Navigating the amino acid sequence space between functional proteins using a deep learning framework

Tristan Bitard-Feildel[1, 2]

[1] Sorbonne Université, CNRS, IBPS, Laboratoire de Biologie Computationnelle et Quantitative, France.

[2] Sorbonne Université, Institut des Sciences du Calcul et de des Données (ISCD), Paris, France.

Corresponding Author:

Tristan Bitard-Feildel [1, 2]

Email address: tristan.bitard-feildel@upmc.fr (previous), tristan@bitardfeildel.fr (current)

Navigating the amino acid sequence space between functional proteins using a deep learning framework

Tristan Bitard-Feildel[1, 2]

[1]**Sorbonne Université, CNRS, IBPS, Laboratoire de Biologie Computationnelle et Quantitative, Paris, France.**

[2]**Sorbonne Université, Institut des Sciences du Calcul et de des Données (ISCD), Paris, France.**

Corresponding author:

Tristan Bitard-Feildel[1,2]

Email address: tristan.bitard-feildel@upmc.fr (previous), tristan@bitardfeildel.fr (current)

ABSTRACT

Motivation: Shedding light on the relationships between protein sequences and functions is a challenging task with many implications in protein evolution, diseases understanding, and protein design. The protein sequence space mapping to specific functions is however hard to comprehend due to its complexity. Generative models help to decipher complex systems thanks to their abilities to learn and recreate data specificity. Applied to proteins, they can capture the sequence patterns associated with functions and point out important relationships between sequence positions. By learning these dependencies between sequences and functions, they can ultimately be used to generate new sequences and navigate through uncharted area of molecular evolution.

Results: This study presents an Adversarial Auto-Encoder (AAE) approached, an unsupervised generative model, to generate new protein sequences.AAEs are tested on three protein families known for their multiple functions : the sulfatase, the HUP and the TPP families. Clustering results on the encoded sequences from the latent space computed by AAEs display high level of homogeneity regarding the protein sequence functions. The study also reports and analyzes for the first time two sampling strategies based on latent space interpolation and latent space arithmetic to generate intermediate protein sequences sharing sequential properties of original sequences linked to known functional properties issued from different families and functions. Generated sequences by interpolation between latent space data points demonstrate the ability of the AAE to generalize and produce meaningful biological sequences from an evolutionary uncharted area of the biological sequence space. Finally, 3D structure models computed by comparative modelling using generated sequences and templates of different sub-families point out to the ability of the latent space arithmetic to successfully transfer protein sequence properties linked to function between different sub-families.

All in all this study confirms the ability of deep learning frameworks to model biological complexity and bring new tools to explore amino acid sequence and functional spaces.

Availability: Code and data used for this study are freely available at https://github.com/T-B-F/aae4seq.

Contact: tristan@bitardfeildel.fr

Supplementary information: Supplementary data are available at online.

INTRODUCTION

Protein diversity, regarding sequence, structure, or function, is the result of a long evolutionary process. Protein fitness and natural selection lead to the current observation of only a fraction of all possible amino acid sequence combinations and therefore, structures, and functions Dryden et al. (2008). These observed sequences are also referred to the amino acid sequence space. The sequence space is difficult to computationally explore due to its huge size and thus the constrains between sequence positions are hard to understand Axe (2004); Luisi et al. (2006); Dryden et al. (2008); Marchi et al. (2019). The classification of amino acid sequences into protein domain families allows to organize the sequence space and reduce

47   its complexity.

48

49   Many resources have been developed over the years to group amino acid sequences into families
50   whose members share sequence and structural similarities Dawson et al. (2016); Pandurangan et al. (2018);
51   El-Gebali et al. (2018). Thus, protein families permit to organize the sequence space. The sequence
52   space area between these families is however mostly uncharted Das et al. (2015) in spite of very remote
53   evolutionary relationships between families Alva et al. (2010). Navigating the sequence space with respect
54   to the functional diversity of a family is therefore a difficult task. This difficulty is even increased by the
55   low number of proteins with experimentally confirmed function. In this regard, computer models are
56   needed to explore the relationships between sequence space and functional space of the protein families
57   Goldstein and Pollock (2016); Tian et al. (2018); Copp et al. (2018); Salinas and Ranganathan (2018);
58   Tubiana et al. (2019); Poelwijk et al. (2019); Russ et al. (2020). Perfect modeling of the sequence space
59   could have applications in molecular engineering, functional annotation, or evolutionary biology. It may
60   for example be possible to understand completely the relationships between amino acid positions of a
61   family responsible of a molecular function or to navigate the sequence space between families of different
62   functions.
63   In this study, tools and strategies based on an unsupervised deep learning approach are proposed to
64   model and navigate the current evolutionary uncharted area of the amino acid sequence space.

65

66   Previous deep learning generative models such as variational autoencoders (VAE) have been applied
67   on biological and chemical data. They have for example been used to explore and classify gene expression
68   in single-cell transcriptomics data Lopez et al. (2018), or to explore the chemical space of small molecules
69   for drug discovery and design Rampasek et al. (2017); Gómez-Bombarelli et al. (2018). Their ability to
70   reduce input data complexity in a latent space and perform inference on this reduced representation make
71   them highly suitable to model, in an unsupervised manner, complex systems. Regarding protein science,
72   VAE have been able to accurately model amino acid sequence and functional spaces Sinai et al. (2017),
73   to predict mutational impact Hopf et al. (2017); Riesselman et al. (2018), to decipher protein evolution
74   and fitness landscape Ding et al. (2019) or to design new proteins Greener et al. (2018). In this study,
75   Adversarial AutoEncoder (AAE) network Makhzani et al. (2015) is proposed as a new and efficient way
76   to represent and navigate the functional space of a protein family. Unlike VAE, AAE networks constrain
77   the latent space over a prior distribution. Prior distribution allows better inference to explore the whole
78   latent distribution which is particularly useful for travelling through uncharted areaKadurin et al. (2017).
79   Like VAE and other autoencoder architectures, AAEs reduce high dimensional data by projection,
80   using an encoder, into a lower dimensional space. This space is known as a latent space, or embedding
81   representation. The latent space can, in turn, be used by the decoder to reconstruct the initial data. AAE
82   Makhzani et al. (2015) architecture corresponds to a probabilistic autoencoder but with a constraint on the
83   latent space of the encoder. The latent space is constrained to follow a defined prior distribution. This
84   constraint is applied using a generative adversarial network (GAN) Goodfellow et al. (2014) trained to
85   discriminate between the latent space and the prior distribution. It ensures that meaningful samples can
86   be generated from anywhere in the latent space defined by the prior distribution. Applied to biological
87   sequences of a protein domain family, it is then possible to encode the sequence diversity to any prior
88   distribution. Thus, the model is able to sample and generate new amino acid sequences of the family
89   from any point of the prior distribution. Ideally, the learned latent space should be representative of
90   the functions of the protein domain family and even able to dissociate protein sequences with different
91   sub-functions.

92

93   Protein sequences can cluster in the latent space of AAE network. These clusters were analyzed to
94   verify their ability to group sequences according to function as observed with VAE networks. Three
95   protein families including different sub-families were used to train AAE models. The protein functional
96   annotations of these families were used to analyze the clustered sequences. The three different protein
97   families selected were the sulfatases, the HUP (HIGH-signature proteins, UspA, and PP-ATPase) and the
98   TPP (Thiamin diphosphate (ThDP)-binding fold, Pyr/PP domains) families. The sulfatases are a group of
99   proteins acting on sulfated biomolecules. This family have been manually curated into sub-family with
100   specific functions according to substrate specificity Barbeyron et al. (2016). They are found in various
101   protein family databases, such as in Pfam (PF00884). The SulfAtlas database Barbeyron et al. (2016) is a

102 collection of curated sulfatases centered on the classification of their substrate specificity. The majority of
103 Sulfatases (30,726 over 35,090 Version 1.1 September 2017) is found in family S1 and is sub-divided
104 into 73 sub-families corresponding to different substrate specificities. Sub-families S1-0 to S1-12 possess
105 proteins with experimentally characterized EC identifiers.

