CSC

Machine Learning for Chemistry

Shreyas Kaptan
Department of Physics
University of Helsinki

Outline

- Introduction
- Methods
- An example application
- Acknowledgements

Introduction

What is Machine Learning (ML)?

The “Machine” part

1. Automated algorithms
2. Reproducible results
3. Handling large datasets

The “Learning” part

1. Statistical models of the process
2. Inferring the parameters
3. Validating the results

Why use Machine Learning?

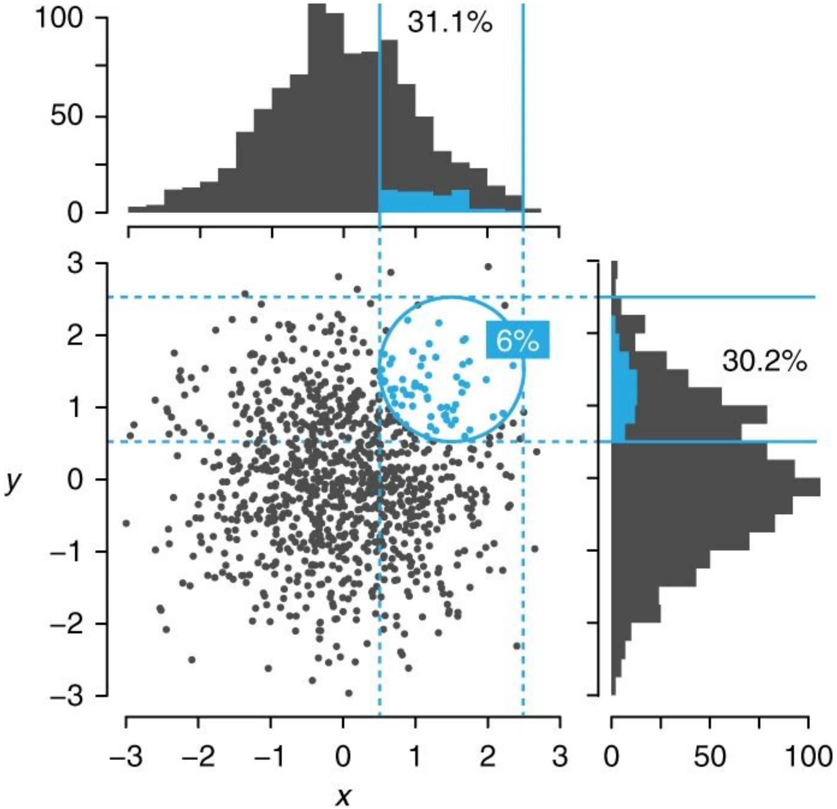
Or

The Many Problems of Drawing Statistical Inference in High Dimensionality

The many curses of dimensionality…

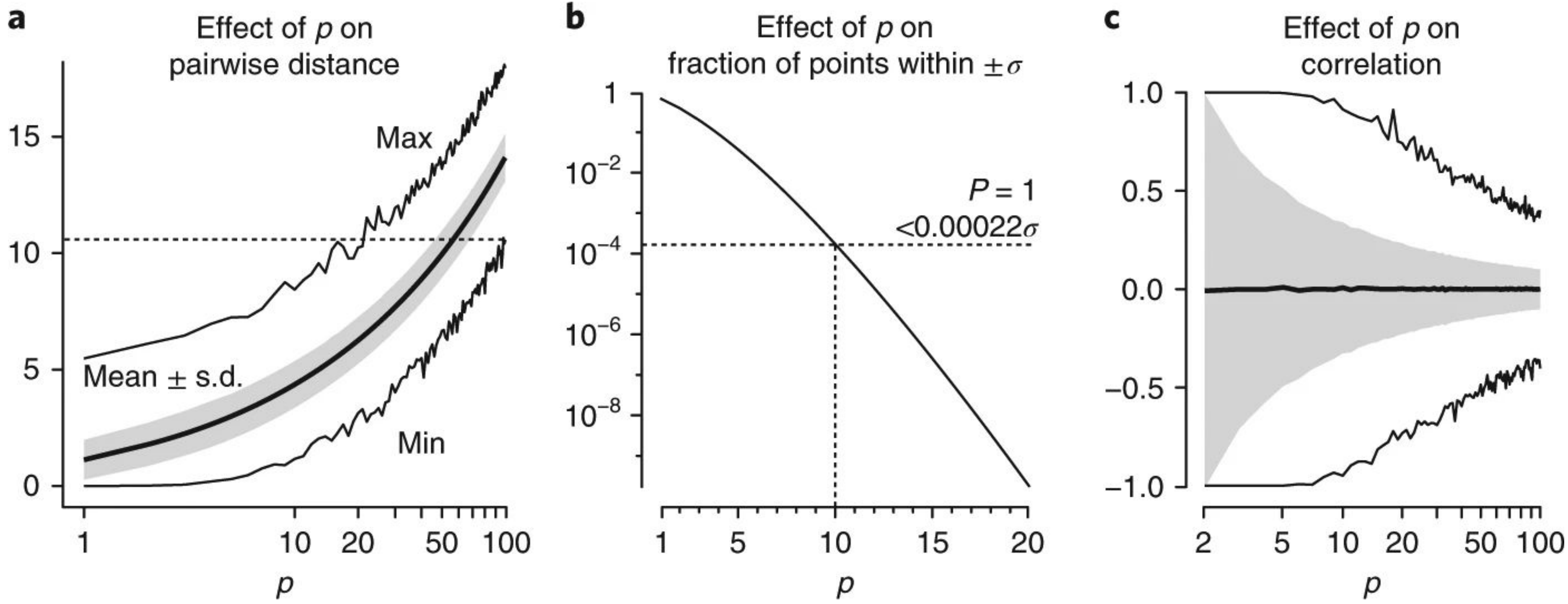
Data tend to be sparse

in higher dimensions.



Altman, N., Krzywinski, M. The curse(s) of dimensionality. *Nat Methods* 15, 399–400 (2018). https://doi.org/10.1038/s41592-018-0019-x

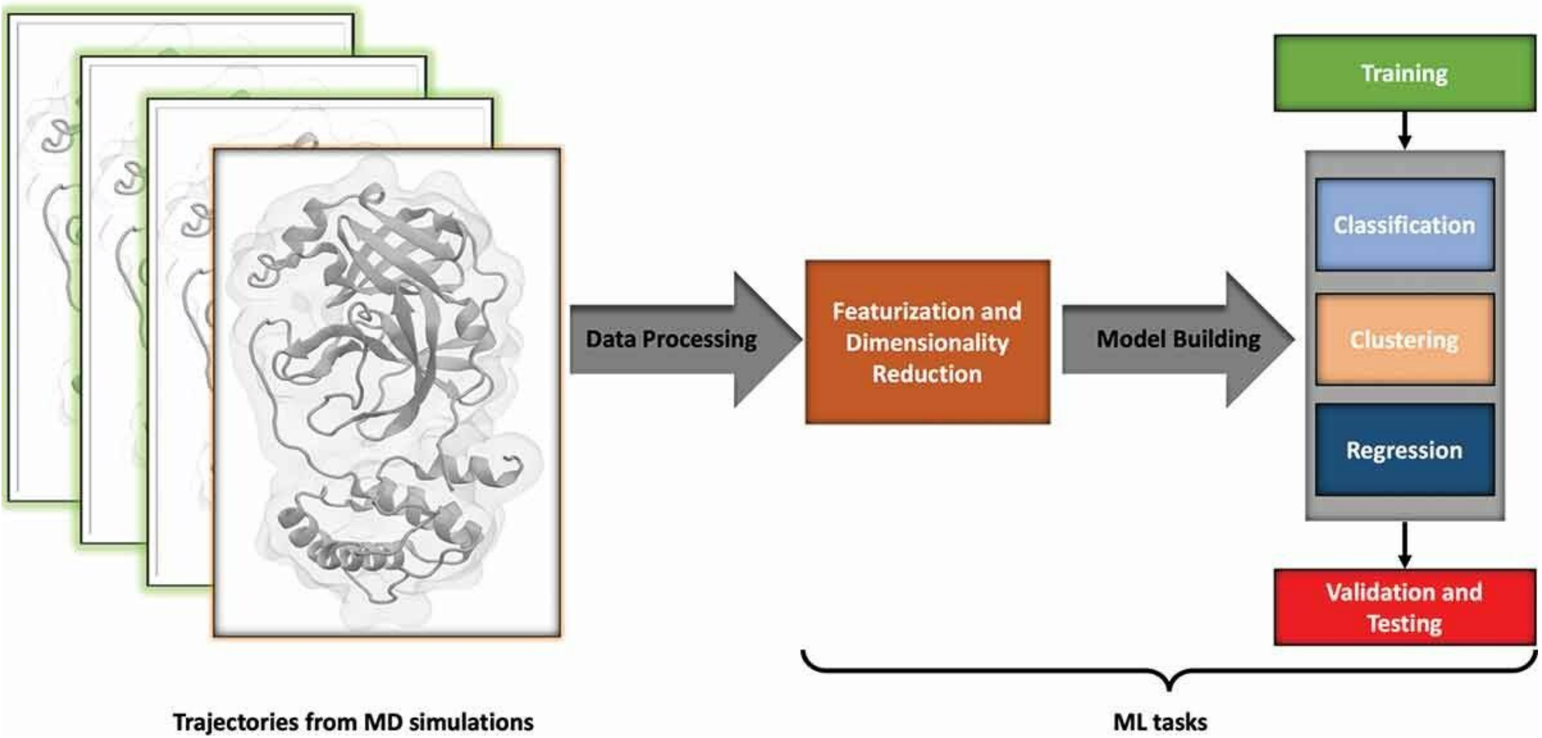
The many curses of dimensionality…



As the number of variables *p* increases, distances between points grow rapidly and correlations decrease.

