Analyzing Structural Features of Proteins from Deep-Sea Organisms

Deep-Sea Protein Structure Features

Jochen Sieg,[†] Chris Claudius Sandmeier,[†] Julia Lieske,[‡] Alke Meents,[‡] Christian Lemmen,[¶] Wolfgang R. Streit,[§] and Matthias Rarey*[,†]

[†]*Universität Hamburg, ZBH - Center for Bioinformatics, Bundesstraße 43, 20146 Hamburg, Germany*

[‡]*Center for Free-Electron Laser Science, Deutsches Elektronen-Synchrotron DESY, Notkestraße 85, 22607 Hamburg, Germany*

[¶]*BioSolveIT GmbH, An der Ziegelei 79, 53757 Sankt Augustin, Germany*

[§]*Universität Hamburg, Department of Microbiology and Biotechnology, Ohnhorststraße 18, 22609 Hamburg, Germany*

E-mail: matthias.rarey@uni-hamburg.de

Abstract

Protein adaptations to extreme environmental conditions are drivers in biotechnological process optimization and essential to unravel the molecular limits of life. Most proteins with such desirable adaptations are found in extremophilic organisms inhabiting extreme environments. The deep sea is such an environment and a promising resource that poses multiple extremes on it's inhabitants. Conditions like high hydrostatic pressure and high or low temperature are prevalent and many deep-sea organisms tolerate multiple of these extremes. While molecular adaptations to high temperature are comparatively good described, adaptations to other extremes like high pressure

are not well understood yet. To fully unravel the molecular mechanisms of individual adaptations it is probably necessary to disentangle multifactorial adaptations. In this study we evaluate differences of protein structures from deep-sea organisms and their respective related proteins from non-deep-sea organisms. We created a data collection of 1,281 experimental protein structures from 25 deep-sea organisms and paired them with orthologous proteins. We exhaustively evaluate differences between the protein pairs with machine learning and Shapley Values to determine characteristic differences in sequence and structure. The results show a reasonable discrimination of deep-sea and non-deep-sea proteins from which we distinguish correlations previously attributed to thermal stability from other signals potentially describing adaptions to high pressure. While some distinct correlations can be observed the overall picture appears intricate.

Introduction

Exploiting the properties of proteins from extremophilic microorganisms is a highly active area of research.[1–5] Understanding molecular protein adaptations towards extreme conditions would enable effective design and engineering of proteins with specific properties, which would have important implications for biotechnological processes in many fields like pharmacology, agriculture and biofuels production.[1,4–10] While this research objective is around for multiple decades, in recent years, the understanding of extreme environments and extremophiles has increased tremendously.[4,11] Increasing efforts in metagenomics for a variety of environments provides a continuously growing number of genomic data from extreme environments,[4,5,12,13] which in turn yields a rich source of protein data from extremophiles. Not surprisingly, there is a great interest to systematically analyze the data currently available.

Most extreme environments on earth are characterized by multiple extremes.[11] One of the largest and a particular interesting extreme environment is the deep-sea. It poses multiple extreme conditions on its inhabitants and for this reason, many deep-sea organisms likely exhibit multiple adaptations.[4,9,11,14,15] The extremes in the deep-sea are a temperature range

from high temperature at hydrothermal vents with up to 120°C to low temperature at the sediment of around 2°C.[8] Especially, elevated hydrostatic pressure is inherent in this biom.[16] The average pressure at the ocean floor is 38 MPa[8] and reaches a maximum of approximately 110 MPa at the Challenger Deep of Mariana Trench.[17] In addition, other stressors like an extreme salt range can be found in the deep-sea sediment and at hydrothermal vents.[18]

Within this study we aim to disentangle the multi-factorial aspects of several adaptations in proteins from deep-sea organisms. We are specifically interested to decipher potential protein adaptions to high hydrostatic pressure. The organisms that live under elevated pressure or even need high pressure to grow are called piezophiles (or barophiles).[9,15,19,20] Protein adaptations to high pressure are not well described[9,16,17] and the identification of a molecular signature for high pressure is complicated through other prevailing extremes for example the temperature differences of most high pressure environments.[4,15] Different studies also suggest that pressure adaptations might be challenging to detect, because they might be rather subtle and pronounced differently in different protein classes.[4,15] In addition, like for other extremes,[21] protective mechanisms on the cellular level also seem to play a role as an adaptive strategy in some piezophiles,[4,15] which demonstrates that not all pressure adaptions need to be encoded in the protein.

In contrast to pressure, the adaptations to high temperature are by far the most well-described extreme adaptations.[4,9,16] Numerous studies are comparing thermophiles and mesophiles.[22–24] Even diverse protein engineering efforts demonstrated the thermal stabilization of proteins.[25] Comparison based studies investigating correlating protein properties between homologous proteins of thermophiles and mesophiles even suggest that a global or "nearly universal" signature of protein thermal adaptations exists.[24] However, while intensively studied a fully precise and global physical picture sufficient to enable large-scale rational protein design is not yet derived.[26] Despite general trends it seems that an essential bottleneck is that a complex context-dependent combinations of multiple factors determines the stability toward extreme temperature.[26]

Equipped with the insights of the last decades on protein adaptations to high temperature we aim to delineate high temperature protein adaptations from potential high pressure adaptations in proteins from deep-sea extremophiles inhabiting both, a high pressure and high temperature environment. By taking this perspective not only potential pressure adaptations might be deciphered, but even further insights into the still intricate facets of thermal adaptations might be provided.

Currently, not many studies are taking a data-driven perspective to compare characteristics of proteins of piezophiles or deep-sea organisms with their homologs from other environments. Of the existing studies most are comparing amino acids preferences in the proteom based on genome data[16,27–31] while analysis of protein structures on a larger scale are becoming available only recently.[32] Consequently, it becomes interesting to assess the state of experimental protein structures from deep-sea organisms currently available and comprehensively analyze their features regarding adaptations.

In this study we first establish a data set of protein structures from deep-sea organisms. We collect names of organisms living in the deep-sea from literature and map those names to the Protein Data Bank (PDB).[33] Based on the collected protein structures we assess the current state of available experimental structural protein data of deep-sea organisms. Using the deep-sea protein data we further collect protein structures from organisms not from the deep-sea to compile a data set of orthologous pairs. Protein pairs are selected such that they are related, meaning they are reasonably similar in sequence and structure. This selection aims to enable the isolation of correlating protein features involved in adaptation mechanisms by minimizing evolutionary changes unrelated to extreme adaptations. Subsequently, we analyze a wide range of protein features in a comprehensive top-down machine learning-based feature selection process. The goal is to isolate sequence and structure features that differentiate proteins of deep-sea organisms from proteins of organisms inhabiting different environments. In these experiments we (i) evaluate if there are distinguishing differences between deep-sea proteins and proteins from other environments, like mesophilic and ther-

mophilic organisms and in different protein classes. Then (ii) we determine which features are characteristic and important for differentiation. Finally, (iii) we compare the relevant features derived to already described protein characteristics of thermophiles to assign the observed signals to the individual extremes.

Materials and Methods

Collection of Deep-Sea Protein Structures

The names of microorganisms found in the deep-sea were collected from literature (see[17,34–36] for an overview). The literature for each organism was reviewed manually. The resulting list of organism names was matched to the source organism annotation in the PDB to retrieve protein structures (using the binomial nomenclature and manual review). The list of PDB entries resulting from this protocol can be seen as the currently available experimental deep-sea protein structure data in the PDB. The list of deep-sea organism names collected from literature and the corresponding PDB entries can be found in the supplementary information (deep_sea_species.tsv and deep_sea_pdbs.tsv).

Generation of Potential Orthologous Protein Pairs

Based on the collected deep-sea protein structures potential orthologous protein structures in the PDB were searched. Three protein similarity methods are used to identify protein chain pairs that are related in sequence and structure. All sequences from the deep-sea PDB entries are selected based on the `_entity_poly.pdbx_seq_one_letter_code` data field in the CIF files. Exact sequence duplicates from the same PDB entry were removed, e.g. from homo-meric assemblies to avoid redundant computation. The remaining deep-sea chains are subject to the following protocol to generate orthologous structure pairs from each protein chain.

First, the deep-sea protein chains were used as input to HH-suite[37] (version v3.3.0). HH-suite performs profile-profile alignments of Hidden Markov Models (HMM) to assess the

relationship of protein sequences. HH-suite is able to sensitively identify remote homologous with low sequence identity.[38] We used HHblits[37] with `UniRef30_2020_06`[39] to generate HMM profiles for the deep-sea protein chains (using 3 iterations). The created profiles are then used as input to HHsearch[37] to search PDB70 (`http://wwwuser.gwdg.de/~compbiol/data/hhsuite/databases/hhsuite_dbs/` version 210115) for homologous sequences. From the resulting hits those with a probability > 50 and E-value $< 10^{-3}$ are kept as potential orthologous partner to a deep-sea chain.

The second phase aims to enrich the protein collection found with HHsearch. Since HH-suite comes with PDB70 a pre-computed search database of profiles from a clustered and redundancy reduced version of the PDB a substantial number of sequences from the PDB are removed. However, in this study we are interested in highly similar proteins as long as they are from different organisms. Instead of generating a non-redundant profile database of the whole PDB, which is highly computational intensive, we further enriched our collection of potential protein pairs using the needle tool from EMBOSS suite[40] (version 6.6.0). needle is an implementation of the Needleman-Wunsch algorithm for global sequence alignments. We use needle to compute all pairwise sequence alignments between each deep-sea protein chain and the sequences from the entire PDB. The gapopen and gapextend parameters were set to 10.0 and 0.5, respectively. From the resulting hits all pairs with a sequence identity $> 25\%$ are kept as potential partner.

In an intermediate step all PDB entries from the potential hits that are present in the deep-sea set are removed from the potential hit list.

Finally, we use TM-align[41] (version 20190822) to compute the final protein pairs by filtering the potential sequence hits from HH-suite and needle by fold similarity of the structure. TM-align computes a 3D-structural protein alignment by minimizing the TM-score. The TM-score is a length dependent measure for global fold similarity of two protein chains. We use a TM-score cutoff of 0.5, which has been reported as a criterion to identify protein chains of the same fold.[42] Any protein chain matched based on sequence which has at least one of

the two resulting TM-scores below the threshold is removed.

This protocol is designed to collect protein chain pairs that are at least remotely related in sequence and at the same time share the same overall 3D-fold, which finally should provide a high probability that these proteins are evolutionary related and orthologous. The non-deep-sea data set collected by this protocol will be called decoy data set in the following.

Filtering Protein Structures

PDB entries containing DNA, RNA or chimeric entries or entities, i.e. protein chains from different organisms, were removed. Only structures solved with X-ray crystallography, with a resolution better than 3.0Å are kept. In addition, protein chains with a sequence length < 50 are removed. We also remove PDB entries with suspicious source organism names that contain any of the words "synthetic", "uncultured", "undefined", "unidentified", "artificial", the symbol "?" or where the organism name is empty.

