bioexcel
Centre of Excellence for Computational Biomolecular Research

Horizon 2020
European Union Funding
for Research & Innovation

BioExcel HPC Workflows: predictive power and its applications in pharmacology

BioExcel Webinar, 2022-04-26

Adam Hospital, Miłosz Wieczór, Federica Battistini

Molecular Modeling and Bioinformatics, IRB Barcelona

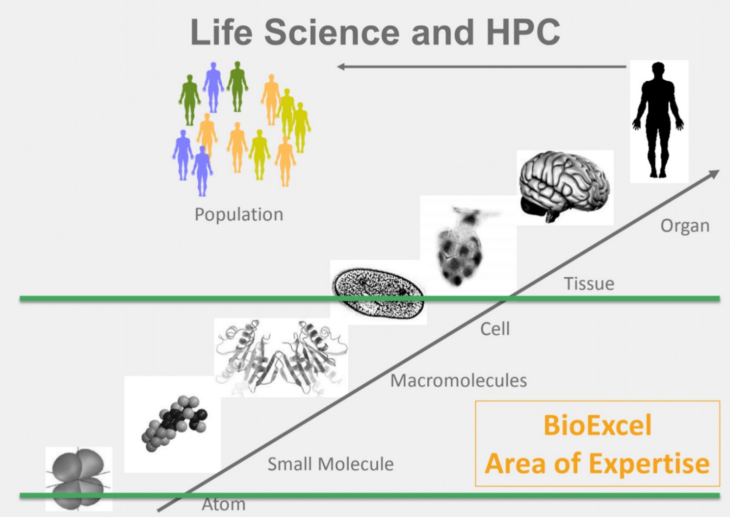
IRB
BARCELONA
INSTITUTE
FOR RESEARCH
IN BIOMEDICINE

bioexcel
Centre of Excellence for Computational Biomolecular Research

HORIZON 2020

Centre of Excellence for Computational Biomolecular Research

A central hub for biomolecular modelling and simulations



Enabling **better science** by:

- **Improving** the **performance** and **functionality** of **key applications**

FAST. FLEXIBLE. FREE. GROMACS

HADDOCK High-Ambiguity Driven Docking

CP2K

CPMD QM/MM

pmx: generate hybrid protein structure and topology
Computational Biomolecular Dynamics Group

- Developing **user-friendly computational workflows**

KTH VETENSKAP OCH KONST

IRB BARCELONA INSTITUTE FOR RESEARCH IN BIOMEDICINE

MMB

EMBL-EBI

MANCHESTER 1824 The University of Manchester

MAX-PLANCK-GESELLSCHAFT

THE UNIVERSITY OF EDINBURGH

JÜLICH FORSCHUNGSZENTRUM

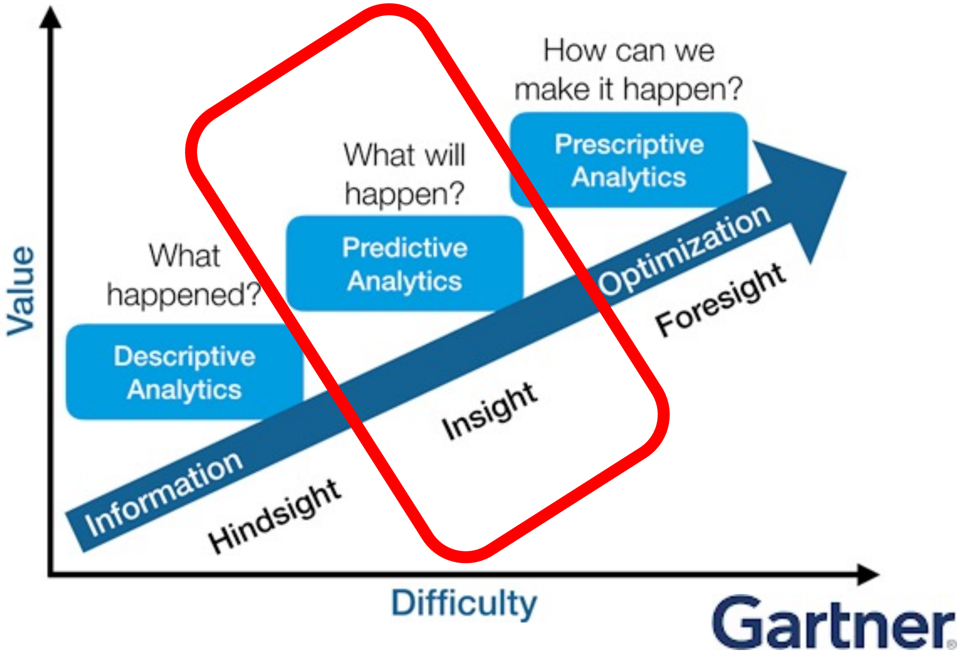
acrosslimits

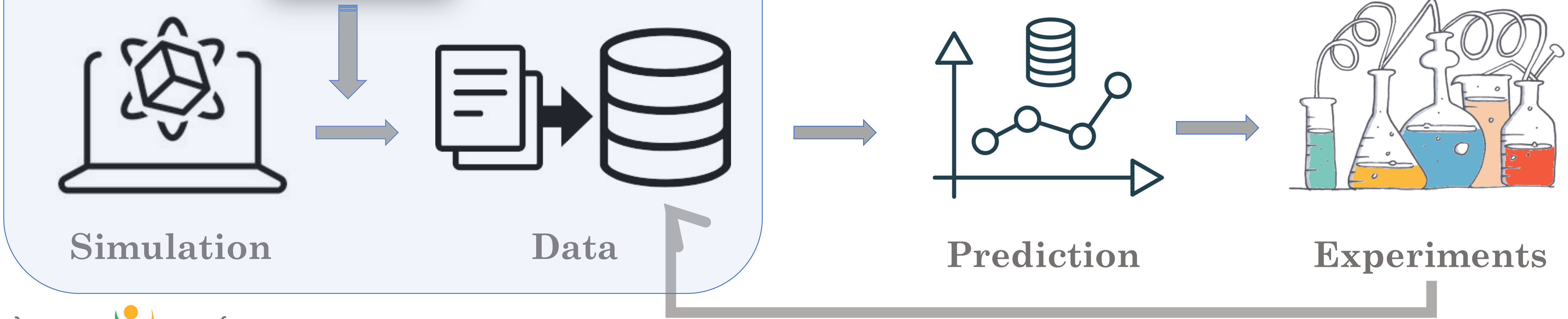
BSC Barcelona Supercomputing Center Centro Nacional de Supercomputación

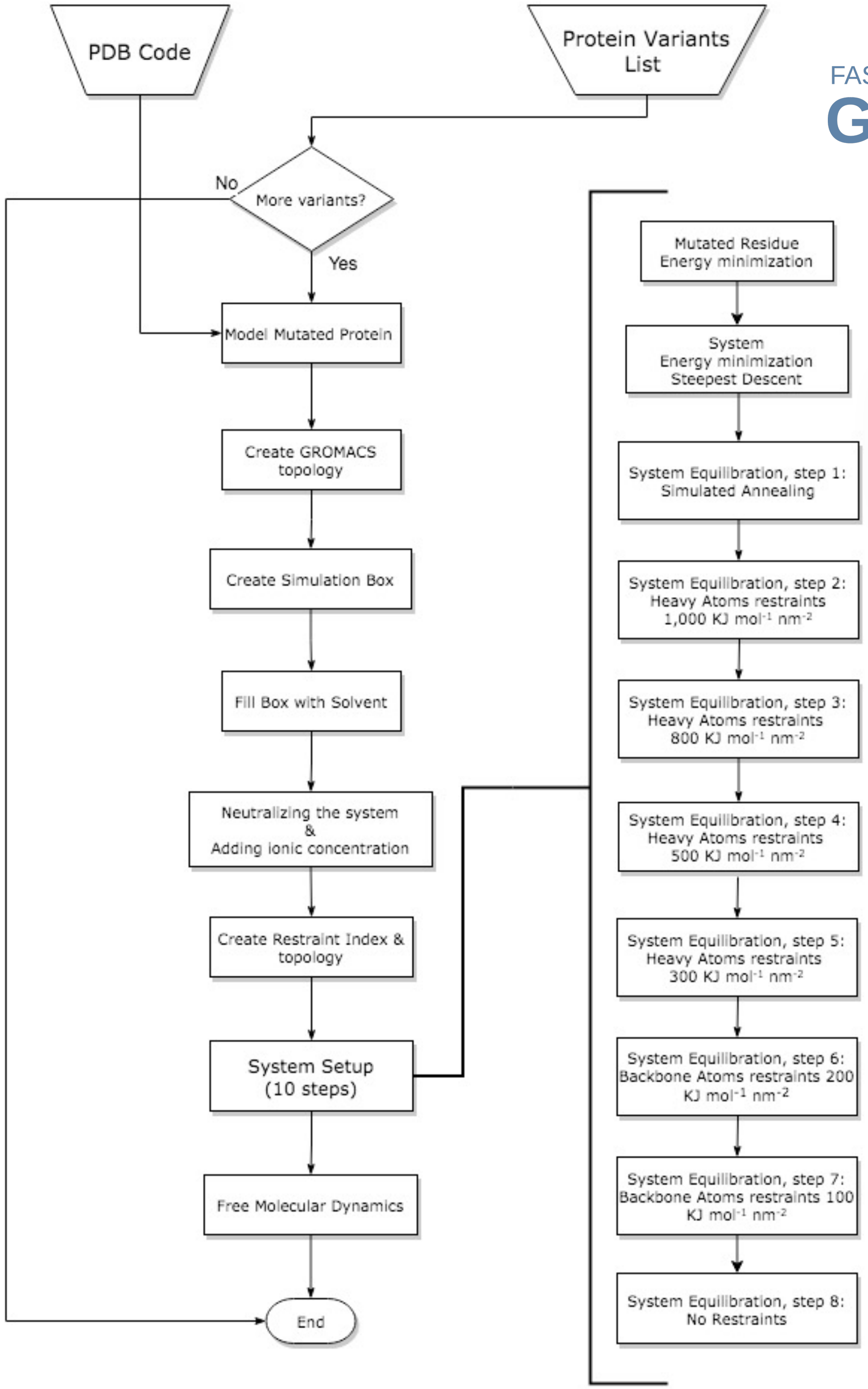
Universiteit Utrecht

ih Ian Harrow Consulting

Data-Driven Science







PDB Code

Protein Variants List

More variants?

No

Yes

Model Mutated Protein

Create GROMACS topology

Create Simulation Box

Fill Box with Solvent

Neutralizing the system & Adding ionic concentration

Create Restraint Index & topology

System Setup (10 steps)

Free Molecular Dynamics

End

Mutated Residue Energy minimization

System Energy minimization Steepest Descent

System Equilibration, step 1: Simulated Annealing

System Equilibration, step 2: Heavy Atoms restraints 1,000 KJ mol^{-1} nm^{-2}

System Equilibration, step 3: Heavy Atoms restraints 800 KJ mol^{-1} nm^{-2}

System Equilibration, step 4: Heavy Atoms restraints 500 KJ mol^{-1} nm^{-2}

System Equilibration, step 5: Heavy Atoms restraints 300 KJ mol^{-1} nm^{-2}

System Equilibration, step 6: Backbone Atoms restraints 200 KJ mol^{-1} nm^{-2}

System Equilibration, step 7: Backbone Atoms restraints 100 KJ mol^{-1} nm^{-2}

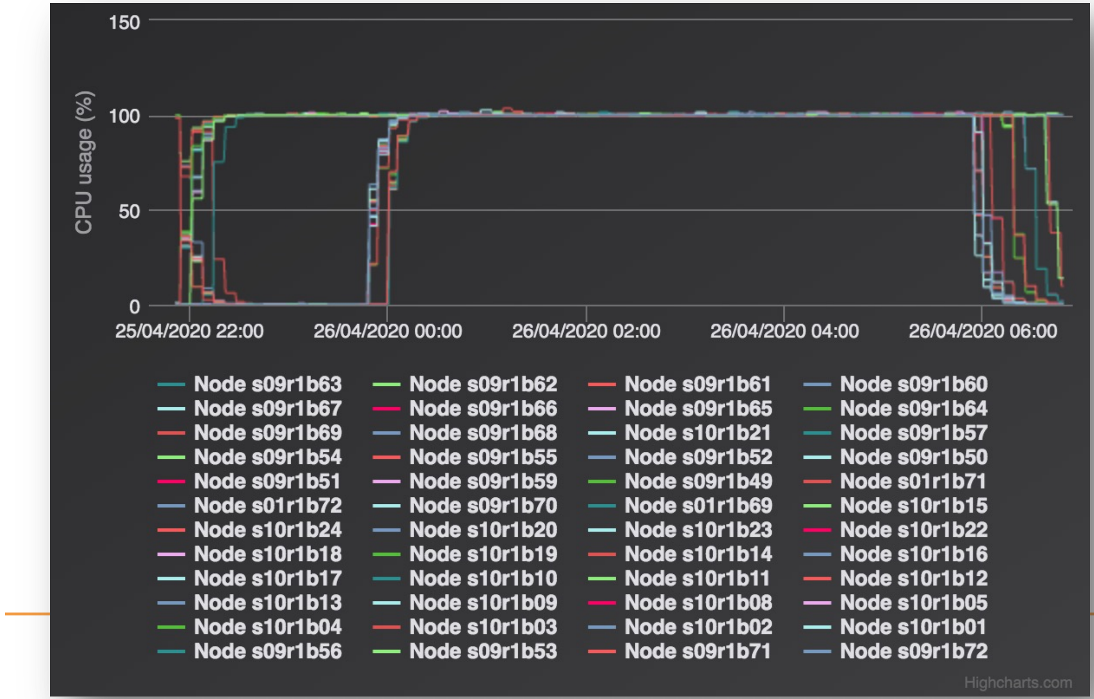
System Equilibration, step 8: No Restraints

FAST. FLEXIBLE. FREE. GROMACS

biobb

pyCOMPSs

Mutation Modeling + MD Setup + MD Run

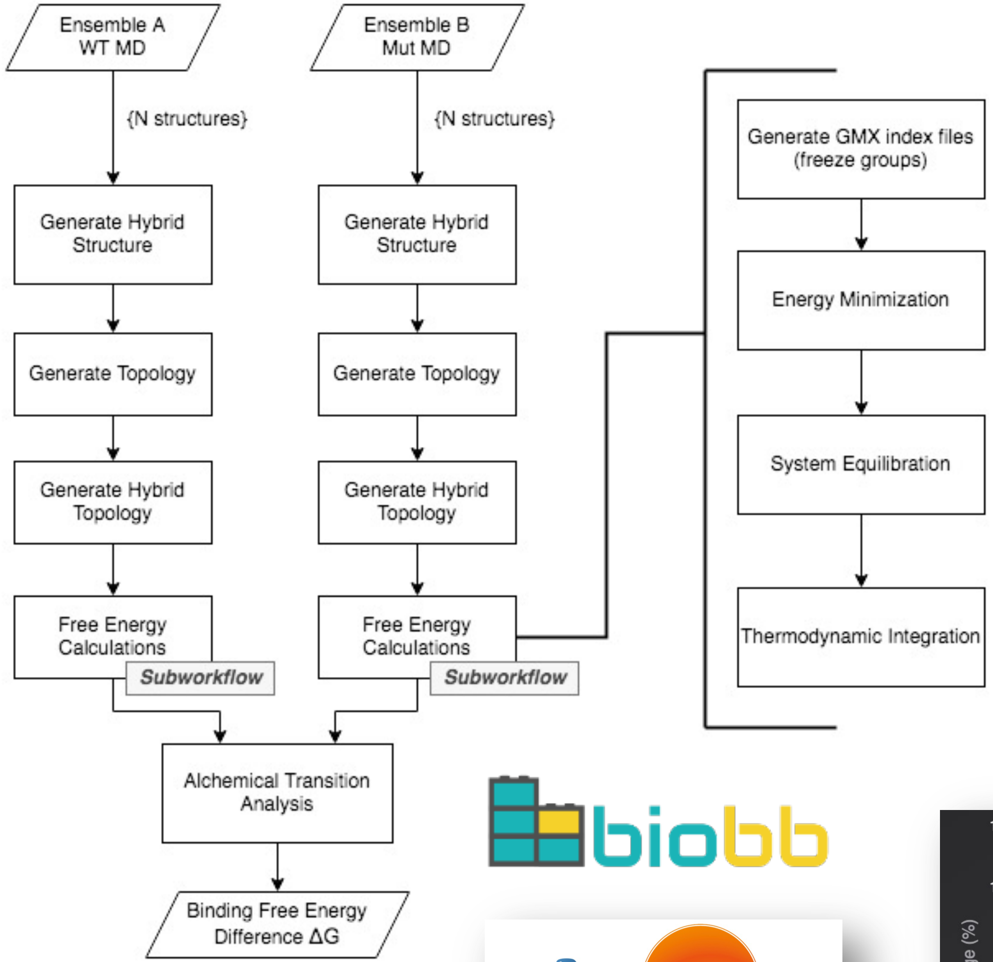


BSC Barcelona Supercomputing Center
Centro Nacional de Supercomputación

48 MareNostrum nodes
2,304 cores → 1 job

12 mutations
10ns-length MDs
GROMACS 4 nodes MPI

Time: 8h



Ensemble A
WT MD

{N structures}

Generate Hybrid Structure

Generate Topology

Generate Hybrid Topology

Free Energy Calculations

Subworkflow

Ensemble B
Mut MD

{N structures}

Generate Hybrid Structure

Generate Topology

Generate Hybrid Topology

Free Energy Calculations

Subworkflow

Alchemical Transition Analysis

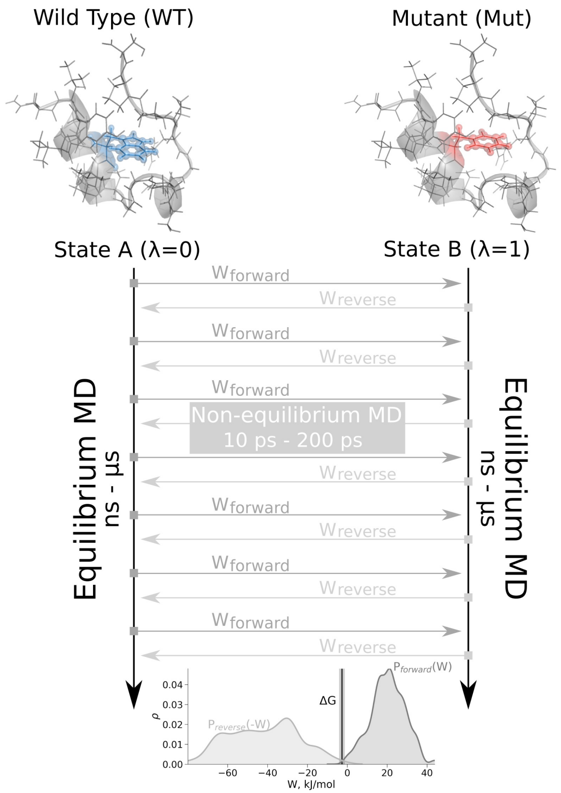
Binding Free Energy Difference ΔG

Generate GMX index files (freeze groups)

Energy Minimization

System Equilibration

Thermodynamic Integration



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Non-equilibrium free energy calculation

> J Comput Chem. 2015 Feb 15;36(5):348-54. doi: 10.1002/jcc.23804. Epub 2014 Dec 8.

pmx: Automated protein structure and topology generation for alchemical perturbations

Vytautas Gapsys [1], Servaas Michielssens, Daniel Seeliger, Bert L de Groot

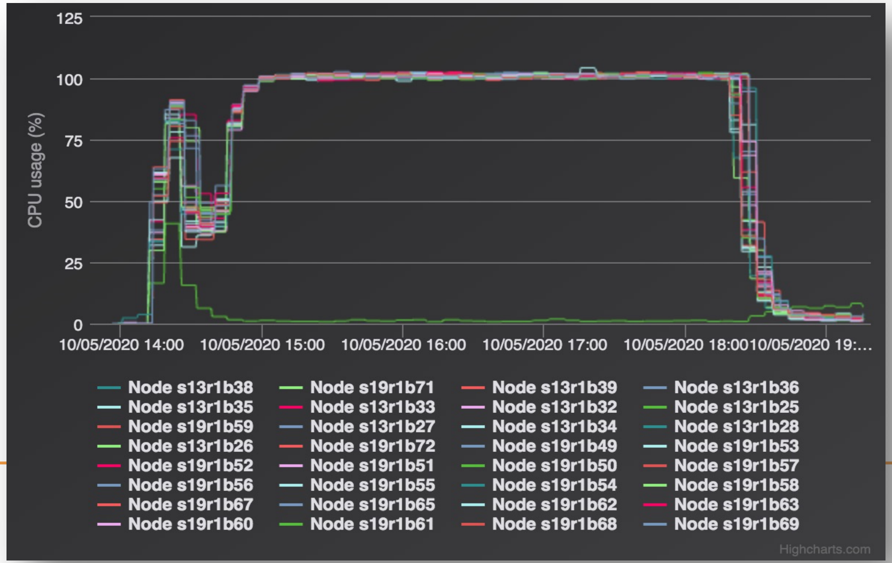
biobb

pyCOMPSs

FAST. FLEXIBLE. FREE.
GROMACS

pmx: generate hybrid protein structure and topology
Computational Biomolecular Dynamics Group

bioexcel



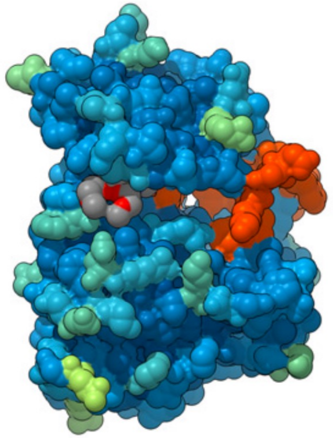
32 MareNostrum nodes
1,536 cores → 1 job

1000 short TI MDs (50ps)
500 forward
+
500 reverse

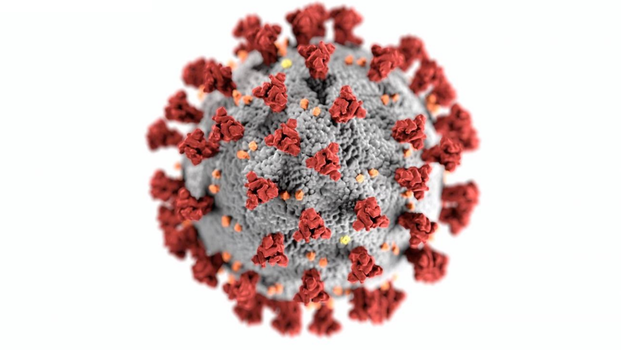
Time: 5h

Projects

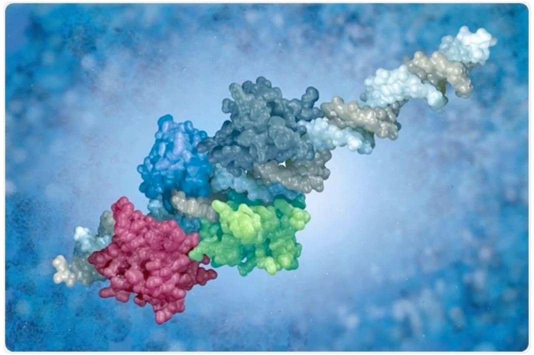
- High-throughput **prediction** of the **impact** of **genetic variability** on **drug sensitivity** and **resistance patterns** for clinically relevant **EGFR mutations** from atomistic simulations.



- Large-scale **SARS-CoV2 mutation** analysis, including a study on the **evolutionary path** and **host-selection mechanism** of **SARS-CoV-2**.

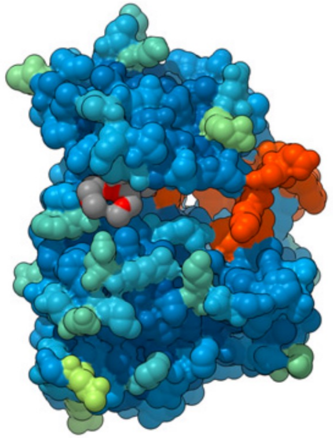


- **DNAffinity**: A **Machine-Learning** approach to **predict DNA Binding affinities** of **Transcription Factors**.

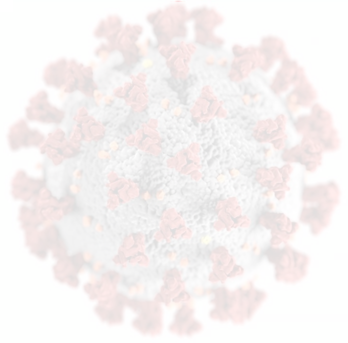


Projects

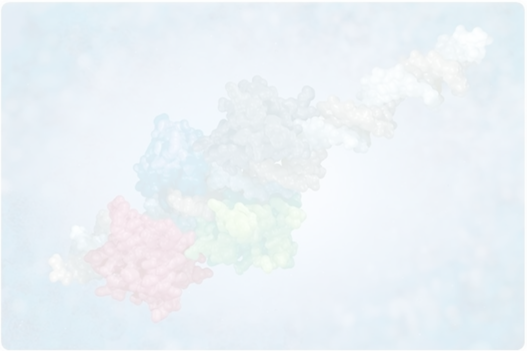
- High-throughput **prediction** of the **impact** of **genetic variability** on **drug sensitivity** and **resistance patterns** for clinically relevant **EGFR mutations** from atomistic simulations.



- Large-scale **SARS-CoV2 mutation** analysis, including a study on the **evolutionary path** and **host-selection mechanism** of **SARS-CoV-2**.

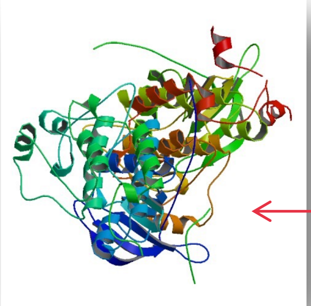


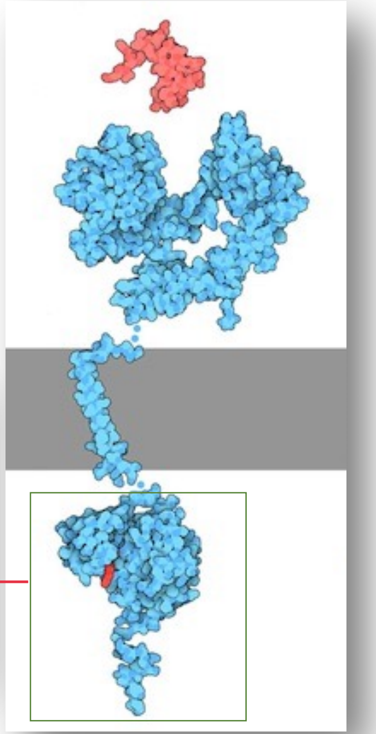
- **DNAffinity**: A **Machine-Learning** approach to **predict DNA Binding affinities** of **Transcription Factors**.

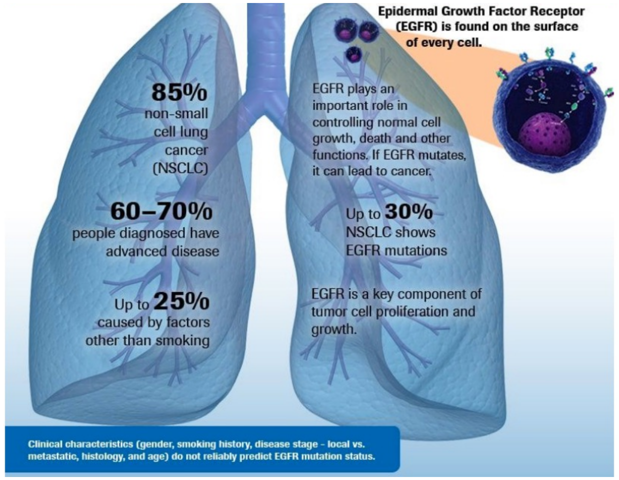


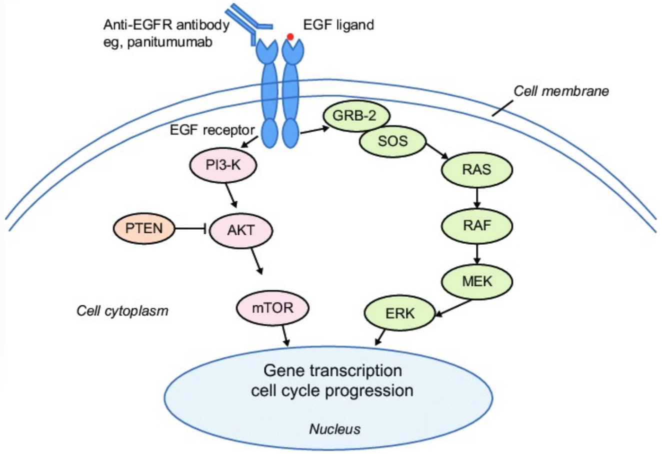
- ❑ **Epidermal Growth Factor Receptor (EGFR) -** (Kinase Domain)
- ❑ **EGFR mutations** drive some types of cancers, like **carcinoma, glioblastoma or NSCLC.**
- ❑ **Key component** of **tumor cell proliferation and growth.**
- ❑ **Two therapeutic approaches:**
  - ❑ **Monoclonal antibodies** (extracellular domain)
  - ❑ **ATP competitive inhibitors** (intracellular domain)

- ❑ Selected mutations from literature:
  - ○ **T790M** (**gatekeeper**) confers resistance to ***Erlotinib*** and ***Gefitinib*** by increasing ATP binding.
  - ○ L718Q, L747F, L747H kill ***Osimertinib***
  - ○ G719S, S768I, L833V enhances **Gefitinib**






Epidermal Growth Factor Receptor (EGFR) is found on the surface of every cell.
85% non-small cell lung cancer (NSCLC)
EGFR plays an important role in controlling normal cell growth, death and other functions. If EGFR mutates, it can lead to cancer.
60–70% people diagnosed have advanced disease
Up to 30% NSCLC shows EGFR mutations
Up to 25% caused by factors other than smoking
EGFR is a key component of tumor cell proliferation and growth.
Clinical characteristics (gender, smoking history, disease stage - local vs. metastatic, histology, and age) do not reliably predict EGFR mutation status.

