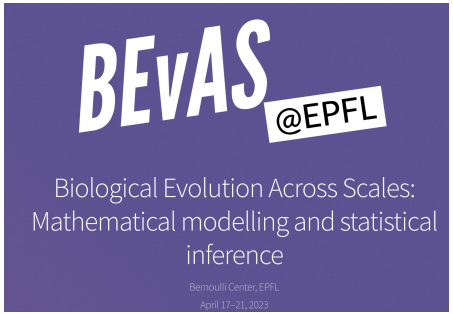
# Exploiting evolutionary patterns in homologous protein sequences to predict short-term polymorphisms: applications to *E. coli* and SARS-CoV-2




**Giancarlo Croce**

Biological Evolution Across Scales: Mathematical modelling and statistical inference

Bernoulli Center, EPFL
April 19, 2023

UNIL | Université de Lausanne

LUDWIG CANCER RESEARCH
Lausanne

SIB
Swiss Institute of Bioinformatics

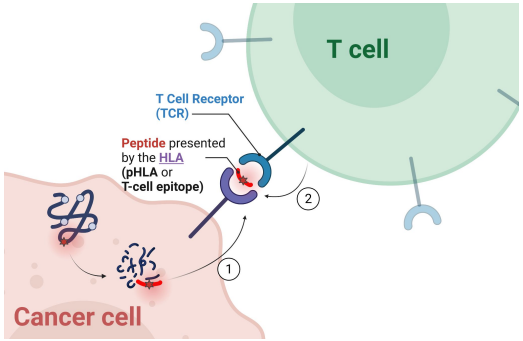
# Giancarlo Croce

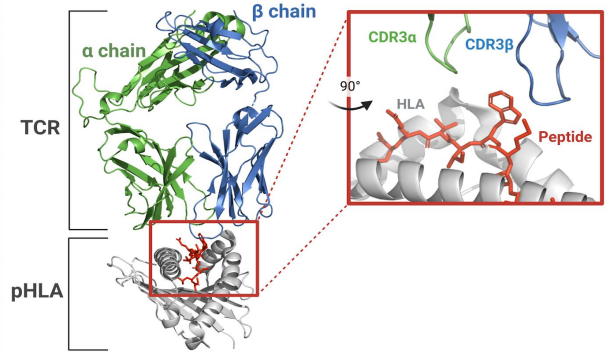
**PostDoc:** *D. Gfeller* - Computational Cancer Biology Lab - UNIL

Computational methods to better understand interaction between cancer and immune cells

**PhD:** *M. Weigt* - Computational and quantitative biology Lab - Paris Sorbonne University

Statistical-physics inspired method (Direct coupling analysis) to model and predict protein evolution



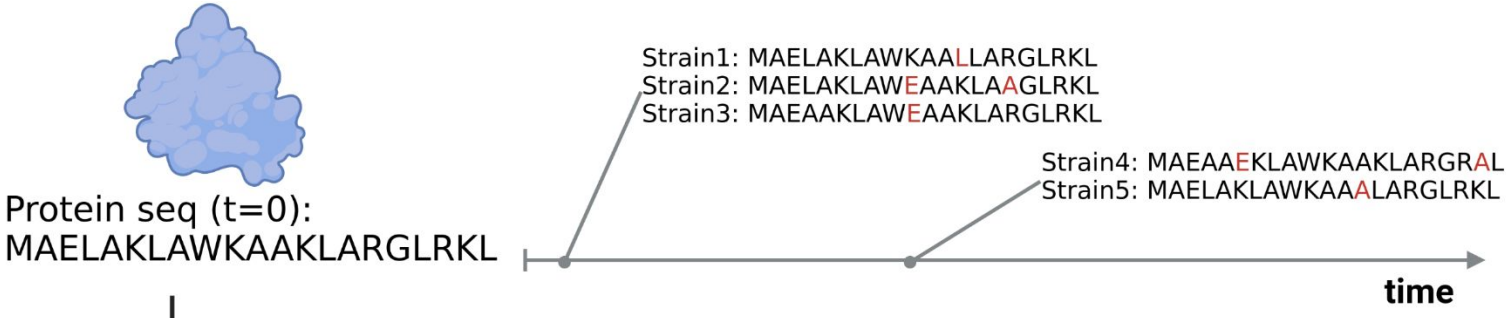



# Giancarlo Croce

**PhD**: *M. Weigt* - Computational and quantitative biology Lab - Paris Sorbonne University

Statistical-physics inspired method (Direct Coupling Analysis) to *model and predict protein evolution*

**Predicting polymorphic positions**

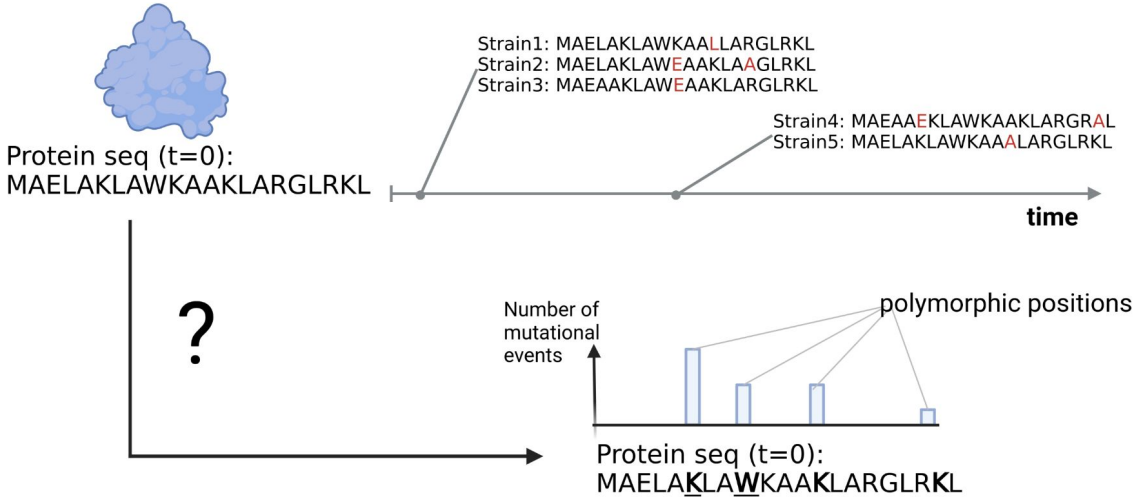


# Giancarlo Croce

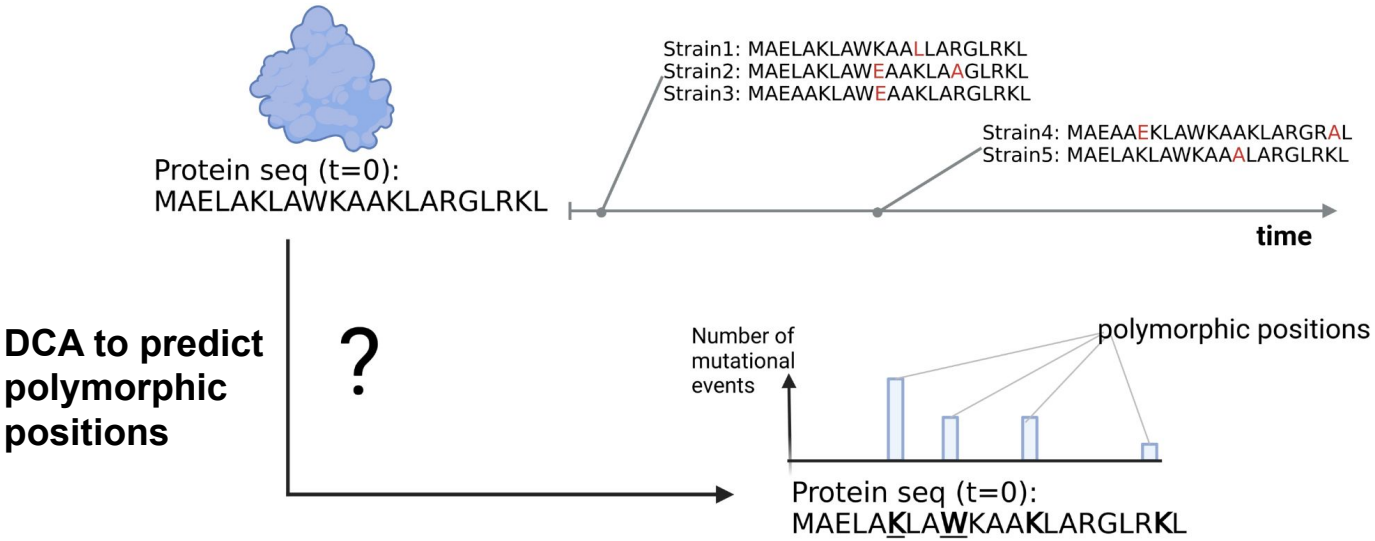
**PhD**: *M. Weigt* - Computational and quantitative biology Lab - Paris Sorbonne University

Statistical-physics inspired method (Direct Coupling Analysis) to *model and predict protein evolution*

**DCA to predict polymorphic positions**



# DCA to predict polymorphic positions




RESEARCH ARTICLE | BIOPHYSICS AND COMPUTATIONAL BIOLOGY

## Epistatic models predict mutable sites in SARS-CoV-2 proteins and epitopes

Juan Rodriguez-Rivas, Giancarlo Croce, Maureen Muscat, and Martin Weigt  Authors Info & Affiliations

Edited by John Barton, Physics and Astronomy, University of California, Riverside, CA; received July 16, 2021; accepted December 13, 2021 by Editorial Board Member Mehran Kardar

**January 12, 2022** | 119 (4) e2113118119 | https://doi.org/10.1073/pnas.2113118119

nature > nature communications > articles > article

Article | Open Access | Published: 12 July 2022

## Deciphering polymorphism in 61,157 *Escherichia coli* genomes via epistatic sequence landscapes

Lucile Vigué, Giancarlo Croce, Marie Petitjean, Etienne Ruppé, Olivier Tenaillon & Martin Weigt

*Nature Communications* **13**, Article number: 4030 (2022) | Cite this article

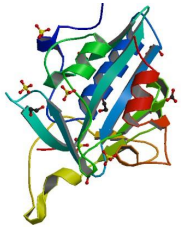
# Direct Coupling Analysis (DCA)

## Sequence

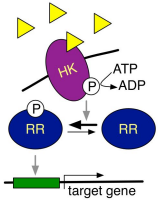
Human TPVNILKGKNQVMHLSAQERSAEEYQQALVADNIEELEGLSRLTENILFLAR

From the *genotype* (the protein sequence) to the *phenotype*

- protein structure



- protein function



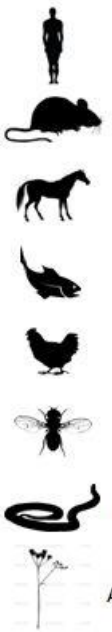

- mutational effects

M
ALG↓MLDHIMHQW
I

- And many more..

# Direct Coupling Analysis (DCA)

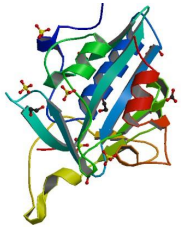
## Multiple sequence alignment (MSA) of homologous proteins



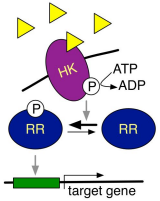
| | |
|---|---|
| Human | TPVNILKGKNQVMHLSAQERSAEEYQQALVADNIEELEGLSRLTENILFLAR |
| Mouse | TPIAIIKANTEVLHEI----TMGK-NQWTEKDILKQVKRLSGLVNDMVALAK |
| Horse | NMLTGVWGSLDLIHKLS----GRLVERFMDAYALISAQRLASLTDRLLAFSR |
| Zebrafish | QPINSIKLIAQDMHADYGELTDGDVQTTIDKDMSLLEHLSQTLVDVFRGFYR |
| Chicken | NPNAVIWLNVDLVHKKWSEMSEEL-PLLLTEYEEGAGRLKRILVDDLKDFAR |
| Fruit Fly | NILQIIWGNTQILHQYQTNPDPP----QLLEYLKAVERLTALLTRSMLAFSR |
| Nematode | TPLNAIKGFIQVLHKD-AEMKPKD-REYLELDDESSKNLLSLLVNDIIEIDL |
| Arabidopsis | TPVATLKGYLEAVHEDVRPLDAST----IAVDRDQAVRLTRLLAQDLADVTH |

Sequence identity ~20,30%

- protein structure



- protein function

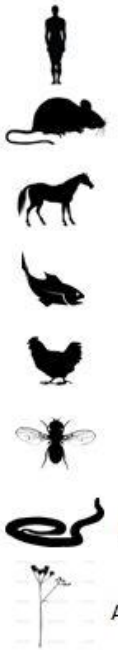



- mutational effects


- And many more..

# Direct Coupling Analysis (DCA)

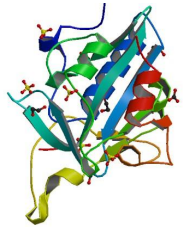
## Multiple sequence alignment (MSA) of homologous proteins



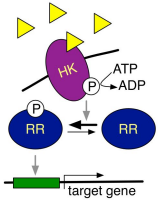
```
Human      TPVNILKGKNQVMHLSAQERSAEEYQQALVADNIEELEGLSRLTENILFLAR
Mouse      TPIAIIKANTEVLHEI----TMGK-NQWTEKDILKQVKRLSGLVNDMVALAK
Horse      NMLTGVWGSLDLIHKLS----GRLVERFMDAYALISAQRLASLTDRLLAFSR
Zebrafish  QPINSIKLIAQDMHADYGELTDGDVQTTIDKDMSLLEHLSQTLVDVFRGFYR
Chicken    NPNAVIWLNVDLVHKKWSEMSEEL-PLLLTEYEEGAGRLKRILVDDLKDFAR
Fruit Fly  NILQIIWGNTQILHQYQTNPDPP----QLLEYLKAVERLTALLTRSMLAFSR
Nematode   TPLNAIKGFIQVLHKD-AEMKPKD-REYLELDDESSKNLLSLLVNDIIEIDL
Arabidopsis TPVATLKGYLEAVHEDVRPLDAST----IAVDRDQAVRLTRLLAQDLADVTH
```

Conservation patterns

- protein structure



- protein function




- mutational effects

```
   M
ALG↓MLDHIMHQW
   I
```

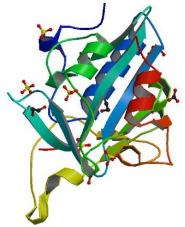
- And many more..

