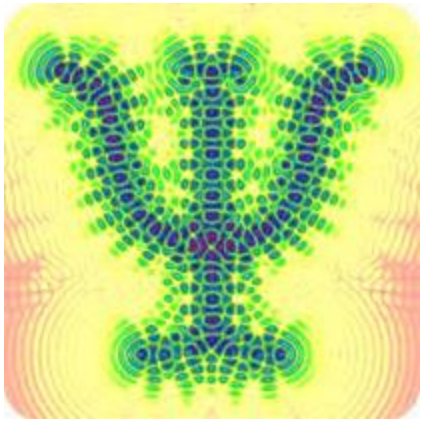
Understanding color tuning rules in the Arch family with a fully computational QM/MM approach

RESMOL GROUP



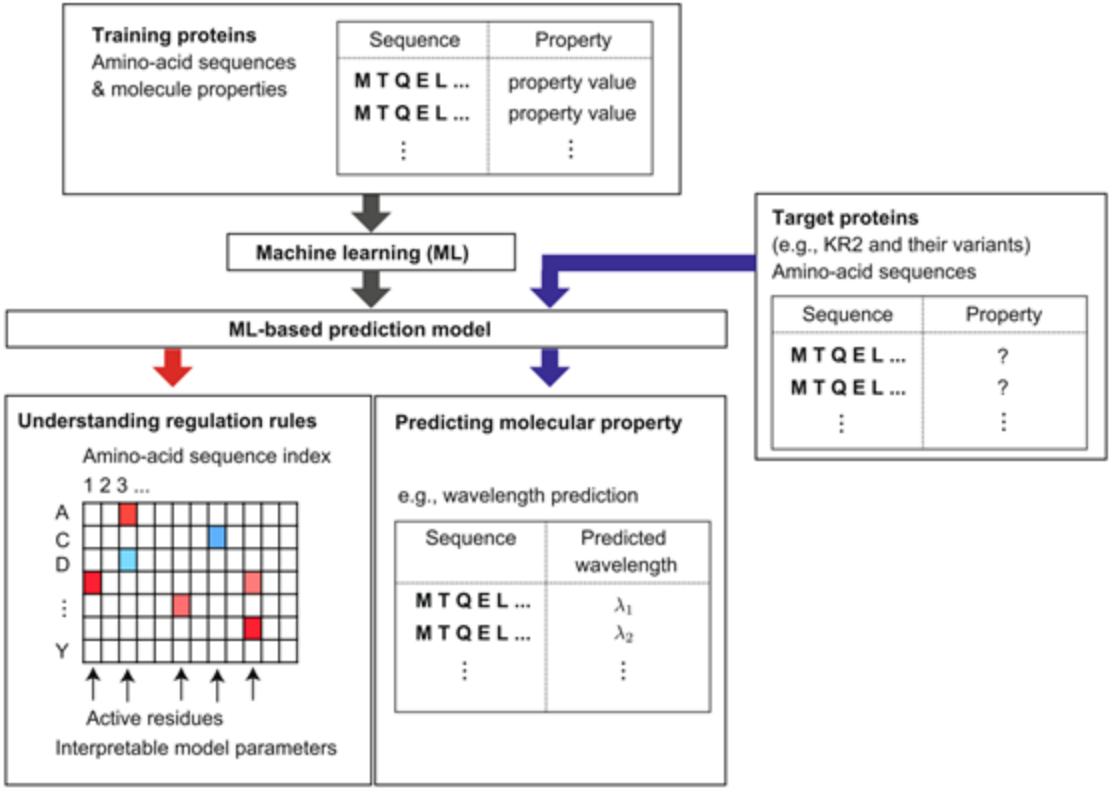
HPC project

UNIVERSITÀ DI SIENA 1240

- Data-driven determination of Arch3 variants excitation energies from gene sequences
- Self-consistent computational approach
- Understanding color tuning determinants in the Arch family
- Identifying Arch red-shifted candidates for optogenetics