The list of filtered pairs can be found in the supporting information (protein_pairs.tsv).

For the removal of highly redundant protein chains from our data sets we used the MMseqs2[43] software suite (version 13-45111). The PDB contains many highly similar protein chains. On the one hand, we exploit this redundancy to deduce subtle difference between proteins from different organisms. On the other hand, for proteins from the same organism we want to avoid highly similar chains to avoid skewing the distribution in the subsequent evaluation. For this purpose MMseqs2 is applied with default parameters (greedy set cover strategy, coverage=0.8, min_seq_id=0.0) to the sequences of each organism separately. The source organism names were normalized by converting them to lower case and stripping off all words of the name after the first two. For each generated cluster a single representative is selected based on resolution, R-free and largest proportion of resolved residues in the structure. First, the protein chains of the deep-sea data set were clustered for each source organism separately. Second, the protein chains of the decoy set were clustered also for each source organism separately. After this step we removed identical sequences within the decoy set across all organisms.

Protein Features and Structure Preparation

In total 25 sequence and 45 structure features were computed and used in the experiments (see Table 1).

For sequence features the relative frequency of the 20 amino acids for each protein is computed as well as the relative frequency of amino acids with the physico-chemical properties: polar (SER, THR, TYR, ASN, GLN), hydrophobic (ALA, VAL, LEU, ILE, PRO, TRP, PHE, MET), positively charged (LYS, ARG), negatively charged (ASP, GLU) and aromatic (PHE, TRP, TYR).

Structure features can be grouped into the categories non-covalent molecular interactions, secondary structure features, features of the protein's solvent-accessible-surface (SAS) and buried residues, buried waters, volume as well as rigidity/flexibility. Furthermore, features within these categories are combined to create new features.

Table 1: List of computed protein structure and sequence features. Counts of these features are computed per protein structure/sequence and used in the machine learning experiments. 'non-interacting' denotes potential interaction sites that are able to form a specific interaction but in the given state of the structure do not participate in an interaction.

Structure features		Sequence features
Hydrogen bonds	**Secondary structure**	**Amino acid proportions**
Hbonds backbone-backbone	residues in helix	ALA
Hbonds sidechain-sidechain	residues in strand	ARG
Hbonds backbone-sidechain	residues in loop	ASN
acceptors backbone, non-interacting	**Solvent-Accessible Surface (Å^2)**	ASP
donors backbone, non-interacting	hydrophobic	CYS
acceptors sidechain, non-interacting	polar	GLN
donors sidechain, non-interacting	sulfur	GLU
Hbonds surface	pos. charged	GLY
acceptors surface, non-interacting	neg. charged	HIS
donors surface, non-interacting	aromatic	ILE
Ionic interactions	**Buried residue mass (Da)**	LEU
salt bridges	hydrophobic	LYS
anions, non-interacting	polar	MET
cations, non-interacting	sulfur	PHE
salt bridges surface	pos. charged	PRO
anions surface, non-interacting	neg. charged	SET
cations surface, non-interacting	aromatic	THR
Aromatic interactions	**Water**	TRP
cation π	buried waters	TYR
cation π surface	**Volume**	VAL
aromatic π-π	packing density	hydrophobic residues
aromatic π-π surface	**Flexibility**	polar residues
aromatic, non-interacting	torsional constraints	pos. charged residues
aromatic surface, non-interacting	independent hinge joints	neg. charged residues
Hydrophobic interactions		aromatic residues
hydroph. interactions		
hydroph., non-interacting		
hydroph. interactions surface		
hydroph. interactions surface, non-interacting		

All structure features except those in the category volume and flexibility were computed using the NAOMI[44] library. The protonation states of each protein chain is determined with Protoss.[45]

Hydrogen bonds and ionic interactions were computed using the definition and scoring within Protoss.[45] Salt bridges were only computed between the residues ASP, GLU with LYS and ARG.

Cation-π and aromatic π-π interactions were computed with the NAOMI library. Cation-π interactions were considered between LYS and ARG with PHE, TYR or TRP with a distance threshold of < 6Å between cation and ring center, as well as, a maximal deviation of 2Å of the cation from the normal defined at the ring center on the ring plane. π-π interactions were calculated between PHE, TYR and TRP with a maximal distance of 5.5Å between ring centers.

Hydrophobic/lipophilic atoms were identified using the same definition as the JAMDA scoring function.[46] A hydrophobic contact is predicted between two hydrophobic atoms if for their distance d applies $vdW_{sum} < d < 1.75vdW_{sum}$. Hydrophobic contacts were only considered if they are between the side chains of ALA, VAL, LEU, ILE, PRO, TRP, PHE and MET.

Besides the number of observed interactions we also consider potential interaction sites such as atoms, lone pairs and π electron systems that are able to form a specific interaction but in the given state of the structure do not participate in an interaction. We term these 'non-interacting' in the following. An example are hydrogen bond acceptors or donors which are not involved in a hydrogen bond. We also consider interactions and non-interacting sites at certain locations in the protein structure, like on the protein surface or in sidechain and backbone for hydrogen bonds. All features in the category non-covalent interactions are represented by counts of each feature and are normalized by the number of all residues in the structure.

Secondary structure elements were computed with an implementation of the DSSP al-

gorithm[47] within the NAOMI framework. Residues were assigned to the structure elements helix or strand based on the computation or as loop if neither helix nor strand was predicted. Secondary structure features are also normalized by the number of residues in the whole protein structure.

An SAS representation was calculated using the respective algorithm in HYDE[48] from which features of different atoms on the surface could be derived (for example non-covalent interaction features located at the surface). The SAS is computed based on heavy atoms only. From the surface representation we derive the proportion of surface area made up by residues with the physico-chemical properties: hydrophobic, polar, positively and negatively charged, aromatic and sulfur containing residues (MET, CYS). The definitions of those properties are the same as for the sequence features, except sulfur containing residues. All surface area-based features are normalized by the surface area of the whole protein.

Analogously we compute the physico-chemical property distribution of residues buried within the protein (without surface contact). In addition to simply counting we weight the counts by the molecular weight (MW) to capture the size differences of single amino acids. These features are normalized by the MW of the whole protein.

The number of buried waters was used as a descriptor. A water molecule was considered buried if $< \frac{1}{3}$ of it's oxygen's surface is part of the surface that is defined by the heavy atoms of the protein and waters of the complex. The surface was computed with the respective algorithm in HYDE.[48]

Packing density was computed with ProteinVolume[49] (version 1.3) as the van der Waals volume divided by the total volume of the structure in solution.

Rigidity/Flexibility descriptors were used from MSU ProFlex (`https://github.com/psa-lab/ProFlex` version 5.2, formally called FIRST[50]). Specifically, we used the predicted number of torsional constraints and hinge joints as features to describe the global rigidity of the structure. Both features are normalized by the number of residues in the respective protein.

Machine Learning-based Feature Evaluation

Feature Selection Scheme

To evaluate the collected protein structure pairs for expressive differences we use machine learning-based feature selection. Supervised machine learning methods optimize mathematical functions to learn the discrimination of labeled data points. In our case the goal is to learn a model that differentiates between protein structures from deep-sea organisms and protein structures from non-deep-sea organisms. Correspondingly, the labels we use in our experiments are "deep-sea" and "decoy" representing the deep-sea and decoy protein data set respectively. Machine learning algorithms are effective for capturing correlations not only in single features, but in combinations of features.

Figure 1 shows the workflow of our feature evaluation experiments. Initially, we split the collected protein pairs based on the optimal growth temperature (OGT) of their deep-sea source organism. The protein pairs with proteins of (hyper)thermophilic deep-sea organisms (called HT-group) will be used for feature selection. In contrast, pairs with proteins of deep-sea psychrophiles and mesophiles (called PM-group) will be used as an external test set.

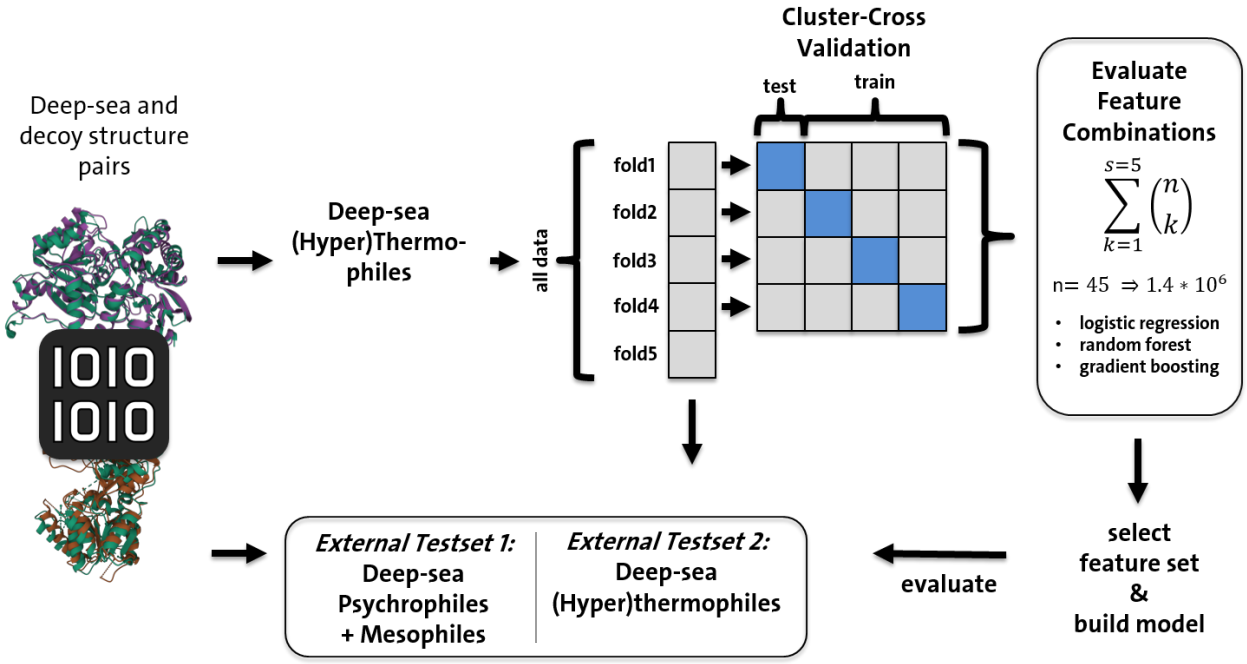

Figure 1: The workflow of the feature selection experiments. Initially, the structure pair data is split based on the deep-sea source organism. The pairs of deep-sea (hyper)thermophiles are used to select important protein features with machine learning. Exemplary, for the $n = 45$ structure features there are 1.4 million feature combinations which are evaluated. The selected features are evaluated on two external test set.