# Direct Coupling Analysis (DCA)

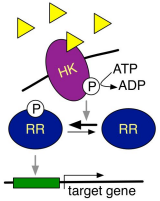
## Multiple sequence alignment (MSA) of homologous proteins


Correlation patterns

- protein structure



- protein function

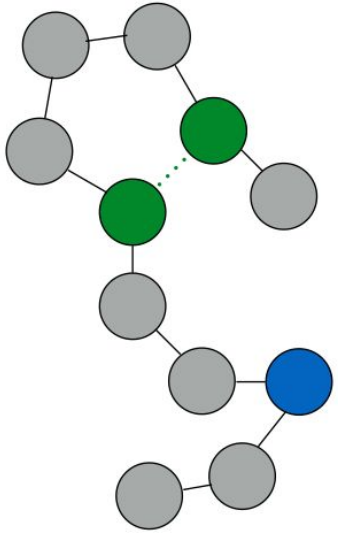



- mutational effects

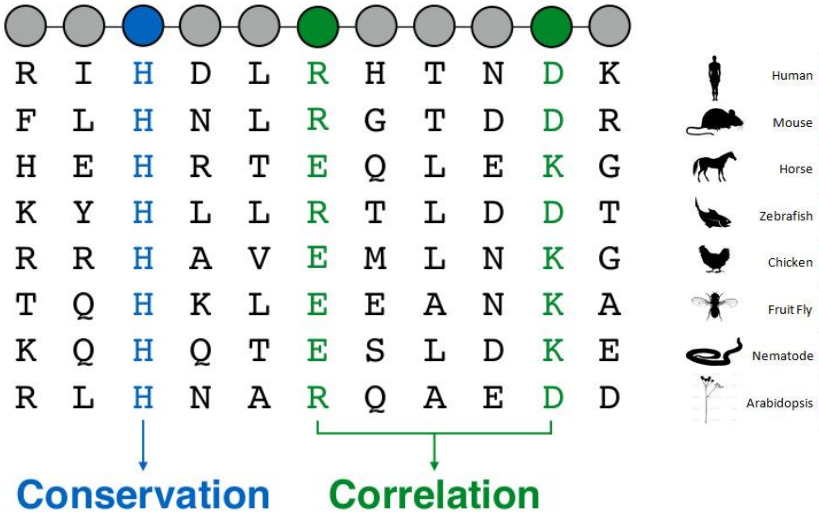

- And many more..

# Direct Coupling Analysis (DCA)

**conservation** of structure and function → **imposes constraints** on the sequence variability

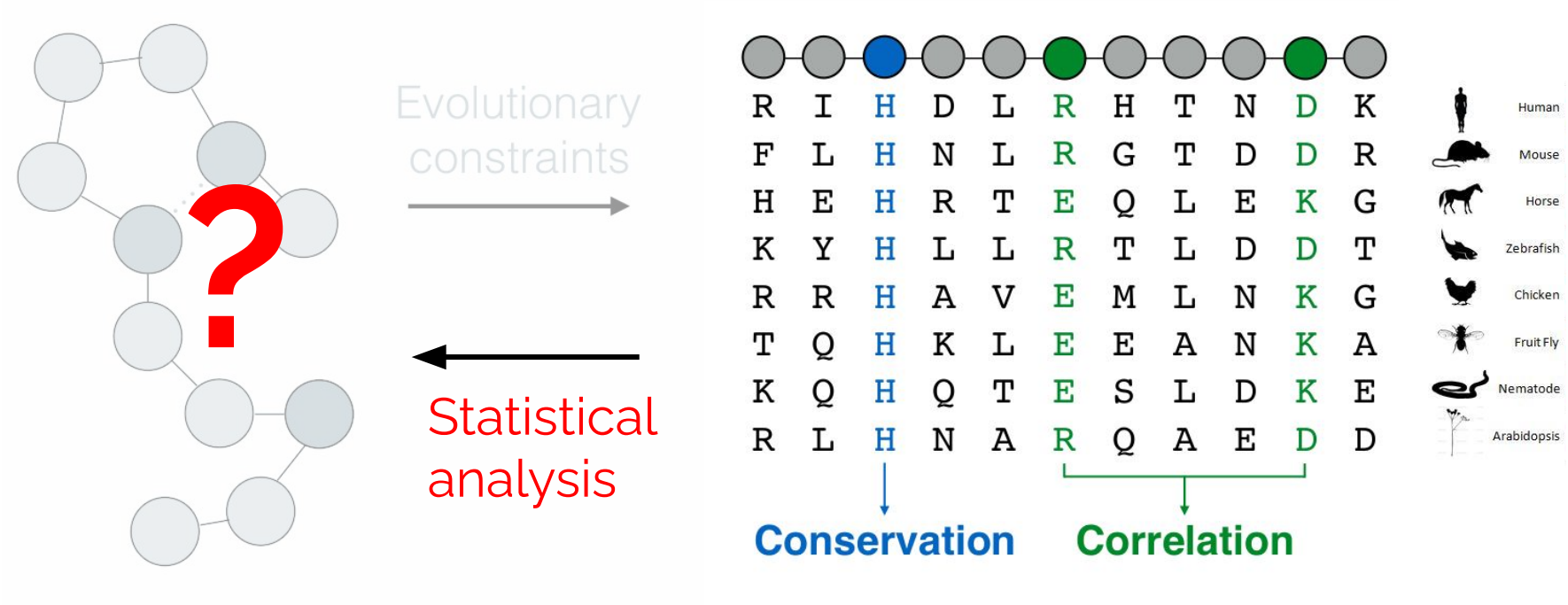


Evolutionary constraints →




- Functionally or structurally **important residues -> conservation** in the MSA
- **Epistatic interactions** between residues **-> correlation** in the MSA

# Direct Coupling Analysis (DCA)



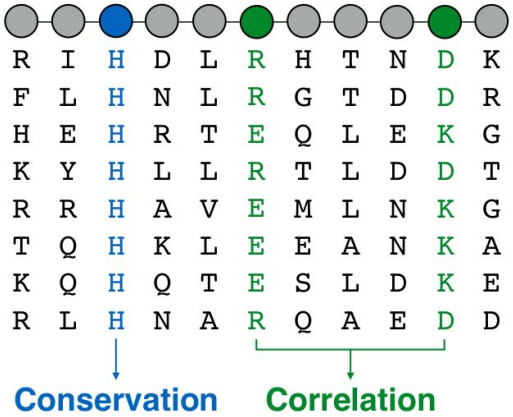

DCA: exploiting the **statistical patterns** of the MSA to computationally characterize the protein

# Direct Coupling Analysis (DCA)

[Weigt et al., PNAS 2009]

$$P(a_1, \ldots, a_N) = \frac{1}{Z} \exp\left( \sum_{i<j}^{N} \underbrace{J_{ij}(a_i, a_j)}_{\text{couplings}} + \sum_{i=1}^{N} \underbrace{h_i(a_i)}_{\text{fields}} \right)$$

**Direct Coupling Analysis (DCA)**




$f_i(a)$ $f_{ij}(a, b)$

- **Sequence of the MSA:** results of a sampling of an unknown probability distribution $P(a_1, \ldots, a_N)$
- **Inference**: fit $J$ and $h$ such that

from the model

$$P_i(a) = f_i(a)$$

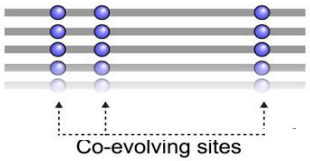
$$P_{ij}(a, b) = f_{ij}(a, b)$$

from the MSA

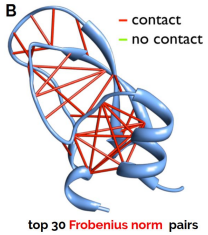
- Use it to **infer the phenotype**

# Direct Coupling Analysis (DCA): some applications

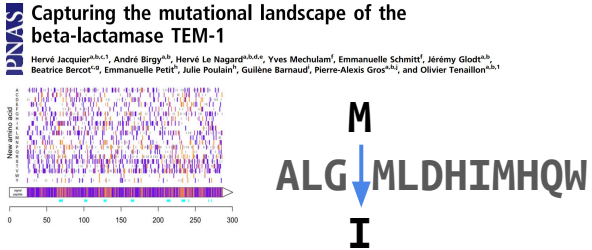
## Contact predictions




$$P(a_1,\ldots,a_N) = \frac{1}{Z}\exp\left(\sum_{i<j}^{N} J_{ij}(a_i,a_j) + \sum_{i=1}^{N} h_i(a_i)\right) \rightarrow F_{ij} = \sqrt{\sum_{a,b} J_{ij}(a,b)^2}$$

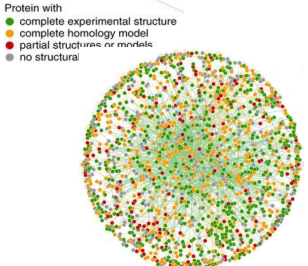


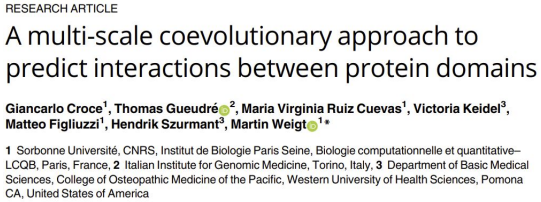

## Predict Mutational effect




## Generate new functional proteins

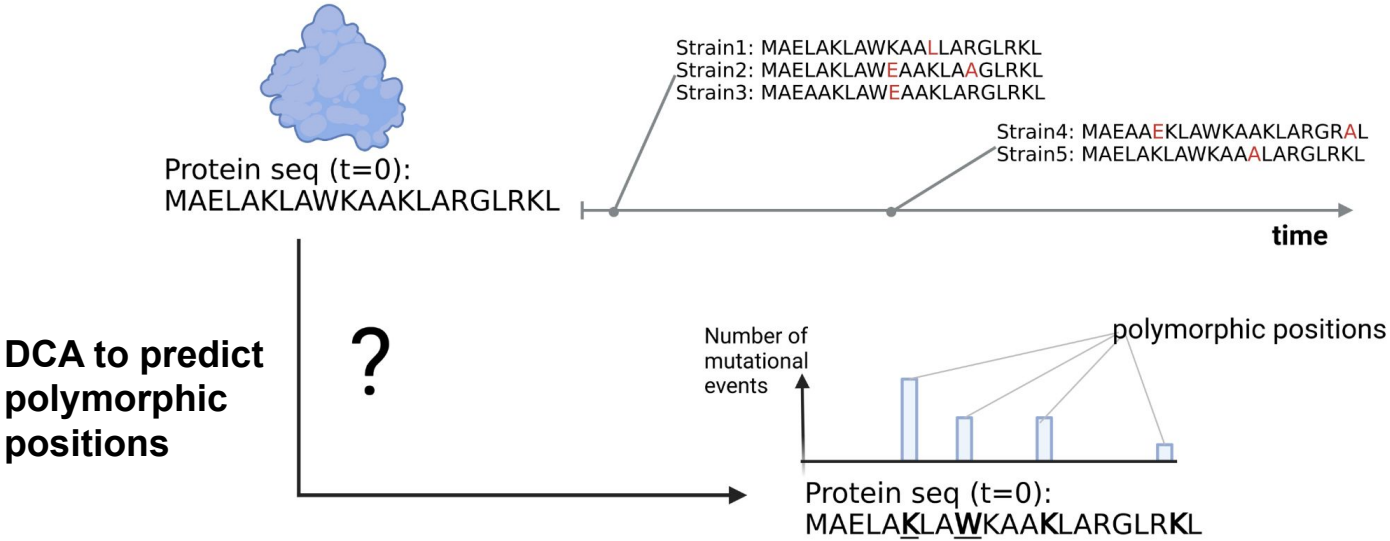

## Predicting Protein-protein interaction network




and many others...

# Can we use Direct Coupling Analysis (DCA) to model and predict protein evolution?



# DCA to model and predict protein evolution: SARS-CoV-2

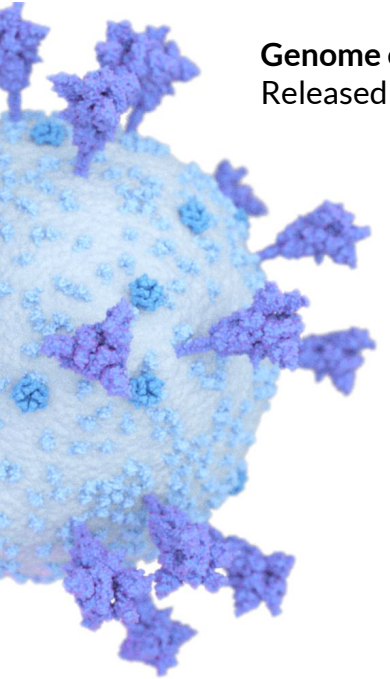
RESEARCH ARTICLE | BIOPHYSICS AND COMPUTATIONAL BIOLOGY

## Epistatic models predict mutable sites in SARS-CoV-2 proteins and epitopes

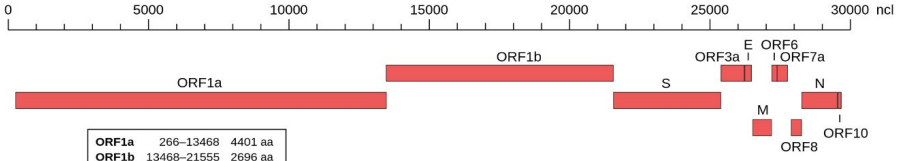
Juan Rodriguez-Rivas, Giancarlo Croce, Maureen Muscat, and Martin Weigt Authors Info & Affiliations

Edited by John Barton, Physics and Astronomy, University of California, Riverside, CA; received July 16, 2021; accepted December 13, 2021 by Editorial Board Member Mehran Kardar

**January 12, 2022** | 119 (4) e2113118119 | https://doi.org/10.1073/pnas.2113118119



**Genome of the first SARS-CoV-2 strain - Wuhan-Hu-1**
Released on Dec 30, 2019




| | | |
|---|---|---|
| **ORF1a** | 266–13468 | 4401 aa |
| **ORF1b** | 13468–21555 | 2696 aa |
| **S** | 21563–25384 | 1273 aa |
| **ORF3A** | 25393–26220 | 275 aa |
| **E** | 26245–26472 | 75 aa |
| **M** | 26523–27191 | 222 aa |
| **ORF6** | 27202–27387 | 61 aa |
| **ORF7a** | 27394–27759 | 121 aa |
| **ORF8** | 27894–28259 | 121 aa |
| **N** | 28274–29533 | 419 aa |
| **ORF10** | 29558–29674 | 38 aa |

**Wuhan-Hu-1 (GenBank MN908947.3)**

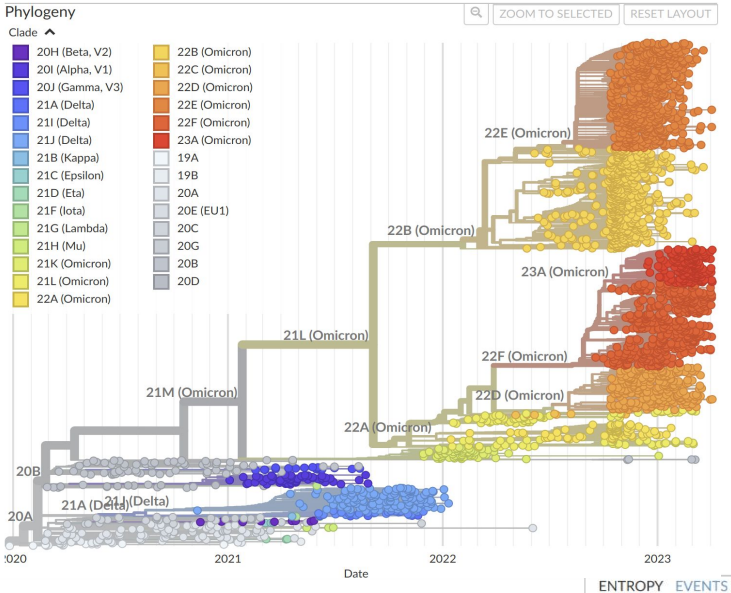
**29 proteins** including 4 structural, 16 nonstructural, and 9 accessory proteins