Altman, N., Krzywinski, M. The curse(s) of dimensionality. *Nat Methods* 15, 399–400 (2018). https://doi.org/10.1038/s41592-018-0019-x

So.. How do we address this problem?

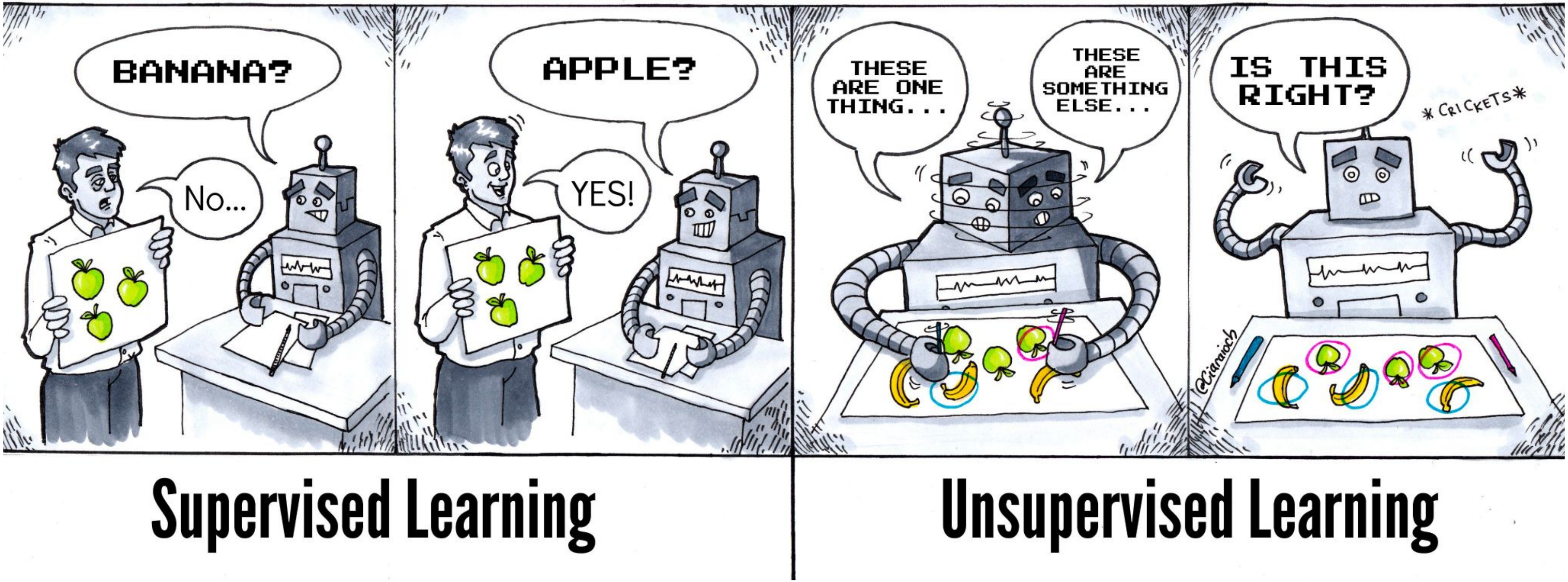


Shreyas Kaptan & Ilpo Vattulainen (2022) Machine learning in the analysis of biomolecular simulations, Advances in Physics: X, 7:1, DOI: 10.1080/23746149.2021.2006080

Vacabulary

1. **Features**: Variables present in your raw data
2. **Feature selection**: Choosing a subset of the Features or their functions for further analysis
3. **Training**: Learning the statistical parameters associated with a ML model
4. **Validation**: Testing the learnt models on a previously unseen and uncorrelated data to prevent overfitting

Generically, there are two kinds of ML tasks…



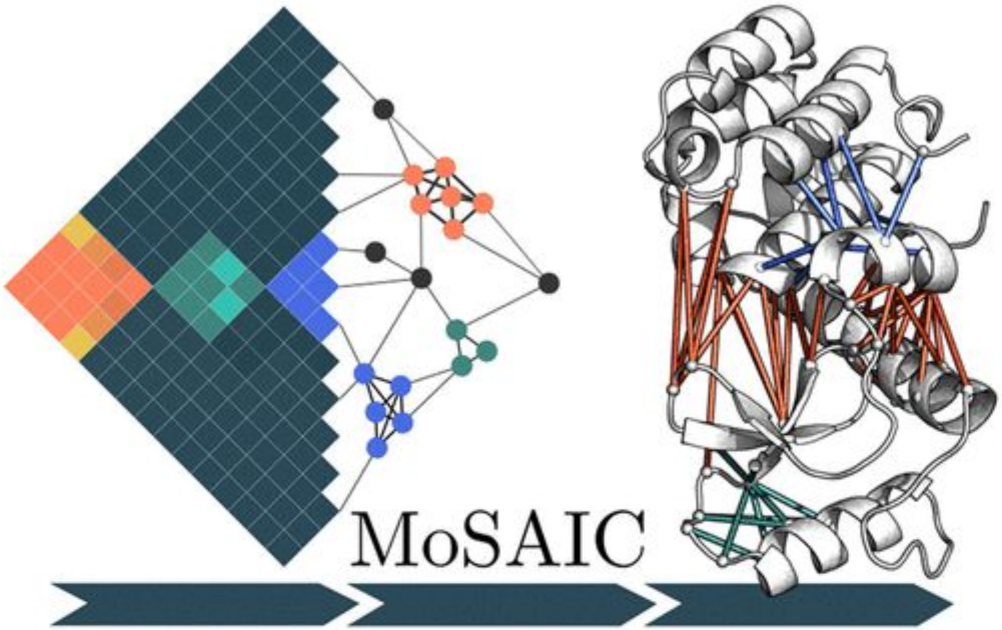
Comic by **Ciaraíoch**

Methods

Problem #0: I want to choose which part of my data I should use…

Feature selection tasks

1. Discarding variables with low variance
2. Testing for significance of individual features e.g. through significance tests (works for supervised tasks)
3. Chemical Insight (e.g. C-alpha atoms; heavy atoms, dihedrals etc)

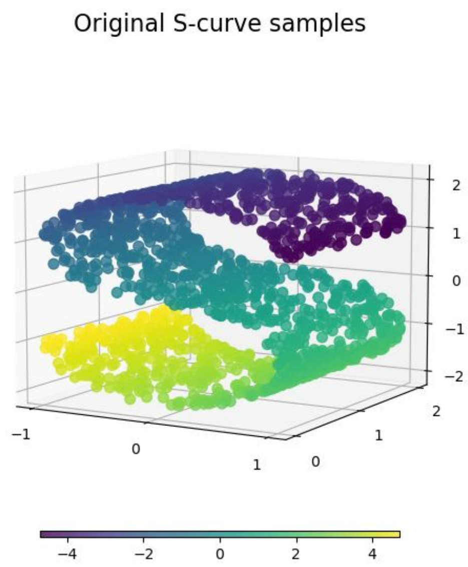


Correlation-Based Feature Selection to Identify Functional Dynamics in Proteins
Georg Diez, et al (2022)

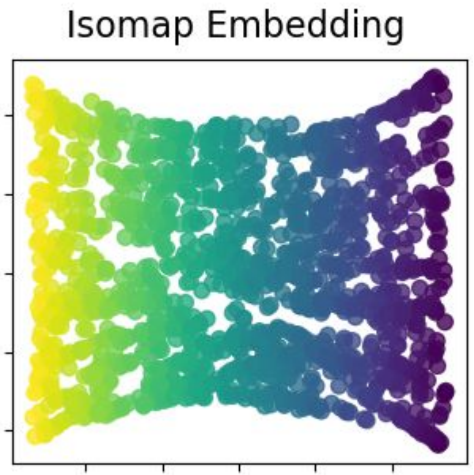
Problem #1: I want to find structure in my data…