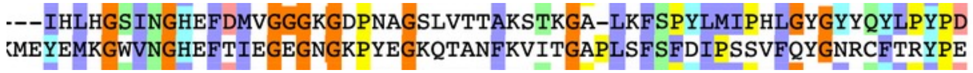


NBD
NOSTRUM BIODISCOVERY

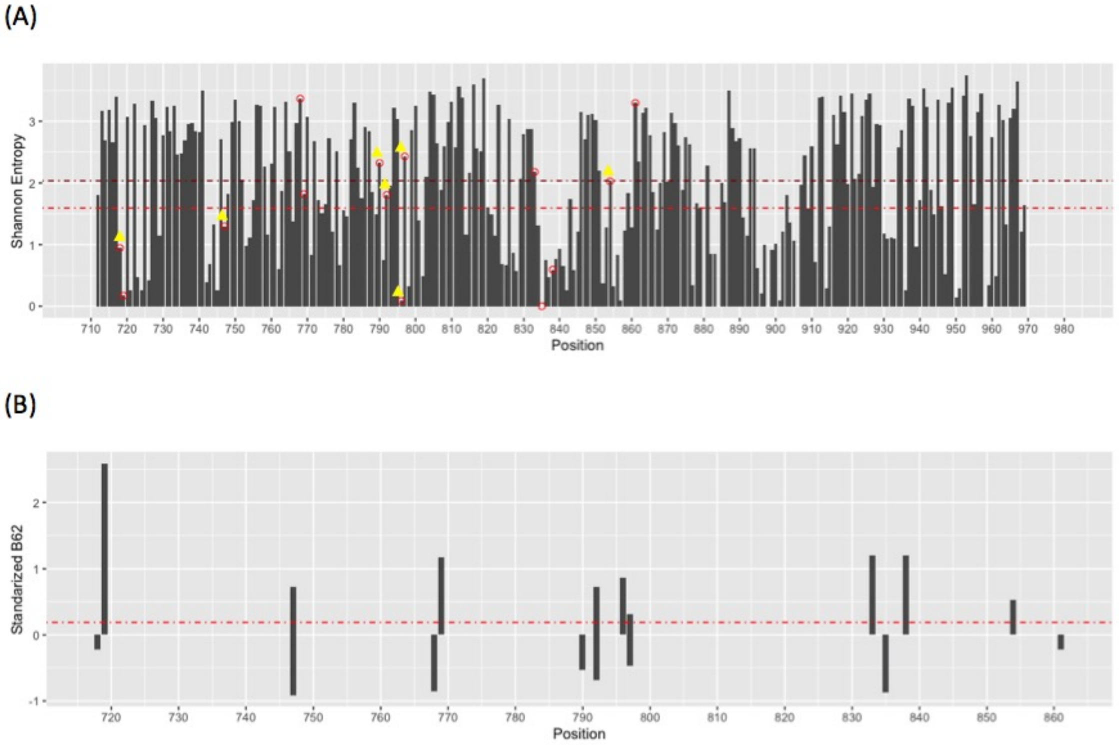
Could we predict the effect of the mutations?

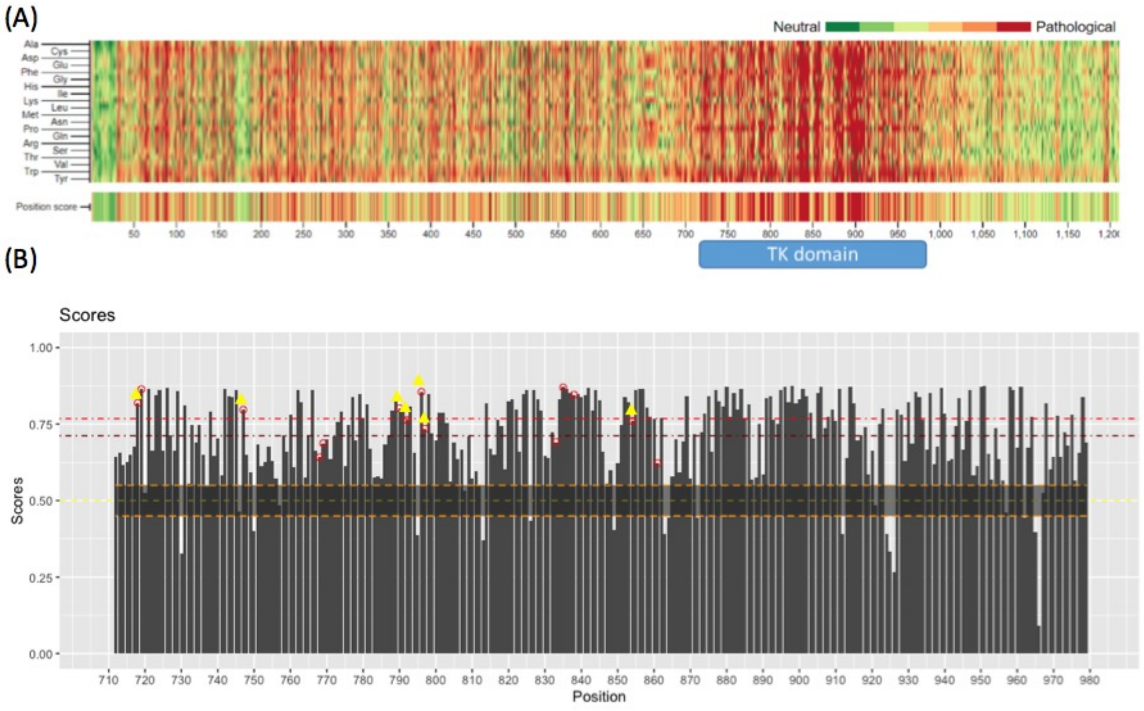
bioexcel

Sequence



```
---IHLHGSINGHEFDMVGGGKGDPNAGSLVTTAKSTKGA-LKFSPYLMIPHLGYGYYQYLPYPD
KMEYEMKGWVNGHEFTIEGEGNGKPYEGKQTANFKVITGAPLSFSFDIPSSVFQYGNRCFTRYPE
```

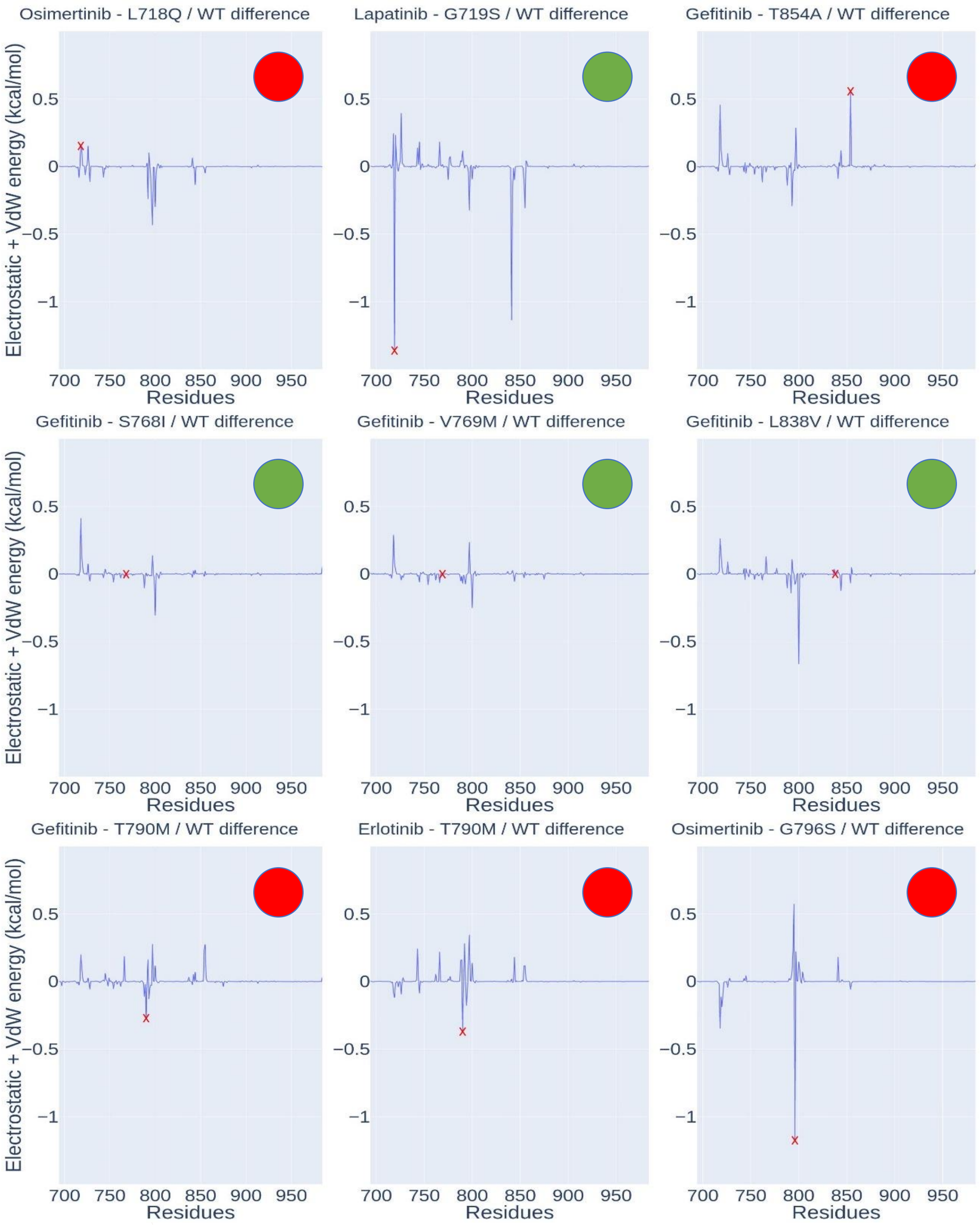


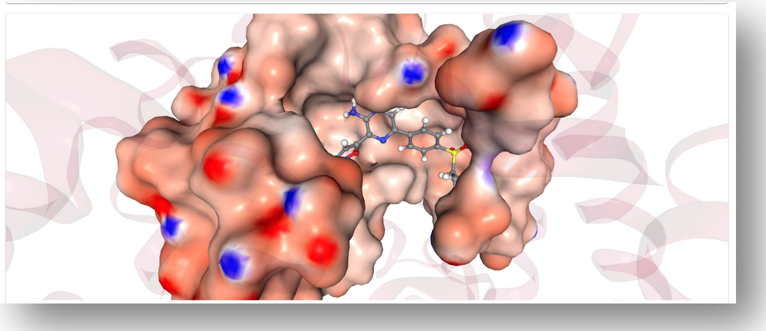


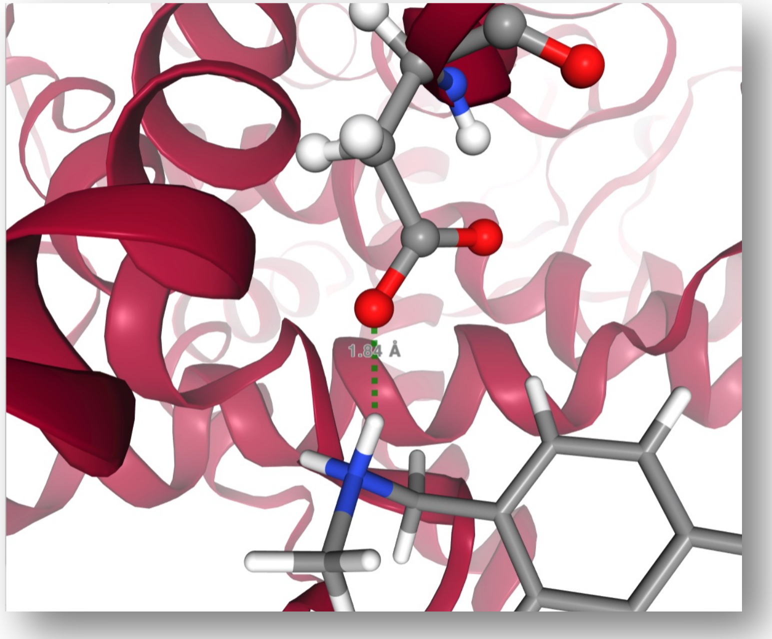
PMut: a web-based tool for the annotation of pathological variants on proteins, 2017 update

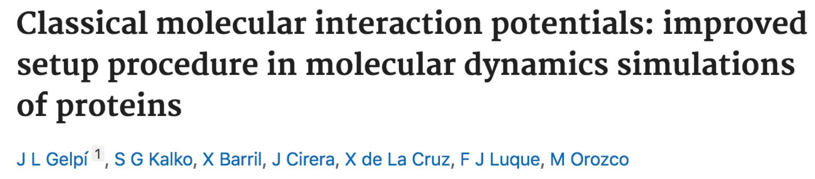
Víctor López-Ferrando, Andrea Gazzo, Xavier de la Cruz, Modesto Orozco ✉, Josep Ll Gelpí ✉

Nucleic Acids Research, Volume 45, Issue W1, 3 July 2017, Pages W222–W228, https://doi.org/10.1093/nar/gkx313



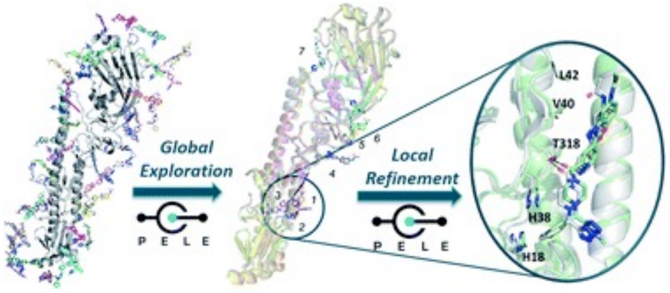


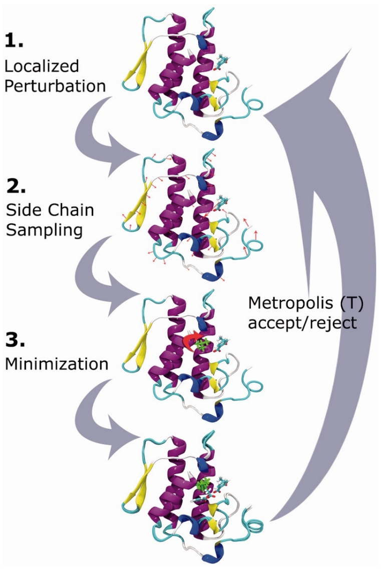


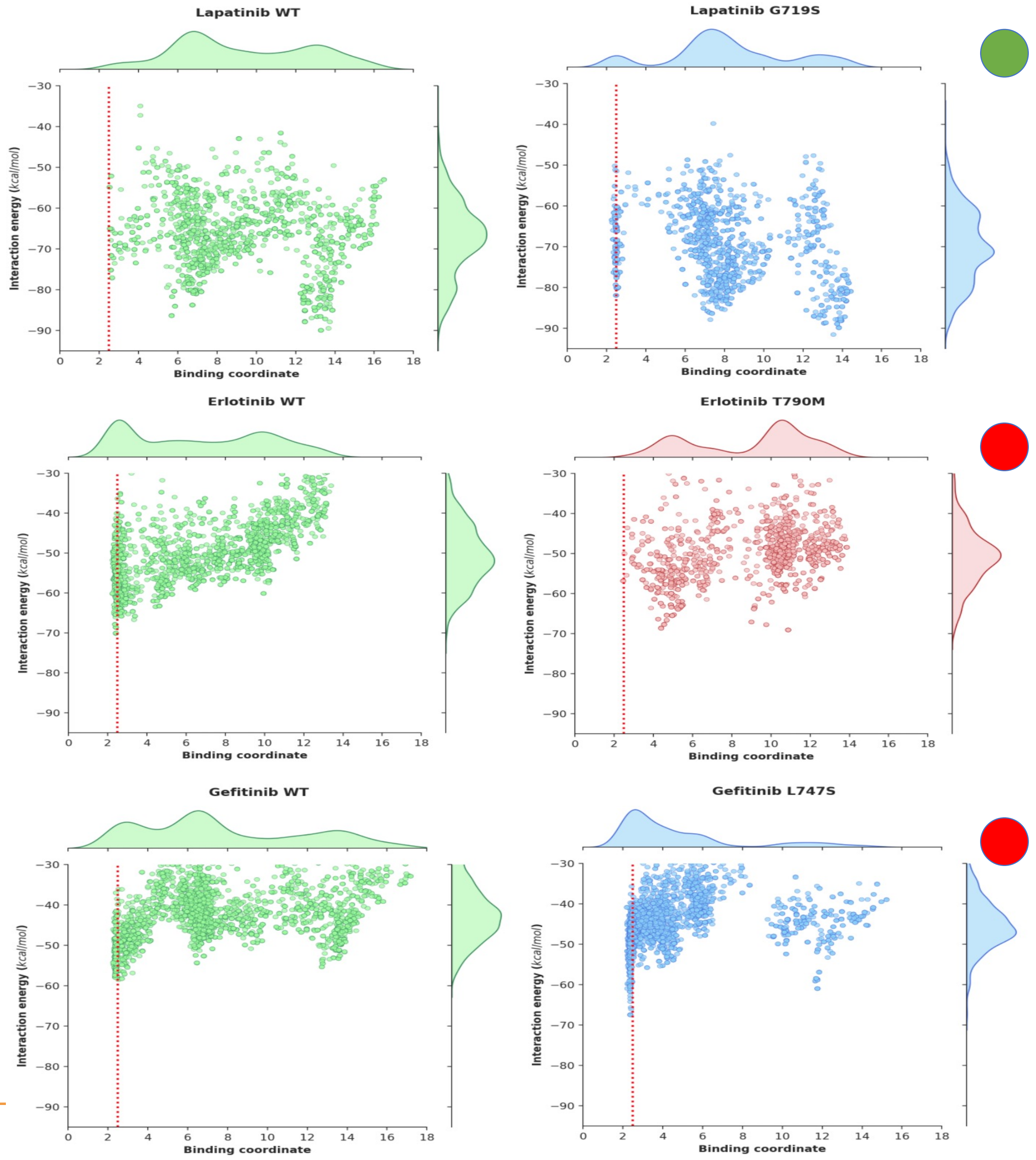


Classical molecular interaction potentials: improved setup procedure in molecular dynamics simulations of proteins

J L Gelpí[1], S G Kalko, X Barril, J Cirera, X de La Cruz, F J Luque, M Orozco



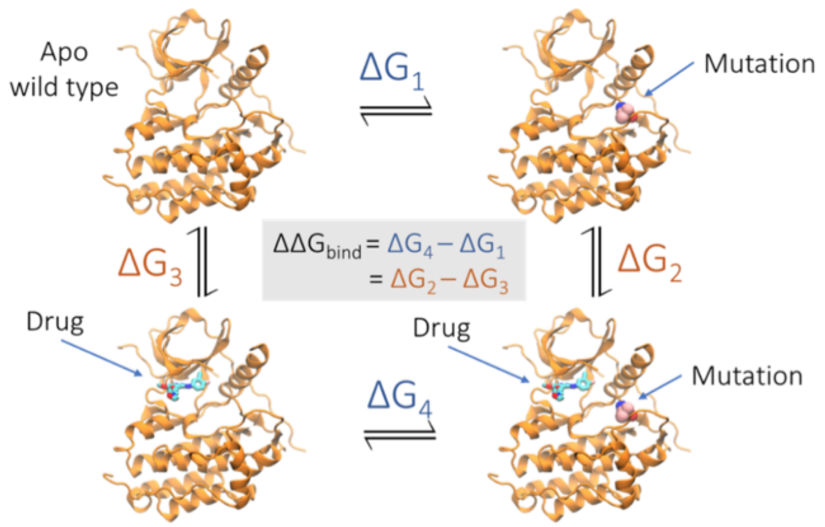


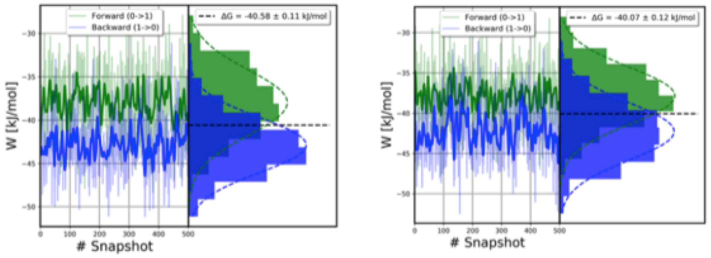


> Nucleic Acids Res. 2013 Jul;41(Web Server issue):W322-8. doi: 10.1093/nar/gkt454.
Epub 2013 May 31.

PELE web server: atomistic study of biomolecular systems at your fingertips

Armin Madadkar-Sobhani [1], Victor Guallar





JE: $e^{-\beta\Delta G_{AB}} = \left\langle e^{-\beta W}\right\rangle$

CGI: $\frac{P_f(W)}{P_b(-W)} = e^{\beta(W-\Delta G)}$

BAR: $\Delta G_{AB} = -\frac{1}{\beta} ln\frac{n_B}{n_A} + C$

Prediction Of The Impact Of Genetic Variability On Drug Sensitivity For Clinically Relevant EGFR Mutations

Aristarc Suriñach, Adam Hospital, Yvonne Westermaier, Luis Jordà, Sergi Orozco-Ruiz, Daniel Beltrán, Francesco Colizzi, Pau Andrio, Robert Soliva, Martí Municoy, Josep Ll. Gelpi, Modesto Orozco

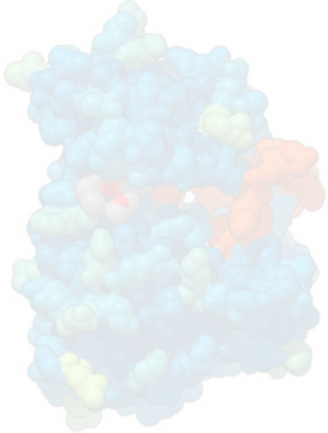
doi: https://doi.org/10.1101/2022.04.25.489389

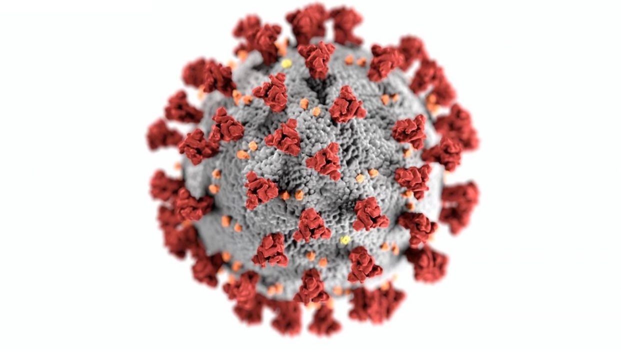
Could we apply the method to different systems?

Mutation	Drug	Eprof (pred)	PELE[*] (pred)	Exp. Impact[Δ]
L718Q	Osimertinib	R	-	Resistance[1]
G719S	Gefitinib	S	S	Sensitive[2]
G719S	Icotinib	S	-	Sensitive[2]
G719S	Erlotinib	S	S	Sensitive[3]
G719S	Lapatinib	S	S	Sensitive[4]
L747S	Gefitinib	*S*	*S*	Resistance[5]
L747F	Osimertinib	*S*	-	Resistance[6]
L747H	Osimertinib	*S*	-	Resistance[6]
S768I	Gefitinib	S	*R*	Sensitive[7]
V769M	Gefitinib	S	S	Sensitive[8]
T790M	Gefitinib	*S*	*S*	Resistance[9]
T790M	Erlotinib	*S*	R	Resistance[9]
T790M	Lapatinib	*S*	R	Resistance[10]
T790M	Osimertinib	S	-	Sensitive[11]
T790M	Icotinib	*S*	-	Resistance[12]
L792F	Osimertinib	*S*	-	Resistance[13]
L792H	Osimertinib	R	-	Resistance[13]
G796S	Osimertinib	*S*	-	Resistance[14]
C797G[&]	Osimertinib	R	R	Resistance[15]
C797S[&]	Osimertinib	R	R	Resistance[16]
L833V	Gefitinib	S	S	Sensitive[17]
H835L	Gefitinib	S	S	Sensitive[17]
L838V	Gefitinib	S	S	Sensitive[18]
T854A	Gefitinib	R	*S*	Resistance[5]
L861Q	Gefitinib	S	S	Sensitive[19]
T790M/C797S	Erlotinib	R	R	Resistance[20]

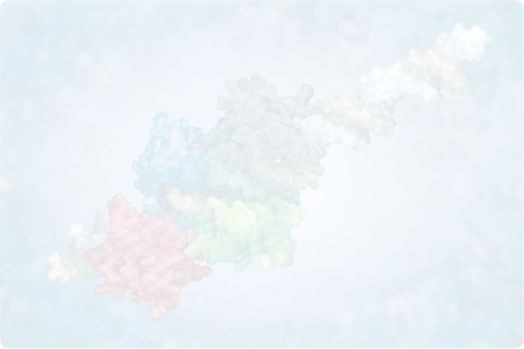
Projects

- High-throughput **prediction** of the **impact** of **genetic variability** on **drug sensitivity** and **resistance patterns** for clinically relevant **EGFR mutations** from atomistic simulations.





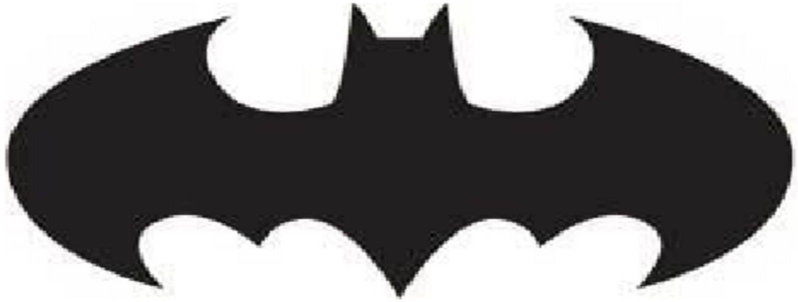
- Large-scale **SARS-CoV2 mutation** analysis, including a study on the **evolutionary path** and **host-selection mechanism** of **SARS-CoV-2**.



- **DNAffinity**: A **Machine-Learning** approach to **predict DNA Binding affinities** of **Transcription Factors**.

