Engineering Dehalogenase Enzymes using Variational Autoencoder-Generated Latent Spaces and Microfluidics

Pavel Kohout[#,a,b], Michal Vasina[#,a,b], Marika Majerova[a,b], Veronika Novakova[a,b], Jiri Damborsky[a,b], David Bednar[a,b], Martin Marek[a,b], Zbynek Prokop[a,b,*], Stanislav Mazurenko[a,b,*]

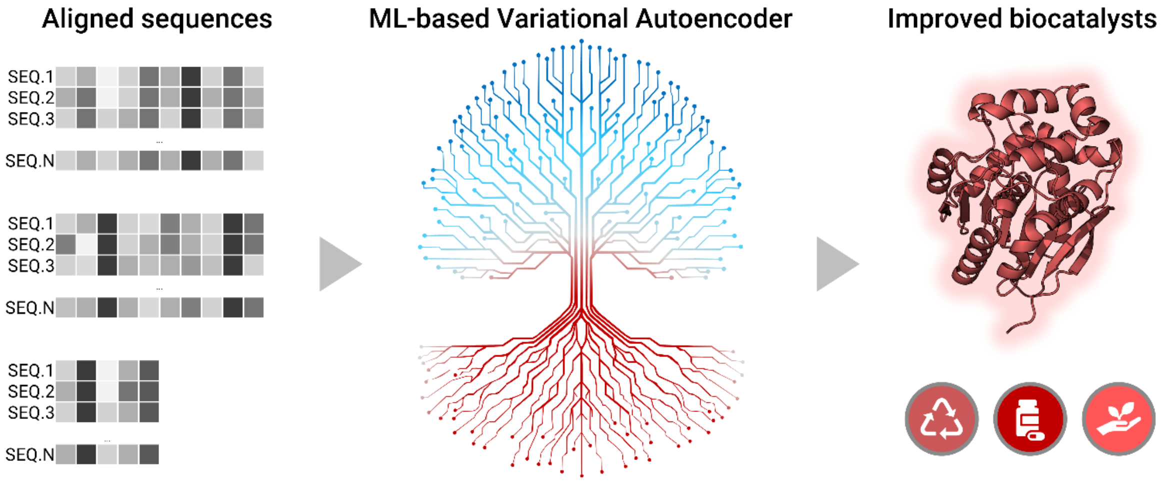
[a] Loschmidt Laboratories, Department of Experimental Biology and RECETOX, Faculty of Science, Masaryk University, Brno, Czech Republic

[b] International Clinical Research Centre, St. Anne’s Hospital, Brno, Czech Republic

[#] These authors contributed equally.

[*] Corresponding authors: Stanislav Mazurenko - mazurenko@mail.muni.cz; Zbynek Prokop - zbynek@chemi.muni.cz.

Graphical abstract



Abstract

Enzymes play a crucial role in sustainable industrial applications, with their optimization posing a formidable challenge due to the intricate interplay among residues. Computational methodologies predominantly rely on evolutionary insights, leveraging homologous sequences to pinpoint conserved and functionally critical regions. However, despite their notable advancements, deciphering the evolutionary variability and complex dependencies among residues presents substantial hurdles. Here, we present a new machine-learning method based on variational autoencoders in combination with a simple evolutionary sampling strategy to address those limitations. We customized our method to generate novel sequences of haloalkane dehalogenases, enzymes widely used in biodegradation, biocatalysis, and biosensing. Three consecutive design-build-test cycles improved the solubility of variants from 11% to 75%. Thorough experimental validation using the state-of-the-art microfluidic device MicroPEX resulted in 20 multiple-point variants. Nine of them, sharing as little as 67% sequence similarity with the template, showed a melting temperature increase of up to 9°C and an average improvement of 3°C. The most stable variant demonstrated a 3.5-fold increase in activity compared to the template, while five variants exhibited average dehalogenase activities. High-quality experimental data collected with 20 variants represent a valuable dataset for the critical validation of novel protein design approaches and scoring functions. Python scripts and data sets are available on GitHub (https://github.com/loschmidt/vae-dehalogenases), and interactive calculations will be possible via an easy-to-use website: https://loschmidt.chemi.muni.cz/fireprotasr/.

Keywords

dehalogenase, protein engineering, machine learning, microfluidics, protein, variational autoencoder

Highlights

- Variational Autoencoder deciphers evolutionary patterns for protein sequences
- Novel sampling strategy generates dehalogenases for biocatalysis and biosensing
- Twenty designs yielded soluble proteins thoroughly characterized by microfluidics
- Solubility success rate was improved more than six-fold over three protein design iterations
- The best variant showed improved stability by 9°C and 3.5-fold higher activity

1. Introduction

Biocatalysis is a promising field that offers sustainable and environmentally friendly solutions for industries increasingly driven by the remarkable capabilities of enzymes. Thanks to millions of years of evolution, enzymes are fine-tuned to carry out specific chemical reactions with high efficiency. This makes them attractive alternatives to traditional catalysis, which often relies on harsh conditions and toxic chemicals [1]. Thus these biocatalysts find application across various industries, including pharmaceuticals, food production, and sustainability efforts aimed at reducing waste and energy consumption [2]. Since natural enzymes often exhibit suboptimal performance in non-native environments, enzyme engineering is usually required to unlock their full potential [3,4]. In addition to commonly used experimental approaches such as directed evolution, scientists can also expedite the process and reduce associated development costs by incorporating computational methods [5,6]. These methods help navigate the vast sequence space, as it is estimated that only a fraction of all possible sequences fold into functional protein structures [7]. Most natural proteins have marginal stability [8], thus posing a significant risk for any manipulations with their sequences. Many computational methods aiming to reduce the search space rely on homologous sequences [9,10]. These sequences of different but related proteins stemming from a common ancestor contain rich evolutionary information [11]. Homologous protein sequences can be employed to identify conserved and functionally important regions, suggest beneficial mutations, and create phylogenetic trees to reconstruct ancestral sequences [12]. Ancestral sequence reconstruction has proven to be a promising strategy for enhancing protein stability [13–16].

Despite the recent progress in extracting evolutionary information from multiple sequence alignments (MSA) of homologous proteins, analyzing this variability is challenging. Historically, this data was used primarily by looking at only one or two positions at a time [17,18]. More recent approaches extract patterns by deep neural networks, in particular algorithms that map the sequence space onto their internal low-dimensional representation, also referred to as latent spaces. Generative models trained on large datasets of tens of thousands of sequences have shown excellent results in producing highly interpretable embeddings and generating novel protein variants [19–22]. Variational autoencoders (VAEs) show a particular promise in this domain due to the explicit modeling of the latent space [23]. They have already proven useful in several applications, including predicting protein structures [24], discovering novel drugs [25], and predicting protein functions [26]. By learning the latent space representation of a specific family, VAEs provide valuable insights into the evolution of protein families, as demonstrated in recent studies exploring the phylogenetic relationships within the latent space [27–29]. In particular, Ding et al. showed that the latent space of the variational autoencoders can capture the biophysical properties of protein variants and the phylogenetic relationships within protein families [27]. However, the study did not offer a strategy that would allow exploiting these relationships to generate new proteins from the latent space.

Here we suggest a simple strategy to leverage the geometry of the VAE-learned latent space to produce novel variants of haloalkane dehalogenases (HLDs; EC 3.8.1.5). These enzymes cleave the carbon-halogen bonds [30] and are widely used in biocatalysis, biosensing, cell imaging, and protein analysis [31]. The proposed workflow is based on a small number of proteins with known functions and aims to produce new variants that preserve catalytic function and improve stability (**Fig. 1**). First, we mined sequences with preserved catalytic residues using the EnzymeMiner [32] to obtain an MSA of functionally related proteins. Second, we trained a VAE and specified several metrics to measure its capacity to generate protein sequences and capture the phylogeny in the constructed latent representations. Third, based on the geometry of the latent space, we developed a sampling strategy and produced a statistical profile of candidate sequences to select promising variants from the evolutionary trajectory. Fourth, we over-expressed and characterized variants experimentally using advanced microfluidics. Three consecutive rounds of experimental characterization and workflow optimization resulted in 20 variants, sharing as little as 67% sequence similarity to known HLDs. Obtained enzymes showed up to a 9°C increase in melting temperatures and an average improvement of 3°C across all soluble variants. We also observed a boost in activity, up to 3.5-fold for the most stable variant, whereas most of the other expressed variants showed activity levels comparable to benchmark enzymes.

2. Materials and Methods

MSA and data preprocessing

The HLD I-IV dataset is created by the EnzymeMiner tool [32]. As an input sequence, 3.8.1.5 - Haloalkane dehalogenase was used, and from the sequence selection table, all 33 provided sequences together with their essential residues were selected. In advanced options, only the maximum number of hits in PSI-BLAST was changed to 50000. The resulting dataset was composed of 22567 sequences (job ID xvmwa7). For a given query sequence, the MSA was preprocessed through several steps: (i) the gap positions in the query were removed for every input sequence except for the columns where less than 20% of sequences had gaps; (ii) sequences with gaps in more than 40% of positions were removed; (iii) sequences were clustered by 90% identity to support the diversity, and only one sequence for each cluster was picked for the training dataset, together with the query sequence, (iv) sequences with less than 50% overlap with the query were excluded. The sequence P59336_S14 of DhaA (Haloalkane dehalogenase from *Rhodococcus sp*) was chosen as a query for the training process. This resulted in 12053 sequences with 299 positions in the MSA left for the training after the preprocessing.

The HLD I-II dataset for training models used in the third round of experiments was created by a similar protocol. As input sequences, only DhlA and LinB were selected together with all 19 sequences in the “Other known sequences” input field. The resulting dataset was composed of 4053 sequences (job ID d6cv1o). As a query for MSA preprocessing, the DhaA sequence (Uniprot ID P0A3G3) was selected. Experimental results from rounds 1 and 2 indicated that the poor solubility might stem from many deletions suggested in the reconstructed sequences. Therefore, we adjusted parameters of MSA preprocessing to lower the number of gaps. In addition to the previous preprocessing steps, we implemented a further refinement by excluding columns from the MSA where amino acid symbols were present in the query if the column contained more than 80% gaps (>70% for Model 3). These excluded query amino acids were stored separately and reintegrated into their original positions in the generated sequence. Also, we did not perform 90% identity sequence clustering.

Two sets of experimentally measured stability values were mapped to the latent space of Model 1. The first set consisted of six ancestral sequences from the previous ancestral campaign of the thoroughly characterized dehalogenases DbjA, DbeA, DhaA, DmxA, and DmmA [33]. These sequences were realigned with the MSA profile of the original input sequences and only retained the corresponding columns that remained after the MSA processing procedure. The second set consisted of 24 previously engineered DhaA variants (the DhaA115 dataset) incorporating both evolutionary and energy-based mutations based on the FireProt method [15,34,35]. Similarly, for the DhaA115 dataset, we performed individual sequence alignment with the query sequence P59336_S14, preserving the indices that were retained after the MSA preprocessing step.

Variational autoencoders and training

VAE is a generative machine learning algorithm based on the classic autoencoder architecture composed of two parts denoted as encoder and decoder. To make the algorithm more suitable for the inference of unseen sequences, the VAE framework adds the regularization term into the loss function and Monte Carlo sampling to the learning process. The sampling parameters are determined by the encoder learning a mean and deviation for input sequences. We used the one hidden layer in the encoder and decoder, both composed of N neurons, where N is the width of preprocessed MSA (number of positions), the latent space dimensionality of 2, and zero weight decay. The final model had 3 million parameters. We used the Adam optimizer with a learning rate of 0.001 and stopped training after not improving the loss function for more than 3 consecutive rounds.

