

MAPK1 and MAPK3 CAGI challenges with MuSiC

Dr. Fabrizio Pucci

Computational Biology and Bioinformatics,
École polytechnique de Bruxelles, ULB
Interuniversity Institute of Bioinformatics in Brussels (IB²)

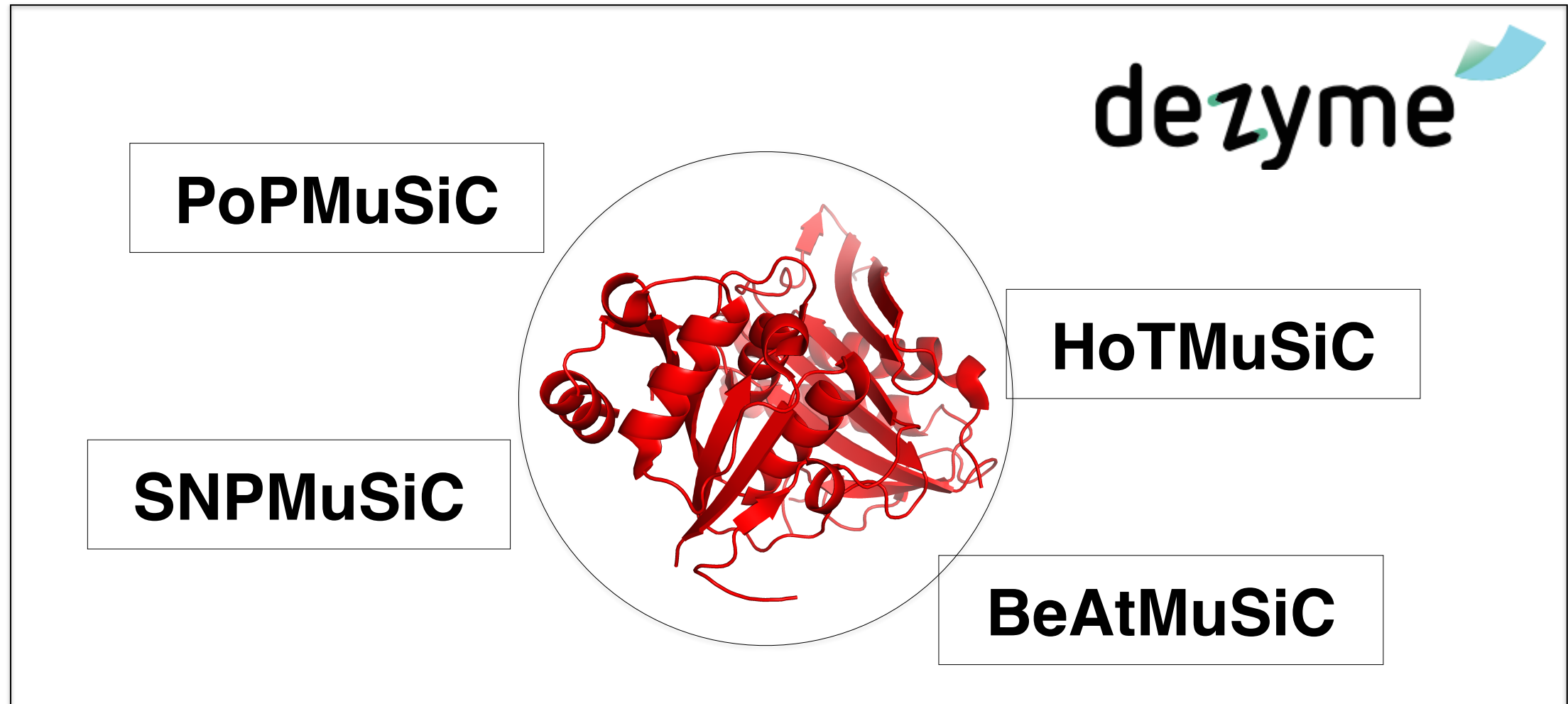
with G. Cia, M. Schwersensky, M. Rooman



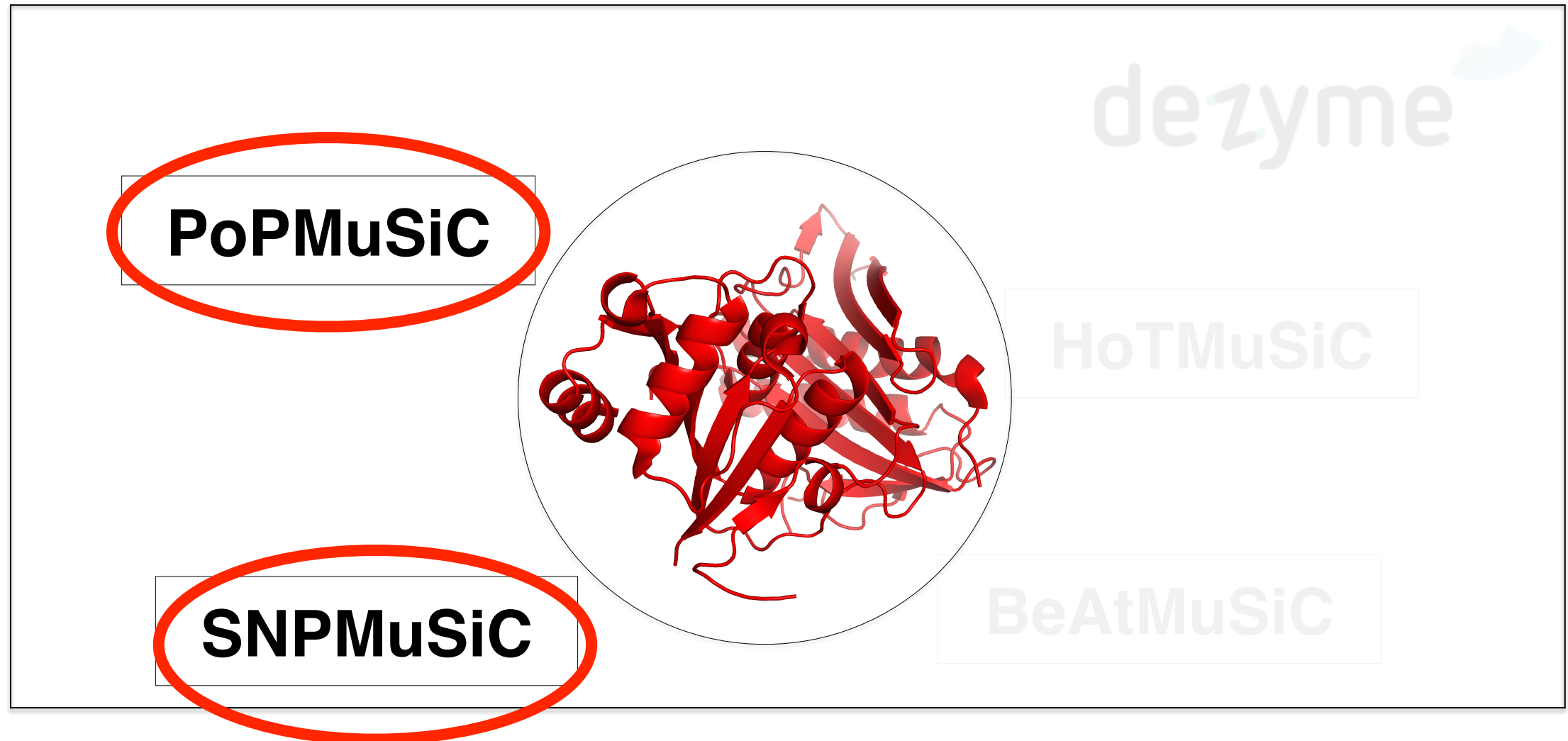
COMPUTATIONAL BIOLOGY
AND BIOINFORMATICS LAB



