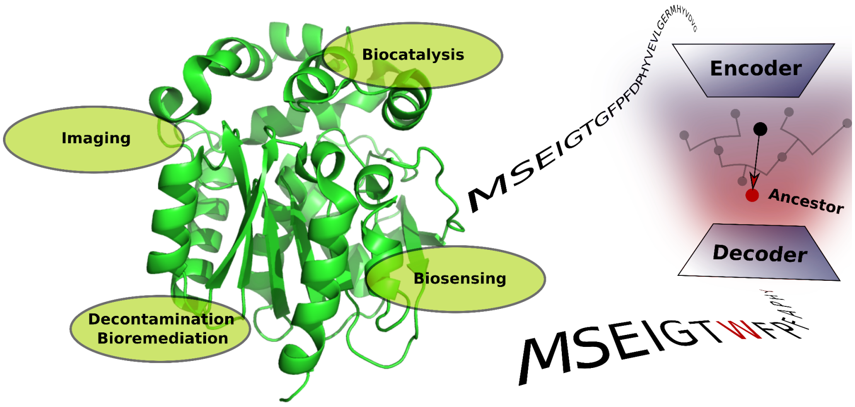
Design of Enzymes for Biocatalysis, Bioremediation, and Biosensing using Variational Autoencoder-Generated Latent Spaces

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
*Corresponding author: Stanislav Mazurenko - mazurenko@mail.muni.cz

Graphical abstract



Abstract

Enzymes enable sustainable and environmentally friendly solutions in industrial biocatalysis, bioremediation, and biosensing. Evolutionary data have proven pivotal for narrowing down the vast sequence space during enzyme optimization. However, capturing all important dependencies among residues is challenging due to the nonlinear influence of coevolution at individual positions. To overcome this challenge, deep learning methods are being actively trained on protein sequence data. While they have demonstrated incredible capacity to grasp protein evolution, the strategies to leverage this information for the design of promising biocatalysts remain largely unexplored. Here, we introduce evolutionary trajectories generated by a generative deep-learning framework of variational autoencoders. We optimized and utilized this framework and the latent space geometry to produce a set of deep-learning-based ancestral sequences of model enzymes haloalkane dehalogenases. The generated novel proteins were expressed and experimentally characterized, showing stability and activity at the level of the wild type for soluble variants. We also identified a major limitation: the sequences distant from the template tend to accumulate many insertions and deletions, known to compromise protein solubility. Taking this limitation into account, we demonstrate that the geometry of the latent space, together with the generative potential of variational autoencoders, can be used for diversification of natural protein sequences.

Introduction

Biocatalysis is a promising field that offers sustainable and environmentally friendly solutions for industry driven by the remarkable capabilities of enzymes. The power of enzymes comes from millions of years of evolution, fine-tuning them to carry out specific chemical reactions with high efficiency. That makes them attractive alternatives to traditional catalysis, which often relies on harsh conditions and toxic chemicals [Wu 2021]. Thus biocatalysts have various applications in industries ranging from pharmaceuticals to food production or sustainable processes to reduce waste and energy consumption [Bell 2021]. They also play a vital role in biodegradation, breaking down pollutants and enabling advanced biosensing technologies [Koudelakova 2013].

One of the major limitations of natural enzymes is that they often exhibit suboptimal performance under non-native environments. Enzyme engineering becomes essential for unlocking the full potential of enzymes in creating more sustainable and efficient processes [Silvestre 2021, Tiso 2022]. In addition to experimental approaches, the incorporation of computational methods for rational enzyme design can further expedite the process and reduce associated development costs [Planas-Iglesias 2021, Marques 2021]. These methods help navigate the vast sequence space, as it is estimated that only a fraction of all possible sequences folds into functional protein structures [Baker 2019]. In addition, the majority of natural proteins have marginal stability [Taverna 2002] posing a significant risk for any manipulations with their sequences, as even a single mutation may result in deleterious effects and compromise protein function. Therefore, the identification of key residues as the source of beneficial mutations for particular protein functions or finding a sequence with desired properties in this space is still a challenging task. Nonetheless, the vastness of sequential space makes exhaustive search still intractable for current computational methods.

One of the strategies to reduce the search space is by inferring desired amino acid sequence patterns *in silico* from homologous sequences [Yang 2019, Wu 2019]. These sequences contain rich information about protein evolution that can be used for further generating candidates with improved properties [McLaughlin 2012]. The study of protein evolution provides important insights into how proteins change over time and across species. This field helps us understand the molecular basis of evolution, the acquisition of new protein functions, and the origin of diseases. In protein engineering studies, homologous protein sequences can be used to identify conserved, functionally important regions, create evolutionary trees, trace ancestral sequences, and detect beneficial mutations for the system. The use of evolutionary information has been shown important for the range of protein engineering tasks, from predicting protein structures [Jumper 2021] to improving protein function and stability [Sumbalova 2018].

A promising way of leveraging evolutionary information to produce more stable proteins is ancestral sequence reconstruction [Furukawa 2020, Musil 2021, Livada 2023]. It is based on the hypothesis that far ancestors of current organisms were thermophiles and evolved from harsh prehistoric conditions, possessing proteins that were more thermostable than extant homologs [Wijma 2013]. The basis of this method is multiple sequence alignment

(MSA), shown to be an efficient way to examine dependencies in individual protein families based on conserved regions and the variability of amino acids in aligned positions [Sievers 2020]. The analysis of this variability is challenging, as it may include complicated patterns of relationships among positions. Historically, this data was used by primarily looking at only one-two positions at a time [Lehmann 2000, Morcos 2011]. More recent approaches extract patterns by deep neural networks, in particular the networks that map the sequence space onto their internal low-dimensional representation, also referred to as latent space. 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 [Lin 2023, Elnaggar 2023, Alley 2019, Hawkins-Hooker 2021].

One type of generative models that shows significant promise in this domain is Variational autoencoders (VAE) [Kingma 2013]. They have already proven useful in several applications related to protein families, such as predicting protein structures [Eguchi 2022], discovering novel drugs [Bombarelli 2018], and predicting protein functions [Lian 2022]. By training a VAE on a dataset of proteins from a specific family, the model can uncover the patterns in the structure and sequence of these proteins. This allows VAE to provide valuable insights into the evolution of protein families, as demonstrated in recent studies exploring the phylogenetic relationships within the latent space [Ding 2019][Ziegler 2023][Detlefsen 2022]. In particular, Ding *et al.* showed that the latent space of the variational autoencoders can capture the full variation of sequences in MSA by machine learning [Ding 2019]. The learned protein representations were shown to capture the biophysical properties of protein variants and the phylogenetic relationships in corresponding protein families. However, the study did not investigate how these relationships can be exploited for generating new proteins from the latent space, which would require additional optimization of architecture and the dataset.

In this study, we explore how the geometry of the latent space of the variational autoencoder can be used directly to produce novel variants of haloalkane dehalogenases (HLDs; EC 3.8.1.5), enzymes that cleave the carbon-halogen bonds using a S_N2 mechanism [Janssen 2004]. These enzymes produce alcohols, halide ions, and protons as products and are thus valuable in the synthesis of bioactive compounds, facilitating the production of key intermediates [Westerbeek 2011][Prokop 2010] essential for the development of pharmaceuticals and other biologically active substances [Leeuwen 2012][Patel 2006]. Moreover, HLDs demonstrate significant potential in biodegrading toxic chemicals and finding diverse applications in various biotechnological processes such as biosensing, cell imaging, and protein analysis [Koudelakova 2013].

The main contribution of our work is the full workflow for generating latent-space-based ancestral sequences (Figure 1). This workflow is based on a small number of catalytically-related proteins, and it aims to produce new variants that preserve protein function and improve stability. First, we mined the sequences with preserved catalytic residues using the EnzymeMiner tool [Hon 2020] to obtain an MSA of functionally related proteins. Second, we built a variational autoencoder and specified several metrics to measure its capacity to generate protein sequences and capture the phylogeny in the constructed latent representations. The final model generates sequences similar to the input dataset while maintaining enough flexibility in the generative model to infer completely new sequences. Third, based on the geometric properties of the latent space, we employed a sampling strategy and produced a statistical profile of candidate sequences to select promising variants from the evolutionary trajectory. We implemented two rounds of comprehensive experimental characterisation of the sampled sequences, including temperature and substrate specificity profiles. The first round included 9 sequences as far as 245 mutations away from the query and showed a drastically reduced solubility. The second round consisted of 3 sequences closer to the wild type in the latent space. The suggested ancestors mainly from the second round showed stability and activity at the level of the benchmark dehalogenase DhaA. Overall, we demonstrate that the structure of the latent space together with the generative potential of variational autoencoders can be used to lead the sequence search and design novel active proteins. At the same time, we identified a negative relationship between protein expression and distance from the query in the latent space and suggested promising strategies to address this limitation in the future.

Results

To leverage the power of the variational autoencoder and its latent spaces for the design of promising biocatalysts, we developed the pipeline depicted in Figure 1. This pipeline was inspired by the previous studies reporting the connections between latent space geometry and phylogeny for a given multiple sequence alignment [Ding 2019, Detlefsen 2022]. However, those studies did not attempt to exploit this connection in generating new protein sequences, which was our main motivation for the current study.

Processing the multiple sequence alignment

The data collection is optimized to preserve catalytic activities. The first step of our pipeline is to construct a multiple-sequence alignment. Instead of using Pfam alignments as in [Ding 2019], we decided to narrow down the search of relevant sequences to those likely to preserve the dehalogenation activity. Generic multiple-sequence alignments can sometimes be too broad, introducing large gapped regions and making it difficult to design proteins with desired functions [Wong 2008]. To overcome this challenge, we used the EnzymeMiner web tool [Hon 2020], which generates alignments specifically selected for function and catalytic site similarity (Figure 1A).

The search query of Haloalkane dehalogenase from Rhodococcus with UniProt ID *P59336* (see “Methods”) yielded 22567 sequences. 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 12053 sequences and 299 positions, which were used for training. Similar to [Ding 2019], 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.

