

Predicting the effect of mutation on enzymatic kinetic parameters

Yvan Rousset, Alexander Kroll, Martin Lercher

Yvan.rousset@hhu.de

Background

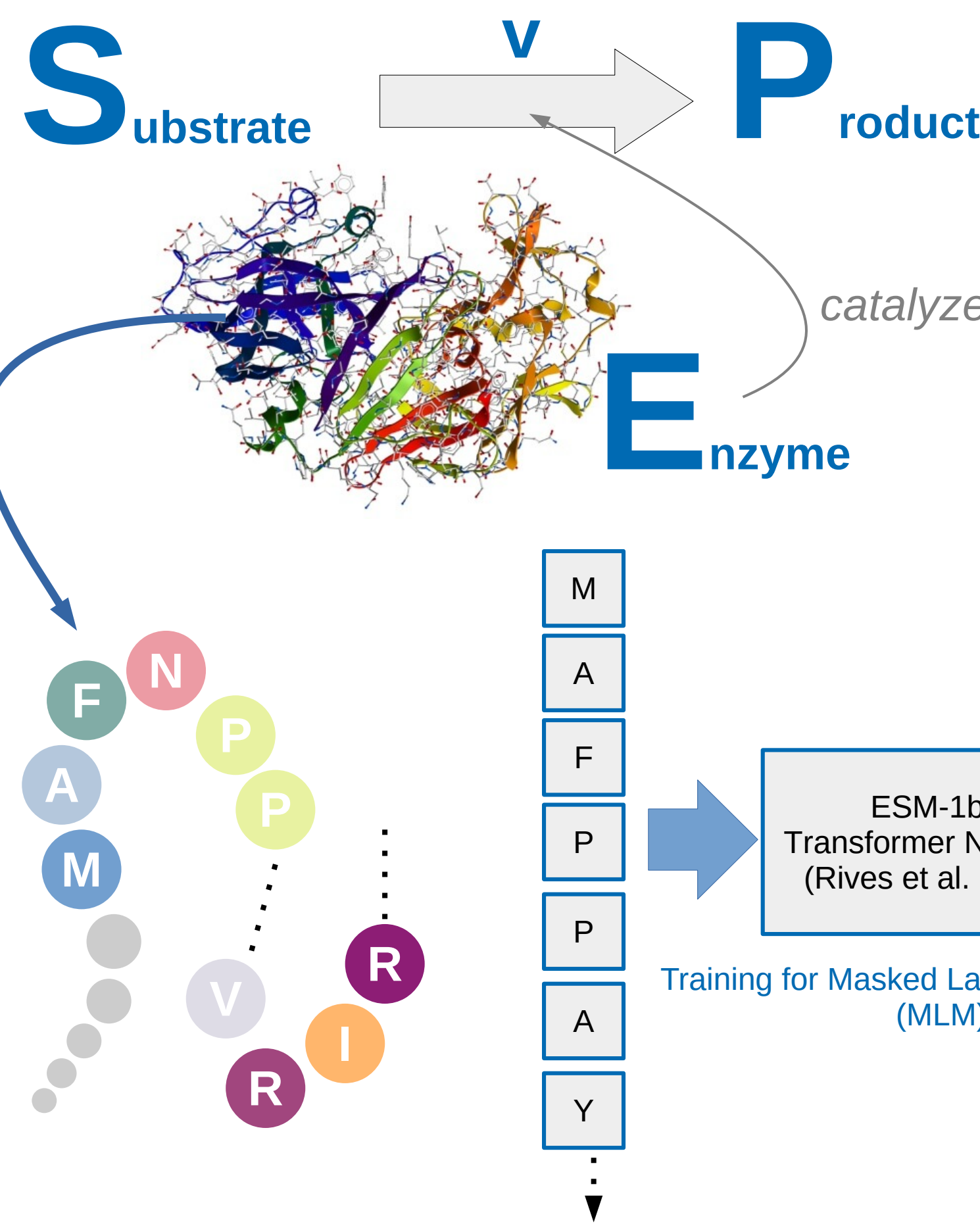
Accessing biochemical kinetic parameters is crucial in systems biology and bio-engineering. Recently, machine learning (ML) methods have shown promising results in predicting unknown parameters.