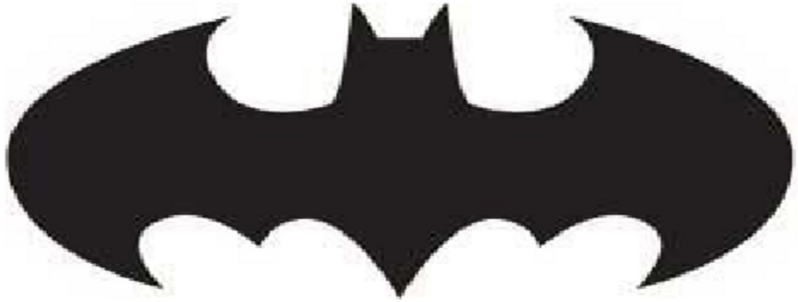
Overview:

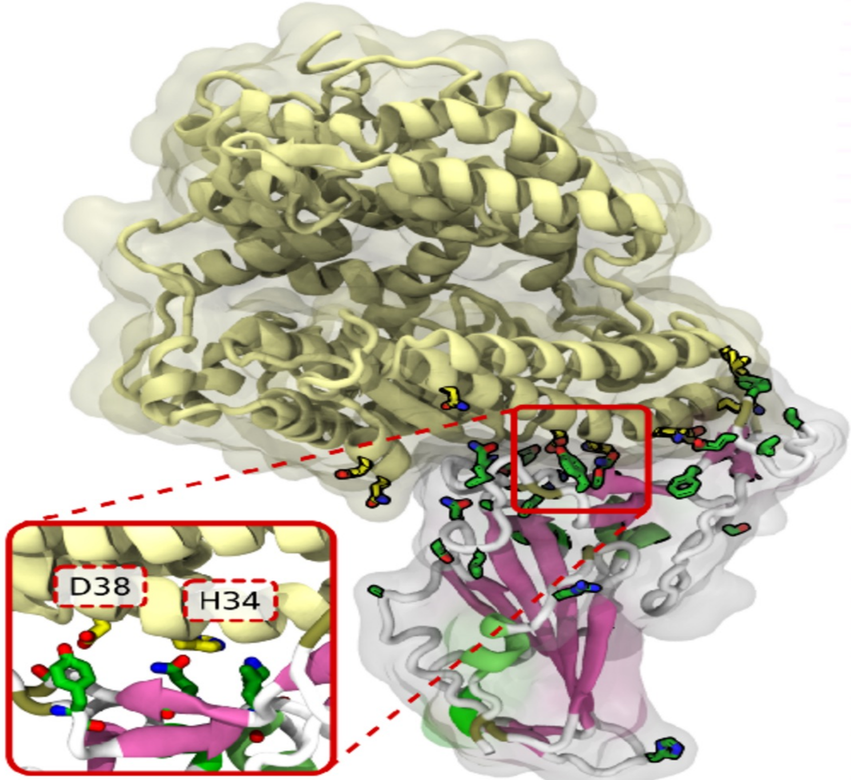
- The bat-to-human zoonotic transition
- "Humanized" bat polymorphism
- The "Spanish mutant" or A222V



Overview:

- The bat-to-human zoonotic transition
- "Humanized" bat polymorphism
- The "Spanish mutant" or A222V



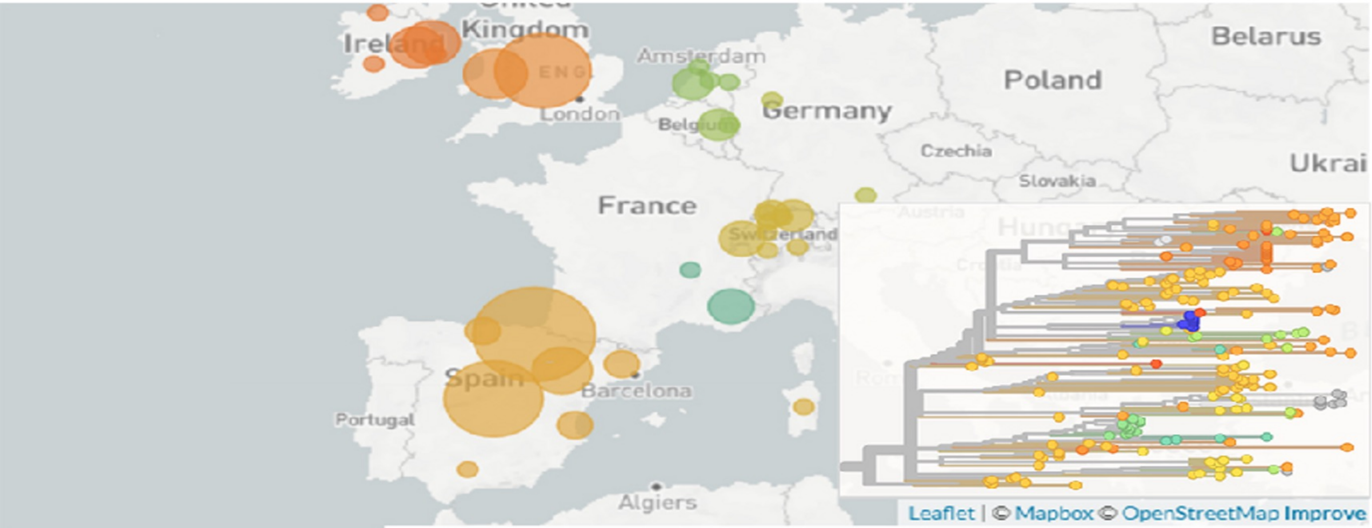


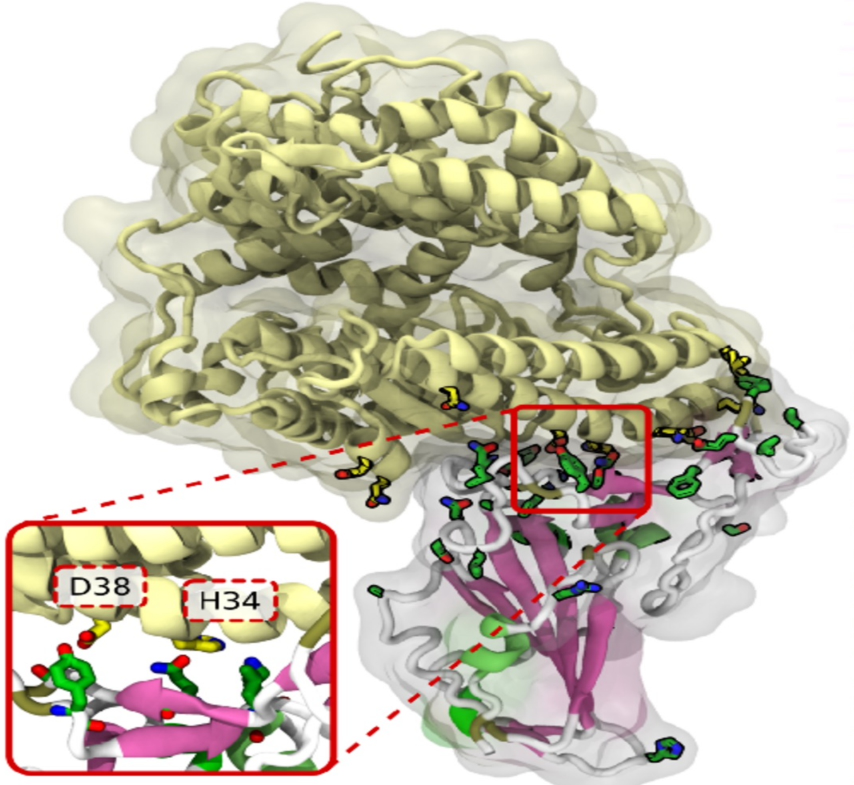
Position	affiACE2	hACE2
21	T	I
24	R	Q
27	I	T
31	N	K
34	R/H	H
38	E/D	D
49	E	N
82	N	M
325	E	Q
329	N	E

Position	RaTG13	SC2
346	T	R
372	T	A
403	T	R
439	K	N
440	H	N
441	I	L
443	A	S
445	E	V
449	F	Y
459	A	S
478	K	T
483	Q	V
484	T	E
486	L	F
490	Y	F
493	Y	Q
494	R	S
498	Y	Q
501	D	N
505	H	Y

Overview:

- The bat-to-human zoonotic transition
- "Humanized" bat polymorphism
- The "Spanish mutant" or A222V

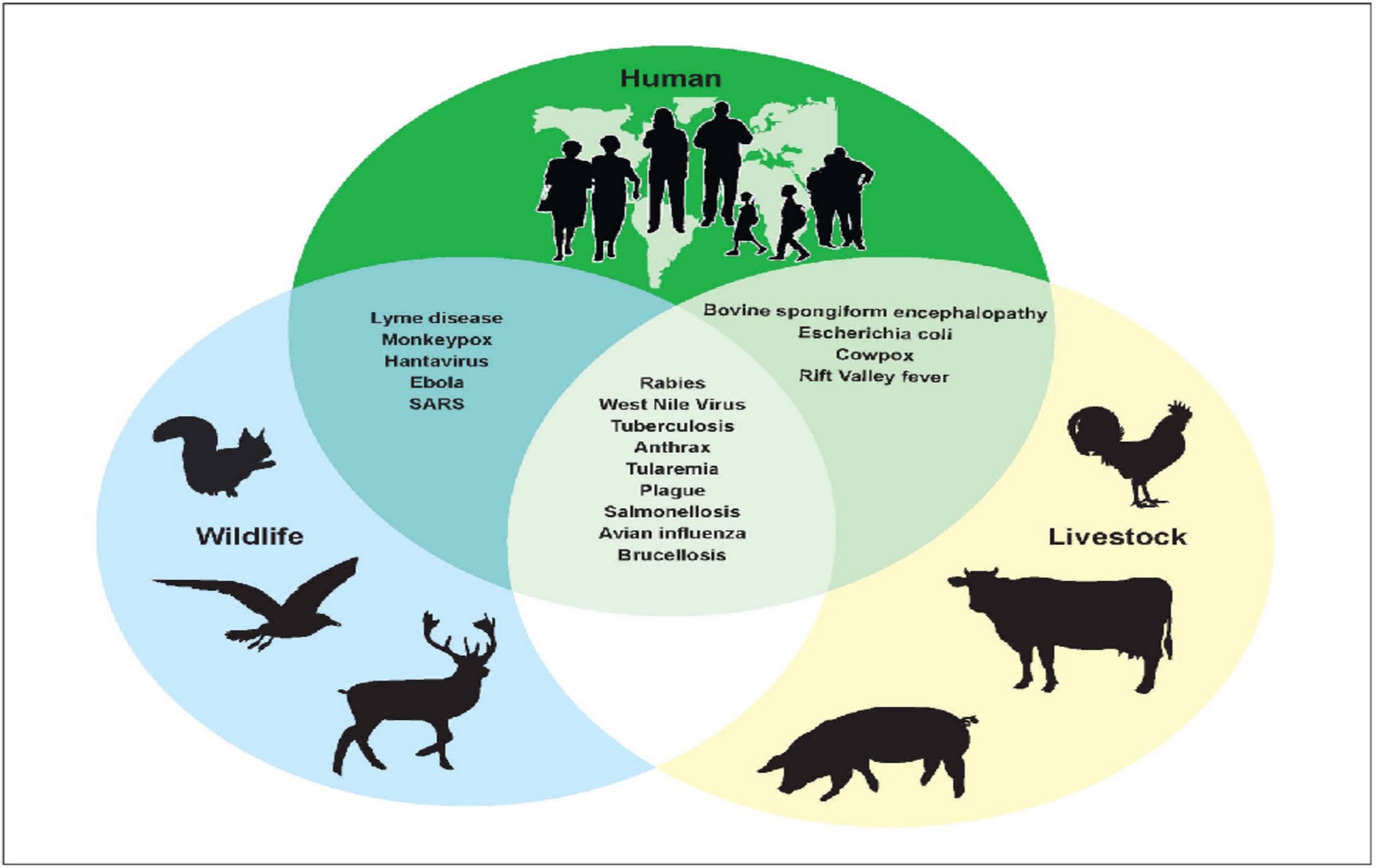




Position	affiACE2	hACE2
21	T	I
24	R	Q
27	I	T
31	N	K
34	R/H	H
38	E/D	D
49	E	N
82	N	M
325	E	Q
329	N	E

Position	RaTG13	SC2
346	T	R
372	T	A
403	T	R
439	K	N
440	H	N
441	I	L
443	A	S
445	E	V
449	F	Y
459	A	S
478	K	T
483	Q	V
484	T	E
486	L	F
490	Y	F
493	Y	Q
494	R	S
498	Y	Q
501	D	N
505	H	Y

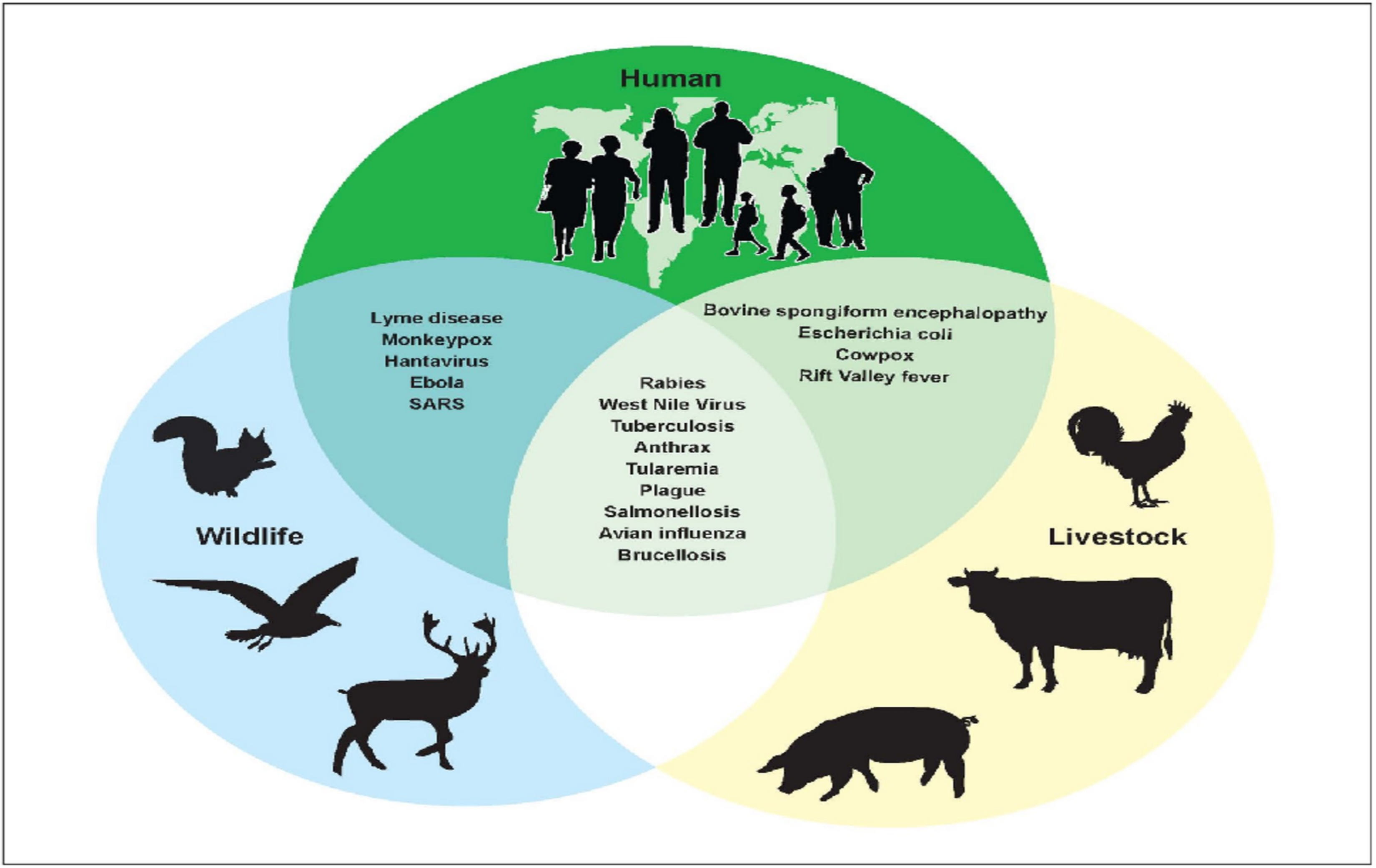
From bats to humans



Source: GAO analysis of USGS data (data); Art Explosion (images).

- Thanks to deforestation and agriculture, more and more pathogens cross the interspecies barrier

From bats to humans



Human

Lyme disease
Monkeypox
Hantavirus
Ebola
SARS

Bovine spongiform encephalopathy
Escherichia coli
Cowpox
Rift Valley fever

Rabies
West Nile Virus
Tuberculosis
Anthrax
Tularemia
Plague
Salmonellosis
Avian influenza
Brucellosis

Wildlife

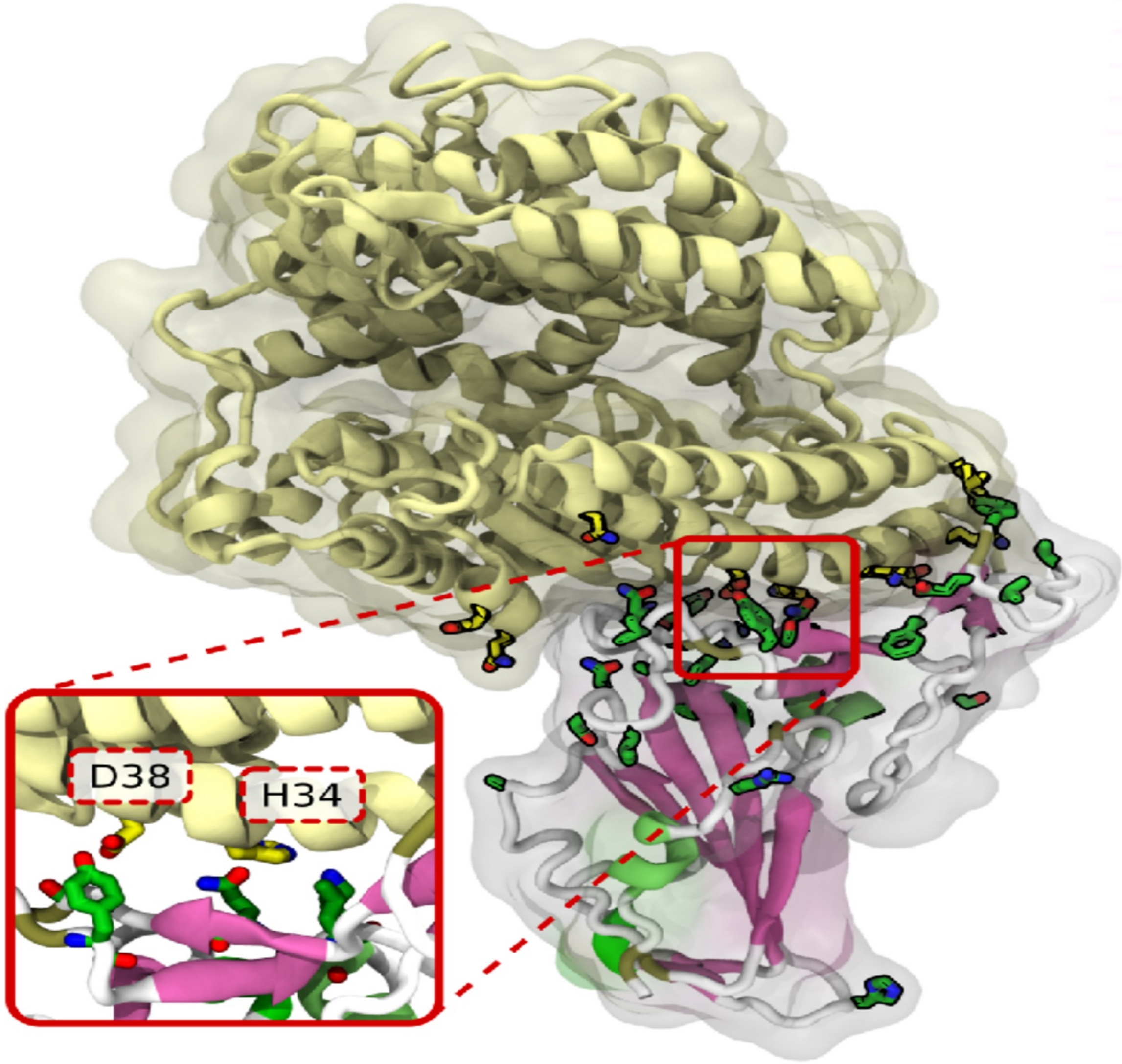
Livestock

Source: GAO analysis of USGS data (data); Art Explosion (images).

- Thanks to deforestation and agriculture, more and more pathogens cross the interspecies barrier
- For SARS-CoV-2, the closest known relative was RaTG13, a virus isolated from *Rhinolophus affinis* in 2013 (a new one found recently!)

From bats to humans

- The receptor-binding domains (RBDs) of both viruses differ by 21 amino acids

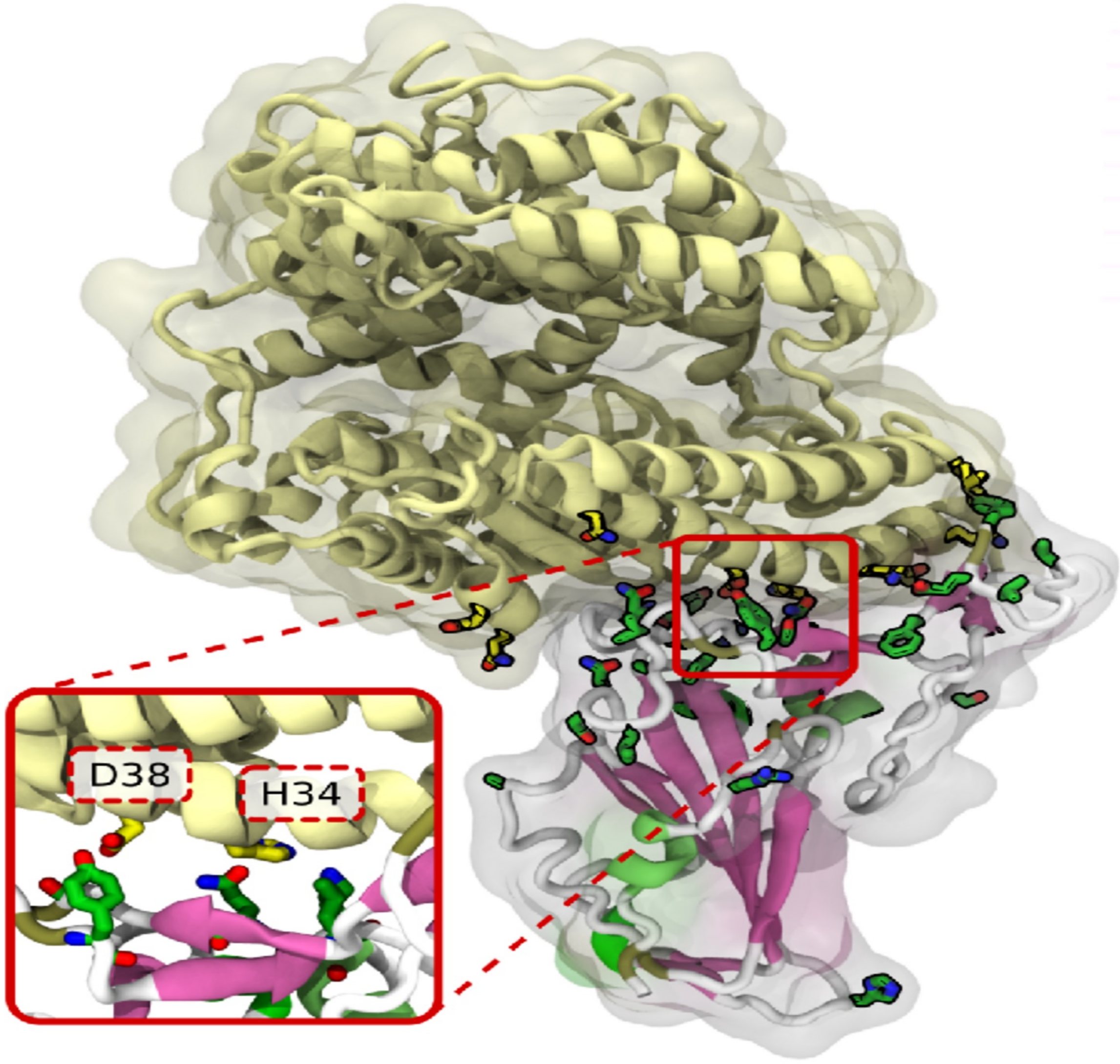


Position	affiACE2	hACE2
21	T	I
24	R	Q
27	I	T
31	N	K
34	R/H	H
38	E/D	D
49	E	N
82	N	M
325	E	Q
329	N	E

Position	RaTG13	SC2
346	T	R
372	T	A
403	T	R
439	K	N
440	H	N
441	I	L
443	A	S
445	E	V
449	F	Y
459	A	S
478	K	T
483	Q	V
484	T	E
486	L	F
490	Y	F
493	Y	Q
494	R	S
498	Y	Q
501	D	N
505	H	Y

From bats to humans

- The receptor-binding domains (RBDs) of both viruses differ by 21 amino acids
- Challenge: identify the most important mutations that enabled infecting a new host

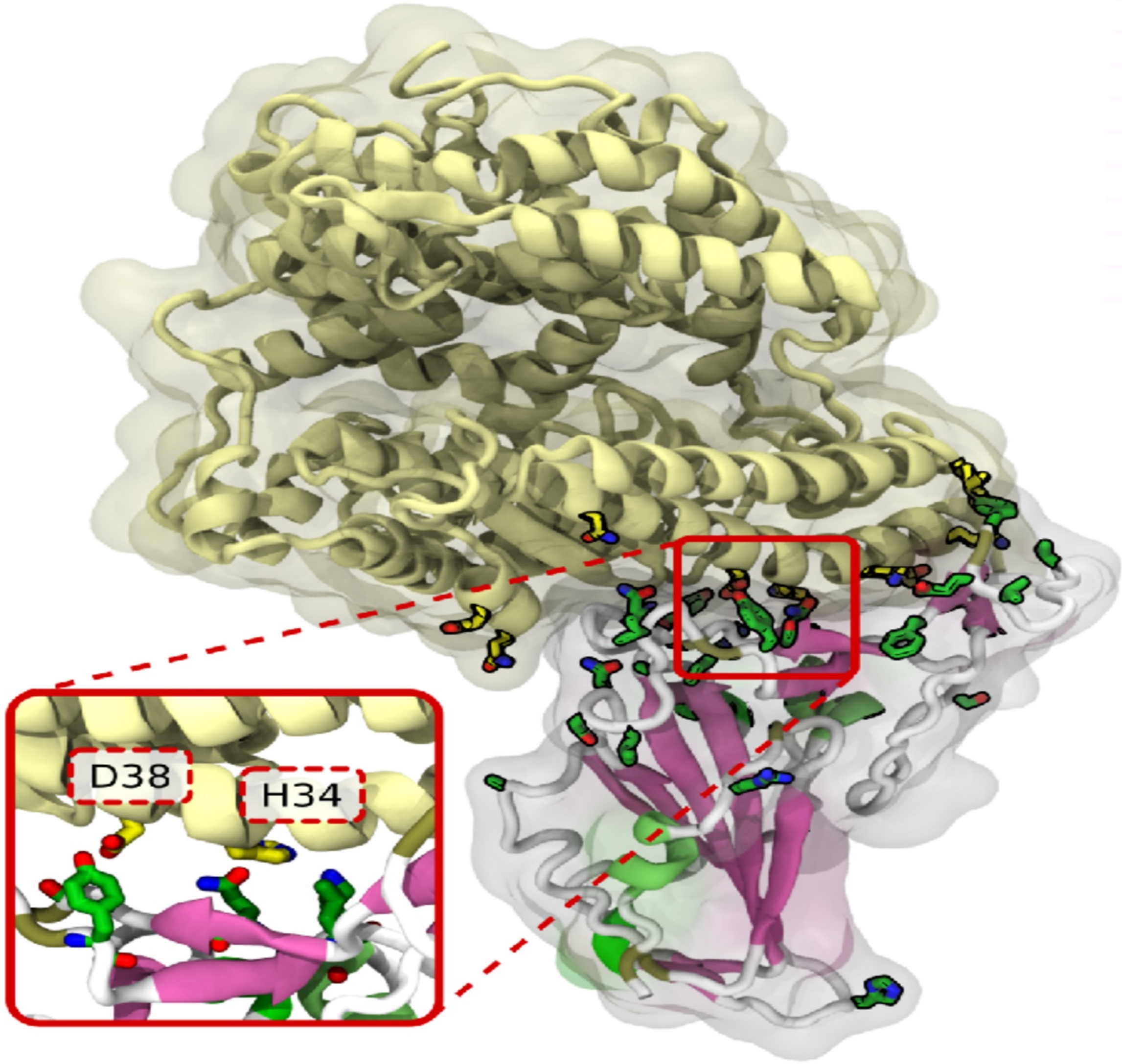


Position	affiACE2	hACE2
21	T	I
24	R	Q
27	I	T
31	N	K
34	R/H	H
38	E/D	D
49	E	N
82	N	M
325	E	Q
329	N	E

Position	RaTG13	SC2
346	T	R
372	T	A
403	T	R
439	K	N
440	H	N
441	I	L
443	A	S
445	E	V
449	F	Y
459	A	S
478	K	T
483	Q	V
484	T	E
486	L	F
490	Y	F
493	Y	Q
494	R	S
498	Y	Q
501	D	N
505	H	Y