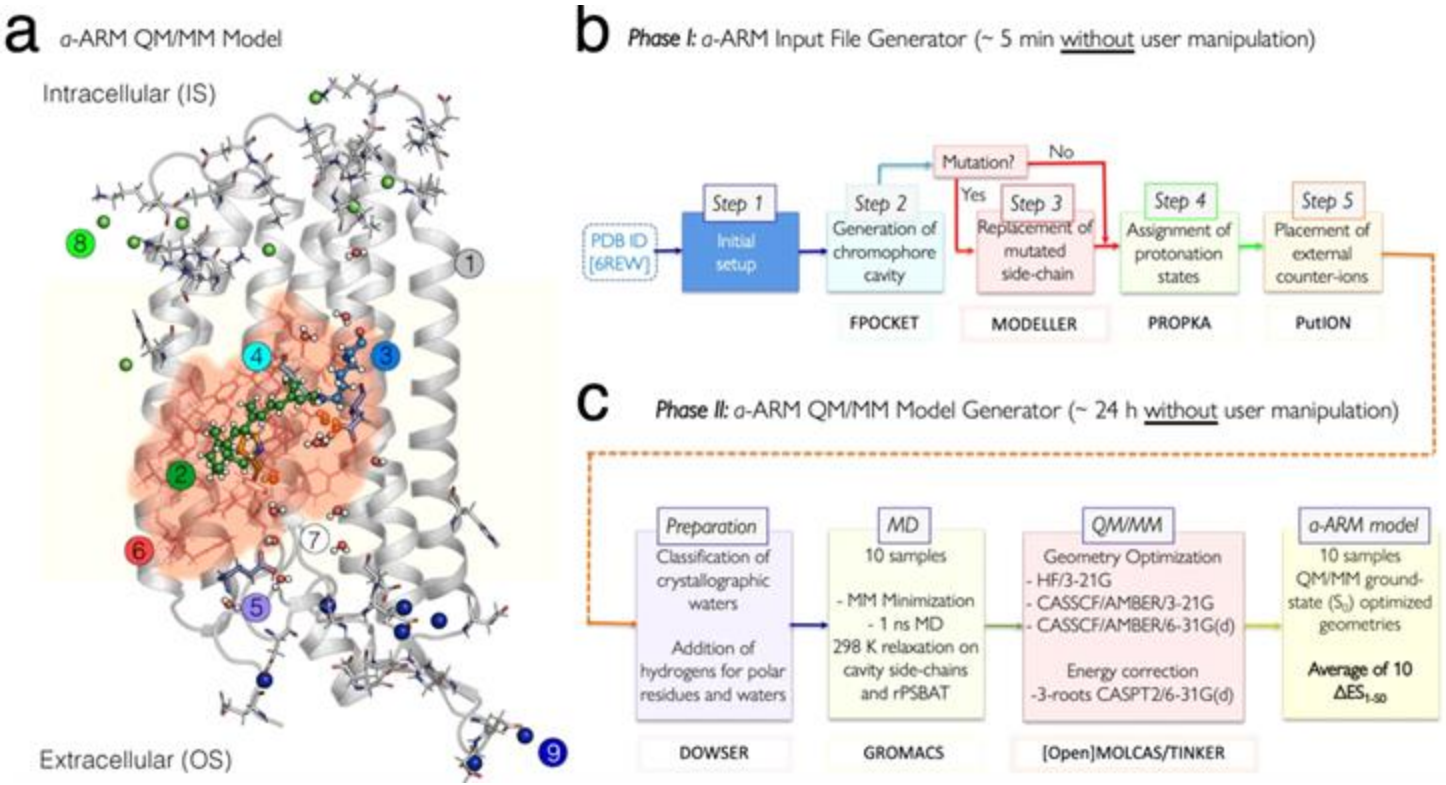
HPC project

- Data-driven determination of excitation energies from gene se



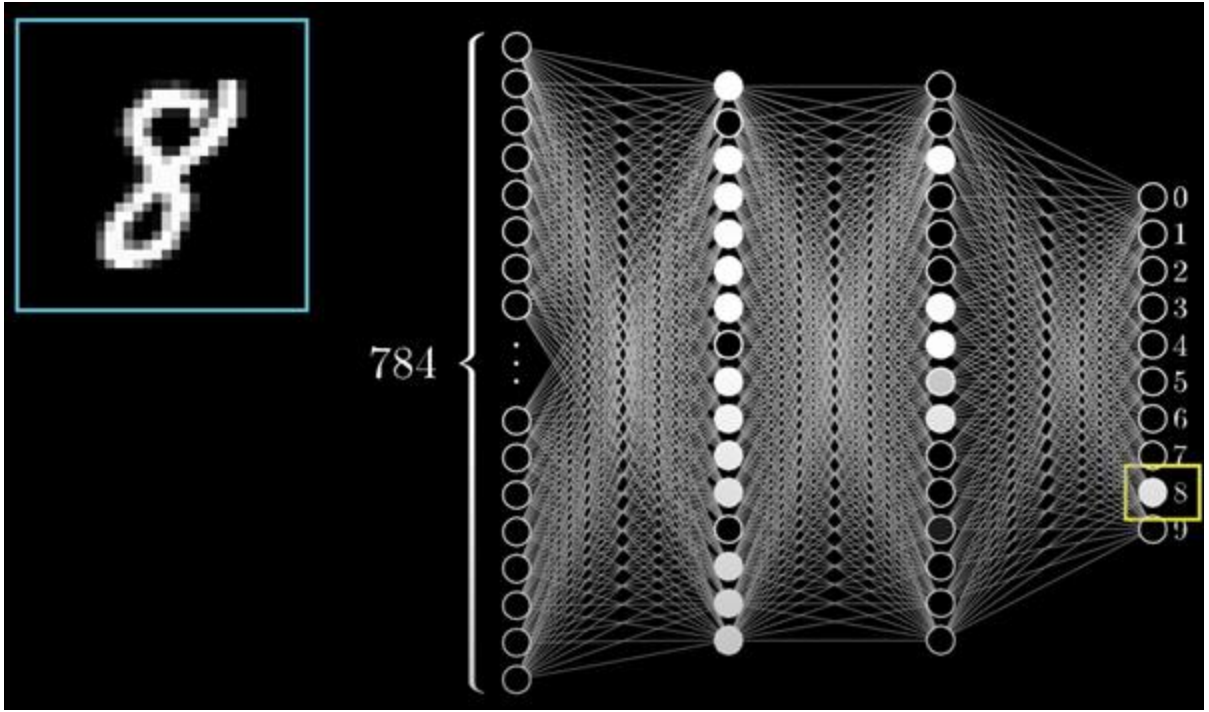
HPC project

- Self-consistent computational approach



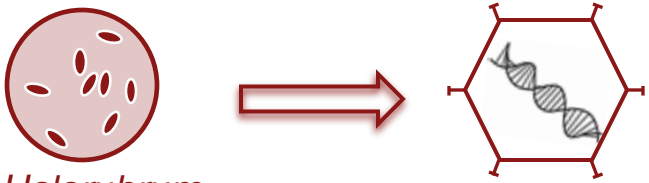
HPC project

- Understanding of color tuning determinants in the Arch family



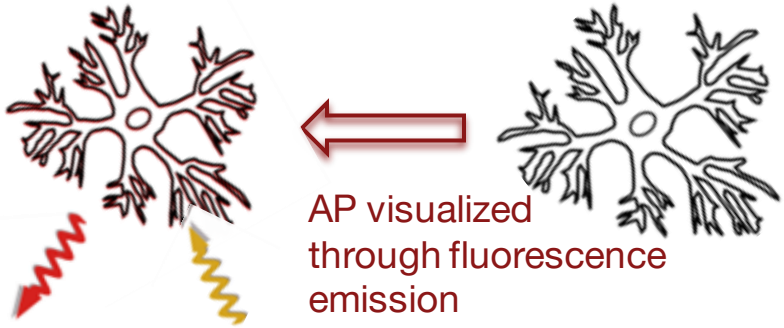
HPC project

- Identifying Arch3 red-shifted candidate for optogenetics

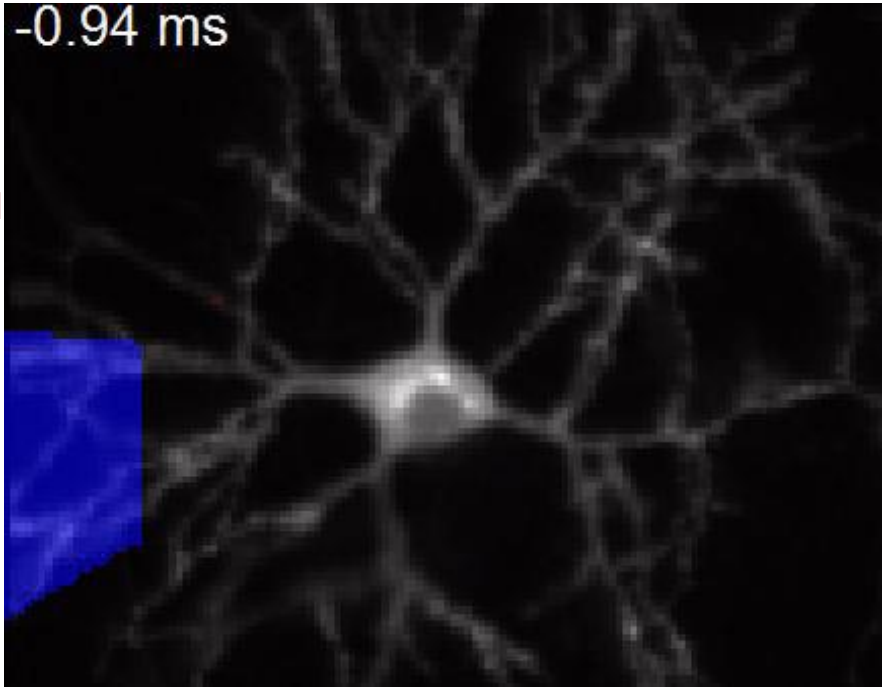


Halorubrum sodomense expressing Arch3

Gene introduced and expressed in neurons

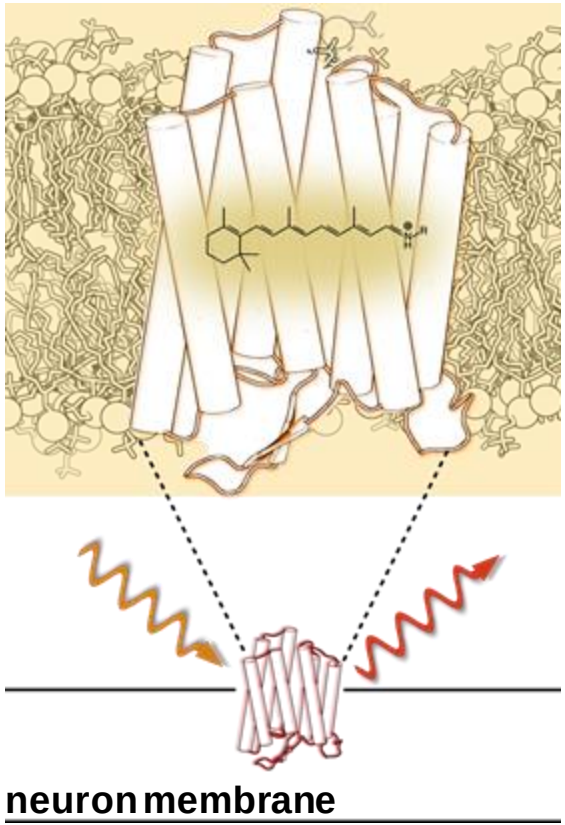


AP visualized through fluorescence emission



Hochbaum, D *et al*. *Nature methods* 11, 825, **2014**.

Optogenetics

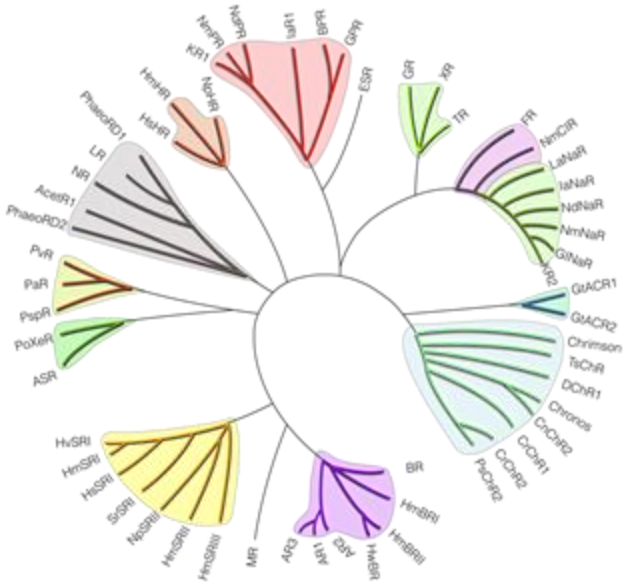


Sensors

- Spectral orthogonality with actuators
- Deep tissue penetration
- Low phototoxicity
- Long fluorescence lifetime
- High voltage sensitivity

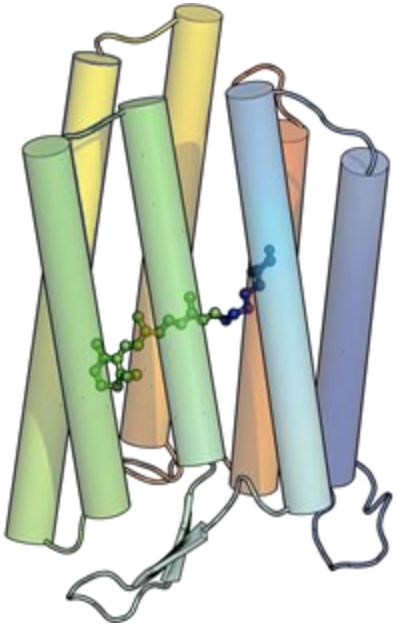
Rhodopsin family

- Huge phylogenetic variability

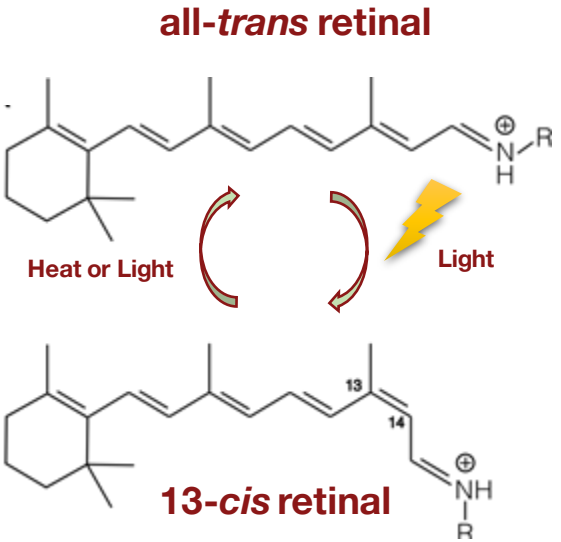


Karasuyama, M *et al. Scientific reports* 8,15580, 2018.

- Opsin family
- 7 transmembrane helices
- Retinal chromophore



- Light-induced photoisomerization



Arch3

UNIVERSITÀ DI SIENA 1240

Gradinaru, V. (CalTech) McIsaac, R. S et al. Directed evolution of a far-red fluorescent rhodopsin. *Proceedings of the National Academy of Sciences* **2014**, *111*, 13034-13039.

Arnold, F. H. (CalTech) McIsaac, R. S. et al. Recent advances in engineering microbial rhodopsins for optogenetics. *Curr. Opin. Struct. Biol.* **2015**, *33*, 8-15.

Cohen, A. E. (Harward) Hochbaum, D. R. et al. All-optical electrophysiology in mammalian neurons using engineered microbial rhodopsins. *Nat. Methods* **2014**, *11*, 825-825.

Boyden, E. S. (MIT) Piatkevich, K. D. et al. A robotic multidimensional directed evolution approach applied to fluorescent voltage reporters. *Nat. Chem. Biol.* **2018**, *14*, 352-360.

Protein	$\Delta E^{a,Exp}_{S1\text{-}S0}$ (kcal mol^{-1})	$\lambda^{a,Exp}_{max}$ (nm)	$\Delta E^{f,Exp}_{S1\text{-}S0}$ (kcal mol^{-1})	$\lambda^{f,Exp}_{max}$ (nm)	FQY	Ref.
Arch2	51.5	555	NR[a]	NR[a]	NR	22
Arch3	51.4	556	41.6 [b]	687[b]	1–9 × 10^{-4}	23
QuasAr1	49.3	580	40.0, 38.6	715, 740	6.5 × 10^{-3}, 8.0 × 10^{-3}	24, 25
Archon2	48.8	586	38.9	735	1.1 × 10^{-2}	26
QuasAr2	48.5	590	40.0	715	4.0 × 10^{-3}	24
Arch7	46.4	616	39.3	727	1.2 × 10^{-2}	23
Arch5	46.0	622	39.1	731	8.7 × 10^{-3}	23

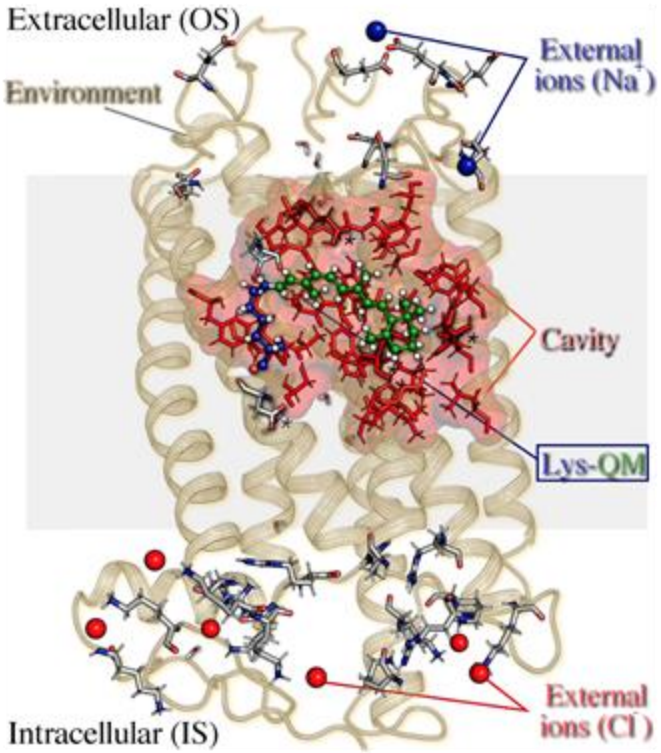
[a] Not Reported

[b] The λ^{f}_{max} reported is likely coming from the Q-intermediate rather than from the dark-adapted state (see ref. [27])

The a-ARM protocol

UNIVERSITÀ DI SIENA 1240

Phase I: Model Structure setup AUTOMATED setup



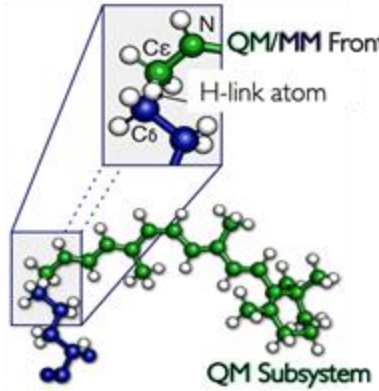
Phase II: QM/MM calculation setup AUTOMATED setup

MM Subsystem
Protein environment (*aa*, ions and waters)
Fixed at crystallographic/homology geometry

MM Subsystem
Chromophore cavity (*aa* and waters)
MM optimization(AMBER ff)

GROMACS FAST. FLEXIBLE. FREE.