Dimensionality reduction tasks

1. You believe that the data in your possession can be mapped to a lower dimensional embedding



2. Need to preserve certain properties
   a. Total variance (PCA)
   b. Total Autocorrelation (tICA)
   c. Pairwise distances between points (Multidimensional Scaling (MDS))
   d. Geodesic distances (Diffusion maps / Isomaps)
   e. Reconstruction (Autoencoders)

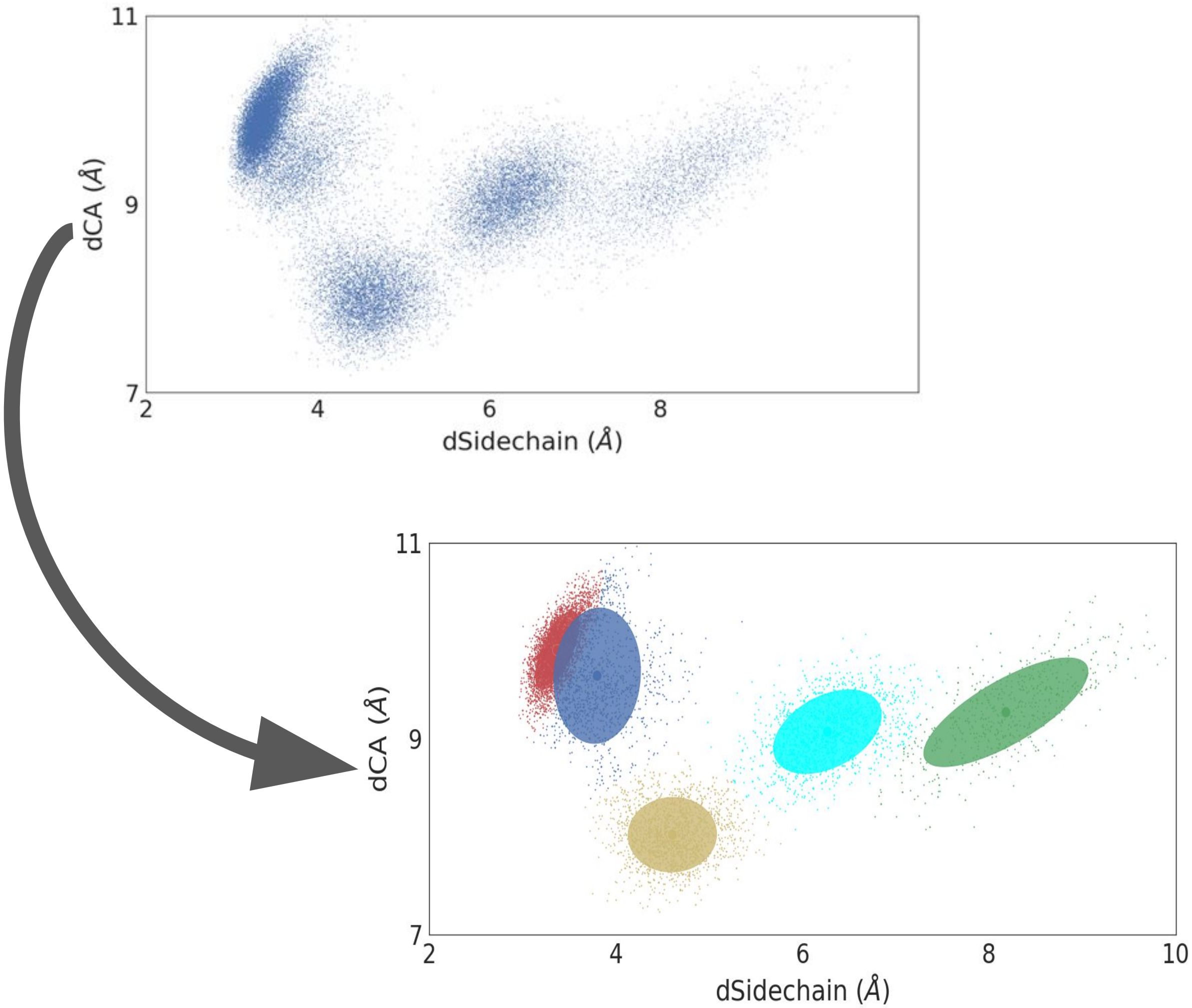


*scikit-learn

Problem #2: I want to find distinct groups in my data / “segmentation”…

Clustering tasks

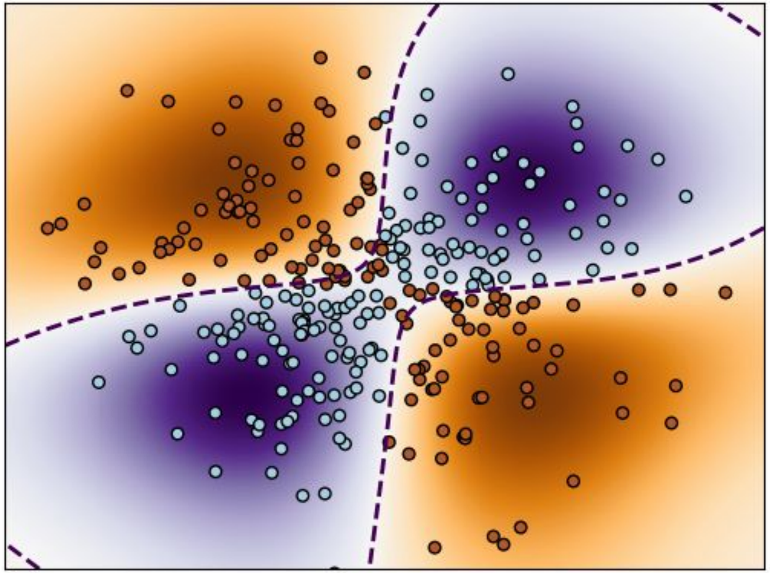
1. When you observe separate groups in your data distribution with similar properties
2. Methods:
    a. K-means clustering and cousins
    b. Gaussian Mixture models (GMMs)
    c. Density based methods (DBSCAN, OPTICS, HDBSCAN)



Problem #3: I know there are categories in my data, but I want to find what separates them…

Classification tasks

1. Class labels for the data are known and the goal is to find a ‘decision boundary’ i.e. a function that separates the labels from each other.
2. Methods:
   a. Naive Bayes classifier
   b. Support vector machines
   c. Neural Nets based classification

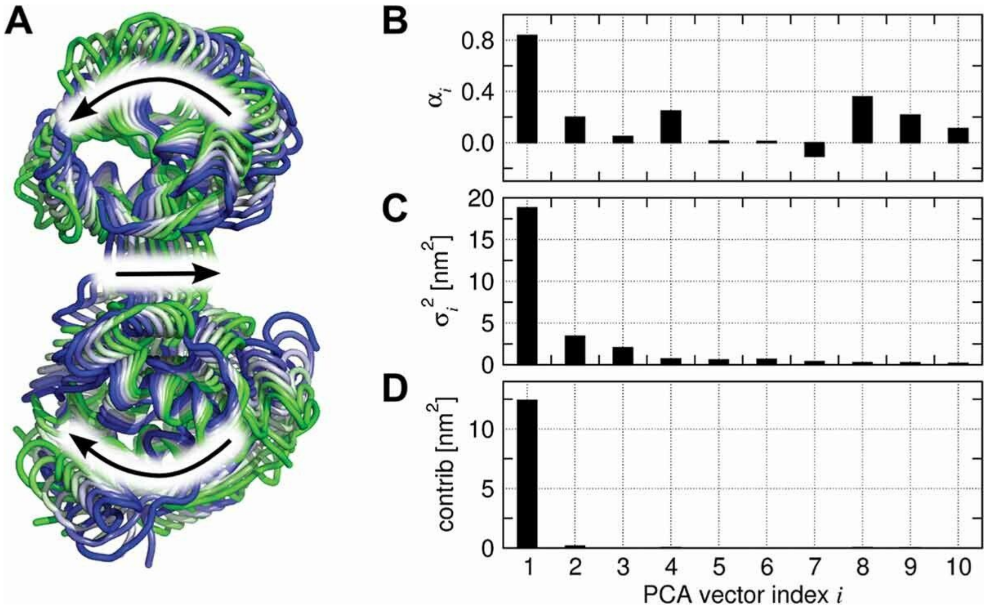


Classification with Support Vector machines *scikit-learn

Problem #4: I want to know how my data connects to another variable…

Regression tasks

1. For each member of the data a continuous label is known
2. Methods:
    a. Multilinear regression (MLR)
    b. Principal Component Regression (PCR)
    c. Partial Least Squares regression (PLS)
    d. Neural Nets based regression etc