106   The two other protein families, HUP and TPP families are not manually curated but were selected as
107 they are known to have multiple functions Das et al. (2015). Proteins of the HUP family are a very diverse
108 group with functions linked to particular motifs such as HIGH and KMSKS (nucleotidyl transferases
109 and t-RNA synthetases activities), ATP PyroPhosphatase motif, or sequence motifs responsible of the
110 hydrolysis of the alpha-beta phosphate bond of ATP Bork and Koonin (1994); Wolf et al. (1999); Aravind
111 et al. (2002). The TPP family is made of very similar protein domains which are probably evolutionary
112 related Muller et al. (1993); Berthold et al. (2005). They have pyruvate dehydrogenases, decarboxylate,
113 and binding functions Muller et al. (1993).

114

115   The VAE architecture has previously been used to cluster protein sequences and interpret the resulting
116 clusters regarding their function or evolutionary history Sinai et al. (2017); Hopf et al. (2017); Riesselman
117 et al. (2018); Ding et al. (2019); Greener et al. (2018). These experiments have not studied the quality of
118 the architecture generative ability for protein sequences. Particularly, the performances of the architecture
119 is not known for the tasks of navigating the sequence space and transferring features between clusters. In
120 this study, two experiments were carried out in this direction using latent space interpolation and latent
121 space arithmetic operations. These experiments were designed as new tools and frameworks for the amino
122 acid sequence space exploration.

123   Data point interpolations between protein sequences of different sulfatase sub-families was used to
124 analyze the latent space coverage of the protein domain family functional space. The interpolated data
125 points correspond therefore to unseen proteins, *i.e.* evolutionary uncharted area between groups of amino
126 acid sequences. A good model should be able to produce realistic protein sequences from these data
127 points.

128   This study also explored arithmetic operations with protein sequences encoded in their latent space to
129 generate new protein sequences. Arithmetic operations on latent space have previously been reported
130 to transfer features between images of different classes Radford et al. (2015). These operations may
131 therefore have interesting potential for molecular design and for exploration of the amino acid sequence
132 space. Four different strategies were explored to combine latent spaces of different sulfatase sub-families.
133 The generated proteins from the combined latent spaces were analysed in term of sequences and structures,
134 after being built by comparative modelling.

METHODS

Protein Families

The sulfatase family.   An initial seed protein multiple sequence alignment (MSA) was computed from
sequences of the protein structures of SulfAtlas Barbeyron et al. (2016) database sub-families 1 to 12.
This seed was used to search for homologous sequences on the UniRef90 Suzek et al. (2014) protein
sequence database using hmmsearch Eddy (2011) with reporting and inclusion e-values set at $1e^{-3}$.
A label was assigned to each retrieved protein if the protein belonged to one of the 12 known sub-families.
The MSA computed with hmmsearch was filtered to remove columns and sequences with more than 90%
and 75% gap characters respectively. Proteins with multiple hits on different parts of their sequences
were also merged into a single entry. From 105181 initial protein sequences retrieved by hmmsearch, the
filtering steps led to a final set of 41901 proteins.

HUP and TPP protein families.   A similar protocol was followed for the HUP and TPP protein families.
Instead of using an initial seed alignment made of sequences with known 3D structures, the CATH protein
domain HMM Orengo et al. (1997); Sillitoe et al. (2018) was used to search for homologous sequences
in the UniRef90 database. CATH models 3.40.50.620 and 3.40.50.970 correspond to the HUP and TPP
protein families, respectively. A sequence filtering pipeline identical to the one used for the sulfatase
family was applied to each of the resulting MSAs.
The final numbers of proteins in each dataset were: 25041 for the HUP family (32590 proteins before
filtering) and 33693 for the TPP family (133701 before filtering).

Deep learning model

Generative Adversarial Network.   A complete description of Generative Adversarial Network can be found in Goodfellow et al. (2014). To summarize, the GAN framework corresponds to a min-max adversarial game between two neural networks: a generator (G) and a discriminator (D). The discriminator computes the probability that an input x corresponds to a real point in the data space rather than coming from a sampling of the generator. Concurrently, the generator maps samples z from prior $p(z)$ to the data space with the objective to confuse the discriminator. This game between the generator and discriminator can be expressed as:

$$min_G \ max_D \ E_{x \sim p_{data}}[\log D(x)] + E_{z \sim p(z)}[log(1 - D(G(z)))] \tag{1}$$

Adversarial auto-encoder.   Adversarial autoencoders (AAEs) were introduced by Makhzani et al. (2015). The proposed model was constructed using an encoder, a decoder networks, and a GAN network to match the posterior distribution of the encoded vector with an arbitrary prior distribution. Thus, the decoder of the AAE learns from the full space of the prior distribution. A Gaussian prior distribution was used in this study to compute the aggregated posterior $q(z|x)$ (the encoding distribution). The mean and variance of this distribution was learned by the encoder network: $z_i \sim N(\mu_i(x), \sigma_i(x))$. The re-parameterization trick introduced by Kingma and Welling (2014) was used for back-propagation through the encoder network.

Three different architectures were evaluated. The general architecture was as follows (see Table S1 and Fig. S1 for a representation of architecture number 3). The encoder was made of one or two 1D convolutional layers with 32 filters of size 7 and a stride of length 2, and one or two densely connected layers of 256 or 512 units. The output of the last layer was passed through two stacked densely connected layers of hidden size units to evaluate μ and σ of the re-parameterization trick Kingma and Welling (2014).

The decoder was made of two or three densely connected layers of the length of the sequence family time alphabet units for the last layers and of 256 or 512 units for the first or the two first layers. The final output of the decoder was reshaped to match the input shape. A softmax activation function was applied, corresponding to the amino acid probabilities at each position. To convert the probability matrix of the decoder into a sequence, a random sampling according to the probability output was performed at each position. The selected amino acid at a given position was therefore not necessarily the amino acid with the highest probability but reflect the biological distributions. The discriminator network was made of two or three densely connected layers. The last layer had only one unit and corresponds to the discriminator classification decision using a sigmoid activation function.

Model training   The network was trained for each protein family independently. Amino acids and gap symbol of sequence input data were transformed using one-hot-encoding. A batch size of 32 was used to train the network. The autoencoder was trained using a categorical cross-entropy loss function between the input data and the predicted sequences by the autoencoder. The discriminator was trained using binary cross-entropy loss function between the input data encoded and the samples from the prior distribution.

Generated sequences and structures analyses

Dimensionality reduction.   The AAE model can be used to reduce the dimensionality of the sequence space by setting a small latent size. Two dimensionality reductions were tested with latent size of 2 and 100. Latent size of 2 can be easily visualized and a larger latent size of 100 should represent the input data more efficiently as more information can be stored.

Clustering.   HDBSCAN Campello et al. (2013); McInnes and Healy (2017) was used to cluster the sequences in the latent space due to its capacity to handle clusters of different sizes and densities and its performances in high dimensional space. The Euclidean distance metric was used to compute distances between points of the latent space. A minimal cluster size of 60 was set to consider a group as a cluster as the number of protein sequences is rather large. The minimal number of samples in a neighborhood to consider a point as a core point was set to 15 to maintain relatively conservative clusters.

Functional and taxonomic analyses.   Enzyme functional annotation (EC ids) and NCBI taxonomic identifiers were extracted when available from the Gene Ontology Annotation portal (January 2019) using

197 the UniProt-GOA mapping Huntley et al. (2014). Proteins without annotation were discarded.

198 The annotation homogeneity was computed for each cluster. Considering a cluster, the number of different

199 EC ids and taxonomic ids were retrieved. The percentage of each EC id (taxonomic id) was computed

200 by cluster. An EC id (taxonomic id) of a cluster with a value of 90% indicates that 90% of the cluster

201 members have this EC id (taxonomic id). A cluster with a high homogeneity value corresponds to

202 functionally or evolutionary related sequences.