TINKER



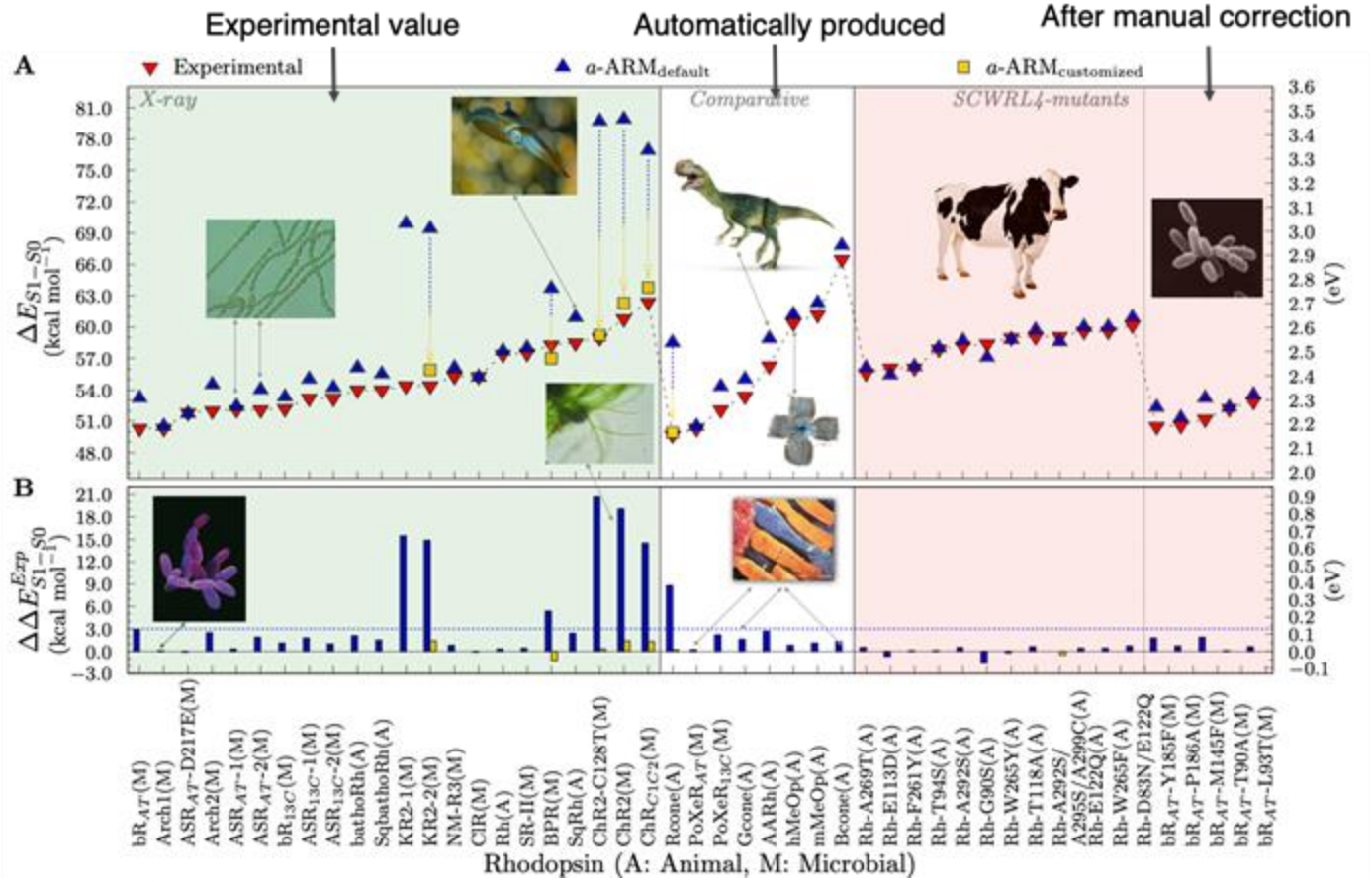
MOLCAS

SA-CASPT2/CASSCF(12,12)/ 6-31G(d)//AMBER

Electrostatic Embedding (ESPF), AMBER ff (except frontier)

Pedraza-González, L.; De Vico, L.; Marín, M. D. C.; Fanelli, F.; Olivucci, M.. *J. Chem. Theory Comput.* **2019**, *15*, 3134-3152.

The a-ARM protocol



Pedraza-González, L.; De Vico, L.; Marín, M. D. C.; Fanelli, F.; Olivucci, M.. *J. Chem. Theory Comput.* **2019**, *15*, 3134-3152.

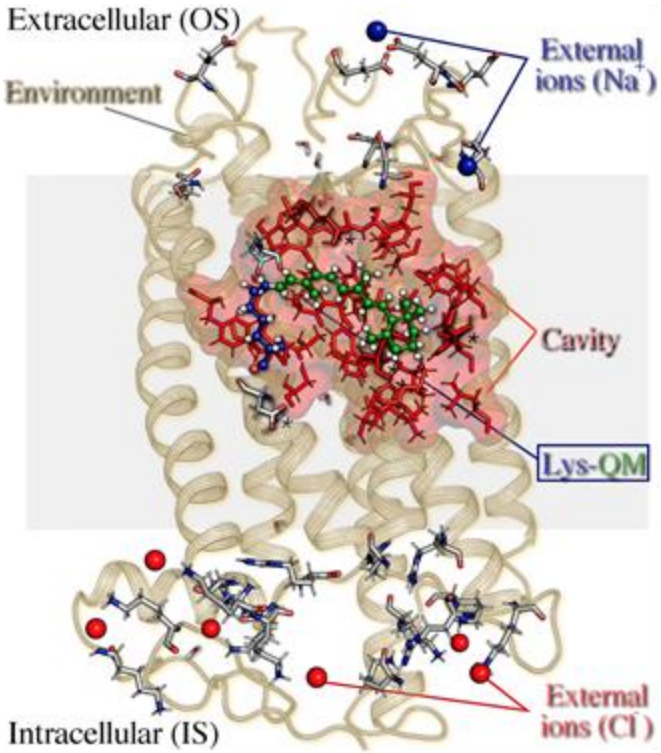
$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} = \frac{\sum_{i=1}^{n} |e_i|}{n}$$

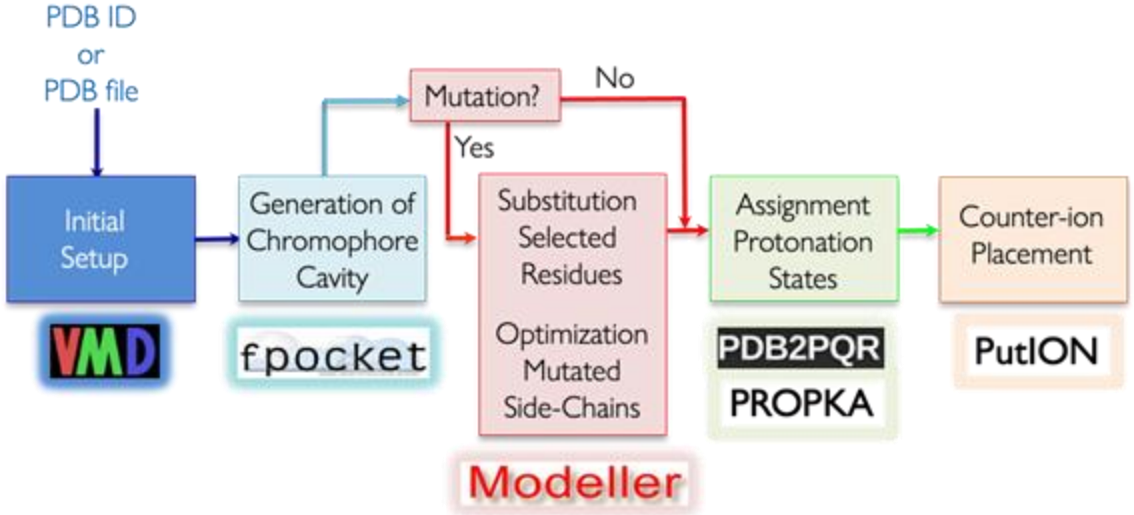
MAE for a-ARM is 0.9 kcal/mol

The a-ARM protocol

UNIVERSITÀ DI SIENA 1240

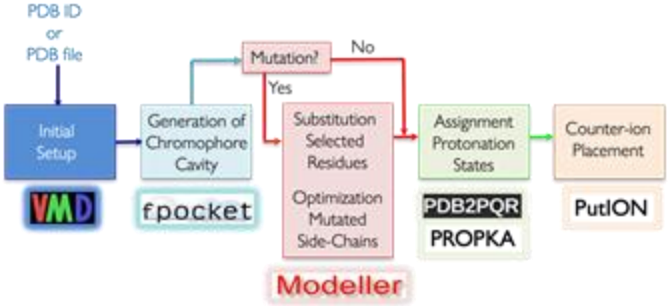
Phase I: Model Structure setup | **AUTOMATED setup**