From bats to humans

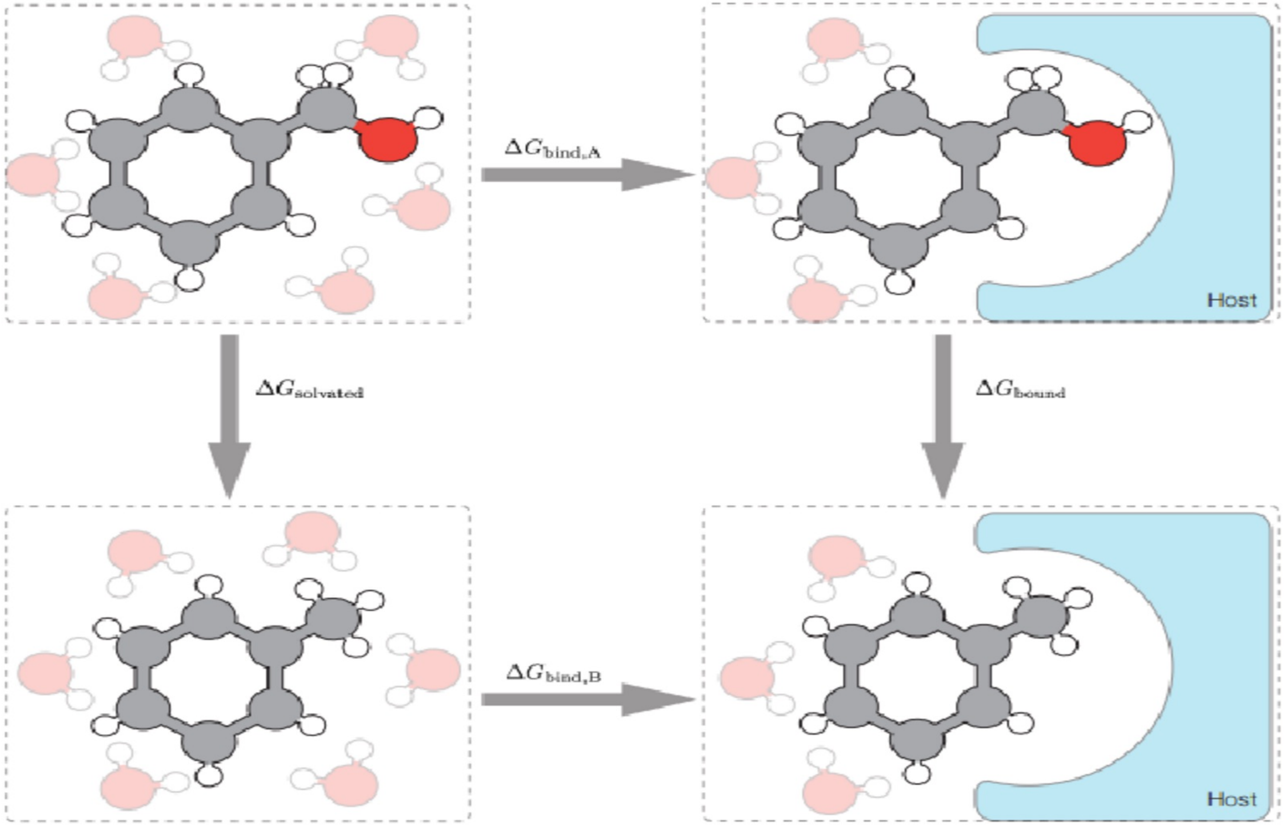
- The receptor-binding domains (RBDs) of both viruses differ by 21 amino acids
- Challenge: identify the most important mutations that enabled infecting a new host
- Constraint: hACE2 shows experimentally a preference for SARS-CoV-2 of ca. 3 kcal/mol



Position	affiACE2	hACE2
21	T	I
24	R	Q
27	I	T
31	N	K
34	R/H	H
38	E/D	D
49	E	N
82	N	M
325	E	Q
329	N	E

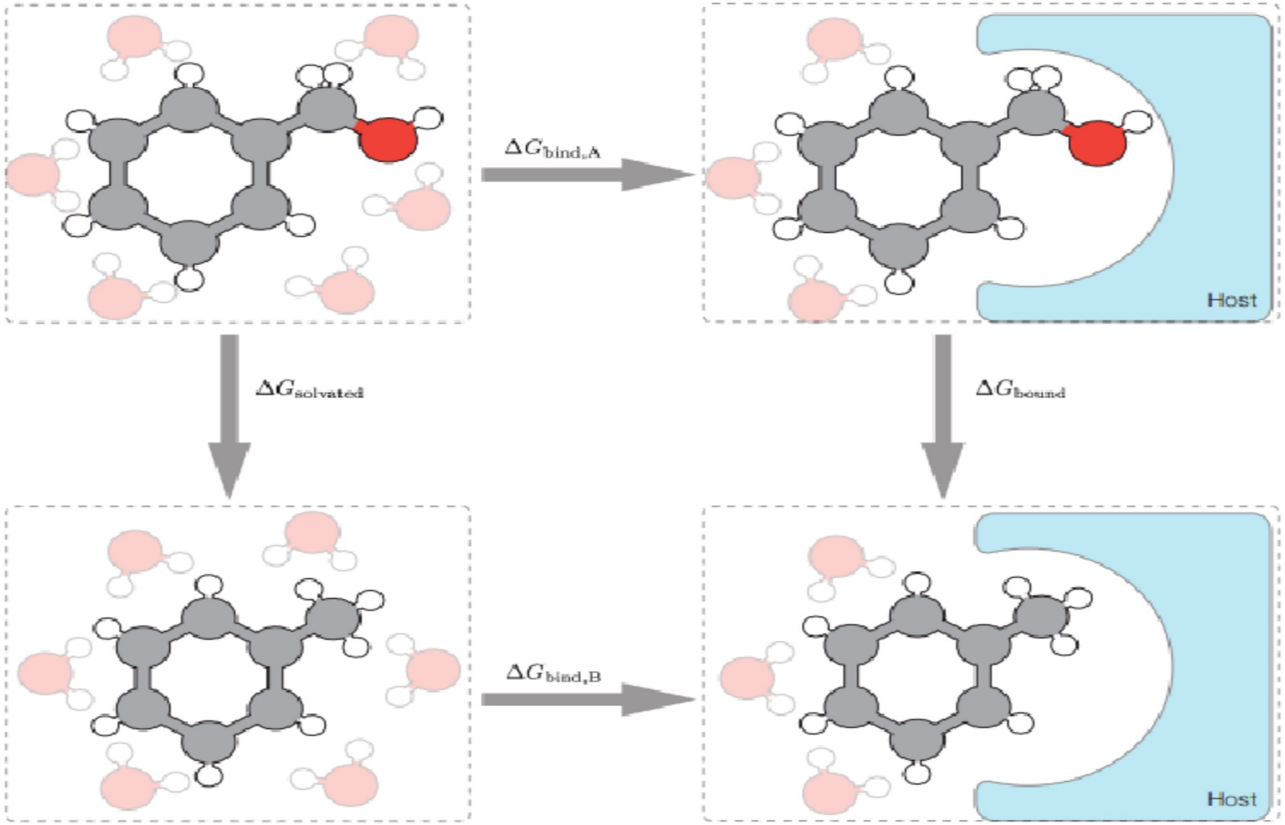
Position	RaTG13	SC2
346	T	R
372	T	A
403	T	R
439	K	N
440	H	N
441	I	L
443	A	S
445	E	V
449	F	Y
459	A	S
478	K	T
483	Q	V
484	T	E
486	L	F
490	Y	F
493	Y	Q
494	R	S
498	Y	Q
501	D	N
505	H	Y

Mutations through alchemistry

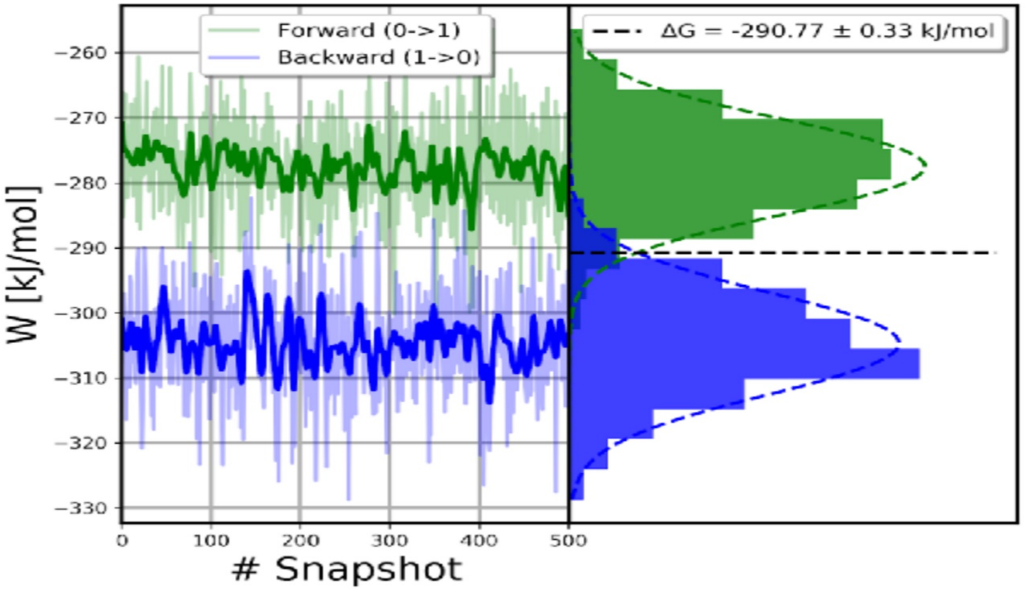

$\Delta G_{\text{bind,A}}$
$\Delta G_{\text{solvated}}$
ΔG_{bound}
$\Delta G_{\text{bind,B}}$
Host
Host

Principle of alchemical simulations: calculate the chemical change (vertical) to obtain the difference in binding energies (horizontal)

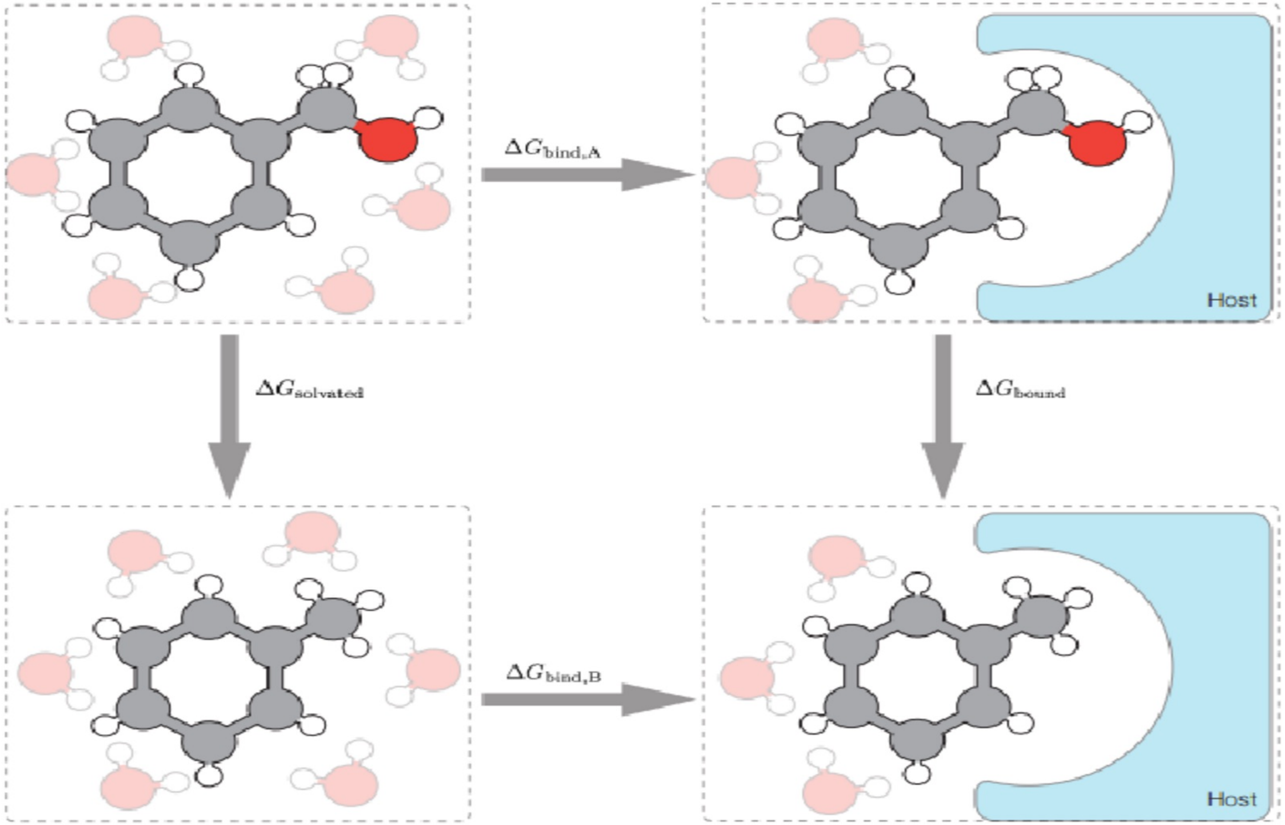
Mutations through alchemistry


$\Delta G_{\text{bind,A}}$
Host
$\Delta G_{\text{solvated}}$
ΔG_{bound}
$\Delta G_{\text{bind,B}}$
Host

Principle of alchemical simulations: calculate the chemical change (vertical) to obtain the difference in binding energies (horizontal)

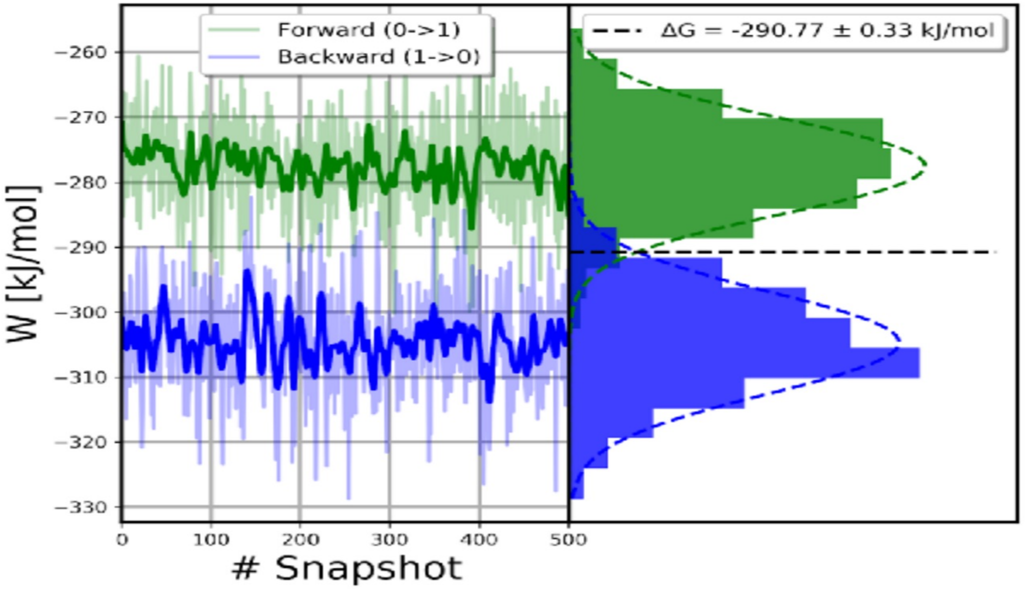


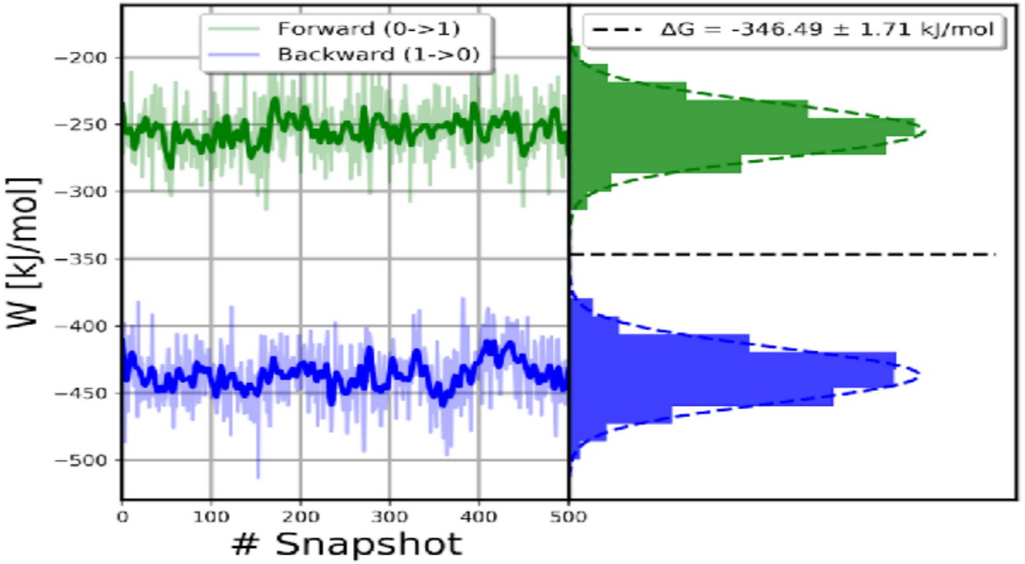
Mutations through alchemistry



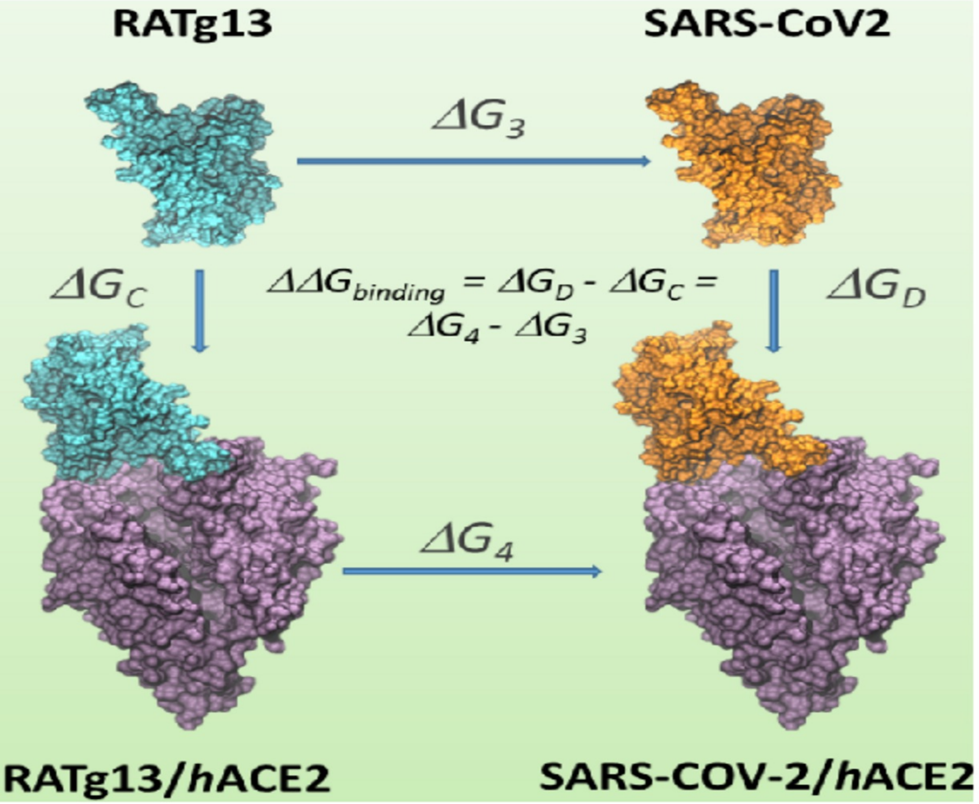
$\Delta G_{\mathrm{bind,A}}$
Host
$\Delta G_{\mathrm{solvated}}$
$\Delta G_{\mathrm{bound}}$
$\Delta G_{\mathrm{bind,B}}$
Host

Principle of alchemical simulations: calculate the chemical change (vertical) to obtain the difference in binding energies (horizontal)



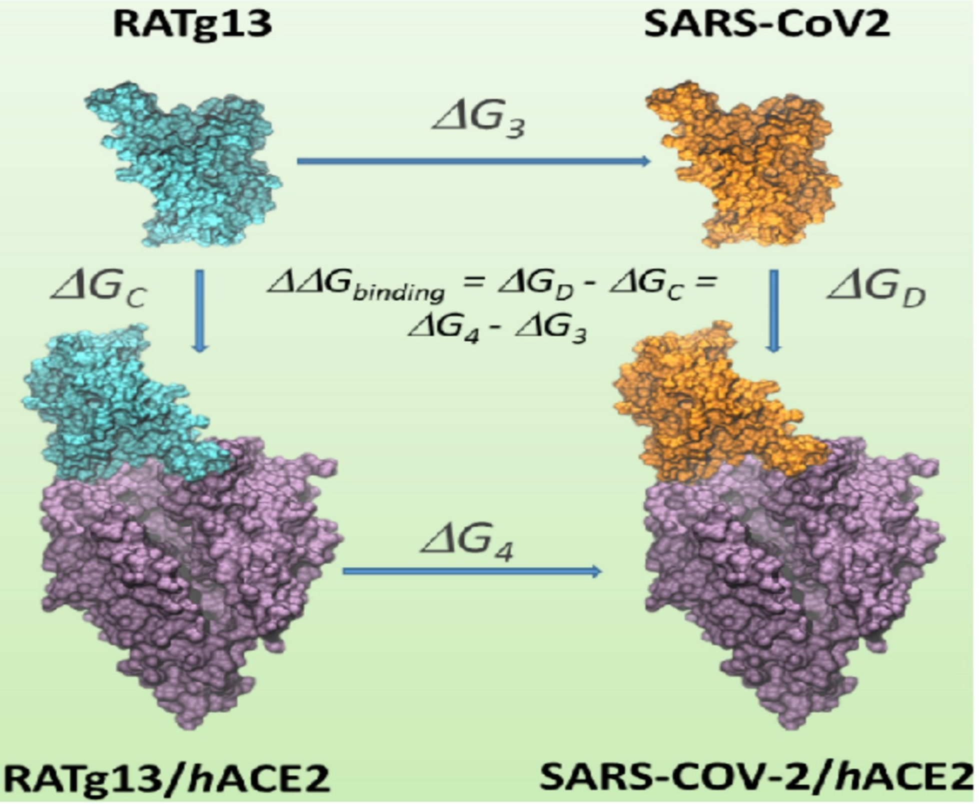


Mutations through alchemistry

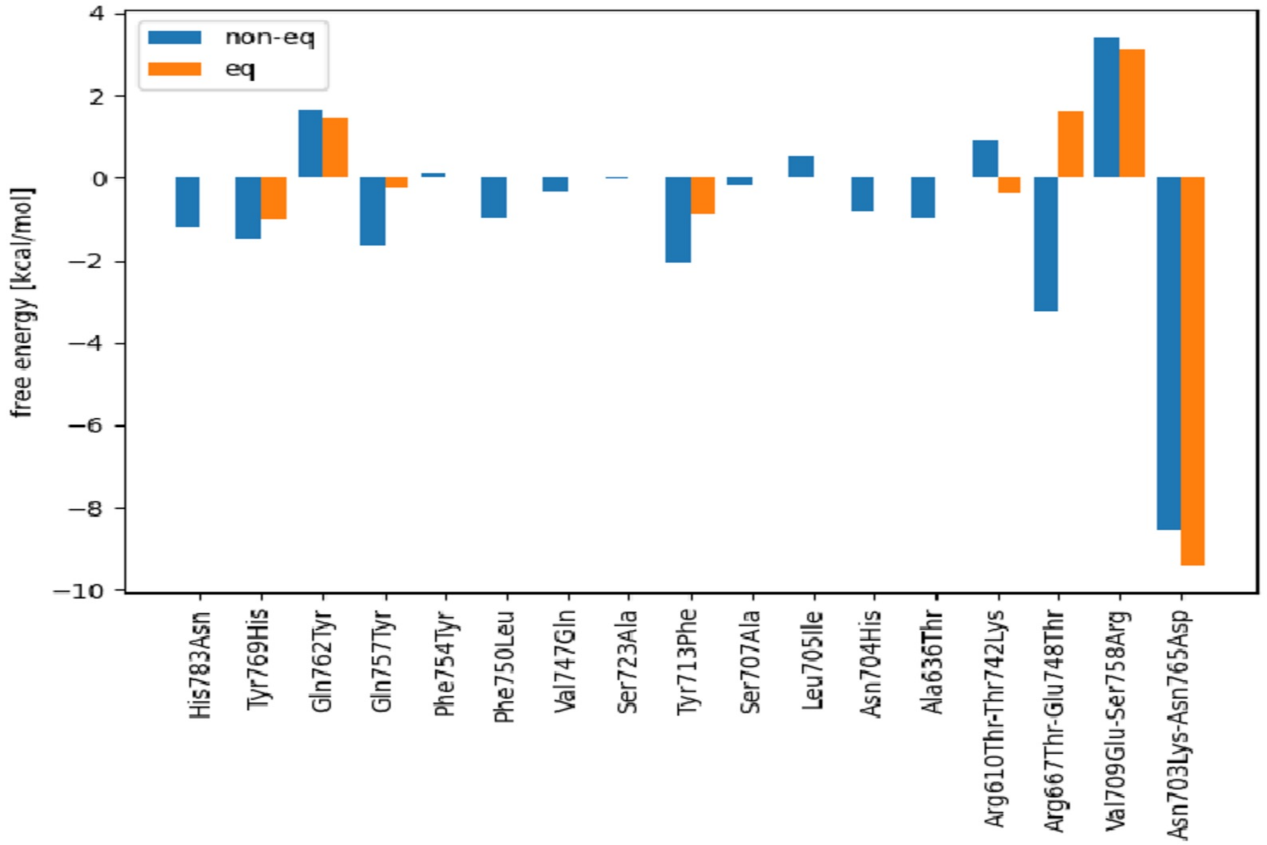


- Strategy: start with "cheap" non-eq, validate selected with expensive equilibrium protocol if numbers don't match up

Mutations through alchemistry

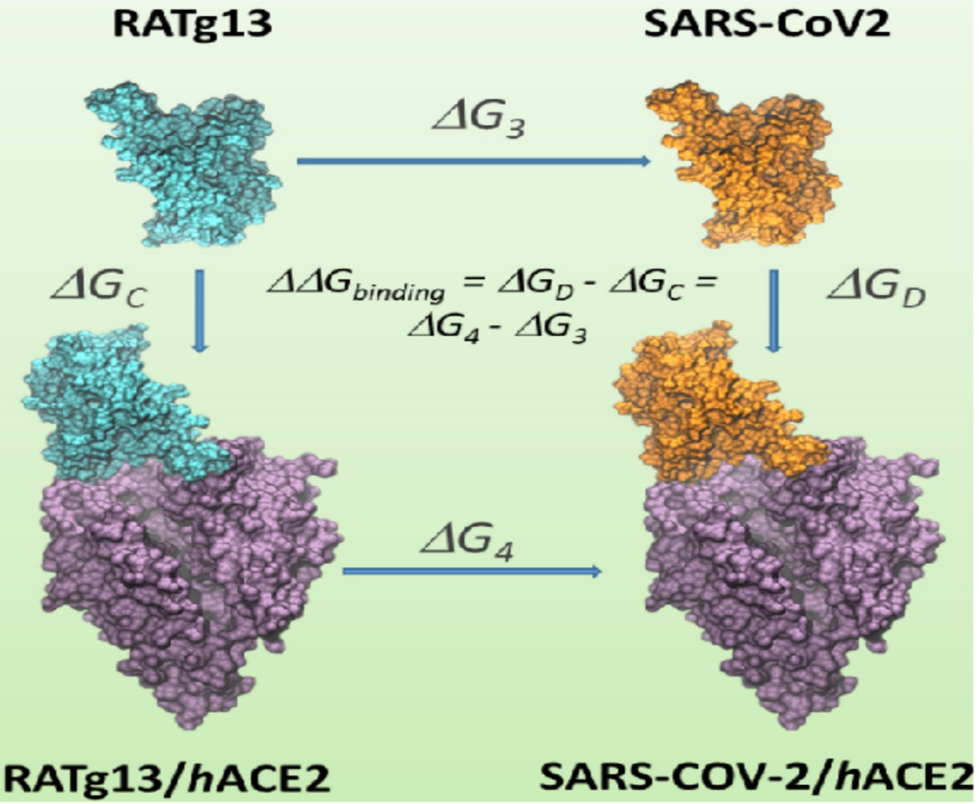


- All contributions sum up to -9.5 kcal/mol (expt ca. -3.0) - failure?