203 Homogeneous clusters will point out the ability of the AAE model to capture and distinguish protein

204 sequences with functionally or evolutionary relevant features without supervision.

205 **Latent space interpolation.**   Twenty pairs of protein sequences were randomly chosen between all

206 combinations of sulfatases sub-families with at least 100 labeled members but with less than 1000

207 members (to avoid pronounced imbalance between classes): S1-0 (308 proteins), S1-2 (462 proteins),

208 S1-3 (186 proteins), S1-7 (741 proteins), S1-8 (290 proteins) and S1-11 (669 proteins). The coordinates

209 of the selected sequences in the encoded latent space with 100 dimensions were retrieved. Spherical

210 interpolations using 50 steps were performed between the pairs. Spherical interpolation has previously

211 been reported to provide better interpolation for the generation of images White (2016). The interpolated

212 points were given to the decoder to generate new sequences. Statistical analyses were carried out on the

213 sequence transition from one family to an other. A model able to learn a generalized latent space should

214 generate new sequences with smooth transitions between families. Analyses at the amino acid level were

215 also performed on the interpolated sequences of two Sulfatase sub-families encoded far from one-another

216 in the latent space.

217 **Shannon entropy computation.**   Shannon entropy is computed to measure the degree of variability at

218 each position (column) of the MSA Jost (2006).

$$H(X) = -\sum_{i=1}^{n} P_i \log P_i \tag{2}$$

219 with P_i the frequency of symbol i and n the number of characters (20 amino acids and a gap symbol). The

220 mean entropy per amino acid is computed for MSAs of biological sequences and generated sequences.

221 Low entropy indicates that the analyzed sequences have low amino acid variability between each other.

222 High entropy indicates high amino acid variability.

223 **Latent space arithmetic.**   It is possible to transfer features between data such as images by using

224 subtraction or addition between projected data into a latent space Radford et al. (2015). This latent space

225 property was tested on seven Sulfatase sub-families (S1-0, S1-1, S1-2, S1-3, S1-7, S1-8 and S1-11)

226 selected on the basis of their number of protein sequences. Different arithmetic strategies (Fig. S2)

227 were tested between latent spaces. The sub-family whose features are transferred is named the source

228 sub-family. The sub-family receiving the transferred feature is named the query sub-family.

229 A first strategy consists in the addition of the mean latent space of the source sub-family to the encoded

230 sequences of the the query sub-family. The second strategy differs from the first one by subtracting from

231 the mean background latent space of all sub-families the latent space of the query sub-family. The third

232 strategy differs from the second by the mean background strategy being computed using all sub-families

233 except the source and query sub-families. Finally, in the fourth strategy, the subtraction is performed

234 using a local KD-tree to only remove features shared by the closest members of a given query and the

235 addition is performed by randomly selecting a member of the source family and its closest 10 members.

236 For each strategy, new sequences were generated using the latent spaces of all query proteins in the

237 sub-families. The generated sequences by latent space arithmetic are compared to the initial query and

238 source sub-families in terms of sequence and structural properties.

239 The protein sequence similarities were computed between the generated sequences by latent space arith-

240 metic and the biological sequences of the two initial sub-families using a Blosum 62 substitution matrix.

241 The sequence similarities were also computed inside a sub-family, between sub-families, and between

242 generated sequences. The distributions of sequence similarities allow to explore the abilities of the latent

243 space arithmetic operations and of the decoder to produce meaningful intermediate protein sequences

244 from data points not corresponding to biological sequences. These data points correspond to an uncharted

245 sequence space.

Protein structural models were computed using the structures of the initial sub-families as templates for MODELLER Webb and Sali (2014) and evaluated using the DOPE score Shen and Sali (2006). Models were computed using the generated sequences by latent space arithmetic on template structures from their source and query sub-families. The DOPE energies of the modeled structures were compared to structural models computed as references. The first structural model references were computed using the sequences and template structures belonging to the same sub-families, which should provide the best DOPE energies. The second structural model references were computed using the sequences and template structure belonging to different sub-families (ex: sequences from source sub-family and template structures from the query sub-family or inversely, sequences from query sub-family and template structures from the source sub-family), which should provide the worst DOPE energy. If the generated sequences by latent space arithmetic correspond to intermediate proteins with properties from two sub-families, they should have intermediate DOPE energies when compared to the others evaluated models.

Glossary

The following glossary defines the different terms and techniques used in this manuscript.

Adversarial Auto Encoder. A neural network architecture used for generative tasks. The architecture combine an auto-encoder and a generative adversarial network.

Encoder. Part of the AAE used to project the input data to a latent space.

Latent space. Input data point representation in a lower dimension.

Decoder. Part of the AAE used to reconstruct the input data from the latent space.

Query sub-family. Starting sub-family in interpolation experiments. Sub-family whose individual sequence latent spaces have been used in combination with the mean latent space of sequences from a source sub-family in latent space arithmetic strategies.

Source sub-family. Sub-family whose mean latent space has been used in combination with individual sequence latent space of a query sub-family in latent space arithmetic strategies.

Target sub-family. sub-family used as end point in interpolation experiments.

RESULTS

A structurally constrained MSA was computed using Expresso from T-Coffee webserver Armougom et al. (2006); Di Tommaso et al. (2011) between sequences of S1 sulfatases structures. This MSA was processed into a Hidden Markov Model and hmmsearch was used to retrieve aligned sequence matches against the UniRef90 sequence database. A total of 76,427 protein sequence hits were found to match the sulfatase HMM in UniRef90. The sequences were filtered to remove columns and hits with more than 90% and 75%, respectively, of gap characters. The final MSA comprised 41,901 sequences. The sulfatases protein dataset was separated into a training, a validation, and a test sets with a split ratio of: 0.8, 0.1, and 0.1.

The three different AAE architectures (see Method section) were trained on the training set and evaluated on the validation set. The test set was only used on the final selected architecture. Models were evaluated by computing top k-accuracy, corresponding to the generation of the correct amino acid in the first k amino acids. Table S2 shows the top k accuracy metric for k=1 and k=3 computed for the different AAEs. The accuracy scores scaled down with the number of parameters, but without any large difference. The architecture with the fewest number of parameters (architecture 3) was therefore selected to avoid over-fitting the data. The final accuracy scores on the test set were computed and were similar to the values observed during the model training: 62.5% and 80.2% (k=1 and k=3). The selected architecture was separately trained using the protein sequences of the HUP and TPP families with identical train, validation, and test set splits.

Latent space projection

AAE can be used as a dimensional reduction and visualization techniques by fixing the dimension of the latent space to two or three for plotting purpose. In this section, AAE network ability to create meaningful projection is tested on Sulfatase, HUP and TPP families by clustering and analysing protein sequences in terms of enzymatic activity and phylogenetic diversity.