Protein language models are used to transform amino acid sequences into a more useful deep representation called ESM1b [1]. However, the ESM1b model is trained on wild-type (wt) sequences and don't generalize well to mutated sequences (mut), making predictions for mutants difficult.

Here we propose to predict the difference in parameters given a pair (wt-mut), by using a customized representation of this pair from ESM1b.

General workflow [2]:

for wild-type only



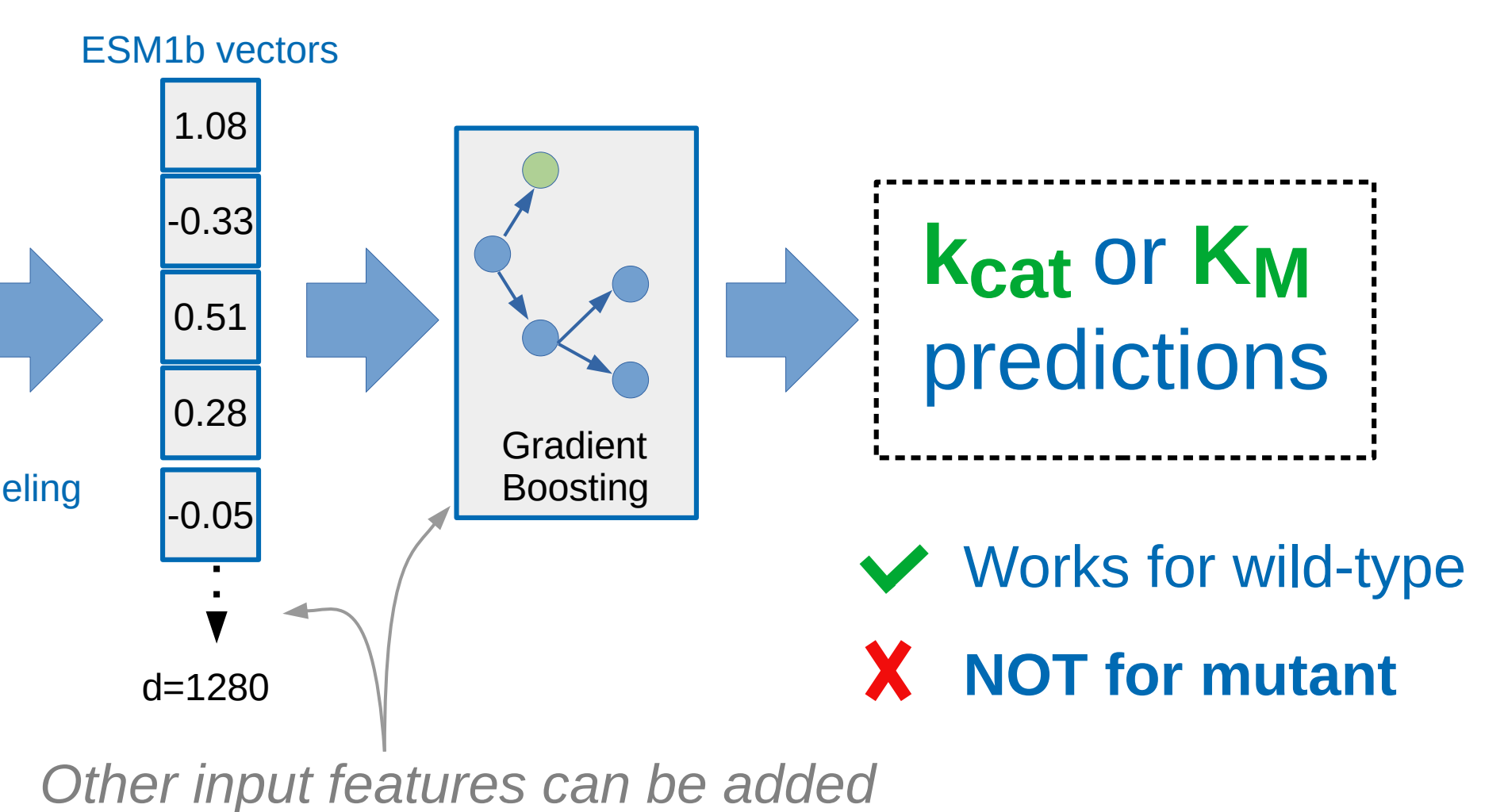
Reaction kinetic

(What we are interested in predicting...)

$$v = \frac{k_{cat} \cdot [E] \cdot [S]}{K_M + [S]}$$

$[s^{-1}]$
 $[M]$

"The higher k_{cat}/K_M is, the more efficient a reaction is."