Model generative capacity

The first-order statistic is a comparison of the frequency of occurrence of each amino acid on each position between two sets of data. The frequency of amino acid α at position *i* in MSA with *N* sequences is given by the following formula:

$$f^i_\alpha = \frac{count_i(\alpha)}{N}, \qquad \textbf{(Eq. 1)}$$

where $count_i(\alpha)$ is the number of occurrences of the amino acid α in the MSA. The second-order statistics are similar but for two positions i, j:

$$f^{ij}_{\alpha\beta} = \frac{count_{i,j}(\alpha,\beta)}{N}, \qquad \textbf{(Eq. 2)}$$

where $count_{i,j}(\alpha,\beta)$ is the number of occurrences of the amino acid pair (α,β) in the MSA. To gain more insight into the distribution of features in the generated dataset, we further compute the pairwise covariance score given by the following formula:

$$C^{ij}_{\alpha\beta} = f^{ij}_{\alpha\beta} - f^i_\alpha f^b_\beta, \qquad \textbf{(Eq. 3)}$$

where $f^{ij}_{\alpha\beta}$ and f^{i}_{a}, f^{b}_{β} are the second and first-order statistics for columns *i, j* of the alignment, respectively. Each covariance term measures the difference between the joint frequency for pairs of amino acids and the product of the frequencies of the residues at each site, i.e. the expected counts under the statistical hypothesis of independence. Therefore, zero $C^{ij}_{\alpha\beta}$ for all $\alpha\beta$ would imply frequences one would observe if positions i, j were independent. Reproducing pairwise covariance in protein alignments is an important aspect of generative models as it measures how well the model captures interactions between distant amino acids, an essential indicator for the likely stability and function of the generated proteins [36].

For each model, we compared the pairwise covariance scores for all positions and residuals in generated $(\hat{C}^{ij}_{\alpha\beta})$ and the input $(C^{ij}_{\alpha\beta})$ alignment using the Pearson correlation coefficient $\rho(\{C^{ij}_{\alpha\beta}\},\{\hat{C}^{ij}_{\alpha\beta}\})$. In the case of this study, the statistics were calculated for the input and synthetically generated datasets having 3,000 randomly selected samples. Synthetic data were reconstructed from the latent space points sampled according to the a priori chosen distribution, i.e., a Gaussian distribution with the zero mean and the variance of 2.

Average reconstruction accuracy and controls

The average reconstruction accuracy of the sequence was approximated as an average reconstructed sequence identity for 5,000 samples around the original sequence coordinates of its latent space embedding. The parameters of sample distribution are determined by the encoder (mean, variance) for a given sequence. The negative control subset was generated by sampling sequences from the profile of input MSA only based on the amino acid frequencies in each position. The positive control subset comprised 5% of preprocessed sequences randomly selected from the MSA and excluded from the training. The training control subset was built from a training dataset. Finally, the ancestral subset was composed of 100 reconstructed sequences by the straight evolutionary strategy. All controls contained the number of sequences corresponding to 5% of the preprocessed dataset.

Phylogeny mapping and evaluation

We generated 13 phylogenetic trees using our input MSA dataset for the latent space phylogenetic analysis. Each tree consisted of approximately 100 randomly sampled leaf nodes from the preprocessed MSA. Using our fully automated in-house ancestral sequence reconstruction tool FireProtASR [15], we obtained an average of 10 levels in each phylogenetic tree. To explore the relationship between the tree branches and the latent space, we mapped each individual branch, along with its reconstructed ancestral sequences, into the latent space. We then measured the correlation between the depth of a node (i.e., the distance between the root and the node) and the distance of the corresponding ancestral node's latent embedding from the origin of the latent space, following an approach similar to [27]. In addition, we sought to gain insights into the reconstruction strategy by evaluating the direction in which the tree branches were mapped in the latent space. Specifically, we calculated the vector representing the first principal component of each branch and computed the dot product with the vector pointing from the embedding of the leaf node sequence to the origin of the latent space. This analysis allowed us to compare the trajectory directions of the different tree branches, and we reasoned that the straight evolutionary strategy would generate ancestral-like sequences.

Conditional variational autoencoders and determination of solubility bin ranges

In the conditional variational autoencoders (CVAE) [37], the VAE mechanism was extended with a tag added to the input of the encoder and decoder. Three tag values corresponding to the LOW, MEDIUM, and HIGH bins of sequence solubility values were used. The learning objective for CVAE can be interpreted as creating different representations of the latent space for each tag but with shared network parameters. Generating a new sequence from the latent embedding space can then be done for a desired tag (e.g. HIGH to maximize solubility), and the CVAE is forced to introduce the patterns observed for the corresponding solubility bin into the generated design.

The output of EnzymeMiner includes solubility labels predicted by SoluProt [38] for all output sequences in MSA. These values range from 0 to 1, with larger values corresponding to higher probability of soluble expression in *E. coli*. The queried DhaA sequence (DhaA_S19) had a predicted solubility of 0.87. The solubility distribution of the HLD I-II dataset can be seen in **Fig. S5**. Most of the sequences had predicted solubility of less than 0.55. Therefore, the SoluProt tags were divided into 3 bins with thresholds set to make the distribution of sequences in each bin as uniform as possible (LOW: < 0.35, MEDIUM: 0.35-0.55, HIGH: > 0.55). A nearly uniform distribution for the solubility bins was chosen to encourage the CVAE to extract sequential patterns based on a balanced number of samples [39].

AlphaFold structure prediction and manual analysis of the suggested mutations

For structural predictions of ancestral sequences, the AlphaFold2 Google Colaboratory implementation, ColabFold, using MMseqs2, was used [40]. We used amber relaxation for top-ranked structure, no templates provided, unpaired+paired pair_mode, model_type auto, and 3 cycles in the pipeline settings. The relaxed first-ranked structure was used as the result of the prediction.

In round 3, the proposed mutations by VAE were also curated manually (see **Section 2 Manual curation of designs in round 3 of SI** for more detail). The visual inspection of the modeled AlphaFold variants was performed by Pymol [41], and the MutCompute web server [42] was used to calculate the score per residue as the log-likelihood ratio of the substitution residue to the original residue type. Thus, a positive score indicates that MutCompute assesses the substituted residue as more likely to occur in the given structural microenvironment than the wild-type residue.

Cell transformation

E. coli BL21(DE3) cells (NEB, USA) were transformed with expression plasmid vector pET21b containing the corresponding gene and plated on LB-agar containing 100 µg/ml ampicillin and then incubated at 37°C overnight (12-16 h). The cells transformed with pET21b::DhaAwt, pET21b::RLuc and empty pET21b were used as controls.

Small-scale protein over-expression and affinity purification test

Several *E. coli* colonies were streaked to inoculate 2 ml of starting media (2xLB supplemented with 0.5% glucose and 100 µg/ml ampicillin) in a 24-deep-well plate (GE Healthcare, UK). The plate was covered with an air-pore membrane and incubated at 37°C for 4 hours, 200 rpm. After incubation, 2 ml of induction media (2xLB supplemented with 0.6% lactose, 50 mM HEPES (pH 7.4), 0.5 mM IPTG, and 100 µg/ml ampicillin) was added. The plate was covered with an air-pore membrane and incubated at 22°C for 16 hours, 200 rpm. Cells were harvested by centrifugation using Sigma 6K-15 centrifuge (SciQuip, UK) for 10 min, 1519 g, and 4°C, and resuspended in 1.3 ml of a purification buffer (16.4 mM K_2HPO_4, 3.6 mM KH_2PO_4, 400 mM NaCl, 10 mM imidazole, pH 7.5). After cell disruption (Sonic Dismembrator Model Q700S, FisherBrand, USA) the whole soluble fraction was clarified by centrifugation for 20 min, 3572 g, and 4°C. Soluble fraction was added to TALON SuperFlow Metal Affinity Resin (Takara) pre-equilibrated with sterile water and incubated for 2 hours on a roller (40 rounds/min) at 4°C. Unbound proteins were washed twice by and centrifuged at 94 x g for 2 min followed by resuspending in a purification buffer. After the second wash, 40 µl of SDS-PAGE loading buffer (2x Laemmli Sample buffer containing DTT) was added to each protein/resin sample. Samples and a marker (Color Prestained Protein Standard, Broad Range 10–250 kDa, New England Biolabs, USA) were loaded on SDS-PAGE gel with run conditions: 400 mA, 200 V, 40 min. After staining with InstantBlue™ (Missouri, USA) for 20 min, the gel was washed with water for 40 min.

Cell cultivations for enzymatic screenings and HOX assay

Single colonies of transformed cells were transferred into sterile 96-well plates (MTP) containing 100 µl of LB medium supplemented with ampicillin (100 µg/ml). The plates were covered with air-pore membrane and cultivated for 3 h at 37°C and 200 rpm. After that, an additional 100 µl of LB medium with ampicillin (100 µg/ml) and IPTG (1 mM) were added to the mini-cultures and MTP was afterward incubated at 20°C, 200 rpm, for 18 hours. Cell cultures were harvested by centrifugation at 4°C, 1600 x g, 20 min. The supernatant was discarded, and the cell pellets were washed with 200 µl of reaction buffer (1 mM orthovanadate, 20 mM phosphate buffer, pH = 8.0). MTP was centrifuged again at 4°C, 1600 g, 20 min and the washing step was repeated twice. Finally, the pellets were resuspended in 200 µl of the reaction buffer, and optical density (OD_{600}) was determined spectrophotometrically. Into each well of a new black bottom 96-well MTP plate, 100 µl of MasterMix was dispensed: 25 µM aminophenyl fluorescein, 26 mM H2O2, 1.1 U *Curvularia inaequalis* vanadium chloroperoxidase with additional His tag, 1 mM orthovanadate, 20 mM phosphate buffer, pH = 8.0. Also, 4 µl of resuspended cells were added into each well, followed by the addition of 96 µl of 0.3 mM 1,2-dibromoethane in the reaction buffer. Fluorescence was measured using Synergy™ H4 Hybrid Microplate Reader (BioTek, USA) (Excitation at 488 nm; emission detection at 525 nm; 30°C). Data for all tested variants were measured in four biological replicates and average activity with a standard deviation of four measurements was determined.

Large-scale protein over-expression

Several colonies of *E. coli* were incubated in 10 ml 1x LB medium supplemented with 100 µg/ml ampicillin. The pre-culture was incubated at 37°C for 4 hours. After incubation, the pre-culture was added to 1 liter of 1x LB medium supplemented with 100 µl/ml ampicillin. The culture was incubated to OD_{600} = 0,8 and expression was induced by the addition of IPTG to a final concentration of 0.5 mM. The cell culture was incubated at 20°C, 150 rpm, 16 h and harvested by centrifugation at 4000 rpm, 4°C, 25 min. The cell pellet was resuspended in approximately 30 ml of harvesting purification buffer A: 16.4 mM K_2HPO_4, 3.6 mM KH_2PO4, 400 mM NaCl, 10 mM imidazole, pH 7.5, and frozen at -70°C.