Starting from the final MSA of the Sulfatase family, an AAE network was trained to project the sequences in a latent space with two dimensions. A PCA of the MSA was computed for comparison purpose with the AAE projection using the PCA first two principal components.
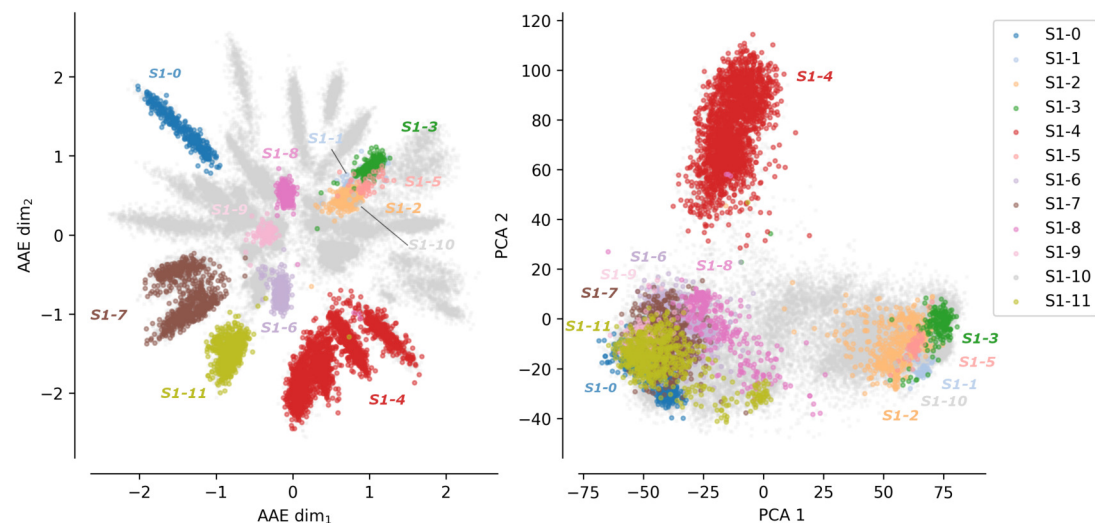
Figure 1. Projections of the MSA sequences of the SulfAtlas family. Left: projection in the encoding learned using an AAE (number of latent dimensions: 2). Right: projection using a PCA (two first components). Gray data points correspond to protein sequences not part of the curated 12 sub-families. This analysis is also performed for the HUP and TPP families. Results can be found in Fig. S3.

Figure 1 shows the protein sequences encoded by the AAE and the PCA projection. Each dot corresponds to a protein sequence. The dots are colored according to their sub-family. Gray dots correspond to protein sequences not belonging to any of the 12 curated sulfatases sub-families. The AAE displays in this figure a better disentanglement of the S1 family sequence and functional spaces than the PCA. Well-separated gray spikes can also be observed in the AAE projection.These spikes may correspond to groups of enzymes sharing common substrate specificity but not yet experimentally characterized.

In some cases, sub-families with identical functions are projected closely on the encoded space. For instance, sub-families S1-6 (light magenta) and S1-11 (yellow) have both the EC 3.1.6.14 activity (N-acetylglucosamine-6-sulfatase) and are closely located in the encoded space. Moreover, some sub-family projections appear entangled such as the S1-1 sub-family (light blue, Cerebroside sulfatase activity, EC 3.1.6.8), the S1-2 (orange) and the S1-3 (green) sub-families (Steryl-sulfatase activity, EC 3.1.6.2), the S1-5 (pink) sub-family (N-acetylgalactosamine-6-sulfatase activity, EC 3.1.6.4), and the S1-10 (gray) sub-family (Glucosinolates sulfatase activity EC 3.1.6.-). The five families correspond to four different functions but are made of Eukaryotic protein sequences only and their entanglement may be due to their shared common evolutionary history. This separation based on the sequence kingdoms can clearly be visualized in the PCA projections with Eukaryotic sequences on the right side on sub-families with a majority of Bacteria sequences on the left side. The PCA projections failled to finely separate protein sub-families based on their functions. The example of protein B6QLZ0_PENMQ is also interesting. The protein is projected (yellow dot corresponding to the S1-11 sub-family) at coordinates (0.733, -1.289), inside the space of the S1-4 family (red). This may look like an error but a closer inspection shows that this protein is part of both the S1-4 and S1-11 sub-families of the SulfAtlas database.

Projections of sequences into latent spaces using AAE with two dimensions were also tested on the

HUP and TPP families. The AAE projections can be visualized on Fig. S3. There are fewer functional annotations for these two families than for the sulfatase family. A strong separation can however clearly be observed between the major functions of the two families .

Latent spaces were evaluated for each protein family based on enzyme classification (EC) and taxonomic homogeneity. Given a set of protein sequences, the encoded sequences in a latent space of a 100 dimensions were clustered using HDBSCAN.
For the sulfatase family, 27 clusters were found, for which taxonomic and EC annotations could be extracted (Fig. S4 and Table S3). All these clusters displayed either strong taxonomic or EC homogeneity. Enzymatic homogeneity was higher than taxonomic homogeneity for 16 clusters, found equal in one cluster and lower for 10 clusters.
In the HUP family, all clusters had very high EC homogeneity (Table S4). Only two clusters out of 47 could be found with higher taxonomic homogeneity than EC homogeneity. For these two clusters enzymatic homogeneity values were high and only marginally different (cluster 5, taxonomic homogeneity of 100% an EC homogeneity of 99% and cluster 31, taxonomic homogeneity of 99 % and EC homogeneity of 97%). Five clusters were found with equal taxonomic and EC homogeneity.
In the TPP family all clusters had also very high EC homogeneity (Table S5). Five clusters out of 51 could be found with higher taxonomic homogeneity than EC homogeneity. For these 5 clusters the differences between taxonomic homogeneity and EC homogeneity were higher than the differences observed for the HUP clusters. Six clusters were found with equal taxonomic and EC homogeneity.

The differences between the AAE and the PCA projections together with the general cluster enzymatic homogeneity highlight the ability of the encoding space to capture amino acid functional properties.

Protein latent space interpolation

Interpolation between encoded sequences can be used to "navigate" between proteins of two sub-families. After the selection of a query sub-family, the sequences of this sub-family are projected to the latent space and used as the starting points of the interpolation. The end points of the interpolation correspond to sequences of a target sub-family, different from the query sub-family, and projected into the latent space. Twenty pairs of protein sequences were randomly selected between all combinations of protein sub-families to test the capacity of the encoded space and 50 intermediates, *i.e. interpolated,* data points were generated between each query / target pair. The sequence similarities were computed between the generated protein sequences from the interpolated latent space and the biological query and target protein sequences of the sub-families. It is thus possible to measure the amino acid sequence drift from one protein to another one.

The observed amino acid transitions from the query sub-family to the target sub-family are very smooth for all combinations of sub-families. The sequence similarity distributions display a logistic function shape as shown in Fig. 2-A. The smooth transition between points demonstrates the ability of AAE network to encode the sequences into a smooth latent space and thus to correctly "fill" the gap between projected protein sequence sub-families.

The Shannon entropy was computed for each group of sequences: interpolated sequences between query and target sub-families, sequences of the query sub-families, and sequences of the target sub-families. Figure 2-B shows the Shannon entropy distribution for the S1-0 and S1-11 sequences and their interpolated sequences. Interestingly, the figure shows lower entropy for interpolated sequences than for original sequences. Lower entropy indicates fewer amino acid variation at each position of the interpolated sequences than in biological sequences. Fewer amino acid variation at each position for the interpolated sequences could corresponds to restricted paths to travel between sub-families. This trend is true for all interpolated sequences between all sub-families as reported in the Table S6. This is in agreement with molecular evolution theory and experiments that describe protein families as basins in fitness landscape Bornberg-Bauer and Chan (1999); Sikosek et al. (2012); Boucher et al. (2014).