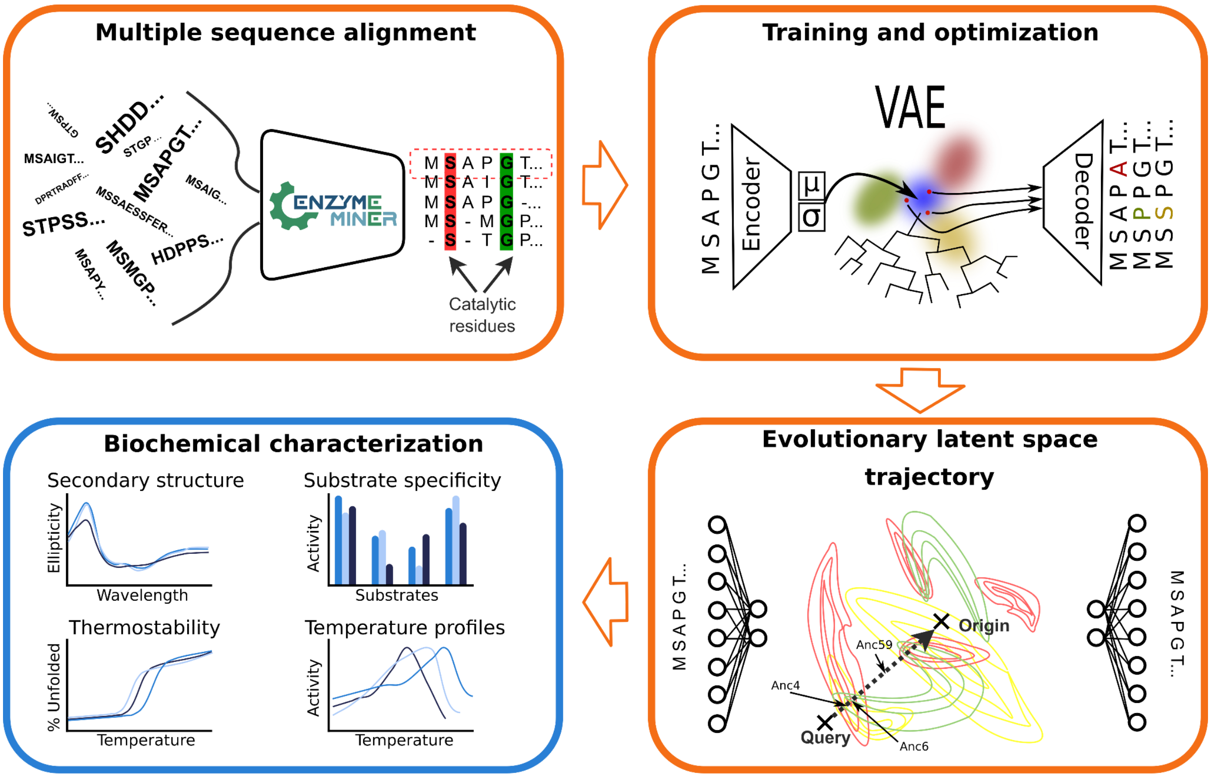


Figure 1: The scheme of the variational autoencoder-based pipeline for the design of novel sequences. The first step is an advanced sequence search of active-related proteins using the EnzymeMiner tool. **The second step** optimizes the architecture of the variational autoencoder to reproduce the sequence distribution of the MSA and capture phylogenetic dependencies within the latent space. **The third step** uses 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. **In the fourth step**, the experimental characterization of the proposed designs is conducted. The orange and blue frames represent the computational and experimental steps respectively.

Optimization of the network architecture

Variational autoencoders are a suitable framework for deciphering hidden protein dependencies. Variational autoencoders (VAEs) [Kingma 2013] are a type of deep generative learning models whose goal is to generate new data that is similar to the input. VAEs consist of two main components: an encoder and a decoder (Figure 1B). The encoder takes the input data and maps it to a lower dimensional representation called a latent representation. Within this latent space, the encoded input is modeled as a normal distribution by two parameters, mean and variance. Subsequently, the decoder operates by drawing 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 that consists of two parts. The first part called the reconstruction term, penalizes incorrect reconstruction of the input data, helping the model learn the latent space rich enough for the decoder to reconstruct the input sequences. The second one 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 and those are enforced to be close to normal by measuring 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.

VAEs have previously been shown to have the ability to capture phylogenetic relationships between proteins [Ding 2019]. *Ding et al.* revealed that within the latent space, sequences extracted from MSA do not exhibit random distribution but instead form a star-like structure characterized by multiple spikes extending from the center outward in a specific direction. As a result, sequences within separate latent space clusters tend to group together if they belong to the same branch at a particular reference evolutionary time point in the phylogenetic tree. The study also showed that as sequences evolve from a root node to a leaf in the phylogenetic tree, their positions in latent space move from the proximity of the origin to the outer regions along a defined trajectory. Based on this, we hypothesized that the latent space coordinates could be utilized to navigate the search in the sequence space for generating ancestral-like sequences, thus potentially improving query protein stability while preserving its function.

Variational autoencoder optimization for ancestral sequence reconstruction. Before testing our hypothesis, we embarked on selecting the best model architecture based on the implementation provided by [Ding 2019]. This involved optimizing the encoder, decoder, and training procedure to minimize the difference between the generated and input sequence distributions (generative capacity) while also preserving the relationship between phylogeny and the latent space (geometric properties).

In order to evaluate the generative capacity of our model, we devised 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 3000 randomly sampled MSA input sequences with those generated by our VAE model (Figure 2A). The first-order statistics compare the frequencies of amino acids in every position. The VAE model successfully reproduced the first-order statistics, 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 [Russ 2020]. 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 [Johnson 2021, Morcos 2011]. The VAE model reproduced the empirical second-order statistics of natural sequences (Figure 2B). We integrated query reconstruction accuracy [Costello 2019] as an additional metric into our analysis. This addition was crucial because relying solely on order statistics couldn't ensure that the model generated the query sequence with minimal mutations from its latent space embedding, which was essential for capturing strong evolutionary signals. Notably, the final model demonstrated the ability to reconstruct sequences with as few as 4 mutations.

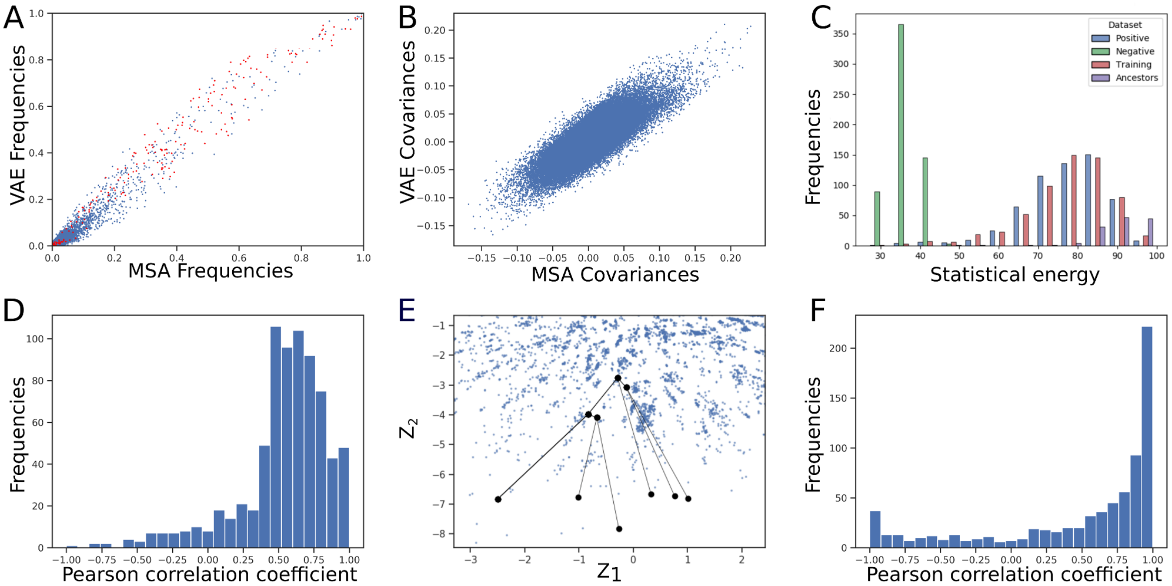


Figure 2: Showcases of the statistics used to measure the generative capacity of the final VAE model and the geometric properties of its latent space. A-B the first and second-order statistics for 3000 sequences randomly selected for the input MSA or VAE generated. The first-order statistics with red dots represent the gap symbol frequencies in sequence positions ρ=0.98. The second-order statistics demonstrate that our model can reconstruct pairwise amino acid occurrences ρ=0.68. **C** The average reconstruction accuracy for the "negative" (green), "training" (red), "positive" (blue), and "ancestral" (violet) control sets. 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-F** The relationship between phylogeny in the input MSA and the geometry of the latent space. **D** The positive correlation between the depth and the latent space origin distance for most of the sequences in the tree branches. **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 sampling energies of sequences from various control sets. 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" control consisted of the same number of sequences sampled randomly from the training dataset. For the "positive" control set, we evaluated the energy of 5% MSA sequences removed from the input before training. We ended up with 612 sequences for the "negative", "positive", and "training" controls. The statistical properties of our model can distinguish random sequences from those in the MSA and identify sequences from the given family (Figure 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, enabling the generation of ancestral sequences of the desired protein family. Thus, while optimizing model properties we kept track of the relationship between phylogeny and the geometry of latent space (Figure 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 encoded latent space point distances to the latent space origin of protein sequences and their corresponding positions in the phylogenetic tree (Figure 2D). To gain insight into the general patterns of embeddings in individual tree branches, we calculated the dot product of angles between two vectors: the one going from the latent space position of each leaf node to the origin of the latent space, and the other is defined by the first principal component of the latent coordinates of all nodes within the phylogenetic branch, extending from the extant sequence to the root of the tree (Figure 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. The final width of the dense layer was set to the length of the protein sequence and the latent space dimensionality of 2 was sufficient to reconstruct the statistics ofthe input sequences and capture phylogenic relationships in the geometry of the latent space.

Construction of the evolutionary trajectory

Latent space is able to capture stability in its structure. The ancestral sequences are often associated with greater stability than their extant counterparts [Spence 2021, Livada 2023]. We hypothesized that the structure of the VAE latent space may 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. The first set consisted of six ancestral sequences from the previous ancestral campaign of the thoroughly characterized dehalogenases DbjA, DbeA, DhaA, DmxA, and DmmA [Babkova 2017]. The second set consisted of previously engineered DhaA variants incorporating both evolutionary and energy-based mutations based on the FireProt method [Beerens 2018, Musil 2021]. Sequences were aligned to those in the training MSA (see Methods). Sequences from both sets were mapped closer to the origin of the latent space (Figure 3A, B). Our results support the notion that the latent space encodes the stability landscape.

Specification of a latent space strategy to guide the search and selection of sequences. The regular distribution of stable sequences in the latent space (Figure 3B) led us to develop a *straight-line evolutionary strategy*. This strategy encodes the query sequence into its latent representation and then follows the straight line connecting that point to the origin of the latent space (Figure 3A), mimicking the mapping of ancestral dependencies into the latent space. The line is divided into equal intervals, whose boundaries are selected latent space points 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 inserted/deleted characters from the original sequence. The values obtained were plotted and visually inspected to identify points with interesting statistical values, which were selected for further characterization in the laboratory. The generated profile was then used to select suitable variants for subsequent laboratory experiments (Figure 3C).