MuSiC suite www.dezyme.org



MuSiC suite www.dezyme.org



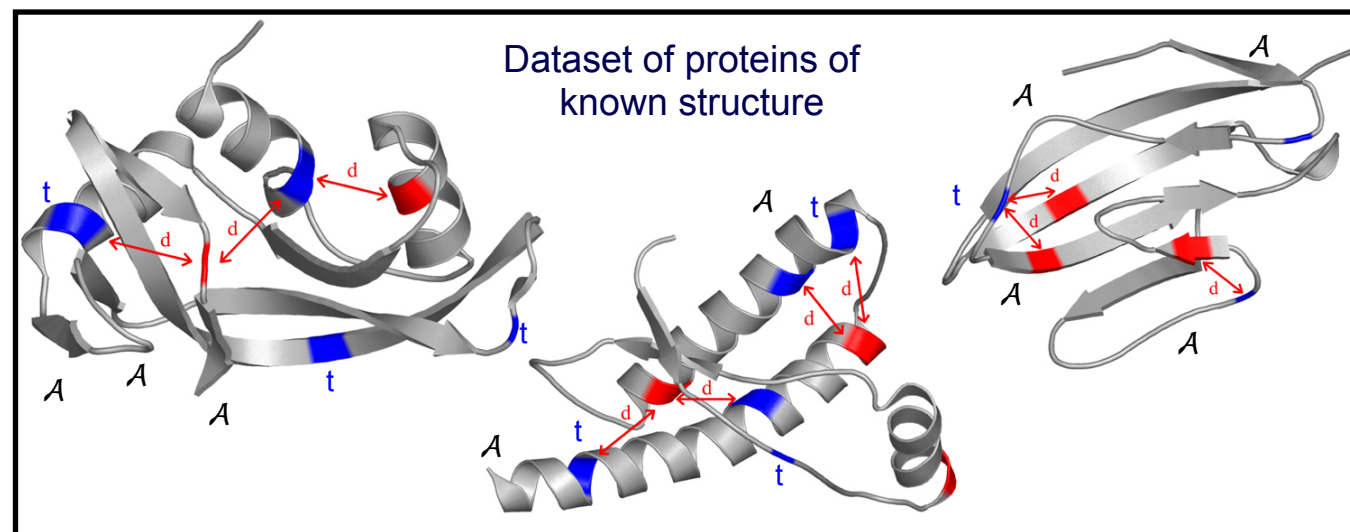
► Protein 3D structure as input

Based on a **simplified representation** of protein structure!

- Only the main chain atoms + average side chain center $\mathbf{C}_\mu$
- Neglecting the side chain degrees of freedom

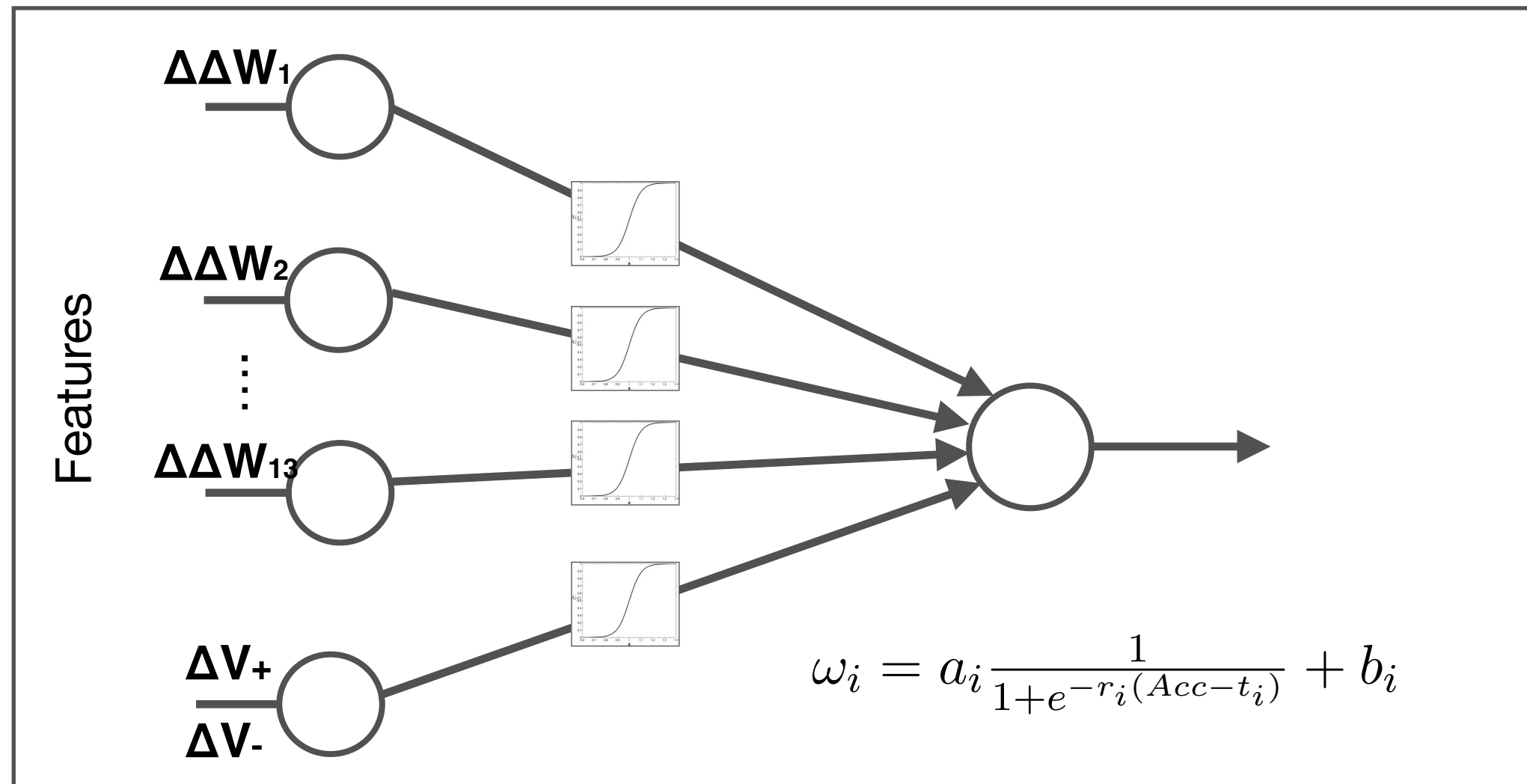
► Use of mean force potentials as features