To be able to measure whether combination of features are predictive across different protein families we employ a cluster cross validation strategy. For the creation of the folds we cluster the protein sequences of both deep-sea and decoy samples of the HT-group data set with MMseqs2. We use the connected component clustering mode with a coverage of 0.5 and a min_seq_id of 0.3, which is more suited to capture transient graph connections and therefore should assign more remote homologs in the same cluster. Subsequently, we ensure that all orthologous pairs are kept in the same cluster by adding new edges representing the orthologous pairs to the graph representing the generated clustering of MMSeqs2. On this new graph we compute the connected components again to obtain the final clusters. A fixed number of folds is then generated by assigning the clusters to five folds by trying to keep the size of the folds equal. The folds can be found in the supporting information (folds.tsv).

While four of the five cluster-folds are used for feature selection the remaining cluster-

fold is used as another external test set. With this we evaluate selected features on a fold of deep-sea proteins sequentially dissimilar to those used in feature selection but which are also from (hyper)thermophiles. The performance on the external test sets is determined by models trained on all four folds of the cross validation with the five best performing features from the feature selection. For the PM-group, any decoy chains present in both the PM-group and the four folds used for feature selection are removed from the PM-group before evaluation.

To further investigate the relevance of the decoys' origin we split the folds based on the decoy source organism. We will call the set of all pairs the DecoyAll set (equivalent to the HT-group). Other decoy sets are specifically selected subsets. The MesoModel set contains pairs with mesophilic model organisms like *Homo sapiens* and *Escherichia coli*. The ThermoAll set contains the structures from thermophilic organisms from literature[24,51] of which the ThermoModel set is a subset that only contains proteins of well studied thermophilic organisms for example *Thermus thermophilus* and *Thermotoga maritima*. The feature selection workflow illustrated in Figure 1 is separately conducted for the four different data sets. A list of the decoys source organisms in each group can be found in the supporting information (decoy_subsets.tsv).

We perform feature selection with wrapper methods[52] to evaluate different feature combinations and find the optimal feature set for the binary classification task within a large fraction of all possible feature sets. In this feature selection scheme all combinations of a list of single features are enumerated and machine learning models are trained and evaluated with each feature set in the cluster cross validation. The number of possible feature sets is $2^n - 1$, where n is the number of features. Enumerating all possible feature sets is infeasible. Therefore, we only evaluate feature sets up to a size of $s = 5$ features (see Figure 1). Correspondingly, for the $n = 45$ structure features the number of feature sets is $\sum_{k=1}^{s=5} \binom{45}{k} = 1,385,979$. The threshold of 5 was chosen as a trade-off between computation time and expected predictive power.

Machine learning algorithms are employed from scikit-learn[53] (version 0.23.2). We use the linear method logistic regression (solver='lbfgs', max_iter=10000) as well as the non-linear methods random forest classifier (n_estimators=200) and the gradient boosting classifier (n_estimators=200) which are both based on ensembles of decision trees. The linear method is comparably simple and will be used as baseline method. The two non-linear ensemble methods are able to capture more complex relationships between features.

The measure of choice to assess the prediction performance of the trained machine learning models is to compute the area under the receiver operating characteristic curve (ROC AUC)[54] on the test data sets. The ROC AUC is a threshold free measure assessing the ability of a model to rank positive instances relative to negative instances. This metric provides a value between 0 and 1, where 0.5 represents the random baseline. A useful statistical property is that a ROC AUC of a classification model is equivalent to the probability to rank a randomly selected positive sample higher than a randomly selected negative sample.[54] Consequently, in our experiments, a trained model achieving a ROC AUC of 0.70 would correspond to a 70% probability to rank a randomly selected deep-sea protein before a randomly selected decoy protein based on the given test set.

For the experiments training and test sets were normalized column-wise using the respective training set. We used $z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}$ to compute the normalized feature value z_{ij} from each feature value x_{ij}. i denotes the row and j the column in the feature matrix. z_{ij} is computed by subtracting the mean μ_j of the column from x_{ij} and divide by the column's standard deviation σ_j.

Feature Attribution Scheme

We follow two approaches to not only select predictive features, but to attribute relevance through prediction performance to single features. With this we want to validate our approach and interpret features in the context of protein adaptations to extreme conditions. The basis for these interpretation approaches is the enumeration and evaluation of feature combinations.

In the first approach we will simply evaluate which feature combinations are sufficient to achieve a notable performance in the validation scenario. Small feature subsets, even single features, achieving a comparable or better performance relative to larger sets, like the set of all features indicate highly relevant features in the smaller set.

For the second approach we use the framework of Shapley values[55] from cooperative game theory. Shapley values provide a concept to attribute contributions single features make in combination with other features to the individual single features. The Shapley value of a feature i is defined as the weighted sum of the marginal contributions i makes when i is included in a feature set S:

$$Sh_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!}(v(S \cup \{i\}) - v(S))$$

where N is the set of all features, S is a subset of N, n is the total number of features and v is a function that maps a feature set to a real number. In our experiments we define v to map a feature set to the ROC AUC value the feature set generates in our experiment. Shapley values can be computed in polynomial time, for example through sampling.[56] Here in this study, similar to a sampling approach, we compute the contribution of each feature i by considering only the marginal contributions from a sample of all possible coalitions (in our case the subset of all feature combinations we enumerate). Specifically, we use the resulting mean ROC AUC values from the cluster cross validation experiments of all enumerated feature combinations to attribute contributions to each feature i in terms of ROC AUC. In other words, we will compute the Shapley value for each feature i using the mean ROC AUC in the cluster cross validation experiments of all enumerated feature sets. Features with high resulting contribution values would indicate that these features hold valuable information for the classification model.

Results & Discussion

Deep-Sea Protein Data in the PDB

The data set created contains protein structures from 25 deep-sea organisms. In total, 1,281 PDB entries could be retrieved. A comprehensive overview of the distribution of organisms and number of proteins is shown in Table 2.

Table 2: Deep-sea organisms from literature with protein structures in the PDB. The number of PDB entries corresponds to the number after filtering and with redundancy. The Depth column shows the sample depth in the sea. The T column shows the optimal growth temperature (OGT) or the preferred temperature range if not indicated differently. The P column shows the optimal growth pressure or preferred pressure range if not indicated differently. The T-phile column indicates the classification in hyperthermophile (HT) ($T \geq 75°$), thermophile (T) ($50 \leq T < 75°$), mesophile (M) ($24 \leq T < 50°$) and psychrophile (P) ($T < 24°$).

Species Name	**Depth** (m)	**T** (°C)	**P** (MPa)	**Domain**	**T-phile**	**PDB Entries**
Pyrococcus horikoshii[57]	1395	98	30[19]	Archaea	HT	562
Methanocaldococcus jannaschii[58]	2600	85	75[59]	Archaea	HT	359
Geobacillus sp. HTA-462[60]	10897	55-75		Bacteria	T	113
Pyrococcus abyssi[61]	2000	96	20-40	Archaea	HT	91
Methanopyrus kandleri[62]	2000	98	20[17]	Archaea	HT	39
Shewanella loihica[63,64]	1325	18		Bacteria	P	20
Methanothermococcus thermolithotrophicus[35,65]	0.5	65	50	Archaea	T	17
Thermococcus thioreducens[66]	2300	83–85		Archaea	HT	16
Oceanobacillus iheyensis[67]	1050	30	30 (max)	Bacteria	M	14
Persephonella marina[68]	2507	73		Bacteria	T	7
Photobacterium profundum[69–71]	2551[69,71]/5110[70]	15[69,71]/8-12[70]	28[69,71]/10[70]	Bacteria	P	6
Idiomarina loihiensis[72]	1296	4–46		Bacteria	M	6
Marinactinospora thermotolerans[73]	3865	28		Bacteria	M	5
Shewanella benthica[74]	10898	cold	70	Bacteria	P	4
Pyrococcus yayanosii[75]	4100	98	52	Archaea	HT	4
Moritella profunda[76]	2815	2	22	Bacteria	P	4
Shewanella violacea[77]	5110	8	30	Bacteria	P	3
Thermovibrio ammonificans[78]	2500	75		Bacteria	HT	3
Thermococcus chitonophagus[79]	2600	85	23	Archaea	HT	2
Caldithrix abyssi[80]	3000	60		Bacteria	T	1
Thermosipho melanesiensis[81]	1832-1887	70		Bacteria	T	1
Cryptococcus liquefaciens N6[82,83]	6500			Eukarya		1
Shewanella piezotolerans WP3[84]	1914	15–20	20	Bacteria	P	1
Palaeococcus ferrophilus[85]	1338	83	30	Archaea	HT	1
Methanocaldococcus vulcanius[86]	2600	80		Archaea	HT	1
						1281

The organisms listed in Table 2 were collected in depths greater 1,000m or an elevated optimal growth pressure was reported. Based on the collection depth and the linearly increasing pressure through the water column the approximate pressure range the retrieved organisms inhabit is 10MPa to roughly 110MPa (starting from 1,000m depth). There are 14 Bacteria, 10 Archaea and 1 Eukarya in the data set. The reported optimal growth temperatures (or preferred temperature range) of the organisms are between 2°C and 98°C. This illustrates both extremes of hyperthermophilic and psychrophilic organisms that inhabit the deep-sea. Following the definition of Hait et al.[24] for hyperthermophilic (HT) ($T \geq 75°$), thermophilic (T) ($50 \leq T < 75°$), mesophilic (M) ($24 \leq T < 50°$) and psychrophilic (P) ($T < 24°$) there are 10 hyperthermophilic, 5 thermophilic, 3 mesophilic and 6 psychrophilic organisms in the data set.

The distribution of protein structures collected from the PDB is imbalanced between the organisms. This reflects the imbalanced organism distributions in the PDB itself and is not surprising since that research interest, accessibility and cultivation conditions are also different for different organisms. Most PDB entries that have been retrieved are from *Pyrococcus horikoshii* with 562 proteins structures (44%) and *Methanocaldococcus jannaschii* with 359 structures (28%). Besides the proportions of proteins of the data set that come from individual organisms it is interesting to look at proportions of groups of organisms. The 10 hyperthermophilic Archaea for example make up the majority of proteins in the data set (1078 PDB entries, 84%). In addition, 139 protein entries are from thermophilic organisms which means that 95% of proteins are from organisms living under elevated temperature. In contrast, only 38 proteins (3%) are from psychrophilic organisms and 25 (2%) from mesophilic organisms.

These results show that the current state of available protein data of deep-sea organisms in the PDB is skewed towards individual organisms and towards hypterthermopilic Archaea. Therefore, it is unlikely that the currently available experimental protein structure data on deep-sea proteins is representative for the whole population of proteins from the deep-sea

habitat. However, the data available provides a reasonable basis to compare the proteins of deep-sea (hyper)thermophiles to those of organisms from other environments.

Protein Pair Generation

Protein chains of the retrieved deep-sea proteins were used to identify related protein chains from non-deep-sea organisms from the PDB named decoys in the following (see methods section).

