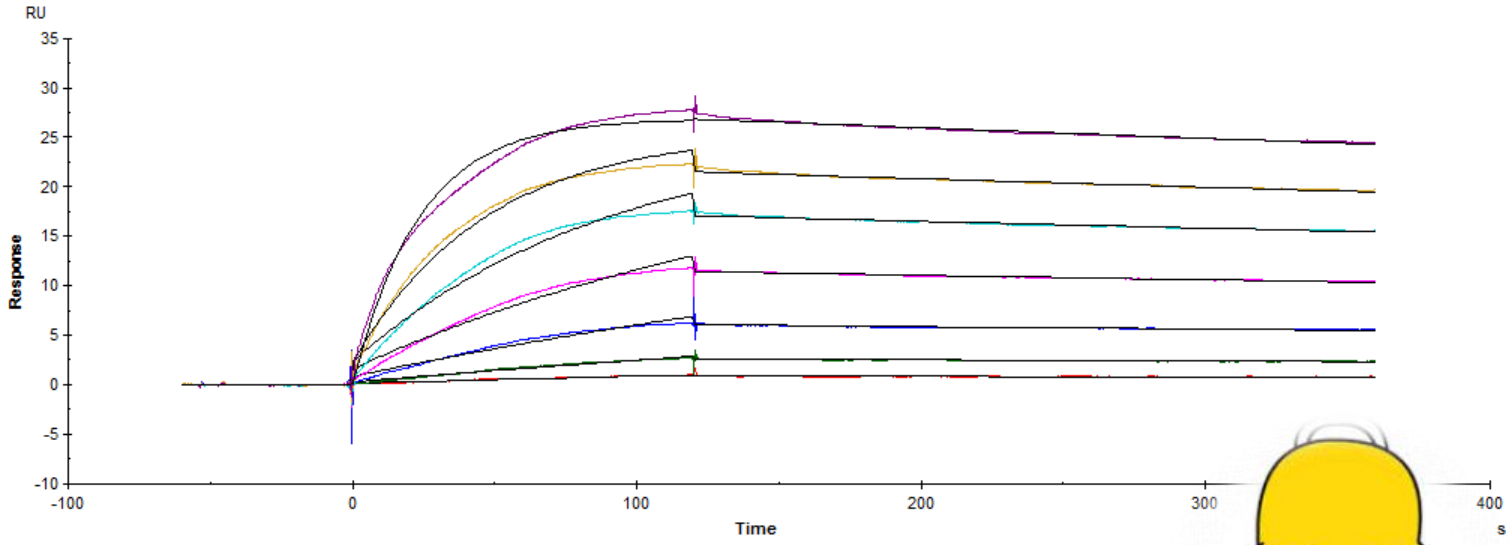
$\Delta\Delta G_{binding}$ (RG13 ➢ SC2)/Affi+ = **-1.6 ± 1.1** Kcal/mol $\Delta\Delta G_{binding}$ (RG13 ➢ SC2)/Human = **-0.4 ± 0.2** Kcal/mol