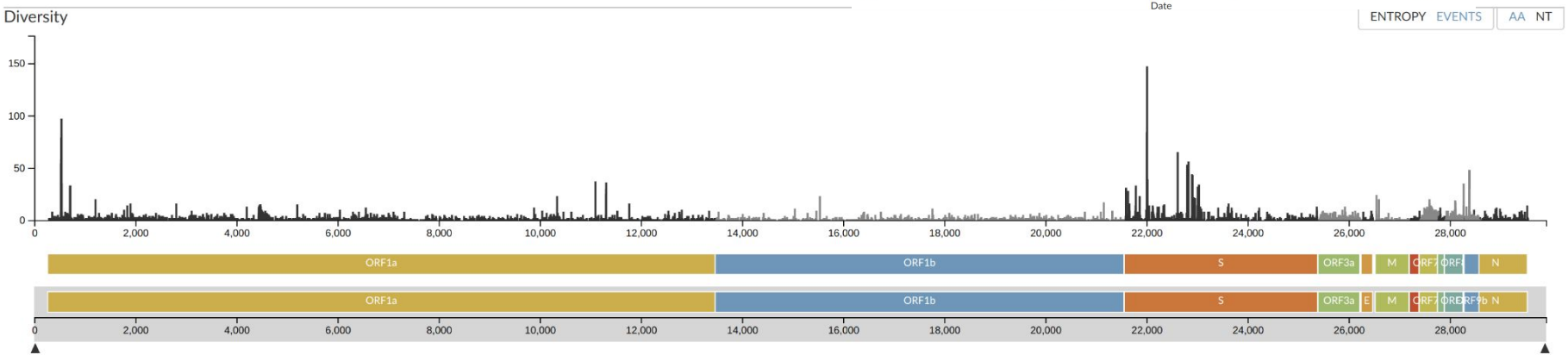
# DCA to model and predict protein evolution: SARS-CoV-2

Nextstrain

Real-time tracking of pathogen evolution

**Nextstrain**: Showing 2810 genomes sampled between Dec 2019 and Apr 2023



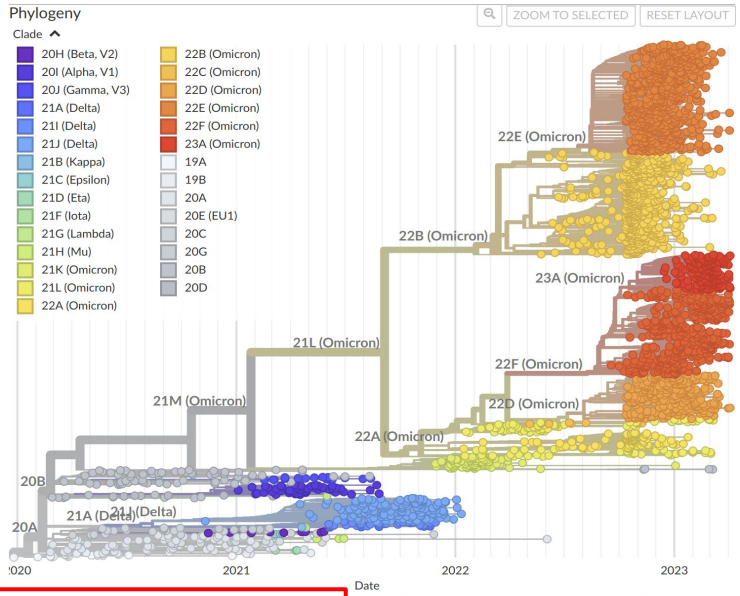



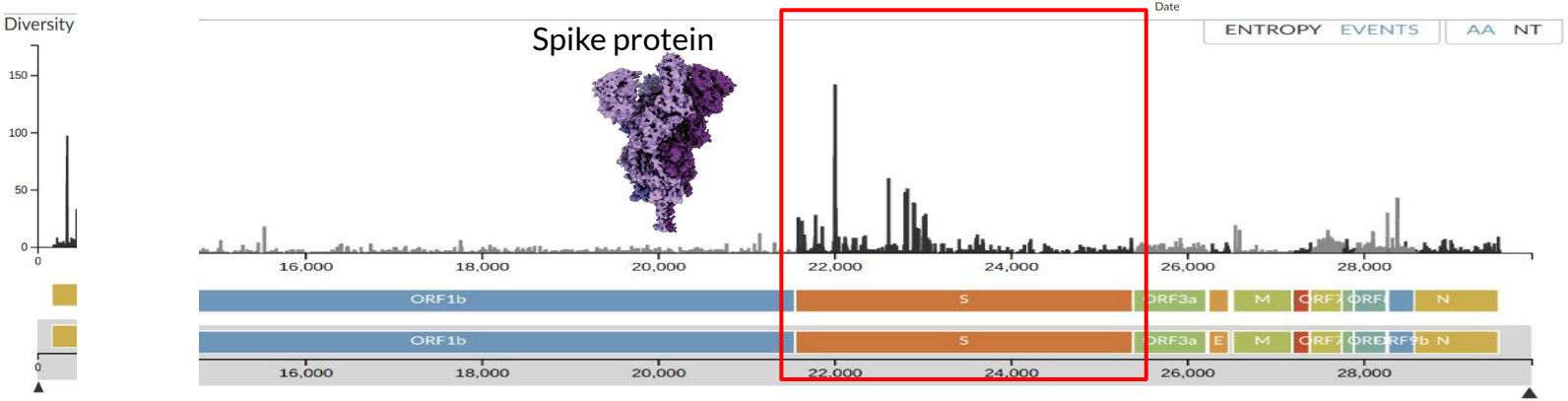
# DCA to model and predict protein evolution: SARS-CoV-2

Nextstrain

Real-time tracking of pathogen evolution

**Nextstrain**: Showing 2810 genomes sampled between Dec 2019 and Apr 2023



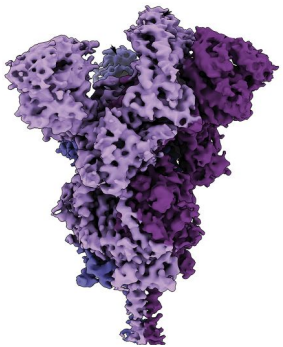
# DCA to model and predict protein evolution: SARS-CoV-2

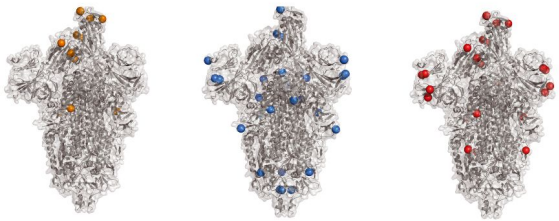
**Wuhan-Hu-1**
(30 Dec 2019)

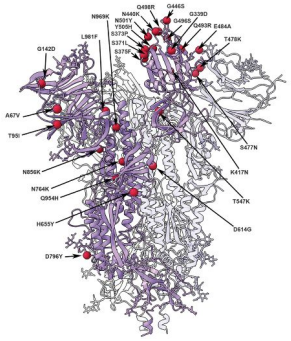
$t_0$

~ 3 years of evolution

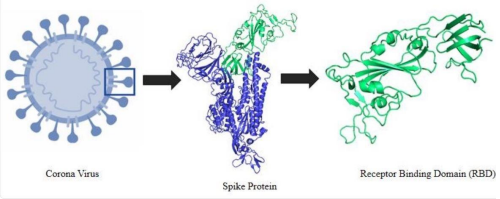
t








**Receptor binding domain (RBD)**



# DCA to model and predict protein evolution: SARS-CoV-2

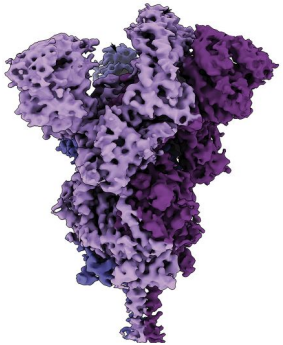
**Can we anticipate which positions are more likely to be polymorphic?**

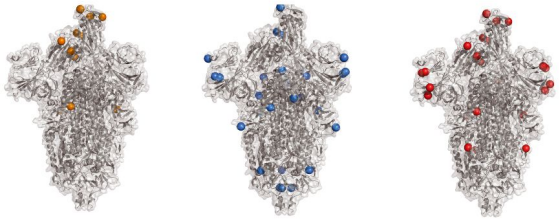
**Wuhan-Hu-1**
(30 Dec 2019)

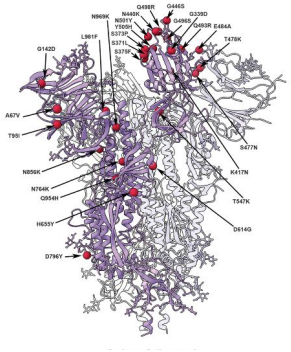
$t_0$

~ 3 years of evolution

t

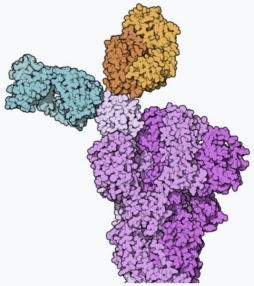




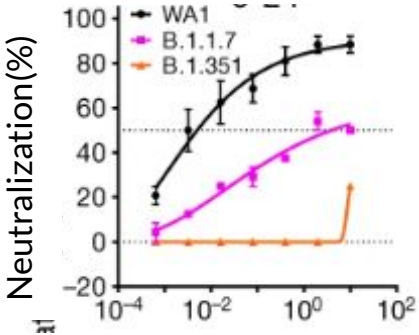


Monoclonal Antibody

**Casirivimab/imdevimab**



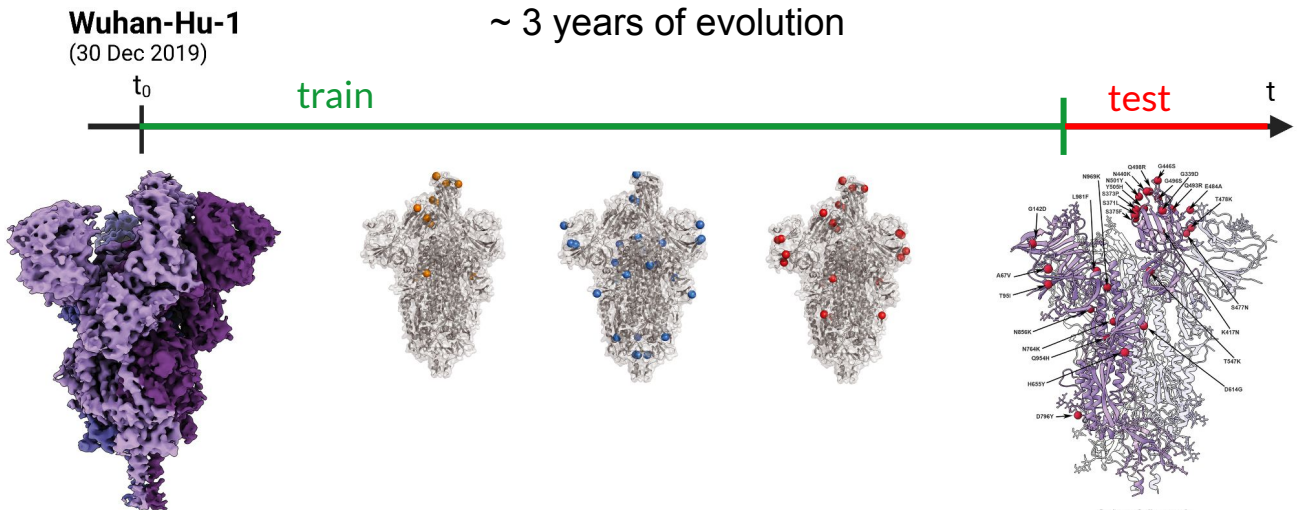
REGN10933 (blue) and REGN10987 (orange) bound to SARS-CoV-2 spike protein (pink).
From PDB: 6VSB, 6XDG.




Wang, Pengfei, et al. "Antibody resistance of SARS-CoV-2 variants B. 1.351 and B. 1.1. 7." *Nature*, 2021

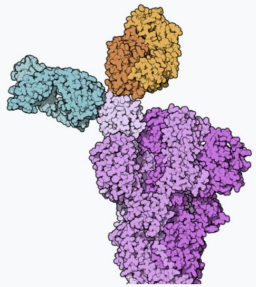
# DCA to model and predict protein evolution: SARS-CoV-2

**Can we anticipate which positions are more likely to be polymorphic?**




Monoclonal Antibody

**Casirivimab/imdevimab**



REGN10933 (blue) and REGN10987 (orange) bound to SARS-CoV-2 spike protein (pink). From PDB: 6VSB, 6XDG.

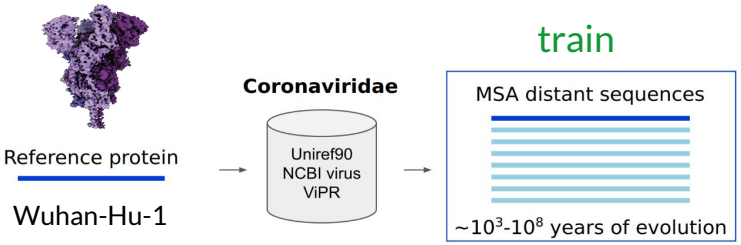

[Hie, Brian, et al. "Learning the language of viral evolution and escape. *Science* (2021)]
[Maher, M. Cyrus, et al. "Predicting the mutational drivers of future SARS-CoV-2 variants of concern." *Science Translational Medicine* (2022]

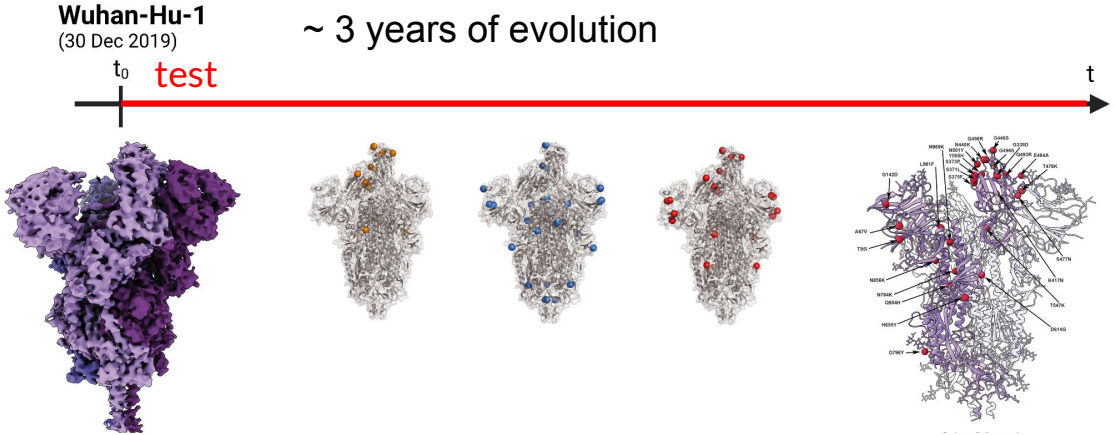
[Telenti, Amalio, Emma B. Hodcroft, and David L. Robertson. "The evolution and biology of SARS-CoV-2 variants." *Cold Spring Harbor perspectives in medicine* 12.5 (2022)]

# DCA to model and predict protein evolution: SARS-CoV-2

**Can we anticipate which positions are more likely to be polymorphic?**

**Learning from pre-pandemic data to anticipate polymorphic residues**

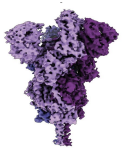



**Pros:** predictions rely exclusively on **data available at the day 0 of the outbreak**, and predictions can be **tested while more data accumulate**