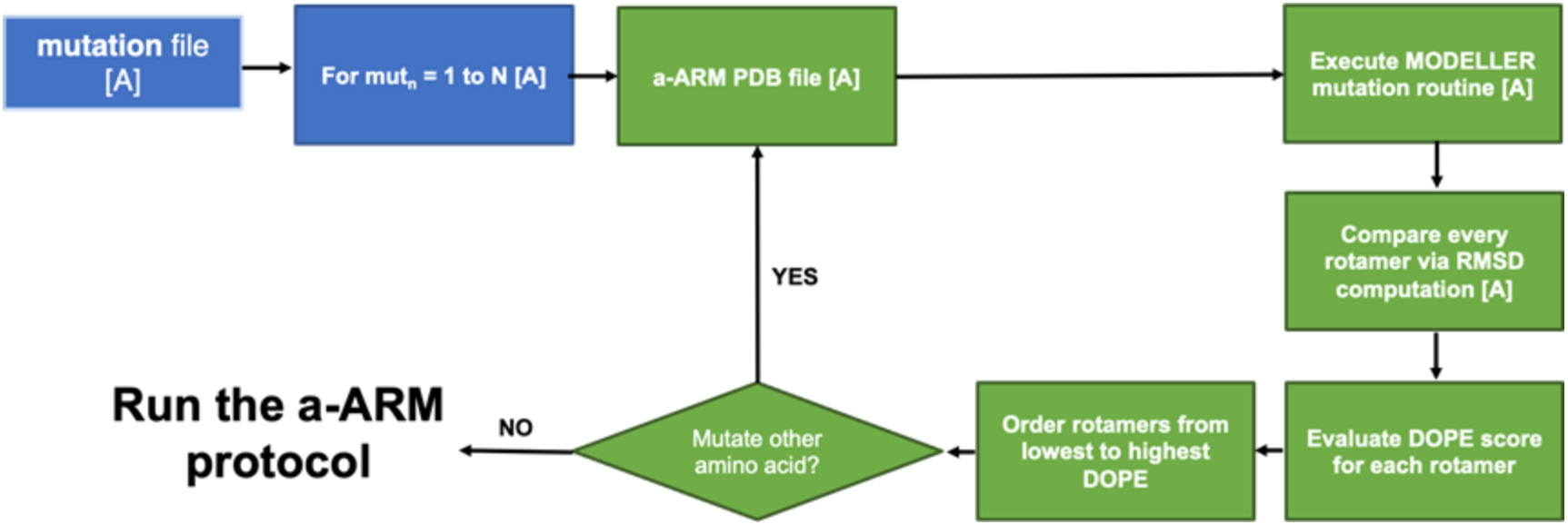




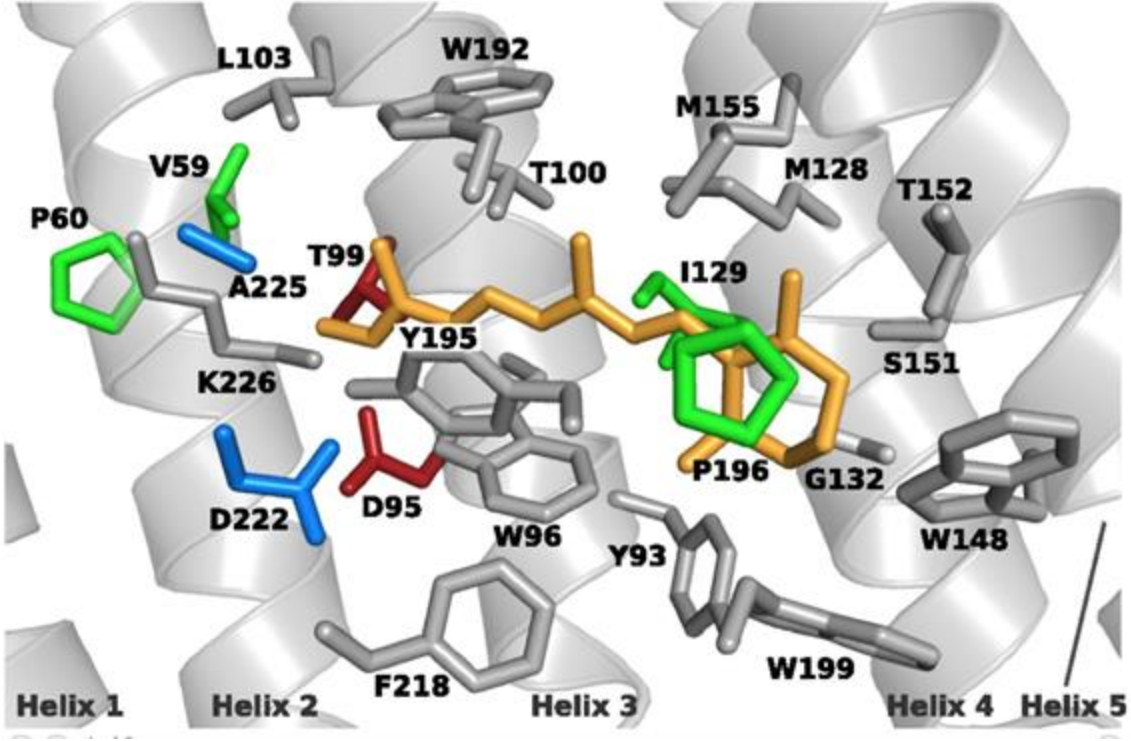
Pedraza-González, L.; De Vico, L.; Marín, M. D. C.; Fanelli, F.; Olivucci, M.. *J. Chem. Theory Comput.* **2019**, *15*, 3134-3152.

Design of the mutation strategy





Design of the mutation strategy

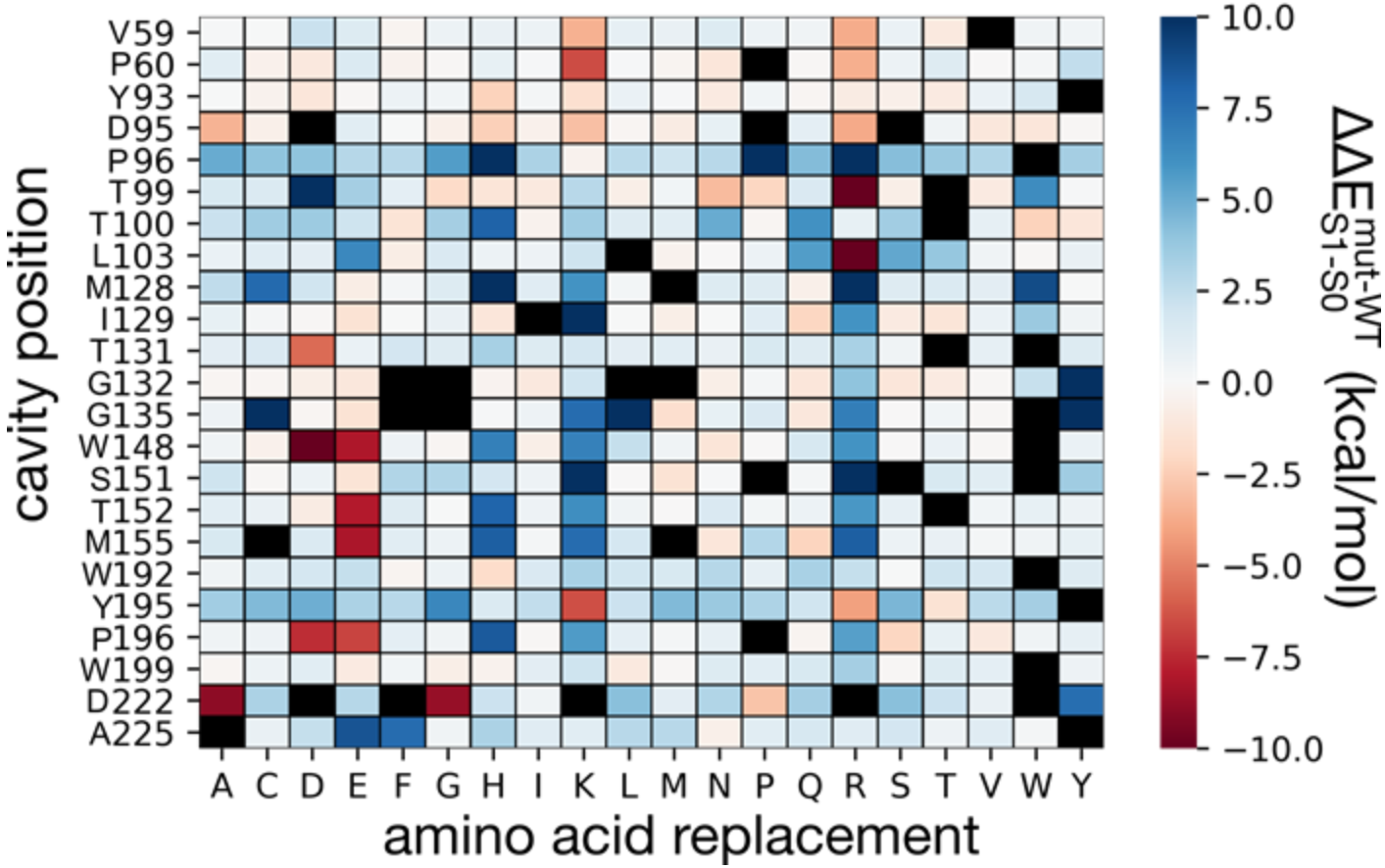


Positions V59, P60, D95, T99, P196, D222, A225 were found to be important in mutagenesis experiments

Arnold, F. H. (CalTech) McIsaac, R. S. et al. Recent advances in engineering microbial rhodopsins for optogenetics. *Curr. Opin. Struct. Biol.* **2015**, *33*, 8-15.

Cohen, A. E. (Harward) Hochbaum, D. R. et al. All-optical electrophysiology in mammalian neurons using engineered microbial rhodopsins. *Nat. Methods* **2014**, *11*, 825-825.

Preliminary results - Single Mutants



Biggest shifts are displayed for substitutions to charged aa

Sequences representation

Protein aa sequence → **Binary representation** → **Mathematical relationship between sequence and vertical excitation energy**

Alanine (A)	10000000000000000000
Cysteine (C)	01000000000000000000
Aspartate (D)	00100000000000000000
Glutamate (E)	00010000000000000000
Phenylalanine (F)	00001000000000000000
Glycine (G)	00000100000000000000
Histidine (H)	00000010000000000000
Isoleucine (I)	00000001000000000000
Lysine (K)	00000000100000000000
Leucine (L)	00000000010000000000
Methionine (M)	00000000001000000000
Asparagines (N)	00000000000100000000
Proline (P)	00000000000010000000
Glutamine (Q)	00000000000001000000
Arginine (R)	00000000000000100000
Serine (S)	00000000000000010000
Threonine (T)	00000000000000001000
Valine (V)	00000000000000000100
Tryptophan (W)	00000000000000000010
Tyrosine (Y)	00000000000000000001

$$f(\boldsymbol{x}) = \beta_0 + \beta_{1,1}\, x_{1,1} + \beta_{1,2}\, x_{1,2} + \ldots + \beta_{1,N}\, x_{1,N} \quad \text{Amino-acid A}$$
$$+ \beta_{2,1}\, x_{2,1} + \beta_{2,2}\, x_{2,2} + \ldots + \beta_{2,N}\, x_{2,N} \quad \text{Amino-acid C}$$
$$\vdots \qquad\qquad \vdots \qquad\qquad \vdots$$
$$+ \beta_{M,1}\, x_{M,1} + \beta_{M,2}\, x_{M,2} + \ldots + \beta_{M,N}\, x_{M,N} \quad \text{Amino-acid Y}$$
$$\text{Residue } 1 \qquad \text{Residue } 2 \qquad \text{Residue } N$$

Sequences representation

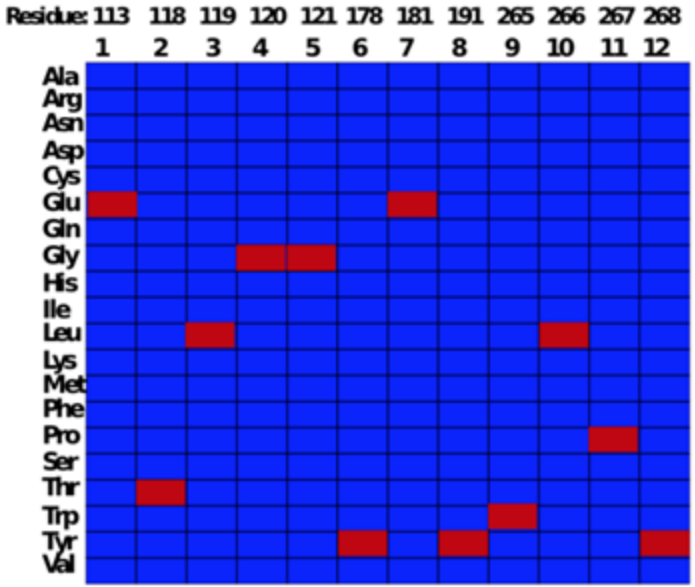
M = 20 (binary representation)
N = sequence length
K = number of variants