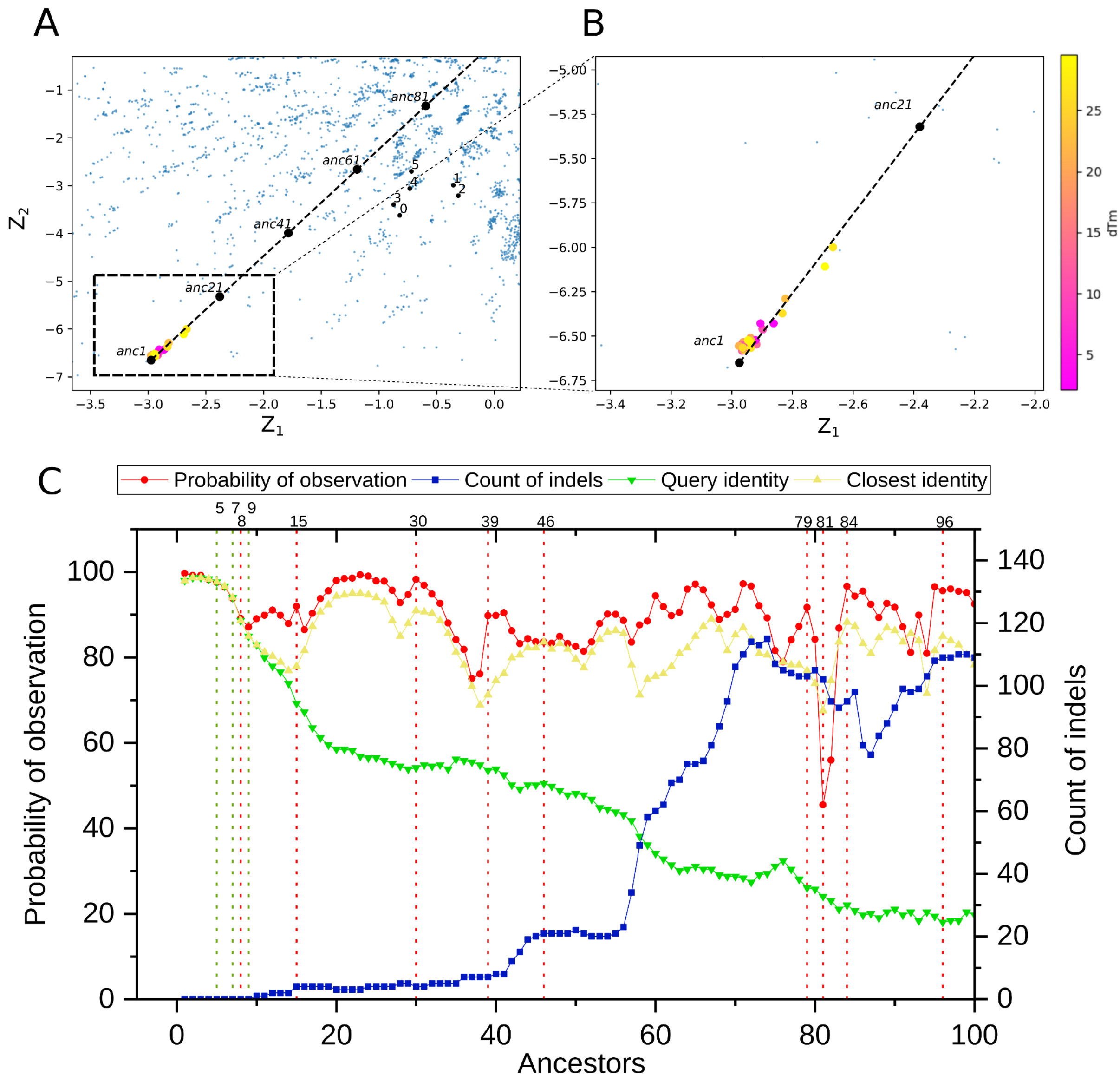


Figure 3: Latent Space Structure-Guided Mining Strategy for Novel Ancestral Proteins. A Straight evolutionary strategy reconstructed 100 sequences along the trajectory from query embedding to the latent space origin. The embeddings of experimentally validated Babkova's ancestors [Babkova 2017] (points 0-5) and engineered DhaA variants [Beerens 2018] 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. **C** The statistical profile of 100 sequences from the straight evolutionary strategy. The dashed vertical lines represent sequences selected for experimental characterization (see Table S1) where green 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 0 indels to 138 substitutions and 109 indels (Table 1). The substitutions covered the entire protein structure and with deletions in most of the cases. In the second round, we focused on the more conserved variants (ancestors 5, 7, and 8) with 7 to 34 substitutions without insertions or deletions. In total, 12 designs were taken to the lab for expression and biophysical characterization.

Experimental characterization of expressed variants

Variant production. Protein overexpression was carried out in two rounds (see Table S1) to assess the solubility and activity of DhaA variants. In the first round, the solubility of DhaA ancestor variants was analyzed by overproducing them in *E. coli* and purifying them using a purification buffer (16.4 mM K_2HPO_4, 3.6 mM KH_2PO_4, 400 mM NaCl, 10 mM imidazole, pH 7.5). As a reference positive control, the soluble DhaA template was also expressed and purified. Among all the tested variants from this round, only Ancestor 9 was purified as a soluble protein in sufficient amounts (Figure S2). Furthermore, a halide oxidation (HOX) assay was used for screening of the dehalogenase activity (Figure S3A)[Aslan‐Üzel 2020]. Apart from the positive control, only one active variant, Ancestor 9, was identified. Solubility issues were expected as several previous ML-based pipelines for protein design have led to soluble designs in around 20% of the generated sequences [Repecka 2021, Anishchenko 2021]. Inspired by the good yield of Ancestor 9, in the second round, the ancestors 5, 7, and 8 were selected for production. Among these variants, Ancestor 5 showed the highest expression, followed by Ancestor 7, while Ancestor 8 exhibited the lowest expression and solubility (Figure S2). Based on the results from both rounds, Ancestors 5, 7, and 9 were chosen for further characterization.

Secondary structure experimental validation. To confirm proper folding of the studied variants, circular dichroism (CD) spectra were collected for all soluble variants. The experimental results from all of these measurements are summarized in **Figure 4A**. The CD spectra of Ancestor 5 highly resemble the one of the Template (typical α/β-hydrolase fold), confirming proper folding. On the contrary, the spectra of Ancestral 9 and also Ancestor 7 (**Figure S4B**) deviate from the template spectrum. To further understand the secondary structure of the variants BeStSel5 [Micsonai 2022] was used for fitting of experimental data and analysis of PDB structures, and PDBMD2CD [Drew 2020] for predicting CD spectra from experimental structures and AlphaFold models. As for the composition of secondary structure (Figure S2), Ancestor 9 and Ancestor 7 exhibited a decreased ratio of helices and an increased ratio of antiparalell β-sheets compared to the other variants. Figure S4B shows that the prediction of CD spectra based on AlphaFold structures does not predict the changes in folding of the particular variant. 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 further assessed by means of CD spectroscopy (**Figure S5A**) and nano-differential scanning fluorimetry (**Figure 2B** and **S5B**). The apparent melting temperatures for the variants were in the range of 47 – 53 °C, showing that the thermostability was not compromised, but neither improved. Protein aggregation has been observed for the variant Ancestor 9, showing aggregation onset at about 45°C.

Temperature profiles. The next step in the biochemical characterization of all variants was the measurement of temperature profiles (**Figure 4C**). All the variants showed the T_{max} (temperature where maximum activity was detected) of 40 °C, which is in agreement with previously determined temperature profiles for DhaA [Buryska 2019]. The temperature profile for Ancestor 7 was not obtained, as the activities were below the limit of detection. Due to compromised activity and folding, Ancestor 7 was excluded from substrate specificity profiling.

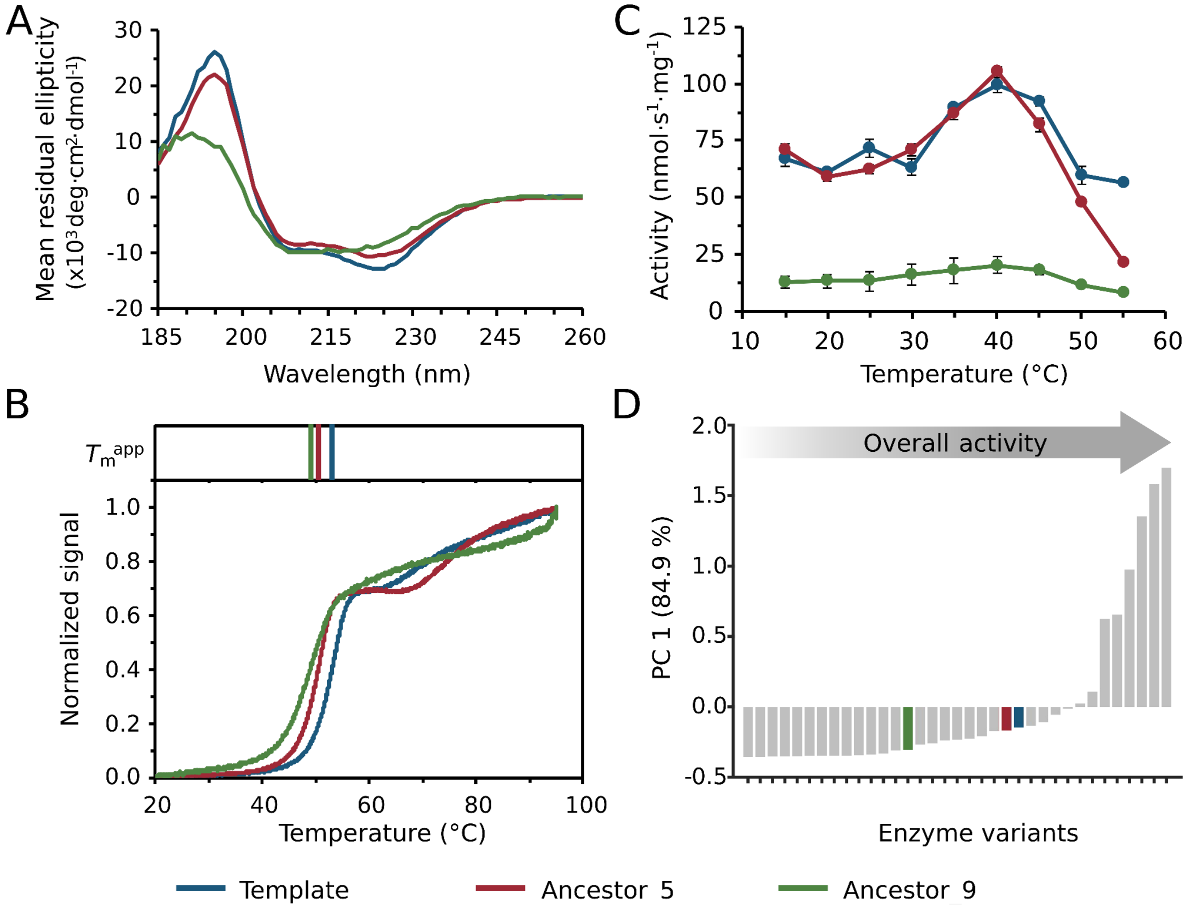


Figure 4 Laboratory testing of selected variants. (**A**) Far-UV circular dichroism spectra probing the correct folding and secondary structure of the variants. (**B**) Normalized thermal denaturation curves from nanoDSF spectroscopy with apparent melting temperatures (T_m^{app}) shown above the curves. (**C**) The dependence of specific activity on temperature. The error bars denote the standard deviations of the data. (**D**) The score plot shows the first principal component (PC 1, explaining 84.9 % of the data variance), which compares characterized haloalkane dehalogenases in terms of their overall activity with 27 substrates. The data for enzymes in grey were collected previously [Vasina 2022]. The tested variants including DhaA Template, Ancestor 5, and Ancestor 9 are shown in blue, red, and green color, respectively, as shown in the bottom.

Substrate specificity. The temperature of 35 °C has been selected for the subsequent specificity characterization as it was below the onset of denaturation for each of the remaining variants and the temperature is very close to T_{max} (**Figure 4D**). The obtained substrate-specificity profile of Ancestor 5 was analogous to the Template DhaA (**Figure S6A and S6B**). Importantly, Ancestor 5 showed improved activity for 19 out of 27 substrates. On the other hand, Ancestor 9 showed a decreased overall activity of roughly 21% (**Figure S6B**).

Probably the high number of substitutions (45) negatively affected the function of this variant.

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 [Vasina 2022] with the newly obtained data (Figure S6). PCA confirmed that in terms of overall activity among HLDs, the template and Ancestor 5 were above average. The analysis of log-transformed activity data (**Figure S7**) further showed that all of the studied variants differed only very slightly in their substrate preferences.