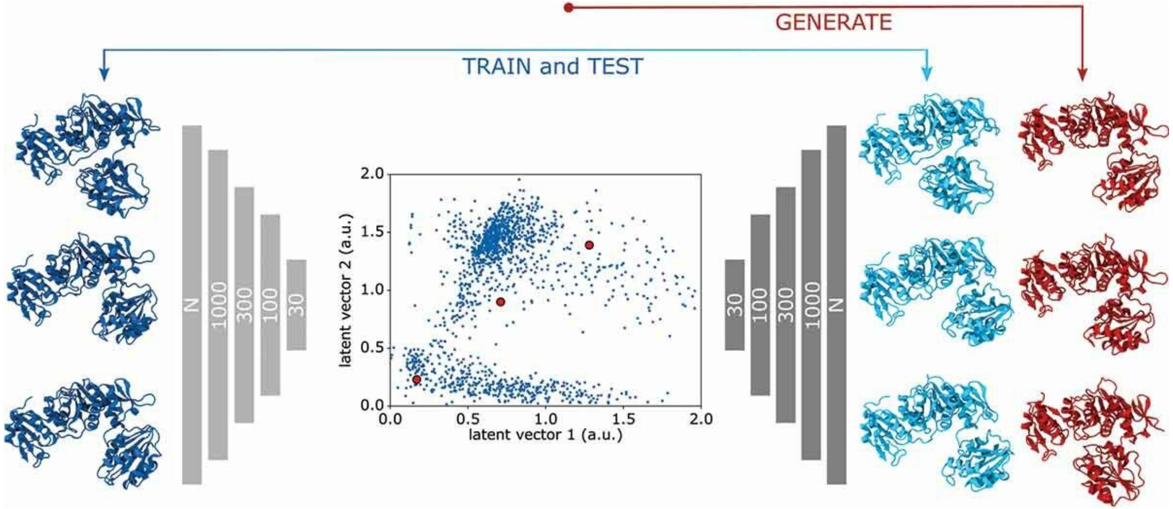
Principal Component Regression (PCR). A. PCR-based ensemble-weighted mode for the Leucine Binding Protein. **B**. Coefficient αi of the contribution to the PCR model from the largest PCA eigenvectors. **C**. Eigenvalues of the PCs used to construct the PCR model. **D**. Contribution of the variance of the PCs to the variance of the collective mode.

* Hub, Jochen S et al “Detection of functional modes in protein dynamics.” (2009)

Deep Learning based methods

Artificial Neural Nets

1. Extremely powerful prediction engines
2. Model relations as a non-linear function
3. Fantastic generative powers
4. Caveat: Might be hard or even impossible to interpret



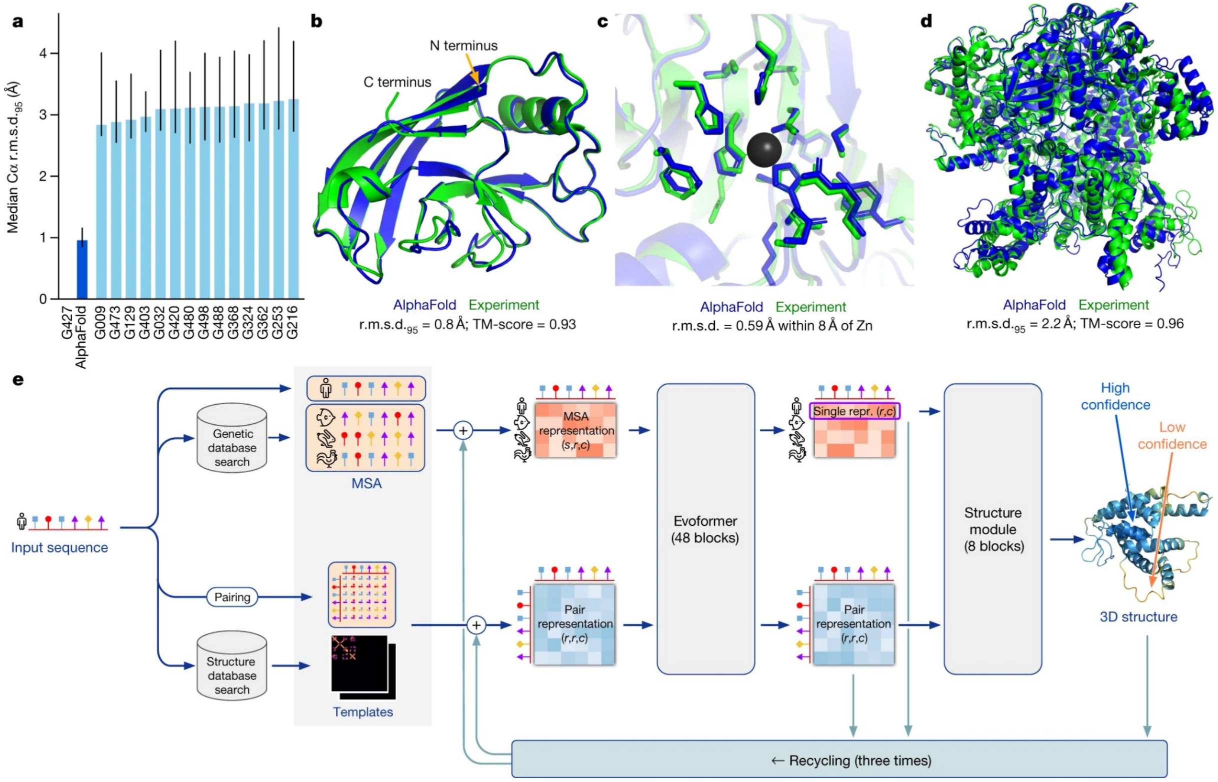
ANNs for Dimensionality Reduction. Autoencoder networks (shown in gray) are used for training a lower-dimensional representation of the simulation data by reconstructing sampled structures (deep blue) with decoded structures (light blue). A trained autoencoder can be used to generate a latent space representation of the data set (blue points), which can used to generate unseen latent space data (red points) to mine unsampled structures (red structures).

Degiacomi M., “Coupling Molecular Dynamics and Deep Learning to Mine Protein Conformational Space,” (2019

AlphaFold (2)

Based on a Transformer architecture, AlphaFold2 can predict structures and oligomeric interfaces with very high confidence starting only from the sequence/s.

Jumper, J., Evans, R., Pritzel, A. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583−589 (2021). https://doi.org/10.1038/s41586-021-03819-2


a
Median Cα r.m.s.d.$_{95}$ (Å)
0
1
2
3
4
G427
AlphaFold
G009
G473
G129
G403
G032
G420
G480
G498
G488
G368
G324
G362
G253
G216
b
N terminus
C terminus
AlphaFold Experiment
r.m.s.d.$_{95}$ = 0.8 Å; TM-score = 0.93
c
AlphaFold Experiment
r.m.s.d. = 0.59 Å within 8 Å of Zn
d
AlphaFold Experiment
r.m.s.d.$_{95}$ = 2.2 Å; TM-score = 0.96
e
Input sequence
Genetic database search
MSA
Pairing
Structure database search
Templates
MSA representation (s,r,c)
Pair representation (r,r,c)
Evoformer (48 blocks)
Single repr. (r,c)
Pair representation (r,r,c)
Structure module (8 blocks)
High confidence
Low confidence
3D structure
← Recycling (three times)