Mean-force knowledge-based potentials derived from a non-redundant set of well resolved protein structures.



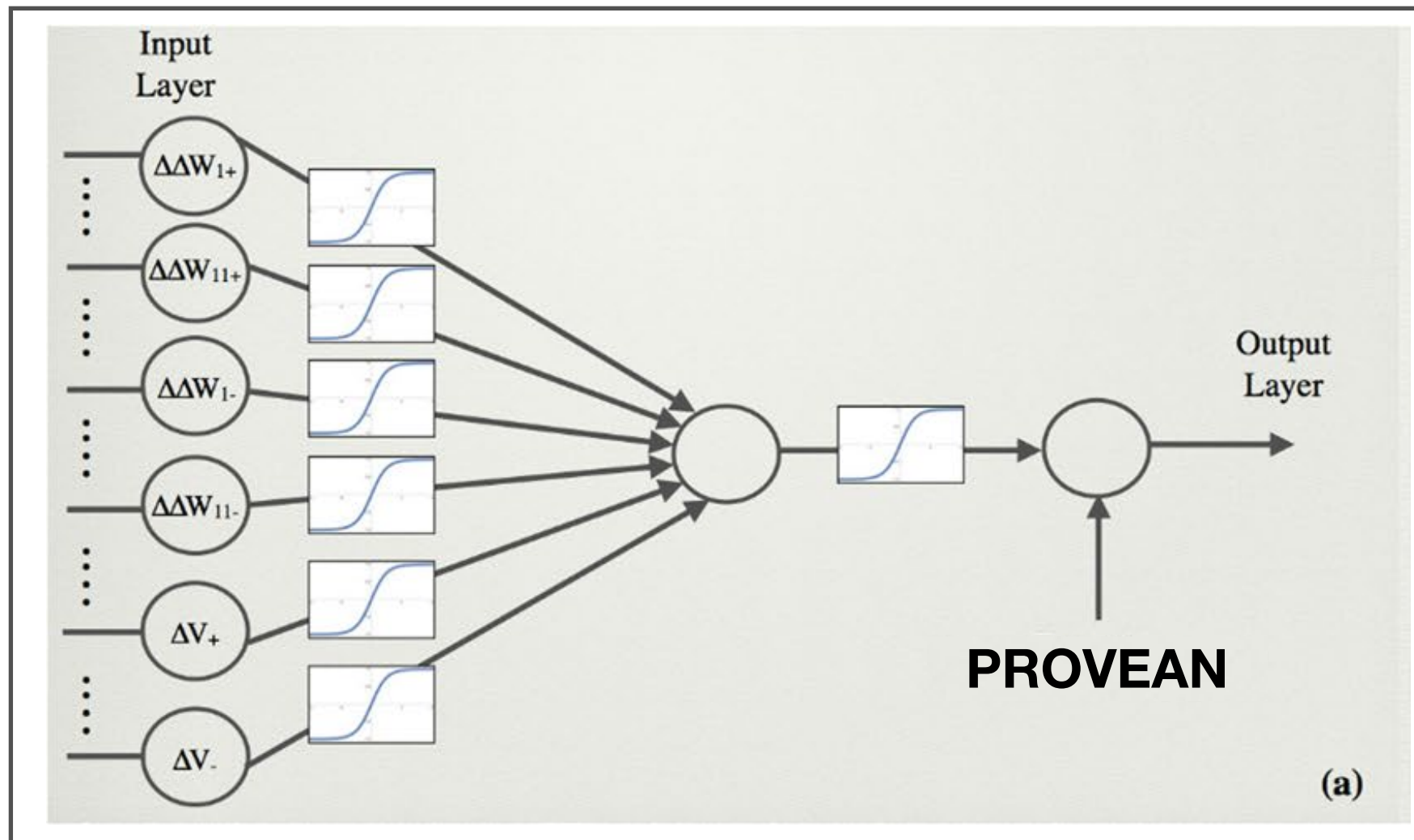
$$\Delta W(c, s) \simeq -kT \ln \frac{P(c|s)}{P(c|s)_{\text{ref}}} \simeq -kT \ln \frac{P(c, s)}{P(c)P(s)}$$