**Cons**: We cannot capture effect specific for the SARS-CoV-2 - human interaction (ACE2 human receptor)

# DCA to model and predict protein evolution: SARS-CoV-2



Reference protein

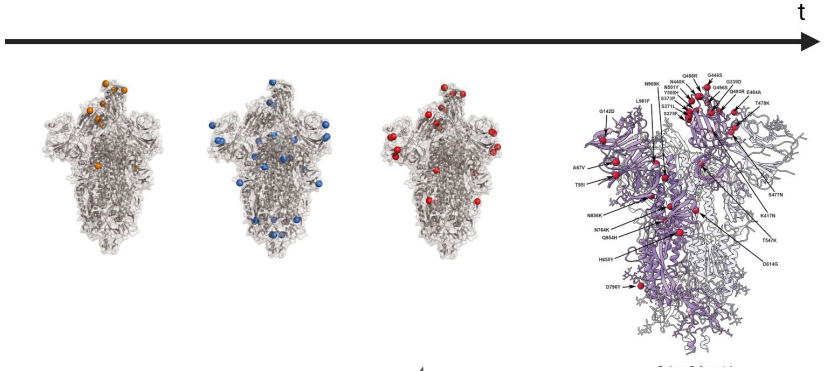
Wuhan-Hu-1

**Coronaviridae**

Uniref90
NCBI virus
ViPR

MSA distant sequences

~$10^3$-$10^8$ years of evolution




Train a Direct Coupling Analysis (DCA) model

$$P(a_1,\ldots,a_N) = \frac{1}{Z}\exp\left(\sum_{i<j}^{N} J_{ij}(a_i,a_j) + \sum_{i=1}^{N} h_i(a_i)\right)$$

We **do not predict which amino acid** is going to appear (**E**484**K**) next, but **which positions are more likely to accumulate mutations** (484 -> polymorphic)

Define a score to predict which positions are more likely to mutate

the effect of a single mutation ai→b can be computed as the difference between a wild-type sequence and single-mutant sequence:

$$\Delta E_{DCA}(i,b) = \log P_{DCA}(a_1,\ldots,a_i,\ldots,a_L) - \log P_{DCA}(a_1,\ldots,b,\ldots,a_L)$$

**Mutability score**

$$S_{DCA}(i) = \frac{1}{q}\sum_{k=1}^{q} \Delta E_{DCA}(i,b_k)$$

Position *i* - > high score -> likely to be polymorphic

# DCA to model and predict protein evolution: SARS-CoV-2

RESEARCH ARTICLE | BIOPHYSICS AND COMPUTATIONAL BIOLOGY |

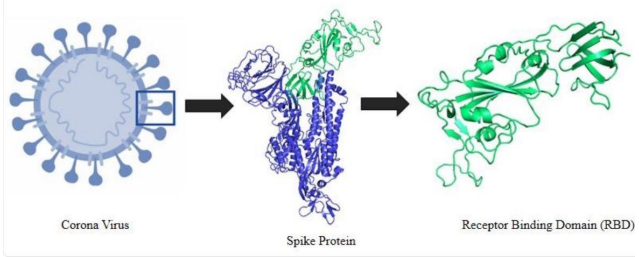
## Epistatic models predict mutable sites in SARS-CoV-2 proteins and epitopes

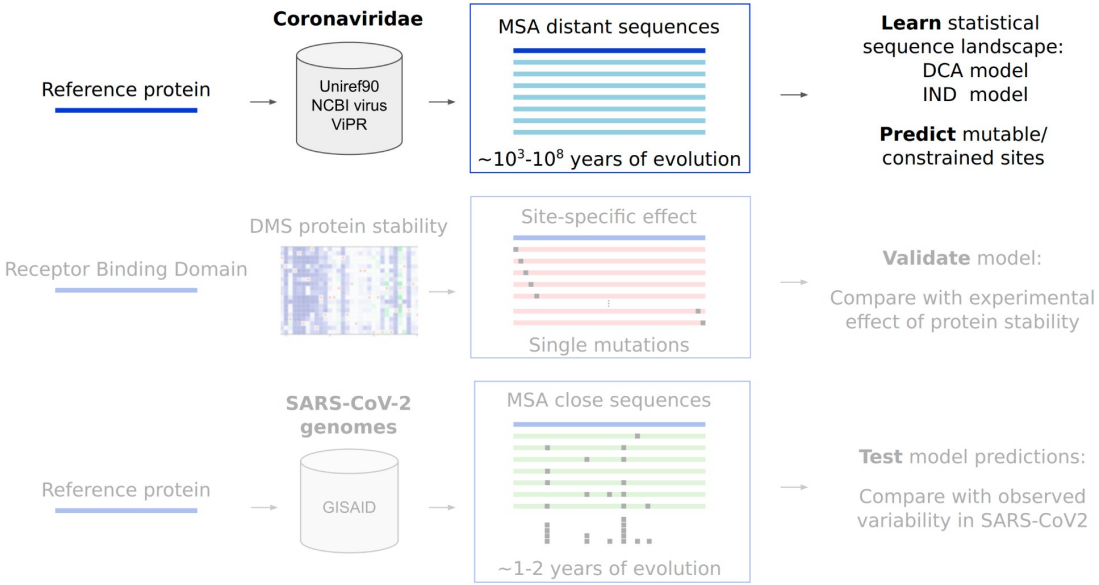
Juan Rodriguez-Rivas, Giancarlo Croce, Maureen Muscat, and Martin Weigt Authors Info & Affiliations

Edited by John Barton, Physics and Astronomy, University of California, Riverside, CA; received July 16, 2021; accepted December 13, 2021 by Editorial Board Member Mehran Kardar

**January 12, 2022** | 119 (4) e2113118119 | https://doi.org/10.1073/pnas.2113118119

**Receptor binding domain (RBD)**
*bCoV_S1_RBD (PF09408)*




**Independent model (IND):**
Baseline model (using only 1-point statistics - frequencies)

**Direct Coupling Analysis (DCA):**
1- and 2-point statistics (epistatic interactions)

# DCA to model and predict protein evolution: SARS-CoV-2

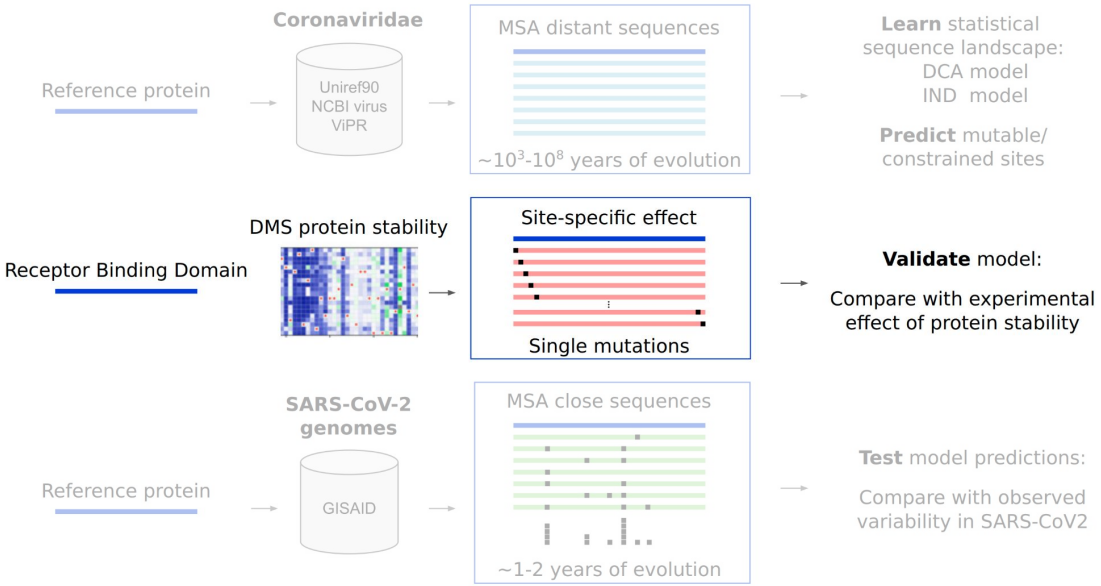
RESEARCH ARTICLE | BIOPHYSICS AND COMPUTATIONAL BIOLOGY |

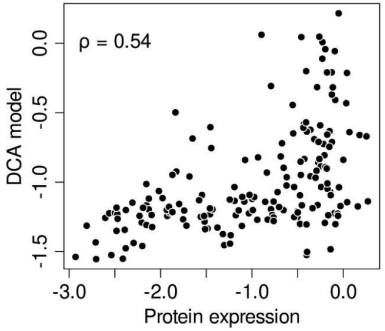
## Epistatic models predict mutable sites in SARS-CoV-2 proteins and epitopes

Juan Rodriguez-Rivas, Giancarlo Croce, Maureen Muscat, and Martin Weigt Authors Info & Affiliations

Edited by John Barton, Physics and Astronomy, University of California, Riverside, CA; received July 16, 2021; accepted December 13, 2021 by Editorial Board Member Mehran Kardar

**January 12, 2022** | 119 (4) e2113118119 | https://doi.org/10.1073/pnas.2113118119

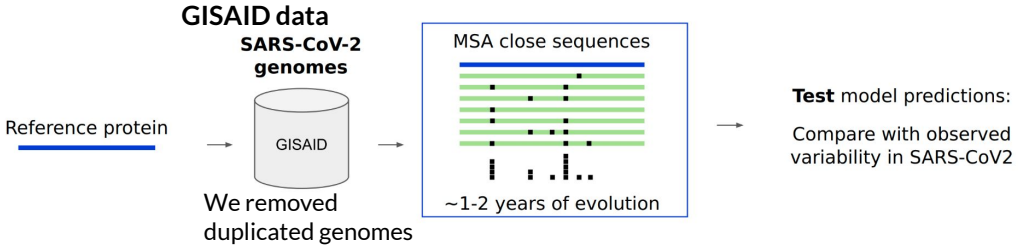



DMS for protein expression data from Bloom's lab

# DCA to model and predict protein evolution: SARS-CoV-2

**Can we anticipate which positions are more likely to be polymorphic?**



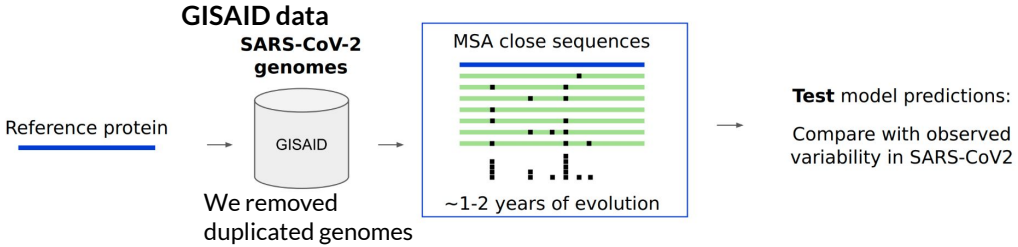

From GISAID data, for each position *i* in the RBD
0 - **constrained** (no mutations)
1 - **mutable** (*x* mutational events)

```
Strain1: MAELAKLAWKAAKLARGLRKL
Strain2: MAELAKLAWEAAKKARGLRKL
Strain3: MAEAAKLAWEAAKLARGLRKL
Strain4: MAEAAKLAWKAAKLARGLRKL
Strain5: MAELAKLAWKAAKLARGLRKL
Pos:     (1,2,3,4,5,6,7,8,9,10,..)
test_set:(0,0,0,1,0,0,0,0,0,..)
dca_pred:(0.2,0.6,0.1,0.9,0.2,0.1,..)
```

**May 2021, 3,883 genomes:** no mutational event has occurred for 58% of the entire proteome, while only 14% has experienced more than two events

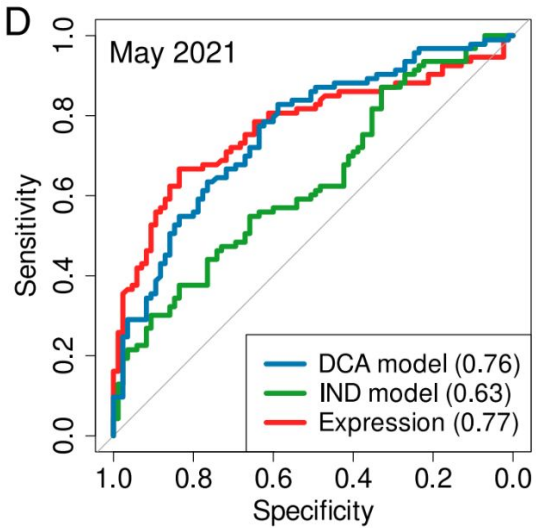
# DCA to model and predict protein evolution: SARS-CoV-2

**Can we anticipate which positions are more likely to be polymorphic?**




From GISAID data, for each position *i* in the RBD
0 - **constrained** (no mutations)
1 - **mutable** (*x* mutational events)

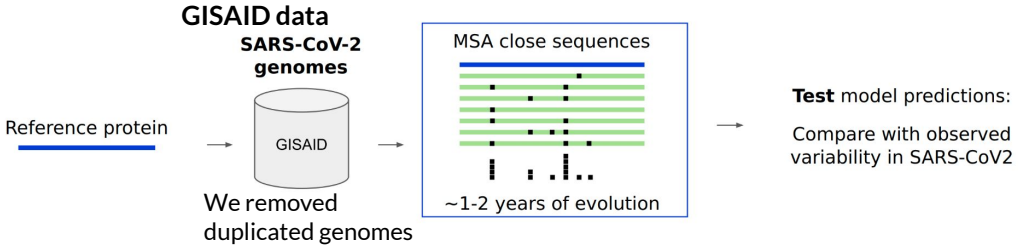
```
Strain1: MAELAKLAWKAAKLARGLRKL
Strain2: MAELAKLAWEAAKKARGLRKL
Strain3: MAEAAKLAWEAAKLARGLRKL
Strain4: MAEAAKLAWKAAKLARGLRKL
Strain5: MAELAKLAWKAAKLARGLRKL
Pos:     (1,2,3,4,5,6,7,8,9,10,..)
test_set:(0,0,0,1,0,0,0,0,0,1,..)
dca_pred:(0.2,0.6,0.1,0.9,0.2,0.1,..)
```




**(Bloom's lab, Deep mutational scans)**

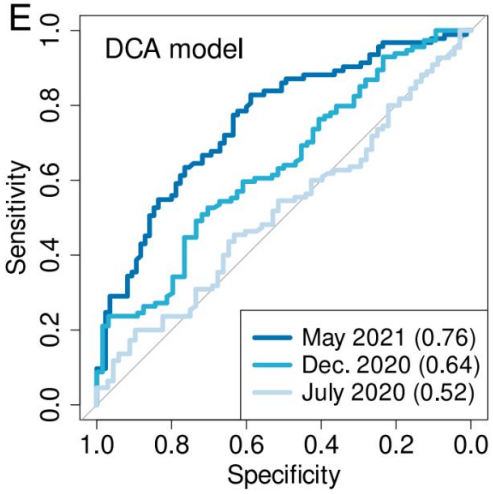
# DCA to model and predict protein evolution: SARS-CoV-2