Experiments by M.Castelli; N.Clementi & M.Nicasio

UNIVERSITÀ DEGLI STUDI DI MILANO

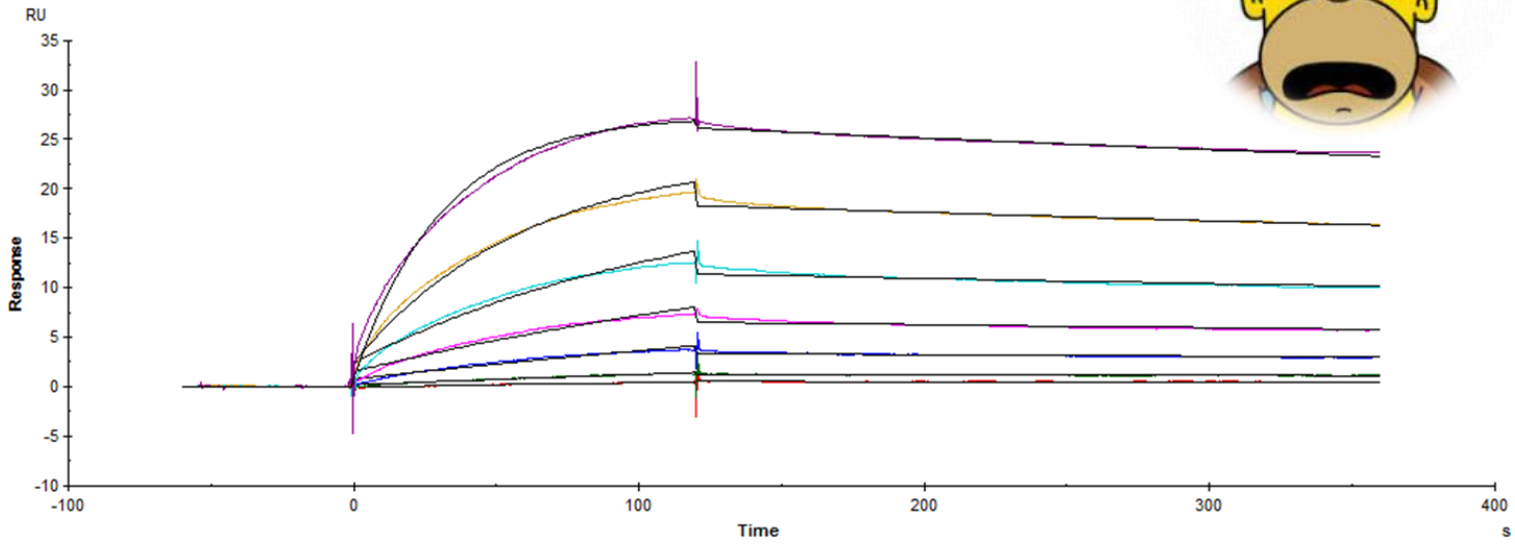
SARS_Cov2

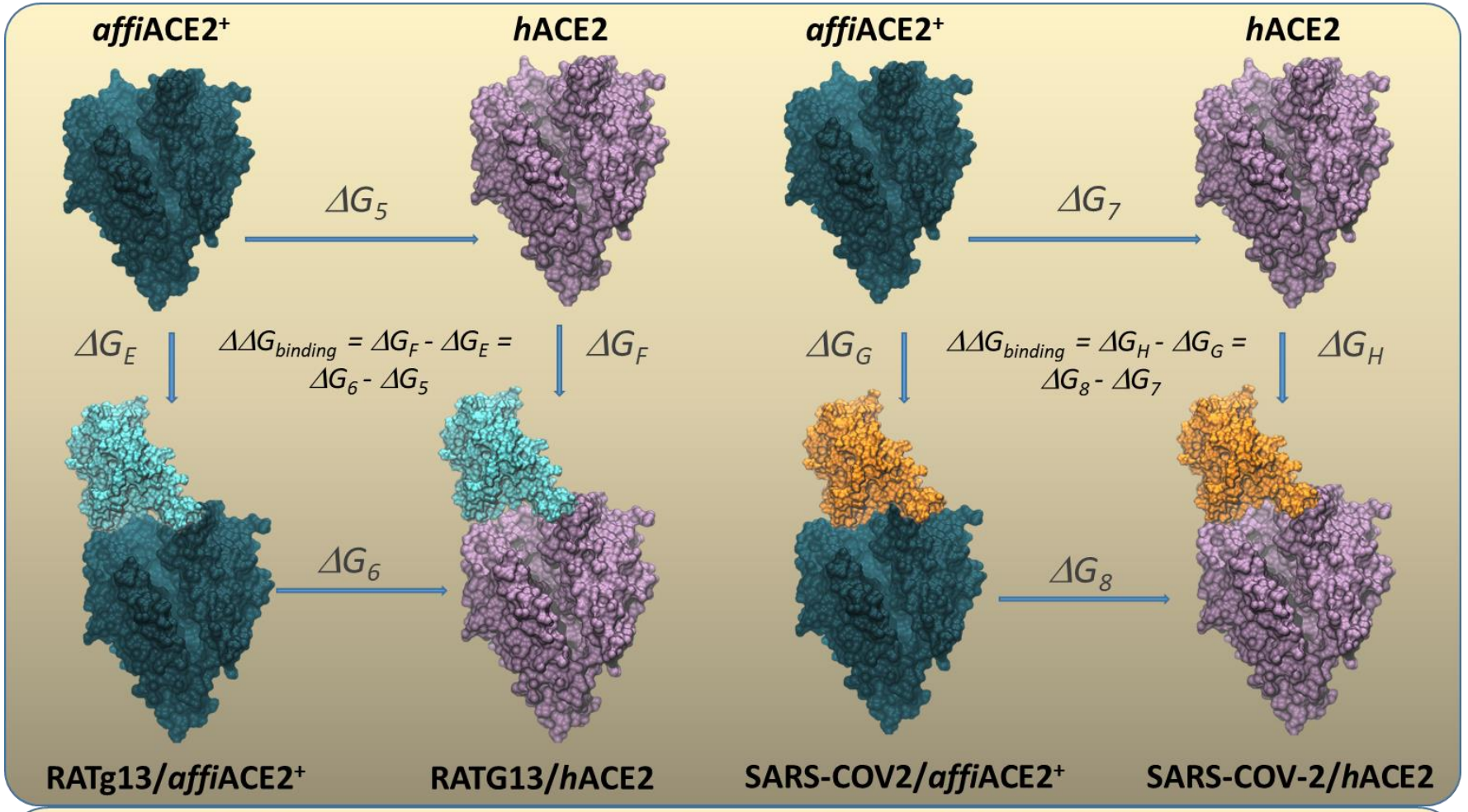




