AI-driven prediction of protein-protein binding trends from atomistic simulation data

Sara Capponi

IBM Almaden Research Center
NSF Center for Cellular Construction

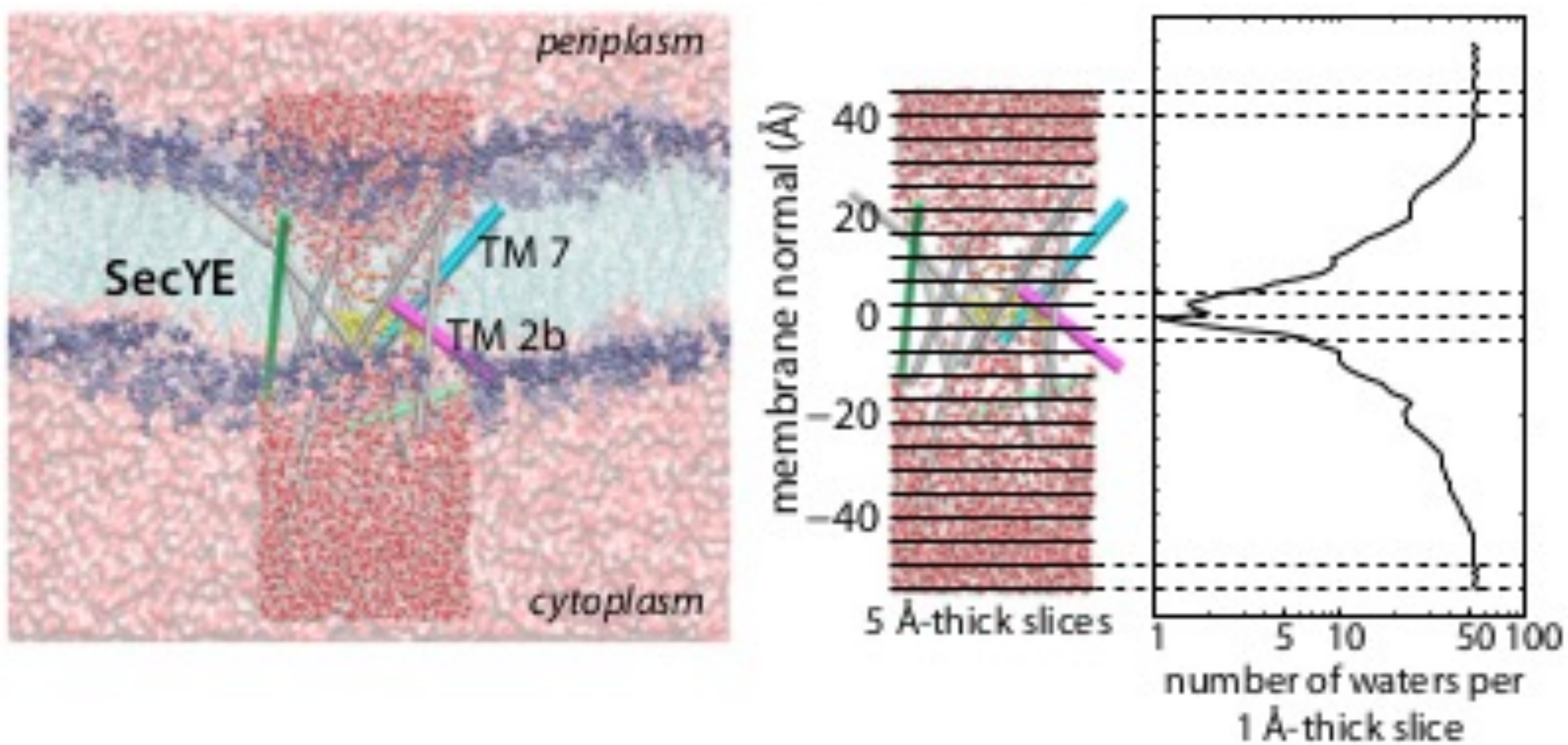
CENTER FOR CELLULAR CONSTRUCTION

This material is based upon work supported by the NSF under Grant No. **DBI-1548297**.

BACKGROUND in COMPUTATIONAL BIOPHYSICS

Molecular Dynamics (MD) Atomistic Simulations → simulate biological processes at atomic resolution (limitations in simulation length)

…unless you can access IBM supercomputers…



Capponi et al., *PNAS* 2015

@ UCI

BACKGROUND in COMPUTATIONAL BIOPHYSICS

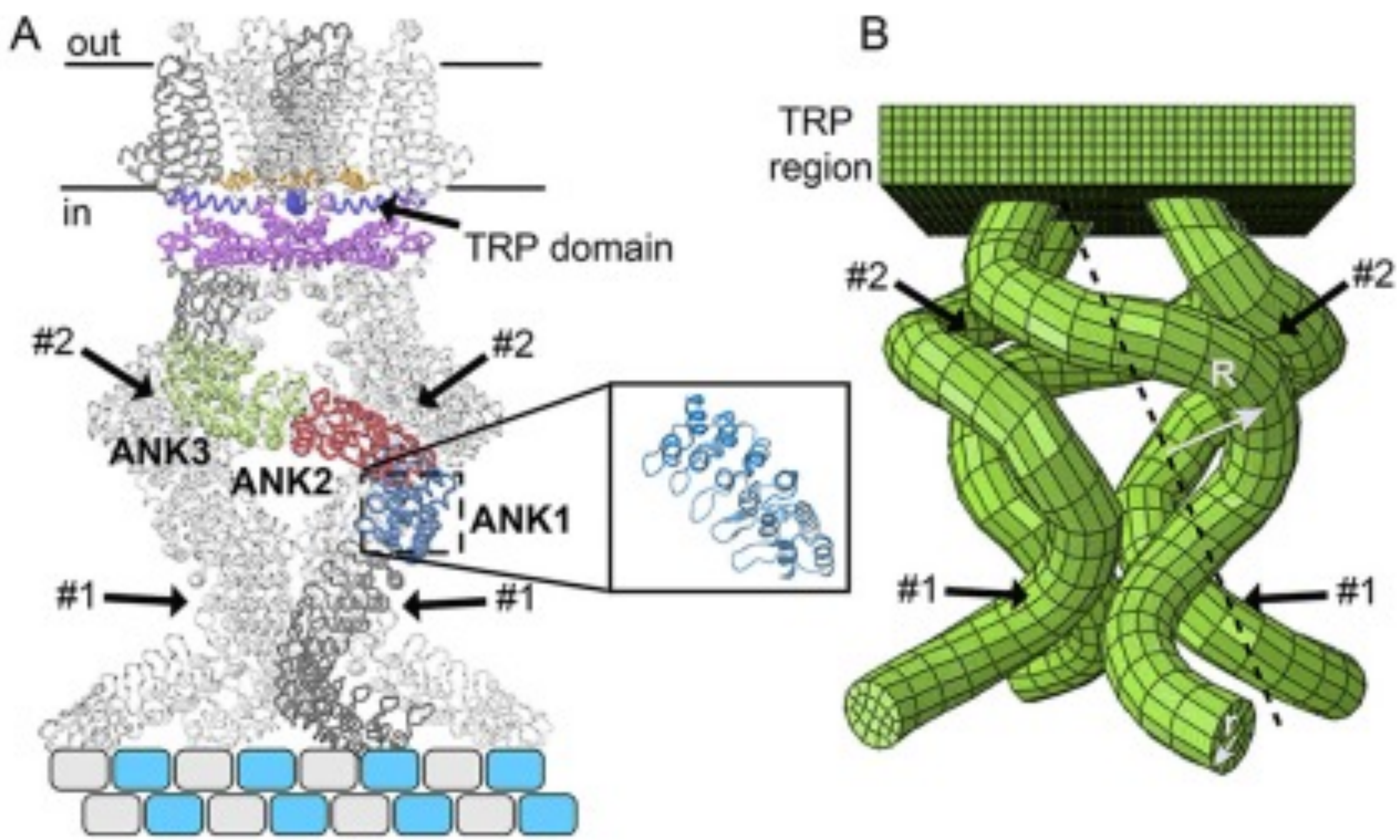
Molecular Dynamics (MD)
Atomistic Simulations
+
Finite Element Model