Protein purification with metal affinity resin

The DNase was added to the cell culture (20 µg/ml) after defrosting from -80°C. The culture was sonicated (Sonic Dismembrator Model 705 Fisher Scientific, USA) in 6 x 2-minute cycles with a 50 % amplitude (5 s pulse, 5 s pause). The cell suspension was centrifuged (21036 g, 4°C, 1 h) using a Sigma 6-16K centrifuge (SciQuip, UK) equipped with the 12166 rotor. 1 liter of cell-free extract was divided into two 50 ml conical centrifuge tubes with washed Resin (TALON® Superflow Metal Affinity Resin, Takara). The mixture was incubated for 1.5 hours at 4°C on a roller (40 rounds/min). After incubation, the resin with bound proteins was centrifuged in a pre-cooled centrifuge for 10 min, 130 g, 4°C using a Sigma 2-16K centrifuge (SciQuip, UK) equipped with the 11192 rotor. The supernatant was discarded, and harvesting buffer A (16.4 mM K_2HPO_4, 3.6 mM KH_2PO_4, 400 mM NaCl, 10 mM imidazole, pH 7.5) was added and transferred to a gravity-flow column equilibrated in the same buffer. The column with mixture resin protein was washed with approximately 200 ml of harvesting buffer A, and a 50 mM phosphate buffer (pH 7.5) was gradually added to the column. The protein was eluted from the resin using an elution buffer of 50 mM phosphate buffer, 300 mM imidazole (pH 7.5).

Purification by gel filtration on FPLC

Affinity-purified proteins were purified in the second step by gel filtration on ÄKTA Pure™ (Cytiva, USA) equipped with HiLoad 16/600 Superdex 75 pg column. After column equilibration with 50 mM phosphate buffer (pH 7.5), proteins were purified using the same buffer and concentrated on an Amicon® Ultra-15 Ultracel-10 gravity flow column 10K (Merck Milipore Ltd.).

Secondary structure experimental validation

The secondary structure of the analyzed variants was experimentally verified using circular dichroism (CD) spectroscopy. CD spectra were measured at 15°C using a spectropolarimeter Chirascan

(Applied Photophysics). The samples were dissolved in 1 mM HEPES buffer or in the 50 mM Phosphate buffer and their concentration was adjusted to approximately 0.18 mg/ml. Data were collected from 185 to 260 nm with 0.25 s integration time and 1 nm bandwidth using a 0.1 cm quartz cuvette. Each spectrum was obtained as an average of five individual repeats. Prediction of CD spectra has been performed by the web tool PDBMD2CD [43] (https://pdbmd2cd.cryst.bbk.ac.uk). For this prediction, either experimental structures from PDB database (1CQW for DhaA) or AlphaFold models were taken as input. The estimation of secondary structure elements from experimental data and PDB database structures (uploading own PDB files – AlphaFold models – was unfortunately not available at the moment of data analysis) was performed additionally by the web tool BeStSel [44] (https://bestsel.elte.hu/).

Thermal denaturation by CD

Thermal unfolding of selected enzyme variants was carried out using a Chirascan spectropolarimeter (Applied Photophysics, UK). Each protein sample was diluted in 50mM Phosphate buffer to the concentration of 0.18 mg·mL^{-1} and measured in a 0.1 cm quartz cuvette. Changes of ellipticity were monitored at three wavelengths, 195 nm, 210 nm, and 227 nm from 15°C to 80°C with a 0.1°C resolution and 1°C·min^{-1} heating rate. Recorded data were fitted using the model "Sigmoid curve + slope" in the Pro Data Viewer software (Applied Photophysics, UK). The apparent melting temperature (T_m) was evaluated as a midpoint of the normalized thermal transition.

Thermal denaturation by nanoDSF

Thermal unfolding was studied using NanoDSF Prometheus (NanoTemper, Germany) by monitoring Trp fluorescence over the temperature range of 20°C to 95°C, at a heating rate of 1°C/min with 20% excitation power. The thermostability parameters (T_{on} and T_{mapp}) were evaluated directly by ThermControl v2.0.2.

Dehalogenase activity measurements on MicroPEX

Activity measurements for the determination of temperature profiles and substrate specificity were conducted on the capillary-based droplet microfluidic platform MicroPEX [45], enabling the characterization of specific enzyme activity within droplets for multiple enzyme variants in one run. A detailed description of the microfluidic method can be found elsewhere [46,47].

Briefly, the droplets were generated using Mitos Dropix (Dolomite, UK). A custom sequence of droplets (150 nl aqueous phase, 300 nl oil spacing) was generated using negative pressure (microfluidic pump). The droplets were guided through a polythene tubing to the incubation chamber. Within the incubation chamber, the halogenated substrate was delivered to the droplets via a combination of microdialysis and partitioning between the oil (FC 40) and the aqueous phase. The reaction solution consisted of a weak buffer (1 mM HEPES, 20 mM Na_2SO_4, pH 8.2) and a complementary fluorescent indicator 8-hydroxypyrene-1,3,6-trisulfonic acid (50 µM HPTS). The fluorescence signal was obtained using an optical setup with an excitation laser (450 nm), a dichroic mirror with a cut-off at 490 nm filtering the excitation light, and a Si-detector. By employing a pH-based fluorescence assay, small changes in the pH were observed and enabling monitoring of the enzymatic activity. Reaction progress was analyzed as an end-point measurement recorded after the passing of 10 droplets/sample through the incubation chamber. The reaction time was 4 min. The raw signal of every single measurement was first processed by the in-house LabView-based (National Instruments, USA) software MicroPEX Data Analyzer 1.0. The peaks were assigned to the particular sample, and the mean signal was calculated for them. The output XLS file gathering mean signal values for every sample type (calibration, enzyme activity, buffer, blank 3 buffer, and blank enzyme) for the dataset served as an input for the MatLab (Mathworks, USA) script to calculate specific activities using the same principle as for previous measurements [45,47]. The activities were classified as “not determined” whenever the measured product concentration was below the limit of detection (LOD – 3 times the standard deviation of the noise signal). Each substrate had a different calibration curve, so the LOD product concentration was in the range of 10-100 μM).

The matrix containing the activity data of 32 previously identified HLDs [45] and the activities obtained for all variants in this study (all measured on MicroPEX) was analyzed by PCA in MATLAB (MathWorks, United States) to uncover the relationships among individual HLDs (objects) based on their activities toward the set of halogenated substrates (variables). Two PCA models were constructed to visualize systematic trends in the dataset. The first one was done on the raw data, which ordered the enzymes according to their total activity. The second PCA was carried out on the log-transformed data after adding 1 to each specific activity to avoid taking the logarithm of zero. The resulting values were then divided by the sum of the values for a particular enzyme. These transformed data were used to calculate principal components, and the components explaining the highest variability in the data were then plotted to identify substrate specificity groups.

3. Results

We developed the pipeline to leverage the power of the variational autoencoder and its latent spaces for the design of promising biocatalysts (**Fig. 1**). This pipeline was inspired by the previous studies reporting the connections between latent space geometry and phylogeny for a given MSA [27,29]. Our main motivation for the current study was to exploit this connection to generate new protein sequences, the direction that was not explored in the previous studies. We iteratively executed our pipeline across three rounds, each iteration followed by experimental validation to improve our workflow (**Appendix section 1**). In both the first and second rounds, we utilized Model 1, with the only difference being a revised selection of VAE ancestors. In the third round, we introduced variations to the MSA preprocessing and additionally manual curation of the generated ancestors by AlphaFold and MutCompute [11,42]. In addition, during the third round of experimental validation, we explored the possibility of conditioning the VAE on solubility scores returned by the ML-based tool SoluProt [38]. In total, three trained VAE models were explored in the third round, referred to as Model 2, Model 3, and Model 4, to better understand the strengths and weaknesses and refine our approach.

3.1 Multiple sequence alignment processing

The data collection is optimized to preserve catalytic activities. The first step of our pipeline is to construct an MSA. Instead of using Pfam alignments as in [27], we narrowed the search of relevant sequences to those likely to preserve the dehalogenation activity. Pfam MSAs are sometimes too broad, introducing large-gapped regions and making it difficult to design proteins with desired functions [48]. To overcome this challenge, we used the EnzymeMiner web tool [32], which generates alignments specifically selected for function and catalytic site similarity (**Fig. 1A**).

The query of haloalkane dehalogenase (DhaA) from *Rhodococcus* strain TDTM0003 with UniProt ID P59336 yielded 22,567 sequences in EnzymeMiner. This extensive search resulted in the creation of a dataset named HLD I-IV. To further refine the results, we preprocessed the resulting MSA against the DhaA query by removing protein sequences and positions with too many gaps. This step narrowed down the size of the alignment to 12,053 sequences and 299 positions, which were used for training.

Similar to [27], each sequence in the MSA was then represented as a matrix of the size L x 21 with one-hot encoded residues, where L stands for the number of positions in the final alignment (299), and 21 columns correspond to 20 amino acids and a gap. The HLD I-IV dataset was utilized in both the first and second rounds of wet lab experiments (**Section 1 Workflow overview of SI, Fig. S1**).

Later on, based on the low solubility of the design observed in the first two rounds of experimental characterization, we applied a stricter protocol for creating the initial MSA. In particular, inspired by the conclusions in Vasina et al. [45] about the low solubility of members from HLD subfamilies III and IV, we created a much smaller MSA, instructing EnzymeMiner to search among HLD subfamilies I and II only. This dataset includes the DhaA enzyme from *Rhodococcus sp.* with UniProt ID P0A3G3, well-characterized in our lab, which we then used as an updated query for the MSA preprocessing with additional filters to further reduce the gap frequencies in comparison to previous rounds: we lowered the threshold for gap column removal and included additional filtering of columns with frequent gaps, even when the query position had an amino acid in that position. These removed positions and corresponding amino acids were then added back to the reconstructed sequences at the end of the pipeline, i.e., we treated them as invariable. By this, we lowered the width of MSA to 293 positions with 4,053 sequences.

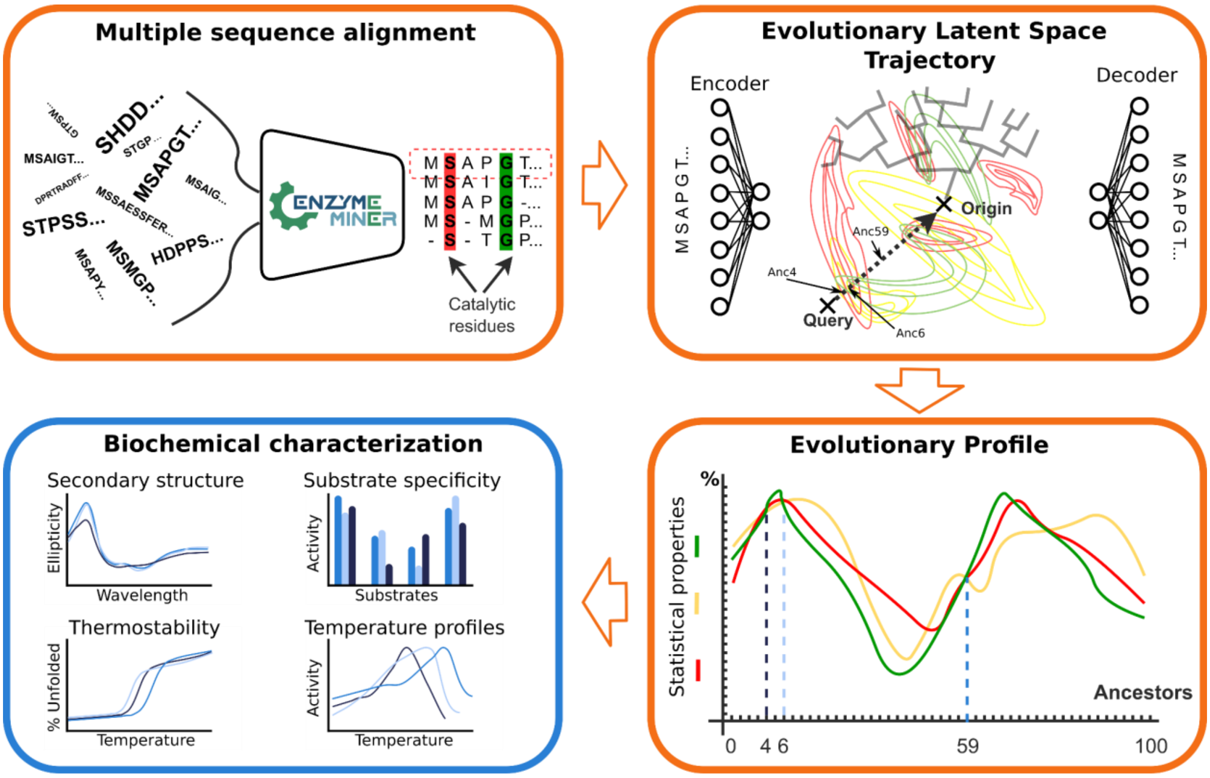


Fig. 1: The scheme of the variational autoencoder-based pipeline for the design of novel sequences. (**A**) Advanced sequence search of homologous proteins using the EnzymeMiner tool [32]. (**B**) Optimization of the variational autoencoder architecture to capture the sequence distribution of the MSA and phylogenetic dependencies within the latent space. (**C**) Exploration of the evolutionary dependencies between the sequences extracted from the variational autoencoder and its low-dimensional latent space. This representation is then used to guide the protein design strategy and generate sequences along the trajectory from the query to the latent space origin. The generated sequences are characterized based on their statistical and sequential properties to produce the evolutionary profile. This profile serves as a guide for selecting designs. **(D**) The experimental characterization of the proposed designs is conducted. The orange frames represent the computational steps and the blue frame is the experimental step.