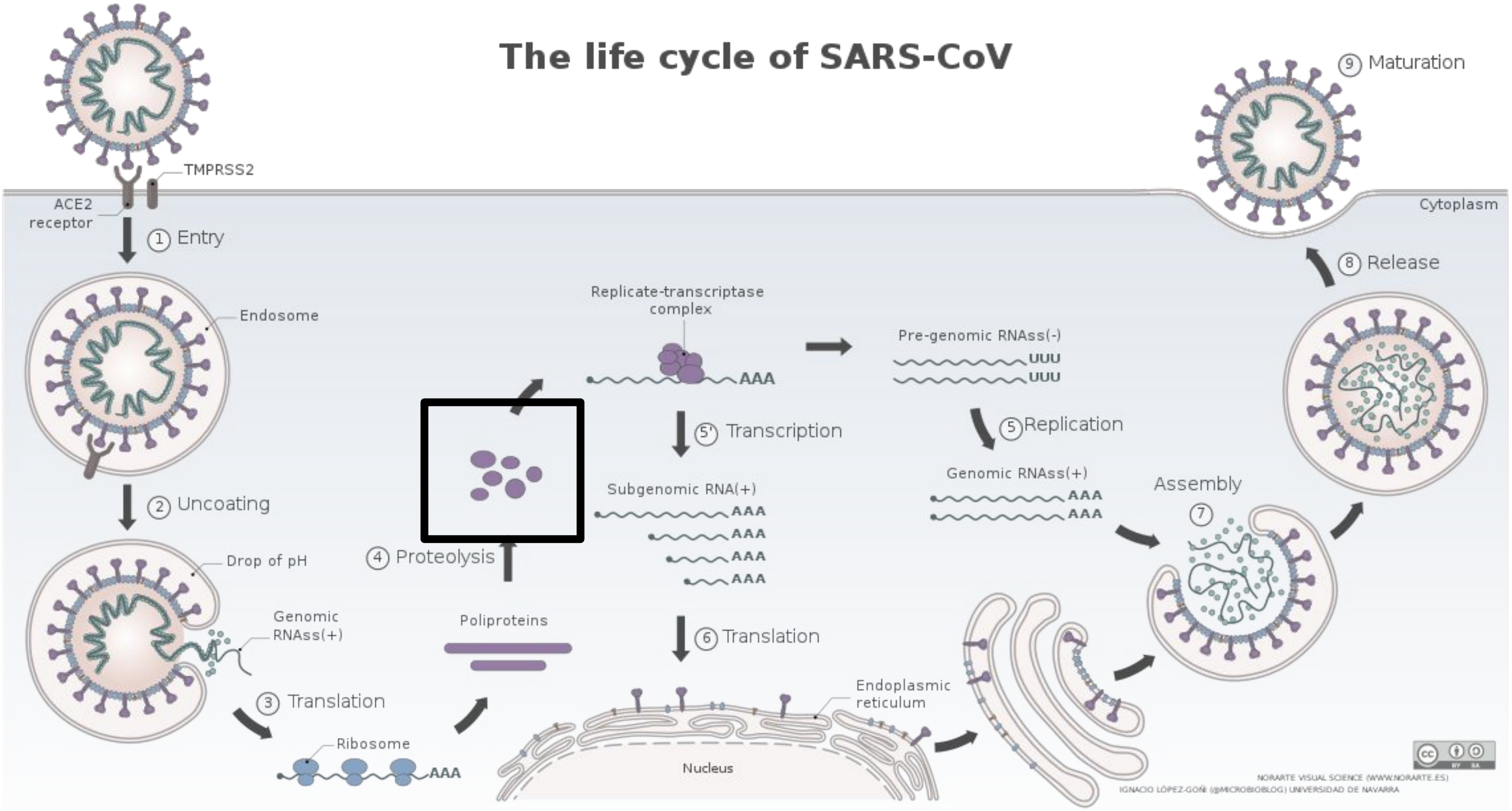
An example application

How to stop the Main Protease of the Covid from working?

Kaptan, Shreyas et al. “Maturation of the SARS-CoV-2 virus is regulated by dimerization of its main protease.” *Computational and structural biotechnology journal* vol. 20 (2022): 3336-3346. doi:10.1016/j.csbj.2022.06.023

How does Covid infect the cell?

* WIKIPEDIA



The life cycle of SARS-CoV

NORARTE VISUAL SCIENCE (WWW.NORARTE.ES)
IGNACIO LÓPEZ-GOÑI (@MICROBIOBLOG) UNIVERSIDAD DE NAVARRA

Introduction: Why M^{pro}?

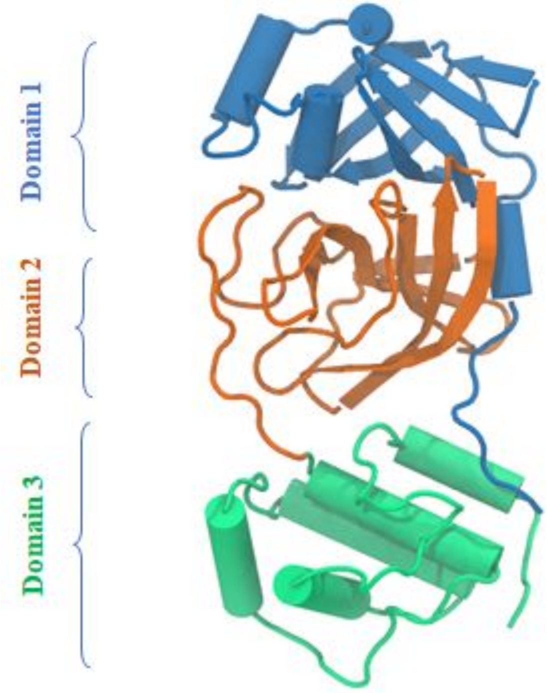
1. Unique sequence identity of the substrate. It is not present in humans.

2. Virus can not mature if the enzyme is disabled.

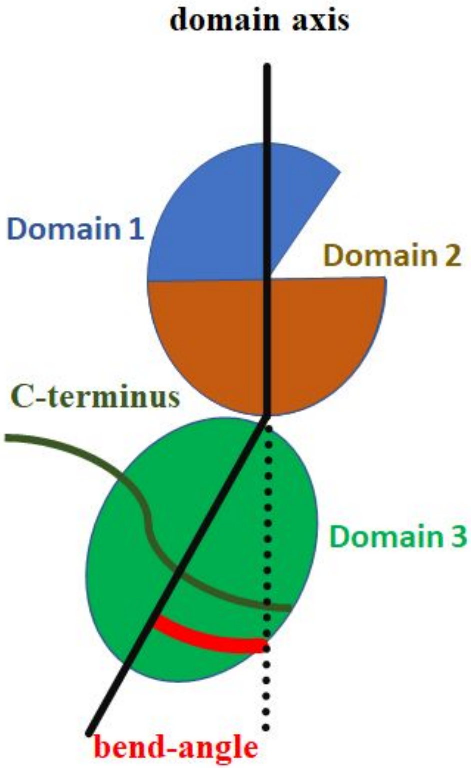
3. Similar enzymes have pre-existing drugs that can be repurposed

What does Mpro look like?

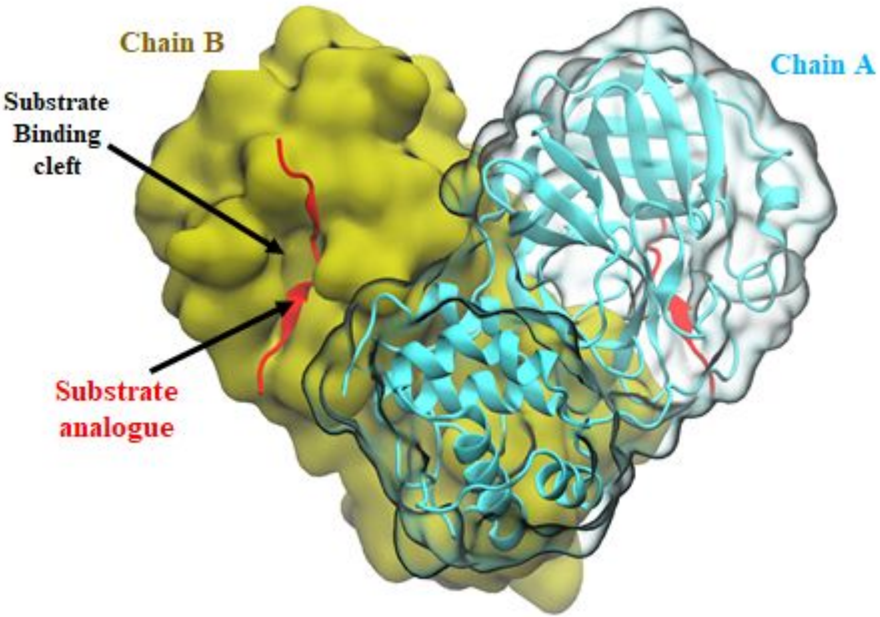
Three domain monomeric structure



What does Mpro look like?



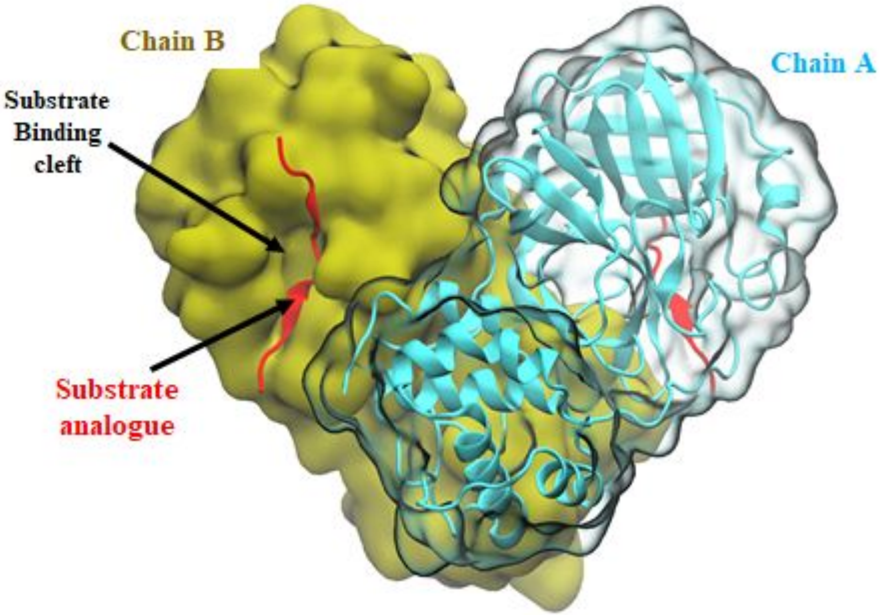
It forms a dimer…



Questions! How do we estimate enzyme activity?