**Can we anticipate which positions are more likely to be polymorphic?**




```
Strain1:  MAELAKLAWKAAKLARGLRKL
Strain2:  MAELAKLAWEAAKKARGLRKL
Strain3:  MAEAAKLAWEAAKLERGLAEL
Strain4:  MAEAAKLAWKAAKLKRGLRKL
Strain5:  MAELAKLAWKAAKLARGLRKL
Strain6:  MAEKAKLAWKAAKLARGLRKL
Strain7:  MAELAKLAWKAAKLARELRKL
Strain8:  MAEAAKLAWKAAKLARGLRKL
Strain9:  MAEEAKLAWKAAKLARGKRKE
Strain10: MAEAAKLAWKAAKLARGLRKL
Strain11: MAELAKLAWKAAKLARKLRKL
Strain12: MAELAKLAWKAAKLELGLRKL
```

From GISAID data, for each position *i* in the RBD
0 - **constrained** (no mutations)
1 - **mutable** (*x* mutational events)




**More data -> more polymorphic positions in the test set**

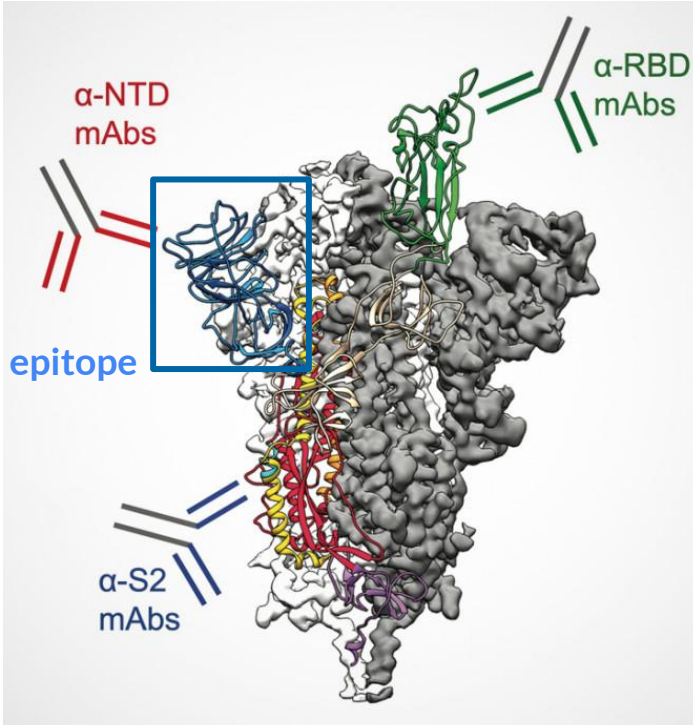
**AUC increases over time (virus has explored more variants = better test set)**

DCA can anticipate which positions will mutate in the future

# DCA to model and predict protein evolution: SARS-CoV-2

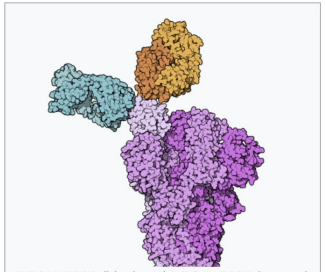
Not all positions are equally important.

**Mutations in B/T cells epitopes can negatively affect the human immune response => more dangerous**




Mutations in B and T cells epitopes
-> not binding antibodies or T cells

**Casirivimab/imdevimab**



REGN10933 (blue) and REGN10987 (orange) bound to SARS-CoV-2 spike protein (pink). From PDB: 6VSB, 6XDG.

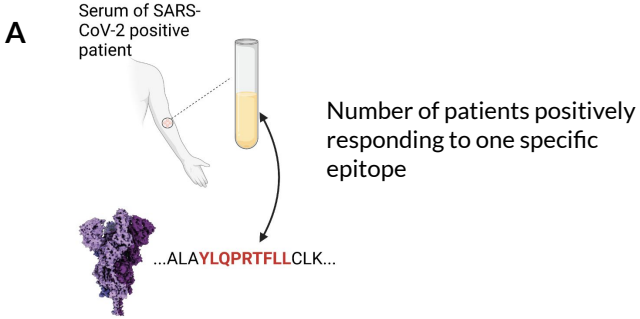
**Immunologically relevant positions**

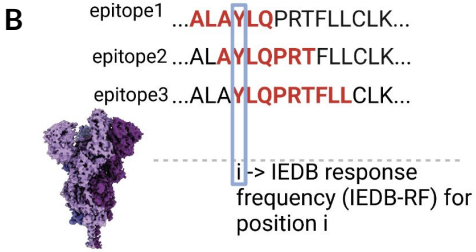
Database of experimentally validated B and T cells epitopes (IEDB)

# DCA to model and predict protein evolution: SARS-CoV-2

Not all positions are equally important.
**Mutations in B/T cells epitopes can negatively affect the human immune response => more dangerous**


## IEDB response frequency (IEDB-RF)



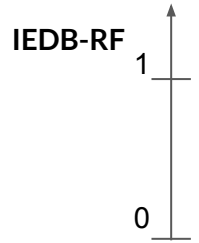




Number of epitopes that share a specific position in the SARS-CoV-2 proteome

**IEDB Response Frequency (IEDB-RF)**
the number of *positively responding subjects* relative to the total number of those tested, *averaged over all epitopes mapped to that position*

and Confidence Interval (C.I.)

```
Strain1: MAELAKLAWKAAKLARGLRKL
Strain2: MAELAKLAWEAAKLARGLRKL
Strain3: MAEAAKLAWEAAKLARGLRKL
Strain4: MAEAAKLAWKAAKLARGLRKL
Strain5: MAELAKLAWKAAKLARGLRKL
Pos:     (1,2,3,4,5,6,7,8,9,10,..)
dca_pred:(0.2,0.6,0.1,0.9,0.2,0.1,..)
IEDB_RF:(0.2±0.1,0.8±0.2,0.2±0.01, …)
```

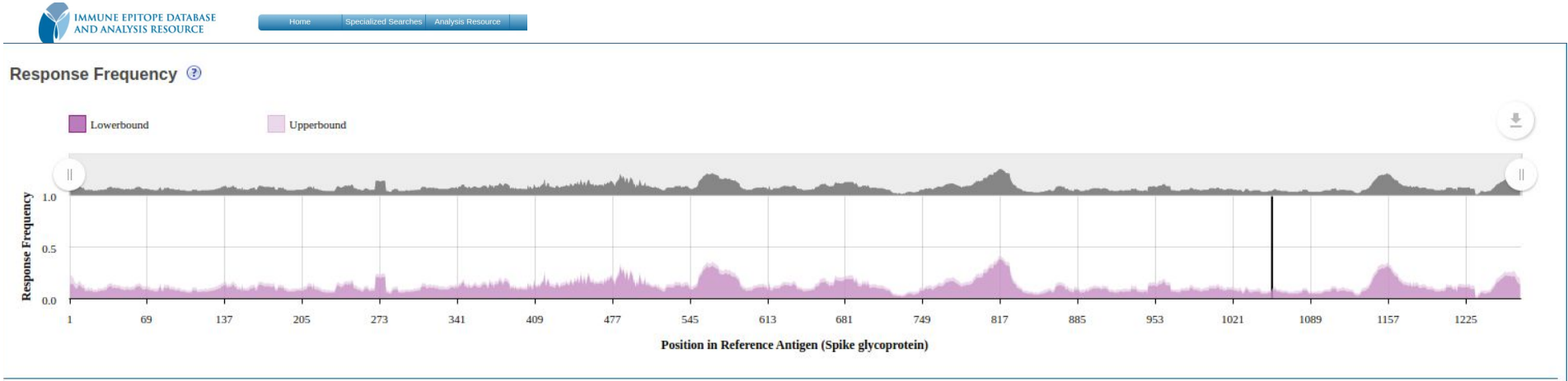



Highly immunogenic position -> if mutated it alters many positively responding B/T epitopes

Low immunogenic position -> not targeted by the human immune system

# DCA to model and predict protein evolution: SARS-CoV-2

Not all positions are equally important.
**Mutations in B/T cells epitopes can negatively affect the human immune response => more dangerous**
Can we predict which **immunologically relevant positions** are more likely to be **polymorphic**?




**IEDB Response Frequency (IEDB-RF)**
the number of *positively responding subjects* relative to the total number of those tested, *averaged over all epitopes mapped to that position*

and Confidence Interval (C.I.)

```
Strain1: MAELAKLAWKAAKLARGLRKL
Strain2: MAELAKLAWEAAKLARGLRKL
Strain3: MAEAAKLAWEAAKLARGLRKL
Strain4: MAEAAKLAWKAAKLARGLRKL
Strain5: MAELAKLAWKAAKLARGLRKL
Pos:     (1,2,3,4,5,6,7,8,9,10,..)
dca_pred:(0.2,0.6,0.1,0.9,0.2,0.1,..)
IEDB_RF:(0.2±0.1,0.8±0.2,0.2±0.01, …)
```

**IEDB-RF**

1 — Highly immunogenic position -> if mutated it alters many positively responding B/T epitopes

0 — Low immunogenic position -> not targeted by the human immune system

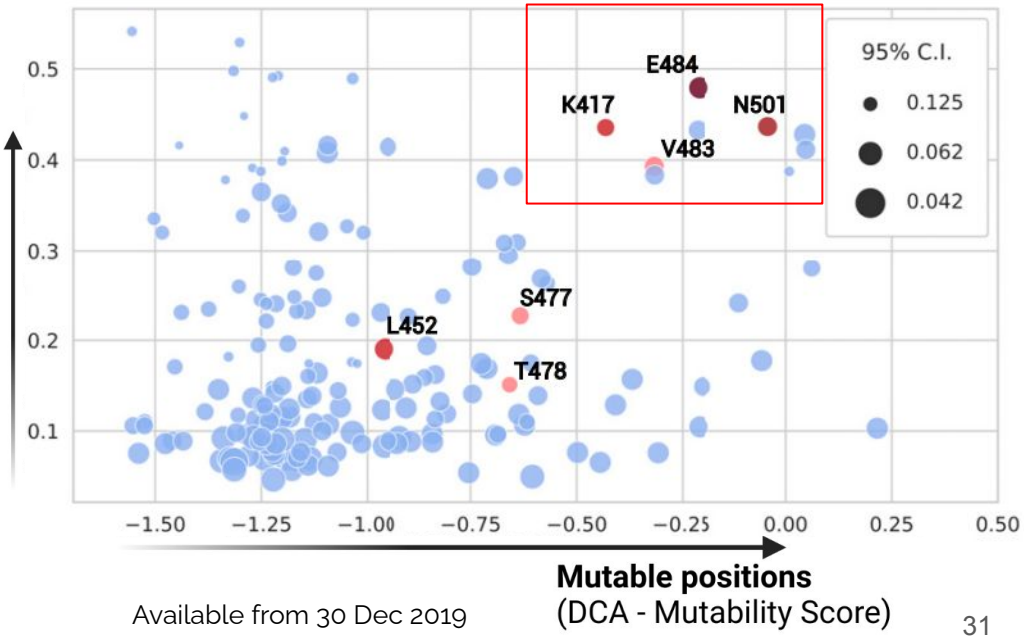
# DCA to model and predict protein evolution: SARS-CoV-2

Not all positions are equally important.
**Mutations in B/T cells epitopes can negatively affect the human immune response => more dangerous**
Can we predict which **immunologically relevant positions** are more likely to be **polymorphic**?


| Pango Lineage | Mutated positions - RBD domain | | | | | | |
|---|---|---|---|---|---|---|---|
| | 417 | 452 | 477 | 478 | 483 | 484 | 501 |
| Wuhan-Hu-1 | K | L | S | T | V | E | N |
| B.1.1.7 | • | • | • | • | • | • | Y |
| B.1.351 | N | • | • | • | • | K | Y |
| B.1.1.28.1, alias P.1 | T | • | • | • | • | K | Y |
| B.1.617 | • | R | • | K* | • | Q* | • |
| B.1.525 | • | • | • | • | • | K | • |
| B.1.427/9 | • | R | • | • | • | • | • |
| B.1.1.28.2, alias P.2 | • | • | • | • | • | K | • |
| B.1.1.28.3, alias P.3 | • | • | • | • | • | K | Y |
| B.1.526 | • | • | N* | • | • | K* | • |
| B.1.616 | • | • | • | • | A | • | • |

**Immunologically relevant positions**
(IEDB - Response Frequency)

(experimental results, May 2021)




Available from 30 Dec 2019

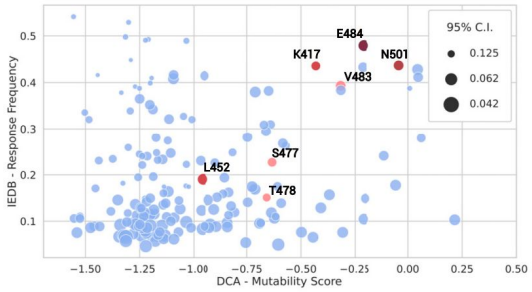
# DCA to model and predict protein evolution: SARS-CoV-2

Not all positions are equally important.
**Mutations in B/T cells epitopes can negatively affect the human immune response => more dangerous**
Can we predict which **immunologically relevant positions** are more likely to be **polymorphic**?

**Table 1. The first 20 predictions, sorted according to the DCA mutability score, with the corresponding IEDB RF and the VOIs and VOIs in which the position has mutated**

| Position | AA Wuhan-Hu-1 | DCA mutability score | IEDB RF (95% CI) | Pango lineage (ref. 38) |
|---|---|---|---|---|
| 519 | H | 0.22 | 0.10 (0.08:0.14) | |
| 403 | R | 0.06 | 0.28 (0.24:0.32) | |
| **490** | **F** | **0.05** | **0.41 (0.38:0.45)** | |
| **493** | **Q** | **0.04** | **0.43 (0.40:0.46)** | |
| **372** | **A** | **0.01** | **0.39 (0.32:0.46)** | |
| **501** | **N** | **−0.04** | **0.44 (0.40:0.47)** | **B.1.1.7; B.1.351; P.1; P.3** |
| 445 | V | −0.06 | 0.18 (0.15:0.21) | |
| 498 | Q | −0.11 | 0.24 (0.21:0.28) | |
| 441 | L | −0.20 | 0.15 (0.12:0.19) | |
| 440 | N | −0.21 | 0.10 (0.08:0.14) | |
| **484** | **E** | **−0.21** | **0.48 (0.45:0.51)** | **B.1.351; P.1; B.1.617; B.1.525; P.2; P.3** |
| **486** | **F** | **−0.21** | **0.43 (0.40:0.47)** | |
| 443 | S | −0.31 | 0.08 (0.05:0.11) | |
| **494** | **S** | **−0.32** | **0.38 (0.35:0.42)** | |
| **483** | **V** | **−0.32** | **0.39 (0.36:0.43)** | **B.1.616** |
| 460 | N | −0.37 | 0.16 (0.13:0.19) | |
| 444 | K | −0.41 | 0.13 (0.10:0.16) | |
| **417** | **K** | **−0.43** | **0.44 (0.40:0.48)** | **B.1.351; P.1** |
| 439 | N | −0.44 | 0.07 (0.04:0.10) | |
| 402 | I | −0.50 | 0.08 (0.05:0.11) | |

Positions with IEDB RF above 0.3 are shown in bold.