Discussion

In this study, we explored the latent space of VAEs for mapping a dataset of functionally related sequences of haloalkane dehalogenases obtained from our in-house tool EnzymeMiner [Hon 2020], onto this latent space. The initial step of the pipeline proposed recently by Ding and coworkers [Ding 2019] led to the observation that the latent space effectively captured the phylogenetic dependencies of our sequences. In particular, the ancestral sequences were positioned closer to the origin of the latent space than corresponding query sequences. These findings motivated us to use the VAE framework as an orthogonal method of ancestral sequence reconstruction to generate novel ancestral-like variants of our target sequence DhaA.

To this end, we focused on optimizing the network hyperparameters to ensure that the model could reproduce the statistical frequencies of the input data while generating variants close to the query sequence. We also aimed to preserve the structural evolutionary properties of latent space. To assess the generative ability of the model, we introduced statistical metrics including the first and second-order statistics, as well as quantitative embeddings of the phylogeny trees. Through our experiments, we found that a small feed-forward neural network showed good reproducibility of these statistics. The final model consisted of a single dense layer with a width equal to the number of columns in the input MSA. Additionally, we found that a two-dimensional latent space was sufficient for our purposes, similar to the recent findings of Ziegler and coworkers [Ziegler 2023].

We then introduced a simple strategy to generate ancestral sequences based on the geometry of the latent space. This strategy exploited the latent space to facilitate the exploration of sequence variants with potential functional properties. We achieved this by reconstructing the embeddings along the trajectory towards the origin of the latent space. Using this approach, we generated a total of 100 candidate sequences, each accompanied by a statistical profile, with the aim of further narrowing down the candidate space. Based on two rounds of experimental characterisation of 9 and 3 variants, respectively, Ancestors 5, 7, and 9, were subjected to the comprehensive analysis of secondary structure, thermostability, and temperature profile analyses. The secondary structure analysis revealed notable discrepancies between the predicted CD spectra obtained from AlphaFold calculations and the experimental validation results. Thermostability analysis showed that the ancestral variants did not improve the thermostability of the template, yet it was not significantly compromised. The substrate specificity was measured only for Ancestors 5 and 9, as the Ancestor 7 showed neglectable activity in the temperature profiling. Ancestor 9 showed moderate activity, while the activities of Ancestors 5 and Template were found to be above average among haloalkane dehalogenases.

The latent space models employed in this study hold considerable promise for protein engineering. However, further advancements and experimental validation of these models are necessary to fully exploit their potential for protein design. Recent developments in transformer-based architectures have demonstrated their efficacy in extracting crucial patterns from amino acid sequences alone [Rao 2021][Castro 2022]. Integrating these architectures with manifold learning techniques applied to the proposed latent space can further enhance their ability to generate highly fit sequences [Detlefsen 2022]. Notably, the work of Ziegler et al. [Ziegler 2023] has drawn attention to the use of a manifold projected onto the latent space, which improves interpretability and may facilitate ancestral design. To bolster the robustness of future studies, adopting a generation protocol for ancestral sequences that incorporates an ensemble of models might also be highly advantageous [Ganaie 2022]. This approach addresses the observed instability of ancestral trajectories within the latent space and could establish a more reliable foundation for subsequent investigations (Figure S1). In summary, our approach provided a systematic framework for exploring the latent space using its geometry, identifying functionally relevant sequences, and reducing the number of candidates through the application of statistical analysis.

Acknowledgments

P. Kohout and M. Vašina are holders of the Brno Ph.D. Talent scholarship funded by the Brno City Municipality and the JCMM. Computational resources were provided by the e-INFRA CZ project (ID:90254), 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 [CETOCOEN Excellence CZ.02.1.01/0.0/0.0/17_043/0009632, ESFRI RECETOX RI LM2023069, ESFRI ELIXIR LM2023055]; Technology Agency of the Czech Republic [TN02000122]; CETOCOEN EXCELLENCE project supported from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857560; by the project National institute for cancer research (Programme EXCELES, No.LX22NPO5102) – Funded by the European Union – Next Generation EU, and by COST (European Cooperation in Science and Technology) [COST Action COZYME CA21162]. 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.

Methods

MSA and data preprocessing

The input MSA is created by the EnzymeMiner tool [3]. 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. The result can be found at the EnzymeMiner web page with job id xvmwa7 or in the supplement. 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 having gaps in less than 20% of sequences; 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.

MSA Alignments of Babkovas and DhaA115 dataset

We aligned the MSA of Babkova's sequences, consisting of six sequences in total, with the MSA profile of the original input sequences. We applied an indice slicing technique. This slicing process allowed us to retain only the columns that remained after the MSA processing procedure. Similarly, for the DhaA115 dataset, we performed individual sequence alignment with the query sequence (DhaA - P59336_S14). Then, we applied an indice slicing operation to match the width of the network input, preserving the indices that were retained after the MSA preprocessing step.

Hyperparameters 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. After hyperparameter search, we used the one hidden layer in the encoder and decoder composed of N neurons where N is the width of preprocessed MSA (number of positions), with latent space dimensionality 2 and no decay factor applied. 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. In the end, the model has 3 million parameters.

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. In the example where at a particular position of the generated and input alignment we obtain values for the relative frequencies of the trait (0.5,0.5), it would indicate that for a given position, there is an amino acid that occurs 50% of the time in the generated samples and 50% of the reference sequences.

The second-order statistics are similar but include two positions instead of one. Value of frequencies at (0,5,0,3) would then indicate that 30% of the generated sequences contain the given X, Y column pair the simultaneous occurrence of a particular pair of amino acids A1 at position X and A2 at Y position, while in the input dataset a given ordered pair of amino acids occurs in identical columns 50% of the time.

In order 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:

$$C^{ij}_{\alpha\beta} = f^{ij}_{\alpha\beta} - f^{i}_{a} f^{b}_{\beta} \qquad (7.3)$$

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 independence. If $C^{ij}_{\alpha\beta}$ equals 0 for all $\alpha\beta$, then positions i, j are not in the covariance. Co-evolving amino acids are an important aspect of sequence variability in protein alignments and the ability of the generative model to reproduce pairwise covariance scores of the training dataset has been used in the past as a basic, non-trivial measure of the ability model to model protein sequence covariance. Thus, we can detect how well the model captures interactions between distant amino acids, an important indicator for the likely stability and function of the generated proteins [McGee 2021].

For each model, we compare 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 $\varrho(\{C^{ij}_{\alpha\beta}\},\{\hat{C}^{ij}_{\alpha\beta}\})$. In the case of my models, the statistics were calculated for the input and synthetically generated datasets having 3000 randomly selected samples. Synthetic data were reconstructed from the latent space points that were sampled according to the chosen a priori distribution, i.e., a Gaussian distribution with mean equal to 0 and variance 2.

The average reconstruction accuracy of the sequence was approximated as an average reconstructed sequence identity for 5000 samples around the original sequence coordinates of its latent space embedding. The parameters of sample distribution are determined by the encoder (mean, variance). The negative subset was generated by sampling sequences from the profile of input MSA based on the amino acid frequencies in each position. The positive subset was composed of 5% of preprocessed sequences randomly selected from the MSA and excluded from the training, the rest is included in the training subset. The ancestral subset was composed of reconstructed sequences by the straight evolutionary strategy.

Phylogeny mapping and evaluation

For the latent space phylogenetic analysis, we generated 13 phylogenetic trees using our input MSA dataset. Each tree consisted of approximately 100 randomly sampled leaf nodes from the pre-processed MSA. Using our fully automated in-house ancestral sequence reconstruction tool, FireProtASR [Musil 2021], 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 a similar approach to [Ding 2019]. 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 using the straight evolutionary strategy to obtain likely ancestral sequences.

Model ensemble trajectories

Model ensembles are important in machine learning because they improve performance, enhance generalization capabilities, provide a more comprehensive representation of the data, and enable model selection and averaging. In our study, we examined the embeddings over an ensemble of four randomly initialized VAE models. Each individual model was used to generate 100 ancestral trajectories using a straight evolutionary strategy. These trajectories were then re-embedded into latent space using the encoder weights of the first model (Figure S1, left top). It is evident that the trajectories in the latent space exhibit scattered patterns, with limited spatial coherence observed at their starting points. While the trajectory of the model 0 is stable along the line to the origin the other models tend to explore different sequential subspaces. These observations are consistent with the cross-embedding of evolutionary trajectories into different VAE models (Figure S1).

Cell transformation

Escherichia coli BL21(DE3) cells (NEB, USA) were transformed with expression plasmid vector pET21b containing 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) was the whole soluble fraction 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 centrifugation 94 g for 2 min followed by resuspending in 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 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 (APF), 26 mM H2O2, 1.1 U *Curvularia inaequalis* vanadium chloroperoxidase with additional His tag (VCPO.His), 1 mM orthovanadate, 20 mM phosphate buffer, pH = 8.0. Also, 4 µl of resuspended cells was added into each well, followed by the addition of 96 µl of 0.3 mM DBEin reaction buffer. Fluorescence was measured for 60 min, in 60 s intervals (Excitation at 488 nm; emission Detection at 525 nm; 30 °C) using Synergy™ H4 Hybrid Microplate Reader (BioTek, USA). Data for all tested variants were measured in triplicate and were normalized to OD_{600} = 1. Average activity with a standard deviation of three 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.

Bench-scale protein purification with resin

The DNase was added to the cell culture (20 µg/ml) after defrosting from -80 °C. The culture was sonicated using a sonicator (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 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 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 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^{TM} (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

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 (AppliedPhotophysics). 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 [Drew 2020] (https://pdbmd2cd.cryst.bbk.ac.uk). For this prediction, either experimental structures from PDB database (1CQW for DhaA) or AlphaFold models were taken as an 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 [Micsonai 2022] (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 either 1 mM HEPES or 50 mM 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, 210, and 227 nm from 15 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^{app}) 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 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 determination of temperature profiles and substrate specificity were conducted on the capillary-based droplet microfluidic platform MicroPEX [Vasina 2022], 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 [Buryska 2019][Vasina 2020].

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 [Vasina 2022], [Buryska 2019]. 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 [Vasina 2022] and the activities obtained for Template and Ancestors 5 and 9 (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. Each specific activity needed to be incremented by 1 to avoid the logarithm of zero values. 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.

References

[Anishchenko 2021] Anishchenko I, Pellock SJ, Chidyausiku TM, Ramelot TA, Ovchinnikov S, Hao J, Bafna K, Norn C, Kang A, Bera AK, DiMaio F. De novo protein design by deep network hallucination. Nature. 2021 Dec 16;600(7889):547-52.

[Alley 2019] Alley EC, Khimulya G, Biswas S, AlQuraishi M, Church GM. Unified rational protein engineering with sequence-based deep representation learning. Nature methods. 2019 Dec;16(12):1315-22.

[Aslan‐Üzel 2020] Aslan‐Üzel AS, Beier A, Kovář D, Cziegler C, Padhi SK, Schuiten ED, Dörr M, Böttcher D, Hollmann F, Rudroff F, Mihovilovic MD. An ultrasensitive fluorescence assay for the detection of halides and enzymatic dehalogenation. ChemCatChem. 2020 Apr 6;12(7):2032-9.

[Babkova 2017] Babkova P, Sebestova E, Brezovsky J, Chaloupkova R, Damborsky J. Ancestral Haloalkane Dehalogenases Show Robustness and Unique Substrate Specificity. Chembiochem. 2017;18(14):1448-1456.

[Beerens 2018] Beerens K, Mazurenko S, Kunka A, Marques SM, Hansen N, Musil M, Chaloupkova R, Waterman J, Brezovsky J, Bednar D, Prokop Z. Evolutionary analysis as a powerful complement to energy calculations for protein stabilization. ACS Catalysis. 2018;8(10):9420-8.