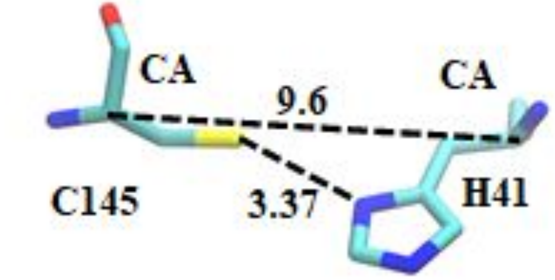
Can the substrate bind tightly?

Are the catalytic residues in the right orientation?



Featurization: catalytic residues

Distance between the catalytic residue-mainchain (C-alpha atoms)

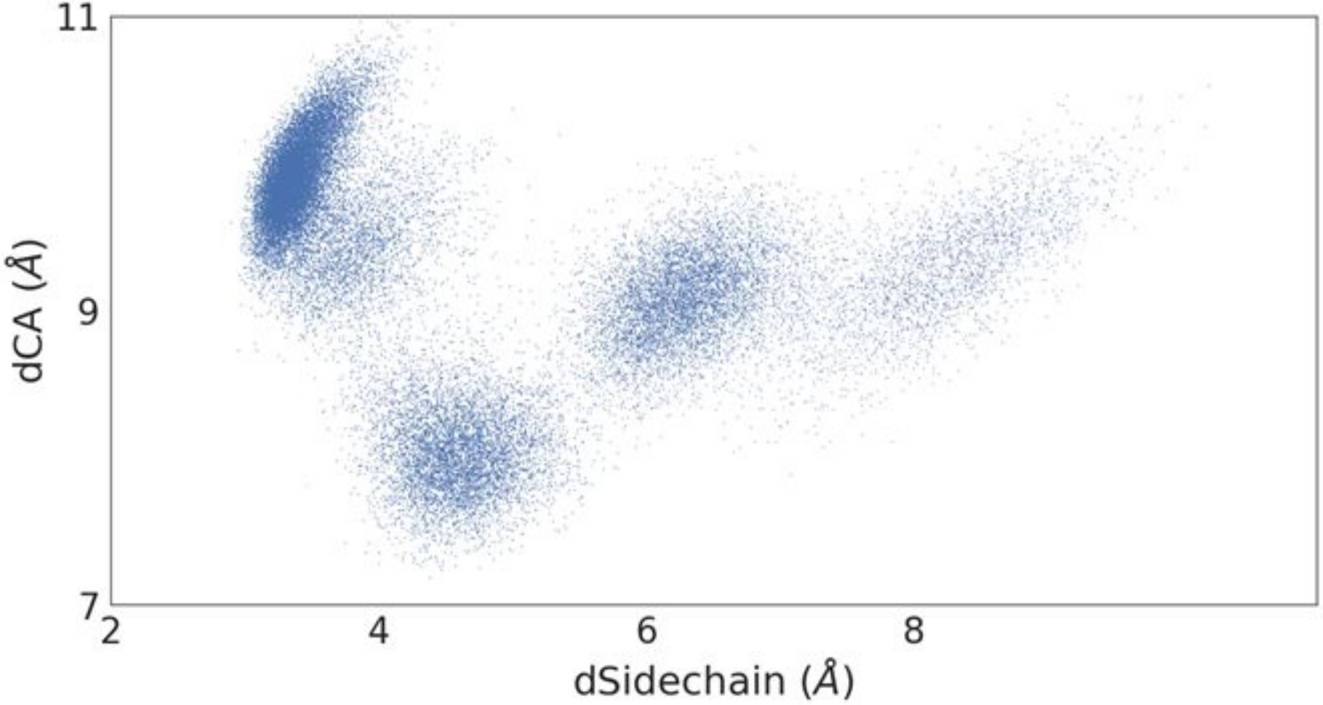


Distance between the catalytic residue-sidechains

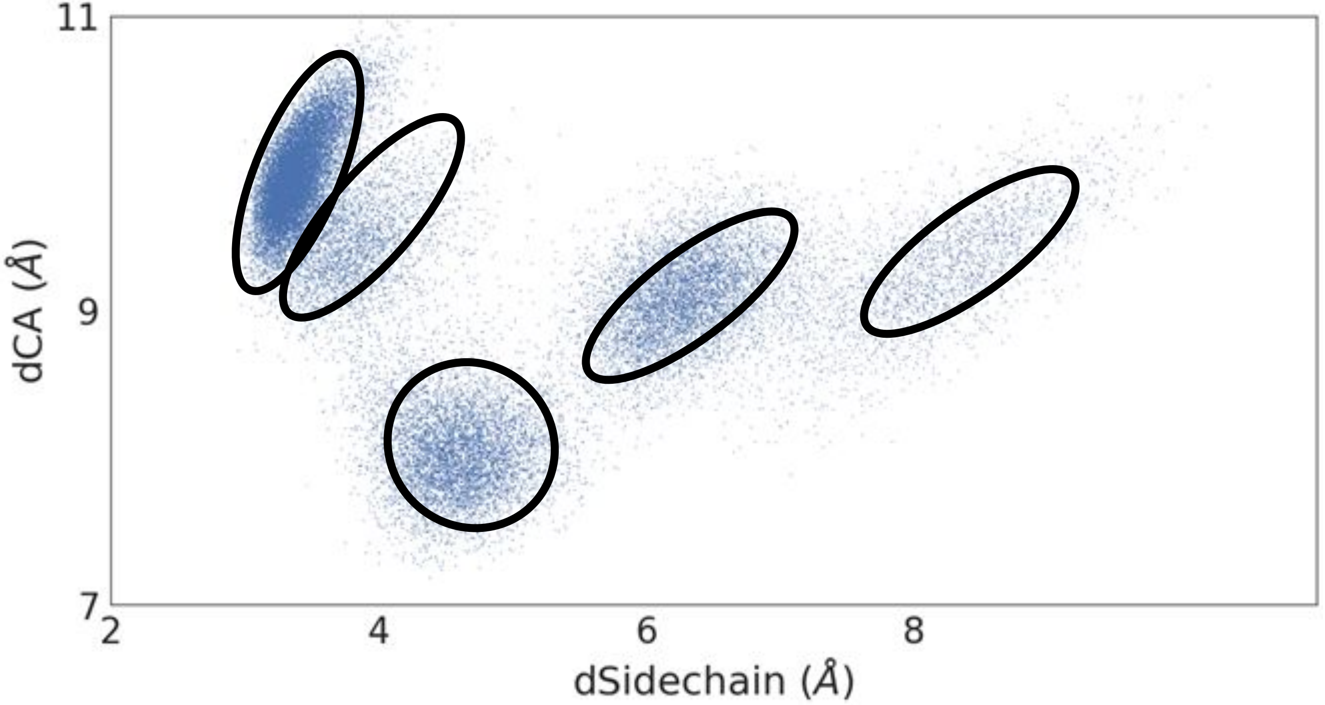
Detects the orientation space in a simple yet sufficient description