For 1,243 deep-sea protein chains (1,204 PDB entries) at least one decoy chain could be identified. In total, 19,173 decoy chains were found in the PDB by the protocol (see protein_pairs.tsv in the supporting information). The matching of deep-sea and decoy chains in this step can be represented by a bipartite graph. In this set of pairs a single deep-sea chain can be paired with multiple decoy chains and a decoy chain can be paired with multiple deep-sea chains.

Highly redundant protein chains were removed with MMseqs2 as described in the methods section. The final data set contains 501 deep-sea chains and 8,200 decoy chains that come from 20 different deep-sea and 1,379 decoy organisms and form 17,148 chain pairs. According to the applied clustering criteria 60% of the deep-sea chains were highly similar and therefore redundant. This data set was then grouped into connected component clusters for cluster cross validation (see method section) and is the basis for the machine learning experiments in the following sections.

In Figure 2A the distribution of sequence and structure similarity of the pairs is depicted. We calculated the mean TM-score (mTM-score) as the mean of the two resulting TM-scores from each alignment. The distribution of this structural measure of similarity shows an expected value of 0.69 for the average chain pair indicating considerable structural similarity.[42] In contrast, the mean sequence identity as calculated by TM-align is 0.19, which alone would not suffice to indicate an evolutionary relation. The ranking in Figure 2B lists the most frequent source organisms in the decoy data set based on the number of protein chains. Figure 3 illustrates the 3D protein structures of two examples of deep-sea protein chains and

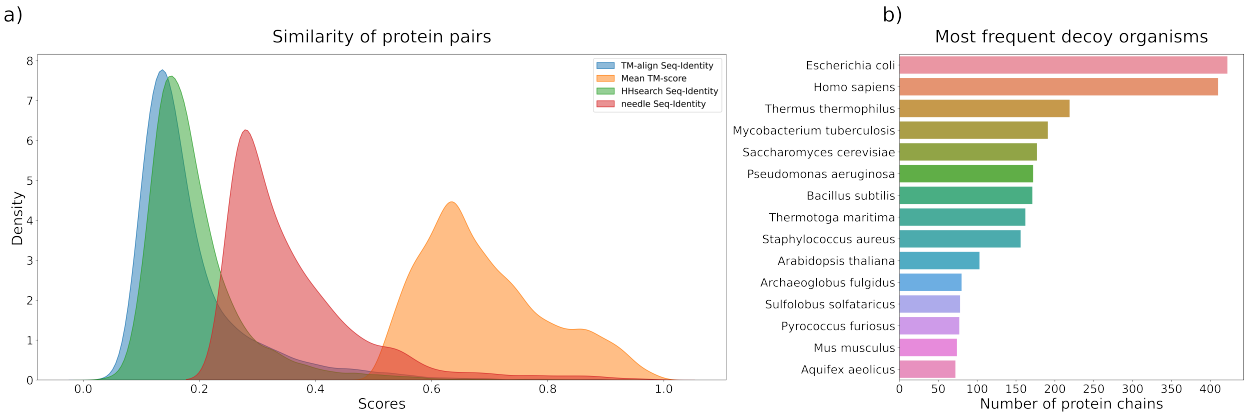

Figure 2: Distributions of the protein pair data set. a) illustrates the distributions of similarity scores between the protein chain pairs of the deep-sea and decoys data set. The mean TM-score (orange) is calculated by taking the mean of the two resulting TM-scores for each protein structure alignment. b) Lists the 15 most frequent source organisms in the decoy data set based on the number of protein chains.

their paired structures.

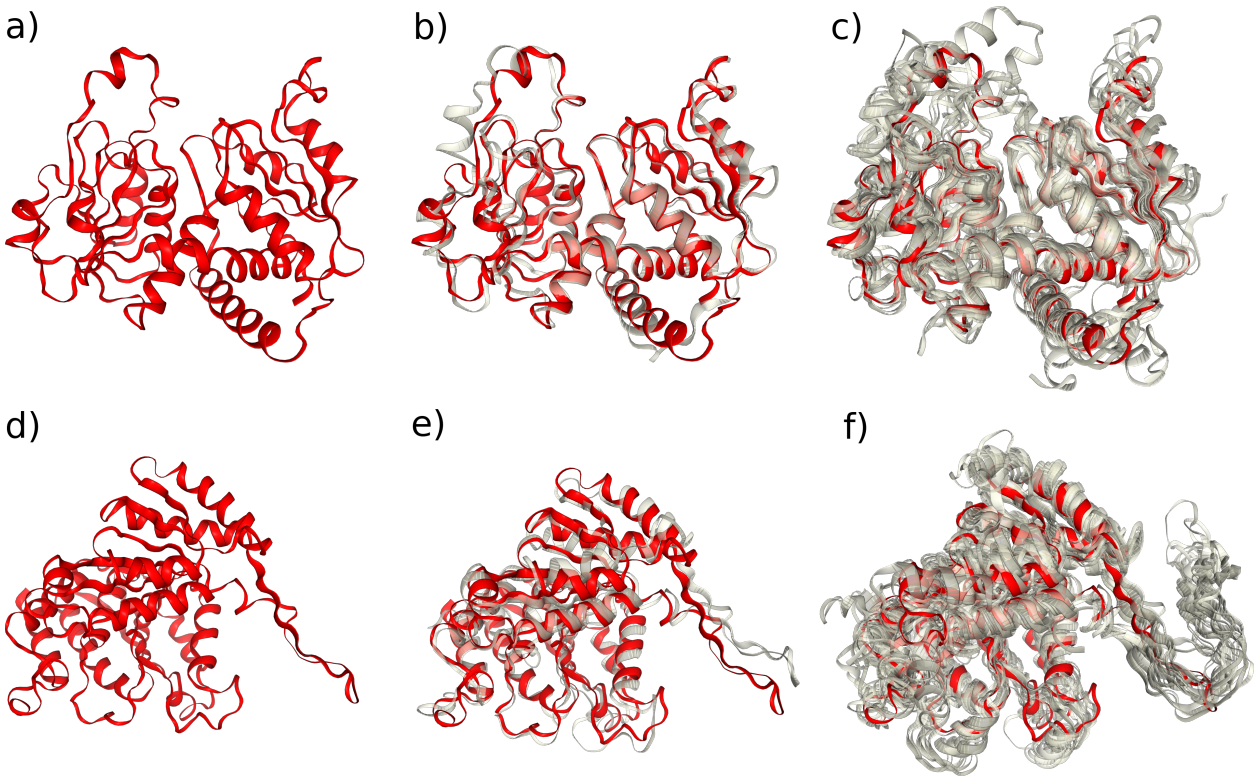

Figure 3: Exemplary structures of protein pairs. Structures from deep-sea organisms are colored in red and decoys in grey. The first row shows pairs generated for the aspartate carbamoyltransferase (1ML4 chain A) from the deep-sea organisms *Pyrococcus abyssi* in red (a)). b) shows the structure paired with an ornithine carbamoyltransferase (1PVV chain A) from *Pyrococcus furiosus*. c) structure ensemble with ten different paired protein chains. The second row shows pairs collected for the 3-isopropylmalate dehydrogenase (3VMK chain A) from the deep-sea organism *Shewanella benthica* in red (d)). e) shows the pair with the 3-isopropylmalate dehydrogenase (1CM7 chain A) from *Escherichia coli*. f) structure ensemble with ten different paired protein chains. Structure alignments have been computed with TM-align and are visualized with NGL.[87] Opacity of decoy structures has been set to 0.6 for visualization purposes.

Machine Learning-based Feature Evaluation

Data Preparation

The compiled protein pair data set was processed for feature selection (see Figure 1). The data was first split based on the deep-sea organisms in the HT-group and PM-group. The HT-group was clustered and grouped into cross validation folds. From the folds subsets were generated based on the source organisms of each decoy in the protein pairs. The composition of the resulting four data subsets of the HT-group are listed in Table 3. Each of these data

sets is evaluated separately with the feature selection workflow in the following.

Table 3: Overview of the protein pair data sets used for feature evaluation. The DecoyAll set contains all deep-sea/decoy protein pairs (equals the HT-group). The other rows are protein pair subsets selected on the decoys source organisms. The MesoModel data sets contains protein pairs with mesophilic model organisms like *Homo sapiens* and *Escherichia coli*. The ThermoAll data set contains pairs with decoy proteins of thermophiles from literature and the ThermoModel data set contains pairs with decoy proteins of model thermophiles like *Thermus thermophilus* and *Thermotoga maritima*.

Decoy Set	**Deep-sea species**	**Decoy species**	**Deep-sea proteins**	**Decoy proteins**
All decoy organisms (DecoyAll)	14	1343	474	7699
Mesophilic model organisms (MesoModel)	12	7	361	1215
All thermophilic organisms (ThermoAll)	11	60	398	931
Thermophilic model organisms (ThermoModel)	9	8	370	684

Can Deep-Sea Proteins be Distinguished?

As a first analysis we investigate the extent deep-sea proteins can be predicted and distinguished from orthologs, but not which specific features are distinguishing them. We will look into the specific features in the next sections.

The prediction performance of feature sets in the cross validation experiment and on the external test sets are depicted in the first and second row in Figure 4, respectively.

The cluster cross validation results in Figure 4 show the distribution of obtained mean ROC AUC over all enumerated feature combinations in the 4-fold cluster cross validation. Feature sets are plotted by their size, meaning at each x position the distribution of mean ROC AUC of all feature sets containing x features is shown. For example a data point in the column $x = 1$ shows the mean ROC AUC achieved by one of the three used machine learning algorithms in the cross validation by using only a single feature, for example the proportion of ALA in the protein. Correspondingly, in the $x = 2$ column performances of feature sets containing two features, for example the proportion of ALA and VAL is depicted. Figure 4 illustrates the performance of all three used machine learning algorithms at once. The performances per algorithm are comparable and can be found in Figure S1, S2 and S3 in the supporting information.