3.2 Network architecture optimization

Variational autoencoders are a suitable framework for capturing sequence spaces. Variational autoencoders (VAEs) [23] are a type of deep generative learning model whose goal is to generate new data that is similar to the input. VAEs consist of two main components: an encoder and a decoder (**Fig. 1B**). The encoder takes the input data and maps it to a lower dimensional representation called a latent space. Within this latent space, the encoded input is modeled as a normal distribution by two parameters, mean and variance. Subsequently, the decoder draws samples from this latent space distribution and maps these samples back to the original input data. The training of a VAE is based on minimizing the loss function which consists of two parts. The first part, called the reconstruction term, penalizes incorrect reconstruction of the input data, thus helping the model to make the latent space rich enough for the decoder to reconstruct the input sequences. The second part, called the regularization term, serves to constrain the latent space distribution of encoded values. The application of the regularization term is made possible by the encoder's ability to encode inputs into distributions, which are enforced to be close to normal distribution by measuring the Kulback-Leibler divergence. As a result, the individual distributions are forced to overlap within the latent space, ensuring that every latent space point, once reconstructed as a sequence by the decoder, aligns with the sequences corresponding to the nearby points in the latent space.

Replicating the methodology described by Ding et al. [27] on the capacity of VAEs to delineate phylogenetic relationships among proteins, we confirmed these relationships for our HLD I-IV dataset (Model 1). We observed a star-like configuration, with multiple spikes radiating from a central point in various directions (**Fig. S3**). This pattern indicates that sequences within distinct latent space clusters tend to aggregate if they share a common lineage at a specific evolutionary juncture on the phylogenetic tree. Furthermore, our study confirmed that as sequences evolve from the root to the leaves of the phylogenetic tree, their latent space coordinates move from near the origin to the periphery (**Fig. 2E**). Leveraging these observations, we posited that the coordinates within the latent space could serve as a navigational tool in the search for ancestral-like sequences within the sequence space, potentially improving query protein stability while preserving its function.

Variational autoencoders are capable of replicating sequence statistics and capturing evolutionary trends. Before testing our hypothesis, we embarked on selecting the best model architecture based on the implementation provided by Ding et al. [27]. This involved optimizing the encoder, decoder, and training procedure to minimize the difference between the generated and input sequence distributions (generative capacity) [36] while also preserving the relationship between phylogeny and the latent space (geometric properties). In order to evaluate the model's generative capacity, we implemented several tests. The first test examined how well our model reproduces the statistics of the input dataset on the output. To this end, we compared the first and second-order statistics of 3,000 randomly sampled MSA input sequences with those generated by our VAE model (**Fig. 2A**). The first-order statistics compare the frequencies of amino acids in every position. The VAE model successfully reproduced them, also maintaining the frequencies of gaps, indicating that it did not introduce extra deletions. Although first-order statistics often provide strong signals for belonging to a given protein family, they cannot secure the generation of viable proteins, as the interaction between distant positions can be critical [49]. The second-order amino acid distribution can serve as an important metric to evaluate the similarity of distant relations in the generated sequences. We employed second-order covariances to assess the ability of the model to reproduce this distribution, as suggested by prior works [18,36,50]. The VAE model reproduced the empirical second-order statistics of natural sequences (**Fig. 2B**). We integrated query reconstruction accuracy [51] as an additional metric in our analysis to ensure that the model was capable of reconstructing the query sequence with minimal mutations. Notably, the final Model 1 demonstrated the ability to reconstruct the query with as few as 6 unknown mutations while trained on HLD I-IV dataset.

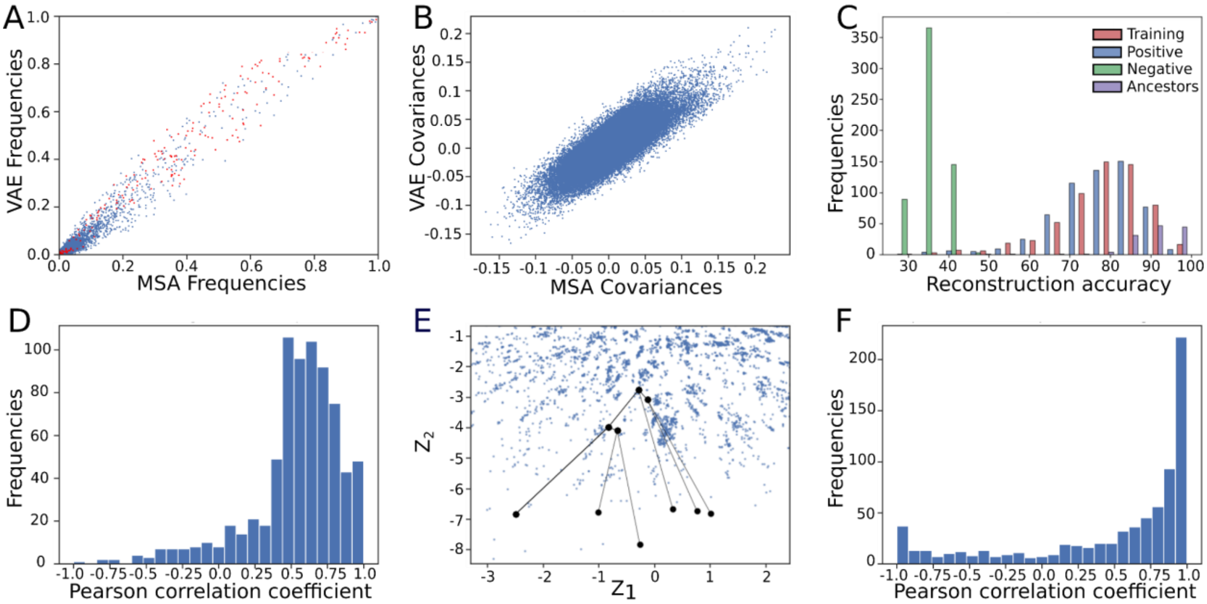


Fig. 2: **Showcases of the statistics used to measure the generative capacity of the final VAE model.** Showed for HLD I-IV (Model 1, see **Appendix section 1**) and the geometric properties of its latent space. (**A**) The first-order statistics for 3000 sequences were randomly selected for the input MSA or VAE generated. The first-order statistics (red dots) represent the gap symbol frequencies in sequence positions, while the blue points denote amino acids. (**B**) The second-order statistics (blue dots) demonstrate that our model can reconstruct pairwise amino acid occurrences fairly well (ρ=0.68). (**C**) The average reconstruction accuracy for the negative (green), training (red), positive (blue), and ancestral (violet) control datasets. The shifts in the histograms between the sets imply that the model can distinguish random sequences (negative) from those in the input MSA (training and positive) and those corresponding to the straight-line strategy of generating ancestors (ancestors). (**D**) Most sequences in tree branches have a positive correlation between depth and latent space origin distance. (**E**) Mapping a small phylogenetic tree onto the latent space. (**F**) Histogram illustrating the directional trends of phylogenetic tree branches projected onto the latent space. In this representation, a value of 1 indicates a straight trajectory towards the latent space origin, while -1 represents the opposite trend. The histogram highlights that the majority of branches tend to align towards the latent space origin.

In our second test, we examined the statistical profile of our model by measuring the model average reconstruction accuracy of sequences from various control sets. The average reconstruction accuracy of the sequence was approximated as an average reconstructed sequence identity for 5,000 samples around the original sequence coordinates of its latent space embedding. For the negative control, we generated a random set of sequences with matching frequencies of amino acids in every position to those in the input MSA. The training subset consisted of the same number of sequences sampled randomly from the training dataset. For the positive control, we evaluated the average reconstruction accuracy of 5% MSA sequences removed from the input before training. We ended up with 612 sequences for the negative control, positive control, and training subsets. The test showed that our model can distinguish random sequences from those in the MSA with the cutoff value of the reconstructed accuracy of 50% (**Fig. 2C**).

As far as the geometric properties of the latent space are concerned, we aimed to preserve the fact that the latent space carries evolutionary information. Thus, while optimizing model properties, we kept track of the relationship between phylogeny and the geometry of latent space (**Fig. 2D-F**). For that purpose, we prepared phylogenetic trees with inferred ancestral sequences. Every tree branch from a leaf to the root was mapped into latent space. We further quantified the relationship between the latent space distances to the origin and the corresponding positions in the phylogenetic tree (**Fig. 2D**). To gain insight into individual tree branches, we also calculated the angle between two vectors: the one going from a leaf node to the origin in the latent space, and the other one defined by the first principal component of the latent coordinates of all the nodes on the corresponding phylogenetic branch (**Fig. 2F**). Our findings indicate that small dense architectures of encoder and decoder are effective in capturing evolutionary dependencies within the latent space structure, whereas deeper architectures disrupt these evolutionary patterns (**Fig. S4**). The final width of the dense layer was therefore set to the length of the protein sequence and the latent space dimensionality of 2 was chosen. This statistical analysis was repeated for Models 2-4 in the third round of experiments (**Appendix section 1**), reconfirming our finding and showing improved secondary statistics with a correlation of 0.8. Consequently, we kept the rules behind the hidden layer width and latent space dimensionality.

3.3 Construction of the evolutionary trajectory

The latent space is able to capture protein stability in its structure. The ancestral sequences are often associated with enhanced stability compared to their extant counterparts [16,52]. We hypothesized that the structure of the VAE latent space might encode stability and predict that the more stable variants of our target protein, DhaA, would be located closer to the origin compared to the wild type. To test this hypothesis, we mapped two sets of experimentally measured stability values to the latent space of Model 1. The first set consisted of six ancestral sequences from the previous protein engineering campaign of the thoroughly characterized dehalogenases: DbjA, DbeA, DhaA, DmxA, and DmmA [33]. The second set consisted of 24 previously engineered DhaA variants incorporating both evolutionary and energy-based mutations based on the FireProt method [15,34,35]. The sequences were aligned to those in the training MSA and mapped to the latent space. Their latent space coordinates were closer to the origin of the latent space (**Fig. 3A, B**), supporting the notion that the latent space captures the information about the stability landscape. The observations were also recapitulated for HLD I-II dataset and Model 2.