→

MD to extract mechanical properties
to describe the motion
of the ankyrin chains of
mechanotransduction channel

MD

Finite Element Model



Argudo et al., *J. Gen Phys.* 2019

@ UCSF

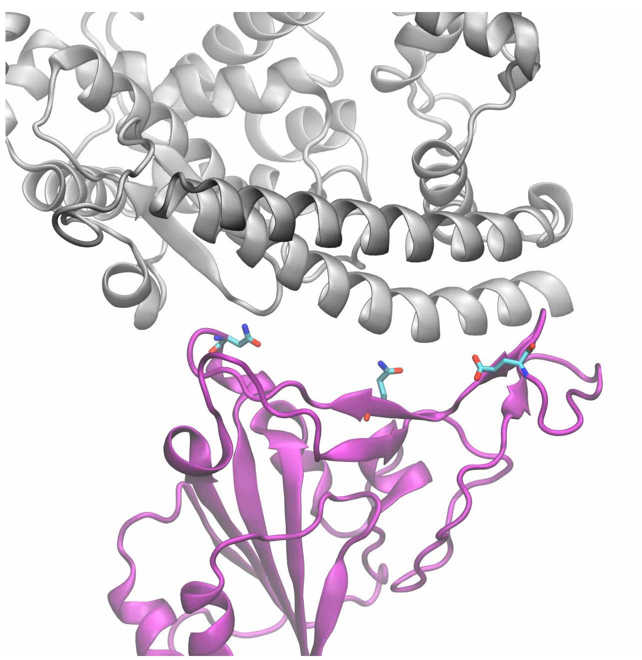
BACKGROUND in COMPUTATIONAL BIOPHYSICS

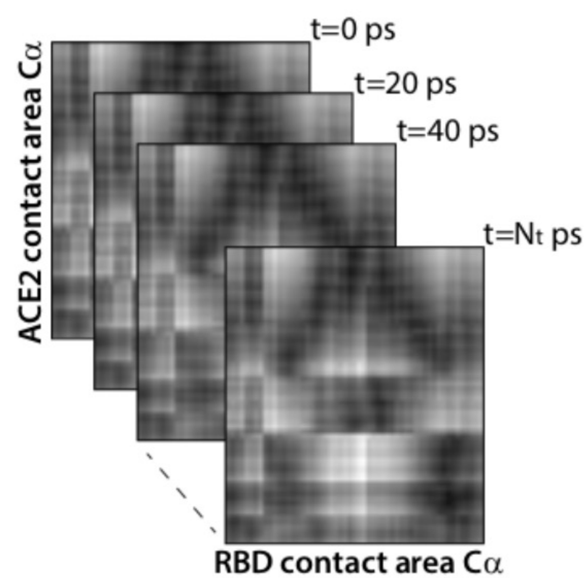
Molecular Dynamics (MD)
Atomistic Simulations
+
Artificial Intelligence

→

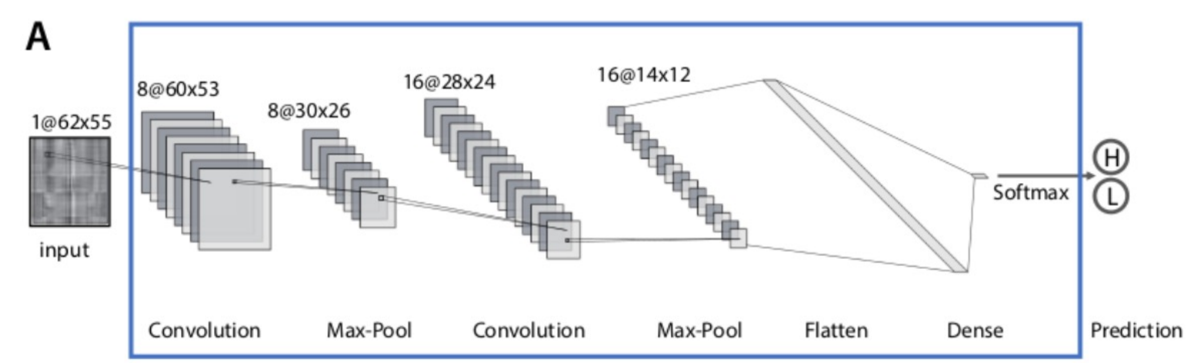
MD to extract dynamical information
to feed AI algorithms and make predictions

MD





AI

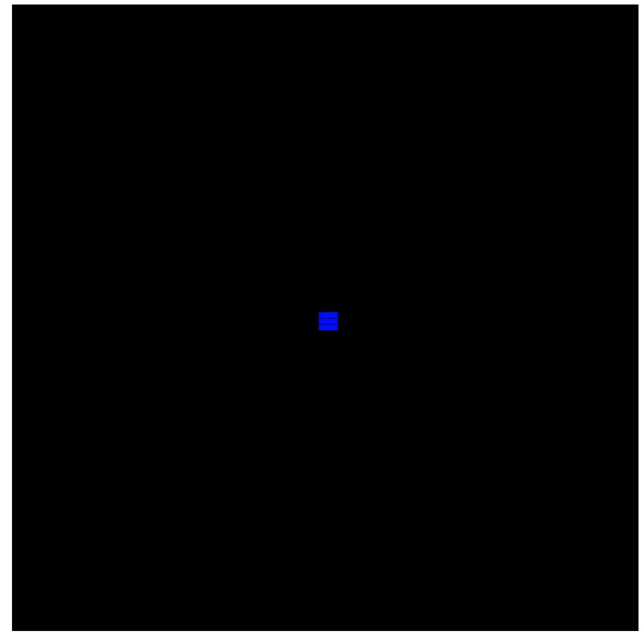


@ IBM

BACKGROUND in COMPUTATIONAL BIOPHYSICS

Agent Based Model (ABM)