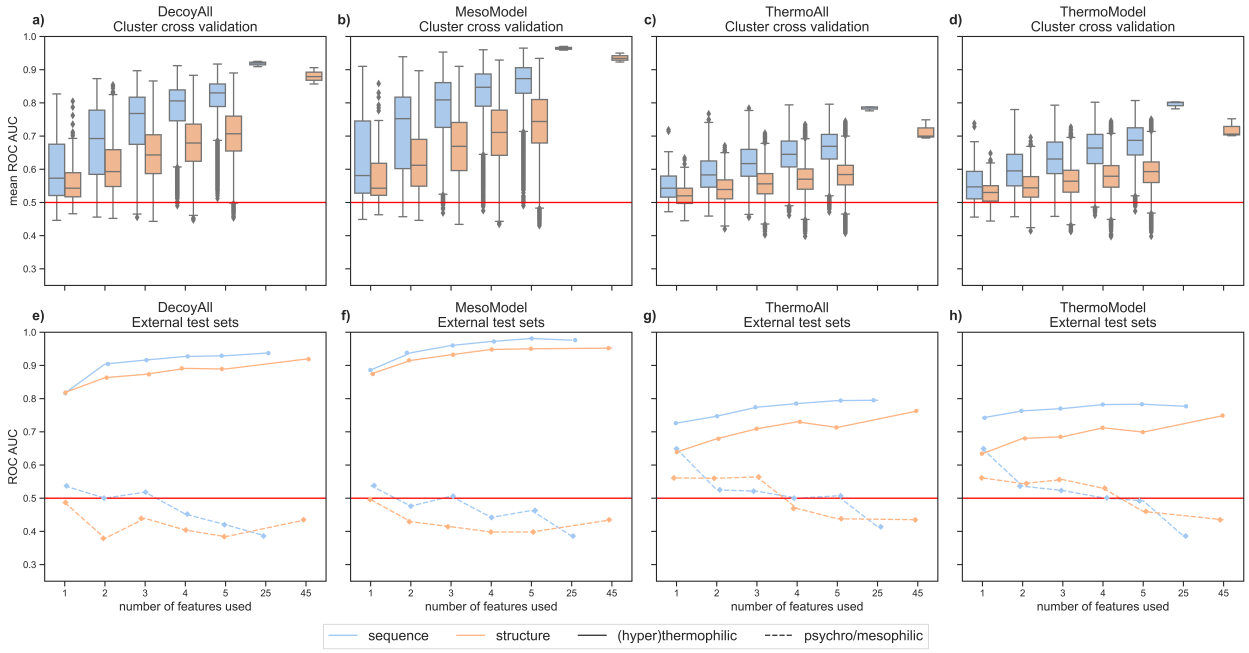

Figure 4: Prediction performance of protein feature combinations on the different protein pair data sets. The first row shows the distribution of mean ROC AUC in the cluster cross validation. Performance is depicted for all used machine learning methods over all enumerated feature combinations for the four different protein pair data sets (a)-d)). The x-axis shows the number of used features in the feature sets. The performance achieved with protein sequence and structure features is depicted separately. The two rightmost entries on the x-axis show the performance with all sequence and structure features, respectively. The second row shows the single best obtained prediction performance on the external test sets from the 5 best features from feature selection for each of the four protein pair data set (e)-h)). The red horizontal line illustrates the random performance baseline of 0.5.

Three general trends of sequence and structure features can be observed from the cross validation results. First, the best and average performance to distinguish deep-sea proteins from their orthologs increases by including more features in almost all cases. Secondly, a small number of ≤ 5 features already yield results similar in comparison to using all features. Even single features yield considerable prediction performance in certain cases. Finally, the prediction performance is observed to be higher when using sequence features instead of structure features. In contrast, to these general trends the prediction performance between decoy data sets differs. For the DecoyAll (a)) and MesoModel (b)) decoy sets the best mean ROC AUC performance is > 0.90 in both sequence and structure. This is an almost perfect class separation. Substantially lower, but also reasonable predictive are the results on the

ThermoAll (c)) and ThermoModel (d)) set with a best mean ROC AUC of 0.81 for sequence and 0.75 for structure features.

The second row of Figure 4 shows the results on the two hold-out external test sets. Models were generated for the five best performing feature sets from the feature selection for each feature set size and for each algorithm, respectively. Only the best performance over all machine learning algorithms and feature sets are shown. Results for all machine learning algorithms and feature sets can be found in Figure S4-S9 in the supporting information.

The results on the hold-out cluster-fold from (hyper)thermophiles, of the HT-group, show a prediction performance that is comparable to the top performance achieved in the cross validation in all four experiments. In contrast the performance on the deep-sea proteins from psychro- and mesophiles, the PM group, is considerably lower and in most cases not better than a random prediction.

The results of both the cluster cross validation and the external test sets show that deep-sea proteins can be successfully separated from orthologous proteins of different environments. However, the extent of this separation depends strongly on the specific source environment of deep-sea organisms and decoy organisms. Noteworthy, on all data sets but the hold-out PM-group data good to perfect prediction performance could be achieved with both sequence and structure features. Consequently, there are systematic differences across the dissimilar protein clusters of (hyper)thermophilic deep-sea organisms. For the DecoyAll and MesoModel data set these differences are global and very easy to capture, they are even encoded in single features. In contrast, systematic differences in the ThermoAll and ThermoModel sets are less obvious and not global. Further, the poor results on the hold-out PM-group suggest that the most relevant features to recognize deep-sea proteins from (hyper)thermophiles are not necessarily relevant to predict proteins from deep-sea psychrophiles and mesophiles. Different adaptation strategies might exists between these groups. However, with only 27 structures the external test data set on proteins from deep-sea psychro/mesophiles is probably not comprehensive enough for conclusions.

Which Features Are Important?

To determine the features important for predictions we use the attributing schemes described in the methods section, mainly the Shapley values analysis. The Shapley values of all features in the different experiments are depicted in Figure 5 and 6. We also provide the standard deviations of the marginals in Figure S10 and S11 in the supporting information. In addition, the distribution of each individual feature in all four data sets can be found in Figure S12-S25 as well as a list of the best performing feature sets from Figure 4 (best_features files) in the supporting information.

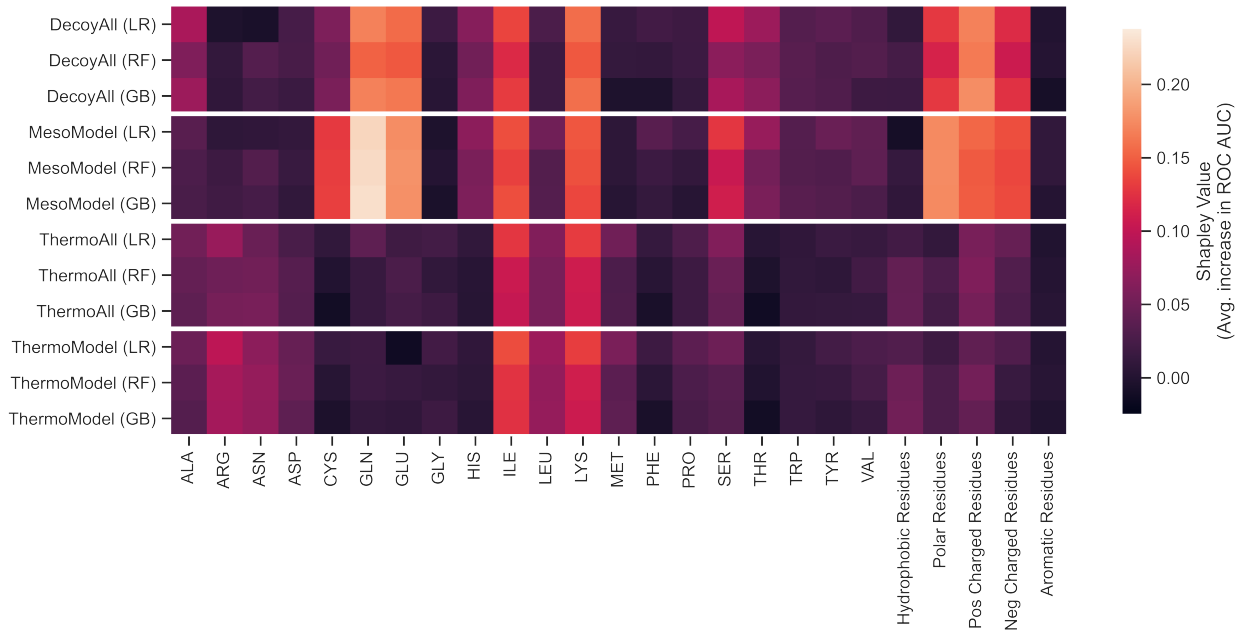

Figure 5: Average ROC AUC contributions of each individual sequence feature over all enumerated and evaluated feature sets in the cluster cross validation. Contributions are computed as the mean of the marginals based on Shapley values. Features are depicted on the x-axis and data sets with the machine learning methods logistic regression (LR), random forest (RF) and gradient boosting (GB) on the y-axis.

The Shapley value plots in Figure 5 and 6 illustrate the average ROC AUC contributions each individual feature makes for sequence and structure features, respectively. More precisely, a cell in the plot shows the average ROC AUC contribution a specific single feature makes on a specific data set for a certain machine learning algorithm in the cluster cross validation.

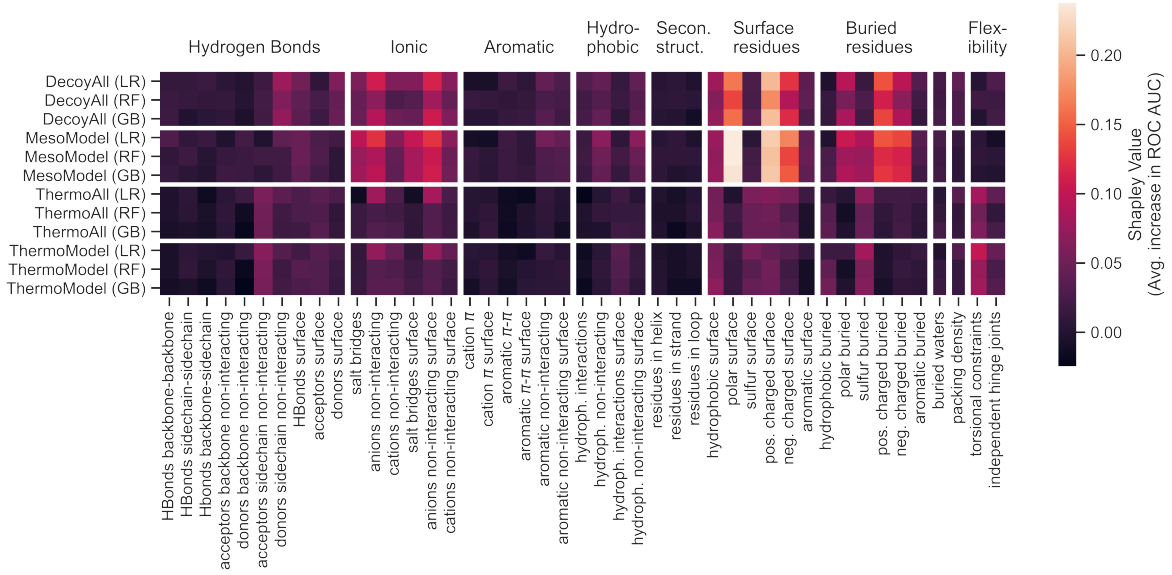

Figure 6: Average ROC AUC contributions of each individual structure feature over all enumerated and evaluated feature sets in the cluster cross validation. Contributions are computed as the mean of the marginals based on Shapley values. Features are depicted on the x-axis and data sets with the machine learning methods logistic regression (LR), random forest (RF) and gradient boosting (GB) on the y-axis.