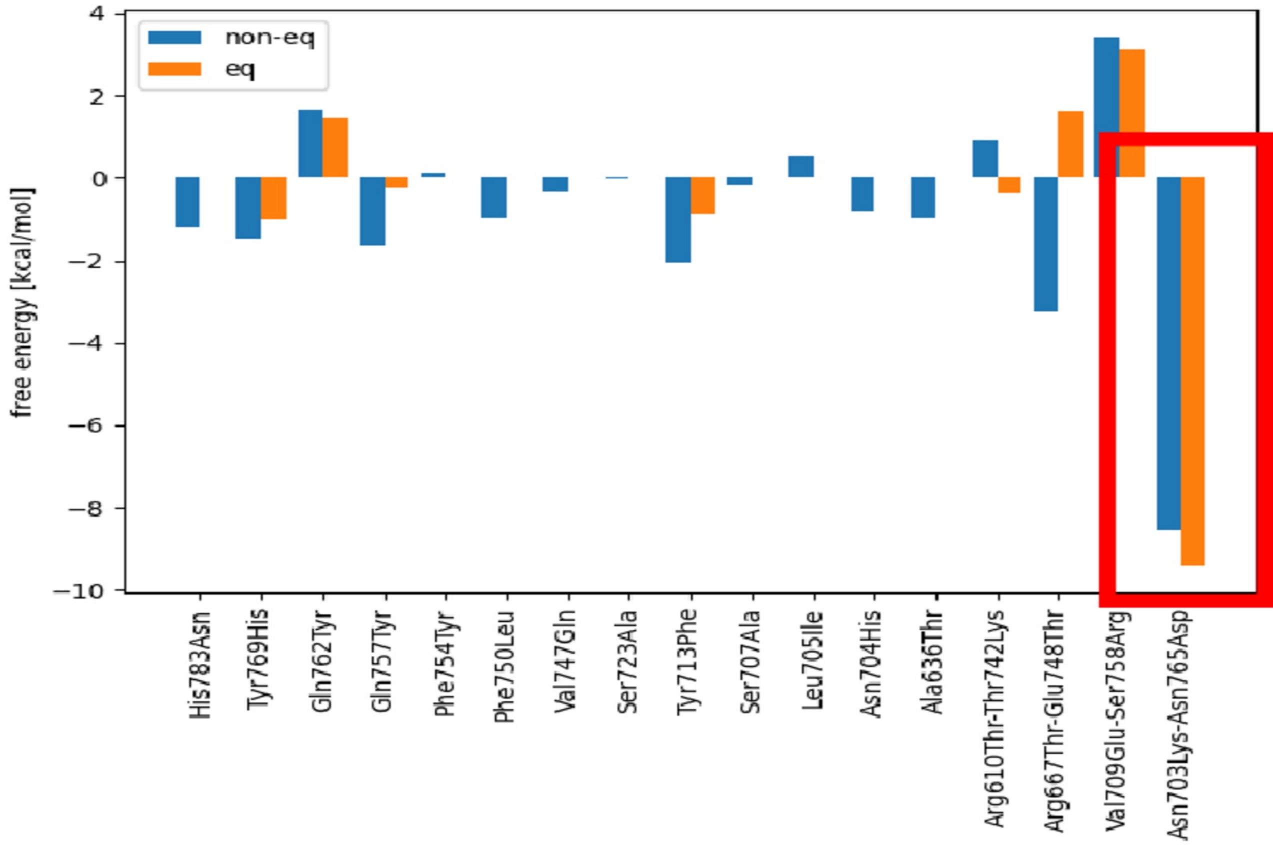
- Strategy: start with "cheap" non-eq, validate selected with expensive equilibrium protocol if numbers don't match up

Mutations through alchemistry



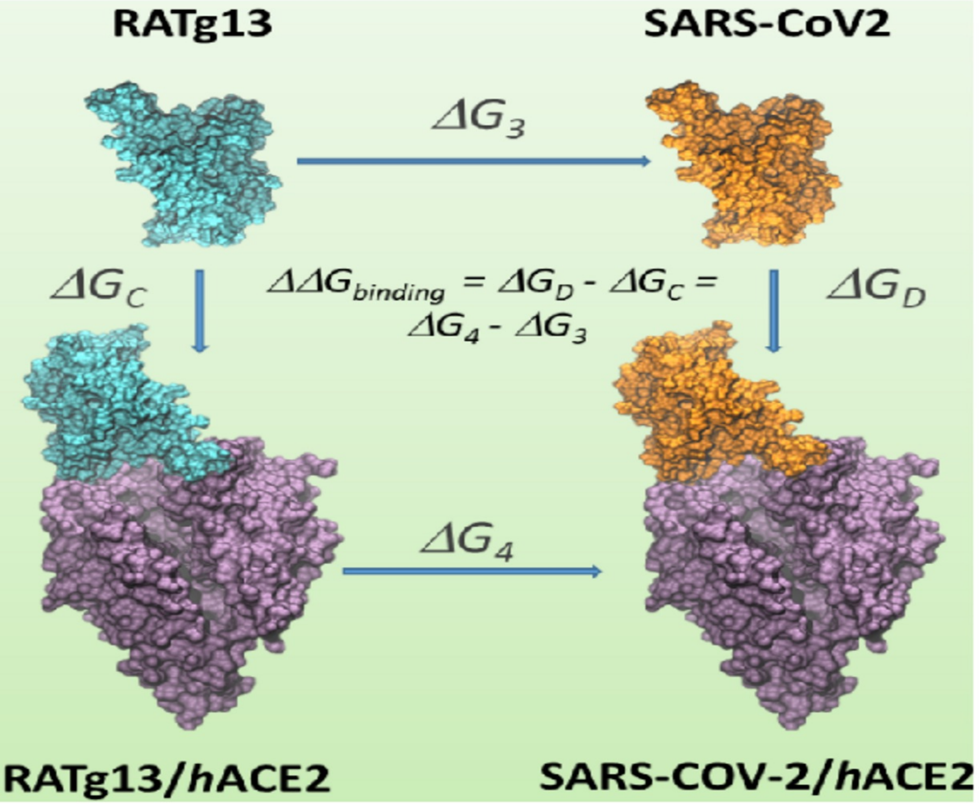
- Strategy: start with "cheap" non-eq, validate selected with expensive equilibrium protocol if numbers don't match up

- All contributions sum up to -9.5 kcal/mol (expt ca. -3.0) - failure?
- BUT the whole error in one mutant

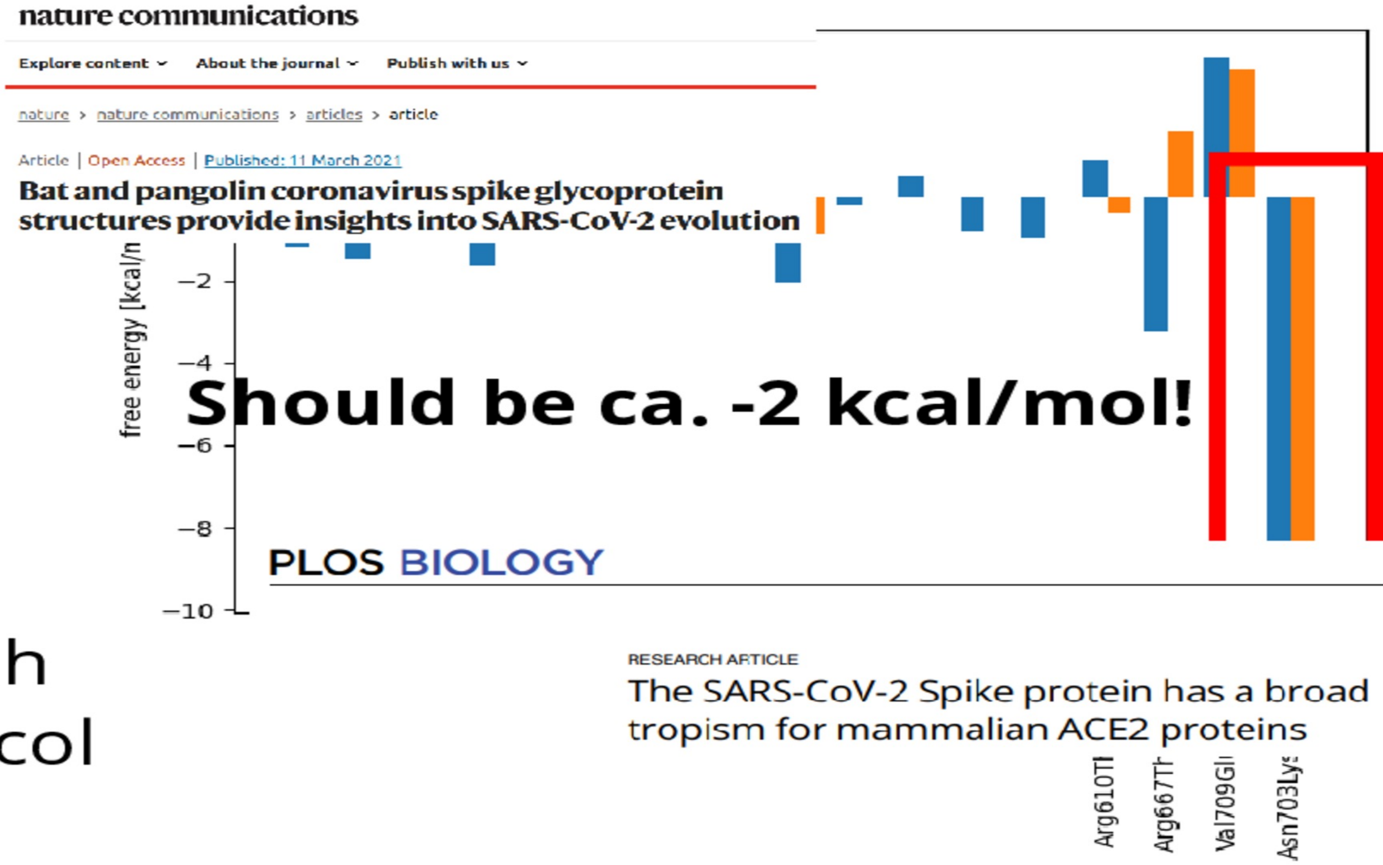


Mutations through alchemistry

- All contributions sum up to -9.5 kcal/mol (expt ca. -3.0) - failure?
- BUT the whole error in one mutant



- Strategy: start with "cheap" non-eq, validate selected with expensive equilibrium protocol if numbers don't match up


nature communications

Explore content ˅ About the journal ˅ Publish with us ˅

nature > nature communications > articles > article

Article | Open Access | Published: 11 March 2021

Bat and pangolin coronavirus spike glycoprotein structures provide insights into SARS-CoV-2 evolution

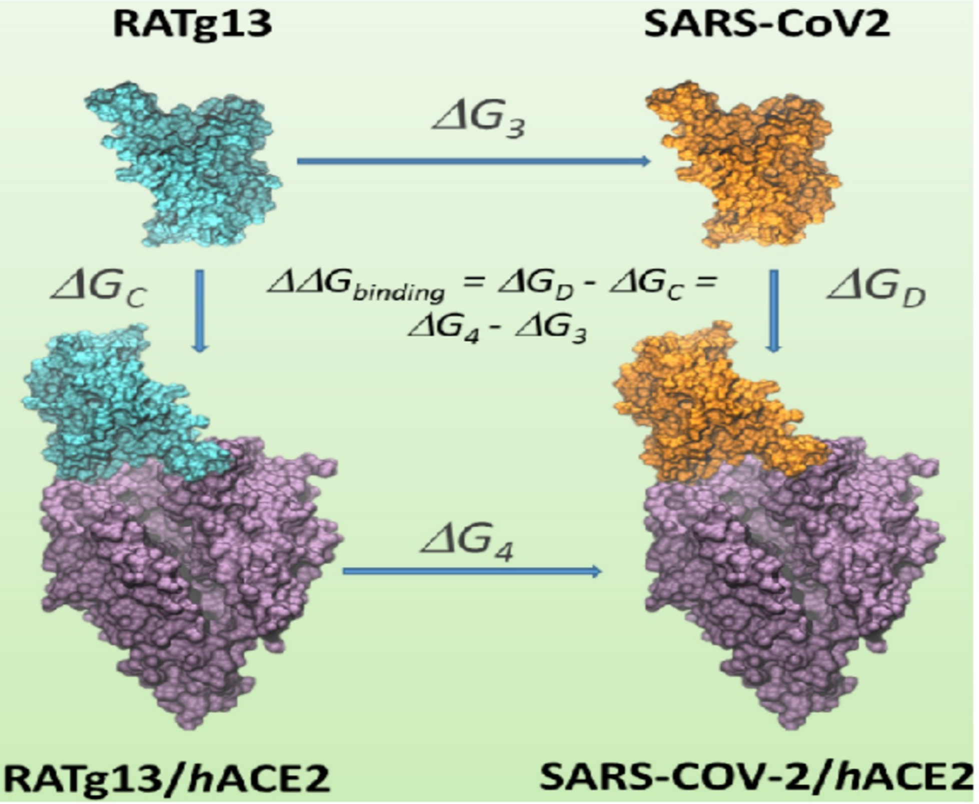
Should be ca. -2 kcal/mol!

PLOS BIOLOGY

RESEARCH ARTICLE

The SARS-CoV-2 Spike protein has a broad tropism for mammalian ACE2 proteins

Mutations through alchemistry



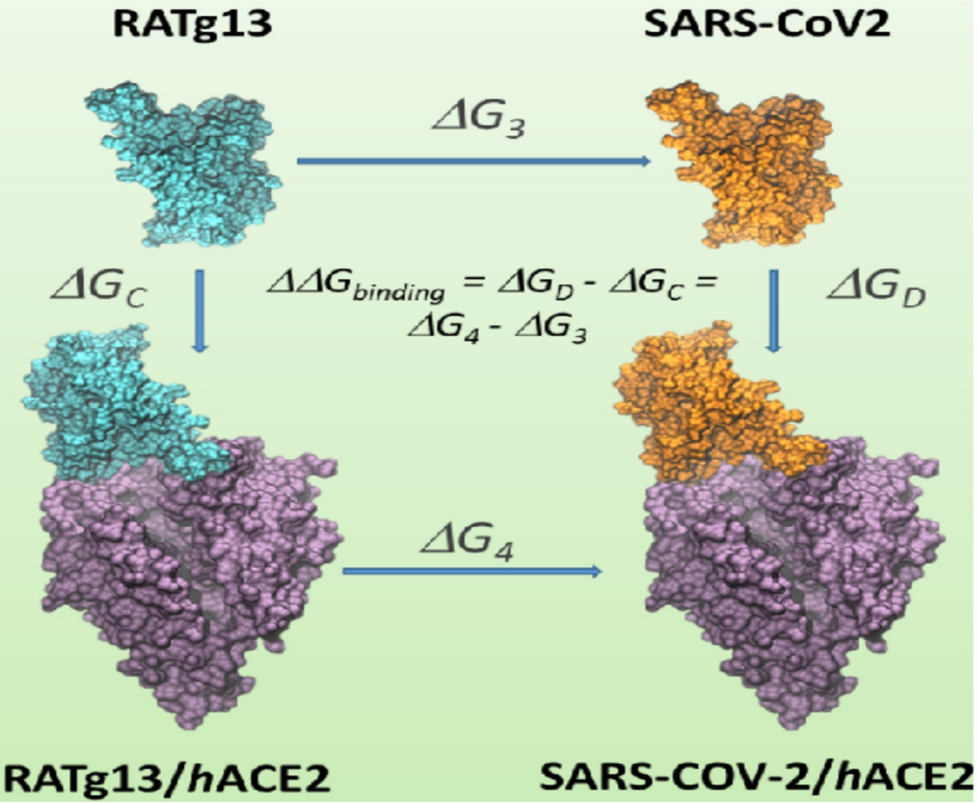
- All contributions sum up to -9.5 kcal/mol (expt ca. -3.0) - failure?
- BUT the whole error in one mutant

So... back to the conceptual side:

```
... propka says:

   ASP 706 B      4.68         3.80
   ASP 731 B      3.50         3.80
   ASP 765 B      7.53         3.80

...
```

- Strategy: start with "cheap" non-eq, validate selected with expensive equilibrium protocol if numbers don't match up

Mutations through alchemistry

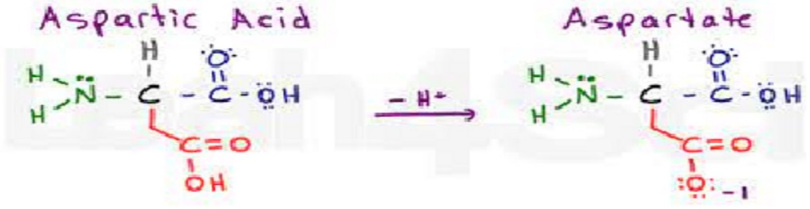


- Strategy: start with "cheap" non-eq, validate selected with expensive equilibrium protocol if numbers don't match up

- All contributions sum up to -9.5 kcal/mol (expt ca. -3.0) - failure?
- BUT the whole error in one mutant

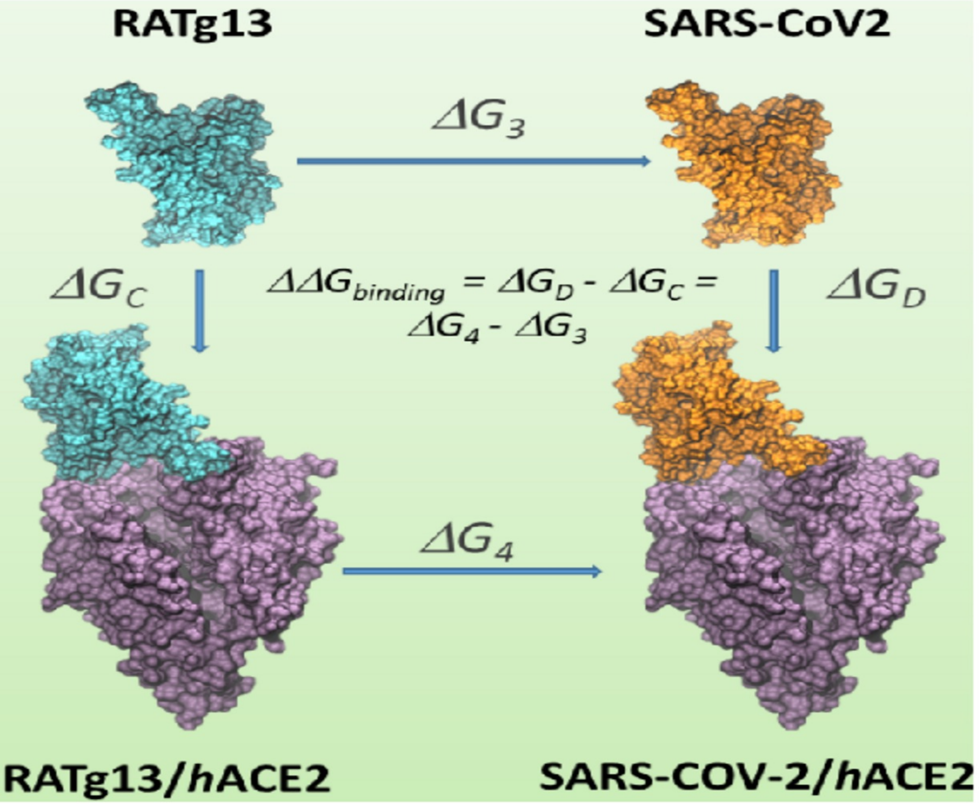
So... back to the conceptual side:

```
... propka says:

   ASP 706 B      4.68         3.80
   ASP 731 B      3.50         3.80
   ASP 765 B      7.53         3.80

...
```



What if we have been simulating the wrong protonation state?

Mutations through alchemistry



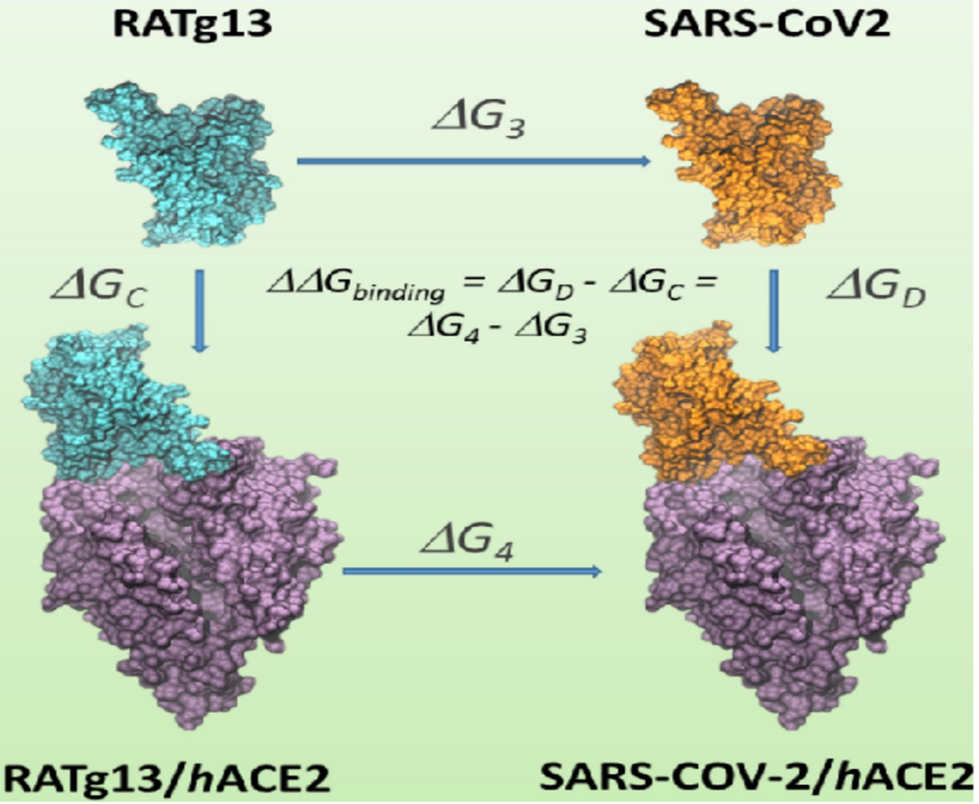
- All contributions sum up to -9.5 kcal/mol (expt ca. **-3.0**) - failure?
- BUT the whole error in one mutant

Corrected:

- +1.5 kcal/mol Asn > Arg
- -4.9 kcal/mol Asn > AspH
- **-3.6 kcal/mol** entire dataset

- Strategy: start with "cheap" non-eq, validate selected with expensive equilibrium protocol if numbers don't match up

Mutations through alchemistry



- All contributions sum up to -9.5 kcal/mol (expt ca. **-3.0**) - failure?
- BUT the whole error in one mutant

Corrected:

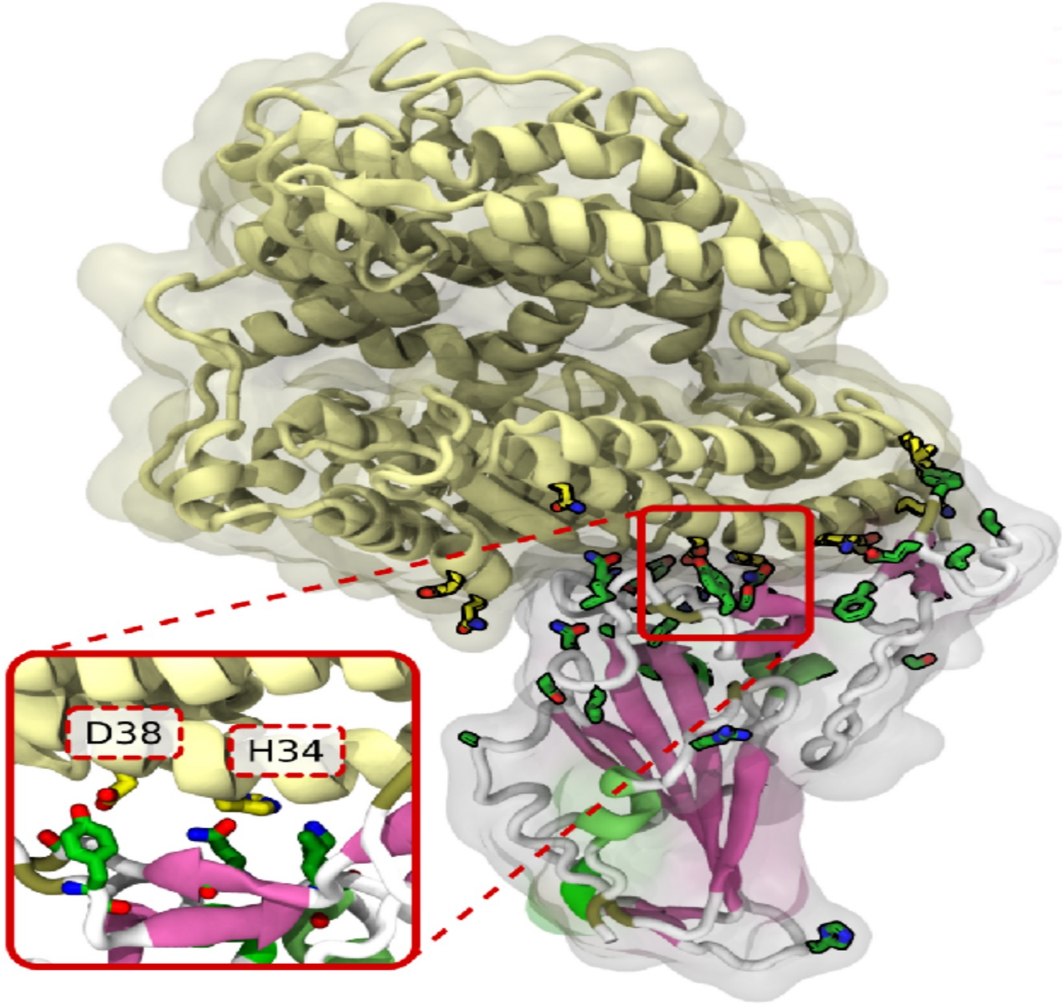
- +1.5 kcal/mol Asn > Arg
- -4.9 kcal/mol Asn > AspH
- **-3.6 kcal/mol** entire dataset

- Strategy: start with "cheap" non-eq, validate selected with expensive equilibrium protocol if numbers don't match up

☑ We're done here!