to study **viral infection**
in presence of DIPs and IFN

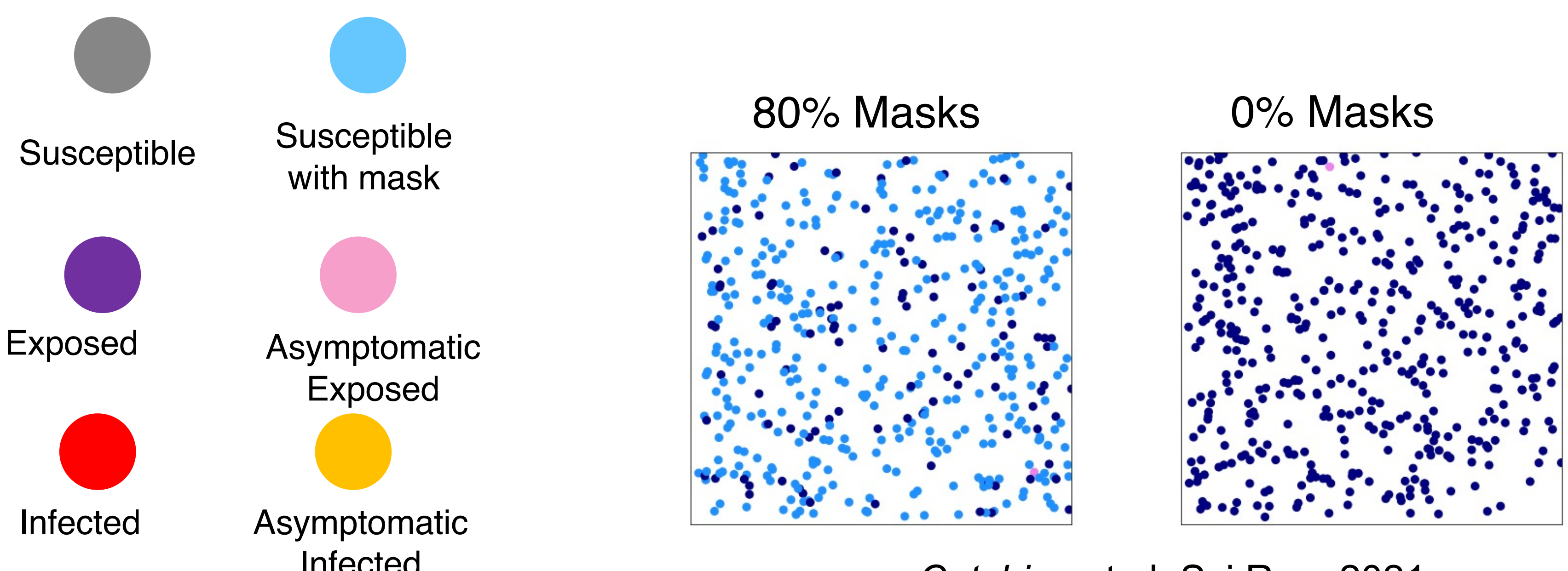


agents = viral particles
cells = lattice squares

DARPA grant @ IBM, S Bianco

to probe **mask effectiveness** for
reducing COVID-19 transmission

IBM COVID-19 Task Force, S Bianco



agents = individuals

Catching et al. Sci Rep, 2021

@ IBM

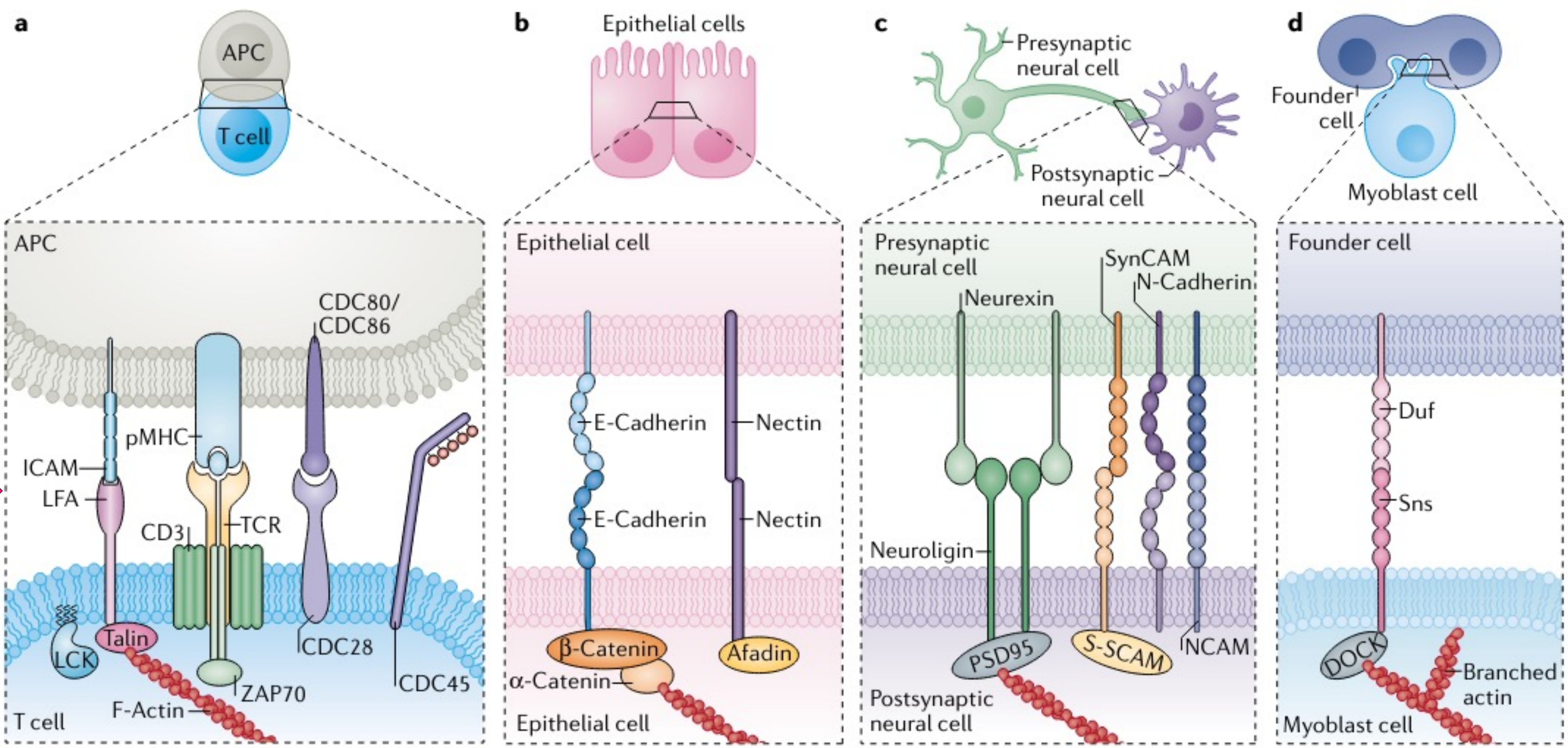
AI-driven prediction of protein-protein binding trends from atomistic simulation data

Motivations

Why do we care about binding affinity?

cell binding → protein-protein binding

between cells



inside cells

Dan Fletcher, *Nature Rev* 2020

Motivations

Can we *efficiently* measure and predict binding affinity?

Experiments and biased atomistic simulations are expensive (money and time)

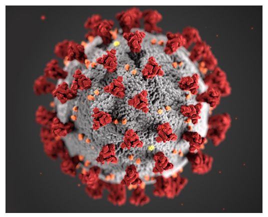
Build a **CNN framework** to:

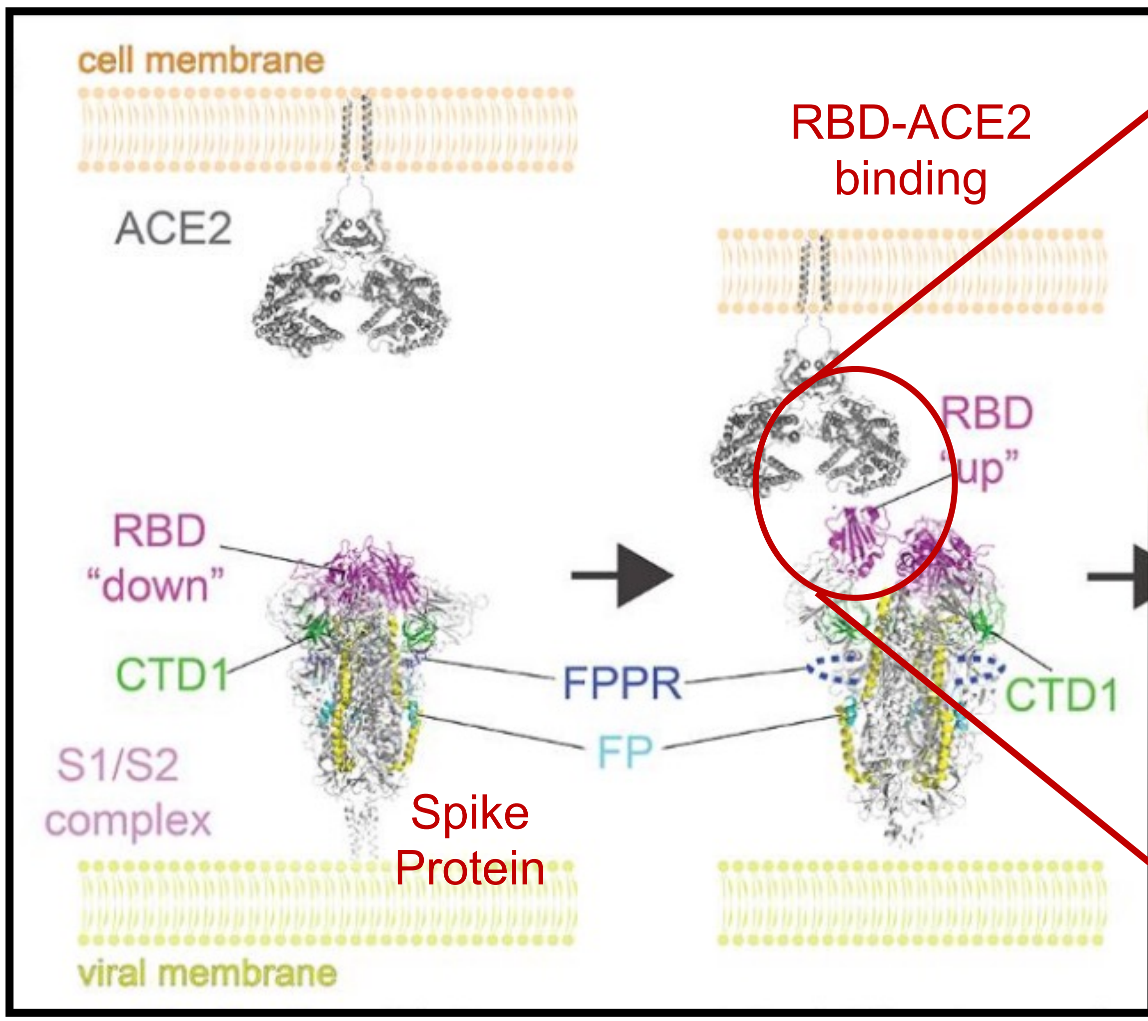
- predict **efficiently protein-protein binding affinity** trend from unbiased atomistic simulation data
- **shortening** the length of **unbiased atomistic simulation** data used for binding affinity estimation

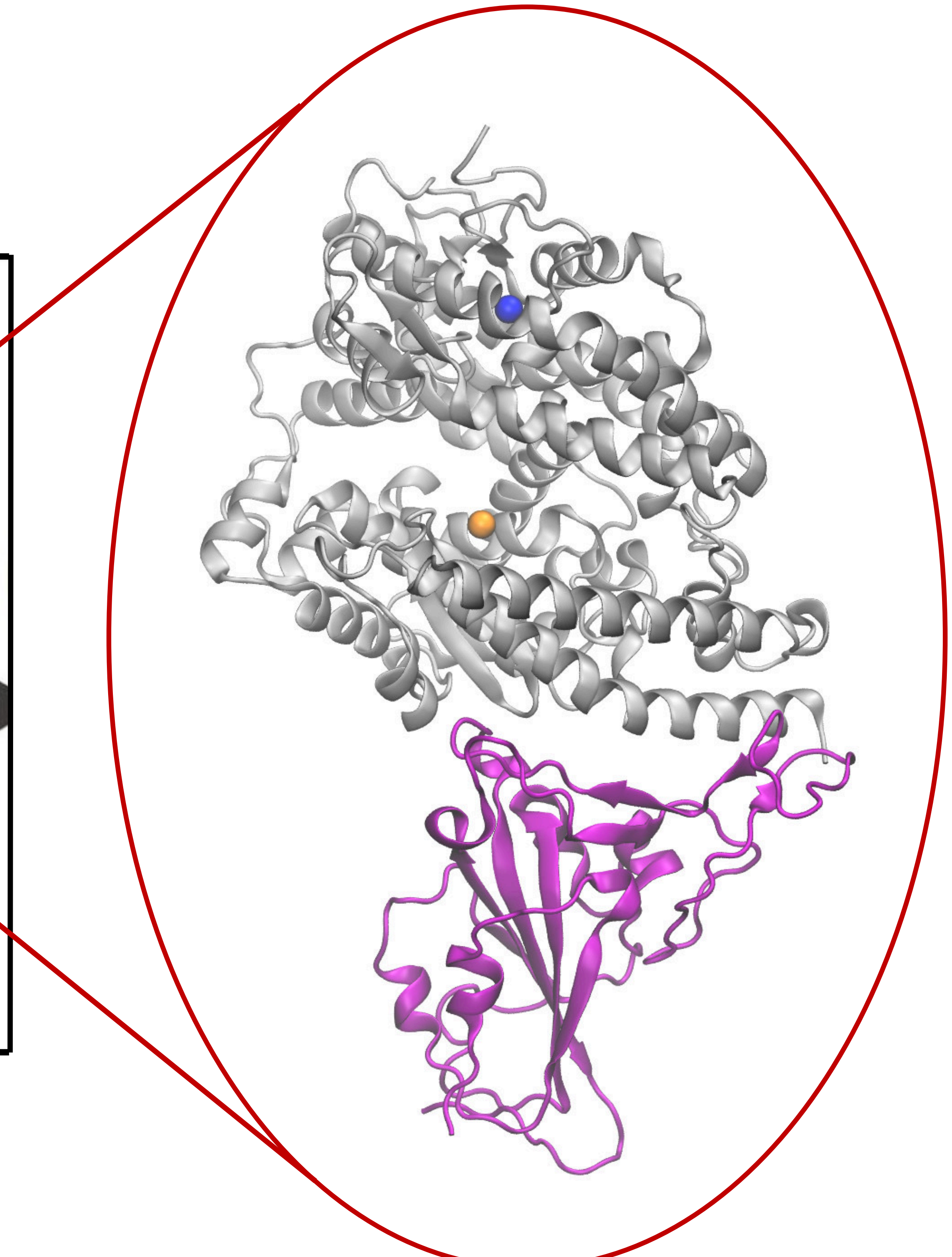
Spike protein S test case:

S responsible for **binding & membrane fusion**

SARS-CoV-2 cell entry mechanism: ACE2-RBD complex







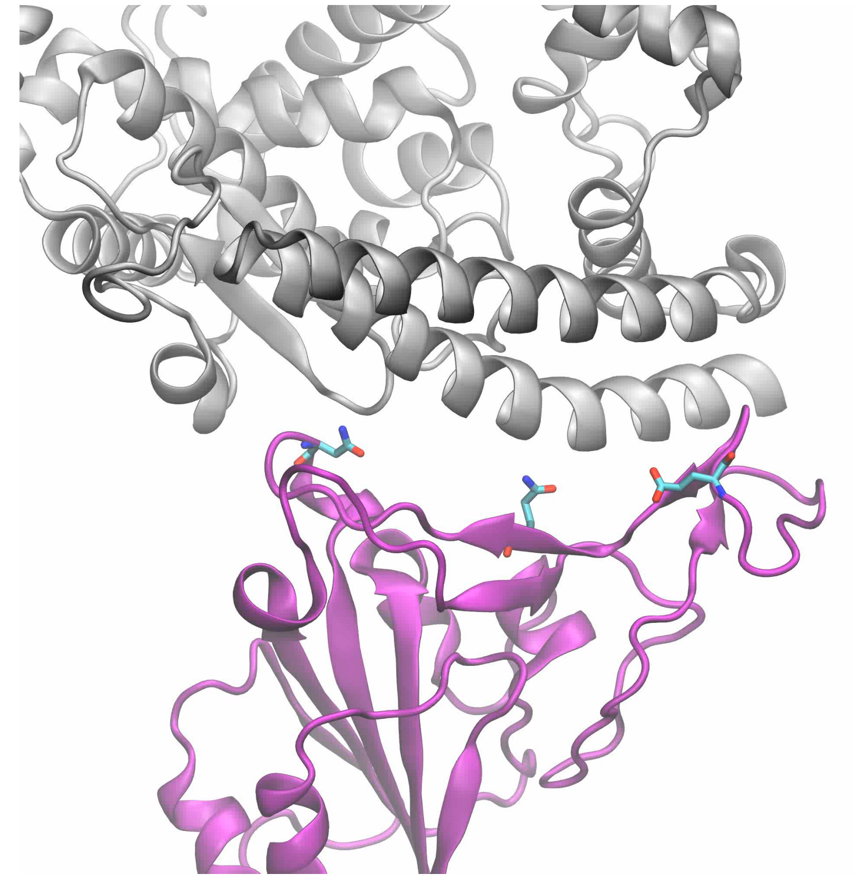
ACE2-RBD binding affinity estimation

Reference data

RBD-ACE2 $\Delta\log_{10}(K_{D,app}) = 0$ (reference)

Mutation	Origin	$\Delta \log_{10}(K_{D,app})$
N501Y	English variant	0.24
Q498Y	SARS-CoV	0.16
N501V	-	0.15
Q493Y	Bat RaTG13	0.12
N501T	SARS-CoV	0.1
E484K	Brazilian variant	0.06
N501S	-	-0.13
Q493N	SARS-CoV	-0.21
Q498N	-	-0.5
Q498K	-	-2.26
N501D	Bat RaTG13	-2.42
G502P	-	-4.55

simulation	lenght (ns)
cRBD-ACE2	270
N501Y-ACE2	187
Q498Y-ACE2	206.84
N501V-ACE2	204.86
Q493Y-ACE2	197.28
N501T-ACE2	186.58
E484K-ACE2	204.56
N501S-ACE2	205.2
Q493N-ACE2	197.18
Q498N-ACE2	198.84
Q498K-ACE2	204.46
N501D-ACE2	210.44
G502P-ACE2	160.58



T Starr et al, *Cell* 2020

Hydrophobic RBD mutations

Mutation	Origin	$\Delta \log_{10}(K_{D,app})$	Length (ns)
V503I	SARS-CoV	0.05	187.8
L492I	–	0.03	261.24
L455I	–	−0.01	272.24
V503A	—	−0.06	188.18
A475V	SARS-CoV	−0.14	226.6
L455A	–	−0.43	280.72
L455V	–	−0.73	280.98
A475L	–	−1.27	279.28
A475P	SARS-CoV	−1.62	281.02

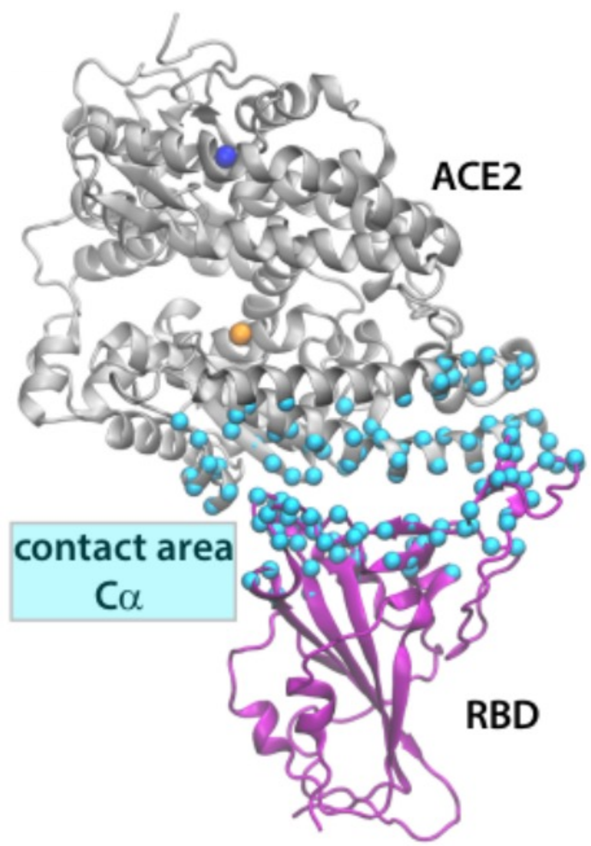
160 ns → ~ 5 days IBM supercomputer
~ 23 days lab server