[Baker 2019] Baker D. What has de novo protein design taught us about protein folding and biophysics?. Protein Science. 2019 Apr;28(4):678-83.

[Bell 2021] Bell EL, Finnigan W, France SP, Green AP, Hayes MA, Hepworth LJ, Lovelock SL, Niikura H, Osuna S, Romero E, Ryan KS. Biocatalysis. Nature Reviews Methods Primers. 2021 Jun 24;1(1):46.

[Bombarelli 2018] Gómez-Bombarelli R, Wei JN, Duvenaud D, Hernández-Lobato JM, Sánchez-Lengeling B, Sheberla D, Aguilera-Iparraguirre J, Hirzel TD, Adams RP, Aspuru-Guzik A. Automatic chemical design using a data-driven continuous representation of molecules. ACS central science. 2018 Feb 28;4(2):268-76.

[Bornscheuer 2012] Bornscheuer UT, Huisman GW, Kazlauskas RJ, Lutz S, Moore JC, Robins K. Engineering the third wave of biocatalysis. Nature. 2012;485(7397):185-94.

[Buryska 2019] Buryska T, Vasina M, Gielen F, Vanacek P, van Vliet L, Jezek J, Pilat Z, Zemanek P, Damborsky J, Hollfelder F, Prokop Z. Controlled oil/water partitioning of hydrophobic substrates extending the bioanalytical applications of droplet-based microfluidics. Analytical chemistry. 2019;91(15):10008-15.

[Castro 2022] Castro E, Godavarthi A, Rubinfien J, Givechian K, Bhaskar D, Krishnaswamy S. Transformer-based protein generation with regularized latent space optimization. Nature Machine Intelligence. 2022;4(10):840-51.

[Chaloupkova 2011] Chaloupkova R, Prokop Z, Sato Y, Nagata Y, Damborsky J. Stereoselectivity and conformational stability of haloalkane dehalogenase DbjA from Bradyrhizobium japonicum USDA110: the effect of pH and temperature. FEBS J. 2011;278(15):2728-38.

[Chaloupkova 2014] Chaloupkova R, Prudnikova T, Rezacova P, Prokop Z, Koudelakova T, Daniel L, Brezovsky J, Ikeda-Ohtsubo W, Sato Y, Kuty M, Nagata Y, Kuta Smatanova I, Damborsky J. Structural and functional analysis of a novel haloalkane dehalogenase with two halide-binding sites. Acta Crystallogr D Biol Crystallogr. 2014;70(Pt 7):1884-97.

[Costello 2019] Costello Z, Martin HG. How to hallucinate functional proteins. arXiv preprint arXiv:1903.00458. 2019.

[Detlefsen 2022] Detlefsen NS, Hauberg S, Boomsma W. Learning meaningful representations of protein sequences. Nature communications. 2022;13(1):1914.

[Ding 2019] Ding X, Zou Z, Brooks III CL. Deciphering protein evolution and fitness landscapes with latent space models. Nature communications. 2019;10(1):5644.

[Drew 2020] Drew ED, Janes RW. PDBMD2CD: providing predicted protein circular dichroism spectra from multiple molecular dynamics-generated protein structures. Nucleic Acids Research. 2020;48(W1):W17-24.

[Elnaggar 2023] Elnaggar A, Essam H, Salah-Eldin W, Moustafa W, Elkerdawy M, Rochereau C, Rost B. Ankh ☥: Optimized Protein Language Model Unlocks General-Purpose Modelling. bioRxiv. 2023:2023-01.

[Eguchi 2022] Eguchi RR, Choe CA, Huang PS. Ig-VAE: Generative modeling of protein structure by direct 3D coordinate generation. PLoS computational biology. 2022 Jun 27;18(6):e1010271.

[Furukawa 2020] Furukawa R, Toma W, Yamazaki K, Akanuma S. Ancestral sequence reconstruction produces thermally stable enzymes with mesophilic enzyme-like catalytic properties. Scientific reports. 2020 Sep 23;10(1):1-3.

[Ganaie 2022] Ganaie MA, Hu M, Malik AK, Tanveer M, Suganthan PN. Ensemble deep learning: A review. Engineering Applications of Artificial Intelligence. 2022 Oct 1;115:105151.

[Gehret 2012] Gehret JJ, Gu L, Geders TW, Brown WC, Gerwick L, Gerwick WH, Sherman DH, Smith JL. Structure and activity of DmmA, a marine haloalkane dehalogenase. Protein Sci. 2012;21(2):239-48.

[Hawkins-Hooker 2021] Hawkins-Hooker A, Depardieu F, Baur S, Couairon G, Chen A, Bikard D. Generating functional protein variants with variational autoencoders. PLoS computational biology. 2021;17(2):e1008736.

[Hon 2020] Hon J, Borko S, Stourac J, Prokop Z, Zendulka J, Bednar D, Martinek T, Damborsky J. EnzymeMiner: automated mining of soluble enzymes with diverse structures, catalytic properties and stabilities. Nucleic acids research. 2020;48(W1):W104-9.

[Janssen 2004] Janssen DB. Evolving haloalkane dehalogenases. Current opinion in chemical biology. 2004 Apr 1;8(2):150-9.

[Johnson 2021] Johnson SR, Monaco S, Massie K, Syed Z. Generating novel protein sequences using Gibbs sampling of masked language models. bioRxiv. 2021:2021-01.

[Jumper 2021] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A. Highly accurate protein structure prediction with AlphaFold. Nature. 2021 Aug;596(7873):583-9.

[Kingma 2013] Kingma DP, Welling M. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114. 2013.

[Koudelakova 2013] Koudelakova T, Bidmanova S, Dvorak P, Pavelka A, Chaloupkova R, Prokop Z, Damborsky J. Haloalkane dehalogenases: biotechnological applications. Biotechnology journal. 2013 Jan;8(1):32-45.

[Sato 2007] Sato Y, Natsume R, Tsuda M, Damborsky J, Nagata Y, Senda T. Crystallization and preliminary crystallographic analysis of a haloalkane dehalogenase, DbjA, from Bradyrhizobium japonicum USDA110. Acta Crystallogr Sect F Struct Biol Cryst Commun. 2007;63(Pt 4):294-6.

[Lehmann 2000] Lehmann M, Pasamontes L, Lassen SA, Wyss M. The consensus concept for thermostability engineering of proteins. Biochimica et Biophysica Acta (BBA)-protein structure and molecular enzymology. 2000 Dec 29;1543(2):408-15.

[Lian 2022] Lian X, Praljak N, Subramanian S, Wasinger S, Ranganathan R, Ferguson AL. Deep learning-enabled design of synthetic orthologs of a signaling protein. bioRxiv. 2022:2022-12.

[Lin 2023] Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, Smetanin N, Verkuil R, Kabeli O, Shmueli Y, dos Santos Costa A. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science. 2023;379(6637):1123-30.

[Livada 2023] Livada J, Vargas AM, Martinez CA, Lewis RD. Ancestral Sequence Reconstruction Enhances Gene Mining Efforts for Industrial Ene Reductases by Expanding Enzyme Panels with Thermostable Catalysts. ACS Catalysis. 2023 Feb 6;13(4):2576-85.

[Leeuwen 2012] van Leeuwen JG, Wijma HJ, Floor RJ, van der Laan JM, Janssen DB. Directed evolution strategies for enantiocomplementary haloalkane dehalogenases: from chemical waste to enantiopure building blocks. ChemBioChem. 2012 Jan 2;13(1):137-48.

[Luo 2021] Luo Y, Jiang G, Yu T, Liu Y, Vo L, Ding H, Su Y, Qian WW, Zhao H, Peng J. ECNet is an evolutionary context-integrated deep learning framework for protein engineering. Nature communications. 2021 Sep 30;12(1):5743.

[Marques 2021] Marques, S. M.; Planas-Iglesias, J.; Damborsky, J. Web-Based Tools for Computational Enzyme Design. Curr. Opin. Struct. Biol. 2021, 69, 19–34.

[McGee 2021] McGee F, Hauri S, Novinger Q, Vucetic S, Levy RM, Carnevale V, Haldane A. The generative capacity of probabilistic protein sequence models. Nature communications. 2021;12(1):6302.

[McLaughlin 2012] McLaughlin Jr RN, Poelwijk FJ, Raman A, Gosal WS, Ranganathan R. The spatial architecture of protein function and adaptation. Nature. 2012 Nov 1;491(7422):138-42.

[Merkl 2016] Merkl R, Sterner R. Ancestral protein reconstruction: techniques and applications. Biol Chem. 2016;397(1):1-21.

[Micsonai 2022] Micsonai A, Moussong E, Wien F, Boros E, Vadászi H, Murvai N, Lee YH, Molnár T, Réfrégiers M, Goto Y, Tantos A. BeStSel: webserver for secondary structure and fold prediction for protein CD spectroscopy. Nucleic Acids Research. 2022;50(W1):W90-8.

[Morcos 2011] Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, Zecchina R, Onuchic JN, Hwa T, Weigt M. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proceedings of the National Academy of Sciences. 2011 Dec 6;108(49):E1293-301.

[Musil 2021] Musil M, Khan RT, Beier A, Stourac J, Konegger H, Damborsky J, Bednar D. FireProtASR: a web server for fully automated ancestral sequence reconstruction. Briefings in bioinformatics. 2021;(4):bbaa337.

[Newman 1999] Newman J, Peat TS, Richard R, Kan L, Swanson PE, Affholter JA, Holmes IH, Schindler JF, Unkefer CJ, Terwilliger TC. Haloalkane dehalogenases: structure of a Rhodococcus enzyme. Biochemistry. 1999;38(49):16105-14.

[Patel 2006] Patel RN. Biocatalysis: synthesis of chiral intermediates for drugs. Curr Opin Drug Discov Devel. 2006 Nov;9(6):741-64. PMID: 17117684.

[Prokop 2010] Prokop, Z., Sato, Y., Brezovsky, J., Mozga, T., Chaloupkova, R., Koudelakova, T., Jerabek, P., Stepankova, V., Natsume, R., van Leeuwen, J. ., Janssen, D., Florian, J., Nagata, Y., Senda, T. and Damborsky, J. (2010), Enantioselectivity of Haloalkane Dehalogenases and its Modulation by Surface Loop Engineering. Angewandte Chemie, 122: 6247-6251. https://doi.org/10.1002/ange.201001753

[Planas-Iglesias 2021] Planas-Iglesias J, Marques SM, Pinto GP, Musil M, Stourac J, Damborsky J, Bednar D. Computational design of enzymes for biotechnological applications. Biotechnol Adv. 2021 Mar-Apr;47:107696. doi: 10.1016/j.biotechadv.2021.107696. Epub 2021 Jan 26. PMID: 33513434.

[Rao 2021] Rao RM, Liu J, Verkuil R, Meier J, Canny J, Abbeel P, Sercu T, Rives A. MSA transformer. International Conference on Machine Learning 2021 Jul 1 (pp. 8844-8856). PMLR.

[Repecka 2021] Repecka D, Jauniskis V, Karpus L, Rembeza E, Rokaitis I, Zrimec J, Poviloniene S, Laurynenas A, Viknander S, Abuajwa W, Savolainen O. Expanding functional protein sequence spaces using generative adversarial networks. Nature Machine Intelligence. 2021 Apr;3(4):324-33.

[Russ 2020] Russ WP, Figliuzzi M, Stocker C, Barrat-Charlaix P, Socolich M, Kast P, Hilvert D, Monasson R, Cocco S, Weigt M, Ranganathan R. An evolution-based model for designing chorismate mutase enzymes. Science. 2020;369(6502):440-5.