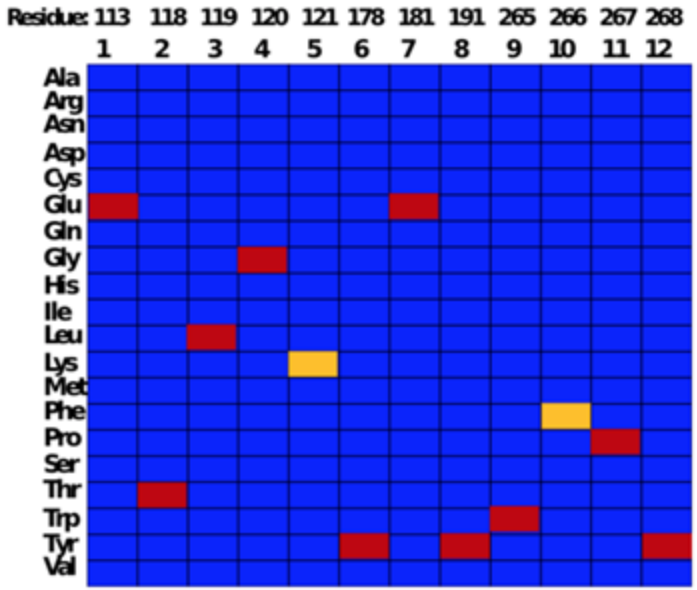
Alanine (A)	10000000000000000000
Cysteine (C)	01000000000000000000
Aspartate (D)	00100000000000000000
Glutamate (E)	00010000000000000000
Phenylalanine (F)	00001000000000000000
Glycine (G)	00000100000000000000
Histidine (H)	00000010000000000000
Isoleucine (I)	00000001000000000000
Lysine (K)	00000000100000000000
Leucine (L)	00000000010000000000
Methionine (M)	00000000001000000000
Asparagines (N)	00000000000100000000
Proline (P)	00000000000010000000
Glutamine (Q)	00000000000001000000
Arginine (R)	00000000000000100000
Serine (S)	00000000000000010000
Threonine (T)	00000000000000001000
Valine (V)	00000000000000000100
Tryptophan (W)	00000000000000000010
Tyrosine (Y)	00000000000000000001

To represent K **rhodopsins** I need a matrix with dimensions:

K, MxN

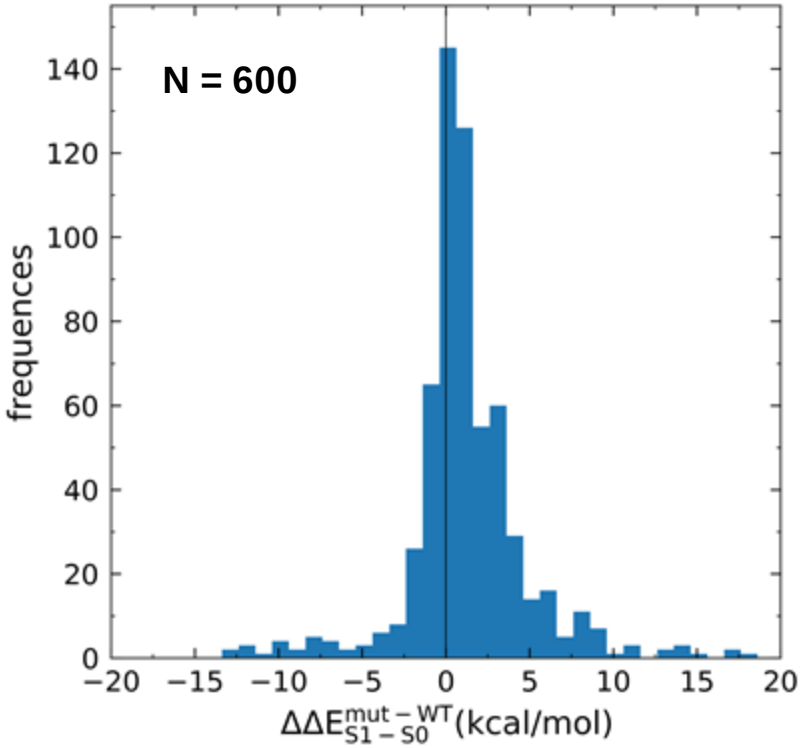
Sequences representation



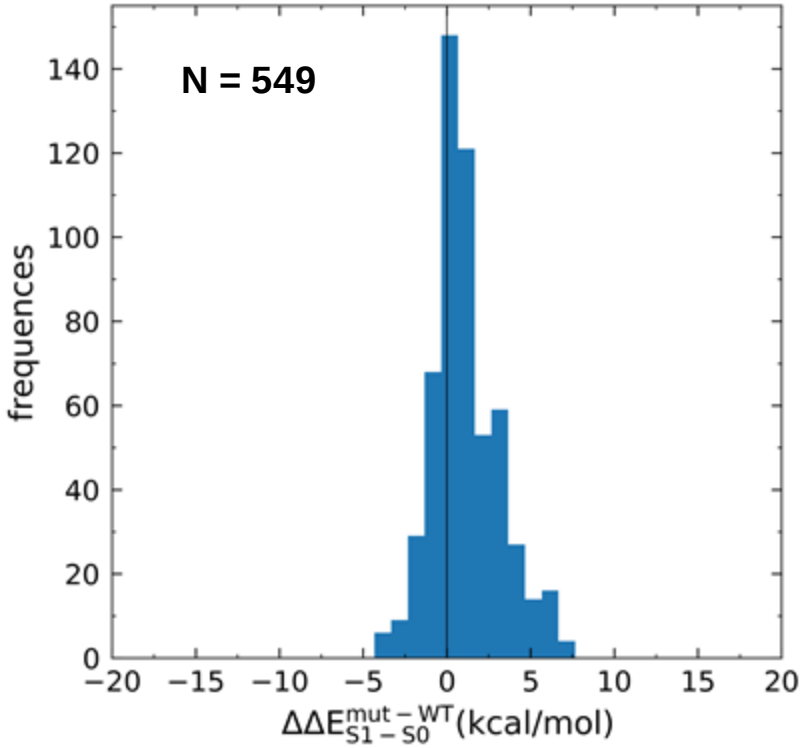


Preprocessing of input data - towards ML

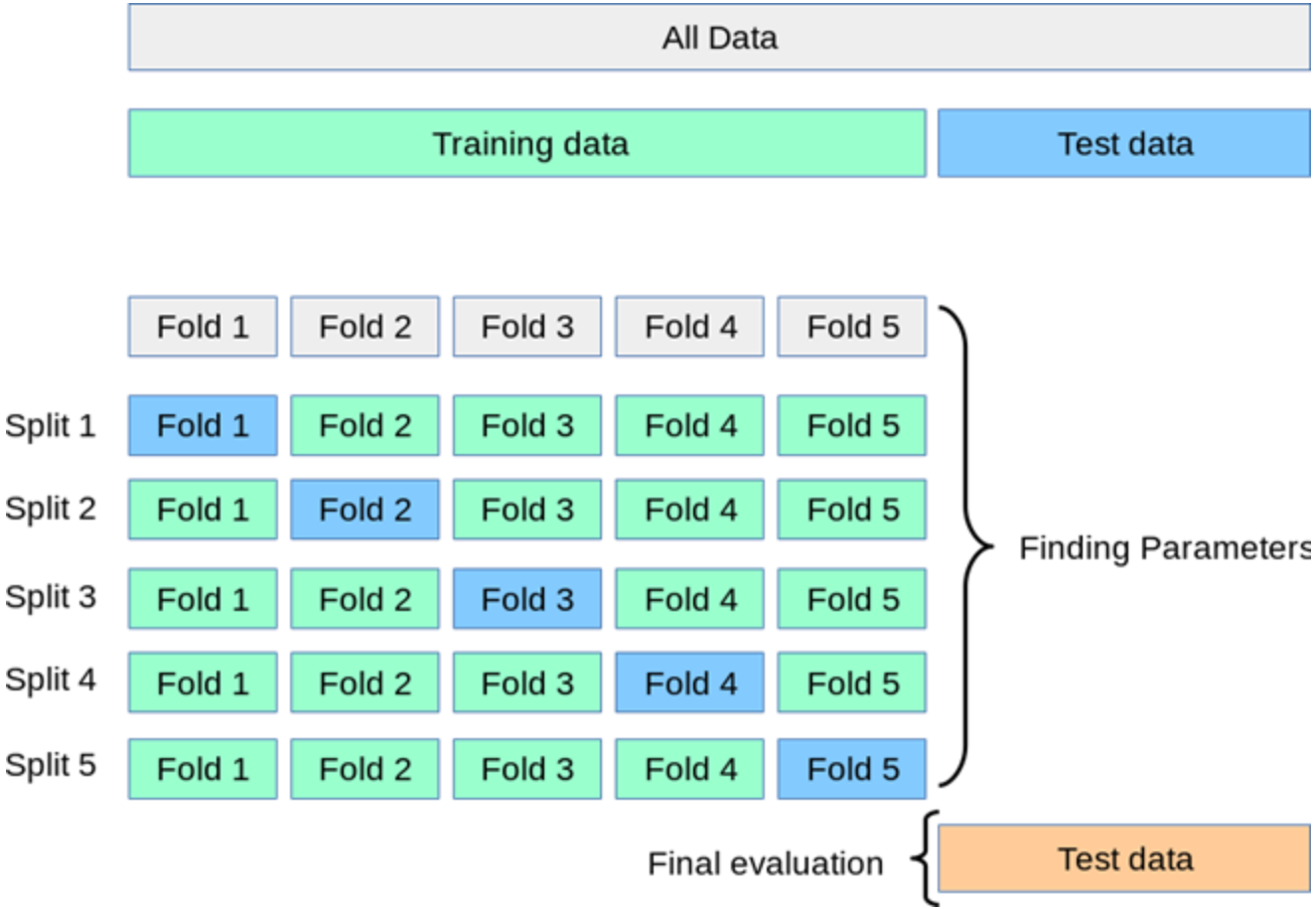
Are outliers actually outliers?



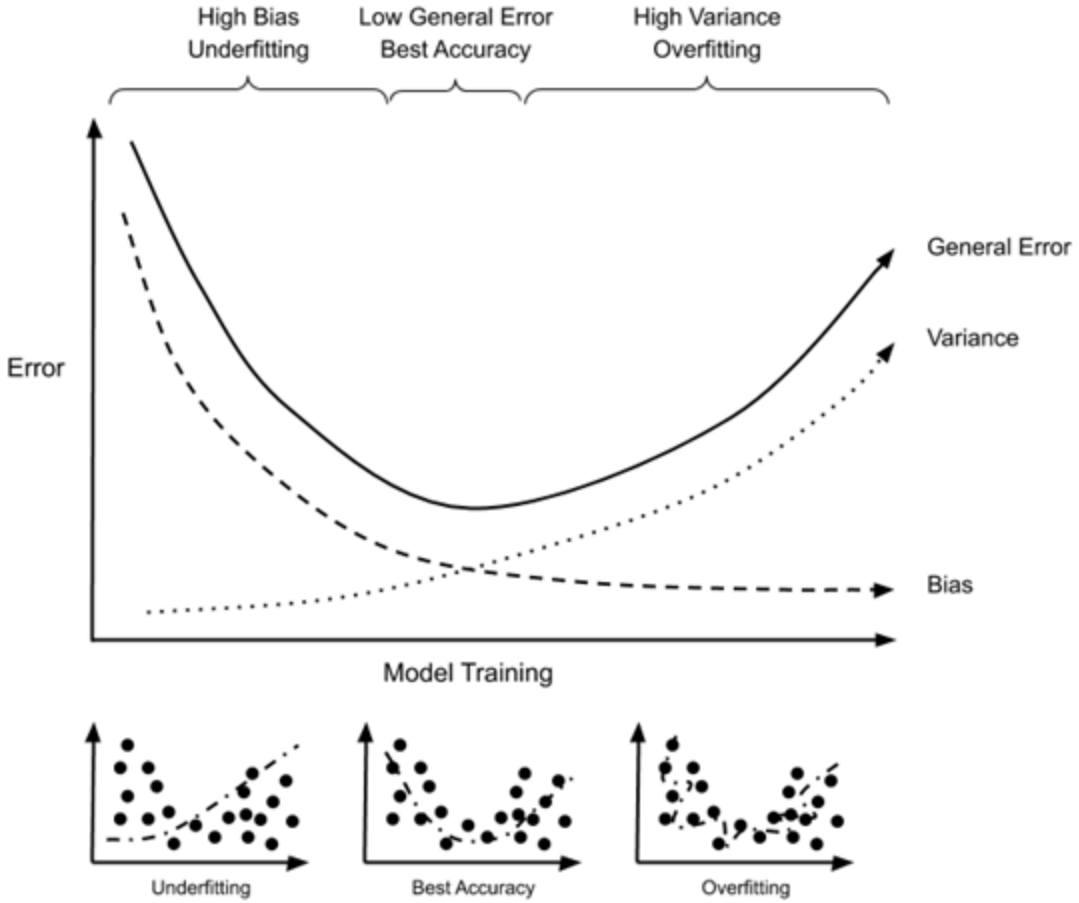
IQR range rule



Preprocessing of input data - towards ML



Preprocessing of input data - towards ML



ML - A linear approach

M = 20 (binary representation of a single AA)
N = sequence length
K = number of variants

$$f(\mathbf{x}) = \beta_0 + \sum_{i=1}^{M}\sum_{j=1}^{N}\beta_{i,j}x_{i,j}$$

$$min_{\beta,\beta_0}\sum_{k=1}^{K}(\Delta E_{S1-S0}^{(k)} - \beta_0 - \sum_{i=1}^{M}\sum_{j=1}^{N}\beta_{i,j}x_{i,j}^{(k)})^2$$

or , in matrix notation:

$\mathbf{y} = \beta_0 + \mathbf{X\beta}$ ⟶ $\|\mathbf{e}\|^2 = \|\,\mathbf{y} - \mathbf{X\beta}\,\|^2$ ⟶ $\min \|\,\mathbf{y} - \mathbf{X\beta}\,\|^2$ ⟶ $\boldsymbol{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\,\mathbf{X}^T\,\mathbf{y}$

Linear Function

Define as least square problem

Solve analytically or via GD

Linear approaches - Ridge Regression

M = 20 (binary representation)
N = sequence length
K = number of variants

$$min_{\beta,\beta_0} \sum_{k=1}^{K} (\Delta E_{S1-S0}^{(k)} - \beta_0 - \sum_{i=1}^{M}\sum_{j=1}^{N} \beta_{i,j} x_{i,j}^{(k)})^2 + \lambda \sum_{i=1}^{M}\sum_{j=1}^{N} \beta_{i,j}^2$$

or , in matrix notation:

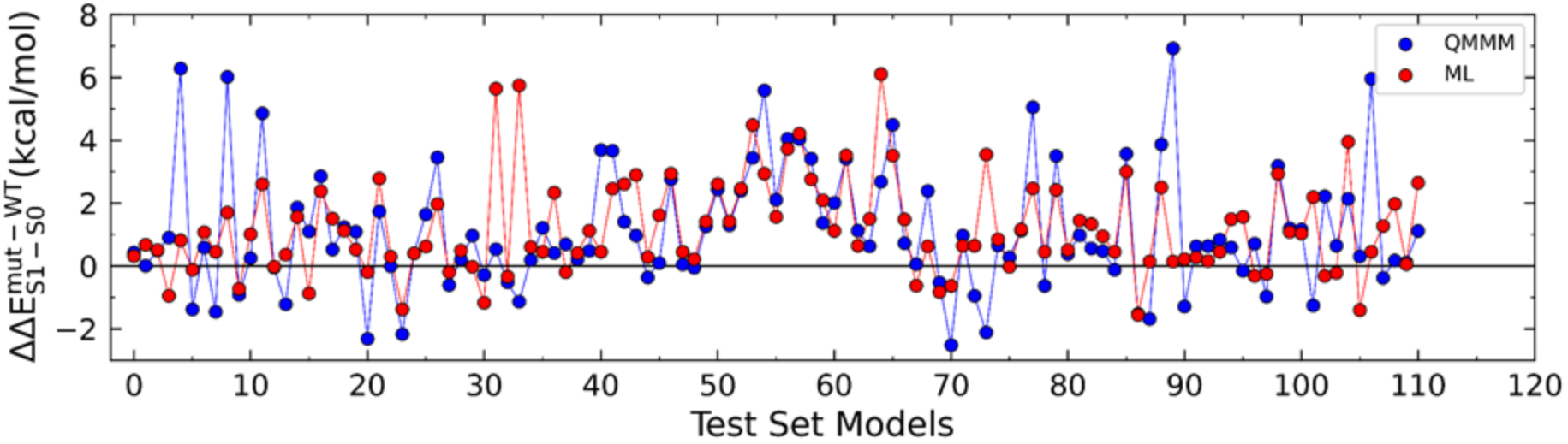
$$\beta = (X^T X + I\lambda)^{-1} X^T y$$

Linear approaches - Ridge Regression

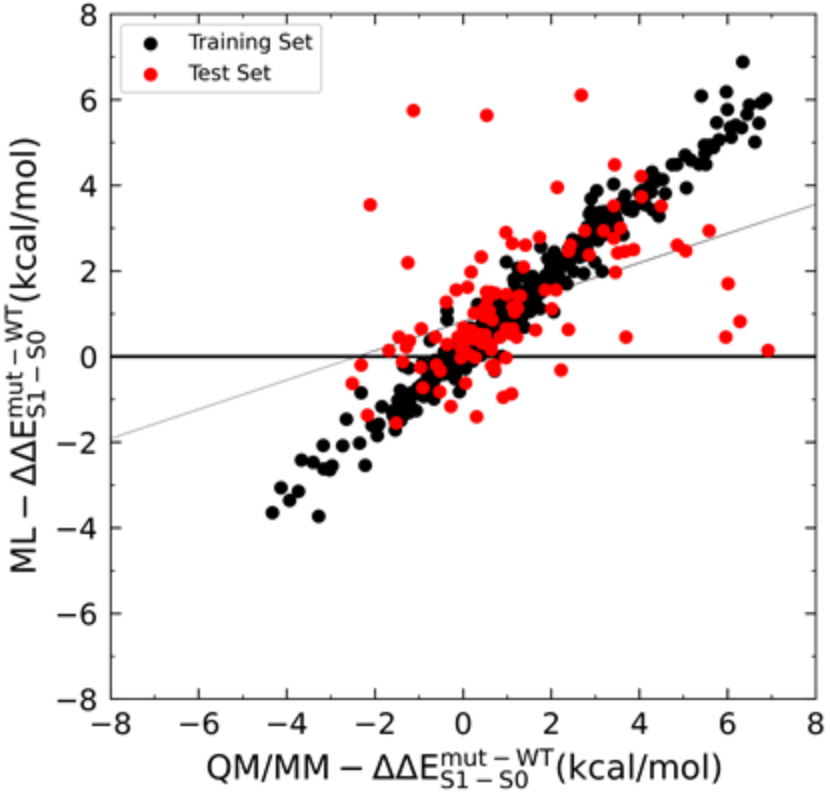
UNIVERSITÀ DI SIENA 1240

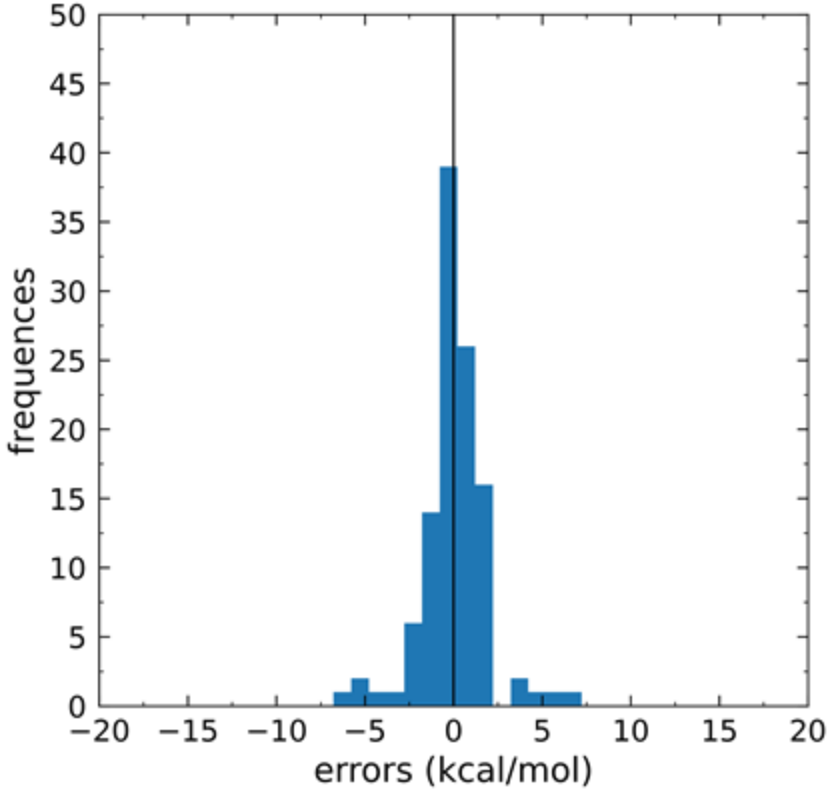
Ridge Regression

MAE -> 1.10 kcal/mol



Linear approaches - Ridge Regression





Linear approaches - Lasso Regression

M = 20 (binary representation)
N = sequence length
K = number of variants

$$min_{\beta,\beta_0} \sum_{k=1}^{K} (\Delta E_{S1-S0}^{(k)} - \beta_0 - \sum_{i=1}^{M} \sum_{j=1}^{N} \beta_{i,j} x_{i,j}^{(k)})^2 + \lambda \sum_{i=1}^{M} \sum_{j=1}^{N} |\beta_{i,j}|$$