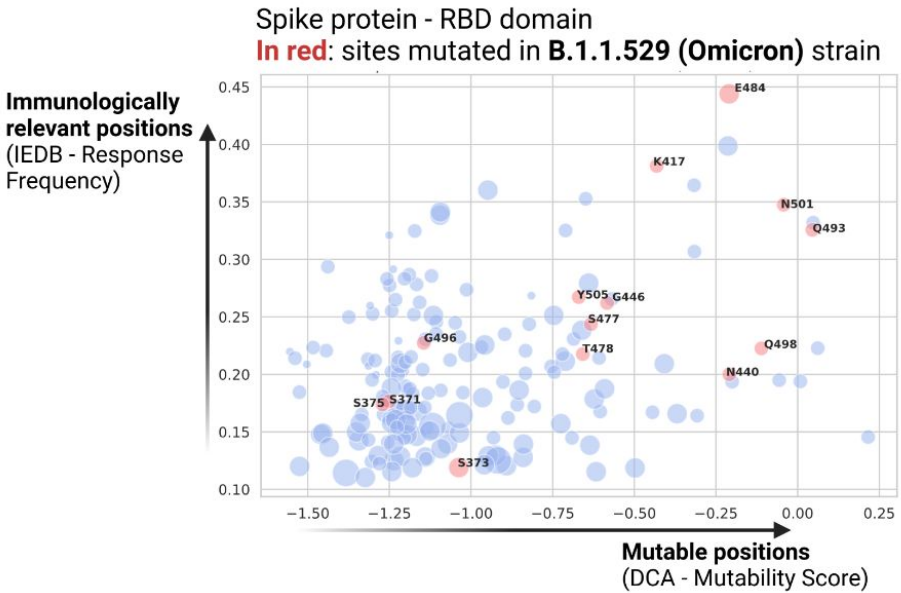


# DCA to model and predict protein evolution: SARS-CoV-2

**Table 1. The first 20 predictions, sorted according to the DCA mutability score, with the corresponding IEDB RF and the VOIs and VOIs in which the position has mutated**

| Position | AA Wuhan-Hu-1 | DCA mutability score | IEDB RF (95% CI) | Pango lineage |
|---|---|---|---|---|
| 519 | H | 0.22 | 0.10 (0.08:0.14) | |
| 403 | R | 0.06 | 0.28 (0.24:0.32) | |
| **490** | **F** | **0.05** | **0.41 (0.38:0.45)** | |
| **493** | **Q** | **0.04** | **0.43 (0.40:0.46)** | |
| **372** | **A** | **0.01** | **0.39 (0.32:0.46)** | |
| **501** | **N** | **−0.04** | **0.44 (0.40:0.47)** | **B.1.1.7; B.1.351** |
| 445 | V | −0.06 | 0.18 (0.15:0.21) | |
| 498 | Q | −0.11 | 0.24 (0.21:0.28) | |
| 441 | L | −0.20 | 0.15 (0.12:0.19) | |
| 440 | N | −0.21 | 0.10 (0.08:0.14) | |
| **484** | **E** | **−0.21** | **0.48 (0.45:0.51)** | **B.1.351; P.1; B.1.617;** |
| **486** | **F** | **−0.21** | **0.43 (0.40:0.47)** | |
| 443 | S | −0.31 | 0.08 (0.05:0.11) | |
| **494** | **S** | **−0.32** | **0.38 (0.35:0.42)** | |
| **483** | **V** | **−0.32** | **0.39 (0.36:0.43)** | **B.1.616** |
| 460 | N | −0.37 | 0.16 (0.13:0.19) | |
| 444 | K | −0.41 | 0.13 (0.10:0.16) | |
| **417** | **K** | **−0.43** | **0.44 (0.40:0.48)** | **B.1.351;** |
| 439 | N | −0.44 | 0.07 (0.04:0.10) | |
| 402 | I | −0.50 | 0.08 (0.05:0.11) | |

Positions with IEDB RF above 0.3 are shown in bold.




Positions mutated in variants of concern after submission (data from https://covariants.org/shared-mutations, 17 Apr 2023)

| 20I (Alpha, V1) (B.1.1.7) | 20H (Beta, V2) (B.1.351) | 20J (Gamma, V3) (P.1) | 21A (Delta) (B.1.617.2) | 21K (Omicron) (BA.1) | 21L (Omicron) (BA.2) | 22A & 22B (Omicron) (BA.4&5) | 22C (Omicron) (BA.2.12.1) | 22D (Omicron) (BA.2.75) | 22E (Omicron) (BQ.1) | 22F (Omicron) (XBB) | 23A (Omicron) (XBB.1.5) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Shared mutations | | | | | | | | | | | |

# DCA to model and predict protein evolution: SARS-CoV-2

Github page (and Google Colab) to reproduce the results

README.md

## Epistatic models predict mutable sites in SARS-CoV-2 proteins and epitopes

GitHub repository accompanying the paper Epistatic models predict mutable sites in SARS-CoV-2 proteins and epitopes Rodriguez-Rivas, Giancarlo Croce, Maureen Muscat, Martin Weigt, PNAS January 25, 2022 119 (4) e2113118119; https://doi.org/10.1073/pnas.2113118119.

**Epistatic models predict mutable sites in SARS-CoV-2 proteins and epitopes**

For each site of the SARS-CoV-2 Wuhan-Hu-1 proteome (Accession NC_045512) included in a PFAM domain, we introduce a Direct Coupling Analysis (DCA) **mutability score to predict mutable and constrained sites**.

Sites of the SARS-Cov-2 Wuhan-Hu-1 proteome with high *DCA-Mutability Scores* are more likely to mutate. We validate our mutability predictions with the mutations observed in SARS-CoV-2 proteomes deposited in the GISAID database.

**Packages**

No packages published
Publish your first package

**Environments** 1

github-pages Active

**Languages**

Jupyter Notebook 95.9% Python 4.0% Roff 0.1%

We collect updated data (novel mutations and most recent IEDB data)

dca_sarscov2.ipynb

File Edit View Insert Runtime Tools Help

+ Code + Text Copy to Drive

```
#################################
#Run this cell if using Google COLAB
#################################

#clone the repository and the data to run the notebook on Google Colab
!git clone https://github.com/GiancarloCroce/DCA_SARS-CoV-2
%cd ./DCA_SARS-CoV-2
#for plotly on google colab
import plotly.io as pio
pio.renderers.default = 'colab'
```

```
Cloning into 'DCA_SARS-CoV-2'...
remote: Enumerating objects: 441, done.
remote: Counting objects: 100% (26/26), done.
remote: Compressing objects: 100% (20/20), done.
^Cceiving objects:   4% (19/441), 28.90 MiB | 7.16 MiB/s
[Errno 2] No such file or directory: './DCA_SARS-CoV-2'
/home/giancarlo/Desktop/DCA_SARS-CoV-2
```

```
[ ] import pandas as pd
    import numpy as np
    import seaborn as sns
    import os
    import pandas as pd
    import datetime
    from sklearn import preprocessing
    import matplotlib.pyplot as plt
    from plotly.offline import init_notebook_mode, iplot
    from plotly.graph_objs import *
    import plotly.graph_objects as go
    from sklearn import metrics
    from utils import load_data_dca, plot_roc, plot_dca_IEDB, plot_dca_IEDB_BTcell
    from utils import get_IEDB_versions, compute_RF, get_updated_IEDB,load_VOC
```

## DCA for SARS-CoV-2

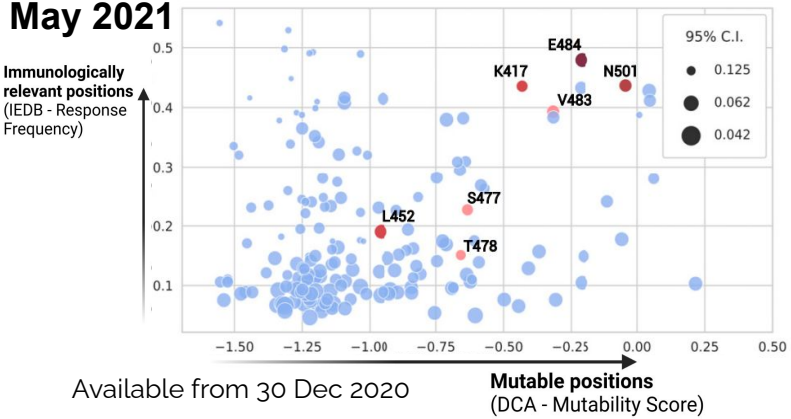
We introduce a DCA mutability score to predict mutable and constrained sites of the SARS-CoV-2 Wuhan-Hu-1 proteome (Accession NC045512). Only sites included in a PFAM domain are considered.Column: **mutability_score(DCA)**.

We also compute the mutability scores using non-epistatic conservation profiles (hereinafter independent models - IND). Column: **mutability_score(IND)**.

# DCA to model and predict protein evolution: SARS-CoV-2

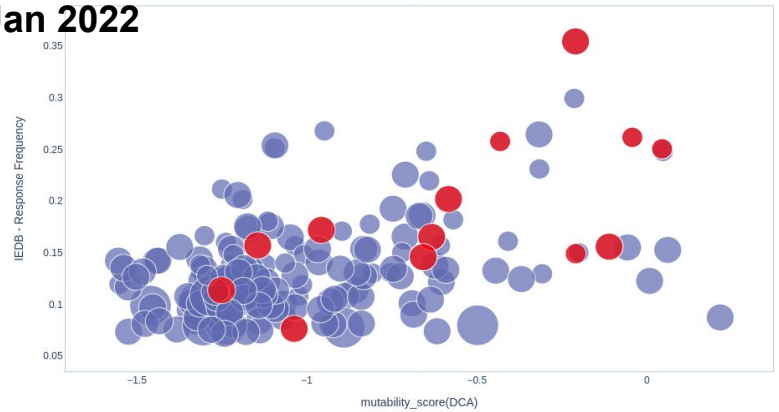
IEDB-DCA **Updated data** of predictions **polymorphic and immunologically relevant sites**

**May 2021**

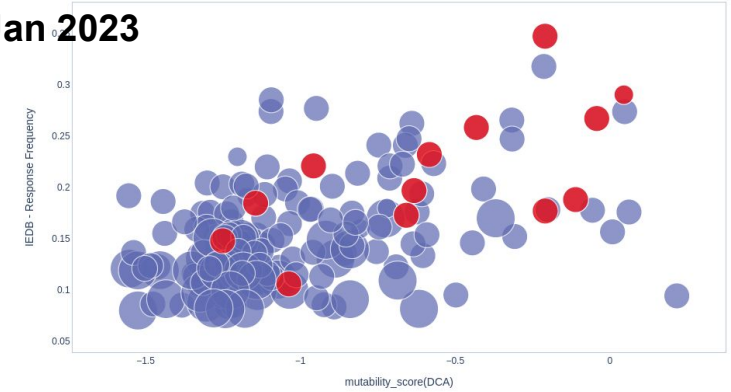



Available from 30 Dec 2020

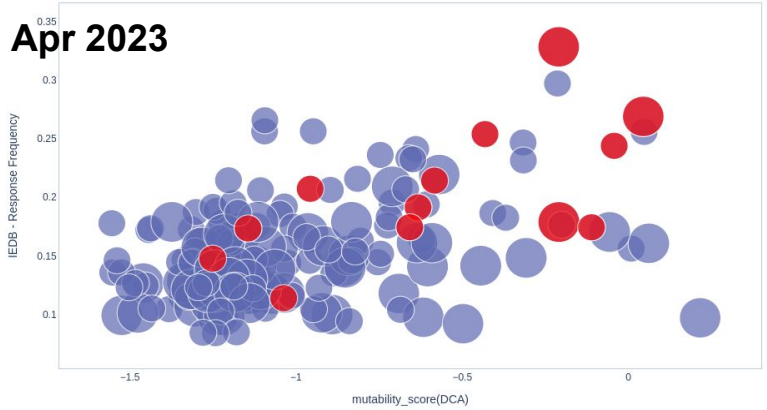
**Jan 2022**




**Jan 2023**

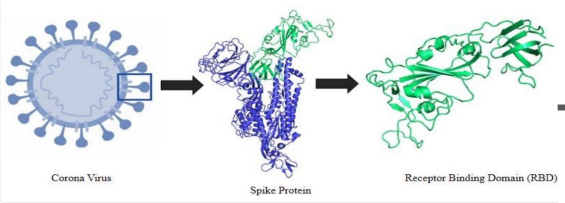


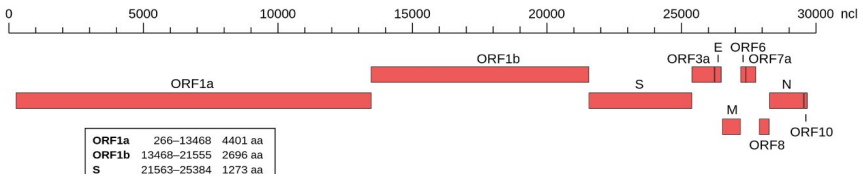

**Apr 2023**



# DCA to model and predict protein evolution: SARS-CoV-2

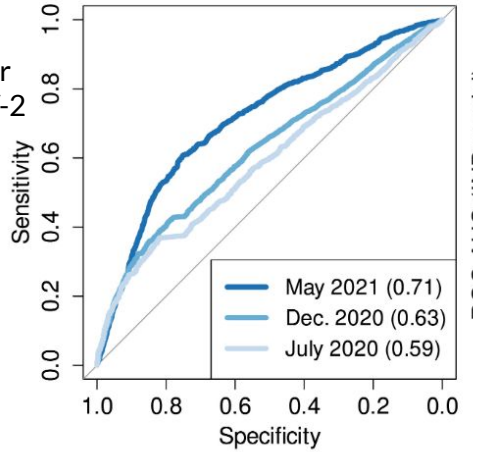
## From the RBD to the whole SARS-CoV-2 proteome



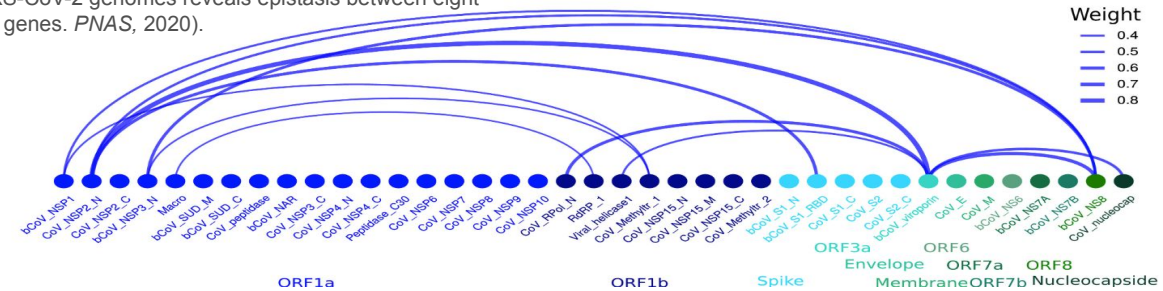

**Receptor binding domain (RBD)**
*bCoV_S1_RBD (PF09408)*

Similar trends for other SARS-CoV-2 domains



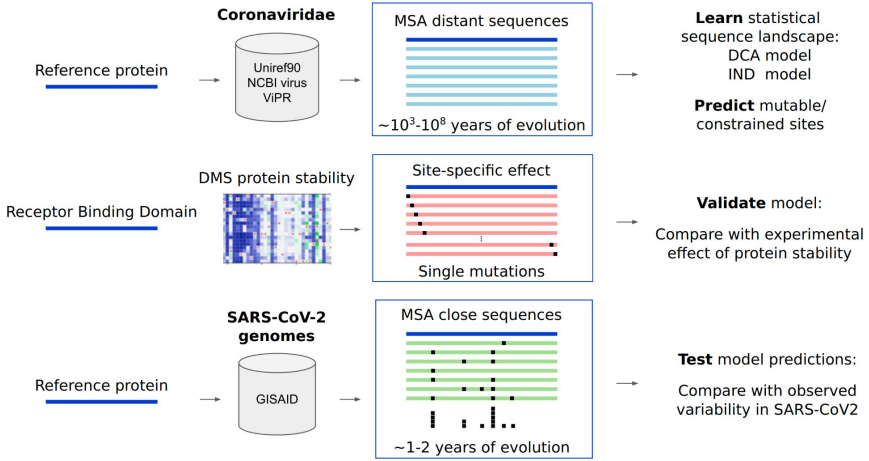

Inter-domain epistatic interactions

[H.-L. Zeng, et al., Global analysis of more than 50,000 SARS-CoV-2 genomes reveals epistasis between eight viral genes. *PNAS*, 2020).