Distinct contributions of certain features can be observed from the results. Notably, these distinct features are in accordance with the most predictive feature sets in the cross validation in the respective experiments (see best_features.csv). While the contributions within each experiment are relatively consistent for all three machine learning algorithms (minimal pearson's correlation coefficient of 0.93 for sequence; 0.78 for structure) the important features differ between the four experiments. The feature contributions in the DecoyAll and MesoModel experiment are similar (minimal pearson's correlation coefficient 0.91 for sequence; 0.90 for structure). Additionally, contributions in ThermoAll and ThermoModel are similar (minimal pearson's correlation coefficient of 0.88 for sequence; 0.77 for structure). However, the contributions between ThermoAll, ThermoModel and DecoyAll, MesoModel are rather dissimilar (maximal pearson's correlation coefficient of 0.30; 0.31 for structure). Given that the same or similar features are important in the two respective data sets we analyse their results separately.

Deep-Sea Proteins vs. Proteins from All Decoys and Mesophilic Model Organisms

In the DecoyAll and MesoModel experiments the most contributing sequence features are the proportion of GLN, GLU, ILE, LYS, SER, positively and negatively charged residues, polar residues, as well as CYS for the MesoModel experiment. There are also slighter contributions from the proportion of ALA, HIS and THR for the DecoyAll experiment and HIS and THR for the MesoModel experiment. Using only the single most contributing features for classification leads already to a mean ROC AUC of 0.91 (MesoModel with GLN) and 0.83 (DecoyAll with pos. charged residues) in the cross validation (see best_features.csv). This illustrates that the distribution of these residue features alone are highly descriptive. Unsurprisingly, the distribution plots of these features show clear differences between deep-sea proteins and corresponding decoy proteins (see Figure S13 and S16). Specifically, on average deep-sea proteins have less GLN and more positively charged residues than their orthologs from other environments.

For structure features the most contributing features are the proportion of polar sur-

face, positively and negatively charged surface as well as the buried polar residues, buried positively and negatively charges residues and the number of non-interacting anions in the whole protein and on the surface. In addition, salt bridges seem to play a role at the surface and in the whole protein. From these polar and charged surface features are by far the most contributing. Using for example positively and negatively charged surface as feature set yields a mean ROC AUC of 0.85 and 0.90 on the DecoyAll and MesoModel data set with logistic regression in the cross validation. The difference in these features can also be well observed in the distribution plots in Figure S23. On average deep-sea proteins show an increased fraction of charged surface and an decreased fraction of polar surface.

While both sequence and structure features are effective predictors, sequence features are more predictive. It is known that the amino acid composition of an organism's proteom correlates with the organism and an organism's environment.[88,89] However, the trends in important sequence features correspond well to the important structure features considering their physicochemical properties. In both the distribution of charged and polar residues is highly important. The differences in these features even seem to be sufficient to almost perfectly distinguish the here used deep-sea and decoy proteins.

To interpret which biological mechanism these correlations describe it is reasonable to consider the organisms from which the used data was derived. We are comparing deep-sea proteins of mostly hyperthermophilic Archaea. Therefore, it is interesting to determine which features are already attributed to thermal stability in the literature. A recent comparison study by Hait et al.[24] aimed to identify generalized molecular principles of thermal adaptation and extracted "nearly universal" signatures from a larger set of prokaryotes with known optimal growth temperature. The signatures were identified over a diverse set of orthologous protein pairs from (hyper)thermophiles and mesophiles similar to our approach. Hait et al. reported that in 94% of the experiments hyperthermophiles preferred charged amino acids, a pattern that is also very prominent in our results. Additionally, it is reported that the small amino acids GLY and ALA (88%) as well as amid amino acids (96%) are disfavored.

While we only observe a moderate contribution from ALA in the DecoyAll (and none in the MesoModel experiment) and none from GLY we can observe a very strong contribution from the proportion of the amid amino acid GLN, but not ASN. On the structure level Hait et al. reported an increased hydrophobic core (73%), higher exposure of charged/polar surface area (79%) and abundant salt-bridges (83%) as well as a higher number of cation-π interactions (74%). In our result the hydrophobic core and the number cation-π interactions seems to have no noteworthy contribution. However we also observe high contributions of charged and polar surface area as well as ionic interactions (salt bridges). Explicitly, the polar surface is reduced on average in our deep-sea proteins and charged surface and salt bridges are increased.

In conclusion, the comparison of the deep-sea proteins from mainly hyperthermophilic Archaea shows a very strong and simple to capture pattern of protein properties to differentiate them from mesophilic proteins and orthologous proteins in general (as measured on the baseline DecoyAll set). Intriguingly, it seems that the patterns observed are very similar to the protein properties typically attributed to protein adaptions for thermal stability, which are specifically a reduced number of polar and an increased number of charged residues.[14,24,26,90] This is not surprising since most available experimental protein structure data from deep-sea organisms comes from hyperthermophilic Archaea. As a consequence, it is complicated to assign these correlations unambiguously to extreme adaptations, like to temperature, pressure or both.

Deep-Sea Proteins vs. Proteins from Thermophiles

In the experiments involving only decoy structures from thermophiles (ThermoAll, ThermoModel) the most prominent sequence features are the proportion of ILE and LYS which by far show the highest contributions (see Figure 5). The third most relevant feature is ARG and we can also observe smaller contributions from ASN and LEU. Using LYS or ILE individually as feature sets results in mean ROC AUC values between 0.71 and 0.74 in the cross validation and a similar performance on the external cluster-fold (see Figure 4, best_feature

files). Deep-sea proteins contain more LYS and more ILE on average than their respective decoys (see Figure S13 and S14). In contrast, ARG is slightly decreased on average in deep-sea proteins (see Figure S12). Noteworthy, the feature set of both ILE and LYS is only marginally better than the two features alone, suggesting that both are correlated.

The most contributing structure features (contribution > 0.05) are the number of torsional constraints, buried sulfur residues and number of non-interacting anions at the surface and in the whole protein, as well as, the number of non-interacting acceptors in side chains, surface area of sulfur residues and hydrophobic surface, the positively and negatively charged surface area and buried hydrophobic residues are contributing. These most contributing features correspond well to the best performing single feature sets in the cross validation experiment (see best_features.csv). Notably, on the external test cluster-fold especially the number of non-interacting anions at the surface and in the whole protein show prediction performance consistent with the cross validation when used as single features (ROC AUC of approx. 0.63).

In the literature there are only a handful of studies exploring differences between deep-sea proteins and thermophiles which focus mainly on sequence composition of proteins from deep-sea piezophiles. Nath et al.[31] determined relevant amino acids to differentiate the protein sequences of piezophilic-thermophilic and thermophilic–nonpiezophilic of *P. yayanosii* and *P. furiosus* as well as *Thermococcus barophilus* and *Thermococcus kodakarensis KOD1*. They ranked ARG, LYS, ASN and ILE for the first pair and ILE, LYS and ARG for the second pair as the most important features. These results are in agreement with the results we obtained. In another sequence of studies, Di Giulio[27,28] also found that especially the frequency of LYS, ILE and ARG are correlated with piezophilic organisms. The author described the hydrostatic pressure asymmetry index for the protein sequences of three pairs of piezophilic-thermophilic and thermophilic–nonpiezophilic organisms, namely *P. abyssis* with *P. furiosus*, *P. yayanosii* with *P. furiosus* and *T. barophilus* with *T. kodakarensis*. Interestingly, depending on the organism pairs the correlation was either positive or negative.[28] The

author reasoned that because both LYS and ARG have similar physicochemical properties at some point in evolution the organisms committed to one or the other. In our experiment we see an increased use of LYS and ILE but an reduced use in ARG meaning that the proteins from organisms we investigate show a positive correlation with LYS and ILE and a negative correlation with ARG.

Again, sequence features are more predictive than structure features. Interestingly, no clear correspondence between the relatively well correlating amino acids LYS and ILE and the properties of important structure features is apparent. A reason for this might be, on the one hand, structural adaptations induced by the sequence adaptations might be simply not well described by our chosen structure features or can not be sufficiently captured from the accuracy or static state of the crystal structure. On the other hand, this discrepancy might be because the amino acid preference is not expressed in structural differences and is therefore potentially unrelated to protein extreme adaptions.

In conclusion, both ILE and LYS and also ARG to a lesser extent are reasonably important in sequence and were also found to be important by others. In contrast, no individual structural feature is contributing very distinctively, except perhaps non-interacting anions. The predictive power of structure features observed in Figure 4 is therefore rather due to combinations of multiple features, instead of a clear preference in one of the single structure feature. The results suggest that this combination is related to non-interacting anions, sulfur containing residues and the flexibility of the protein.

Important Deep-Sea Protein Features

When we compare the prediction performance and important features from all four experiments, we observe that deep-sea protein structures are harder to distinguish from structures of thermophiles (ThermoAll, ThermoModel) than from structures of mesophiles and all decoys (MesoModel, DecoyAll). This is not surprising, considering that most (hyper)thermophilic deep-sea organisms are likely evolutionary more similar to the thermophiles. While the important features in the MesoModel experiments are clear, it is not possible to

assign single correlations to individual (or multiple) extreme conditions on these results alone. For this reason, we compared the deep-sea proteins to proteins of thermophiles to further isolate potential pressure adaptations in proteins from (hyper)thermophilic deep-sea organisms.

An interesting result is that the features most important in the ThermoAll and ThermoModel experiments seem to be relevant also in the MesoModel and DecoyAll experiments (see Figure 5 and 6). Over all four data sets deep-sea proteins contain more LYS and more ILE on average than their respective decoys (see Figure S13 and S14). The only structure features that are reasonably important over all four data sets are the distribution of non-interacting anions in the whole protein and at the surface. Both are increased on average in deep-sea proteins (see Figure S19). These results suggests that the distribution of these features is a rather unique trait of deep-sea proteins.

While our results provide clues about the features characterizing (hyper)thermophilic deep-sea proteins, a clear pattern or mechanism for high pressure adaptations is not apparent. However, effective prediction of deep-sea proteins is possible in all four experiments. These results demonstrate that predictive structural patterns between different deep-sea protein clusters exists. Our most predictive features and feature sets indicate which kind of protein property this hidden pattern might be related to.

Conclusion

Molecular adaptations to the Deep-Sea Environment

The result that proteins of deep-sea (hyper)thermopiles are nearly perfectly separable from proteins from mesophiles likely illustrates the obvious differences between the proteins from thermophiles and mesophiles which have been explored heavily in the past.[14,24,26,90] However, these obvious differences alone are not sufficient to fully enable engineering of proteins towards high temperature[26] and probably other extreme conditions. Our results

on the ThermoAll and ThermoModel data sets show that in addition to the general trends already analyzed in detail, there are other, more complicated patterns in protein sequence and structure correlating with the deep-sea source environment. However, these correlations are not global for the whole population of deep-sea proteins. On the one hand these non-global correlations are in accordance with current believe on pressure adaptations[4,15] which are stating that pressure adaptations are only present in a subset of deep-sea proteins. Yet, some of the relevant features, most importantly LYS, ILE and charged atoms (especially anions), are shared across protein clusters and different decoy sets, which indicates the same adaptations in different protein classes. On the other hand, it seems that the features characteristic for proteins of deep-sea (hyper)thermophiles differ from those of deep-sea psychrophiles. However, the available structure data on deep-sea psychrophiles is scarce, which make these results not conclusive.