… if not in parallel
→ 5 days x 22 simulations ~ 4 months

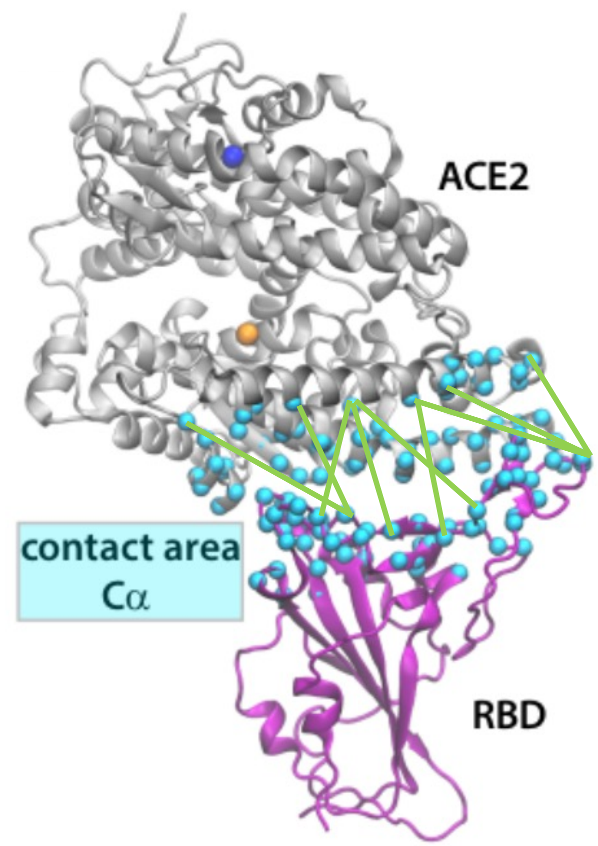
ACE2-RBD binding affinity estimation

training set preparation

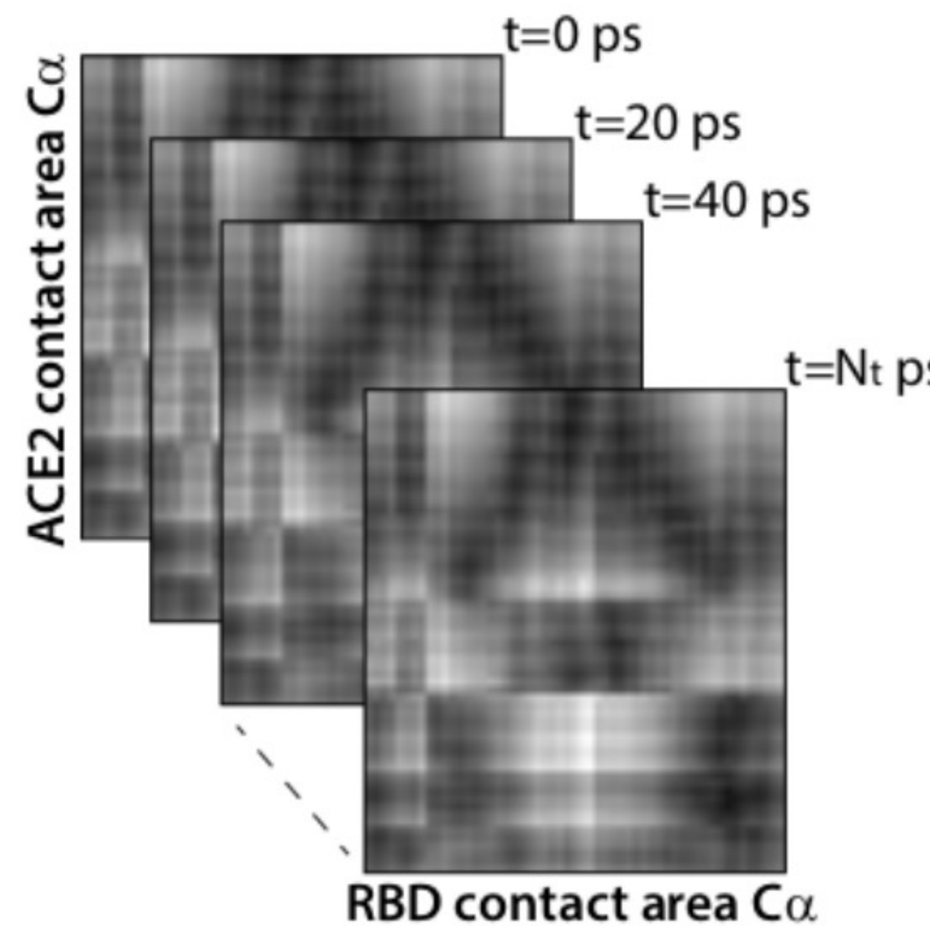
Conversion of MD data to inputs for machine learning algorithm



contact area



calculation all distances
between all Ca atoms
belonging to the *contact area*
(62 ACE x 55 RBD contact matrix)



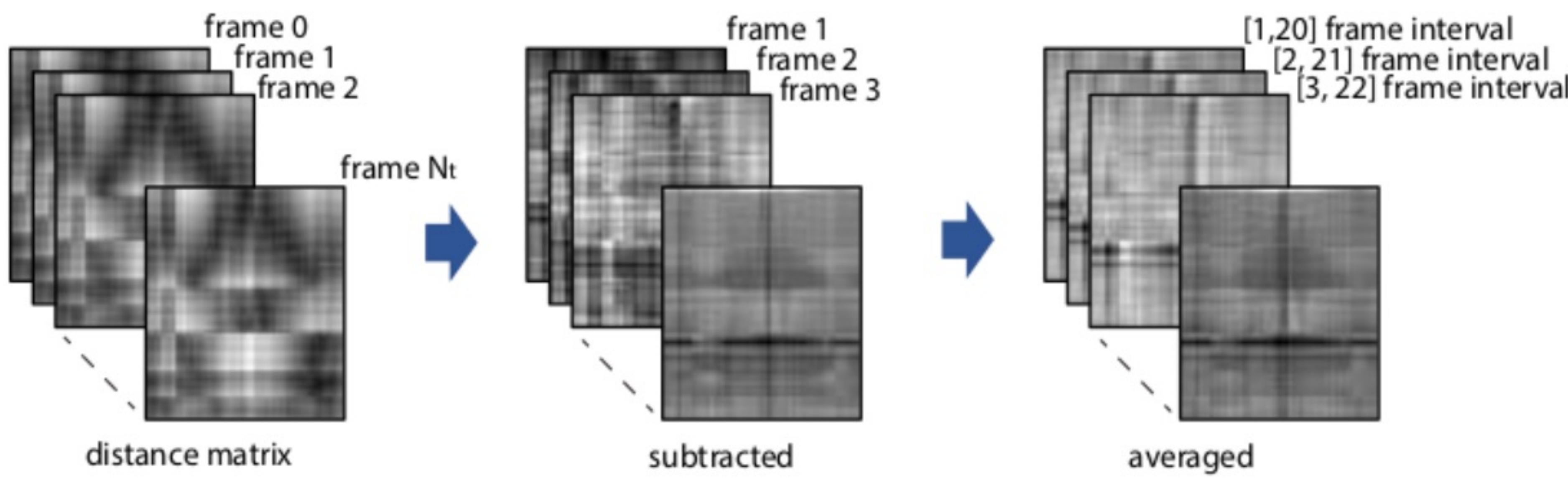
contact matrices
=
gray-scale images

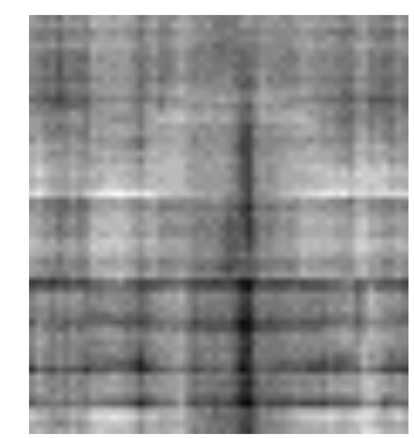
Capponi, et al, *EPJE* 2021

ACE2-RBD binding affinity estimation

training set preparation

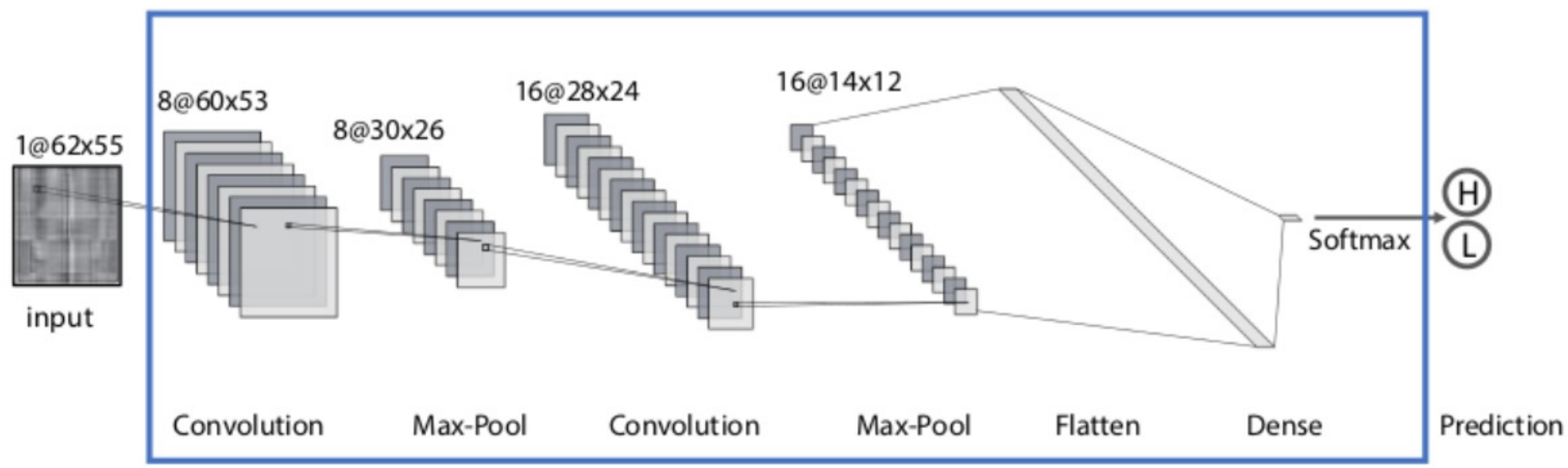
Data processing

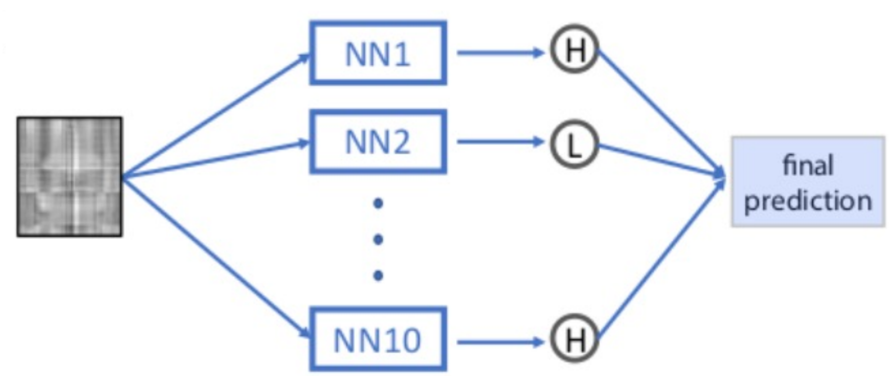




ACE2-RBD binding affinity estimation

Convolutional Neural Network Framework





ensemble prediction strategy (10 NNs)