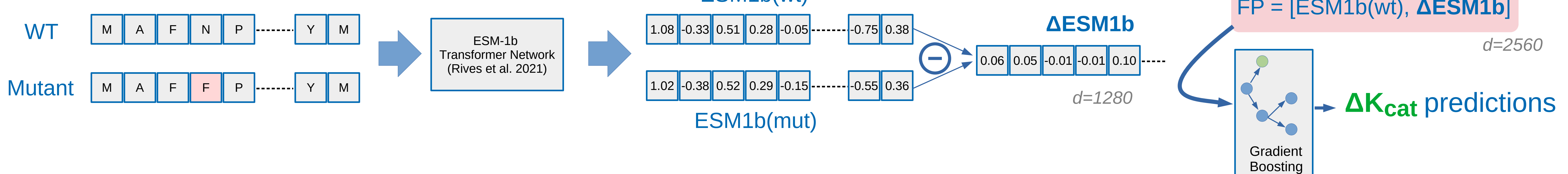


Predict the parameter variation Δp from a wild-type sequence to its mutant.

$$\Delta p = p(wt) - p(mut)$$

"Here we are interested in Δk_{cat} and ΔK_M "

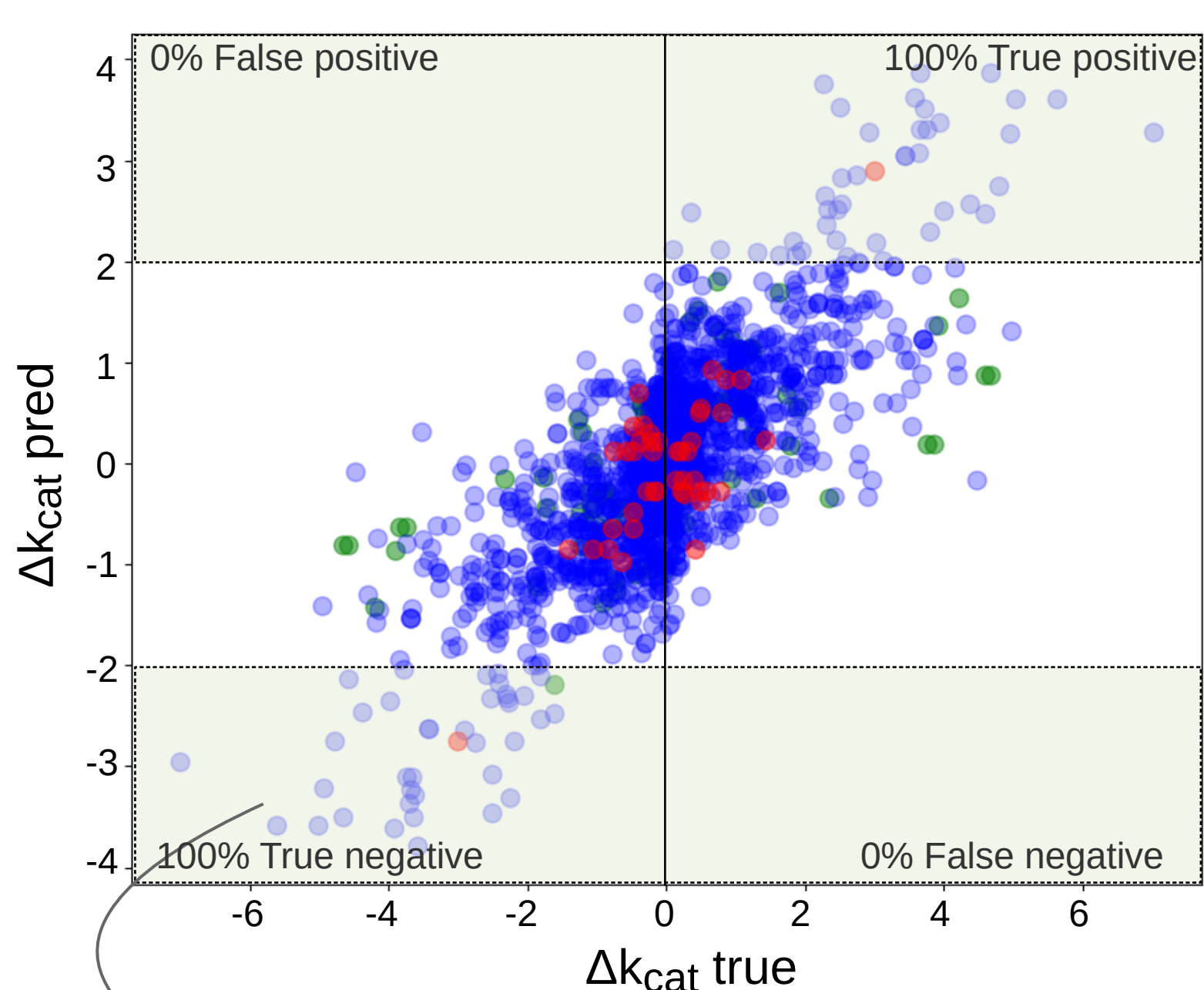
New workflow:



Methods

- 17,010 k_{cat} entries were extracted from the BRENDA database (7481 mutants, 9529 wt).
- The entries were grouped based on the same E.C. number and the same substrate.
- All the wild-type entries were paired with all the mutants in those groups.
- The final data set consists in 9,406 pairs (with >90 % sequence similarity)
- Splitting into training and test sets, along with cross-validation, ensures that a mutant cannot be found in both the training set and the test set.

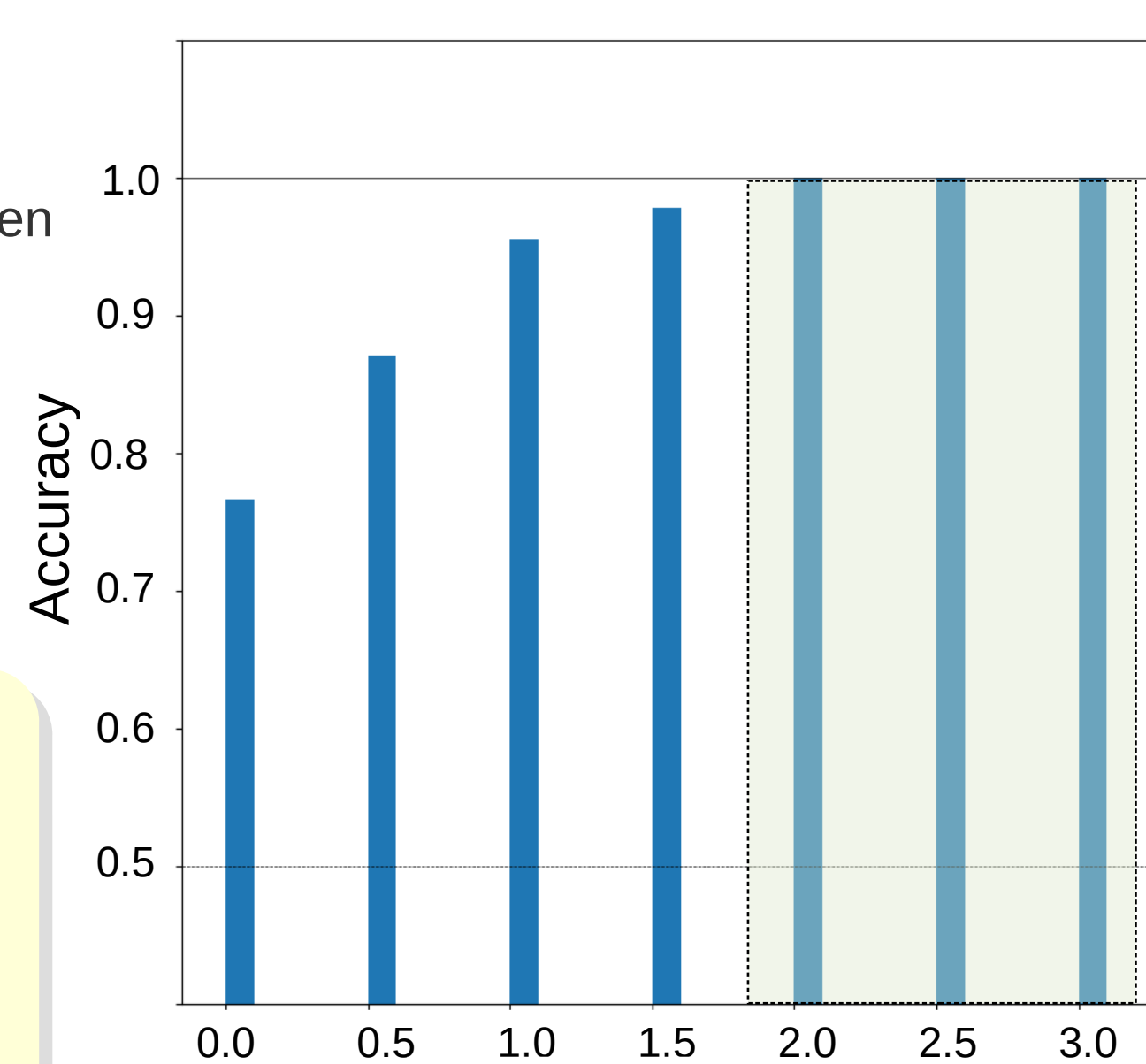
Results and Perspectives



● Fully unseen pair
R2 = 0.34
● WT seen, mutant unseen
R2 = 0.58
● Fully seen
R2 = 0.60
□ Area for $\delta = 2$
($|\Delta k_{cat} \text{ pred}| > 2$)

Definition
 δ is the distance from $\Delta k_{cat} \text{ pred} = 0$.

"The larger the predicted change, the higher the confidence in the direction of that change."



Here, the accuracy is defined as the ratio of true positives (TP) and true negatives (TN) to the total number of points in the region where $|\Delta k_{cat} \text{ pred}| > \delta$.