From bats to humans

- The receptor-binding domains (RBDs) of both viruses differ by 21 amino acids
- Challenge: identify the most important mutations that enabled infecting a new host
- Curiosity: there is a subspecies of *R. affinis* that is closer to humans by 2 residues of the receptor

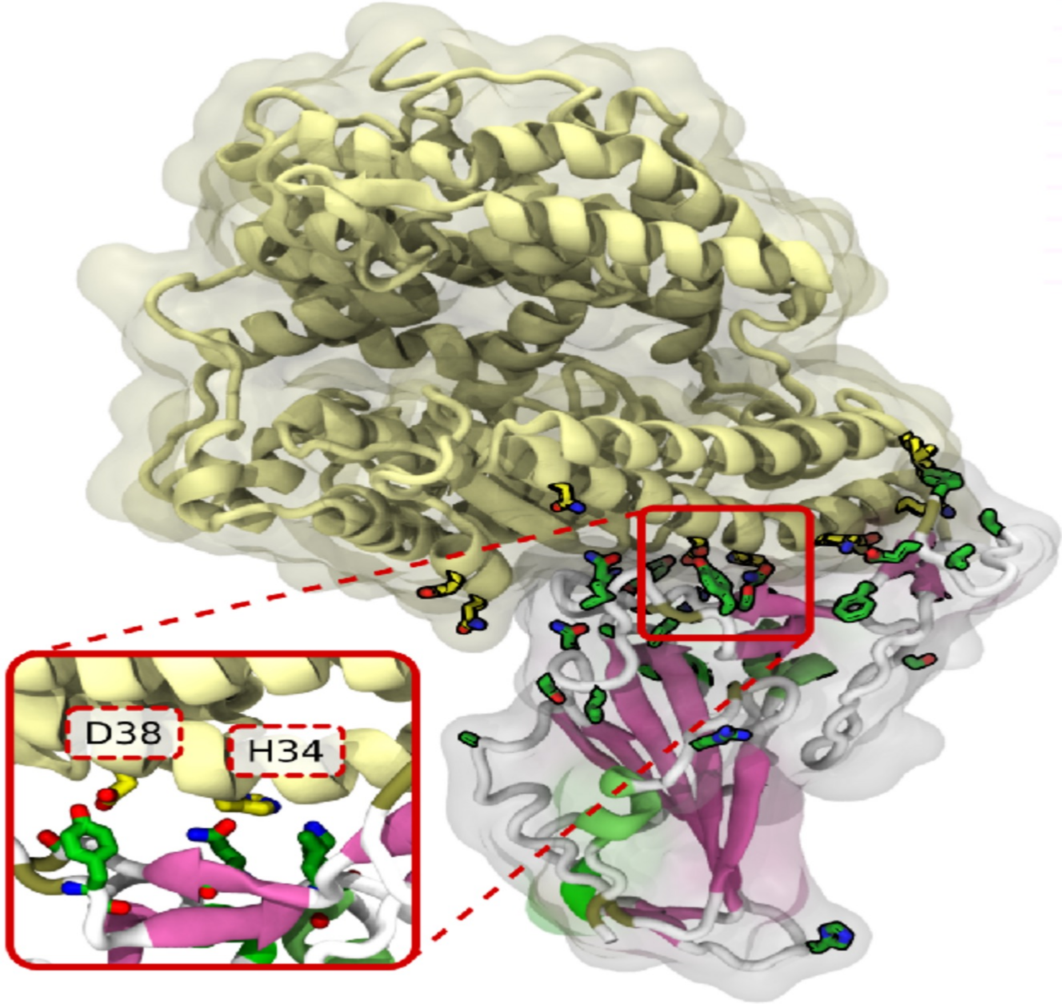


Position	affiACE2	hACE2
21	T	I
24	R	Q
27	I	T
31	N	K
34	R/H	H
38	E/D	D
49	E	N
82	N	M
325	E	Q
329	N	E

Position	RaTG13	SC2
346	T	R
372	T	A
403	T	R
439	K	N
440	H	N
441	I	L
443	A	S
445	E	V
449	F	Y
459	A	S
478	K	T
483	Q	V
484	T	E
486	L	F
490	Y	F
493	Y	Q
494	R	S
498	Y	Q
501	D	N
505	H	Y

Curiosity: bat polymorphism

- The double mutant (RE/HD) lowers the affinity of the bat virus by 1.4 kcal/mol

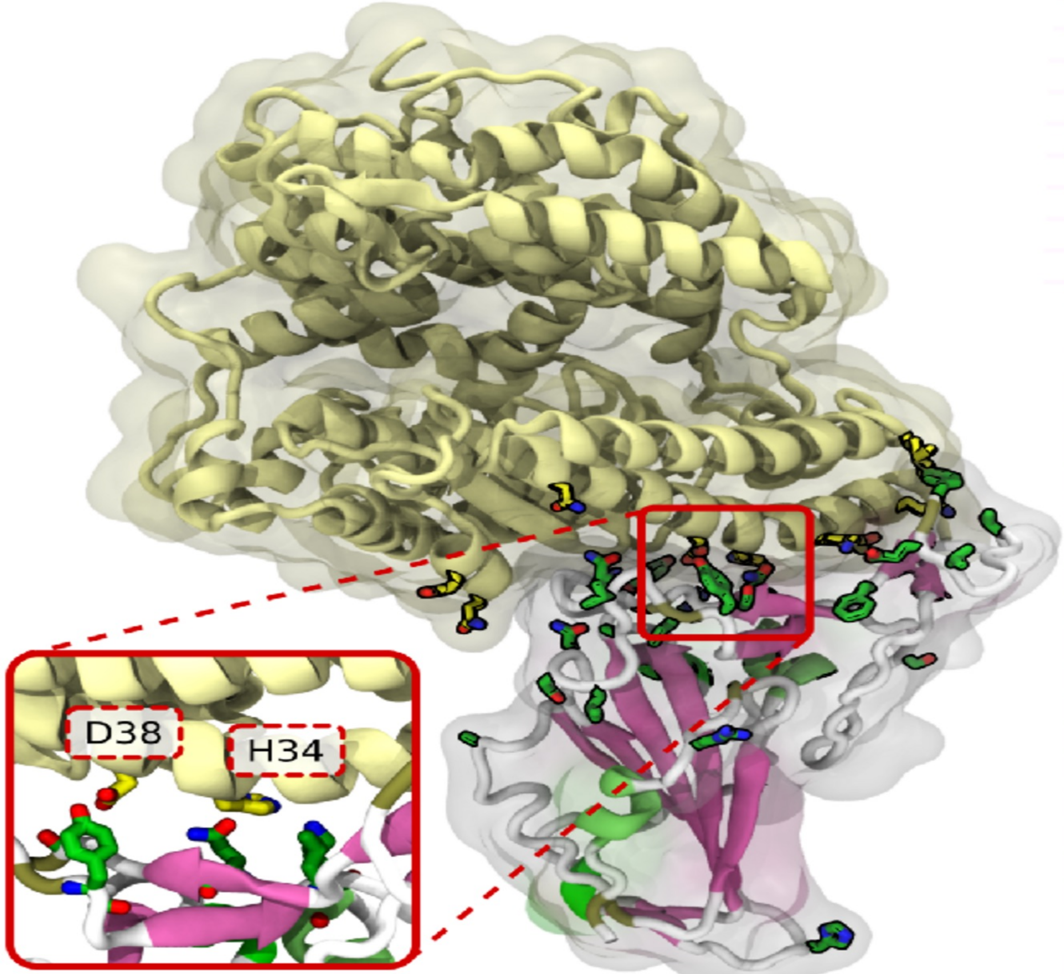


Position	affiACE2	hACE2
21	T	I
24	R	Q
27	I	T
31	N	K
34	R/H	H
38	E/D	D
49	E	N
82	N	M
325	E	Q
329	N	E

Position	RaTG13	SC2
346	T	R
372	T	A
403	T	R
439	K	N
440	H	N
441	I	L
443	A	S
445	E	V
449	F	Y
459	A	S
478	K	T
483	Q	V
484	T	E
486	L	F
490	Y	F
493	Y	Q
494	R	S
498	Y	Q
501	D	N
505	H	Y

Curiosity: bat polymorphism

- The double mutant (RE/HD) lowers the affinity of the bat virus by 1.4 kcal/mol
- In turn, the human virus (SARS-CoV-2) prefers the HD pair by 0.7 kcal/mol

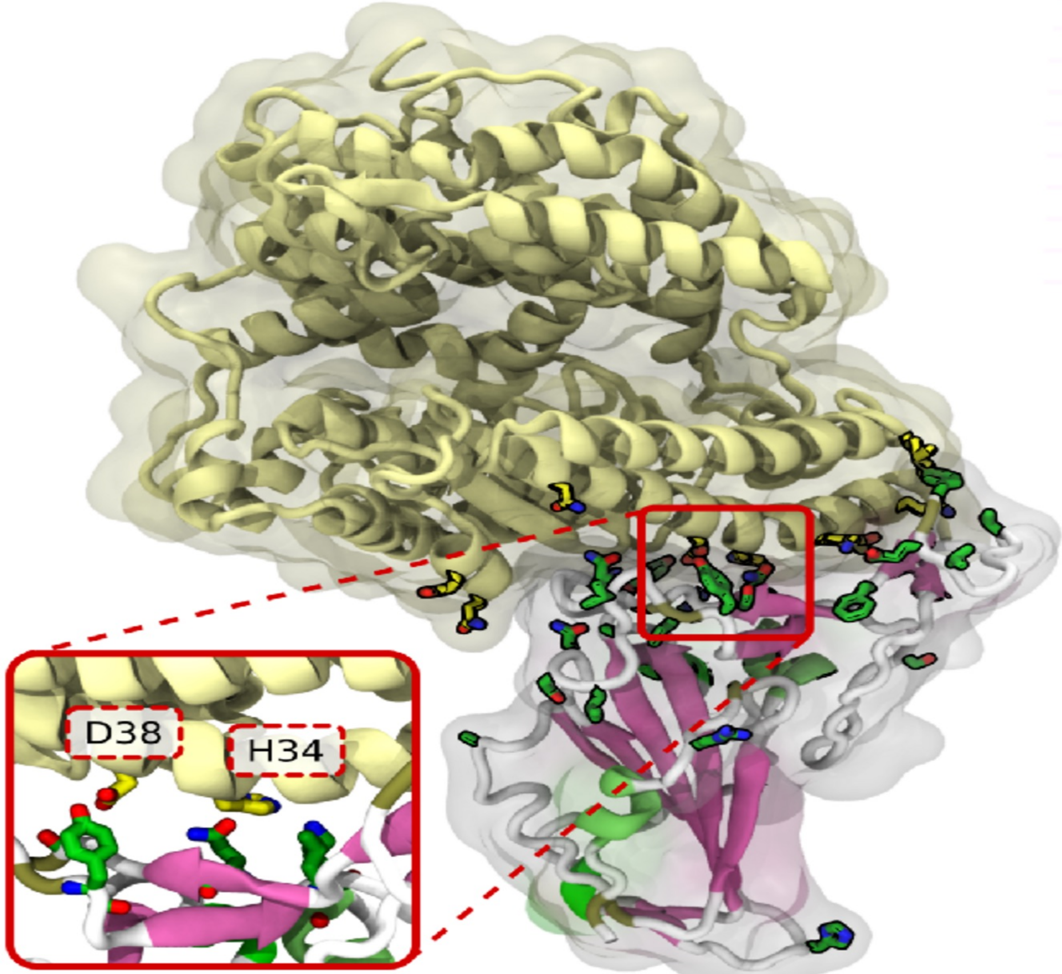


Position	affiACE2	hACE2
21	T	I
24	R	Q
27	I	T
31	N	K
34	R/H	H
38	E/D	D
49	E	N
82	N	M
325	E	Q
329	N	E

Position	RaTG13	SC2
346	T	R
372	T	A
403	T	R
439	K	N
440	H	N
441	I	L
443	A	S
445	E	V
449	F	Y
459	A	S
478	K	T
483	Q	V
484	T	E
486	L	F
490	Y	F
493	Y	Q
494	R	S
498	Y	Q
501	D	N
505	H	Y

Curiosity: bat polymorphism

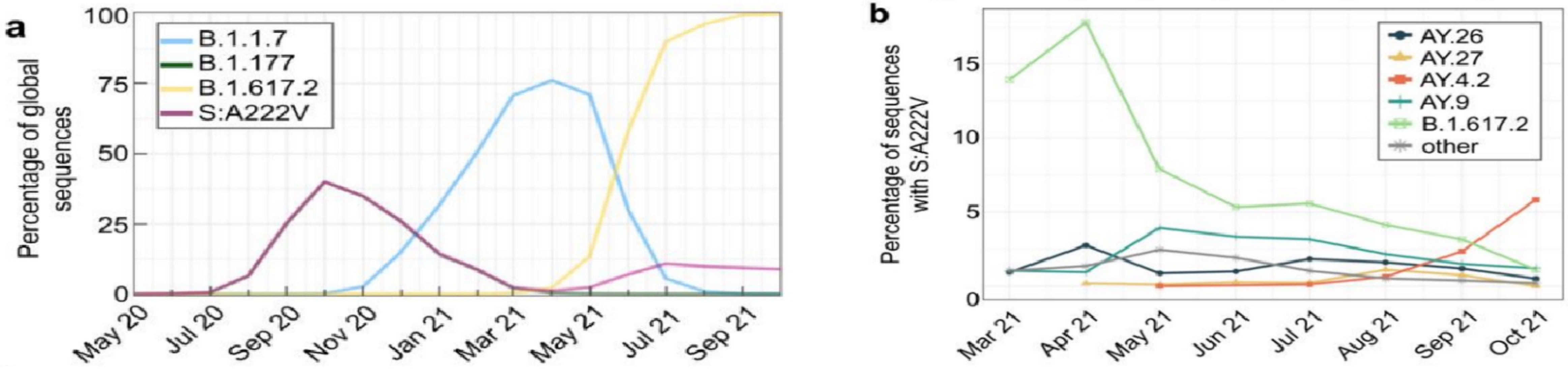
- The double mutant (RE/HD) lowers the affinity of the bat virus by 1.4 kcal/mol
- In turn, the human virus (SARS-CoV-2) prefers the HD pair by 0.7 kcal/mol
- Possible evolutionary driving force for optimization of the local interface?
- Speculative but not improbable (hopefully!)



Position	affiACE2	hACE2
21	T	I
24	R	Q
27	I	T
31	N	K
34	R/H	H
38	E/D	D
49	E	N
82	N	M
325	E	Q
329	N	E

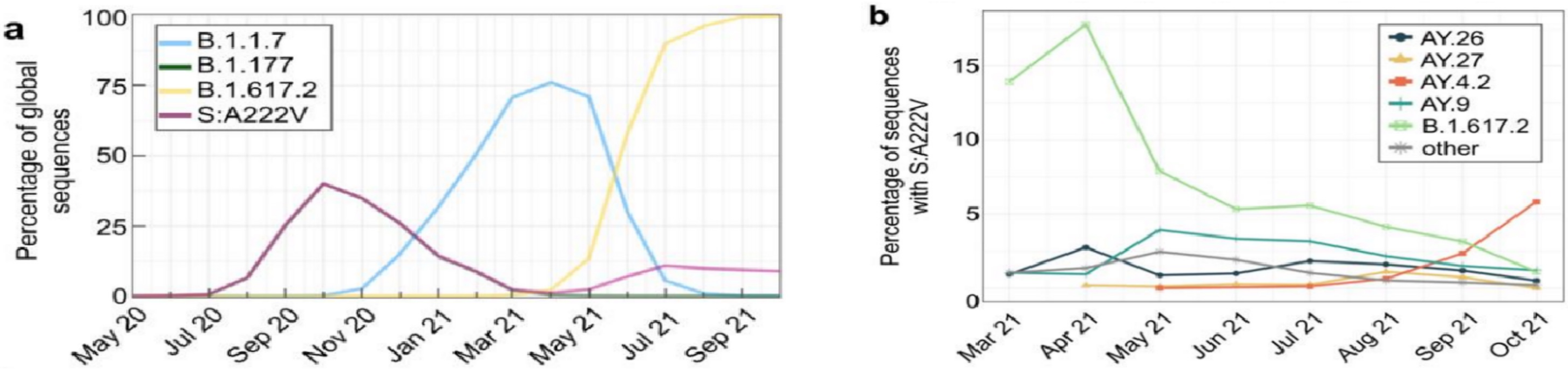
Position	RaTG13	SC2
346	T	R
372	T	A
403	T	R
439	K	N
440	H	N
441	I	L
443	A	S
445	E	V
449	F	Y
459	A	S
478	K	T
483	Q	V
484	T	E
486	L	F
490	Y	F
493	Y	Q
494	R	S
498	Y	Q
501	D	N
505	H	Y

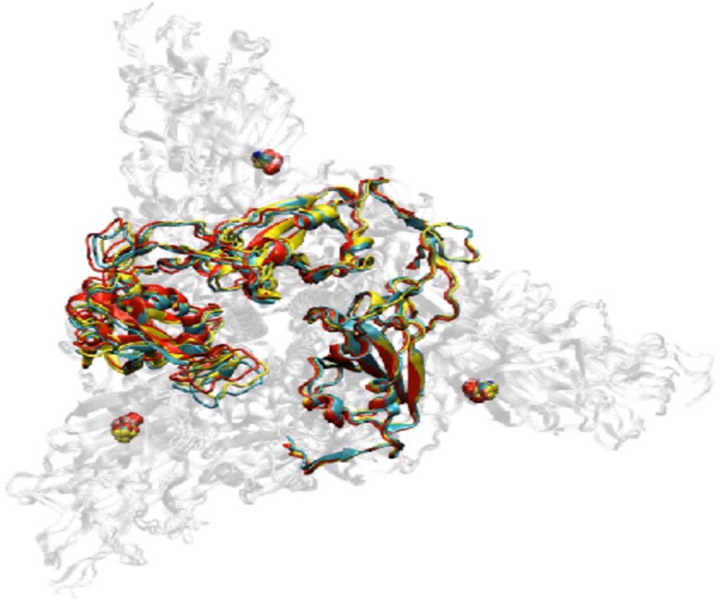
A222V: the "Spanish" mutant



- First appeared in Spain in summer 2020
- Reappeared in "Delta+" (AY.4.2) in late summer 2021, suggesting an advantage

A222V: the "Spanish" mutant





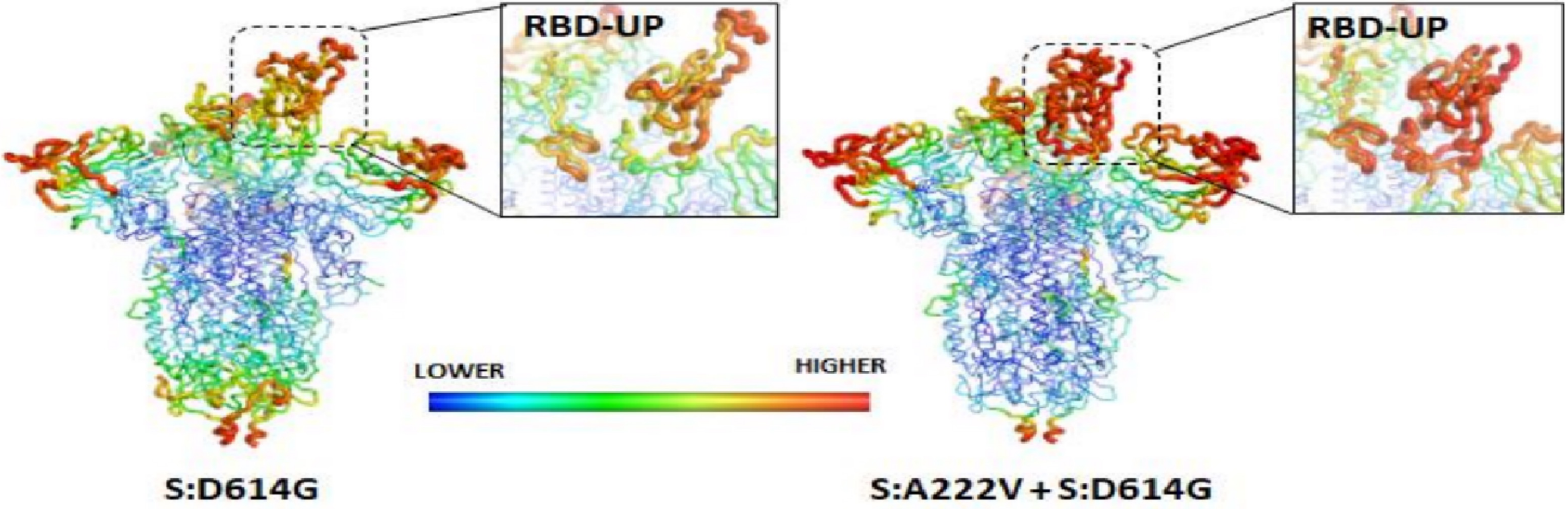
- First appeared in Spain in summer 2020
- Reappeared in "Delta+" (AY.4.2) in late summer 2021, suggesting an advantage
- Located in the N-terminal domain (NTD)
- No obvious functional role (glycosylation, antibody binding, receptor binding, ...) from structure alone

A222V: the "Spanish" mutant

- Alchemical simulations (mutating in open vs closed chain) show no alterations in the preference for opening

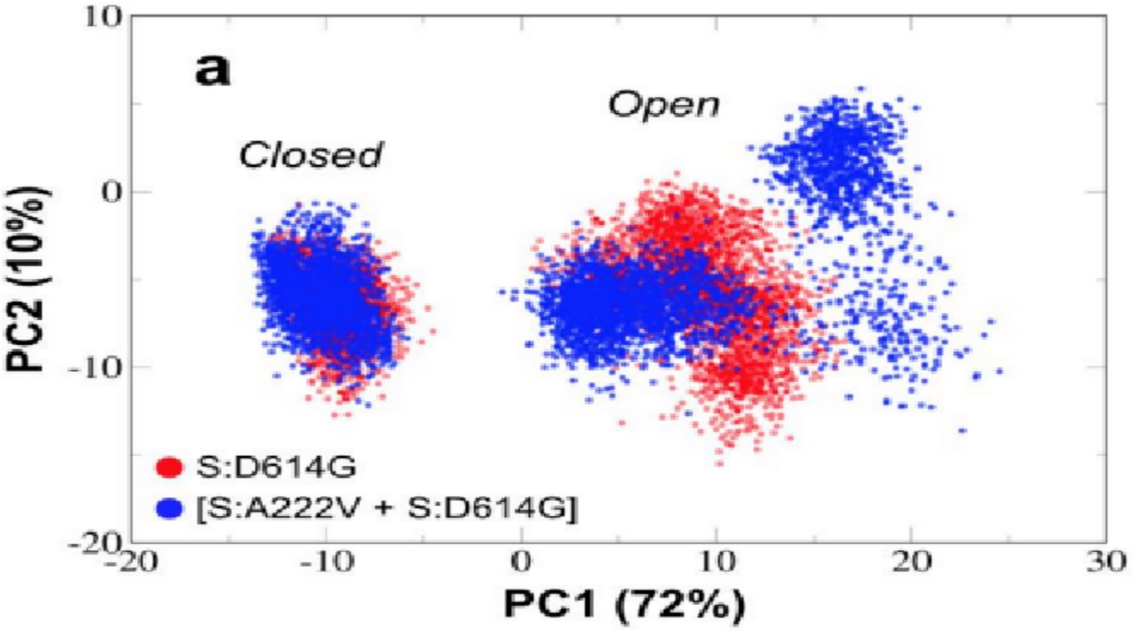
A222V: the "Spanish" mutant

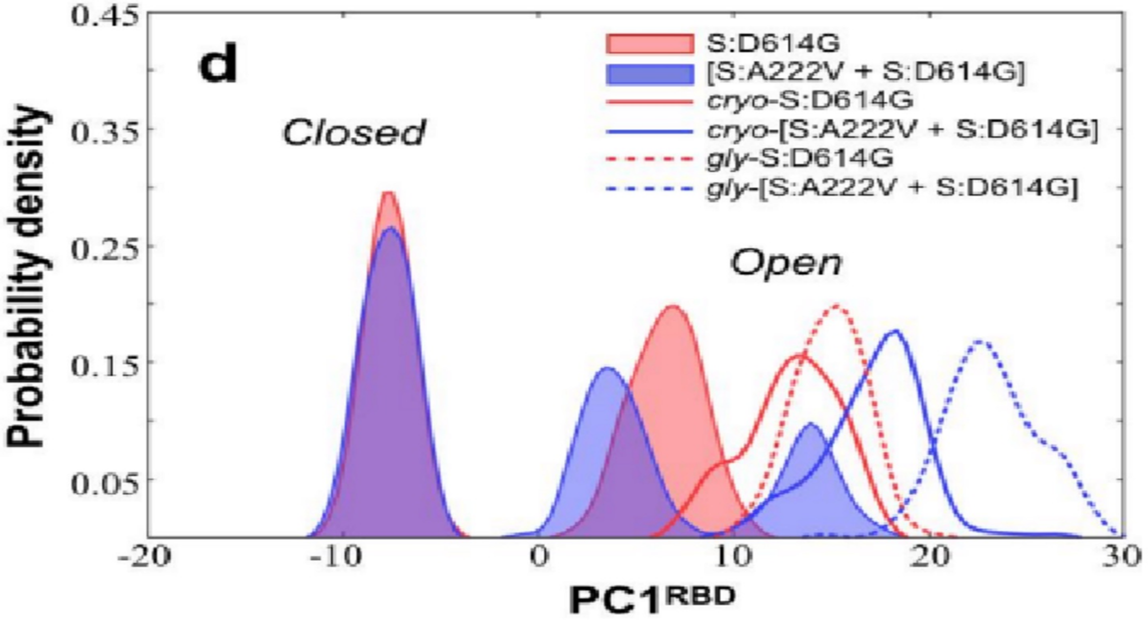
- Alchemical simulations (mutating in open vs closed chain) show no alterations in the preference for opening
- Hints from cryo-EM B-factors:



Can simulations reproduce it?

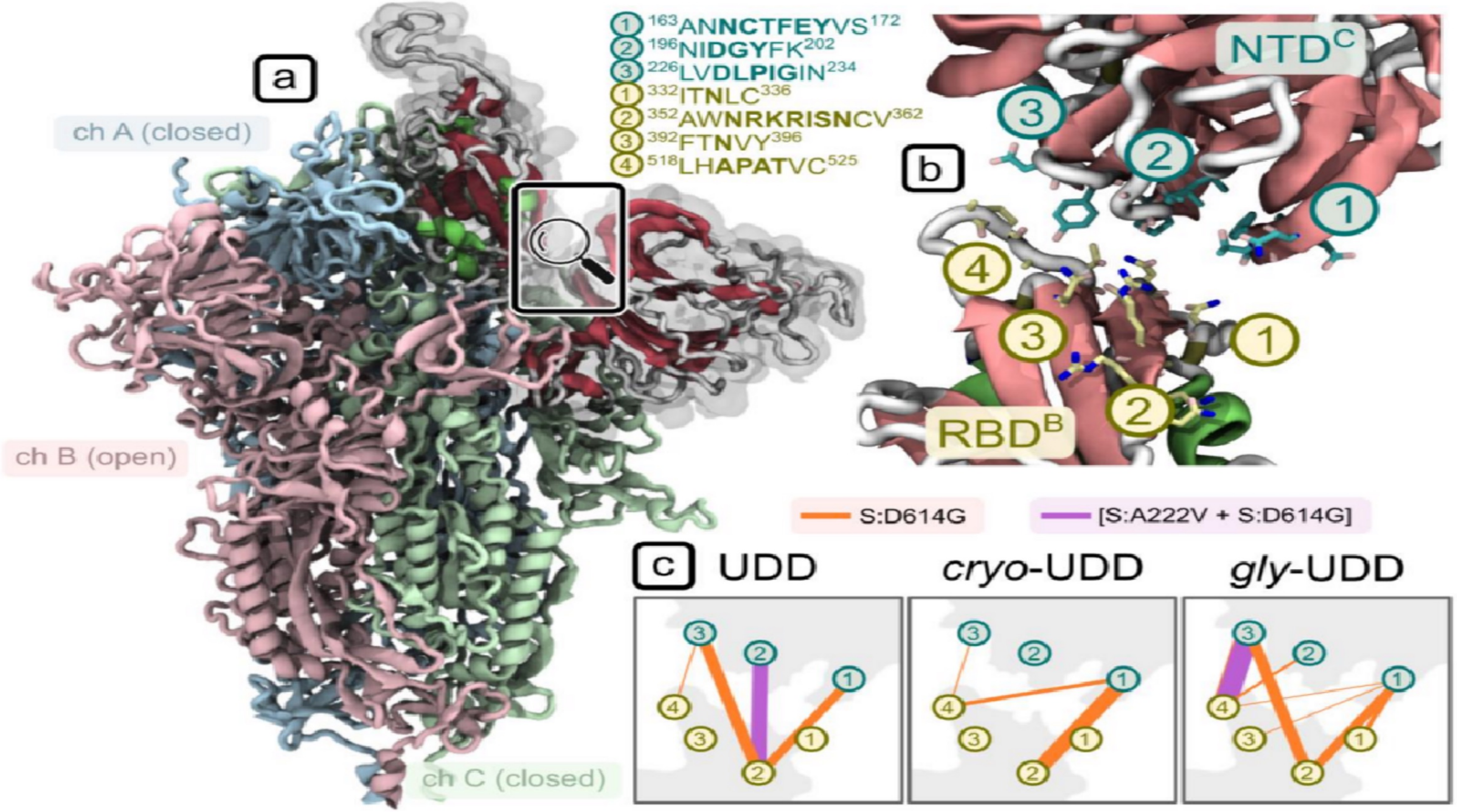
- Turns out they can: multiple simulations show enhanced flexibility of the RBD (sampling more conformational states)
- Apparent bimodal behavior





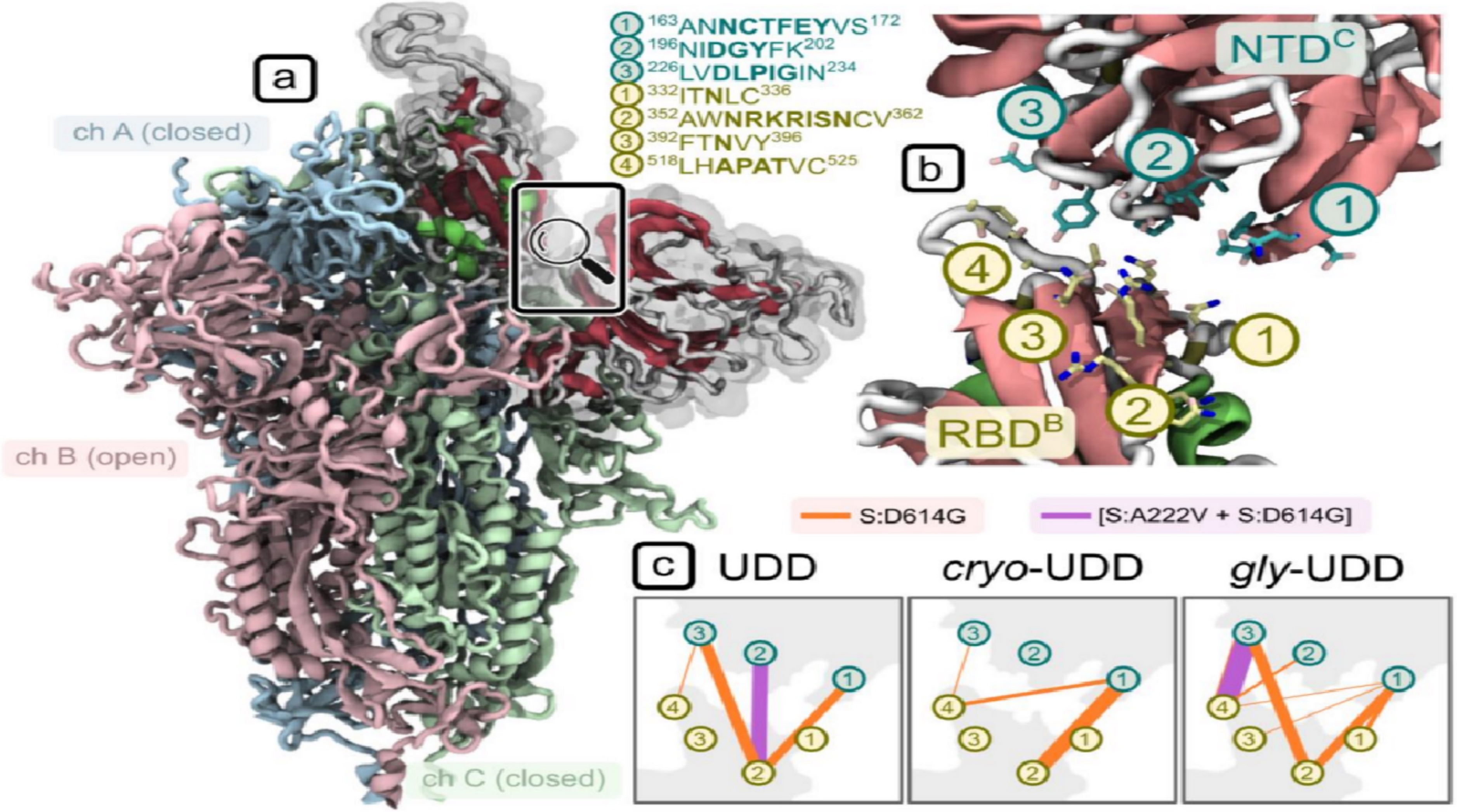
Can simulations explain it?