After a lot of simulations…

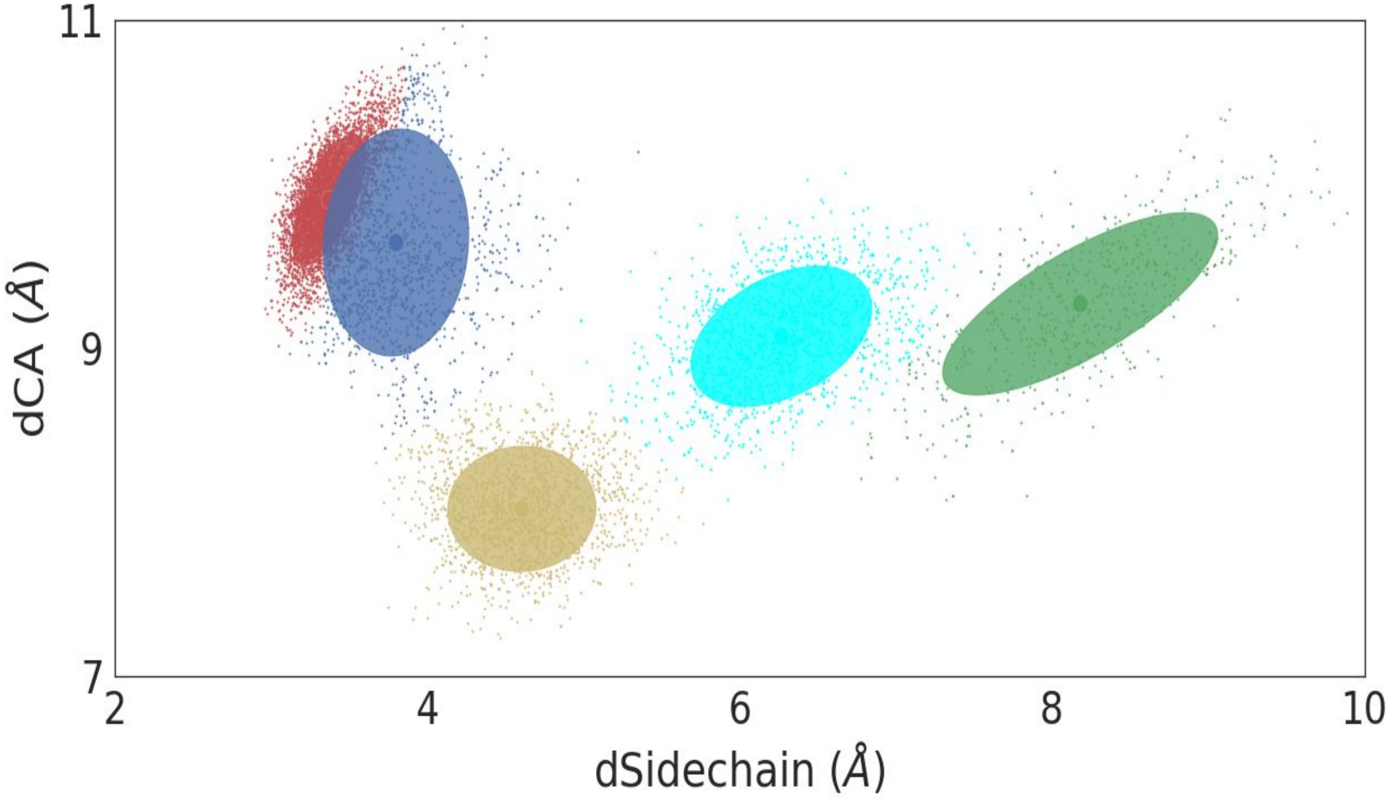
A random sample of above features in 2D



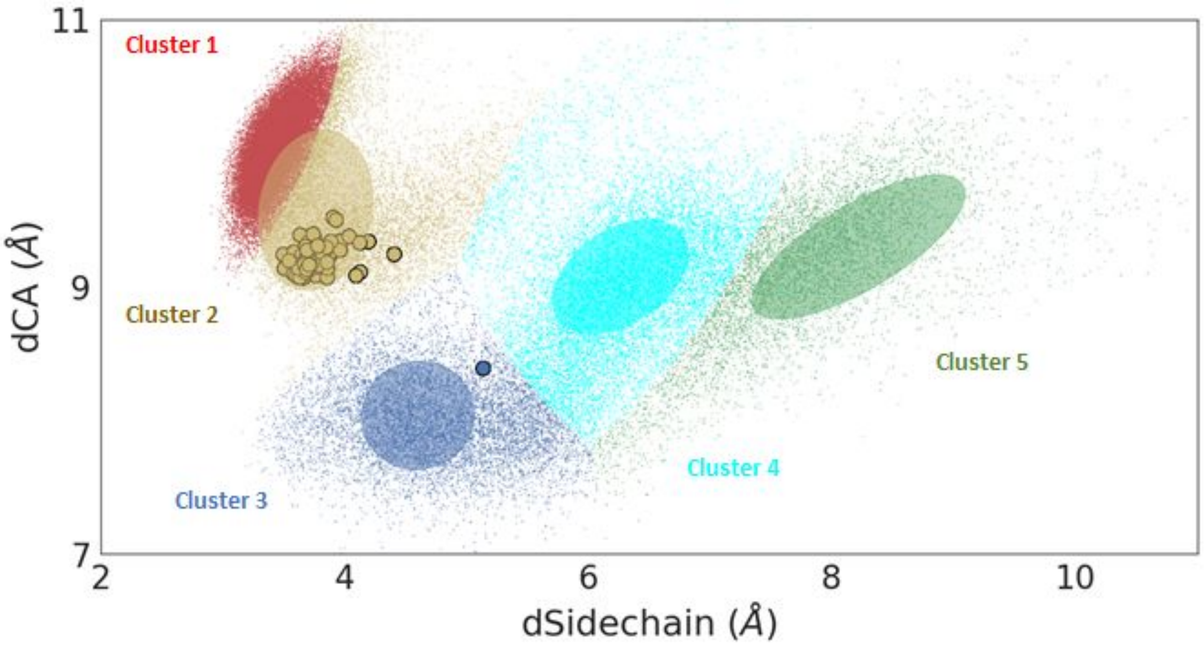
Gaussian mixture model: Clusters?

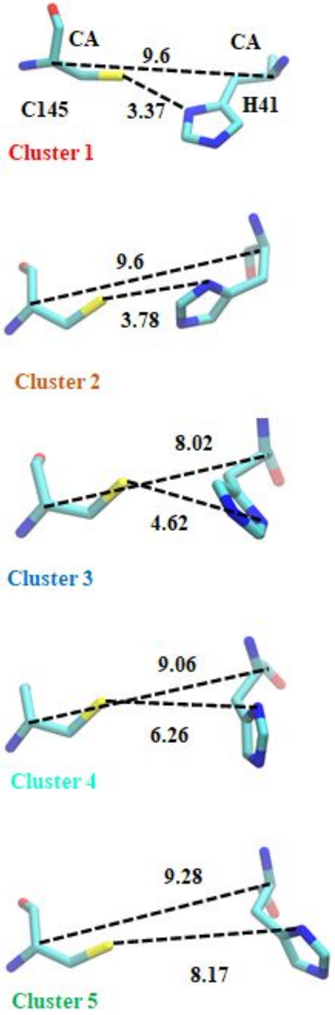


Gaussian mixture model: Clusters?



GMM interpretation





Functional Mode Analysis

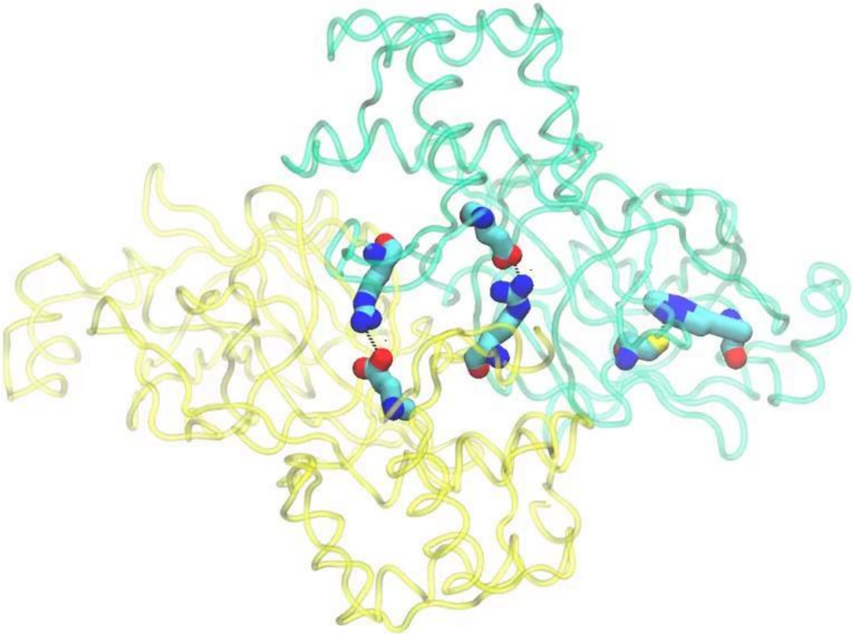
Partial Least Squares based Functional Mode Analysis (PLS-FMA)

Goal: Build a linear model ($\boldsymbol{f} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$) to express a function ($\boldsymbol{f}$) in terms of the Coordinates $\boldsymbol{X}$ with linear coefficients $\boldsymbol{\beta}$ and error/residuals $\boldsymbol{\epsilon}$.

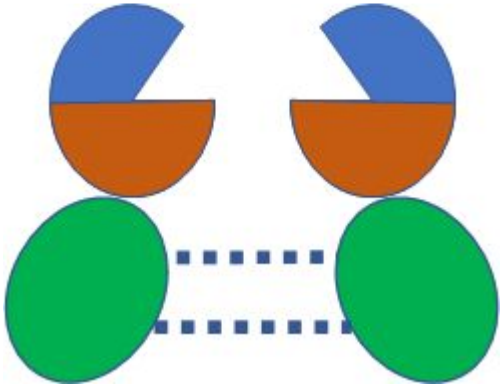
In PLS , k new regressors $\boldsymbol{T}k$ are defined iteratively such that each coordinate is a linear combination of the original coordinates, $\boldsymbol{X}$ ($\boldsymbol{T}k = \boldsymbol{XW}k$) with maximal covariance with $\boldsymbol{f}$, while being uncorrelated to each previous coordinate in $\boldsymbol{T}k$. Subsequently, the regression problem $\boldsymbol{f} = \boldsymbol{XW}k\boldsymbol{\alpha}k + \epsilon$ is solved using $\boldsymbol{XW}k$ as basis.