[Saito 2021] Saito Y, Oikawa M, Sato T, Nakazawa H, Ito T, Kameda T, Tsuda K, Umetsu M. Machine-learning-guided library design cycle for directed evolution of enzymes: the effects of training data composition on sequence space exploration. ACS Catalysis. 2021;11(23):14615-24.

[Sevgen 2023] Sevgen, Emre, Joshua Moller, Adrian Lange, John Parker, Sean Quigley, Jeff Mayer, Poonam Srivastava et al. ProT-VAE: Protein Transformer Variational AutoEncoder for Functional Protein Design. bioRxiv: 2023-01 (2023).

[Sievers 2020] Sievers F, Barton GJ, Higgins DG. Multiple sequence alignments. Bioinformatics. 2020;227:227-50.

[Silvestre 2021] Silvestre, B. S.; Ţîrcă, D. M. Innovations for Sustainable Development: Moving toward a Sustainable Future. J. Clean. Prod. 2019, 208, 325–332.Biocatalysis. Nature Reviews Methods Primers 2021, 1 (1), 1–21.

[Spence 2021] Spence MA, Kaczmarski JA, Saunders JW, Jackson CJ. Ancestral sequence reconstruction for protein engineers. Current opinion in structural biology. 2021;69:131-41.

[Sumbalova 2018] Sumbalova L, Stourac J, Martinek T, Bednar D, Damborsky J. HotSpot Wizard 3.0: web server for automated design of mutations and smart libraries based on sequence input information. Nucleic acids research. 2018 Jul 2;46(W1):W356-62.

[Taverna 2002] Taverna DM, Goldstein RA. Why are proteins marginally stable? Proteins: Structure, Function, and Bioinformatics. 2002;46(1):105-9.

[Tiso 2022] Tiso, T.; Winter, B.; Wei, R.; Hee, J.; de Witt, J.; Wierckx, N.; Quicker, P.; Bornscheuer, U. T.; Bardow, A.; Nogales, J.; Blank, L. M. The Metabolic Potential of Plastics as Biotechnological Carbon Sources - Review and Targets for the Future. Metab. Eng. 2022, 71, 77–98.

[Tratsiak 2013] Tratsiak K, Degtjarik O, Drienovska I, Chrast L, Rezacova P, Kuty M, Chaloupkova R, Damborsky J, Kuta Smatanova I. Crystallographic analysis of new psychrophilic haloalkane dehalogenases: DpcA from Psychrobacter cryohalolentis K5 and DmxA from Marinobacter sp. ELB17. Acta Crystallogr Sect F Struct Biol Cryst Commun. 2013;69(Pt 6):683-8.

[Yang 2019] Yang KK, Wu Z, Arnold FH. Machine-learning-guided directed evolution for protein engineering. Nature methods. 2019 Aug;16(8):687-94.

[Vasina 2022] Vasina M, Vanacek P, Hon J, Kovar D, Faldynova H, Kunka A, Buryska T, Badenhorst CP, Mazurenko S, Bednar D, Stavrakis S. Advanced database mining of efficient haloalkane dehalogenases by sequence and structure bioinformatics and microfluidics. Chem Catalysis. 2022;2(10):2704-25.

[Vasina 2020] Vasina M, Vanacek P, Damborsky J, Prokop Z. Exploration of enzyme diversity: High-throughput techniques for protein production and microscale biochemical characterization. InMethods in Enzymology 2020 (Vol. 643, pp. 51-85). Academic Press.

[Westerbeek 2011] Westerbeek A, Szymanski W, Feringa BL, Janssen DB. Dynamic kinetic resolution process employing haloalkane dehalogenase. Acs Catalysis. 2011 Dec 2;1(12):1654-60.

[Wijma 2013] Wijma HJ, Floor RJ, Janssen DB. Structure-and sequence-analysis inspired engineering of proteins for enhanced thermostability. Current opinion in structural biology. 2013;23(4):588-94.

[Wong 2008] Wong KM, Suchard MA, Huelsenbeck JP. Alignment uncertainty and genomic analysis. Science. 2008;319(5862):473-6.

[Wu 2019] Wu Z, Kan SJ, Lewis RD, Wittmann BJ, Arnold FH. Machine learning-assisted directed protein evolution with combinatorial libraries. Proceedings of the National Academy of Sciences. 2019 Apr 30;116(18):8852-8.

[Wu 2021] Wu, S.; Snajdrova, R.; Moore, J. C.; Baldenius, K.; Bornscheuer, U. T. Biocatalysis: Enzymatic Synthesis for Industrial Applications. *Angew. Chem. Int. Ed Engl.* **2021**, *60* (1), 88–119

[Ziegler 2023] Ziegler C, Martin J, Sinner C, Morcos F. Latent generative landscapes as maps of functional diversity in protein sequence space. Nature Communications. 2023;14(1):2222.

Supplementary Information

Table S1: Overview of characterized DhaA ancestors by the laboratory experiments. DhaA wild type (Uniprot ID P59336) is composed of 294 amino acids. The asterisks indicate the variants that were expressed and experimentally characterized. Color coding is used to indicate type of mutations (indels are in red).

Design name	Mutations compared to the template	Round
Ancestor 5*	7 substitutions, 0 indels	2
Ancestor 7*	18 substitutions, 0 indels	2
Ancestor 8	34 substitutions, 0 indels	2
Ancestor 9*	45 substitutions, 0 indels	1
Ancestor 15	88 substitution, 4 indels	1
Ancestor 30	133 substitution, 4 indels	1
Ancestor 39	132 substitution, 7 indels	1
Ancestor 46	127 substitution, 21 indels	1
Ancestor 79	118 substitution, 103 indels	1
Ancestor 81	125 substitution, 102 indels	1
Ancestor 84	138 substitution, 95 indels	1
Ancestor 96	136 substitution, 109 indels	1

ENSEMBLE ANALYSIS

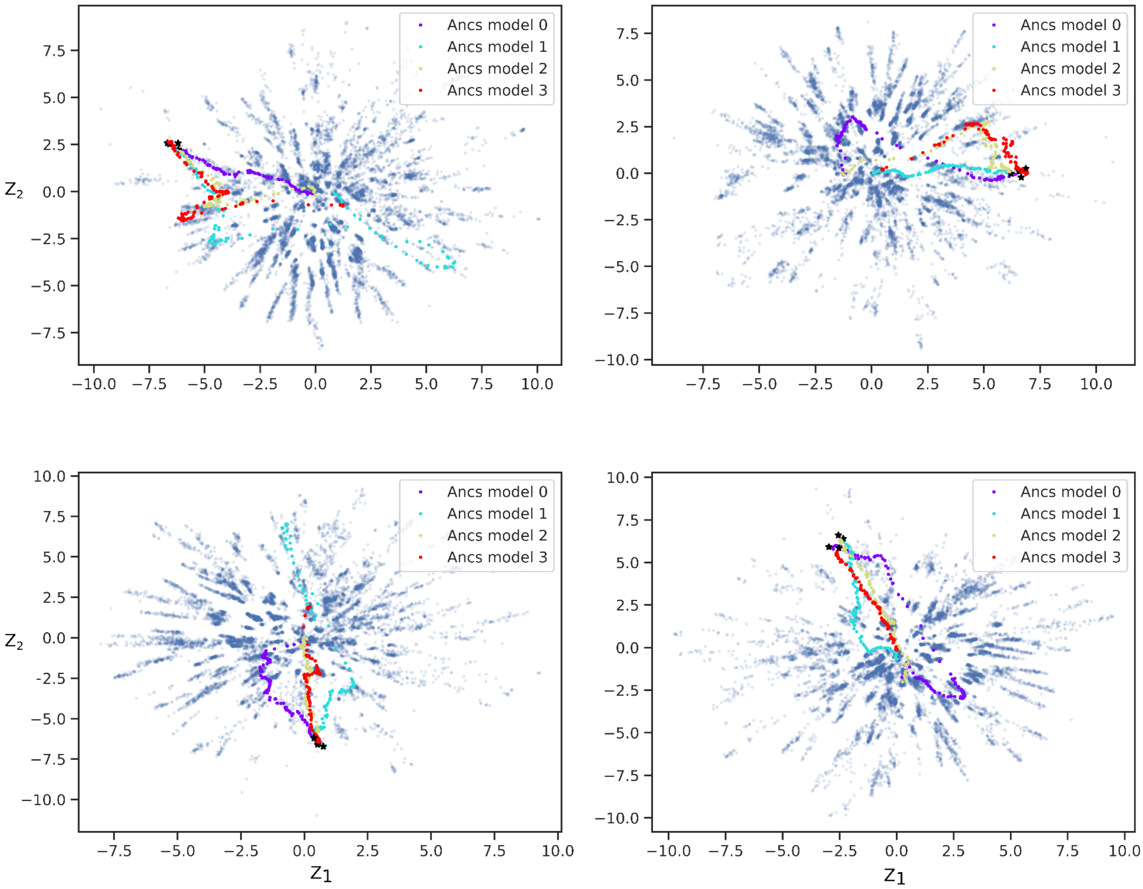


Figure S1: Projection of trajectories onto different latent spaces. Four variational autoencoder (VAE) models were randomly initialized and trained on distinct subsets of the dataset. The dataset itself was divided into five parts, with one part being excluded and the remaining parts used for training. Throughout all scenarios, the query sequence P59336_S14 was included in the training session. Each model generated 100 ancestral sequences using a straight evolutionary strategy, and these sequences were then embedded into the latent space of each respective model. Notably, we observed stability in the sequence embeddings only within the portion of the trajectory near the query embedding. However, discrepancies in the embeddings were observed in the central region, which may correspond to the high entropy of the decoder, as previously noted in [Ziegler 2023]. These discrepancies could potentially serve as a distinct statistical metric in the sequence profile and may be explored further in future studies. Model 0: left top; model 1: right top; model 2: bottom left; model 3: bottom right.