Linear approaches - Lasso Regression

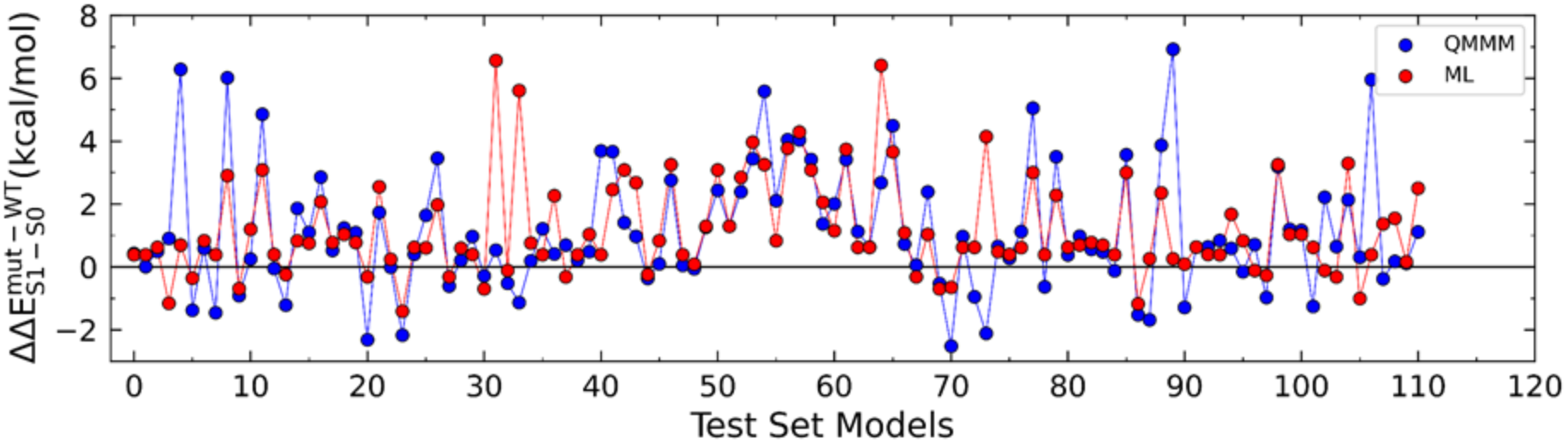
UNIVERSITÀ DI SIENA 1240

Lasso Regression

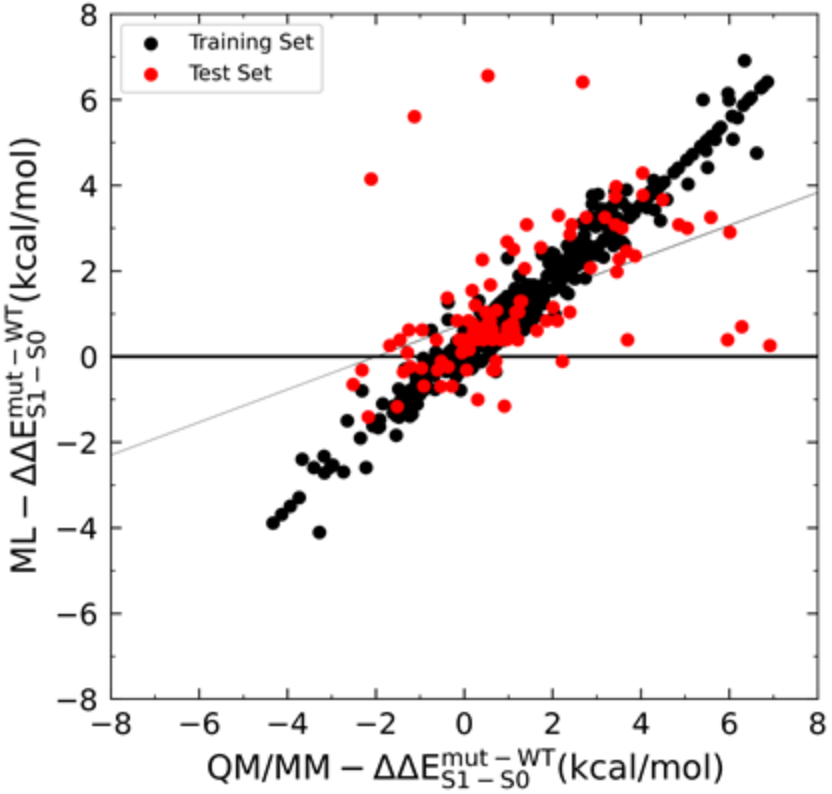
MAE -> 1.05 kcal/mol

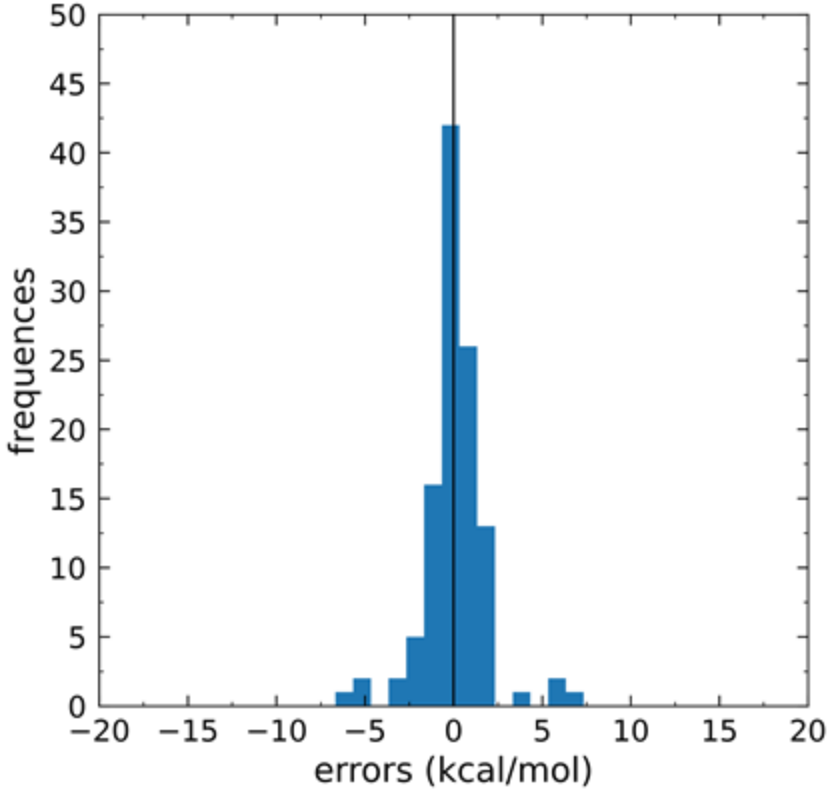
Ridge Regression

MAE -> 1.10 kcal/mol

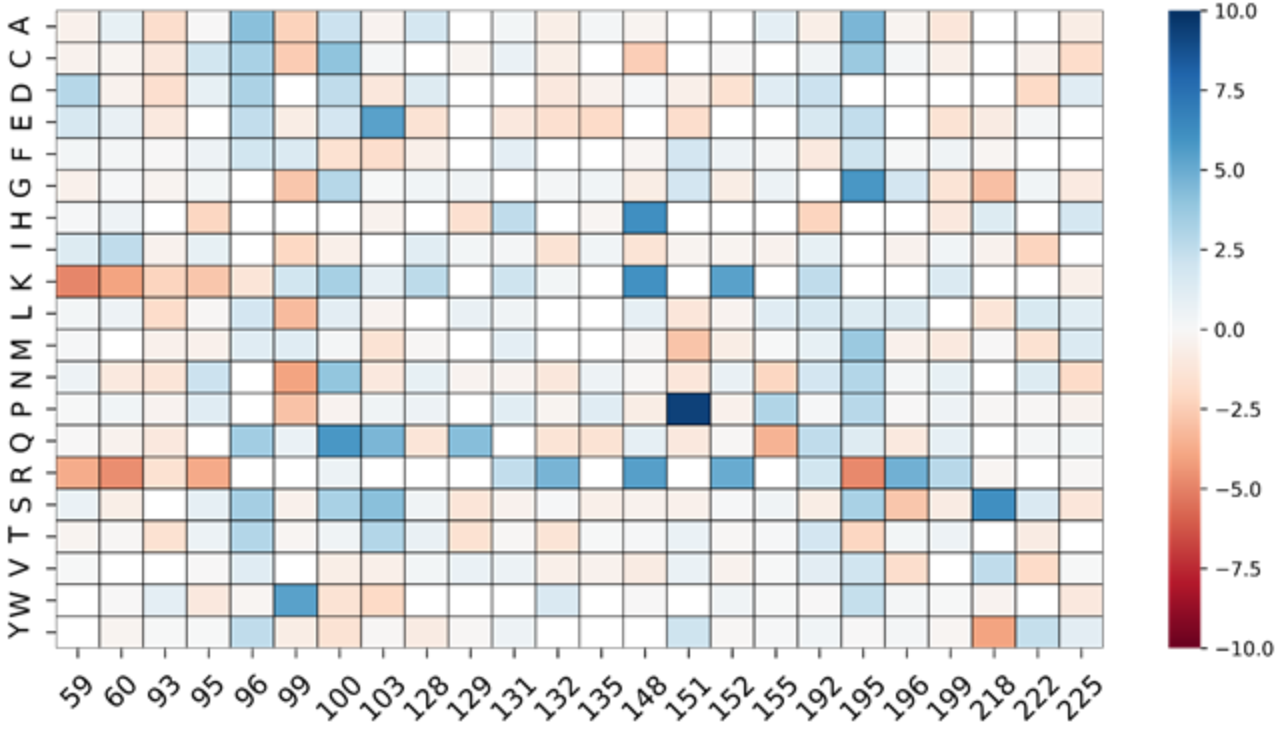


Linear approaches - Lasso Regression



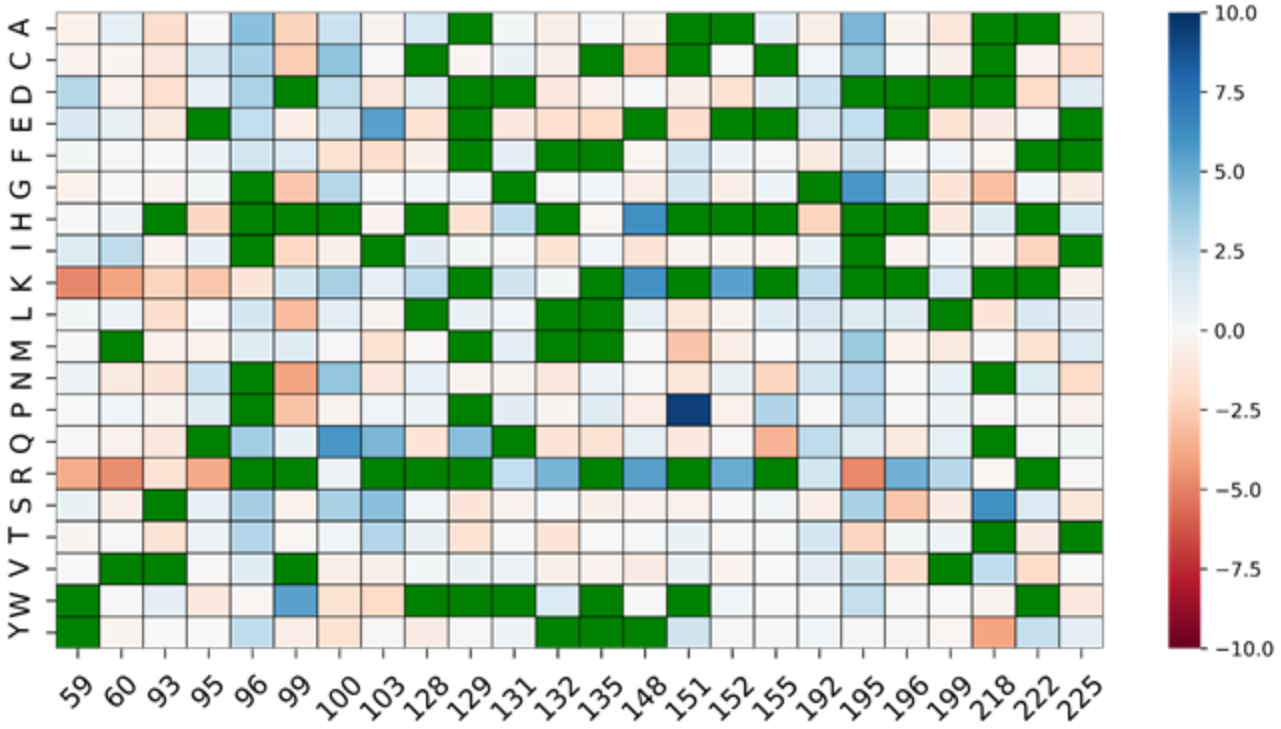


Lasso Regression - coefficients representation



$$\sum_{i=1}^{M}\sum_{j=1}^{N} x_{i,j}\beta_{i,j}$$

Lasso Regression - coefficients representation



$$\sum_{i=1}^{M}\sum_{j=1}^{N} x_{i,j}\beta_{i,j}$$

Non-Linear approaches - Random Forest

UNIVERSITÀ DI SIENA 1240

RF Regression

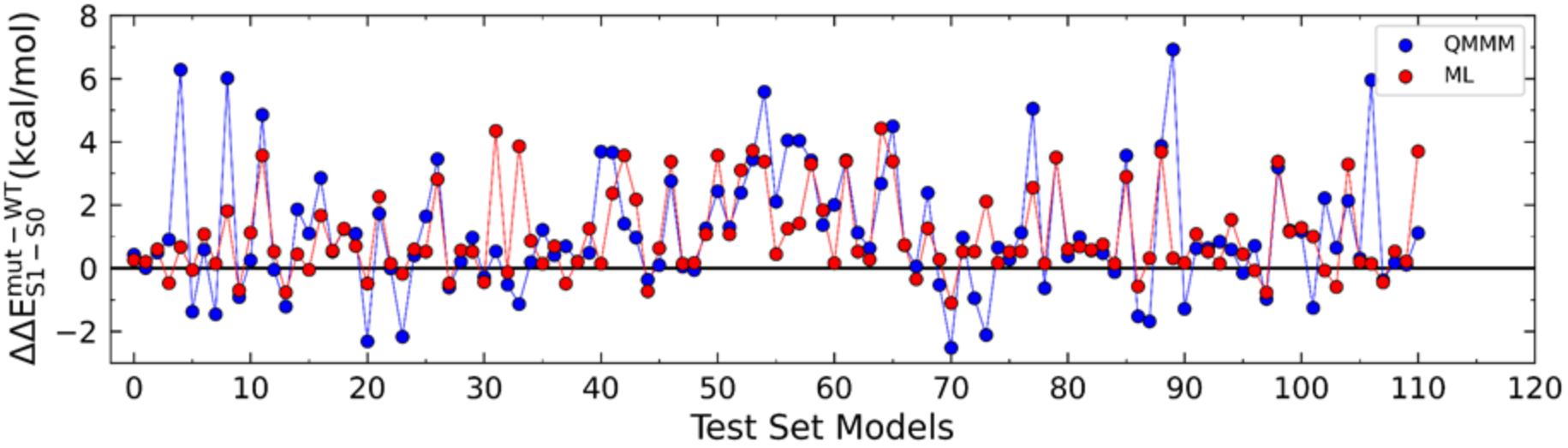
MAE -> 1.00 kcal/mol

Lasso Regression

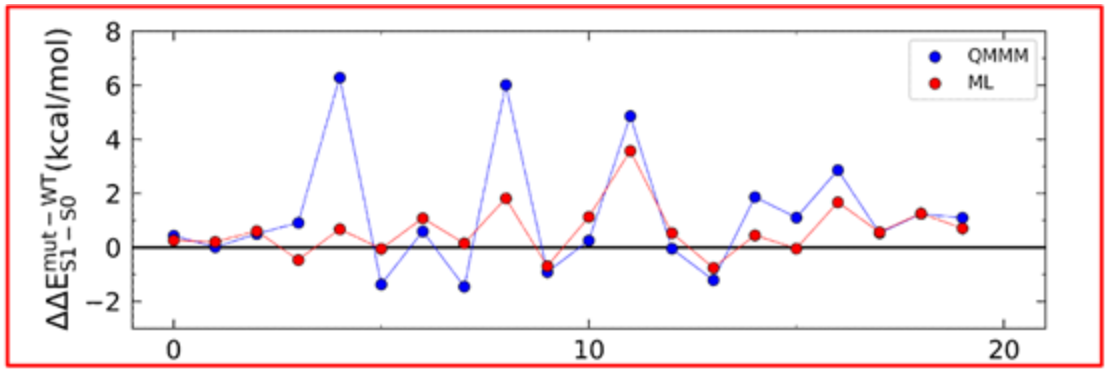
MAE -> 1.05 kcal/mol

Ridge Regression

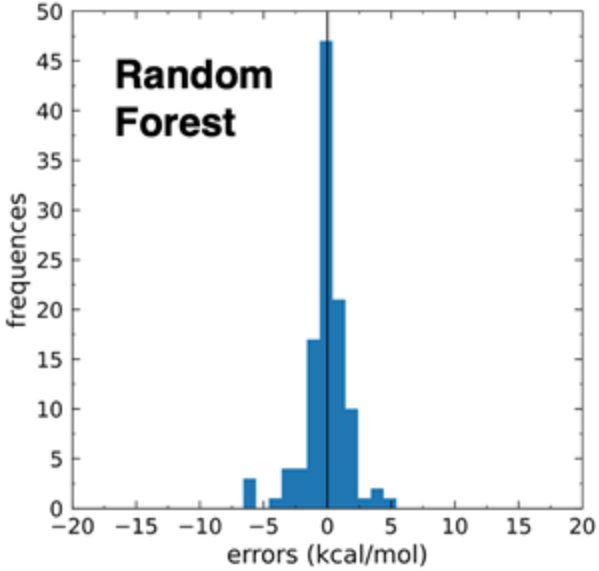
MAE -> 1.10 kcal/mol

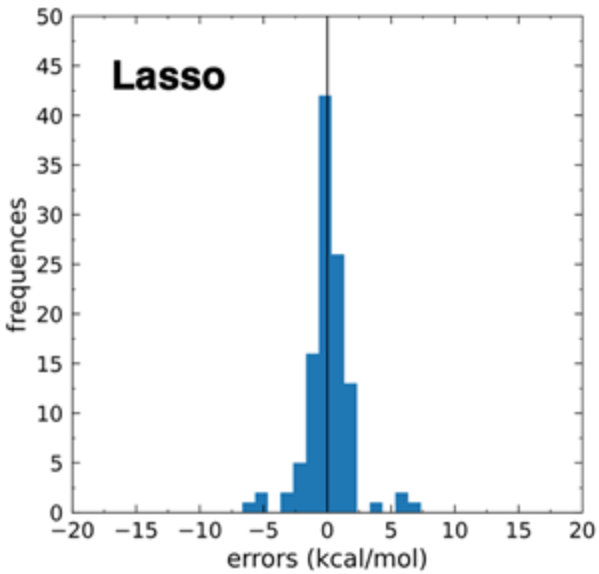


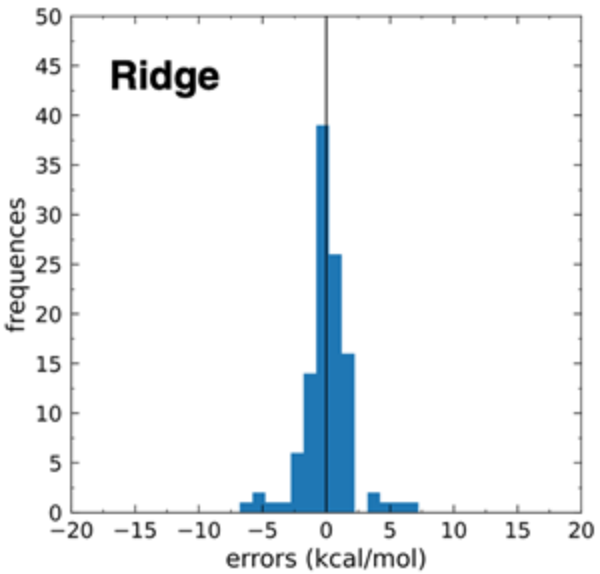
Non-Linear approaches - Random Forest