- Dynamic connectivities from network analysis show disruption of NTD-RBD contacts



Can simulations explain it?

- Dynamic connectivities from network analysis show disruption of NTD-RBD contacts
- Possible synergistic effects with other mutations (epistasis)



To wrap up:

- We are working to design robust strategies to rapidly calculate mutational free energy changes, and identify mutations crucial to crossing the zoonotic barrier
- Combining bioinformatics (polymorphism analysis) with alchemical free energies can be a powerful method for generating new testable hypotheses
- Multiple equilibrium simulations, alchemical free energies and allosteric analyses can provide a multi-angle characterization of single-residue mutants in Spike

To wrap up:

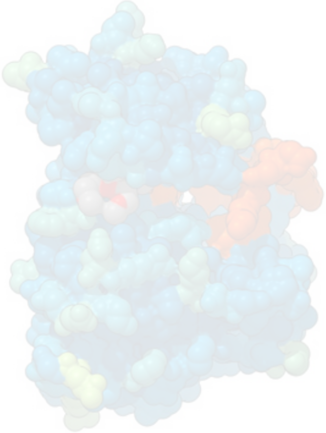
- We are working to design robust strategies to rapidly calculate mutational free energy changes, and identify mutations crucial to crossing the zoonotic barrier
- Combining bioinformatics (polymorphism analysis) with alchemical free energies can be a powerful method for generating new testable hypotheses
- Multiple equilibrium simulations, alchemical free energies and allosteric analyses can provide a multi-angle characterization of single-residue mutants in Spike

To wrap up:

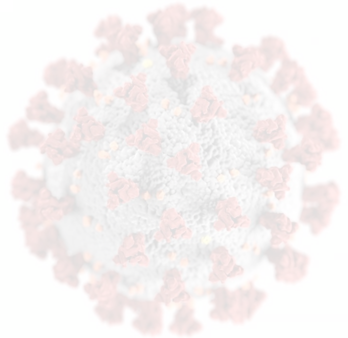
- We are working to design robust strategies to rapidly calculate mutational free energy changes, and identify mutations crucial to crossing the zoonotic barrier
- Combining bioinformatics (polymorphism analysis) with alchemical free energies can be a powerful method for generating new testable hypotheses
- Multiple equilibrium simulations, alchemical free energies and allosteric analyses can provide a multi-angle characterization of single-residue mutants in Spike

Projects

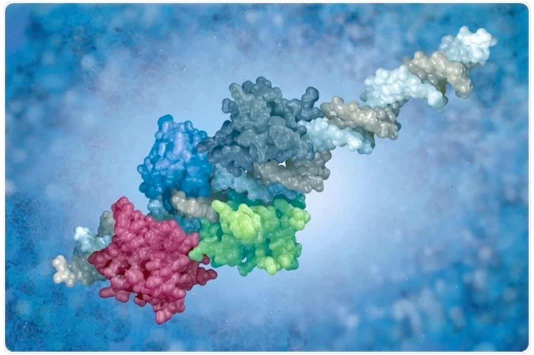
- High-throughput **prediction** of the **impact** of **genetic variability** on **drug sensitivity** and **resistance patterns** for clinically relevant **EGFR mutations** from atomistic simulations.



- Large-scale **SARS-CoV2 mutation** analysis, including a study on the **evolutionary path** and **host-selection mechanism** of **SARS-CoV-2**.

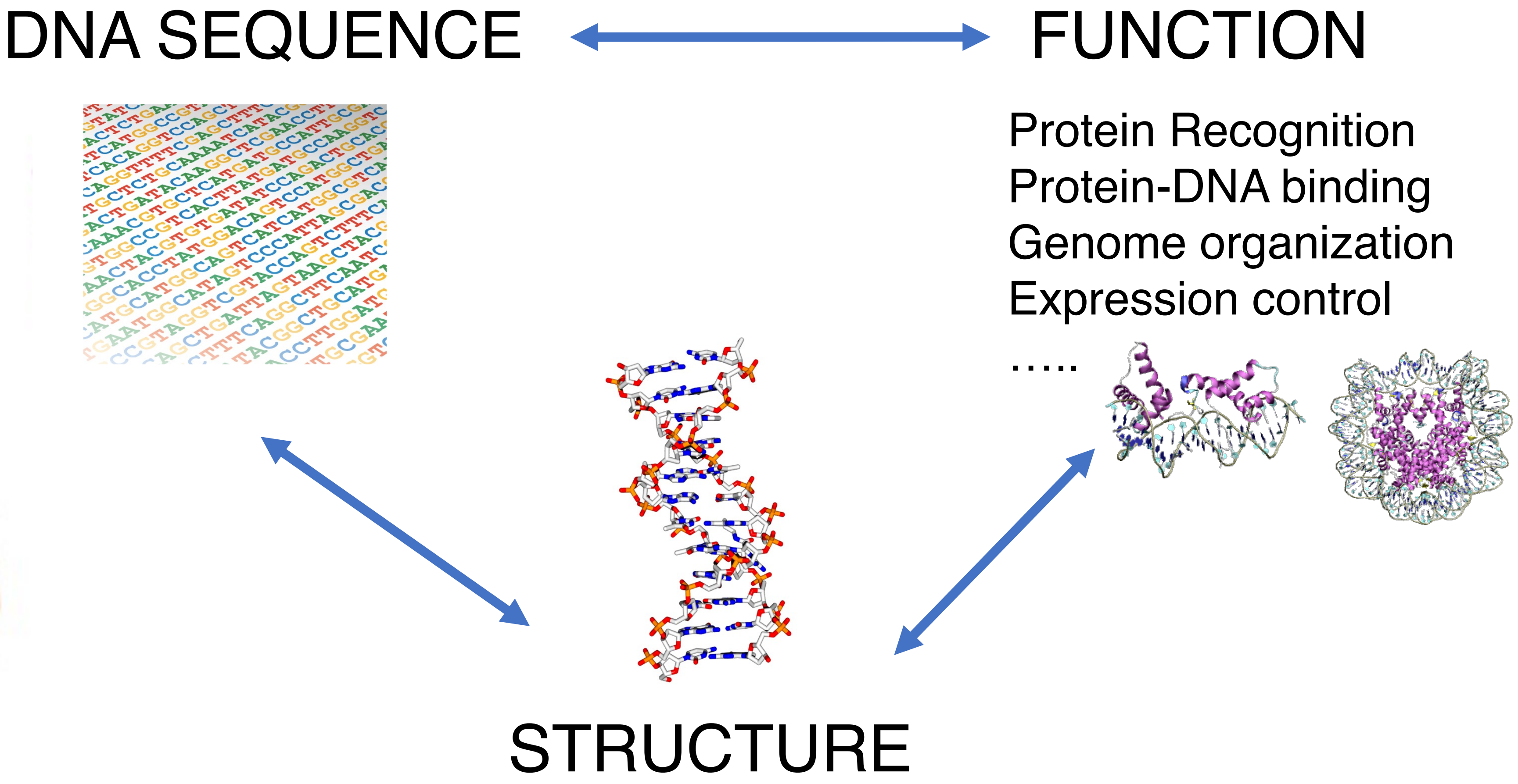


- **DNAffinity**: A **Machine-Learning** approach to **predict DNA Binding affinities** of **Transcription Factors**.



Aim

Prediction of the most likely binding sites for different TFs along a given DNA sequence



DNA SEQUENCE

FUNCTION

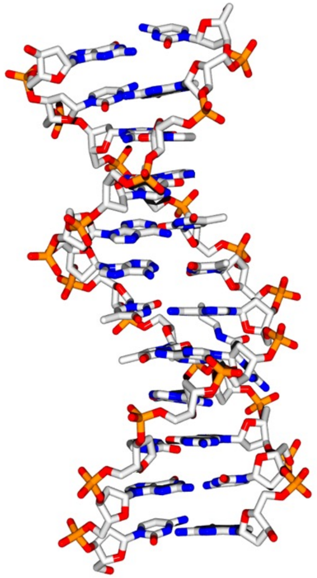
Protein Recognition
Protein-DNA binding
Genome organization
Expression control
…..

STRUCTURE

Methods: Machine learning workflow

- The model takes into account
  - Experimental data
  - Computationally derived structural DNA properties (including neighboring effect)

Scheme ML



DNA sequences

Labels

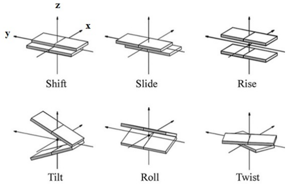
AFFINITIES

In vitro experiments

Features

DNA PROPERTIES (at tetramer level)

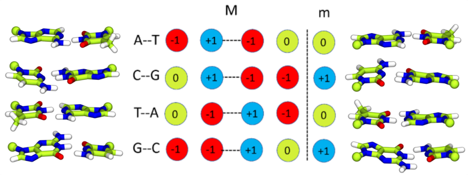
Base pair parameters (AVG) and stiffness (DIAG)



$$\Xi_{\mathrm{h}} = k_{\mathrm{B}} T C_{\mathrm{h}}^{-1} = \begin{bmatrix} k_{\mathrm{twist}} & k_{t-r} & k_{t-l} & k_{t-i} & k_{t-s} & k_{t-d} \\ k_{t-r} & k_{\mathrm{roll}} & k_{r-l} & k_{r-i} & k_{r-s} & k_{r-d} \\ k_{t-l} & k_{r-l} & k_{tilt} & k_{l-i} & k_{l-s} & k_{l-d} \\ k_{t-i} & k_{r-i} & k_{l-i} & k_{\mathrm{rise}} & k_{i-s} & k_{i-d} \\ k_{t-s} & k_{r-s} & k_{l-s} & k_{i-s} & k_{\mathrm{shift}} & k_{s-d} \\ k_{t-d} & k_{r-d} & k_{l-d} & k_{i-d} & k_{s-d} & k_{\mathrm{slide}} \end{bmatrix}$$

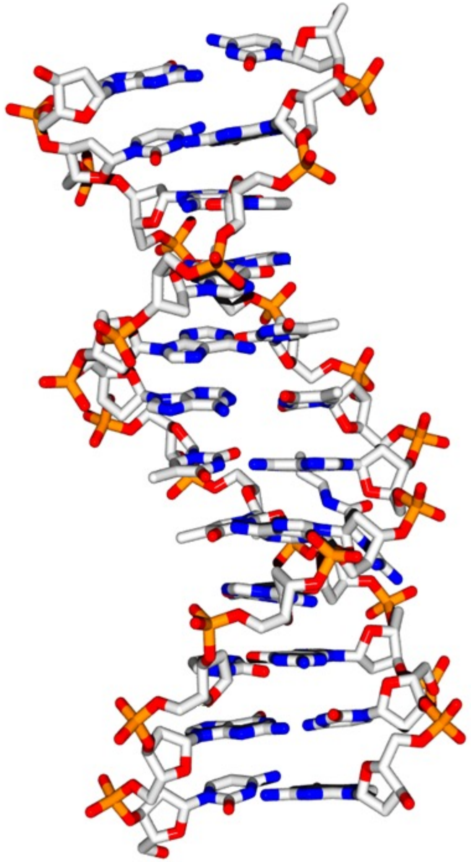
Sequence Pattern (PRESENCE probability)

Electrostatics



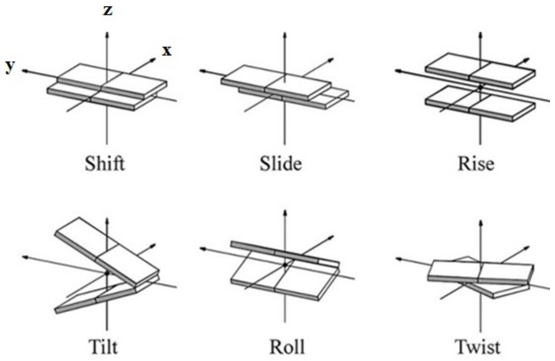
RANDOM FOREST REGRESSOR: Train the method over 80% of the data and test the remaining 20% (R^2)

Features



DNA CONFORMATION AT TETRAMER LEVEL (ParmBSC1):

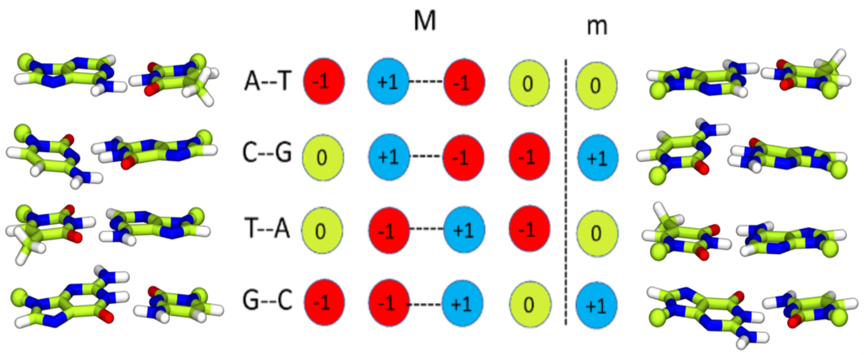
Base pair parameters and stiffness: (INDIRECT READOUT)



$$\Xi_{\mathrm{h}} = k_{\mathrm{B}} T C_{\mathrm{h}}^{-1} = \begin{bmatrix} k_{\mathrm{twist}} & k_{t-r} & k_{t-l} & k_{t-i} & k_{t-s} & k_{t-d} \\ k_{t-r} & k_{\mathrm{roll}} & k_{r-l} & k_{r-i} & k_{r-s} & k_{r-d} \\ k_{t-l} & k_{r-l} & k_{tilt} & k_{l-i} & k_{l-s} & k_{l-d} \\ k_{t-i} & k_{r-i} & k_{l-i} & k_{\mathrm{rise}} & k_{i-s} & k_{i-d} \\ k_{t-s} & k_{r-s} & k_{l-s} & k_{i-s} & k_{\mathrm{shift}} & k_{s-d} \\ k_{t-d} & k_{r-d} & k_{l-d} & k_{i-d} & k_{s-d} & k_{\mathrm{slide}} \end{bmatrix}$$

Sequence Pattern (PRESENCE probability)

Electrostatic potential at base pair level (DIRECT READOUT)



		M			m
A--T	-1	+1	-1	0	0
C--G	0	+1	-1	-1	+1
T--A	0	-1	+1	-1	0
G--C	-1	-1	+1	0	+1

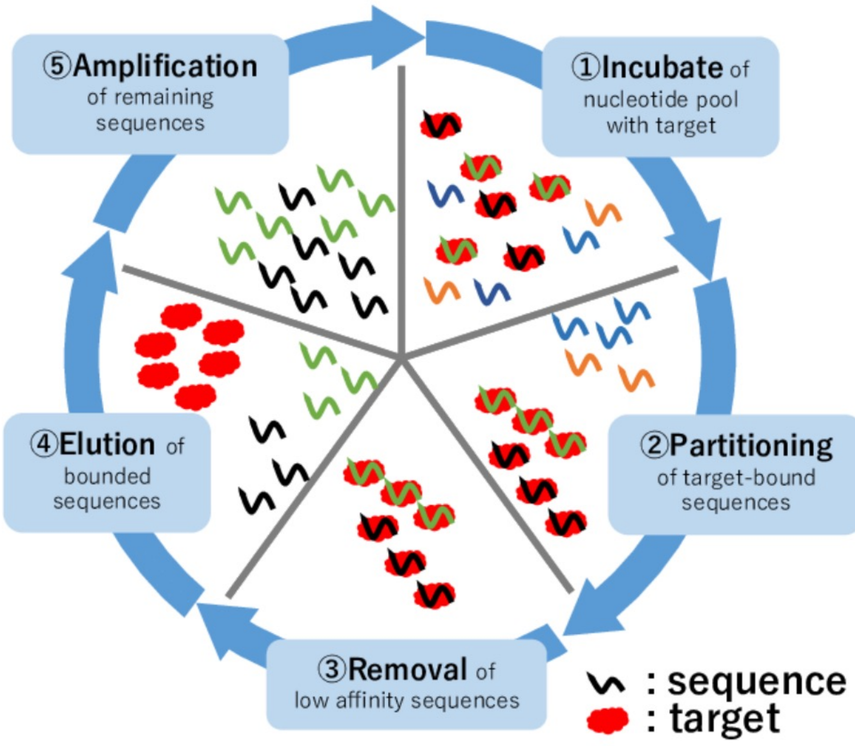
The static and dynamic structural heterogeneities of B-DNA: extending Calladine–Dickerson rules ,Pablo Dans, 2019, NAR

Labels

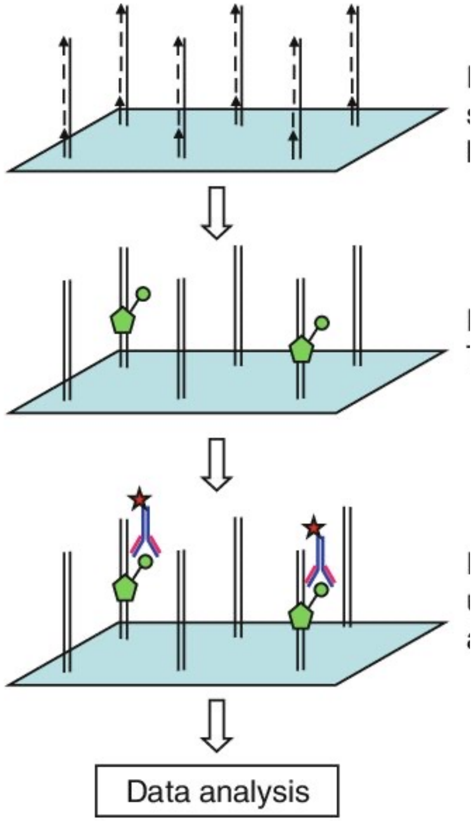
Binding affinity from HT-selex experiments for each TF

IN VITRO EXPERIMENTS

Binding affinity Protein Binding Microarray (PBM) Data for each TF



⑤Amplification of remaining sequences
①Incubate of nucleotide pool with target
②Partitioning of target-bound sequences
③Removal of low affinity sequences
④Elution of bounded sequences
: sequence
: target



Systematic characterization of protein-DNA interactions, Zhi Xie; Cellular and Molecular Life Sciences, 2011

Methods- Preprocessing Data

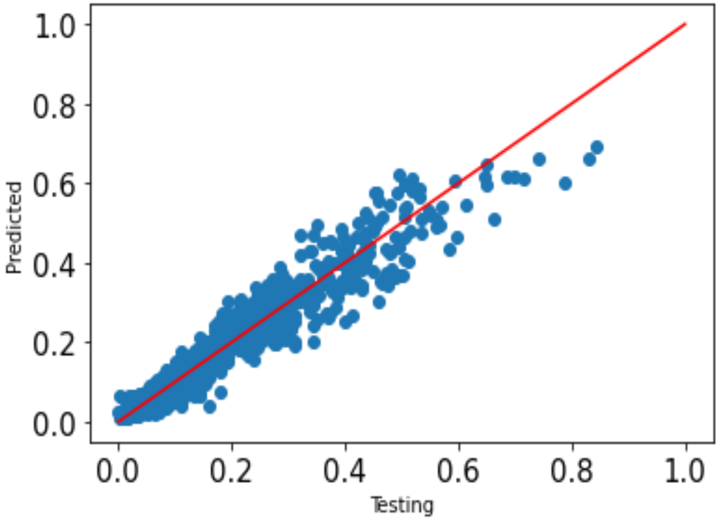
- uPBM (*universal PBM, 36mers*) cut and aligned based on position-weight-matrix (PWM) ofthe highest affinity sequences
Noisy and overrepresentation of low affinity binding sites: Undersampling (removing noise - *uPBM*)

- gcPBM (*genomic PBM*) already centered, removal of sequences with multiple binding site (gcPBM)

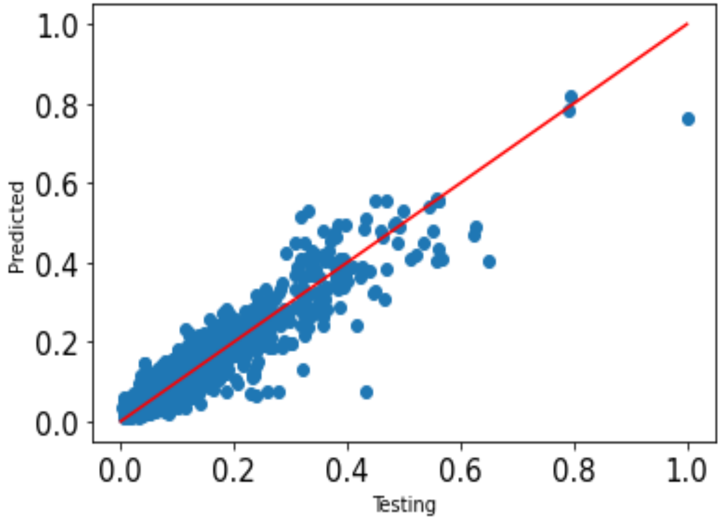
- HT-SELEX data quality assessment: Removing data with low P-value (not reliable) and filtering cases using the correlation between the counts across the different cycles.

Results (gcPBM)

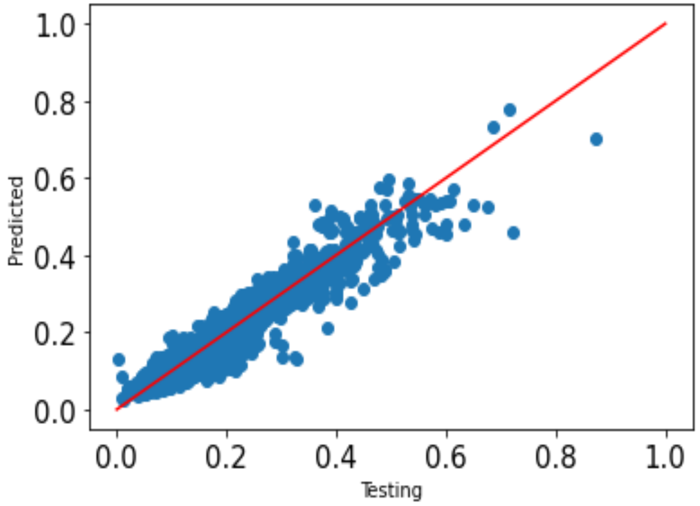
MAD1 (R²=0.951)

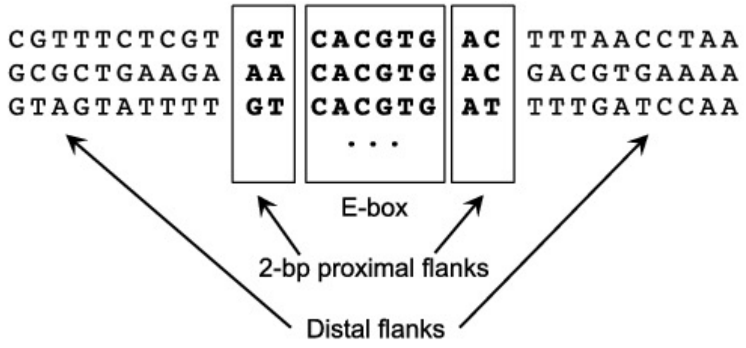


MYC (R²=0.905)



MAX (R²=0.922)





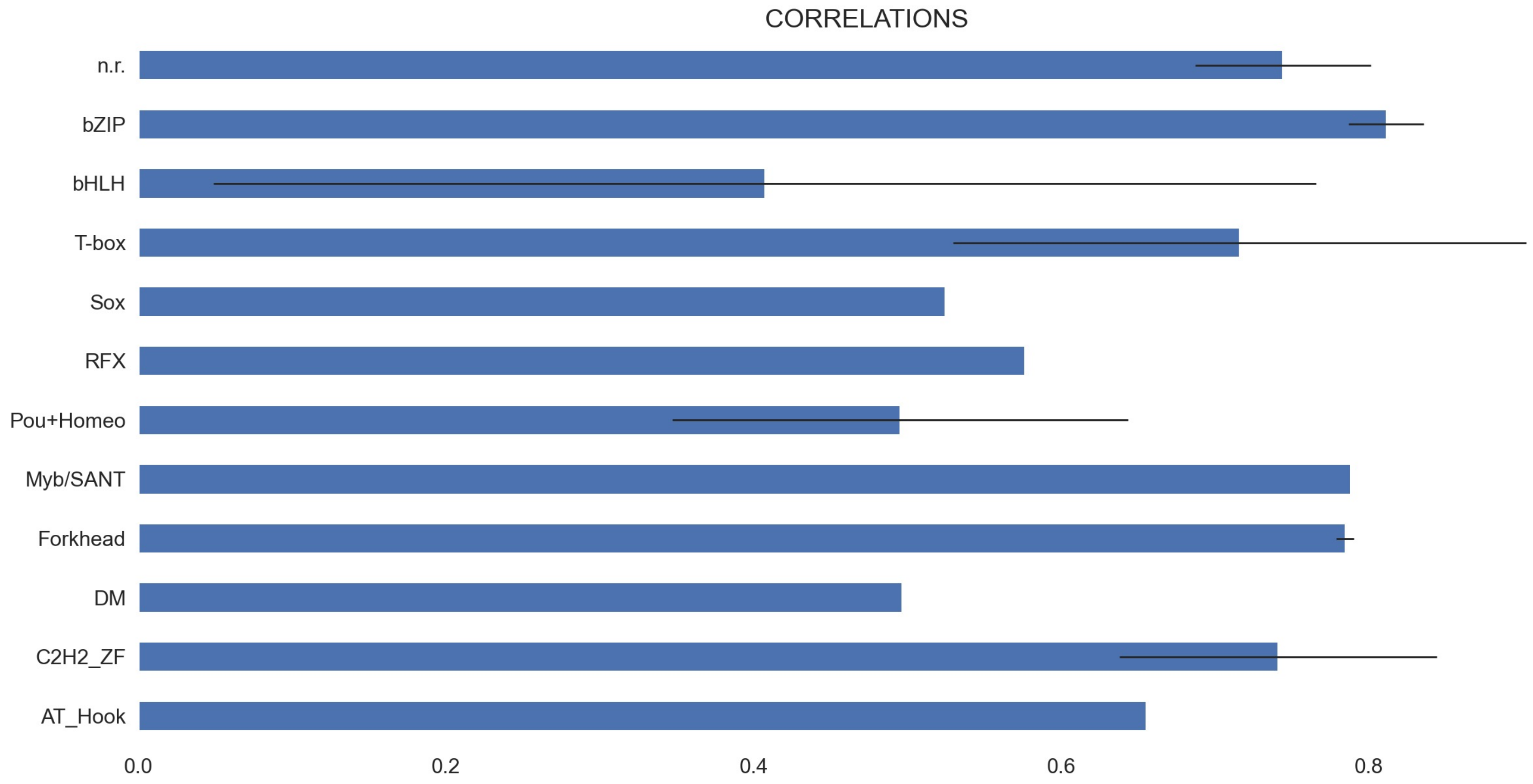
CGTTTCTCGT **GT** **CACGTG** **AC** TTTAACCTAA
GCGCTGAAGA **AA** **CACGTG** **AC** GACGTGAAAA
GTAGTATTTT **GT** **CACGTG** **AT** TTTGATCCAA
...