Specification of a latent space strategy guides the search and selection of sequences. The regular distribution of stable sequences in the latent space (**Fig. 3B**) led us to develop the *straight-line evolutionary strategy*. This strategy encodes the query sequence (DhaA in our case) into its latent representation and then follows the straight line connecting that point to the origin of the latent space (**Fig. 3A**), mimicking the mapping of ancestral dependencies into the latent space. The line is divided into equal intervals, whose boundaries are then selected for reconstruction by the decoder. In the experiments, we selected 100 intervals. To represent the designed sequences, we analyzed several statistical parameters for the individual designed sequences: the average reconstruction match, similarity to the query sequence, similarity to the closest sequence in the latent space from the training set, and the number of insertions/deletions compared to the original sequence. The values obtained were plotted and visually inspected to identify variants with interesting statistical values. The generated profiles were then used to select suitable variants for subsequent experimental characterisation (**Fig. 3C**).

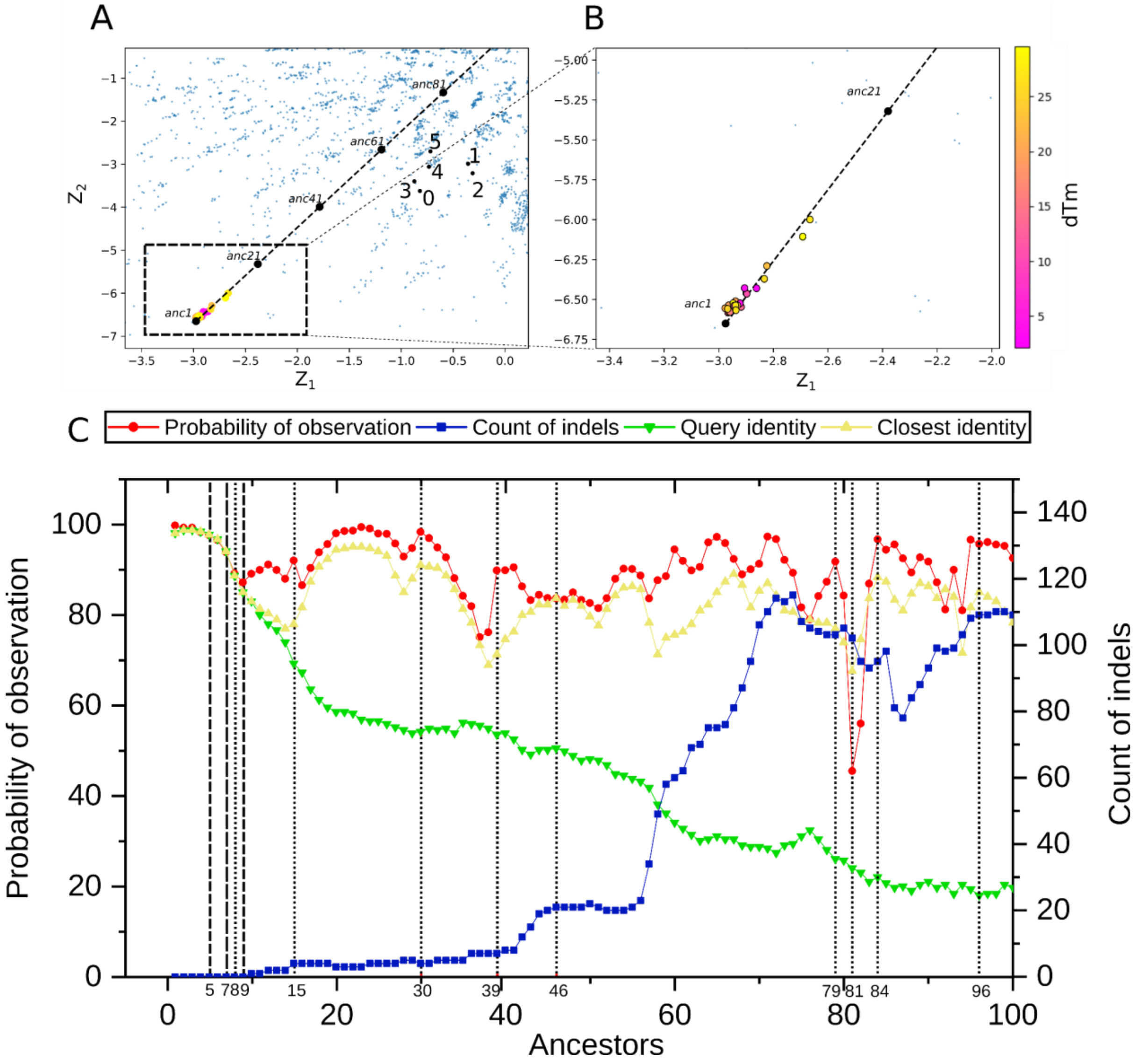


Fig. 3: The straight-line evolutionary strategy for Model 1. (**A**) Straight evolutionary strategy reconstructed 100 sequences along the trajectory from query embedding to the latent space origin (black dashed line). The embeddings of experimentally validated Babkova's ancestors [33] (points 0-5) and engineered DhaA variants [35] (yellow to pink spectrum points) are mapped closer to the latent space origin, supporting the idea behind our ancestral generation strategy. (**B**) A detailed view of DhaA variants engineered to improve their stability almost perfectly aligned along the straight-line trajectory to the origin of the latent space. The degree of stability improvement is depicted by the color bar on the right. While there is no strong correlation between the positions in the latent space and the stability gain of variants up to 28°C, some of the most stable points are situated closer to the origin. (**C**) The statistical profile of 100 sequences from the straight evolutionary strategy. The vertical lines represent sequences selected for experimental characterization for the first and second rounds (**Table S1**) where dashed line variants were successfully expressed.

Using the statistical profile, we identified 9 promising designs in the first round for further laboratory experiments to gain deeper insight into the statistical indicators. These designs exhibited a wide range of sequence variability, ranging from 45 substitutions and no insertions or deletions (indels) to 138 substitutions and 109 indels (**Table S1**) (AncDhaA1-9). The substitutions and indels, with deletions in most cases, covered the entire protein structure. In the second round, we focused on the more conserved variants (ancestors 5, 7, and 8; AncDhaA10-12 for reference in experiments) with 7 to 34 substitutions without indels. Altogether, 12 designs were selected for laboratory expression and biophysical characterization from Model 1.

3.4 Third round of variant selection provided soluble proteins

Solubilization of designs by introducing new knowledge to the dataset. As we observed low solubility in the first round (see **Section 3.5 Experimental characterization of expressed variants** and **Fig. S7A-B**), we incorporated previous findings on the low solubility of HLD subfamilies III and IV [45] in our workflow. To this end, we embarked on a third round of experiments restricted to the HLD subfamilies I and II. This round focused on training additional models (Models 2-3) and designed eight DhaA variants (AncDhaA13-20, **Table S1**). The first candidate VAE (Model 2) achieved up to 97% similarity in query sequence reconstruction with satisfactory second-order statistics. Considering the often-disruptive impact of indels, we selected and curated five designs from the straight-line evolutionary strategy accumulating at most one indel (AncDhaA13-17). Another VAE trained from a different initialization (Model 3) showed similar trends but additionally had a curious pattern for the origin of the latent space: the model demonstrated a significant shift in sequence similarity towards DbjA. Therefore, we selected ancestor 99 (AncDhaA20) from this model for further experimental characterization to explore its unique shared sequence similarity to both DhaA (52%) and DbjA (93%).

Attempt to further solubilize designs by implementing conditional variational autoencoder. Finally, in the third round, we also decided to explore one more solution to low solubility, conditional variational autoencoder (CVAE). To this end we added solubility scores from SoluProt [38] to the training, discretized into three bins for low, moderate, and high solubility values (Model 4). We conditioned the sampling process from Model 4 using a straight-line evolutionary strategy on the highest bin label forcing CVAE to introduce patterns from sequences with predicted high solubility in decoded designs. We took two variants from Model 4: ancestor 0 (AncDhaA18) with 30 mutations as the control of the pattern extracted from highly soluble sequences and ancestor 18 (AncDhaA19), which had 49 mutations and one deletion in the coil region at position 32 (**Fig. S5B**), as it exhibited high confidence in the model (93.28%) and an increased number of high probable residues (85 positions scored above 90%).

Refining the mutations based on manual curation of the structures and stability scores. From Model 2, we selected three ancestors with unique features: ancestor 3 closely mimicked the wild type with 98% similarity; ancestor 16 notably introduced a proline at position 75; and ancestor 23 was the last variant starting with a regular sequence pattern (MSEIGT), suggesting high solubility and expression potential based on its 88% similarity to the PDB sequence 4WCV. To increase our chances of producing soluble variants, we manually curated proposed mutations based on the structural predictions from AlphaFold [11] and stability assessments using the MutCompute tool [42] (**Table S2**). Mutation manual curation led to the classification of VAE-proposed mutations into safe and risky categories, respectively (**Section 2 Manual curation of designs in round 3 of SI**). Starting from ancestor 3, we kept four non-risky mutations. We removed two structurally and statistically risky mutations, P34V and L238F, from the original six-point variant proposed by the VAE, producing the design AncDhaA14. As experimental validation of identified risky mutations, we selected ancestor 0 with eight substitutions (AncDhaA13). For the second manually curated variant (AncDhaA16), we selected ancestor 15 as the template, which included as the last VAE ancestor no insertions, and we removed the risky mutations P34V and L238F, leaving nine mutations compared to the DhaA wild type. To experimentally determine the effect of suggested proline insertion with risky mutations, we selected ancestor 16 as predicted by VAE (AncDhaA15) (**Fig. S2**). As a template for the third manual curation target, we selected ancestor 23. We incorporated all mutations suggested by the VAE except the risky ones (P34V, L238F) and the proline insertions, resulting in a final design carrying 32 mutations (AncDhaA17).

3.5 Experimental characterization of expressed variants

Variant production. Protein overexpression was carried out in three rounds (**Table S1**) to assess the solubility and whole-cell activity of DhaA variants. In the first round, the solubility of DhaA ancestor variants was analyzed after their overproduction in *E. coli* and purification, where only AncDhaA1 was purified as a soluble protein in sufficient yield (**Fig. S7**). This variant was also the only active one within the screening of the dehalogenase activity in whole cells using a halide oxidation (HOX) assay (**Fig. S8A**) [53]. The variants produced in the first round showed overall low solubility, consistent with outcomes from several previous machine learning-based pipelines for protein design, achieving a solubility success rate of around 20% [54,55]. In the second round, the ancestors 5, 7, and 8 (AncDhaA10-12) were selected for production to stay closer to the template. Among these variants, AncDhaA10 showed the highest expression, followed by AncDhaA11, while AncDhaA12 exhibited the lowest expression and solubility (**Fig. S7B**). AncDhaA10 was also shown to be active by HOX assay (**Fig. S8B**). Based on the results from both rounds, AncDhaA1, AncDhaA10, and AncDhaA11 were chosen for further biochemical characterization. From the third round of designs, six variants showed average or high solubility and only AncDhaA17 and AncDhaA19 were poorly soluble in the tested buffer. The HOX assay revealed two highly active variants AncDhaA13 and AncDhaA14 with activities comparable to templates and positive controls (**Fig. S8B**). Variants AncDhaA16 and AncDhaA20 exhibited lower activity, comparable to that of AncHLD-RLuc [56] (**Fig. S8B**), while AncDhaA15 and AncDhaA18 showed low activities (**Fig. S8B**).