A closer inspection of interpolations between sub-families S1-0 and S1-4 (respectively blue and red data points in Fig. 1) was also performed to study changes at the amino acid level. The two sub-families are in "opposite" spaces in the two-dimensional projection. It can be observed in Fig. S5 that gapped
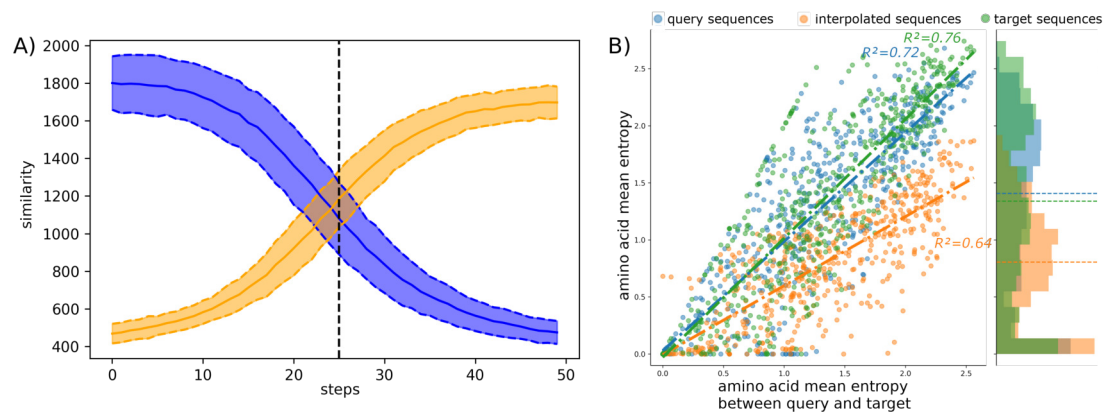
Figure 2. Interpolation analyses between sub-families S1-0 and S1-11. A) Sequence similarity distributions (sum of blosum weights, higher the score higher the similarity) between interpolated sequences and query proteins (blue) or target proteins (orange). B) Distribution of amino acid Shannon entropy for interpolated sequences (orange, $R^2 = 0.64$) between sub-families S1-0 (blue, $R^2 = 0.72$) and S1-11 (green, $R^2 = 0.76$) over the amino acid mean Shannon entropy of query and target sub-families.

378 area found in the query sequence but not in the target sequence (and inversely) are progressively filled (or
379 broken down) starting from the flanking amino acids to the center of the gap (or inversely from the center
380 to the flanking amino acids). This indicates an organized and progressive accumulation of amino acids
381 (or gaps) that extend (or shrink) the region of the sequence previously without (or with) residues. For
382 instance gap reduction can be observed in the generated sequences between sequence ID 2 of the sulfatase
383 S1-0 family (query) and sequence ID 2196 of the sulfatase S1-4 family (target) at positions 75 to 86.
384 Moreover, family-specific amino acids are progressively replaced in key positions. In the previous
385 interpolation between query and target sequences, it can thus be observed at positions 21 and 22 of the
386 MSA a replacement of residues S and C by G and A (Fig. S5). Most transitions are not abrupt, and do not
387 occur at the 25th-generated intermediate sequences but are smooth and correspond to plausible sequences.
388 The ability of the AAE to generate interpolated sequences with emerging or disappearing features of two
389 sub-families, highlights its capacity to generalize the decoding of latent space points not corresponding to
390 encoded sequences and thus never observed during training, and outside the structured organization of the
391 computed latent space.

392 **Protein latent space arithmetic**
393 Latent space arithmetic is able to transfer learned features between different classes Radford et al. (2015).
394 If applied to protein sequence latent space, this technique could permit to transfer features such as
395 enzymatic activity or part of structure between protein families. To test this technique different arithmetic
396 strategies (see Methods and Fig. S2) were tested between latent spaces of two Sulfatase sub-families.After
397 performing the arithmetic operation between latent space coordinates, the protein sequences corresponding
398 to the new coordinates were generated by the decoder. Protein structures of the generated sequences
399 were computed using homology modeling The structure templates correspond to both sub-families. The
400 protein structures of the generated sequences were compared to computed models using sequences and
401 structures of the same sub-families and using sequences from one sub-family and structure templates
402 from the other one. The Sulfatases sub-families S1-0, S1-2, S1-3, S1-7, S1-8 and S1-11 were chosen to
403 test this technique.
404 In the following section, the terminology $\widehat{\text{Seq}}$. S1-XmY will correspond to a generated sequence using a
405 combination of the mean latent space of the sub-family S1-Y added to the latent space of the sub-family
406 S1-X. The X and Y sub-families will be referred to as the query and source sub-families.

407
408 First, two Prosite motifs of the Sulfatase family are analyzed from generated and original sequences.
409 Figure S6 displays logo plots of two regions corresponding to Prosite motifs PS00523 and PS00149 to
410 illustrate the amino acid content of the generated protein sequences by latent space arithmetic. These
411 regions correspond to the most conserved regions of the sulfatase family and have been proposed as

412 signature patterns for all the sulfatases in the Prosite database.

413 Different amino acid patterns can be observed between the sequence groups that can be classified as
414 "competition", "taking over", or "balanced" pattern. A competition pattern of amino acids corresponds
415 to equivalent frequency of two different amino acids in the generated sequences. A taking over pattern
416 corresponds to an amino acid of one of the original sequences being the most frequent in the generated
417 sequences. A balanced pattern corresponds to a maintained equilibrium between amino acids in the
418 generated sequences. Some other positions are displaying much more complex patterns and cannot be
419 summarized as a frequency competition between source and query sub-families. These behaviors can be
420 observed several times through the logo plots but are still position-specific, meaning that the bits scores
421 pattern observed in the source sub-families (Panels A and D of Fig. S6) do not necessary allow to predict
422 the amino acids bits scores in the generated sequences (Panels B and C of Fig. S6).

423

424 Protein sequence similarities were computed to evaluate the diversity of the generated sequences
425 and compare their diversity with the original sub-families. Protein sequence similarities were computed
426 between : the generated sequences, the sequences of a sulfatase sub-family used to generate protein
427 sequences, the generated sequences and their query sulfatase sub-family, the generated sequences and their
428 source sulfatase sub-family, the query and source sequences of sulfatase sub-families. Figure 3 shows the
429 mean and variance distribution of computed protein sequence similarities between these different groups
430 for generated sequences computed using the first strategy. The first, second, and third strategies display a
431 similar pattern and their corresponding figures are available in the Supplementary Information (Fig. S10,
432 S12 and S14).

433 Protein sequence similarities between different sub-families (red upper triangles) have lower similarity
434 scores and lower variances than the other distributions. Protein sequence similarities between sequences
435 of a sub-family (blue circles) have the highest mean and variance values observed. However, since only 6
436 sub-families were kept for analysis (sub-families 0, 2, 3, 7, 8, and 11), trends must therefore be taken with
437 precaution. Generated protein sequences compared to themselves (magenta lower triangles) have mean
438 and variance protein sequence similarities higher than when compared to their query or sub-families. The
439 last two (generated sequences compared to query sequences, orange squares and generated sequences
440 compared to target sequences, green crosses) have mean and variance values spread between the blue and
441 red distributions.

442 These distributions indicate that generated protein sequences by latent space arithmetic have an intrinsic
443 diversity similar to the biological sub-families. Moreover, the generated sequences are less similar to
444 the sequences from their query and source sub-families than to themselves. The generated sequences
445 are also globally as similar to the sequences of their query sub-family as to the sequence of their source
446 sub-family. The generation process is therefore able to capture the features of the selected query and
447 source sub-families and generate a protein sequence diversity similar to the original sub-families.

448

449 Finally, protein structure modeling was performed to assess and compare the properties of the
450 generated sequences by latent space arithmetic and the protein sequences of the natural sub-families.
451 For each sub-family, 100 original sequences were randomly selected along the corresponding generated
452 sequences. All the generated sequences were aligned to protein structures of their corresponding source
453 and query sub-families, and the alignments were used to create 3D structures models by comparative
454 modeling. The quality of models was then evaluated with the DOPE function of MODELLER.