EXPERIMENTAL VALIDATION

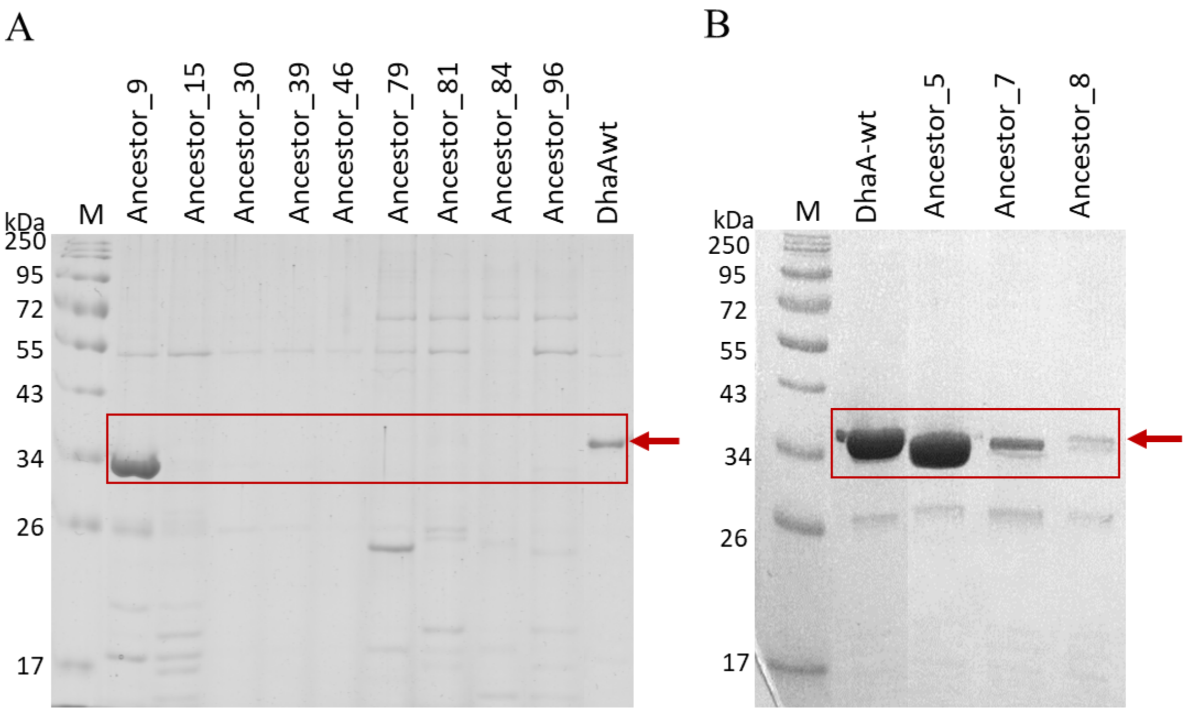


Figure S2: Solubility screening of tested variants: SDS-PAGE gels of affinity-purified Ancestor variants from first round (**A**) and second round of screening (**B**). M indicates markers. DhaAwt is the template protein. The

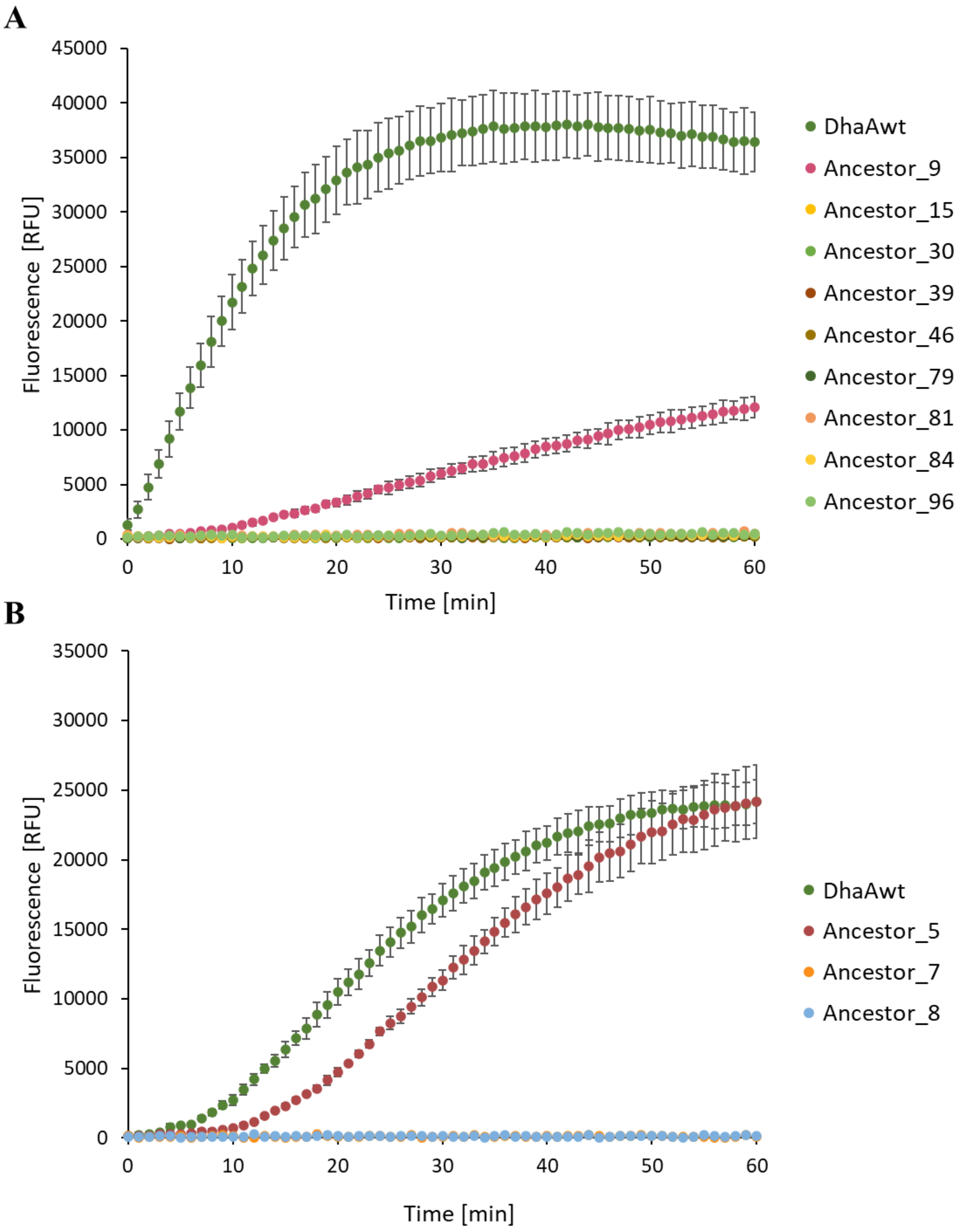


Figure S3: Screening of dehalogenase activity employing HOX assay. A First round of experiments. we identified 9 promising designs for further laboratory experiments to gain deeper insight into the statistical indicators. These designs exhibited a wide range of sequence variability, ranging from 45 substitutions and 0 indels to 138 substitutions and 109 indels. The substitutions covered the entire protein structure and with deletions in most of the cases. Activity was detected only in AncDhaA1(ancestor_9) variant. **B** Second round of experiments. The other more conservative 3 variants (ancestors 5,7, and 8) were selected and examined. Variant

AncDhaA11 (ancestor_5) exhibits dehalogenase activity comparable to positive controls. On the other hand, we were not able to detect any dehalogenase activity in poorly soluble variants ancestor_7 and ancestor_8.

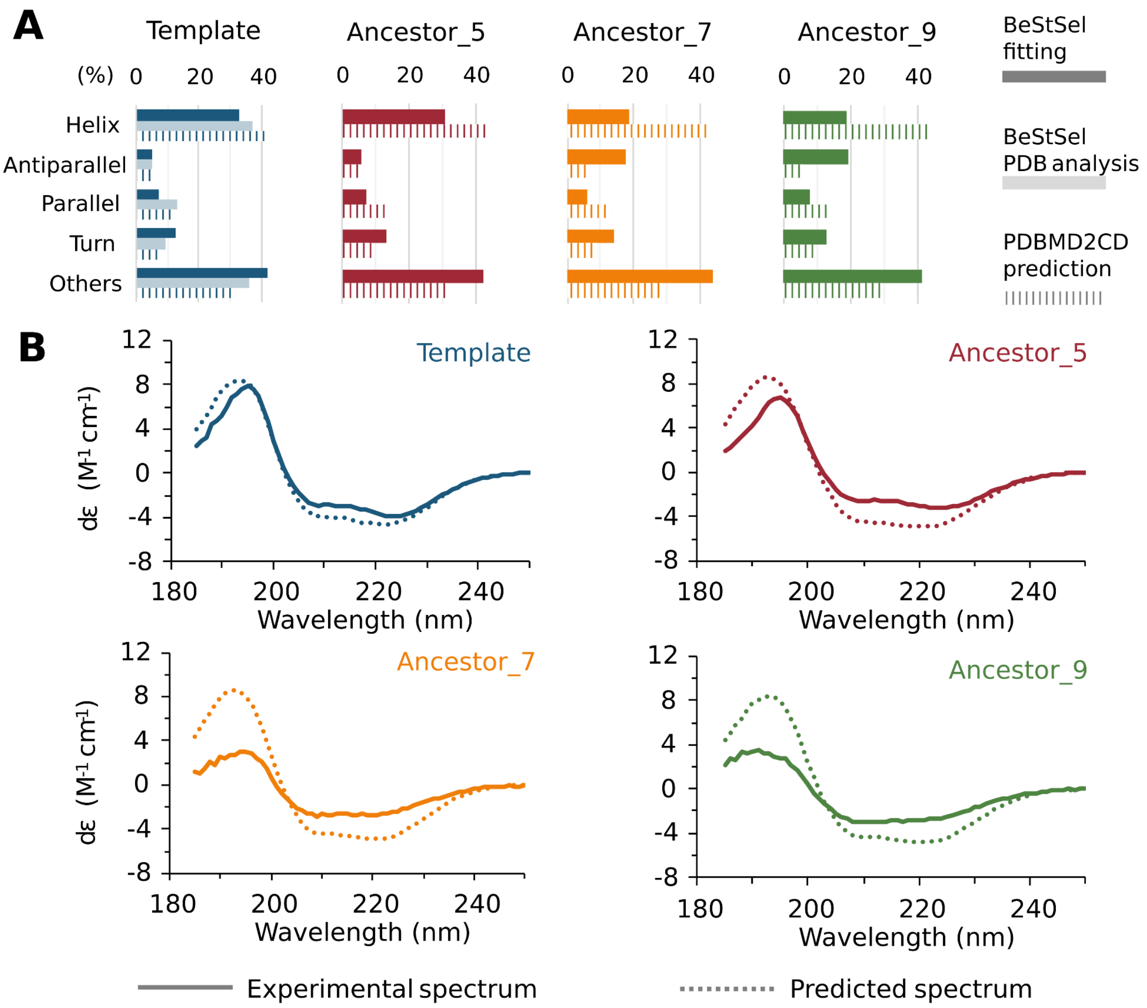


Figure S4: Secondary structure analysis. A The secondary structure composition (in %) of individual variants evaluated by fitting of experimental data (solid fill, full color) or by analysis of PDB structures (solid fill, lighter color, only for Template) by BeStSel [Micsonai 2022] tool, and predicted by the PDBMD2CD [Drew 2020] tool based on PDB structures or AlphaFold models (diagonal stripes). **B** The comparison of predicted (dotted line) and experimentally determined (solid line) CD spectra for individual variants.