Secondary structure experimental validation. To confirm the proper folding of the studied variants, circular dichroism (CD) spectra were collected for all soluble variants: AncDhaA1, AncDhaA10-11, AncDhaA13-16, AncDhaA18, and AncDhaA20 (**Fig. 4A**). Overall, CD spectra of most variants highly resembled those of the templates (typical α/β-hydrolase fold), confirming proper folding. On the contrary, the spectra of AncDhaA1, AncDhaA11, AncDhaA15, and AncDhaA18 (**Fig. S9**) deviated from the templates. To further understand the secondary structure of the variants, BeStSel server [44] was used for fitting experimental data and analysis of PDB structures, and PDBMD2CD [43] was used for predicting CD spectra from experimental structures of templates and AlphaFold models of selected variants. **Fig. S9B** shows that the prediction of CD spectra based on AlphaFold structures did not match the experiments in all the variants. This highlights a limitation in AlphaFold's ability to accurately predict changes in folding and emphasizes the need for further improvements in computational methods. Experimental validation remains essential to address this challenge.

Thermostability. Thermodynamic stability of all variants was assessed by nano-differential scanning fluorimetry (**Fig. 2B, Table S3**). The apparent melting temperatures for the variants were in the range of 45°C–60°C. AncDhaA13, AncDhaA14, AncDhaA16 and AncDhaA20 surpassed Template_2 in terms of apparent melting temperature. The highest ΔT_m of 9°C was measured for AncDhaA20. Protein aggregation was observed for the variants AncDhaA1 and AncDhaA20, showing the onset at 45.5°C and 48.5°C, respectively (**Fig. S10**).

Temperature profiles. We then proceeded to measure temperature profiles (**Fig. 4C**). Most variants showed the T_{max} (temperature at which maximum activity was detected) of 40°C, which is in agreement with previously determined temperature profiles for DhaA [47]. Notably, AncDhaA13 showed T_{max} of 50°C, which aligns with its increased thermostability. The temperature profiles for AncDhaA11 and AncDhaA16 were not obtained, as the activities were below the detection limit. Due to compromised activity and folding, both variants were excluded from the subsequent substrate specificity profiling.

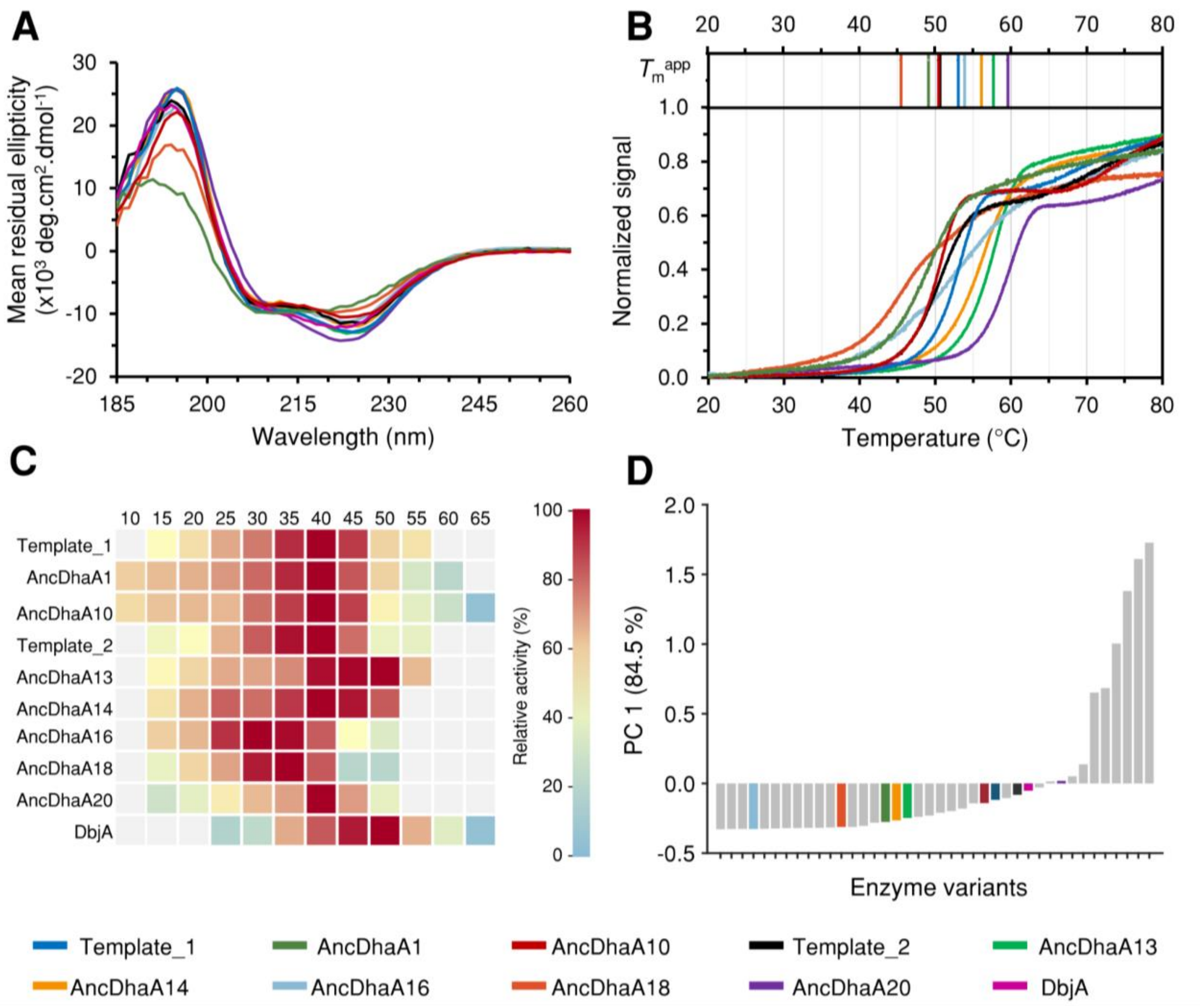


Fig. 4 Experimental characterization of selected variants. (**A**) Far-UV circular dichroism spectra probing the correct folding and secondary structure of the variants. (**B**) Normalized thermal denaturation curves from nanoDSF spectroscopy with apparent melting temperatures (T_m^{app}) are shown above the curves. (**C**) The dependence of specific activity on temperature. The heatmap represents the relative activity of individual variants. (**D**) The score plot shows the first principal component PC 1 explaining 84.9 % of the data variance, which compares VAE designs (in color) with previously characterized haloalkane dehalogenases (gray) in terms of their activity with 27 substrates [45]. The color code for individual variants is shown at the bottom.

Substrate specificity. The temperature of 35°C was selected for the subsequent specificity characterization for being below the onset of denaturation (**Table S3**) for most of the remaining variants and close to their T_{max} values. To explore the obtained substrate specificities in the context of the haloalkane dehalogenase family, the principal component analysis (PCA) has been conducted by augmenting the previously used dataset comprising substrate specificities for 32 wild-type dehalogenases [45] with the newly obtained data. The PCA of raw data showed that the highest overall activity (**Fig. 4D**) was observed for AncDhaA20, surpassing both the Template_2 and DbjA. The overall activity was retained for AncDhaA10, while for the other variants, it was substantially decreased (**Fig. S11**). The analysis of log-transformed activity data (**Fig. S12**) further showed that AncDhaA1, AncDhaA10, AncDhaA13, and AncDhaA14 differed only very slightly from templates in their substrate preferences. The profiles of AncDhaA18 and AncDhaA20 resembled more closely the specificity profile of DbjA, which is not unexpected in the case of AncDhaA20 due to its high sequence similarity. AncDhaA16 differed significantly from all other variants due to the low number of converted substrates

(10 out of 27), which is a unique property for haloalkane dehalogenases known for their broad substrate specificity.

4. Discussion

In this study, we utilized variational autoencoders (VAEs) to map functionally related haloalkane dehalogenase sequences from EnzymeMiner onto VAE latent spaces, following an approach inspired by Ding et al. [27]. This process revealed that the latent space could capture the phylogenetic relationships of the sequences. Motivated by these insights, we employed the VAE framework for reconstructing sequences, aiming to create ancestral-like variants of the haloalkane dehalogenase DhaA. We tuned the network hyperparameters to accurately reflect the statistical frequencies of the input while maintaining the relationship between evolutionary trajectories and latent spaces. We discovered that a simple feed-forward neural network with a single dense layer matched to the input MSA columns and two-dimensional latent space [28] was enough for this task.

We then introduced a simple strategy to generate novel sequences based on the geometry of the latent space. We achieved this by reconstructing the embeddings along the trajectory towards the origin of the latent space. Employing this strategy, we systematically executed the pipeline in three rounds of laboratory experiments and optimization, leading to four VAE models producing 20 designs in total. Similarly to the study of computational filter evaluation for synthetical protein designs from generative models [57], each iteration revealed new insights, allowing for iterative refinement and improvement of our approach. Notably, we increased the success rate of soluble designs from 11% in the first round and 66% in the second round to 75% in the third round, illustrating the effectiveness of applying accumulated knowledge for improvement (**Section 1 Workflow overview of SI**).

In the first two rounds, we were facing solubility issues with our designs since only three designs (AncDhaA1, AncDhaA10, and AncDhaA11) showed expression levels sufficient for in-depth characterization using *in-house* microfluidic devices [58]: (i) secondary structure, (ii) thermostability, (iii) temperature profiles and (iv) substrate specificity. Surprisingly, we observed significant discrepancies between the CD spectra predicted based on the AlphaFold structures and collected experimental data. In particular, both the AlphaFold structures and derived spectra resembled those of the native proteins, contrary to clear signs of misfolding by experiments. This implies a bias in AlphaFold predictions towards native folds and stresses the need for caution when using these predictions for synthetic sequences generated by protein language models [59]. Thermostability analysis showed that the variants neither improved nor compromised thermostability. AncDhaA10 exhibited above-average activity compared to other haloalkane dehalogenases.

In the third round, we addressed the solubility issue by narrowing down the input MSA to HLD subfamilies I-II and implementing stricter preprocessing steps to suppress the frequencies of indels. Moreover, we manually curated the sequences based on AlphaFold [11] structural and MutCompute [42] stability assessments. These steps were successful as the majority of the new designs were soluble. Interestingly, the non-curated AncDhaA13 showed good solubility and catalytic activity, despite our expected risky mutations P34V and L238F. Current computational tools are not particularly strong in predicting the epistasis effects. The curation step rescued poorly soluble AncDhaA17, which was crucial in designing AncDhaA16 and predicting the disruptive effect of the proline insertion on AncDhaA15 activity.

A different initialization of a model training and stricter threshold of column removal with query amino acid positions in the third round led to a second VAE model, which generated sequences with an implicit His-Tag and high similarity to proteins with known experimental structures. Investigating the sequence reconstructed from the origin of the latent space (AncDhaA20) unveiled a notable shift in similarity towards another wild-type dehalogenase, DbjA [60], altered substrate specificity, increased thermostability (60°C) and improved activity (3.5-fold), being a top-performing haloalkane dehalogenase (**Fig. 4D**). To better understand the sensitivity of the *straight-line evolutionary strategy* to different initializations, we examined the embeddings over an ensemble of four randomly initialized VAEs (**Section 3 Model ensemble trajectories of SI**, **Fig. S6**). Although we observed a general trend that the straight-line evolutionary trajectories converged towards the origin of the latent spaces of different VAEs, it was also evident that the trajectories exhibited quite wide scatter. This suggests that ensemble learning [61] might be an interesting direction for follow-up research to improve the robustness of our strategy.