ΔG=-0.3 kcal/mol favoring SARS-Cov2 binding

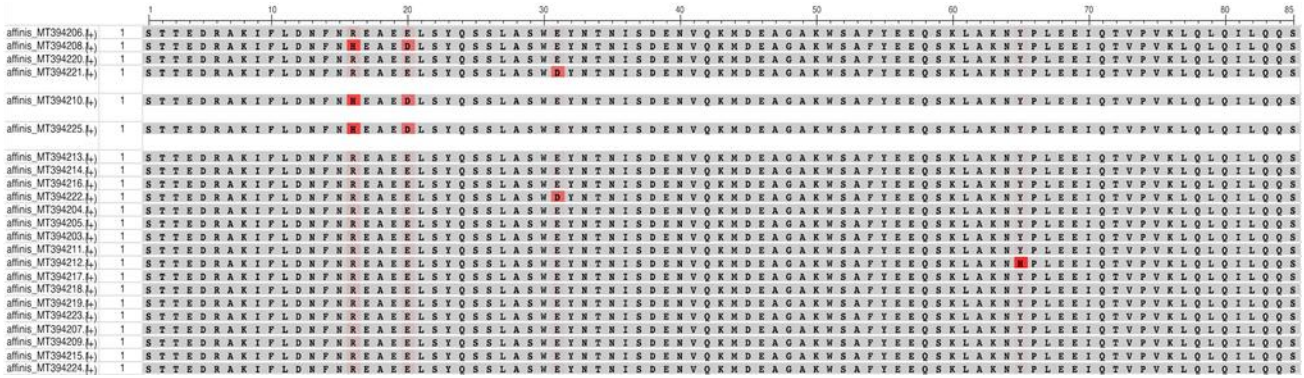
RATG13

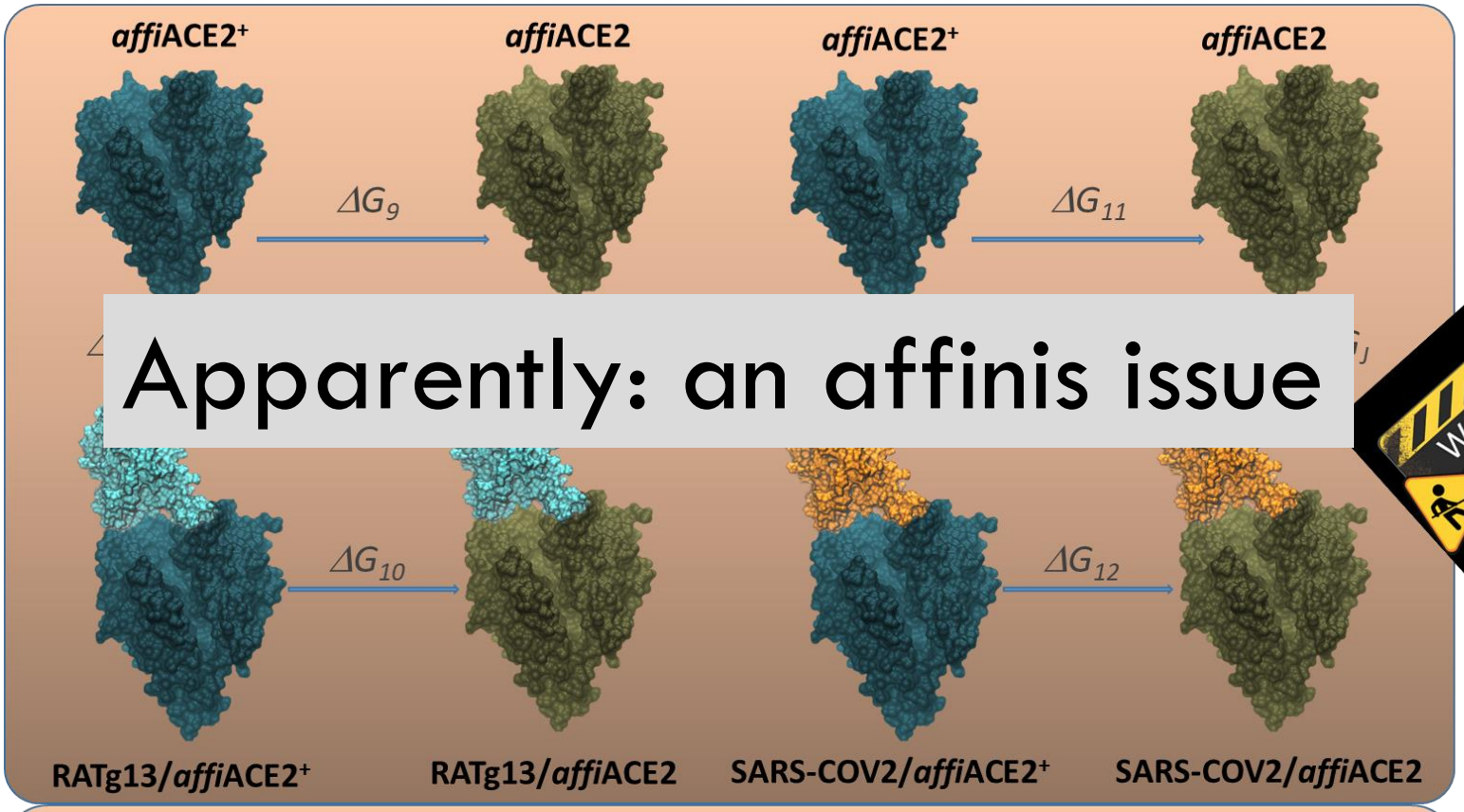




$\Delta\Delta G_{binding}$ (affiACE2$^+$ ➢ Human)/RATg13 = **2.6 ± 1.1** Kcal/mol

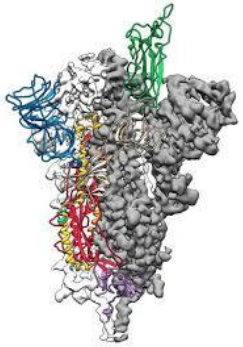
$\Delta\Delta G_{binding}$ (affiACE2$^+$ ➢ hACE2)/SC2 = **2.8 ± 1.3** Kcal/mol