# DCA to model and predict protein evolution: SARS-CoV-2

## Summary




DCA to **predict polymorphic positions**. Accuracies increases as more GISAID data accumulates

Not all positions are equally important. **Mutations in B/T cells epitopes are more dangerous.** We can predict which **immunologically relevant positions that are more likely to mutate**

# DCA to model and predict protein evolution: *E. coli*

RESEARCH ARTICLE | BIOPHYSICS AND COMPUTATIONAL BIOLOGY

Epistatic models predict mutable sites in SARS-CoV-2 proteins and epitopes

Juan Rodriguez-Rivas, Giancarlo Croce, Maureen Muscat, and Martin Weigt Authors Info & Affiliations

Edited by John Barton, Physics and Astronomy, University of California, Riverside, CA; received July 16, 2021; accepted December 13, 2021 by Editorial Board Member Mehran Kardar

January 12, 2022 | 119 (4) e2113118119 | https://doi.org/10.1073/pnas.2113118119

nature > nature communications > articles > article

Article | Open Access | Published: 12 July 2022

**Deciphering polymorphism in 61,157 *Escherichia coli* genomes via epistatic sequence landscapes**

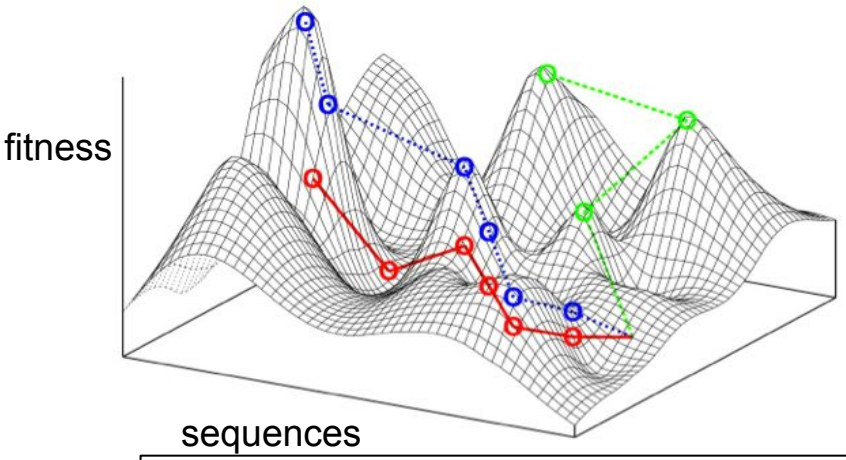
Lucile Vigué, Giancarlo Croce, Marie Petitjean, Etienne Ruppé, Olivier Tenaillon & Martin Weigt

*Nature Communications* **13**, Article number: 4030 (2022) | Cite this article

**Genome scale analysis:** 2053 Pfam domains, 281,513 residues, 2053 core gens

## Fitness landscape

Genotype-phenotype mapping which associates a quantitative phenotype to each possible amino-acid sequence [Wright 1932]




**Experimental characterization is infeasible:**
- **high dimensional space**: impossible to determine the fitness for each genotype variant
- **epistasis**: it may lead to a rugged landscape with many local optima.
- only **extremely local characterization** within **Deep Mutational Scans** experiments

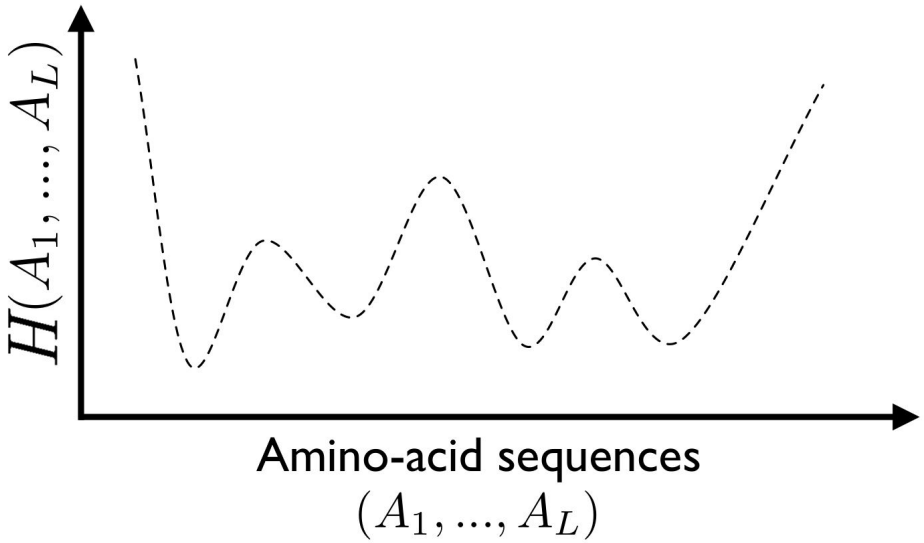
**Darwinian Evolution:** sampling sequences and survival of the fittest

**Predicting evolution ~ inferring the fitness landscape**

# Fitness landscape

Energy (Hamiltonian)

**In DCA framework**: $H(\mathbf{a}) = -\sum_{i<j}^{N} J_{ij}(a_i, a_j) - \sum_{i=1}^{N} h_i(a_i)$ **is it a good proxy for fitness?**

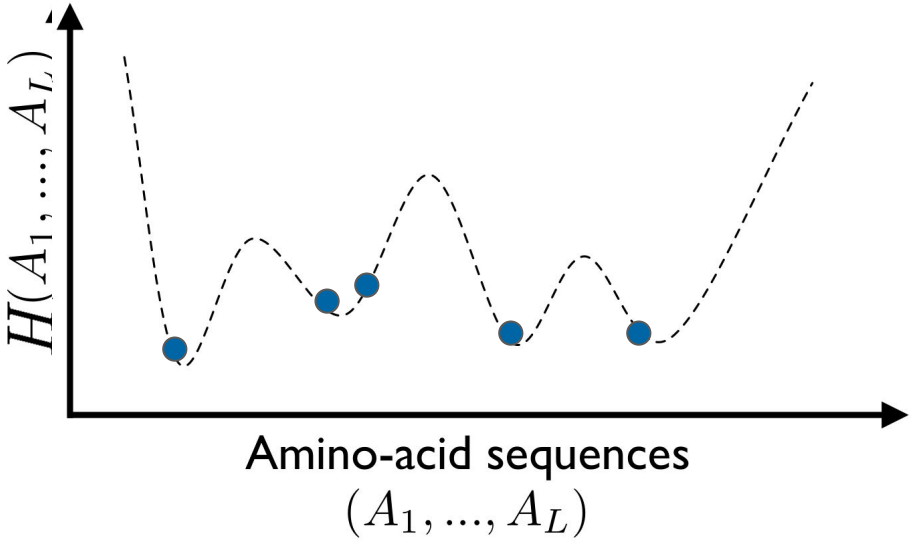



**bad** sequences

↕

**good** sequences

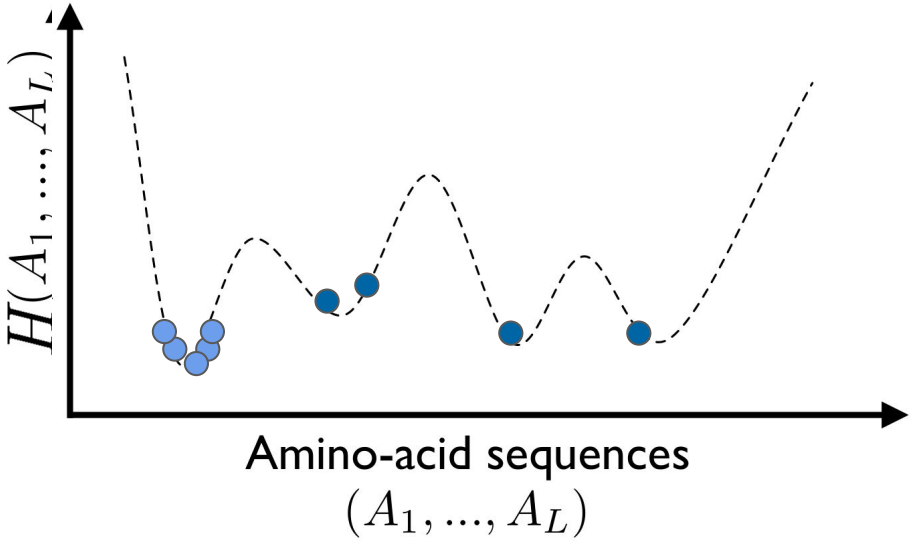
# Fitness landscape




**Homologous sequences (long term evolution)**
**Data in Uniprot/PFAM**

- distinct **species**
- 20-30% sequence ID

# Fitness landscape




**Homologous sequences (long term evolution)**
**Data in Uniprot/PFAM**

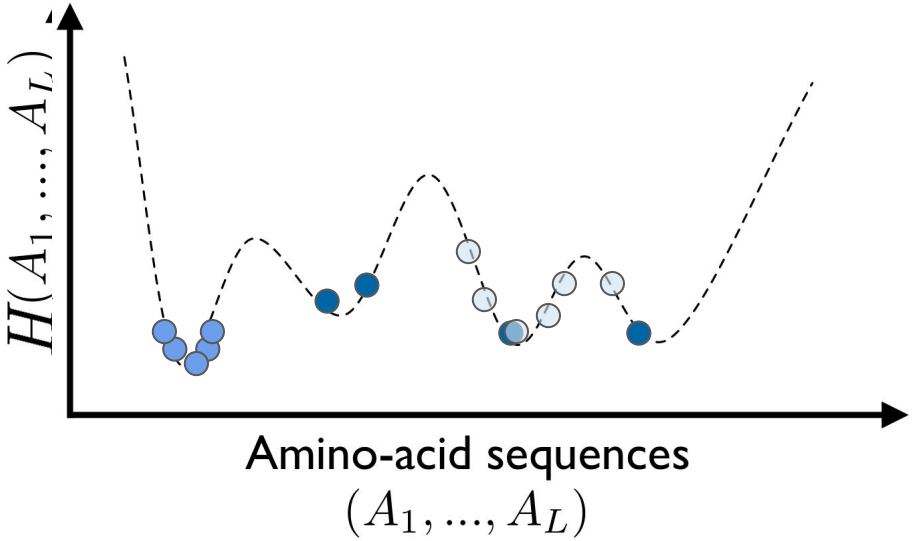
- distinct **species**
- 20-30% sequence ID

**Short term evolution**

- distinct **strains** / same species
  60.000 *E.coli* strains

```
Strain1: MAELKMAKLAAGLRKLAWYAA
Strain2: MAELKAAKLAAGLRKLAWYAA
Strain3: MAELKAAKLAAGLRKLAWKAA
Strain4: MAELKMAKLAAGLRKLAWYAA
Strain5: MAELKAAKLAAGLRKLAWYAA
```

# Fitness landscape




**Homologous sequences (long term evolution)**
**Data in Uniprot/PFAM**

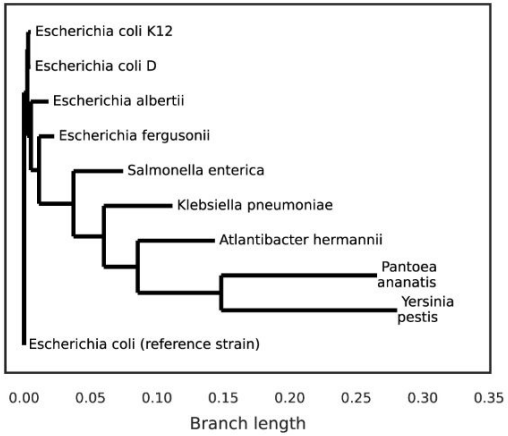
- distinct **species**
- 20-30% sequence ID

**Short term evolution**

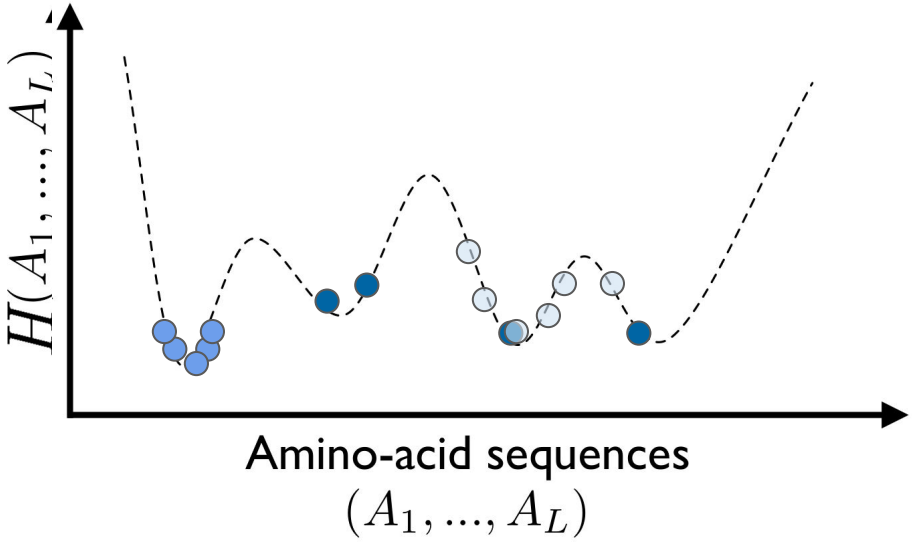
- distinct **strains** / same species
  60.000 *E.coli* strains