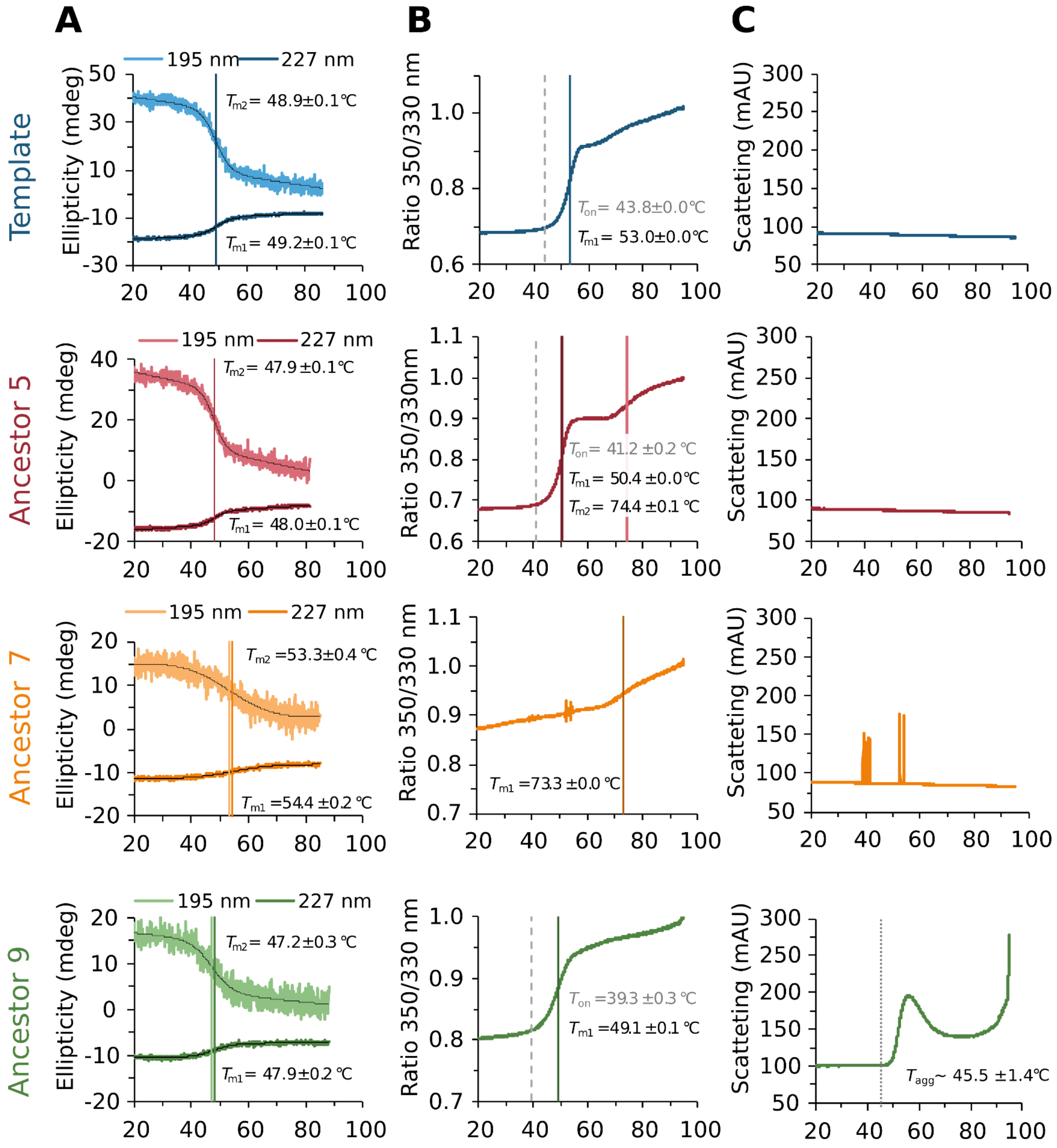


Figure S5: Thermostability analysis. (**A**) Thermal denaturation curves from CD spectroscopy measured at 195 and 227 nm. (**B**) Thermal denaturation curves from nanoDSF spectroscopy. (**C**) Scattering curves from nanoDSF spectroscopy. Relevant thermostability parameters are described within the graphs. Solid, dashed, and dotted lines denote melting temperatures, denaturation onset, and aggregation onset, respectively.

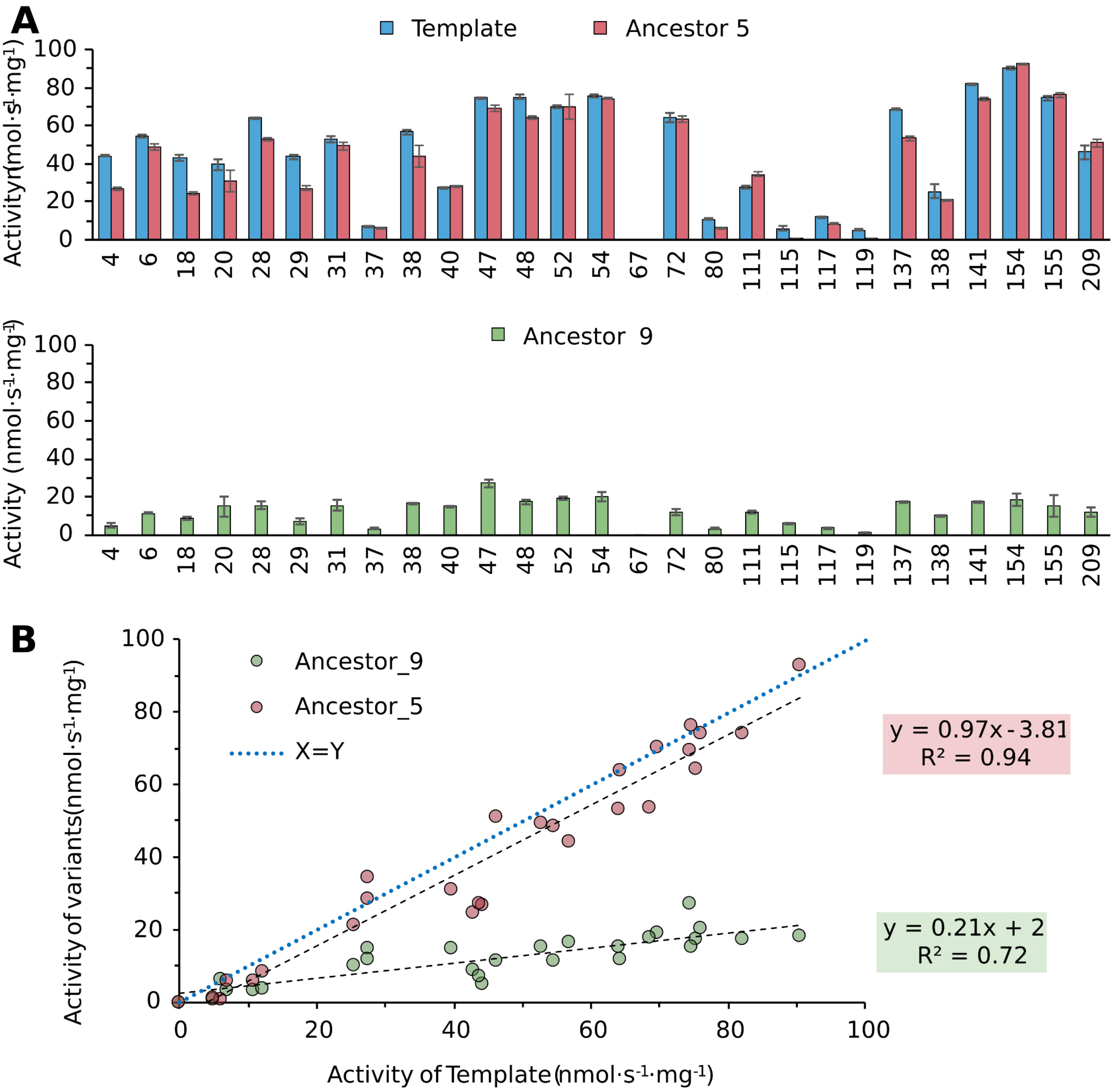


Figure S6: Substrate specificity. (**A**) Substrate specificity profiles for the variants. Numbers on the X-axis denote codes for substrates established previously [Koudelakova 2011]. (**B**) Activities of all variants compared to the activity of Template. The slopes of linear regressions (right) give the estimate of the overall activity in comparison to Template. The theoretical equity of activities is shown by the blue dotted line.

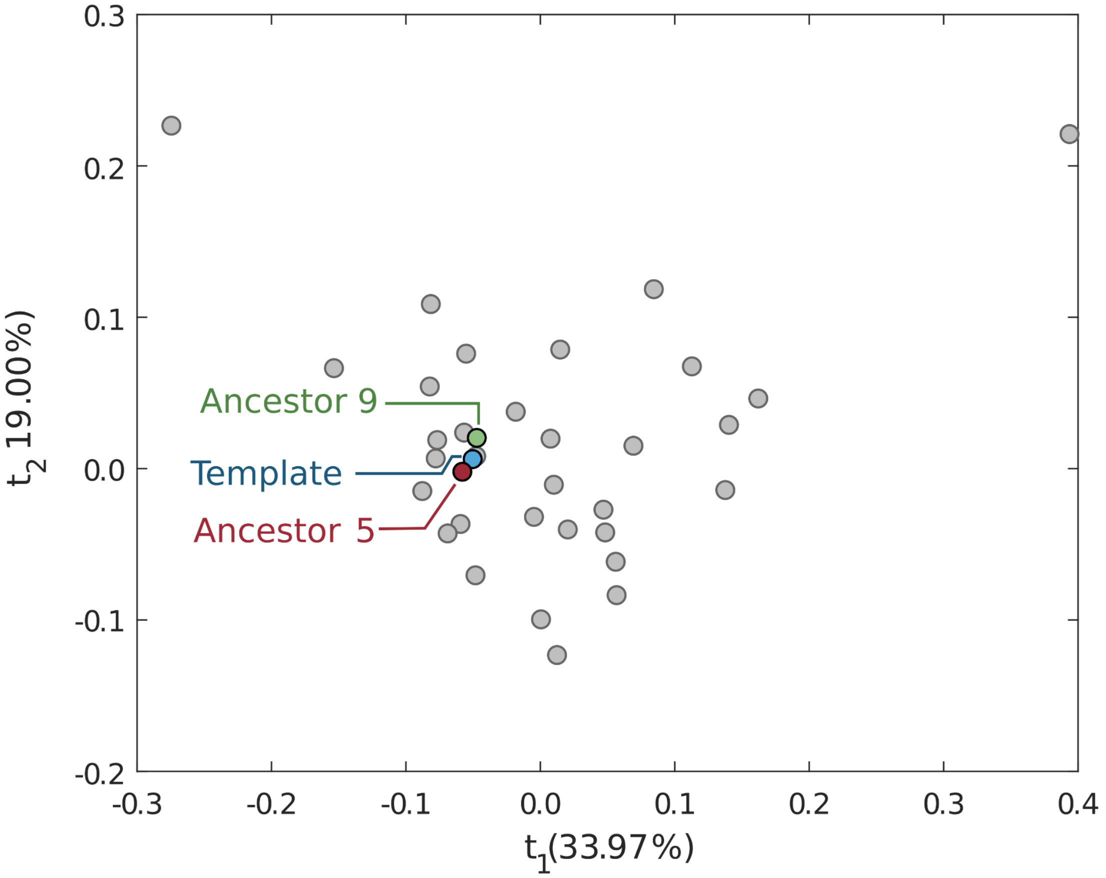


Figure S7: Principal component analysis. The t_1/t_2 score plot (log-transformed data) describes 53.0 % of the variance in the dataset. The benchmark dehalogenases, measured previously [Buryska 2019] [Vasina 2022], are shown and annotated in grey. The newly characterized variants within this study are highlighted.

Supplementary References

[Buryska 2019] Buryska T, Vasina M, Gielen F, Vanacek P, van Vliet L, Jezek J, Pilat Z, Zemanek P, Damborsky J, Hollfelder F, Prokop Z. Controlled oil/water partitioning of hydrophobic substrates extending the bioanalytical applications of droplet-based microfluidics. Analytical chemistry. 2019;91(15):10008-15.

[Drew 2020] Drew ED, Janes RW. PDBMD2CD: providing predicted protein circular dichroism spectra from multiple molecular dynamics-generated protein structures. Nucleic Acids Research. 2020;48(W1):W17-24.

[Koudelakova 2011] Koudelakova T, Chovancova E, Brezovsky J, Monincova M, Fortova A, Jarkovsky J, Damborsky J. Substrate specificity of haloalkane dehalogenases. Biochemical Journal. 2011;435(2):345-54.

[Micsonai 2022] Micsonai A, Moussong E, Wien F, Boros E, Vadászi H, Murvai N, Lee YH, Molnár T, Réfrégiers M, Goto Y, Tantos A. BeStSel: webserver for secondary structure and fold prediction for protein CD spectroscopy. Nucleic Acids Research. 2022;50(W1):W90-8.