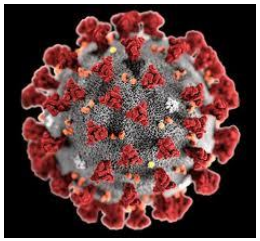
Apparently: an affinis issue

Be ready for further zoonotic transmission



PMX BB Workflow (human polymorphisms)

Modeling studies



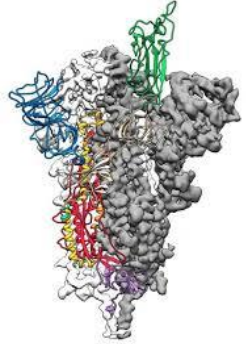
Variants	*Mutation*	*Allele frequency*	*ddG (Kcal/mol)*
rs73635825	S19P	3,13E-04	-
rs1299103394	K26E	5,45E-06	0,89
rs4646116	K26R	3,88E-03	-1,1
rs781255386	T27A	1,09E-05	0,1
rs778500138	E35D	N/A	0,1
rs1348114695	E35K	1,64E-05	4,25
rs146676783	E37K	3,90E-05	7,3
rs755691167	K68E	1,09E-05	2,36
rs766996587	M82I	2,44E-05	1,44
rs759134032	P84T	5,47E-06	-
rs143936283	E329G	3,44E-05	0,36
rs961360700	D355N	1,17E-05	3,48
rs1396769231	M383T	N/A	-1,24
rs762890235	P389H	3,83E-05	-
rs1238146879	P426A	5,47E-06	-
rs1316056737	D427Y	1,09E-05	0,38
rs1016777825	R559S	N/A	-2,14

bioexcel

MAX-PLANCK-GESELLSCHAFT

BSC Barcelona Supercomputing Center
Centro Nacional de Supercomputación

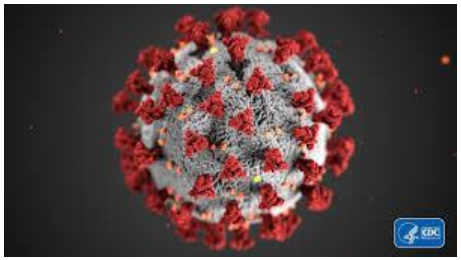
IRB BARCELONA
INSTITUTE FOR RESEARCH IN BIOMEDICINE



Modeling studies

Exploring virus

RBD position	original	variation	frequency	location	Fold-x ddG (Kcal/mol)	pmx ddG (Kcal/mol)	
439	N	K	213	Scotland	-1,88	-3,85	
444	K	R	1		-0,39	-0,77	
446	G	A	3	Australia	2,45	0,45	
445	V	I	1		-0,63	0,07	
446	G	S	3		3,39	0,53	
446	G	V	3	Australia/England	4,07	-0,22	
455	L	F	1		3,22	2,48	
456	F	L	1		2,61	1,55	
475	A	V	4	USA	1,81	-1,71	
476	G	S	19	USA (1 in Belgium)	1,9	0,31	0.2
478	T	I	65	England	-0,85	-0,06	
483	V	A	30	USA	-0,2	0,07	-0.5
483	V	I	2	UK	0,04	0,03	
483	V	F	5	Spain	-0,2	0,21	
484	E	A	1		1,06	-	
490	F	S	3	England	1,11	0,25	
490	F	L	2	Australia	0,64	-0,3	
494	S	P	3	England	-	-	
495	Y	N	1	Luxembourg	5,25	1,39	
503	V	F	1	USA	-0,71	0,19	



SARS-Cov2 is selecting mutations favoring binding!

MAX-PLANCK-GESELLSCHAFT

BSC Barcelona Supercomputing Center
Centro Nacional de Supercomputación

IRB BARCELONA
INSTITUTE FOR RESEARCH IN BIOMEDICINE

bioexcel

BioExcel Partners 2019

BioExcel is funded by the European Union Horizon 2020 program under grant agreements 675728 and 823830.