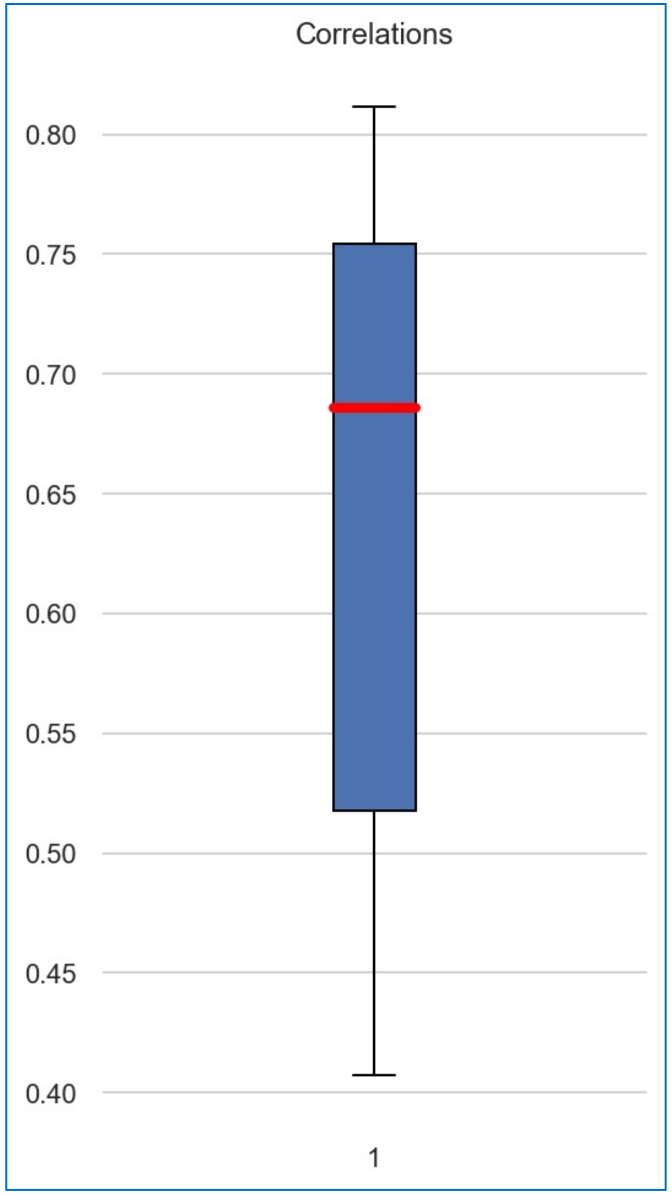
E-box

2-bp proximal flanks

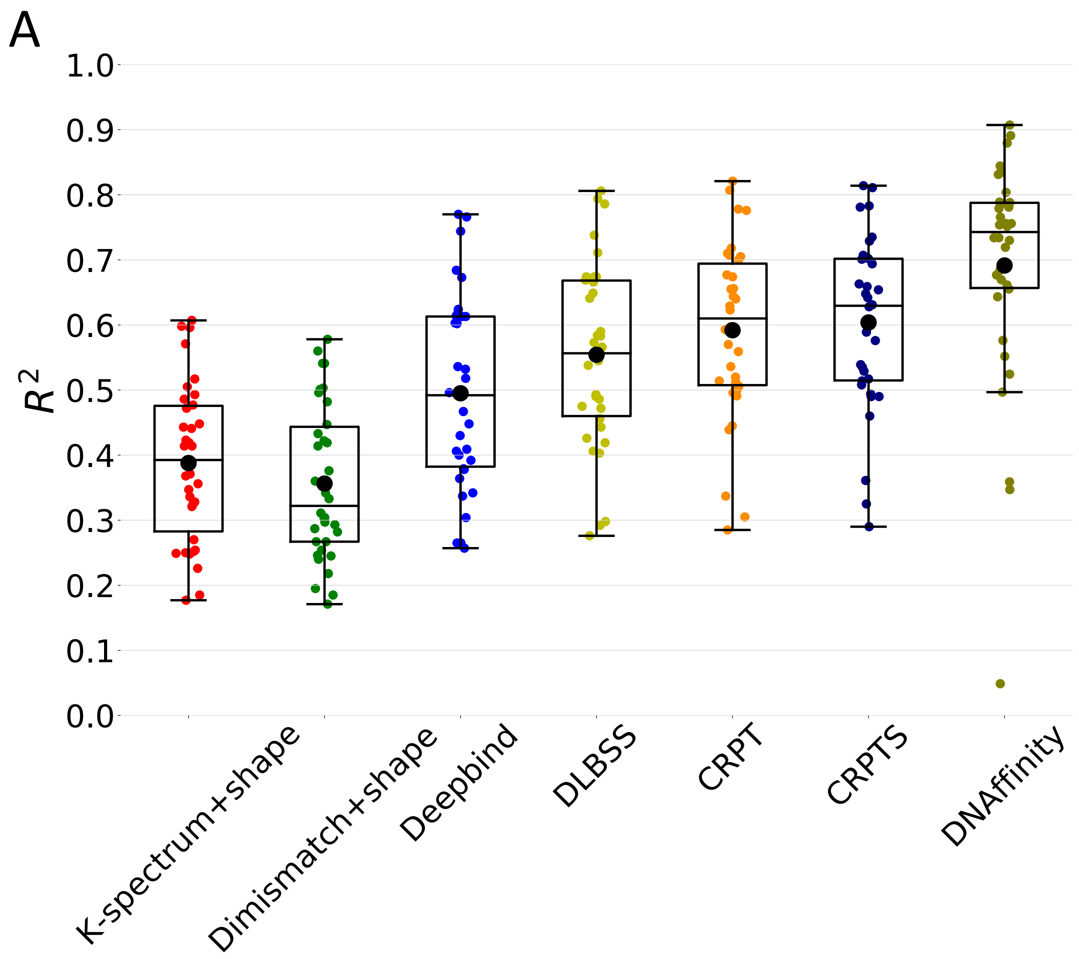
Distal flanks

Results (uPBM)

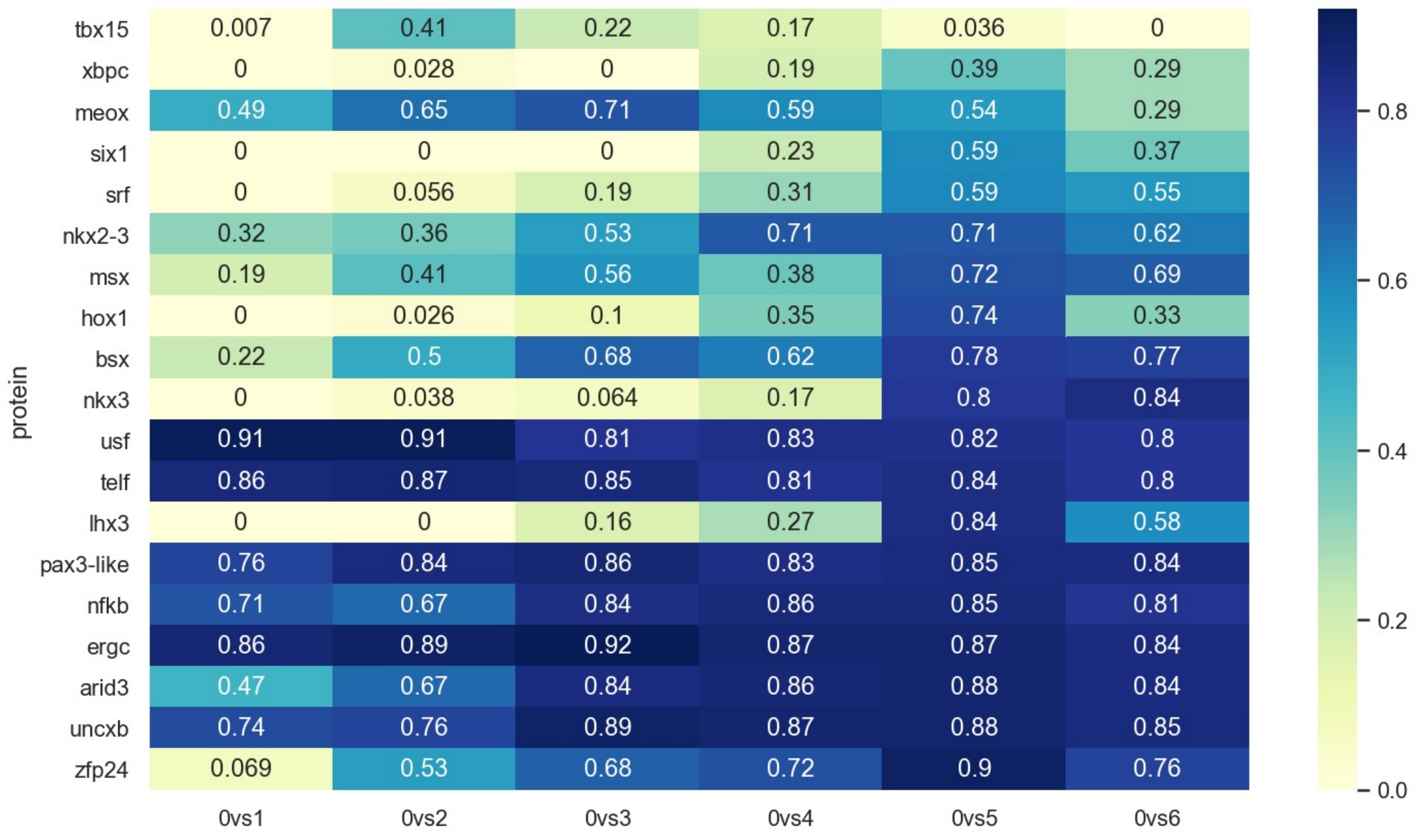




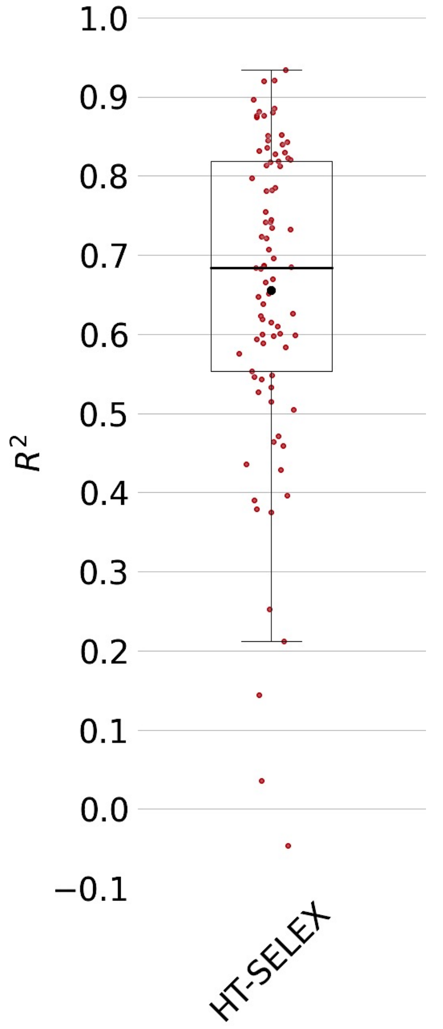
Results (uPBM)-Comparisons



Results (HT-Selex)

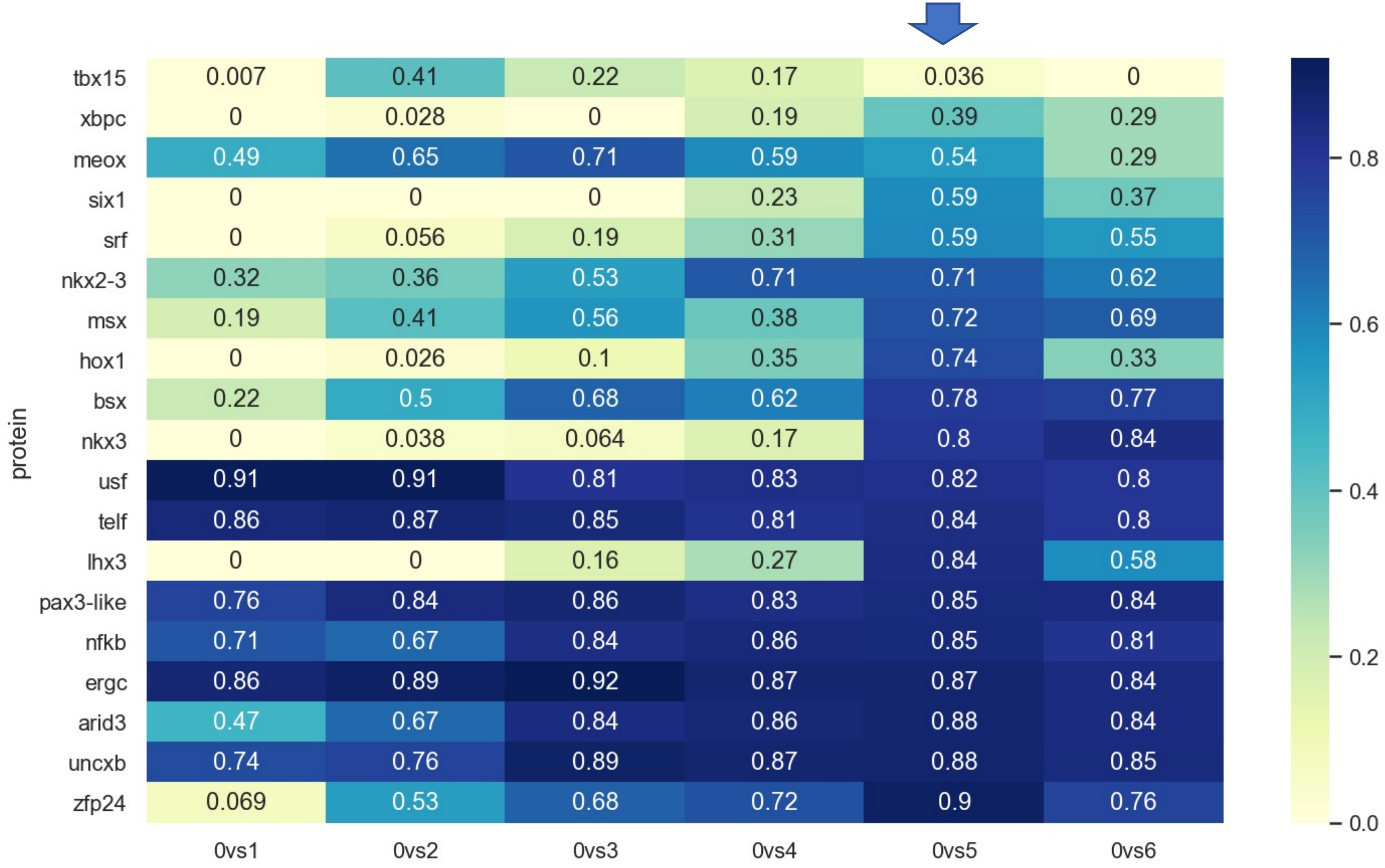


protein	0vs1	0vs2	0vs3	0vs4	0vs5	0vs6
tbx15	0.007	0.41	0.22	0.17	0.036	0
xbpc	0	0.028	0	0.19	0.39	0.29
meox	0.49	0.65	0.71	0.59	0.54	0.29
six1	0	0	0	0.23	0.59	0.37
srf	0	0.056	0.19	0.31	0.59	0.55
nkx2-3	0.32	0.36	0.53	0.71	0.71	0.62
msx	0.19	0.41	0.56	0.38	0.72	0.69
hox1	0	0.026	0.1	0.35	0.74	0.33
bsx	0.22	0.5	0.68	0.62	0.78	0.77
nkx3	0	0.038	0.064	0.17	0.8	0.84
usf	0.91	0.91	0.81	0.83	0.82	0.8
telf	0.86	0.87	0.85	0.81	0.84	0.8
lhx3	0	0	0.16	0.27	0.84	0.58
pax3-like	0.76	0.84	0.86	0.83	0.85	0.84
nfkb	0.71	0.67	0.84	0.86	0.85	0.81
ergc	0.86	0.89	0.92	0.87	0.87	0.84
arid3	0.47	0.67	0.84	0.86	0.88	0.84
uncxb	0.74	0.76	0.89	0.87	0.88	0.85
zfp24	0.069	0.53	0.68	0.72	0.9	0.76

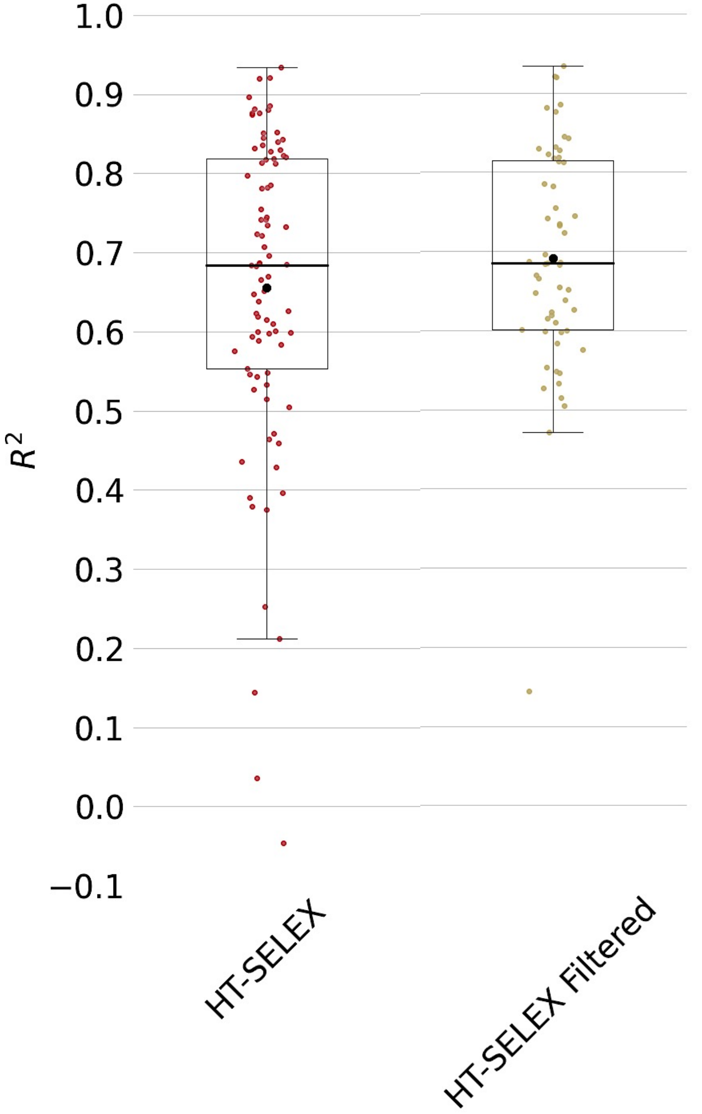


SELEX-Seq: A Method to Determine DNA Binding Specificities of Plant Transcription Factors; Smaczniak C., Methods Mol Biol 2017

Results (HT-Selex)

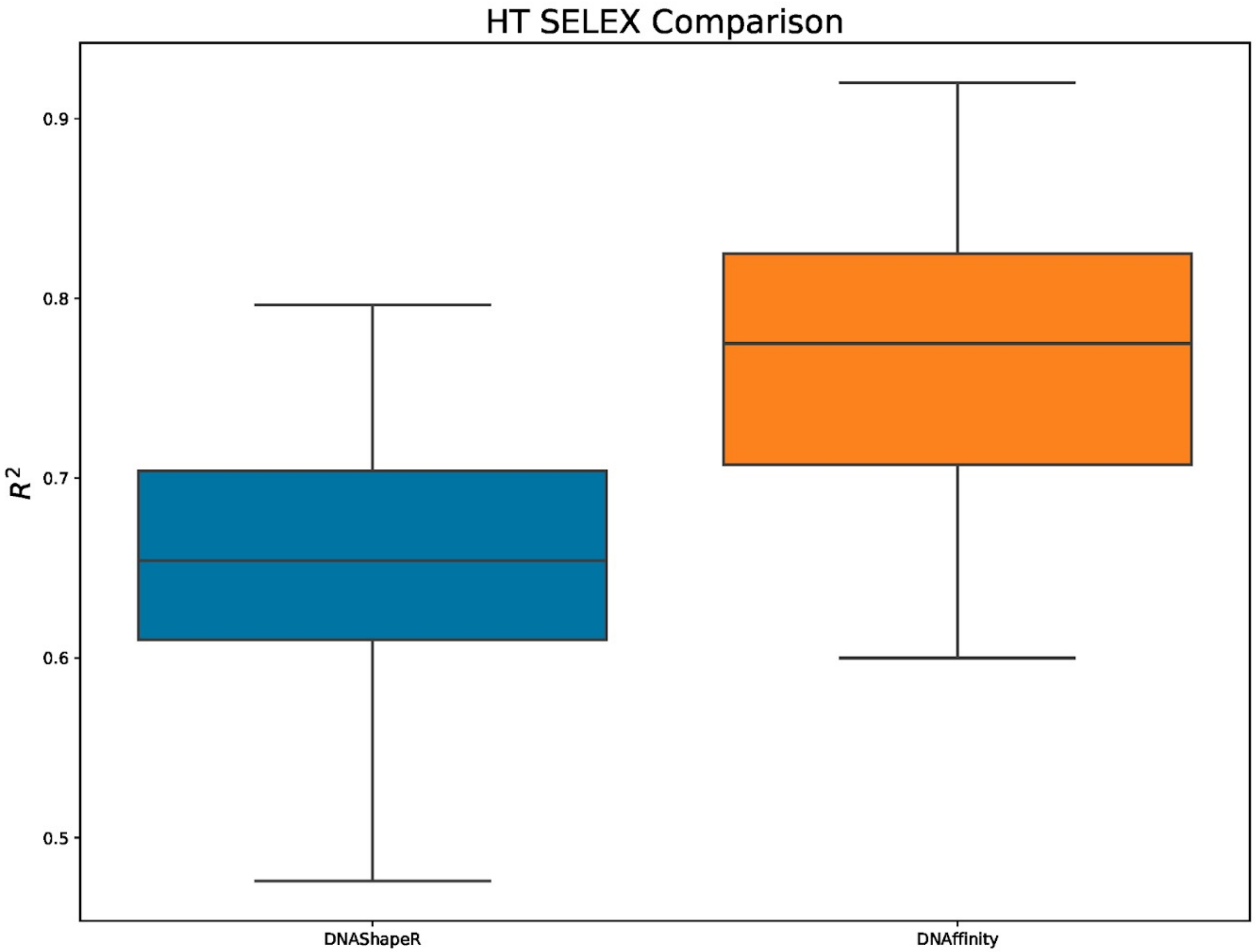


protein	0vs1	0vs2	0vs3	0vs4	0vs5	0vs6
tbx15	0.007	0.41	0.22	0.17	0.036	0
xbpc	0	0.028	0	0.19	0.39	0.29
meox	0.49	0.65	0.71	0.59	0.54	0.29
six1	0	0	0	0.23	0.59	0.37
srf	0	0.056	0.19	0.31	0.59	0.55
nkx2-3	0.32	0.36	0.53	0.71	0.71	0.62
msx	0.19	0.41	0.56	0.38	0.72	0.69
hox1	0	0.026	0.1	0.35	0.74	0.33
bsx	0.22	0.5	0.68	0.62	0.78	0.77
nkx3	0	0.038	0.064	0.17	0.8	0.84
usf	0.91	0.91	0.81	0.83	0.82	0.8
telf	0.86	0.87	0.85	0.81	0.84	0.8
lhx3	0	0	0.16	0.27	0.84	0.58
pax3-like	0.76	0.84	0.86	0.83	0.85	0.84
nfkb	0.71	0.67	0.84	0.86	0.85	0.81
ergc	0.86	0.89	0.92	0.87	0.87	0.84
arid3	0.47	0.67	0.84	0.86	0.88	0.84
uncxb	0.74	0.76	0.89	0.87	0.88	0.85
zfp24	0.069	0.53	0.68	0.72	0.9	0.76



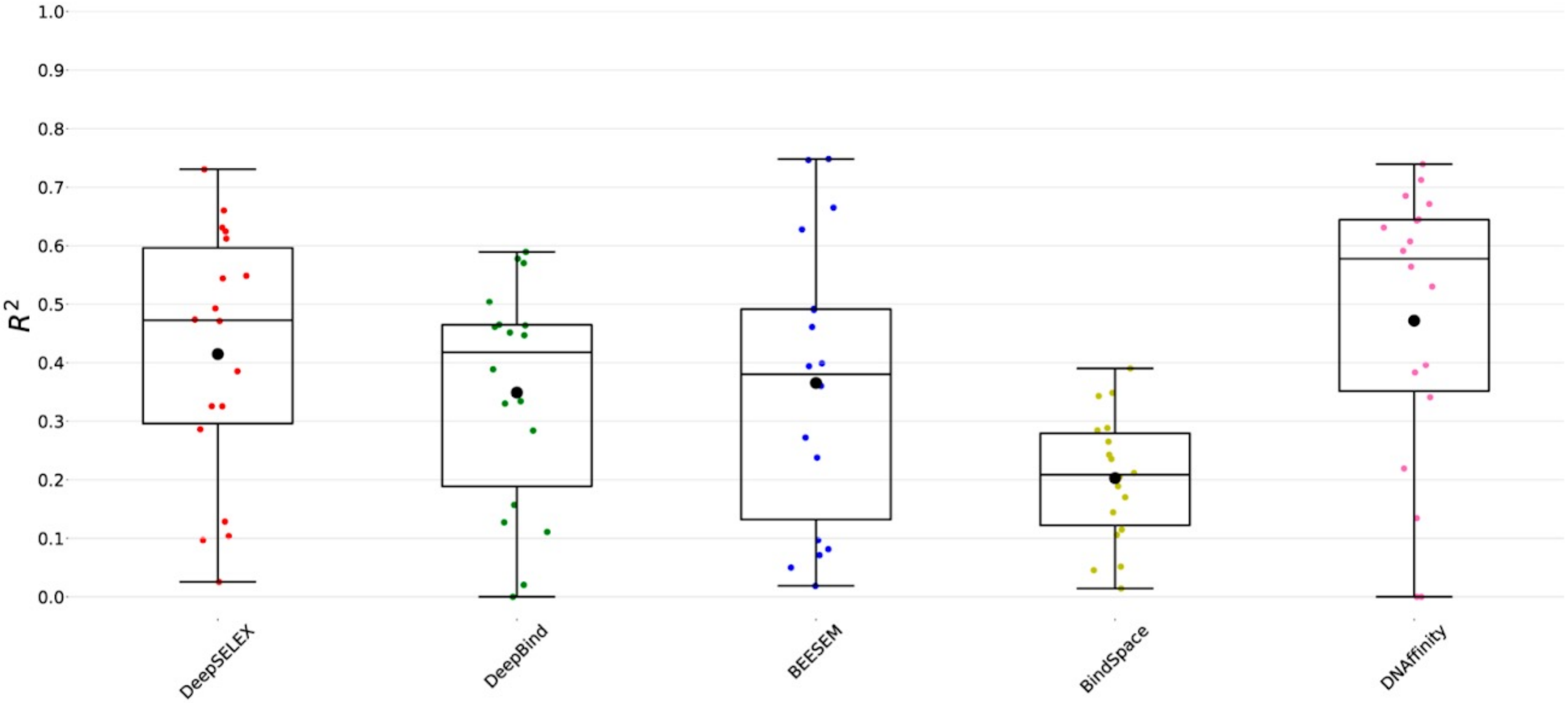
SELEX-Seq: A Method to Determine DNA Binding Specificities of Plant Transcription Factors; Smaczniak C., Methods Mol Biol 2017

Results (HT-Selex)-Comparison



Expanding the repertoire of DNA shape features for genome-scale studies of transcription factor binding, Li et all, NAR (2017)

Results (HT-Selex –> uPBM)



Conclusions

- Using our machine learning algorithm, we were able to predict the experimental TF-DNA affinity with an average correlation of 70%.

- Our method can be applied to data from different experimental techniques.

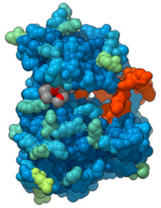
- We can use our trained model to predict *in vivo* transcription factor binding sites -> to be extended to whole genome

Acknowledgements

Prof. Modesto Orozco
Dr. Francesco Colizzi
Daniel Beltrán

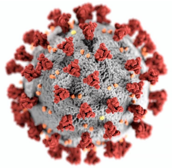
Dr. Robert Soliva
Dr. Yvonne Westermaier
Aristarc Suriñach
Martí Municoy

Prof. Josep Lluís Gelpí
Lluis Jordà
Sergi Orozco-Ruiz
Pau Andrio

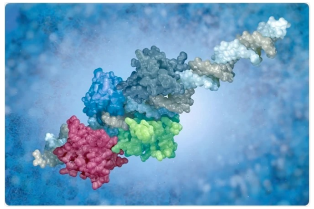


Prof. Modesto Orozco
Dr. Vito Genna
Dr. Adam Hospital

Jose Maria Carazo
James Krieger
Tiziana Ginex
Clara Marco-Marín
Carlos P. Mata
Paula Ruiz-Rodriguez
José Luis Llácer
Mireia Coscolla
Carmen Gil
Iñaki Comas



Prof. Modesto Orozco
Sandro Barissi
Alba Sala
Dr. Milosz Wieczor



and THANK YOU!

Acknowledgements

Molecular Modeling and Bioinformatics

bioexcel

bioexcel

Centre of Excellence for Computational Biomolecular Research

BioExcel Partners 2019

BSC
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Utrecht University

NORMAN CONSULTING

EMBL-EBI

acrosslimits

Ian Harrow Consulting

MAX-PLANCK-GESELLSCHAFT

KTH VETENSKAP OCH KONST

MANCHESTER 1824

THE UNIVERSITY OF EDINBURGH

UNIVERSITY OF JYVÄSKYLA

JÜLICH FORSCHUNGSZENTRUM

NBD NOSTRUM BIODISCOVERY

IRB BARCELONA
INSTITUTO DE INVESTIGACIÓN BIOMÉDICA

Horizon 2020
European Union Funding
for Research & Innovation

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