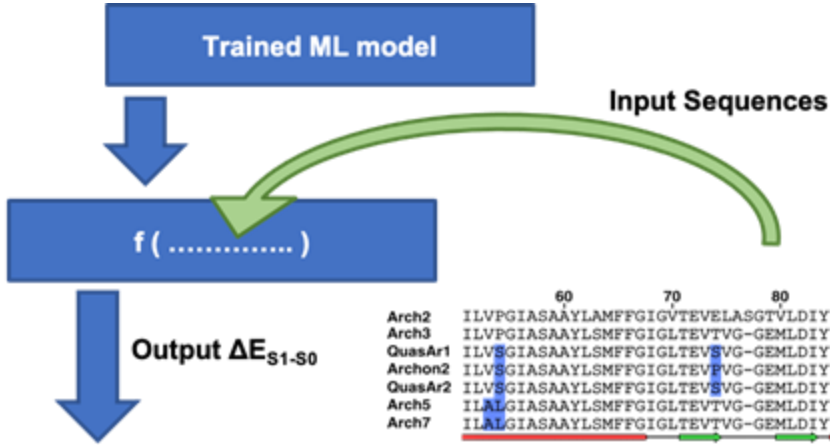
Non-Linear approaches - Random Forest

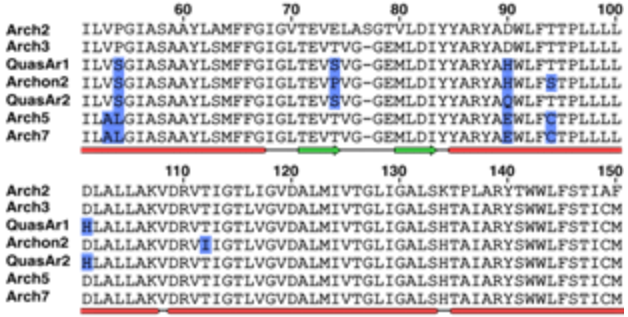






Meaningfulness of the models



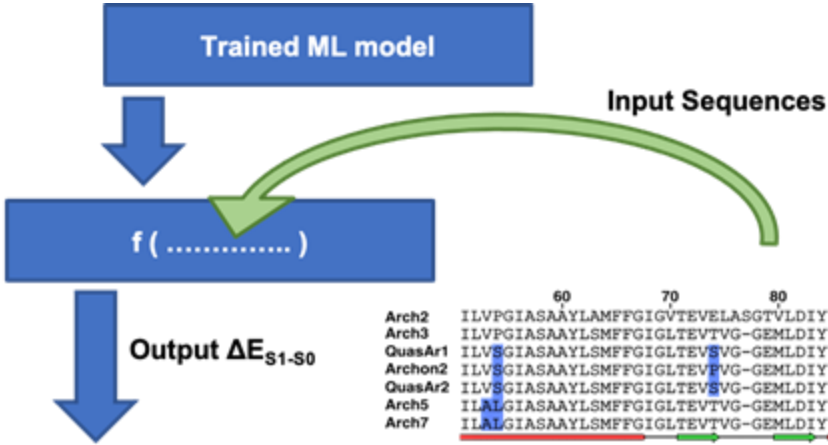


Protein	Exp ΔE_{S1-S0} (kcal/mol)	Exp shift (kcal/mol)	ML shift (kcal/mol)
Arch3	51.4	0	0
D95E/T99C/V59A	46.0	**-5.4**	**-2.9**
D95E/T99C/P60L	45.8	**-5.6**	**-2.7**
D95E/T99C/P196S	45.5	**-5.9**	**-4.3**
Arch5	46.0	**-5.4**	**-5.1**
Arch7	46.4	**-5.0**	**-3.1**

Meaningfulness of the models

ML predicts shifts in the same direction as Exp.

Good QM/MM models?



Protein	Exp ΔE_{S1-S0} (kcal/mol)	Exp shift (kcal/mol)	ML shift (kcal/mol)
Arch3	51.4	0	0
D95E/T99C/V59A	46.0	**-5.4**	**-2.9**
D95E/T99C/P60L	45.8	**-5.6**	**-2.7**
D95E/T99C/P196S	45.5	**-5.9**	**-4.3**
Arch5	46.0	**-5.4**	**-5.1**
Arch7	46.4	**-5.0**	**-3.1**

To Do

- Statistical Analysis of the excitation energies
- Increase the training set by generating more double mutants
- Define the problem as a classification task
- Use non linear-models including NN
- Outliers (?)