► Use of artificial neural networks : PoPMuSiC



Parameters are optimized on the S2648 dataset of mutations with known $\Delta\Delta G$ values.

► Use of artificial neural networks : SNPMuSiC



The model structure is similar but

- 1) Absolute value of statistical potentials are considered
- 2) The PROVEAN score is added to the prediction of the ANN.

Parameters are identified in order to optimize the BACC on a subset of ClinVar entries with known deleteriousness/neutral information.

► Use of artificial neural networks : SNPMSiC

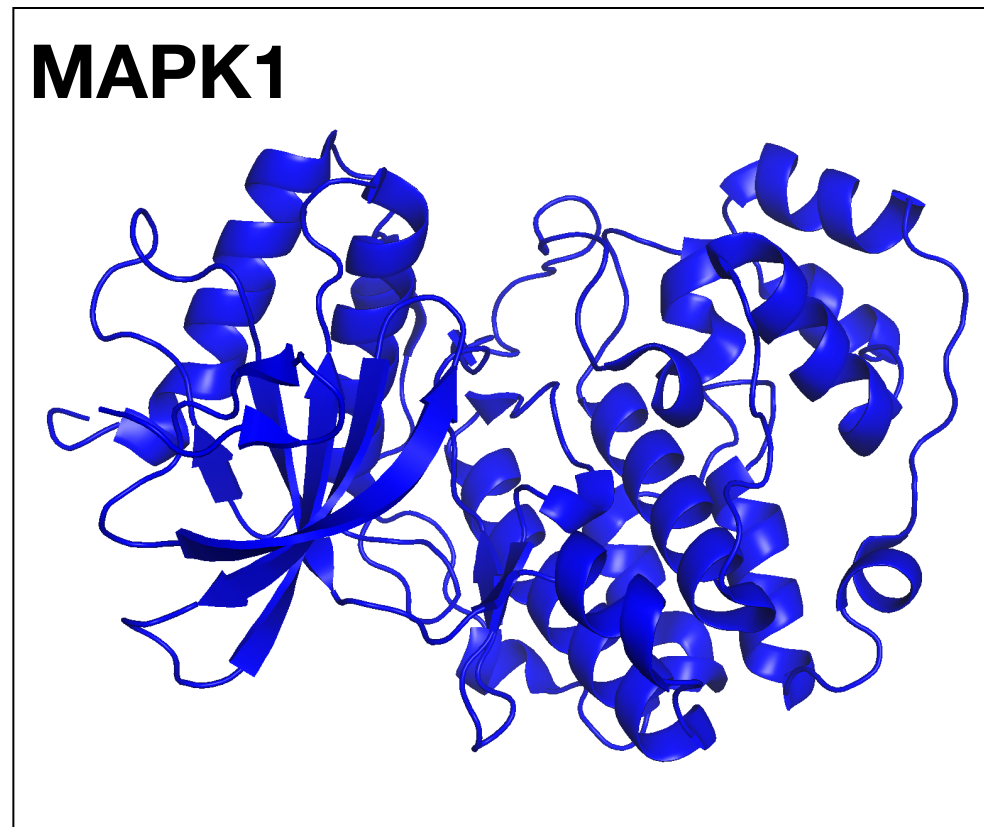


The model structure is similar but

- 1) Absolute value of statistical potentials are considered
- 2) The PROVEAN score is added to the prediction of the ANN.