Another promising directions for further improvement include developing better scoring methods for the sequences generated by protein language models, particularly those that will allow filtering out misfolded or poorly soluble designs *in silico*, and adopting the recent developments in transformer-based architectures, which have demonstrated a better capacity for learning from amino acid sequences [62,63]. Integrating transformer-based architectures with manifold learning can further enhance their ability to generate sequences of stable and soluble proteins [28,29]. To bolster the robustness of future studies, adopting a generation protocol for ancestral sequences that incorporates an ensemble of models might also be advantageous [64]. This approach addresses the observed instability of ancestral trajectories within the latent space and could establish a more reliable foundation for future investigations.

5. Conclusions

Our study demonstrated that the structure of the latent space and the generative potential of VAEs are capable of guiding the sequence search and designing novel soluble and functional proteins with enhanced stability. The workflow underwent systematic improvements through three consecutive design-build-test phases, with each iteration informed by the findings from the previous one. Detailed experimental characterization of designed variants by in-house microfluidics was instrumental for iterative improvement of our computational workflow and optimization of selection strategy. The success rate of soluble designs increased from 11% in the first round to 66% in the second round and 75% in the third round. Such an improvement of the protocol would not be achievable without gaining and implementing new knowledge based on the data collected using the wet lab experiments. Through this process, complemented by manual curation of specific variants, we achieve a notable increase in stability—up to 9°C for the top-performing AncDhaA20 variant, with an average improvement of 3°C across all soluble variants and a significant boost in activity up to 3.5-fold. The frequency and location of indels were the most critical parameters to consider during the design selection for experimental characterization. Using our approach, we recommend selecting designs with a low number of indels or with high protein similarity to natural sequences, preferably to those in PDB. A current limitation of our study is that it was conducted using a single enzyme family. Therefore, validation of designs for other protein families will help understand the generalizability of the developed approach. Overall, our study demonstrates that VAE represents a promising strategy for generating novel soluble, stable, and functional enzymes.

Acknowledgements

Brno Ph.D. Talent Scholarship holders Pavel Kohout and Michal Vašina acknowledge funding from the Brno City Municipality. Computational resources were provided by the e-INFRA CZ and ELIXIR-CZ projects (LM2023055 and LM2018140), supported by the Ministry of Education, Youth and Sports of the Czech Republic. Core Facility Biomolecular Interactions and Crystallography of CEITEC Masaryk University, Brno is gratefully acknowledged for the obtaining of the differential scanning fluorimetry data presented in this paper. This work was also supported by Czech Ministry of Education, Youth and Sports [ESFRI RECETOX RI LM2023069]; Technology Agency of the Czech Republic [TN02000122]; the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No 857560 (CETOCOEN Excellence); by the project National Institute for Cancer Research (No. LX22NPO5102) – Funded by the European Union – Next Generation EU, and by COST (European Cooperation in Science and Technology) [COST Action COZYME CA21162]. Martin Marek thanks the Czech Science Foundation (grant no. GA22-09853S). This publication reflects only the author's view, and the European Commission is not responsible for any use that may be made of the information it contains.

References

[1] S. Wu, R. Snajdrova, J.C. Moore, K. Baldenius, U.T. Bornscheuer, Biocatalysis: Enzymatic Synthesis for Industrial Applications, Angew. Chem. Int. Ed. 60 (2021) 88–119. https://doi.org/10.1002/anie.202006648.

[2] E.L. Bell, W. Finnigan, S.P. France, A.P. Green, M.A. Hayes, L.J. Hepworth, S.L. Lovelock, H. Niikura, S. Osuna, E. Romero, K.S. Ryan, N.J. Turner, S.L. Flitsch, Biocatalysis, Nat. Rev. Methods Primer 1 (2021) 1–21. https://doi.org/10.1038/s43586-021-00044-z.

[3] B.S. Silvestre, D.M. Ţîrcă, Innovations for sustainable development: Moving toward a sustainable future, J. Clean. Prod. 208 (2019) 325–332. https://doi.org/10.1016/j.jclepro.2018.09.244.

[4] T. Tiso, B. Winter, R. Wei, J. Hee, J. de Witt, N. Wierckx, P. Quicker, U.T. Bornscheuer, A. Bardow, J. Nogales, L.M. Blank, The metabolic potential of plastics as biotechnological carbon sources – Review and targets for the future, Metab. Eng. 71 (2022) 77–98. https://doi.org/10.1016/j.ymben.2021.12.006.

[5] J. Planas-Iglesias, S.M. Marques, G.P. Pinto, M. Musil, J. Stourac, J. Damborsky, D. Bednar, Computational design of enzymes for biotechnological applications, Biotechnol. Adv. 47 (2021) 107696. https://doi.org/10.1016/j.biotechadv.2021.107696.

[6] S.M. Marques, J. Planas-Iglesias, J. Damborsky, Web-based tools for computational enzyme design, Curr. Opin. Struct. Biol. 69 (2021) 19–34. https://doi.org/10.1016/j.sbi.2021.01.010.

[7] D. Baker, What has de novo protein design taught us about protein folding and biophysics?, Protein Sci. 28 (2019) 678–683. https://doi.org/10.1002/pro.3588.

[8] D.M. Taverna, R.A. Goldstein, Why are proteins marginally stable?, Proteins Struct. Funct. Bioinforma. 46 (2002) 105–109. https://doi.org/10.1002/prot.10016.

[9] K.K. Yang, Z. Wu, F.H. Arnold, Machine-learning-guided directed evolution for protein engineering, Nat. Methods 16 (2019) 687–694. https://doi.org/10.1038/s41592-019-0496-6.

[10] Z. Wu, S.B.J. Kan, R.D. Lewis, B.J. Wittmann, F.H. Arnold, Machine learning-assisted directed protein evolution with combinatorial libraries, Proc. Natl. Acad. Sci. 116 (2019) 8852–8858. https://doi.org/10.1073/pnas.1901979116.

[11] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S.A.A. Kohl, A.J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A.W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, Highly accurate protein structure prediction with AlphaFold, Nature 596 (2021) 583–589. https://doi.org/10.1038/s41586-021-03819-2.

[12] B.J. Jones, C.N.E. Kan, C. Luo, R.J. Kazlauskas, Chapter Six - Consensus Finder web tool to predict stabilizing substitutions in proteins, in: D.S. Tawfik (Ed.), Methods Enzymol., Academic Press, 2020: pp. 129–148. https://doi.org/10.1016/bs.mie.2020.07.010.

[13] L. Sumbalova, J. Stourac, T. Martinek, D. Bednar, J. Damborsky, HotSpot Wizard 3.0: web server for automated design of mutations and smart libraries based on sequence input information, Nucleic Acids Res. 46 (2018) W356–W362. https://doi.org/10.1093/nar/gky417.

[14] R. Furukawa, W. Toma, K. Yamazaki, S. Akanuma, Ancestral sequence reconstruction produces thermally stable enzymes with mesophilic enzyme-like catalytic properties, Sci. Rep. 10 (2020) 15493. https://doi.org/10.1038/s41598-020-72418-4.

[15] M. Musil, R.T. Khan, A. Beier, J. Stourac, H. Konegger, J. Damborsky, D. Bednar, FireProtASR: A Web Server for Fully Automated Ancestral Sequence Reconstruction, Brief. Bioinform. 22 (2021) bbaa337. https://doi.org/10.1093/bib/bbaa337.

[16] J. Livada, A.M. Vargas, C.A. Martinez, R.D. Lewis, Ancestral Sequence Reconstruction Enhances Gene Mining Efforts for Industrial Ene Reductases by Expanding Enzyme Panels with Thermostable Catalysts, ACS Catal. 13 (2023) 2576–2585. https://doi.org/10.1021/acscatal.2c03859.

[17] M. Lehmann, L. Pasamontes, S.F. Lassen, M. Wyss, The consensus concept for thermostability engineering of proteins, Biochim. Biophys. Acta BBA - Protein Struct. Mol. Enzymol. 1543 (2000) 408–415. https://doi.org/10.1016/S0167-4838(00)00238-7.

[18] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D.S. Marks, C. Sander, R. Zecchina, J.N. Onuchic, T. Hwa, M. Weigt, Direct-coupling analysis of residue coevolution captures native contacts across many protein families, Proc. Natl. Acad. Sci. 108 (2011) E1293–E1301. https://doi.org/10.1073/pnas.1111471108.

[19] E.C. Alley, G. Khimulya, S. Biswas, M. AlQuraishi, G.M. Church, Unified rational protein engineering with sequence-based deep representation learning, Nat. Methods 16 (2019) 1315–1322. https://doi.org/10.1038/s41592-019-0598-1.

[20] A. Hawkins-Hooker, F. Depardieu, S. Baur, G. Couairon, A. Chen, D. Bikard, Generating functional protein variants with variational autoencoders, PLOS Comput. Biol. 17 (2021) e1008736. https://doi.org/10.1371/journal.pcbi.1008736.

[21] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, A. dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, A. Rives, Evolutionary-scale prediction of atomic-level protein structure with a language model, Science 379 (2023) 1123–1130. https://doi.org/10.1126/science.ade2574.

[22] A. Elnaggar, H. Essam, W. Salah-Eldin, W. Moustafa, M. Elkerdawy, C. Rochereau, B. Rost, Ankh: Optimized Protein Language Model Unlocks General-Purpose Modelling, (2023). https://doi.org/10.48550/arXiv.2301.06568.

[23] D.P. Kingma, M. Welling, Auto-Encoding Variational Bayes, (2022). https://doi.org/10.48550/arXiv.1312.6114.

[24] R.R. Eguchi, C.A. Choe, P.-S. Huang, Ig-VAE: Generative modeling of protein structure by direct 3D coordinate generation, PLOS Comput. Biol. 18 (2022) e1010271. https://doi.org/10.1371/journal.pcbi.1010271.

[25] R. Gómez-Bombarelli, J.N. Wei, D. Duvenaud, J.M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T.D. Hirzel, R.P. Adams, A. Aspuru-Guzik, Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules, ACS Cent. Sci. 4 (2018) 268–276. https://doi.org/10.1021/acscentsci.7b00572.

[26] X. Lian, N. Praljak, S.K. Subramanian, S. Wasinger, R. Ranganathan, A.L. Ferguson, Deep learning-enabled design of synthetic orthologs of a signaling protein, (2022) 2022.12.21.521443. https://doi.org/10.1101/2022.12.21.521443.

[27] X. Ding, Z. Zou, C.L. Brooks III, Deciphering protein evolution and fitness landscapes with latent space models, Nat. Commun. 10 (2019) 5644. https://doi.org/10.1038/s41467-019-13633-0.

[28] C. Ziegler, J. Martin, C. Sinner, F. Morcos, Latent generative landscapes as maps of functional diversity in protein sequence space, Nat. Commun. 14 (2023) 2222. https://doi.org/10.1038/s41467-023-37958-z.

[29] N.S. Detlefsen, S. Hauberg, W. Boomsma, Learning meaningful representations of protein sequences, Nat. Commun. 13 (2022) 1914. https://doi.org/10.1038/s41467-022-29443-w.

[30] D.B. Janssen, Evolving haloalkane dehalogenases, Curr. Opin. Chem. Biol. 8 (2004) 150–159. https://doi.org/10.1016/j.cbpa.2004.02.012.

[31] T. Koudelakova, S. Bidmanova, P. Dvorak, A. Pavelka, R. Chaloupkova, Z. Prokop, J. Damborsky, Haloalkane dehalogenases: Biotechnological applications, Biotechnol. J. 8 (2013) 32–45. https://doi.org/10.1002/biot.201100486.