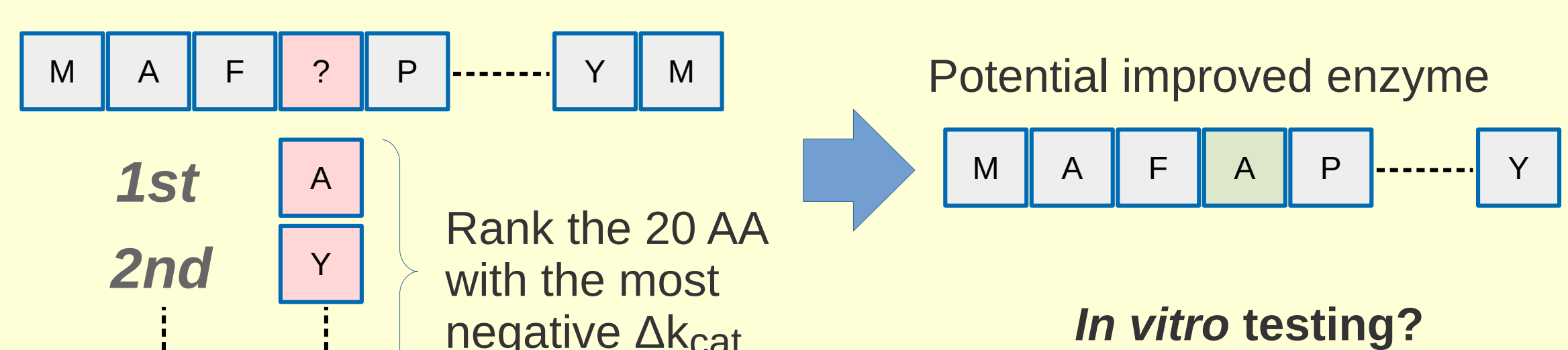
A coefficient of determination of 0.58 was achieved for predicting the change in k_{cat} from a mutant to its wild-type.

In 78% of the cases, the model correctly predicts the direction of the change.

When the model predicts a large Δk_{cat} for a mutation, the direction is likely to be correct (100% of accuracy on the test set when $|\Delta k_{cat} \text{ pred}| > 2$)

Ongoing: Building a model for ΔK_M . Ideally, a final model that can predict increased $k_{cat}/\Delta K_M$, the true "efficiency of an enzyme"

Toward designing sequence with enhanced kinetics



References

- [1] Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, Guo D, Ott M, Zitnick CL, Ma J, Fergus R. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc Natl Acad Sci U S A. 2021
- [2] Kroll A, Rousset Y, Hu XP, Liebrand NA, Lercher MJ. Turnover number predictions for kinetically uncharacterized enzymes using machine and deep learning. Nat Commun. 2023