# DESIGN OF NOVEL PROTEINS FROM LATENT SPACES OF VARIATIONAL AUTOENCODERS

Designing better biocatalysts requires understanding the pathway of protein evolution. We used variational autoencoders (VAEs) to capture the nonlinear epistasis between different positions in protein sequences.

Pavel Kohout, Marika Majerova, Michal Vasina, Martin Marek, Zbynek Prokop, David Bednar, Jiri Damborsky, Stanislav Mazurenko

email: pavel.kohout@recetox.muni.cz

Loschmidt Laboratories, Department of Experimental Biology and RECETOX, Faculty of Science, Masaryk University, Brno, Czech Republic

International Clinical Research Centre, St. Anne's Hospital, Brno, Czech Republic



# INTRODUCTION

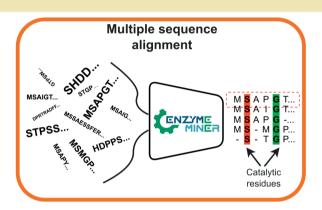
The use of evolutionary information has been shown to be important in the design of new, more stable protein variants for practical applications [1, 2]. However, finding the right sequence for desired properties is challenging due to the vast space of possible sequences. Deep neural networks can capture evolutionary information and the full variation of sequences in multiple sequence alignments, which can guide researchers in producing improved protein variants.

## **RESULTS**

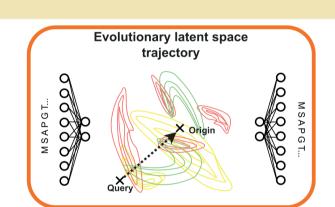
The VAE evolutionary profile identified 14 candidates for laboratory experiments, out of which 4 were stable and soluble. The four successful designs were experimentally characterized for stability, activity, and specificity.

## **OBJECTIVE**

Harness the potential of evolutionary dependencies within VAE latent space to engineer novel stable and active variants of haloalkane dehalogenase DhaA of the  $\alpha/\beta$ -hydrolase 1 family by manipulating latent coordinates of protein sequences.







Ding and co-authors [3] recently showed that

the latent space of VAEs can guide

researchers in producing improved protein

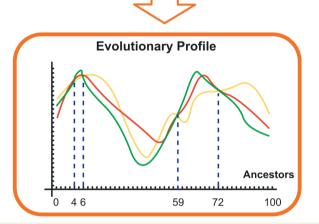
mutants and reconstructing phylogeny. This

provides an alternative to the conventional

method of ancestral sequence reconstruction

[2], which treats positions independently.

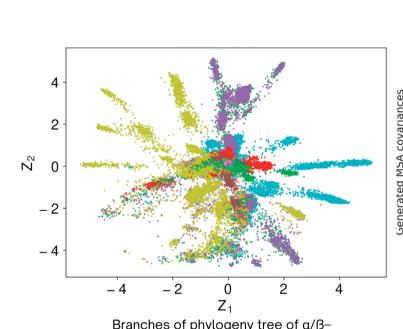
**METHODOLOGY** 



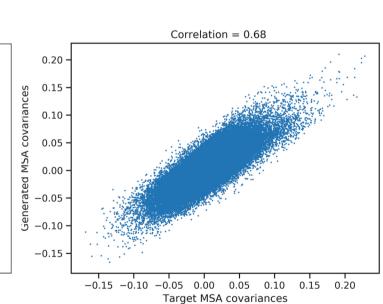
# Laboratory Experiments Ancestor 9 Ancestor 5 Ancestor 7 WT WT

# **ANALYSIS**

- 1. EnzymeMiner identified 22 567 hits, which were aligned, clustered at 90% sequence identity, and filtered for overlaps with the query.
- 2. A shallow neural network with 299 neurons was selected for both the encoder and decoder. The resulting latent space was two-dimensional.
- 3. The reconstruction strategy consisted of selecting 100 points on a straight line from the query to the origin in the latent space.
- 4. Several metrics were retrieved for each point to allow further selection: VAE probability, closest sequence identity, counts of indels, and substitutions.



Branches of phylogeny tree of  $\alpha/\beta$ -hydrolase family (PF00561) mapped into latent space.



Second-order covariances describe
VAEs' ability to design protein sequences
that belong to a particular family

# **CONCLUSIONS**

We examined VAE latent space as an orthogonal method for ancestral sequence reconstruction for generating ancestral variants of haloalkane dehalogenase (DhaA). The soluble variants had up to 45 mutations and showed activity similar to that of the wild type. However, they did not exhibit improved stability. We will further advance our approach by tackling solubility and guiding the search for stable proteins using a novel predictor TmPred.

# Related literature

1. BORNSCHEUER, U. T., G. W. HUISMAN, R. J. KAZLAUSKAS, S. LUTZ, J. C. MOORE and K. ROBINS. Engineering the third wave of biocatalysis. Nature. 2012, 485(7397), 185-194. ISSN 0028-0836.

2. MERKL, R. and R. STERNER. Ancestral protein reconstruction: techniques and applications. Biological Chemistry. 2016, 397(1), 1-21. ISSN 1437-4315.

3. DING, X., Z. ZOU a C. L. BROOKS III. Deciphering protein evolution and fitness landscapes with latent space models. Nature Communications. 2019, 10(1),1-13. ISSN 2041-1723.

4. HON, J., S. BORKO, J. STOURAC, Z. PROKOP, J. ZENDULKA, D. BEDNAR, T. MARTINEK a J. DAMBORSKY. EnzymeMiner: automated mining of soluble enzymes with diverse structures, catalytic properties and stabilities. Nucleic Acids Research. 2020, 48(1), 104-109. ISSN 0305-1048.