[32] J. Hon, S. Borko, J. Stourac, Z. Prokop, J. Zendulka, D. Bednar, T. Martinek, J. Damborsky, EnzymeMiner: automated mining of soluble enzymes with diverse structures, catalytic properties and stabilities, Nucleic Acids Res. 48 (2020) W104–W109. https://doi.org/10.1093/nar/gkaa372.

[33] P. Babkova, E. Sebestova, J. Brezovsky, R. Chaloupkova, J. Damborsky, Ancestral Haloalkane Dehalogenases Show Robustness and Unique Substrate Specificity, ChemBioChem 18 (2017) 1448–1456. https://doi.org/10.1002/cbic.201700197.

[34] A. Kunka, S.M. Marques, M. Havlasek, M. Vasina, N. Velatova, L. Cengelova, D. Kovar, J. Damborsky, M. Marek, D. Bednar, Z. Prokop, Advancing Enzyme's Stability and Catalytic Efficiency through Synergy of Force-Field Calculations, Evolutionary Analysis, and Machine Learning, ACS Catal. 13 (2023) 12506–12518. https://doi.org/10.1021/acscatal.3c02575.

[35] K. Beerens, S. Mazurenko, A. Kunka, S.M. Marques, N. Hansen, M. Musil, R. Chaloupkova, J. Waterman, J. Brezovsky, D. Bednar, Z. Prokop, J. Damborsky, Evolutionary Analysis As a Powerful Complement to Energy Calculations for Protein Stabilization, ACS Catal. 8 (2018) 9420–9428. https://doi.org/10.1021/acscatal.8b01677.

[36] F. McGee, S. Hauri, Q. Novinger, S. Vucetic, R.M. Levy, V. Carnevale, A. Haldane, The generative capacity of probabilistic protein sequence models, Nat. Commun. 12 (2021) 6302. https://doi.org/10.1038/s41467-021-26529-9.

[37] K. Sohn, H. Lee, X. Yan, Learning Structured Output Representation using Deep Conditional Generative Models, in: Adv. Neural Inf. Process. Syst., Curran Associates, Inc., 2015. https://papers.nips.cc/paper_files/paper/2015/hash/8d55a249e6baa5c06772297520da2051-Abstract.html (accessed April 9, 2024).

[38] J. Hon, M. Marusiak, T. Martinek, A. Kunka, J. Zendulka, D. Bednar, J. Damborsky, SoluProt: prediction of soluble protein expression in Escherichia coli, Bioinformatics 37 (2021) 23–28. https://doi.org/10.1093/bioinformatics/btaa1102.

[39] Y. Yao, X. Wangr, Y. Ma, H. Fang, J. Wei, L. Chen, A. Anaissi, A. Braytee, Conditional Variational Autoencoder with Balanced Pre-training for Generative Adversarial Networks, (2022). https://doi.org/10.48550/arXiv.2201.04809.

[40] M. Mirdita, K. Schütze, Y. Moriwaki, L. Heo, S. Ovchinnikov, M. Steinegger, ColabFold: making protein folding accessible to all, Nat. Methods 19 (2022) 679–682. https://doi.org/10.1038/s41592-022-01488-1.

[41] The PyMOL Molecular Graphics System, Version 1.2r3pre, Schrödinger, LLC, (n.d.).

[42] R. Shroff, A.W. Cole, D.J. Diaz, B.R. Morrow, I. Donnell, A. Annapareddy, J. Gollihar, A.D. Ellington, R. Thyer, Discovery of Novel Gain-of-Function Mutations Guided by Structure-Based Deep Learning, ACS Synth. Biol. 9 (2020) 2927–2935. https://doi.org/10.1021/acssynbio.0c00345.

[43] E.D. Drew, R.W. Janes, PDBMD2CD: providing predicted protein circular dichroism spectra from multiple molecular dynamics-generated protein structures, Nucleic Acids Res. 48 (2020) W17–W24. https://doi.org/10.1093/nar/gkaa296.

[44] A. Micsonai, É. Moussong, F. Wien, E. Boros, H. Vadászi, N. Murvai, Y.-H. Lee, T. Molnár, M. Réfrégiers, Y. Goto, Á. Tantos, J. Kardos, BeStSel: webserver for secondary structure and fold prediction for protein CD spectroscopy, Nucleic Acids Res. 50 (2022) W90–W98. https://doi.org/10.1093/nar/gkac345.

[45] M. Vasina, P. Vanacek, J. Hon, D. Kovar, H. Faldynova, A. Kunka, T. Buryska, C.P.S. Badenhorst, S. Mazurenko, D. Bednar, S. Stavrakis, U.T. Bornscheuer, A. deMello, J. Damborsky, Z. Prokop, Advanced database mining of efficient haloalkane dehalogenases by sequence and structure bioinformatics and microfluidics, Chem Catal. 2 (2022) 2704–2725. https://doi.org/10.1016/j.checat.2022.09.011.

[46] M. Vasina, P. Vanacek, J. Damborsky, Z. Prokop, Chapter Three - Exploration of enzyme diversity: High-throughput techniques for protein production and microscale biochemical characterization, in: D.S. Tawfik (Ed.), Methods Enzymol., Academic Press, 2020: pp. 51–85. https://doi.org/10.1016/bs.mie.2020.05.004.

[47] T. Buryska, M. Vasina, F. Gielen, P. Vanacek, L. van Vliet, J. Jezek, Z. Pilat, P. Zemanek, J. Damborsky, F. Hollfelder, Z. Prokop, Controlled Oil/Water Partitioning of Hydrophobic Substrates Extending the Bioanalytical Applications of Droplet-Based Microfluidics, Anal. Chem. 91 (2019) 10008–10015. https://doi.org/10.1021/acs.analchem.9b01839.

[48] K.M. Wong, M.A. Suchard, J.P. Huelsenbeck, Alignment Uncertainty and Genomic Analysis, Science 319 (2008) 473–476. https://doi.org/10.1126/science.1151532.

[49] W.P. Russ, M. Figliuzzi, C. Stocker, P. Barrat-Charlaix, M. Socolich, P. Kast, D. Hilvert, R. Monasson, S. Cocco, M. Weigt, R. Ranganathan, An evolution-based model for designing chorismate mutase enzymes, Science 369 (2020) 440–445. https://doi.org/10.1126/science.aba3304.

[50] S.R. Johnson, S. Monaco, K. Massie, Z. Syed, Generating novel protein sequences using Gibbs sampling of masked language models, (2021) 2021.01.26.428322. https://doi.org/10.1101/2021.01.26.428322.

[51] Z. Costello, H.G. Martin, How to Hallucinate Functional Proteins, (2019). https://doi.org/10.48550/arXiv.1903.00458.

[52] M.A. Spence, J.A. Kaczmarski, J.W. Saunders, C.J. Jackson, Ancestral sequence reconstruction for protein engineers, Curr. Opin. Struct. Biol. 69 (2021) 131–141. https://doi.org/10.1016/j.sbi.2021.04.001.

[53] A.S. Aslan-Üzel, A. Beier, D. Kovář, C. Cziegler, S.K. Padhi, E.D. Schuiten, M. Dörr, D. Böttcher, F. Hollmann, F. Rudroff, M.D. Mihovilovic, T. Buryška, J. Damborský, Z. Prokop, C.P.S. Badenhorst, U.T. Bornscheuer, An Ultrasensitive Fluorescence Assay for the Detection of Halides and Enzymatic Dehalogenation, ChemCatChem 12 (2020) 2032–2039. https://doi.org/10.1002/cctc.201901891.

[54] D. Repecka, V. Jauniskis, L. Karpus, E. Rembeza, I. Rokaitis, J. Zrimec, S. Poviloniene, A. Laurynenas, S. Viknander, W. Abuajwa, O. Savolainen, R. Meskys, M.K.M. Engqvist, A. Zelezniak, Expanding functional protein sequence spaces using generative adversarial networks, Nat. Mach. Intell. 3 (2021) 324–333. https://doi.org/10.1038/s42256-021-00310-5.

[55] I. Anishchenko, S.J. Pellock, T.M. Chidyausiku, T.A. Ramelot, S. Ovchinnikov, J. Hao, K. Bafna, C. Norn, A. Kang, A.K. Bera, F. DiMaio, L. Carter, C.M. Chow, G.T. Montelione, D. Baker, De novo protein design by deep network hallucination, Nature 600 (2021) 547–552. https://doi.org/10.1038/s41586-021-04184-w.

[56] A. Schenkmayerova, G.P. Pinto, M. Toul, M. Marek, L. Hernychova, J. Planas-Iglesias, V. Daniel Liskova, D. Pluskal, M. Vasina, S. Emond, M. Dörr, R. Chaloupkova, D. Bednar, Z. Prokop, F. Hollfelder, U.T. Bornscheuer, J. Damborsky, Engineering the protein dynamics of an ancestral luciferase, Nat. Commun. 12 (2021) 3616. https://doi.org/10.1038/s41467-021-23450-z.

[57] S.R. Johnson, X. Fu, S. Viknander, C. Goldin, S. Monaco, A. Zelezniak, K.K. Yang, Computational Scoring and Experimental Evaluation of Enzymes Generated by Neural Networks, (2023) 2023.03.04.531015. https://doi.org/10.1101/2023.03.04.531015.

[58] M. Vasina, D. Kovar, J. Damborsky, Y. Ding, T. Yang, A. deMello, S. Mazurenko, S. Stavrakis, Z. Prokop, In-depth analysis of biocatalysts by microfluidics: An emerging source of data for machine learning, Biotechnol. Adv. 66 (2023) 108171. https://doi.org/10.1016/j.biotechadv.2023.108171.

[59] K. Amani, M. Fish, M.D. Smith, C.D.M. Castroverde, NeuroFold: A Multimodal Approach to Generating Novel Protein Variants in silico, (2024) 2024.03.12.584504. https://doi.org/10.1101/2024.03.12.584504.

[60] Y. Sato, R. Natsume, M. Tsuda, J. Damborsky, Y. Nagata, T. Senda, Crystallization and preliminary crystallographic analysis of a haloalkane dehalogenase, DbjA, from Bradyrhizobium japonicum USDA110, Acta Crystallograph. Sect. F Struct. Biol. Cryst. Commun. 63 (2007) 294–296. https://doi.org/10.1107/S1744309107008652.

[61] Y. Cao, T.A. Geddes, J.Y.H. Yang, P. Yang, Ensemble deep learning in bioinformatics, Nat. Mach. Intell. 2 (2020) 500–508. https://doi.org/10.1038/s42256-020-0217-y.

[62] R.M. Rao, J. Liu, R. Verkuil, J. Meier, J. Canny, P. Abbeel, T. Sercu, A. Rives, MSA Transformer, in: Proc. 38th Int. Conf. Mach. Learn., PMLR, 2021: pp. 8844–8856. https://proceedings.mlr.press/v139/rao21a.html (accessed April 10, 2024).

[63] E. Castro, A. Godavarthi, J. Rubinfien, K. Givechian, D. Bhaskar, S. Krishnaswamy, Transformer-based protein generation with regularized latent space optimization, Nat. Mach. Intell. 4 (2022) 840–851. https://doi.org/10.1038/s42256-022-00532-1.

[64] M.A. Ganaie, M. Hu, A.K. Malik, M. Tanveer, P.N. Suganthan, Ensemble deep learning: A review, Eng. Appl. Artif. Intell. 115 (2022) 105151. https://doi.org/10.1016/j.engappai.2022.105151.