S. Wang et al, *Nat Comm* (2019)

ACE2-RBD binding affinity estimation

Training and Validation Data Set

prediction on whole simulation length (160 ns)

Mutation	Origin	$\Delta \log_{10}(K_{D,app})$
N501Y	English variant	0.24
Q498Y	SARS-CoV	0.16
N501V	-	0.15
Q493Y	Bat RaTG13	0.12
N501T	SARS-CoV	0.1
E484K	Brazilian variant	0.06
N501S	-	-0.13
Q493N	SARS-CoV	-0.21
Q498N	-	-0.5
Q498K	-	-2.26
N501D	Bat RaTG13	-2.42
G502P	-	-4.55

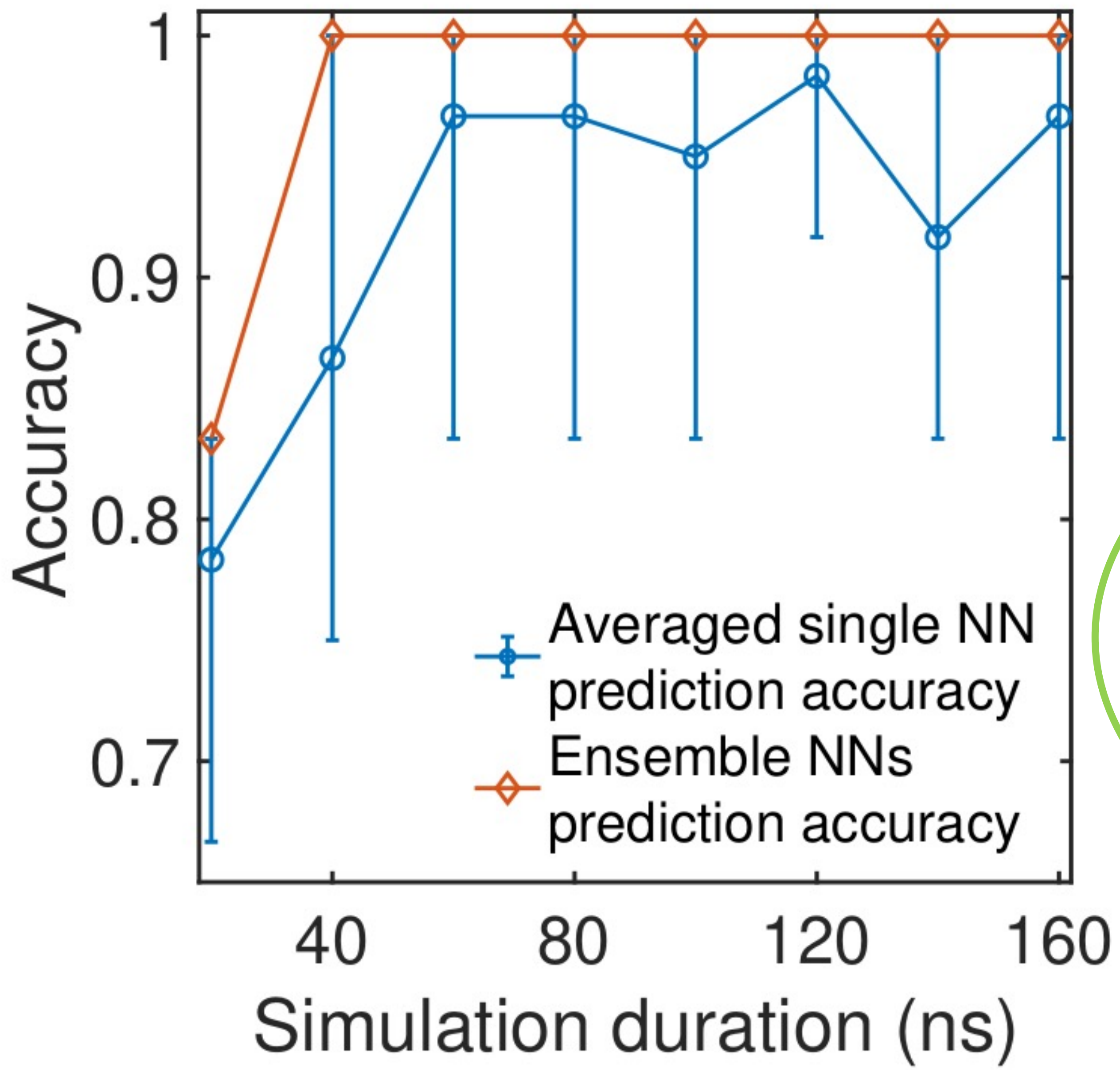
- 100 % NN ensemble accuracy
- 96.67 % averaged accuracy on single NN of the ensemble

CNN ensemble predicts accurately binding affinity trend

ACE2-RBD binding affinity estimation

Predictions on shorter simulated time

prediction on **increasing time window (20ns to 160ns, step 20ns)**



100 % NN ensemble accuracy using 40ns unbiased simulation data

CNN ensemble predicts accurately binding affinity trend using unbiased short atomistic data (40 ns)

Conclusions

- **Ensemble of CNNs** trained on distance matrices predicts with **high accuracy** the binding affinity trend

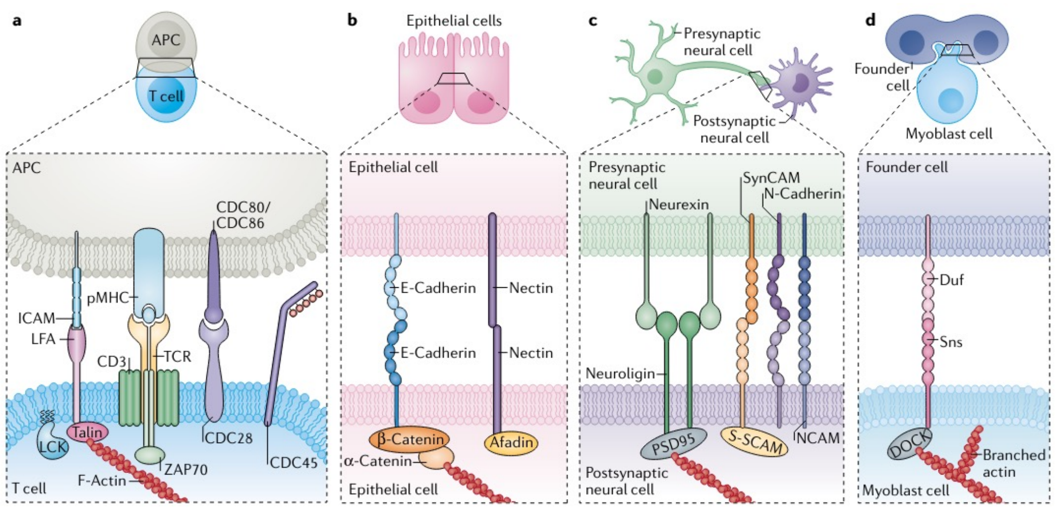
- We proved our AI method predicts binding affinity trends using **short, unbiased atomistic simulations**