455 Figure S7 shows an example of the energy distribution computed from models using the second
456 strategy with query sub-family S1-0 and source sub-family S1-2. The lowest energies (best models)
457 were found on modelled structures using the original protein sequences of a sub-family to the structural
458 templates of the same sub-family (*Struct. 0 Seq. 0* and *Struct. 2 Seq. 2*). Conversely, the highest energies
459 are found on modelled structures using the original protein sequences of a sub-family to the structural
460 templates of another sub-family (*Struct. 0 Seq. 2* and *Struct. 2 Seq. 0*). Interestingly, generated sequences
461 using additions and subtractions of latent spaces have intermediate energy distributions. This can be
462 clearly observed in Fig. 4, where generated sequences are mostly situated between the two dotted lines.
463 Dots on the right side of the vertical line at 0 correspond to modeled structures using sequences of the
464 latent space with lower energy than the modeled structures using sequences from their original sub-family.
465 Dots on the left side of the vertical line at 0 are modeled structures using sequences of the latent space
466 with higher energy than the modeled structures using sequences from their original sub-family. The
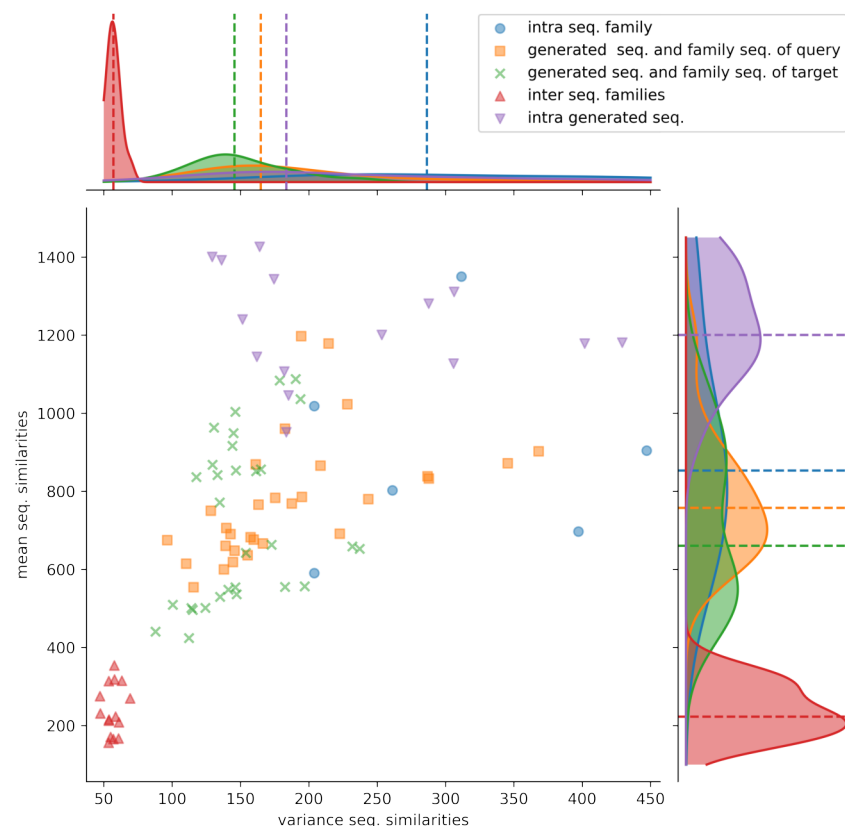
Figure 3. Distributions of protein sequence similarities. Blue dots: protein sequence similarity computed between sequences of the same protein sub-family. Orange squares: similarity computed between generated sequences and the sequences of their query sub-family (ex: S1-0m2 generated sequences and S1-0 sub-family sequences). Green x: similarity computed between generated sequences and the sequences of their target sub-family (ex: S1-0m2 generated sequences and S1-2 sub-family sequences). Red upper triangles: similarity computed between sequences of two different sub-families (ex: S1-0 sequences and S1-2 sequences). Magenta lower triangles: similarity computed between sequences of the same generated sequence group. The variance and the mean of each distribution are displayed on the horizontal and vertical axes.

467 diagonal line on the top-left corner corresponds to the difference in energy between modeled structures
468 using sequences from their original sub-family and modeled structures using sequences from biological
469 sequences of another sub-family. The energy of generated sequences modeled using their query sub-family
470 templates (ex: *Struct. 0 Seq.S1-0m2* and *Struct. 2 Seq. S1-2m0* on Fig. S7 and M_{QS}/Q on Fig. 4) is
471 slightly lower than the energy of models using their source sub-family templates (ex: *Struct. 0 Seq.*
472 *S1-2m0* and *Struct. 2 Seq. S1-0m2* on Fig. S7 and M_{SQ}/Q on Fig. 4). This trend is true for all query /
473 source pairs of sub-families and all strategies except for generated sequences using the fourth strategy
474 (local background subtraction of query latent space using a KD-tree and the addition of source latent
475 space), see Fig. S8, S9, S11, S13 and Methods.
476 In this strategy, the modeled structures using generated sequences do not display energy distributions
477 in-between the energy distributions of the original sequences modeled on structures of the query or of the
478 source sub-families (dotted lines). The energy distribution of generated sequences modeled on structures
479 belonging to the sub-family of their query latent space sub-family (ex: *Struct. 0 Seq.S1-0m2*, blue dots
480 M_{QS}/Q) with the fourth strategy is closer to the energy distribution of the modeled structures using a
481 sequence and a structure template from the same sub-families. The energy distribution of generated
482 sequences modeled on structures corresponding to the sub-family of their source latent space (ex: *Struct.*

483 *2 Seq.S1-0m2*, orange dots M_{SQ}/Q) with the fourth strategy is closer to the energy distribution of the
484 modeled structures using a sequence and a structure template from different sub-families. This indicates
485 that the fourth strategy is less robust to latent space arithmetic than the other three strategies. No clear
486 differences could be observed between the first, second, and third strategy.
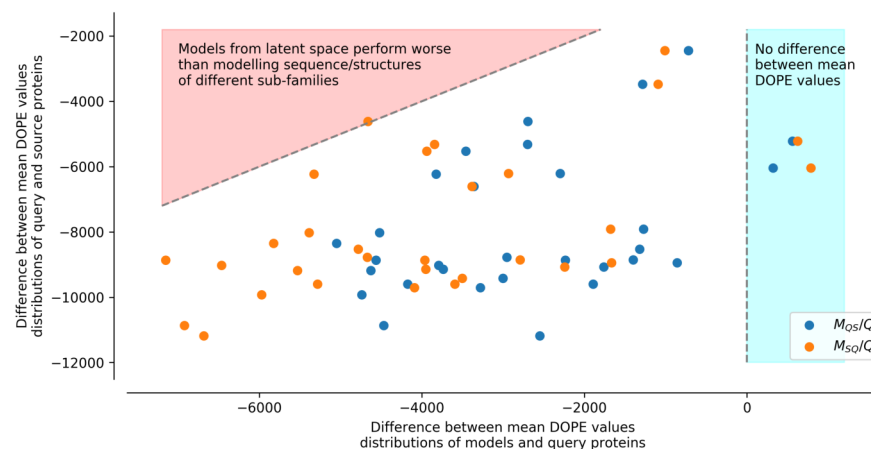
Figure 4. Difference between mean DOPE distributions. Mean value for each distribution, such as
the distributions presented in Fig. S7, were computed. The *y* axis represents the difference between the
mean values computed for query sequences modeled on structures of the same sub-family and mean
values computed for source sequences modeled on structures of the query sub-family (ex: differences
between mean of Struct. 0 Seq. 0 and mean of Struct. 0 Seq. 2 distributions in Fig. S7). The *x* axis
corresponds to the difference between the mean values computed for query sequences modeled on
structures of the same sub-family and mean values computed for query sequences to which latent spaces
of the source sub-family sequences have been added and modeled on structures of the query sub-family
(M_{QS}/Q), or source sequences to which latent spaces of the query sub-family sequences have been added
and modeled on structures of the source sub-family (M_{SQ}/Q) (ex: differences between mean of Struct. 0
Seq. S1-0m2 and mean of Struct. 0 Seq. 0 distributions in Fig. S7).Points in the red area correspond to
mean distribution values from generated sequences whose modeled structures have a higher energy than
models created using pairs of sequences/structures from different sub-families. Points in the blue area
correspond to mean distribution values from generated sequences whose modeled structures have a lower
energy than models created using pairs of sequences/structures from the same sub-family.