Consequently, the next interesting question to address would be in which deep-sea proteins and protein classes do we see molecular adaptations? One approach would be to investigate the proteins for which the determined important features are relevant. It would also be interesting to analyze individual protein classes that are more likely to hold adaptations, like enzymes involved in the energy metabolism.[4] In addition, further experiments with different sub-populations of deep-sea organisms are necessary, for example based on evolutionary relation of organisms or the similarity of their source environments, like the prevailing extreme conditions or the metabolism. Besides that, with our experiments we could provide a picture of the importance of a wide range of different features. However, to pin point single highly important features further features need to be evaluated. An interesting example would be the proteins energetics and dynamics, which are not directly captured by our current descriptors or with the static protein structures. In conclusion, there are still multiple directions little explored yet and which are likely to provide valuable clues to disentangle the multiple protein adaptations to extremes.

The Current Status of Protein Structures from Deep-Sea Organisms

The currently available experimental protein structure data from deep-sea organisms in the PDB is scarce. In this work we could retrieve 1281 experimental protein structures (501 non-redundant) from 25 deep-sea organisms (see Table 2). While this constitutes a first data basis to analyze protein structures from deep-sea organisms and the absolute number of structures is probably sufficient for many analyses, the diversity of the retrieved organisms is limited. Most structures are from hyperthermophilic Archaea and 95% of the proteins are from organisms living under elevated temperature while only 5% are from psychrophilic and mesophilic deep-sea organisms. While the protein structure is more informative, the sequence data that is available in more variety and quantity would foster our understanding given the more tangible signals in sequence features.

In contrast, the available structure data for generating orthologous protein pairs with proteins of organisms from other environments from the PDB seems to be plenty. While we generated structure pairs having the same fold and at least remote homology is detectable in sequence, a more stringent sequence similarity likely provides an even less noisy picture of protein differences. However, this would reduce the number of pairs and leave more deep-sea proteins unpaired. Probably the most important bottleneck in the pair generation process is the annotation of the source organisms environments, optimal growth temperature and pressure or even annotation on the individual protein level. This, however, remains a grand and largely multidisciplinary challenge.[4] Further, while we currently using deep-sea proteins as a proxy for pressure stability (or other extreme adaptations), it would be extremely beneficial to compare protein pairs with experimentally determined low and high pressure stability.

In the future, the ever increasing efforts in environmental metagenomics[4,12] will provide more genome data from extreme environments. Carefully curated metadata annotation of these genomes with the conditions of their natural environment would provide an invaluable resource to comprehend the relationship between protein structure and environmental conditions. At the same time recent advancement made in protein structure prediction from

sequence[91,92] provides a incomparable amount of structural protein information which is detached from what is experimentally solvable. Although, we still need to find out whether these methods model the subtleties in protein structures that we are looking for when we search for protein adaptations. Nevertheless, with more data available more comprehensive evaluation on single protein classes could be conducted. In addition, more expressive methodologies could be applied which allow to explore not only handcrafted features but also derive features from the data itself, which might be a fitting approach given the subtlety and context-dependence molecular adaptations are believed to have. An example of these are deep neural networks which we intentionally set aside in this study because of the limited data. Therefore, the future promises to further advance our understanding of the molecular limits of life and to exploit the full potential of enzymes from extremophiles.

Finally, we hope the compiled data set and our feature evaluation will be useful to the community and a helpful starting point for other studies.

Acknowledgement

This work was supported by the German Federal Ministry of Education and Research as part of protP.S.I. (031B0405B).

Supporting Information Available

Additional data and results are also provided in the supporting information.

References

(1) Vieille, C.; Zeikus, G. J. Hyperthermophilic Enzymes: Sources, Uses, and Molecular Mechanisms for Thermostability. *Microbiology and Molecular Biology Reviews* **2001**, *65*, 1–43.

(2) Feller, G. Psychrophilic Enzymes: From Folding to Function and Biotechnology. *Scientifica* **2013**, *2013*, 1–28.

(3) Acinas, S. G. et al. Deep ocean metagenomes provide insight into the metabolic architecture of bathypelagic microbial communities. *Communications Biology* **2021**, *4*.

(4) Ando, N.; Barquera, B.; Bartlett, D. H.; Boyd, E.; Burnim, A. A.; Byer, A. S.; Colman, D.; Gillilan, R. E.; Gruebele, M.; Makhatadze, G.; Royer, C. A.; Shock, E.; Wand, A. J.; Watkins, M. B. The Molecular Basis for Life in Extreme Environments. *Annual Review of Biophysics* **2021**, *50*, 343–372.

(5) Aevarsson, A. et al. Going to extremes - A metagenomic journey into the dark matter of life. *FEMS Microbiology Letters* **2021**, *368*, 1–16.

(6) Hamilton, P. B.; Van Slyke, D. D. the Gasometric Determination of Free Amino Acids in Blood Filtrates By the Ninhydrin-Carbon Dioxide Method. *Journal of Biological Chemistry* **1943**, *150*, 231–250.

(7) Balny, C. What lies in the future of high-pressure bioscience? *Biochimica et Biophysica Acta - Proteins and Proteomics* **2006**, *1764*, 632–639.

(8) Winter, R.; Lopes, D.; Grudzielanek, S.; Vogtt, K. Towards an Understanding of the Temperature/Pressure Configurational and Free-Energy Landscape of Biomolecules. *Journal of Non-Equilibrium Thermodynamics* **2007**, *32*, 41–97.

(9) Chakravorty, D.; Khan, M. F.; Patra, S. Multifactorial level of extremostability of proteins: can they be exploited for protein engineering? *Extremophiles* **2017**, *21*, 419–444.

(10) Hikida, Y.; Kimoto, M.; Hirao, I.; Yokoyama, S. Crystal structure of Deep Vent DNA polymerase. *Biochemical and Biophysical Research Communications* **2017**, *483*, 52–57.

(11) Harrison, J. P.; Gheeraert, N.; Tsigelnitskiy, D.; Cockell, C. S. The limits for life under multiple extremes. *Trends in Microbiology* **2013**, *21*, 204–212.

(12) Konstantinidis, K. T.; Braff, J.; Karl, D. M.; DeLong, E. F. Comparative metagenomic analysis of a microbial community residing at a depth of 4,000 meters at station ALOHA in the North Pacific Subtropical Gyre. *Applied and Environmental Microbiology* **2009**, *75*, 5345–5355.

(13) Bar-On, Y. M.; Phillips, R.; Milo, R. The biomass distribution on Earth. *Proceedings of the National Academy of Sciences of the United States of America* **2018**, *115*, 6506–6511.

(14) Reed, C. J.; Lewis, H.; Trejo, E.; Winston, V.; Evilia, C. Protein adaptations in archaeal extremophiles. 2013.

(15) Ichiye, T. Enzymes from piezophiles. *Seminars in Cell and Developmental Biology* **2018**, *84*, 138–146.

(16) Peoples, L. M.; Kyaw, T. S.; Ugalde, J. A.; Mullane, K. K.; Chastain, R. A.; Yayanos, A. A.; Kusube, M.; Methé, B. A.; Bartlett, D. H. Distinctive gene and protein characteristics of extremely piezophilic Colwellia. *BMC Genomics* **2020**, *21*.

(17) Salvador-Castell, M.; Oger, P.; Peters, J. *Physiological and Biotechnological Aspects of Extremophiles*; INC, 2020; pp 105–122.

(18) Kaye, J. Z.; Baross, J. A. Synchronous Effects of Temperature , Hydrostatic Pressure , and Salinity on Growth , Phospholipid Profiles , and Protein Patterns of Four Halomonas Species Isolated from Deep-Sea Hydrothermal- Vent and Sea Surface Environments. **2004**, *70*, 6220–6229.

(19) Horikoshi, K. Barophiles: Deep-sea microorganisms adapted to an extreme environment. *Current Opinion in Microbiology* **1998**, *1*, 291–295.

(20) Abe, F.; Horikoshi, K. The biotechnological potential of piezophiles. *Trends in Biotechnology* **2001**, *19*, 102–108.

(21) Jarzab, A. et al. Meltome atlas—thermal proteome stability across the tree of life. *Nature Methods* **2020**, *17*, 495–503.

(22) Kumar, S.; Tsai, C.-J.; Nussinov, R. Factors enhancing protein thermostability. *Protein Engineering, Design and Selection* **2000**, *13*, 179–191.

(23) Razvi, A.; Scholtz, J. M. Lessons in stability from thermophilic proteins. *Protein Science* **2006**, *15*, 1569–1578.

(24) Hait, S.; Mallik, S.; Basu, S.; Kundu, S. Finding the generalized molecular principles of protein thermal stability. *Proteins: Structure, Function and Bioinformatics* **2020**, *88*, 788–808.

(25) Eijsink, V. G.; Bjørk, A.; Gåseidnes, S.; Sirevåg, R.; Synstad, B.; Burg, B. V. D.; Vriend, G. Rational engineering of enzyme stability. *Journal of Biotechnology* **2004**, *113*, 105–120.

(26) Pucci, F.; Rooman, M. Physical and molecular bases of protein thermal stability and cold adaptation. *Current Opinion in Structural Biology* **2017**, *42*, 117–128.

(27) Di Giulio, M. A comparison of proteins from Pyrococcus furiosus and Pyrococcus abyssi: barophily in the physicochemical properties of amino acids and in the genetic code. *Gene* **2005**, *346*, 1–6.

(28) Di Giulio, M. The origin of the genetic code in the ocean abysses: New comparisons confirm old observations. *Journal of Theoretical Biology* **2013**, *333*, 109–116.

(29) Yafremava, L. S.; Di Giulio, M.; Caetano-Anollés, G. Comparative analysis of barophily-related amino acid content in protein domains of Pyrococcus abyssi and Pyrococcus furiosus. *Archaea (Vancouver, B.C.)* **2013**, *2013*, 680436.

(30) Pradel, N.; Ji, B.; Gimenez, G.; Talla, E.; Lenoble, P.; Garel, M.; Tamburini, C.; Fourquet, P.; Lebrun, R.; Bertin, P.; Denis, Y.; Pophillat, M.; Barbe, V.; Ollivier, B.; Dolla, A. The First Genomic and Proteomic Characterization of a Deep-Sea Sulfate Reducer: Insights into the Piezophilic Lifestyle of Desulfovibrio piezophilus. *PLoS ONE* **2013**, *8*.