  (40ns *vs* 160ns simulations → 1 day *vs* 5 days on a HPC cluster)

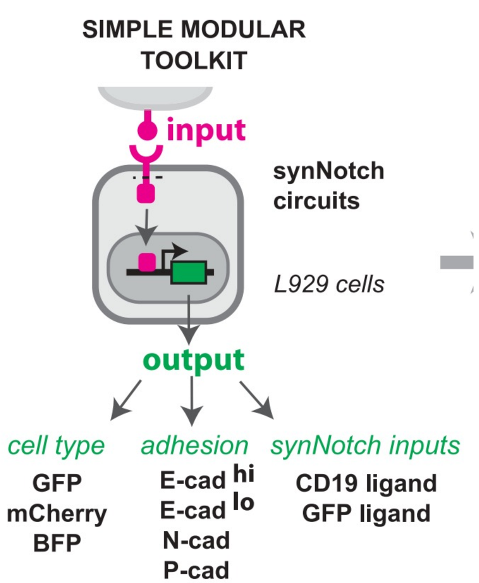
Contribution to CCC: MD-AI

IBM supercomputers

Cell-cell interactions

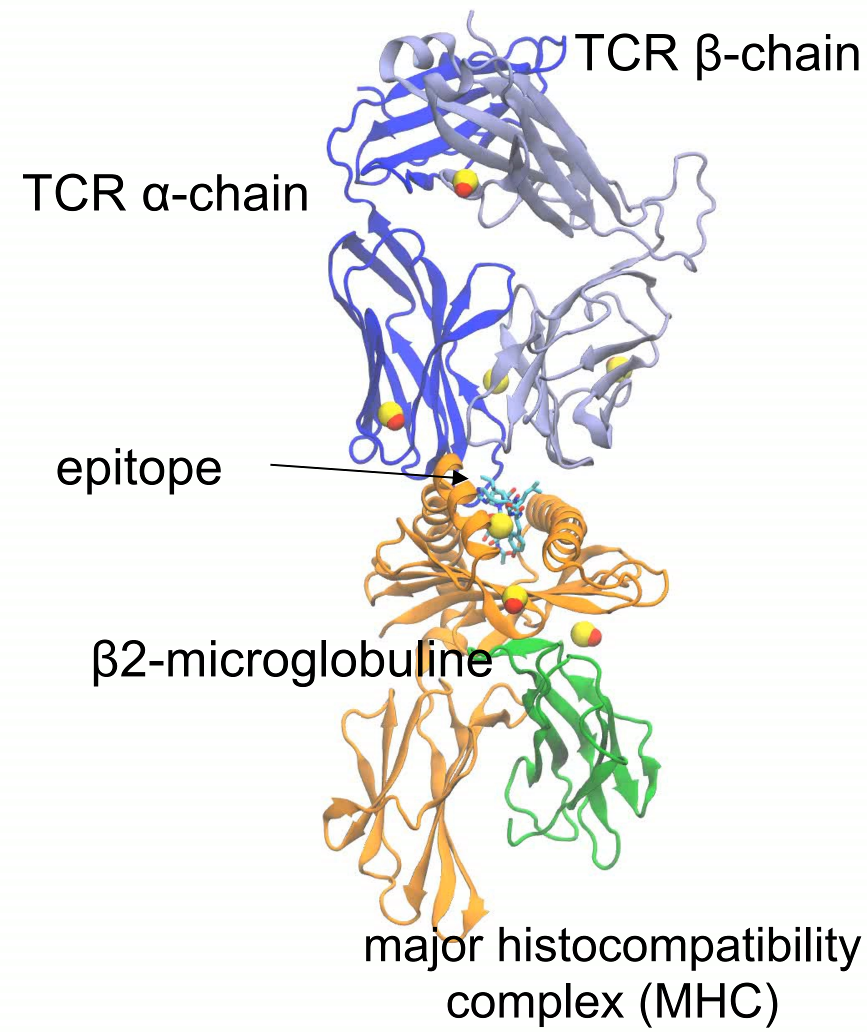


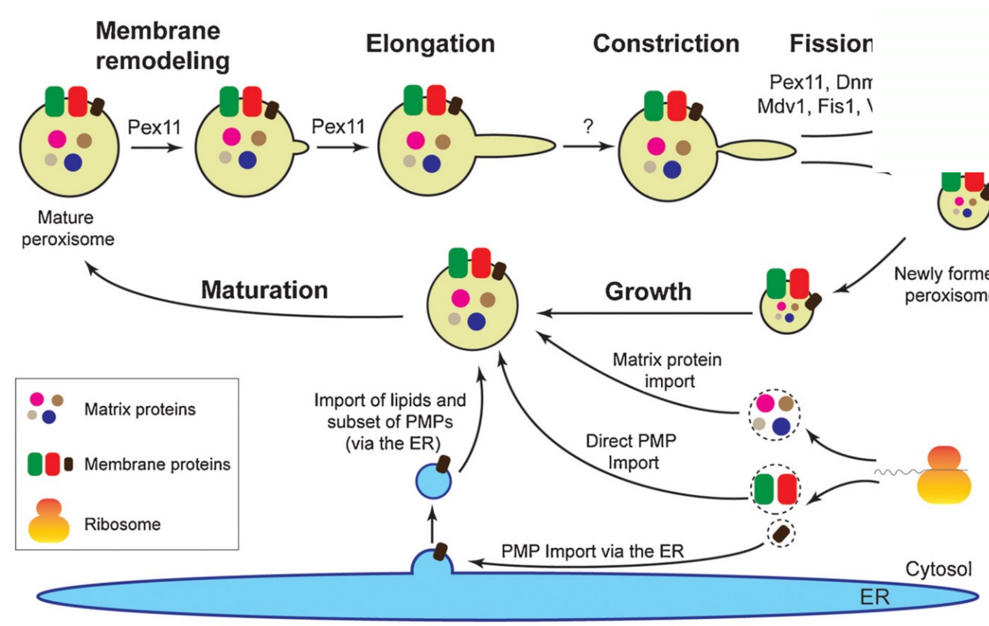
SynNotch
Specific activation of CAR-T



Understanding and programming self-organizing multicellular structures
with synthetic cell-cell signaling

Immune system AI-aided engineering

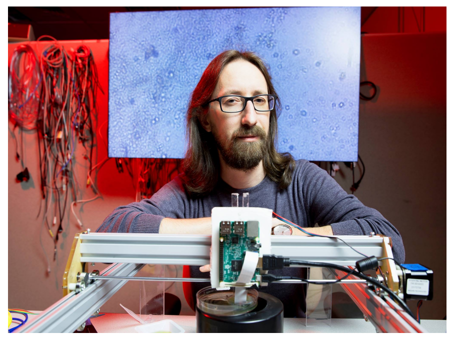




peroxisome proteins and membrane
involved in fusion, fission, maturation

Acknowledgements

IBM Cellular Engineering Group



Simone
Bianco

Shangying
Wang

Tom
Zimmerman

Thank you
for your attention

Open positions:

- IBM summer internship
- MD-AI postdoc

sara.capponi@ibm.com
or CCC slack

CENTER FOR
CELLULAR CONSTRUCTION

This material is based upon work supported by the NSF under Grant No. **DBI-1548297**.