DISCUSSION

488 In this study, an Adversarial Autoencoder (AAE) architecture is proposed to analyze and explore the
489 protein sequence space regarding functionality. Previous works based on Variational Autoencoder (VAE)
490 have successfully reported the ability of this deep learning framework to model protein sequence and
491 functional spaces Sinai et al. (2017), predict amino acid fitness impact Hopf et al. (2017); Riesselman
492 et al. (2018), look into protein evolution Ding et al. (2019) or design new protein Greener et al. (2018).
493 AAE networks have the advantage over VAEs to condition the latent space over a prior distribution which
494 has previously been reported to have better efficiency Kadurin et al. (2017).
495 Similarly to previous works using VAE architectures, this study has analyzed the capacity of the AAE ar-
496 chitecture to correctly disentangle protein functional spaces of different families. The generative capacity
497 of the models have been looked into with two original tasks: protein sequences interpolation between
498 different sub-families and protein sequence arithmetics to mix properties of two sub-families.
499 AAEs have been trained on protein sequences of families known to have different sub-families with
500 specific functions. The results have highlighted the ability of AAEs to separate sequences according to
501 functional and taxonomic properties for the three studied families. This result emphasizes the ability of
502 the AAEs to extract biological relevant features and encode them accordingly into a learned latent space.
503 Interpolations in the latent space between encoded sequences of different sub-families have shown smooth
504 transitions of their amino acids even at the active site positions. The generated sequences along the inter-

505 polation paths can be considered as intermediate sequences with sequence properties linked to functions
506 similar to their closest sub-families forming the start or end points of the path. The generated sequences
507 have Shannon entropy values per amino acid position lower than biological sequences which indicates a
508 lower amino acid diversity at each position. This trend point out the robustness and compactness of the
509 latent space as the interpolation method takes the shortest path between the two points of the sub-families.
510 Finally, three strategies have been explored to generate protein sequences with features from two different
511 sub-families. These strategies are based on latent space arithmetic, a concept previously applied in image
512 generation tasks to produce relevant images with intermediate features Radford et al. (2015). Three out
513 of the four different experiments carried out have been able to generate sequences with intermediate
514 features as measured from their protein sequence similarity distributions and modeling energy assessments.
515 Biological experiments will be needed to confirm the functional relevance of the transferred features, but
516 the strategies could have many applications should it be validated.
517 The absence of measured differences between three out of four strategies used to generate intermediate
518 sequences may also indicate that more optimal approaches could be designed. Similarly, the model
519 architecture could also be improved. Currently the model input is a filtered MSA. An improved model
520 could make use of full protein sequences of different sizes without filtering. Unfiltered protein sequences
521 may benefit the generative model by capturing during training important protein specific motifs for family
522 sub-functions Das et al. (2015) not reaching the filtering thresholds
523 Recent advances have been made regarding the protein sequence universe representation notably using
524 self-supervised approaches, notably with the Transformer architecture Alley et al. (2019); Heinzinger
525 et al. (2019); Rao et al. (2019); Rives et al. (2019); Strodthoff et al. (2020). New models from image
526 synthesis could also provide interesting approaches for the generation of protein sequences Vahdat et al.
527 (2021). The reported techniques in this study can be applied to any latent space projection and it would be
528 interesting to combine them with representation of the protein sequence universe to navigate and perform
529 feature transfer between protein families. These techniques could perhaps lead to the rediscovery of
530 evolutionary sequence paths leading to the current protein families Alva et al. (2010), improving our
531 understanding of the protein sequence universe Dryden et al. (2008).
532

CONCLUSION

534 This study shows that AAE models are able to finely capture the protein functional space of three different
535 protein families with known sub-functions. The presented experiments carried out on the sulfatase family
536 provide new insight on the effectiveness of generative model and protein sequence embedding to study
537 and model protein function and evolution. The proposed methods are robust to artifacts and generate
538 consistent sequences and structures.
539 The results of this study show that AAE, in particular, and deep learning generative models in general,
540 can provide original and promising avenues for protein design and functional exploration.

ACKNOWLEDGEMENTS

542 TBF's work was suported ISCD, Sorbonne Université, Paris, France. TBF thanks the NVIDIA society for
543 providing a TitanXp GPU to perform computations.

REFERENCES

545 Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M., and Church, G. M. (2019). Unified rational protein
546   engineering with sequence-based deep representation learning. *Nature methods*, 16(12):1315–1322.
547 Alva, V., Remmert, M., Biegert, A., Lupas, A. N., and Söding, J. (2010). A galaxy of folds. *Protein
548   Science*, 19(1):124–130.
549 Aravind, L., Anantharaman, V., and Koonin, E. V. (2002). Monophyly of class i aminoacyl trna synthetase,
550   uspa, etfp, photolyase, and pp-atpase nucleotide-binding domains: implications for protein evolution in
551   the rna world. *Proteins: Structure, Function, and Bioinformatics*, 48(1):1–14.
552 Armougom, F., Moretti, S., Poirot, O., Audic, S., Dumas, P., Schaeli, B., Keduas, V., and Notredame, C.
553   (2006). Expresso: automatic incorporation of structural information in multiple sequence alignments
554   using 3D-Coffee. *Nucleic Acids Res.*, 34(Web Server issue):W604–608.

Axe, D. D. (2004). Estimating the prevalence of protein sequences adopting functional enzyme folds. *Journal of molecular biology*, 341(5):1295–1315.

Barbeyron, T., Brillet-Gueguen, L., Carre, W., Carriere, C., Caron, C., Czjzek, M., Hoebeke, M., and Michel, G. (2016). Matching the Diversity of Sulfated Biomolecules: Creation of a Classification Database for Sulfatases Reflecting Their Substrate Specificity. *PLoS ONE*, 11(10):e0164846.

Berthold, C. L., Moussatche, P., Richards, N. G., and Lindqvist, Y. (2005). Structural basis for activation of the thiamin diphosphate-dependent enzyme oxalyl-coa decarboxylase by adenosine diphosphate. *Journal of Biological Chemistry*, 280(50):41645–41654.

Bork, P. and Koonin, E. V. (1994). Ap-loop-like motif in a widespread atp pyrophosphatase domain: implications for the evolution of sequence motifs and enzyme activity. *Proteins: Structure, Function, and Bioinformatics*, 20(4):347–355.

Bornberg-Bauer, E. and Chan, H. S. (1999). Modeling evolutionary landscapes: mutational stability, topology, and superfunnels in sequence space. *Proceedings of the National Academy of Sciences*, 96(19):10689–10694.

Boucher, J. I., Cote, P., Flynn, J., Jiang, L., Laban, A., Mishra, P., Roscoe, B. P., and Bolon, D. N. (2014). Viewing protein fitness landscapes through a next-gen lens. *Genetics*, 198(2):461–471.

Campello, R. J., Moulavi, D., and Sander, J. (2013). Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 160–172. Springer.

Copp, J. N., Akiva, E., Babbitt, P. C., and Tokuriki, N. (2018). Revealing unexplored sequence-function space using sequence similarity networks. *Biochemistry*, 57(31):4651–4662.

Das, S., Dawson, N. L., and Orengo, C. A. (2015). Diversity in protein domain superfamilies. *Curr. Opin. Genet. Dev.*, 35:40–49.

Dawson, N. L., Lewis, T. E., Das, S., Lees, J. G., Lee, D., Ashford, P., Orengo, C. A., and Sillitoe, I. (2016). Cath: an expanded resource to predict protein function through structure and sequence. *Nucleic acids research*, 45(D1):D289–D295.

Di Tommaso, P., Moretti, S., Xenarios, I., Orobitg, M., Montanyola, A., Chang, J. M., Taly, J. F., and Notredame, C. (2011). T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. *Nucleic Acids Res.*, 39(Web Server issue):W13–17.

Ding, X., Zou, Z., and Brooks III, C. L. (2019). Deciphering protein evolution and fitness landscapes with latent space models. *Nature communications*, 10(1):1–13.