**Closely diverged species**

- Evolutionary close sequences



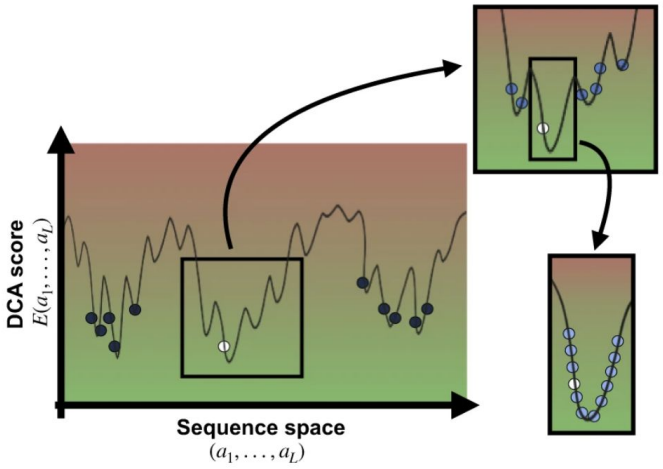
# Fitness landscape



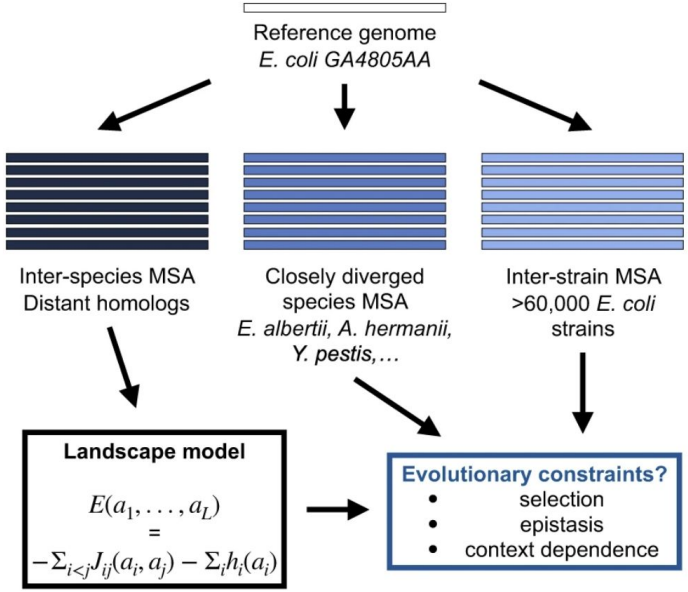

Can DCA models trained on **homologous sequences (long term evolution)** give information about **sequences emerging from short term evolution (different strains or closely related species)?**

Linking the **global** and **local** fitness landscape

# DCA to model and predict protein evolution: *E. coli*




Sequence landscape




Sequence data

**Genome scale analysis:** 2053 Pfam domains, 281,513 residues, 2053 core gens

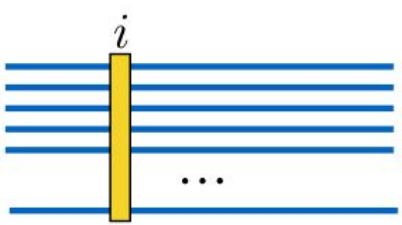
# DCA to model and predict protein evolution: *E. coli*

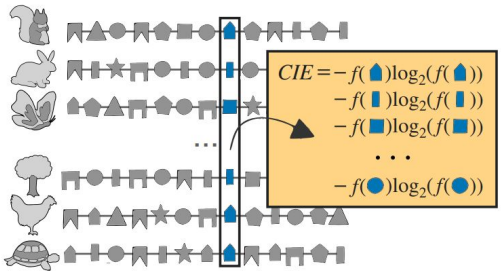
How to predict polymorphic positions?

- **context-independent** site entropy (= column entropy in diverged homologs MSA)

$$P(a_i) = \sum_{\{a_j | i \neq j\}} P(a_1, \ldots, a_N)$$

$$s_i = -\sum_{a_i} P(a_i) \log_2 P(a_i)$$

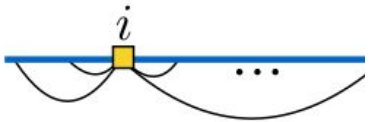


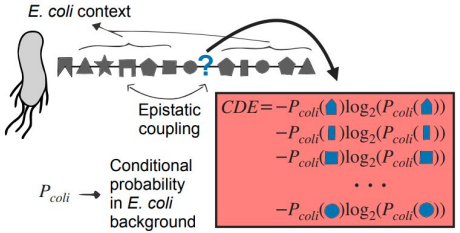




- **context-dependent** site entropy (with DCA model)
  (context of site $i$ : $\mathbf{a}_{-i} = \{a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_N\}$ reference strain)

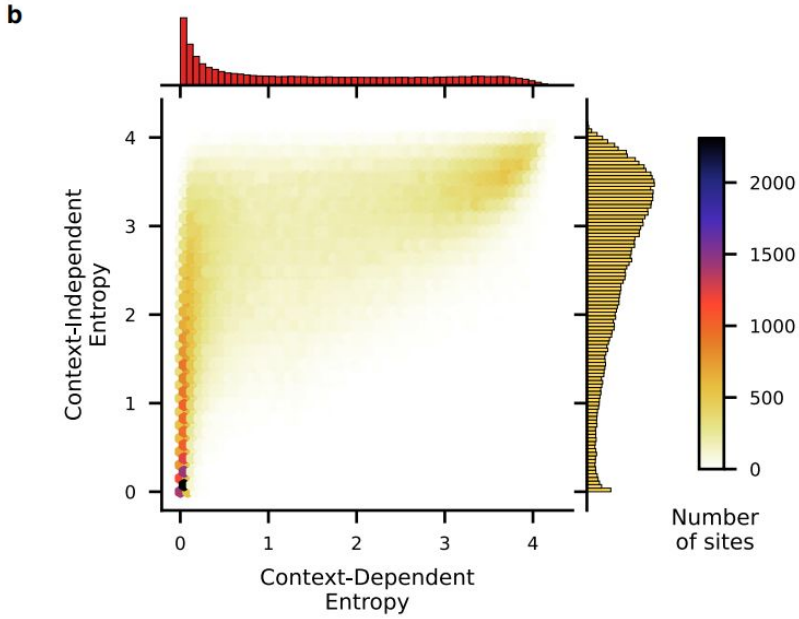
$$P(a_i | \mathbf{a}_{-i}) \sim \exp\left(h_i(a_i) + \sum_{j \neq i} J_{ij}(a_i, a_j)\right)$$

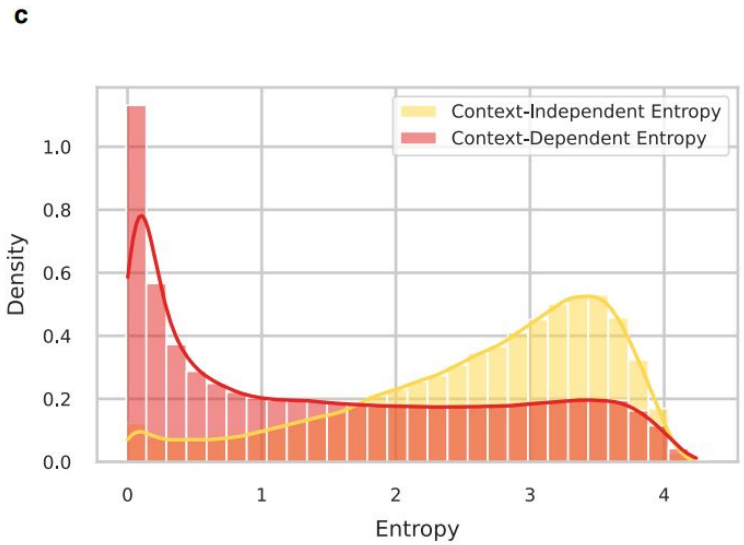
$$s_i(\mathbf{a}_{-i}) = -\sum_{a_i} P(a_i | \mathbf{a}_{-i}) \log_2 P(a_i | \mathbf{a}_{-i})$$



# DCA to model and predict protein evolution: *E. coli*

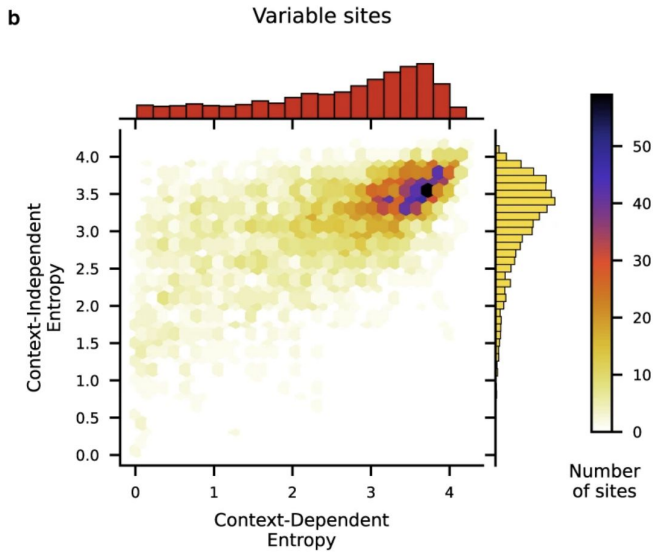


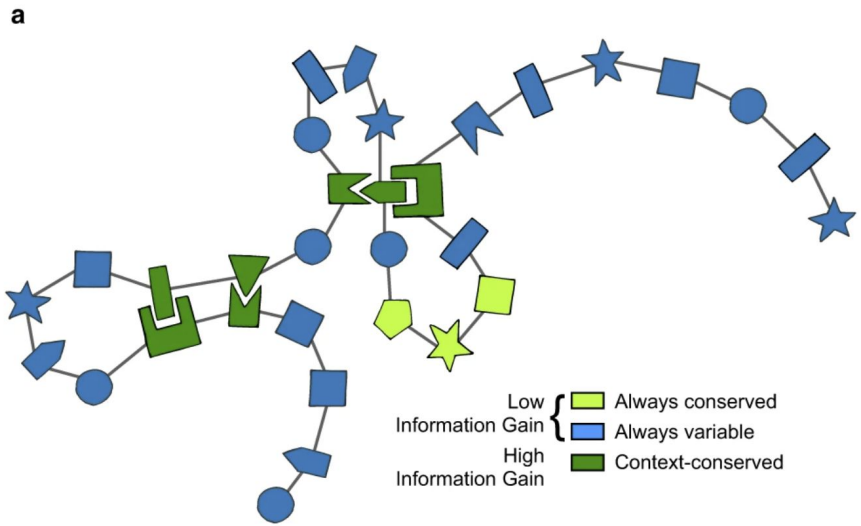




- Different distributions
- Context-independent *higher* than context-dependent
- When we include the specific *E. coli* context, sites tend to be become more constrained (30%-50% of positions)

# DCA to model and predict protein evolution: *E. coli*

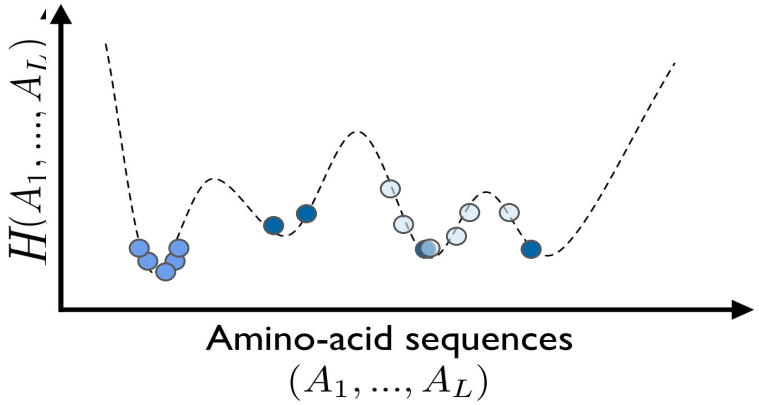
Can we predict polymorphic positions that have mutated in the 60.000 *E. coli* strains?



# DCA to model and predict protein evolution: *E. coli*

Epistatic interactions are weak and a collective effect

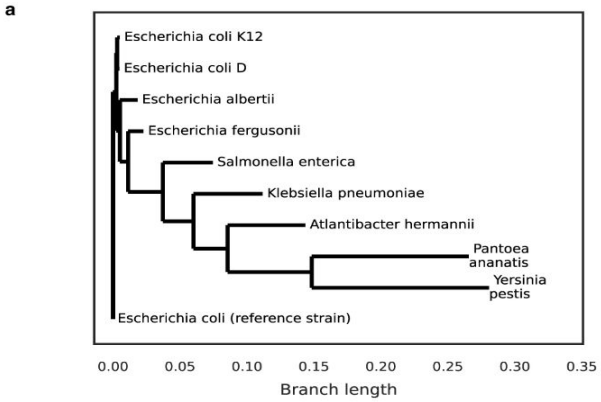


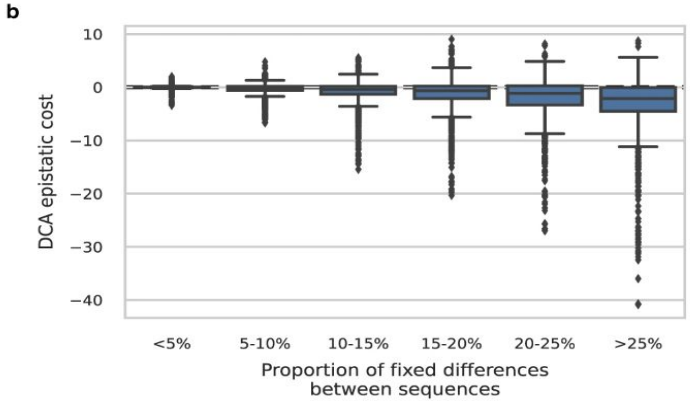

**Short term evolution**

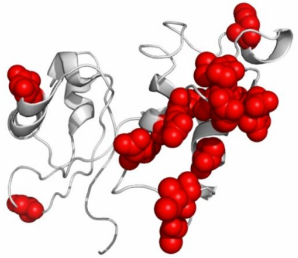
- distinct **strains** / same species
  60.000 *E.coli* strains
  *No clear signal of epistatic interactions*

**Closely diverged species**

- Evolutionary close sequences
  *Epistasis start to matters*



*rplK protein*: residues that differ between *E. coli* and *Y. pestis* in red.
Collective effect -> strong epistatic signal

# Acknowledgments

RESEARCH ARTICLE | BIOPHYSICS AND COMPUTATIONAL BIOLOGY

## Epistatic models predict mutable sites in SARS-CoV-2 proteins and epitopes

Juan Rodriguez-Rivas, Giancarlo Croce, Maureen Muscat, and Martin Weigt Authors Info & Affiliations

Edited by John Barton, Physics and Astronomy, University of California, Riverside, CA; received July 16, 2021; accepted December 13, 2021 by Editorial Board Member Mehran Kardar

**January 12, 2022** | 119 (4) e2113118119 | https://doi.org/10.1073/pnas.2113118119

nature > nature communications > articles > article

Article | Open Access | Published: 12 July 2022

## Deciphering polymorphism in 61,157 *Escherichia coli* genomes via epistatic sequence landscapes

Lucile Vigué, Giancarlo Croce, Marie Petitjean, Etienne Ruppé, Olivier Tenaillon & Martin Weigt

*Nature Communications* **13**, Article number: 4030 (2022) | Cite this article

# Acknowledgments

RESEARCH ARTICLE | BIOPHYSICS AND COMPUTATIONAL BIOLOGY

## Epistatic models predict mutable sites in SARS-CoV-2 proteins and epitopes

Juan Rodriguez-Rivas, Giancarlo Croce, Maureen Muscat, and Martin Weigt Authors Info & Affiliations

Edited by John Barton, Physics and Astronomy, University of California, Riverside, CA; received July 16, 2021; accepted December 13, 2021 by Editorial Board Member Mehran Kardar

January 12, 2022 | 119 (4) e2113118119 | https://doi.org/10.1073/pnas.2113118119

**Thanks for your attention**

nature > nature communications > articles > article

Article | Open Access | Published: 12 July 2022

## Deciphering polymorphism in 61,157 *Escherichia coli* genomes via epistatic sequence landscapes

Lucile Vigué, Giancarlo Croce, Marie Petitjean, Etienne Ruppé, Olivier Tenaillon & Martin Weigt

*Nature Communications* **13**, Article number: 4030 (2022) | Cite this article