Parameters are identified in order to optimize the BACC on a subset of ClinVar entries with known deleteriousness/neutral information.

MuSiC suite at MAPK1



Unphosphorylated form

PDB code = 6oph and 4zxt

Structures bound to different inhibitor
(GDC-0994, catechol)

Results averaged over all the two PDBs

Phosphorylated form

PDB code = 5v60

Structures bound to inhibitor
(AMP-CMP)

RMSD between the two structures 0.3-0.9 Å

MuSiC suite at MAPK1 - MAPK3

How we take into account **phosphorylation** ?

Since **MuSiC** cannot directly take into account the phosphorylation, we modified the phosphorylated residues as

Thr-185 and Tyr-187  Asp-185 and Glu-187

and remodel the 3D structure using homology modeling.

In this way we mimic the effect of the negative charge of the phosphoryl group. These substitutions are common strategies to mimic phosphorylation, which are also used by nature

(Pearlman et al, Cell 147(4), 934–946 (2011)).

Prediction Models for MAPK1-3 change in stability

1) $\Delta\Delta G$ prediction via PoPMuSiC

(Dehouck et al, BMC Bioinformatics 12, 151, 2011).

2) $\Delta\Delta G$ prediction via PoPMuSiCsym

(Pucci et al., Bioinformatics, 34, 3659–3665 (2018))

PoPMuSiC version corrected for inverse-direct symmetry

$$\Delta\Delta G (A \rightarrow B) = - \Delta\Delta G (B \rightarrow A)$$

3) $\Delta\Delta G$ prediction PoPMuSiC + MAESTRO

(Pucci et al., Bioinformatics, 34, 3659–3665 (2018))

(J. Laimer et al., BMC Bioinformatics 16, 116 (2015))

Prediction Models for MAPK1-3 change in activity

Here we did not use stability prediction since we know that stability is only mildly related to activity. Example : catalytic or allosteric sites that are weaknesses of the structures but fundamental for protein activity.

1) Prediction via SNPMuSiC

(Ancien et al, Scientific 8, 4480 (2018))

Following the resealing $r_i = -\text{SNPMUSIC} + 0.5$

2) Prediction via SNPMuSiC

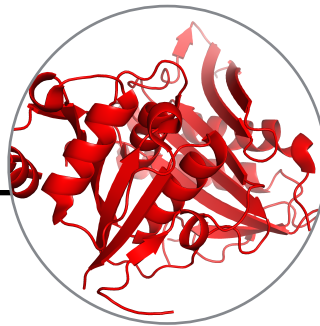
(Ancien et al, Scientific 8, 4480 (2018))

Following the resealing $r_i = -0.75 \text{ SNPMUSIC} + 0.5$

3) Preliminary version of protein fitness predictor FitMuSiC

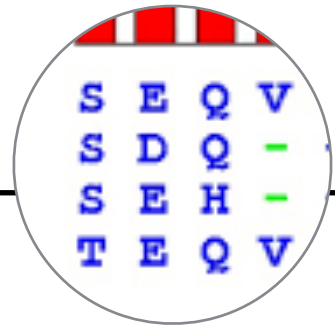
To describe activity one has to consider structural and evolutionary constraints.

Structural Features



- $\Delta\Delta G_i$ computed by PoPMuSiC
- Solvent accessibility of the mutated residues
- $\Delta\Delta G_i$: computed by Maestro
- Deleteriousness score computed by SNPMuSiC

Evolutionary Features



- EVCoupling score
- PROVEAN score
- Conservation Index from DEOGEN2
- Log-Odd ratio score from DEOGEN2

- The features are combined using a **simple linear regression** model
- Coefficients identified by minimizing the mean squared deviation between the predicted scores and the experimental fitness values (UBE2I, SUMO1, CALM1, TPK1 from Weile et al, Mol Syst Biol. 2017 13(12):957).

Thanks Tina, Pedja and Steven for CAGI organization
Thanks Emidio for assessing the challenge
Thanks Valerio and Roberta for providing data

Thank you for the attention
Questions?