Dryden, D. T., Thomson, A. R., and White, J. H. (2008). How much of protein sequence space has been explored by life on earth? *Journal of The Royal Society Interface*, 5(25):953–956.

Eddy, S. R. (2011). Accelerated profile hmm searches. *PLoS computational biology*, 7(10):e1002195.

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A., Smart, A., Sonnhammer, E. L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S. C., and Finn, R. D. (2018). The Pfam protein families database in 2019. *Nucleic Acids Research*, 47(D1):D427–D432.

Goldstein, R. A. and Pollock, D. D. (2016). The tangled bank of amino acids. *Protein Science*, 25(7):1354–1362.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 27:2672–2680.

Greener, J. G., Moffat, L., and Jones, D. T. (2018). Design of metalloproteins and novel protein folds using variational autoencoders. *Scientific reports*, 8(1):1–12.

Gómez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hernández-Lobato, J. M., Sánchez-Lengeling, B., Sheberla, D., Aguilera-Iparraguirre, J., Hirzel, T. D., Adams, R. P., and Aspuru-Guzik, A. (2018). Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Science*, 4(2):268–276.

Heinzinger, M., Elnaggar, A., Wang, Y., Dallago, C., Nechaev, D., Matthes, F., and Rost, B. (2019). Modeling aspects of the language of life through transfer-learning protein sequences. *BMC bioinformatics*, 20(1):1–17.

Hopf, T. A., Ingraham, J. B., Poelwijk, F. J., Schärfe, C. P., Springer, M., Sander, C., and Marks, D. S. (2017). Mutation effects predicted from sequence co-variation. *Nature biotechnology*, 35(2):128–135.

Huntley, R. P., Sawford, T., Mutowo-Meullenet, P., Shypitsyna, A., Bonilla, C., Martin, M. J., and O'donovan, C. (2014). The goa database: gene ontology annotation updates for 2015. *Nucleic acids research*, 43(D1):D1057–D1063.

Jost, L. (2006). Entropy and diversity. *Oikos*, 113(2):363–375.

Kadurin, A., Nikolenko, S., Khrabrov, K., Aliper, A., and Zhavoronkov, A. (2017). drugan: an advanced generative adversarial autoencoder model for de novo generation of new molecules with desired molecular properties in silico. *Molecular pharmaceutics*, 14(9):3098–3104.

Kingma, D. P. and Welling, M. (2014). Auto-encoding variational bayes. *Oral presentation at the International Conference on Learning Representations, Banff, Alberta, Canada*.

Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., and Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. *Nature methods*, 15(12):1053.

Luisi, P. L., Chiarabelli, C., and Stano, P. (2006). From never born proteins to minimal living cells: two projects in synthetic biology. *Origins of Life and Evolution of Biospheres*, 36(5):605–616.

Makhzani, A., Shlens, J., Jaitly, N., and Goodfellow, I. J. (2015). Adversarial autoencoders. *CoRR*, abs/1511.05644.

Marchi, J., Galpern, E. A., Espada, R., Ferreiro, D. U., Walczak, A. M., and Mora, T. (2019). Size and structure of the sequence space of repeat proteins. *PLoS computational biology*, 15(8):e1007282.

McInnes, L. and Healy, J. (2017). Accelerated hierarchical density based clustering. In *Data Mining Workshops (ICDMW), 2017 IEEE International Conference on*, pages 33–42. IEEE.

Muller, Y. A., Lindqvist, Y., Furey, W., Schulz, G. E., Jordan, F., and Schneider, G. (1993). A thiamin diphosphate binding fold revealed by comparison of the crystal structures of transketolase, pyruvate oxidase and pyruvate decarboxylase. *Structure*, 1(2):95–103.

Orengo, C. A., Michie, A., Jones, S., Jones, D. T., Swindells, M., and Thornton, J. M. (1997). Cath–a hierarchic classification of protein domain structures. *Structure*, 5(8):1093–1109.

Pandurangan, A. P., Stahlhacke, J., Oates, M. E., Smithers, B., and Gough, J. (2018). The superfamily 2.0 database: a significant proteome update and a new webserver. *Nucleic acids research*, 47(D1):D490–D494.

Poelwijk, F. J., Socolich, M., and Ranganathan, R. (2019). Learning the pattern of epistasis linking genotype and phenotype in a protein. *Nature communications*, 10(1):1–11.

Radford, A., Metz, L., and Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

Rampasek, L., Hidru, D., Smirnov, P., Haibe-Kains, B., and Goldenberg, A. (2017). Dr.VAE: Drug Response Variational Autoencoder. *arXiv e-prints*, page arXiv:1706.08203.

Rao, R., Bhattacharya, N., Thomas, N., Duan, Y., Chen, X., Canny, J., Abbeel, P., and Song, Y. S. (2019). Evaluating protein transfer learning with tape. *Advances in Neural Information Processing Systems*, 32:9689.

Riesselman, A. J., Ingraham, J. B., and Marks, D. S. (2018). Deep generative models of genetic variation capture the effects of mutations. *Nat. Methods*, 15:816–822.

Rives, A., Goyal, S., Meier, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., and Fergus, R. (2019). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *bioRxiv*.

Russ, W. P., Figliuzzi, M., Stocker, C., Barrat-Charlaix, P., Socolich, M., Kast, P., Hilvert, D., Monasson, R., Cocco, S., Weigt, M., and Ranganathan, R. (2020). An evolution-based model for designing chorismate mutase enzymes. *Science*, 369(6502):440–445.

Salinas, V. H. and Ranganathan, R. (2018). Coevolution-based inference of amino acid interactions underlying protein function. *elife*, 7:e34300.

Shen, M. Y. and Sali, A. (2006). Statistical potential for assessment and prediction of protein structures. *Protein Sci.*, 15(11):2507–2524.

Sikosek, T., Chan, H. S., and Bornberg-Bauer, E. (2012). Escape from adaptive conflict follows from weak functional trade-offs and mutational robustness. *Proceedings of the National Academy of Sciences*, 109(37):14888–14893.

Sillitoe, I., Dawson, N., Lewis, T. E., Das, S., Lees, J. G., Ashford, P., Tolulope, A., Scholes, H. M., Senatorov, I., Bujan, A., Ceballos Rodriguez-Conde, F., Dowling, B., Thornton, J., and Orengo, C. A. (2018). CATH: expanding the horizons of structure-based functional annotations for genome sequences. *Nucleic Acids Research*, 47(D1):D280–D284.

Sinai, S., Kelsic, E., Church, G. M., and Nowak, M. A. (2017). Variational auto-encoding of protein sequences. *arXiv preprint arXiv:1712.03346*.

Strodthoff, N., Wagner, P., Wenzel, M., and Samek, W. (2020). Udsmprot: universal deep sequence models for protein classification. *Bioinformatics*, 36(8):2401–2409.

Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B., Wu, C. H., and Consortium, U. (2014). Uniref clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6):926–932.

Tian, P., Louis, J. M., Baber, J. L., Aniana, A., and Best, R. B. (2018). Co-evolutionary fitness landscapes for sequence design. *Angewandte Chemie International Edition*, 57(20):5674–5678.

Tubiana, J., Cocco, S., and Monasson, R. (2019). Learning protein constitutive motifs from sequence data. *elife*, 8:e39397.

Vahdat, A., Kreis, K., and Kautz, J. (2021). Score-based generative modeling in latent space. *arXiv preprint arXiv:2106.05931*.

Webb, B. and Sali, A. (2014). Comparative protein structure modeling using modeller. *Current protocols in bioinformatics*, 47(1):5–6.

White, T. (2016). Sampling generative networks: Notes on a few effective techniques. *CoRR*, abs/1609.04468.

Wolf, Y. I., Aravind, L., Grishin, N. V., and Koonin, E. V. (1999). Evolution of aminoacyl-trna synthetases—analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfer events. *Genome research*, 9(8):689–710.