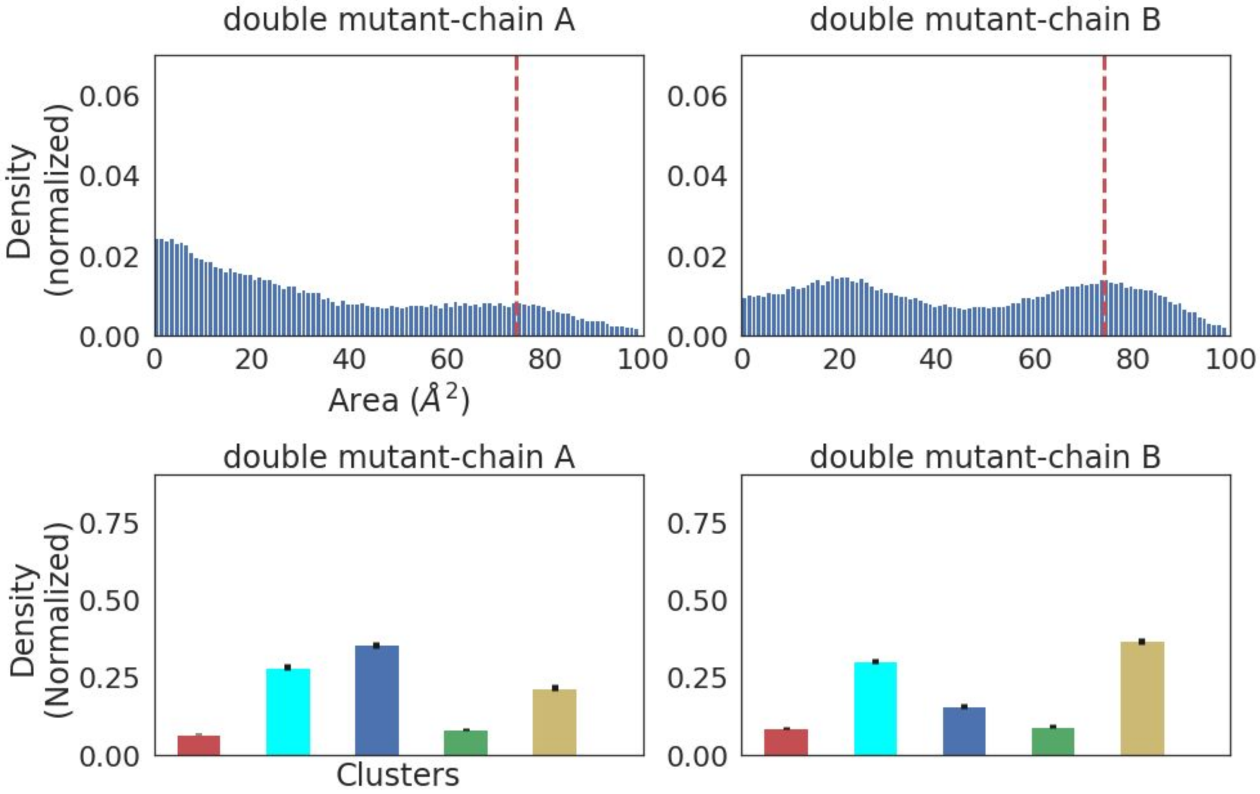
Krivobokova, Tatyana, Rodolfo Briones, Jochen S Hub, Axel Munk, and Bert L de Groot. 2012. “Partial Least-Squares Functional Mode Analysis: Application to the Membrane Proteins {AQP}1, Aqy1, and {CLC}-Ec1.” *Biophysical Journal* 103 (4): 786–96.

Machine model based on dSidechain

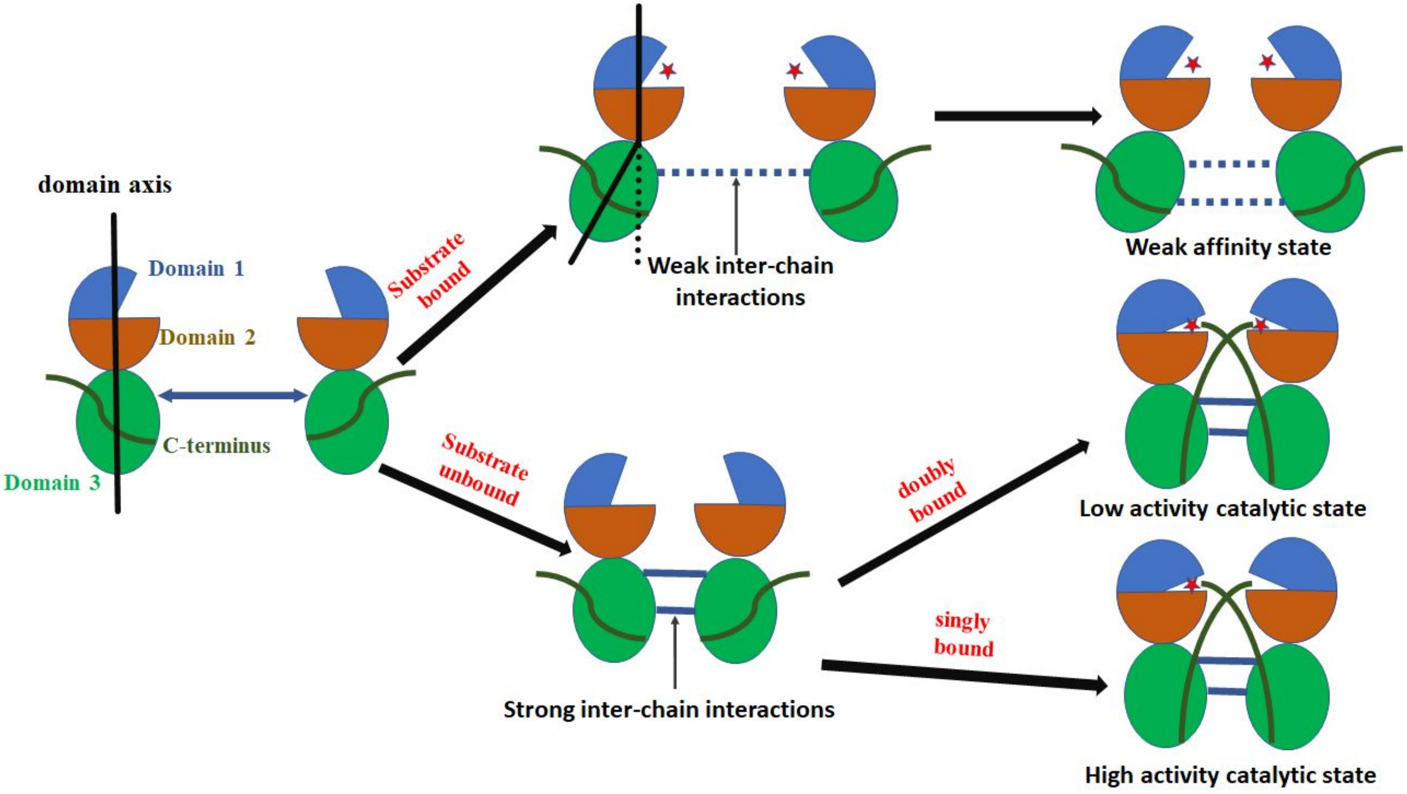


Are machine-model results causative?





Schema for the mechanism of the dimerization



Conclusions

1. M^{pro} monomer is catalytically active and yet has low affinity for the ligand
2. Dimerization makes the enzyme obtain high substrate affinity
3. The singly bound state of the dimer has both high substrate affinity and catalytic activity
4. The buried salt-bridge pair between the dimers regulates the activation of the enzyme

Final Conclusions for the choosing ML models

1. Understand the data generation process really well. It will give you crucial insights into choice of techniques / feature selection.
2. Don’t use ML models as a Black Box. It might easily lead to GIGO issues.
3. Always cross-validate to test for overfitting
4. ML methods are *data driven*. If the model is not “captured” by your data, it can not be extracted from the data. I.e. ML methods are not a substitute for good sampling!

Acknowledgements

CSC

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Prof. Dr. Ilpo Vattulainen

Mikhailo Gyrich for running the simulations

Dr. Vivek Sharma, Dr. Giray Enkavi, Dr. Waldemar Kulig and Dr. Thomasz Rog for discussions and analysis

CSC for the computational time.

University of Helsinki and the MSCA for the funding!

MARIE CURIE ACTIONS

For listening intently…

You!