(31) Nath, A.; Subbiah, K. Insights into the molecular basis of piezophilic adaptation: Extraction of piezophilic signatures. *Journal of Theoretical Biology* **2016**, *390*, 117–126.

(32) Avagyan, S.; Vasilchuk, D.; Makhatadze, G. I. Protein adaptation to high hydrostatic pressure: Computational analysis of the structural proteome. *Proteins: Structure, Function, and Bioinformatics* **2020**, *88*, 584–592.

(33) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Research* **2000**, *28*, 235–242.

(34) Fang, J.; Zhang, L.; Bazylinski, D. A. Deep-sea piezosphere and piezophiles: geomicrobiology and biogeochemistry. 2010.

(35) Jebbar, M.; Franzetti, B.; Girard, E.; Oger, P. Microbial diversity and adaptation to high hydrostatic pressure in deep-sea hydrothermal vents prokaryotes. *Extremophiles* **2015**, *19*, 721–740.

(36) Zhang, Y.; Li, X.; Xiao, X.; Bartlett, D. H. Current developments in marine microbiology: High-pressure biotechnology and the genetic engineering of piezophiles. *Current Opinion in Biotechnology* **2015**, *33*, 157–164.

(37) Steinegger, M.; Meier, M.; Mirdita, M.; Vöhringer, H.; Haunsberger, S. J.; Söding, J. HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics* **2019**, *20*, 1–15.

(38) Söding, J. Protein homology detection by HMM-HMM comparison. *Bioinformatics* **2005**, *21*, 951–960.

(39) Mirdita, M.; Von Den Driesch, L.; Galiez, C.; Martin, M. J.; Soding, J.; Steinegger, M. Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Research* **2017**, *45*, D170–D176.

(40) Rice, P.; Longden, L.; Bleasby, A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics* **2000**, *16*, 276–277.

(41) Zhang, Y.; Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research* **2005**, *33*, 2302–2309.

(42) Xu, J.; Zhang, Y. How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics* **2010**, *26*, 889–895.

(43) Steinegger, M.; Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology* **2017**, *35*, 1026–1028.

(44) Urbaczek, S.; Kolodzik, A.; Fischer, J. R.; Lippert, T.; Heuser, S.; Groth, I.; Schulzgasch, T.; Rarey, M. NAOMI : On the Almost Trivial Task of Reading Molecules from Different File formats. **2011**, 3199–3207.

(45) Bietz, S.; Urbaczek, S.; Schulz, B.; Rarey, M. *Protoss: a holistic approach to predict tautomers and protonation states in protein-ligand complexes*; 2014; Vol. 6; p 12.

(46) Flachsenberg, F.; Meyder, A.; Sommer, K.; Penner, P.; Rarey, M. A Consistent Scheme for Gradient-Based Optimization of Protein - Ligand Poses. *Journal of Chemical Information and Modeling* **2020**, *60*, 6502–6522.

(47) Kabsch, W.; Sander, C. Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers* **1983**, *22*, 2577–2637.

(48) Schneider, N.; Lange, G.; Hindle, S.; Klein, R.; Rarey, M. A consistent description of HYdrogen bond and DEhydration energies in protein-ligand complexes: Methods behind the HYDE scoring function. *Journal of Computer-Aided Molecular Design* **2013**, *27*, 15–29.

(49) Chen, C. R.; Makhatadze, G. I. ProteinVolume: Calculating molecular van der Waals and void volumes in proteins. *BMC Bioinformatics* **2015**, *16*, 101.

(50) Jacobs, D. J.; Rader, A. J.; Kuhn, L. A.; Thorpe, M. F. Protein flexibility predictions using graph theory. *Proteins: Structure, Function and Genetics* **2001**, *44*, 150–165.

(51) Leigh, J. A.; Albers, S. V.; Atomi, H.; Allers, T. Model organisms for genetics in the domain Archaea: Methanogens, halophiles, Thermococcales and Sulfolobales. *FEMS Microbiology Reviews* **2011**, *35*, 577–608.

(52) Guyon, I.; Elisseeff, A. An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research* **2003**, *3*, 1157–1182.

(53) Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.

(54) Fawcett, T. An introduction to ROC analysis. *Pattern Recognition Letters* **2006**, *27*, 861–874.

(55) Shapley, L. S. In *Classics in Game Theory*, ii ed.; Kuhn, H., Tucker, A., Eds.; Princeton University Press: Princeton, 2020; Chapter 17. A Valu, pp 69–79.

(56) Castro, J.; Gómez, D.; Tejada, J. Polynomial calculation of the Shapley value based on sampling. *Computers and Operations Research* **2009**, *36*, 1726–1730.

(57) González, J. M.; Masuchi, Y.; Robb, F. T.; Ammerman, J. W.; Maeder, D. L.; Yanagibayashi, M.; Tamaoka, J.; Kato, C. Pyrococcus horikoshii sp. nov., a hyperthermophilic

archaeon isolated from a hydrothermal vent at the Okinawa Trough. *Extremophiles* **1998**, *2*, 123–130.

(58) Jones, W. J.; Leigh, J. A.; Mayer, F.; Woese, C. R.; Wolfe, R. S. Methanococcus jannaschii sp. nov., an extremely thermophilic methanogen from a submarine hydrothermal vent. *Archives of Microbiology* **1983**, *136*, 254–261.

(59) Miller, J. F.; Shah, N. N.; Nelson, C. M.; Ludlow, J. M.; Clark, D. S. *Pressure and Temperature Effects on Growth and Methane Production of the Extreme Thermophile Methanococcus jannaschii* ; 1988; Vol. 54; pp 3039–3042.

(60) Takami, H.; Inoue, A.; Fuji, F.; Horikoshi, K. Microbial flora in the deepest sea mud of the Mariana trench. *FEMS Microbiology Letters* **1997**, *152*, 279–285.

(61) Godfroy, A.; Lesongeur, F.; Raguénès, G.; Quérellou, J.; Antoine, E.; Meunier, J. R.; Guezennec, J.; Barbier, G. Thermococcus hydrothermalis sp. nov., a new hyperthermophilic archaeon isolated from a deep-sea hydrothermal vent. *International Journal of Systematic Bacteriology* **1997**, *47*, 622–626.

(62) Kurr, M.; Huber, R.; König, H.; Jannasch, H. W.; Fricke, H.; Trincone, A.; Kristjansson, J. K.; Stetter, K. O. Methanopyrus kandleri, gen. and sp. nov. represents a novel group of hyperthermophilic methanogens, growing at 110°C. *Archives of Microbiology* **1991**, *156*, 239–247.

(63) Gao, H.; Obraztova, A.; Stewart, N.; Popa, R.; Fredrickson, J. K.; Tiedje, J. M.; Nealson, K. H.; Zhou, J. Shewanella loihica sp. nov., isolated from iron-rich microbial mats in the Pacific Ocean. *International Journal of Systematic and Evolutionary Microbiology* **2006**, *56*, 1911–1916.

(64) Masanari, M.; Wakai, S.; Ishida, M.; Kato, C.; Sambongi, Y. Correlation between the optimal growth pressures of four Shewanella species and the stabilities of their cytochromes c5. *Extremophiles* **2014**, *18*, 617–627.

(65) Bernhardt, G.; Jaenicke, R.; Lüdemann, H.-D.; König, H.; Stetter, K. O. High Pressure Enhances the Growth Rate of the Thermophilic Archaebacterium Methanococcus thermolithotrophicus without Extending Its Temperature Range. *Applied and Environmental Microbiology* **1988**, *54*, 1258–1261.

(66) Pikuta, E. V.; Marsic, D.; Itoh, T.; Bej, A. K.; Tang, J.; Whitman, W. B.; Ng, J. D.; Garriott, O. K.; Hoover, R. B. Thermococcus thioreducens sp. nov., a novel hyperthermophilic, obligately sulfur-reducing archaeon from a deep-sea hydrothermal vent. *International Journal of Systematic and Evolutionary Microbiology* **2007**, *57*, 1612–1618.

(67) Lu, J.; Nogi, Y.; Takami, H. Oceanobacillus iheyensis gen. nov., sp. nov., a deep-sea extremely halotolerant and alkaliphilic species isolated from a depth of 1050 m on the Iheya Ridge. *FEMS Microbiology Letters* **2001**, *205*, 291–297.

(68) Götz, D.; Banta, A.; Beveridge, T. J.; Rushdi, A. I.; Simoneit, B. R. T.; Reysenbach, A. L. Persephonella marina gen. nov., sp. nov. and Persephonella guaymasensis sp . nov ., two novel , thermophilic , hydrogen-oxidizing microaerophiles from deep-sea hydrothermal. *Microbiology, International Journal of Systematic and Evolutionary* **2002**, *52*, 1349–1359.

(69) DeLong, E. F.; Franks, D. G.; Yayanos, A. A. Evolutionary relationships of cultivated psychrophilic and barophilic deep-sea bacteria. *Applied and Environmental Microbiology* **1997**, *63*, 2105–2108.

(70) Nogi, Y.; Masui, N.; Kato, C. Photobacterium profundum sp. nov., a new, moderately barophilic bacterial species isolated from a deep-sea sediment. *Extremophiles* **1998**, *2*, 1–8.

(71) Allen, E. E.; Facciotti, D.; Bartlett, D. H. Monounsaturated but not polyunsaturated fatty acids are required for growth of the deep-sea bacterium Photobacterium profun-

dum SS9 at high pressure and low temperature. *Applied and Environmental Microbiology* **1999**, *65*, 1710–1720.

(72) Donachie, S. P.; Hou, S.; Gregory, T. S.; Malahoff, A.; Alam, M. Idiomarina loihiensis sp. nov., a halophilic γ-Proteobacterium from the Lō'ihi submarine volcano, Hawai'i. *International Journal of Systematic and Evolutionary Microbiology* **2003**, *53*, 1873–1879.

(73) Tian, X. P.; Tang, S. K.; Dong, J. D.; Zhang, Y. Q.; Xu, L. H.; Zhang, S.; Li, W. J. Marinactinospora thermotolerans gen. nov., sp. nov., a marine actinomycete isolated from a sediment in the northern South China Sea. *International Journal of Systematic and Evolutionary Microbiology* **2009**, *59*, 948–952.

(74) Kato, C.; Li, L.; Nogi, Y.; Nakamura, Y.; Tamaoka, J.; Horikoshi, K. Extremely barophilic bacteria isolated from the Mariana trench, challenger deep, at a depth of 11,000 meters. *Applied and Environmental Microbiology* **1998**, *64*, 1510–1513.

(75) Birrien, J. L.; Zeng, X.; Jebbar, M.; Cambon-Bonavita, M. A.; Quérellou, J.; Oger, P.; Bienvenu, N.; Xiao, X.; Prieur, D. Pyrococcus yayanosii sp. nov., an obligate piezophilic hyperthermophilic archaeon isolated from a deep-sea hydrothermal vent. *International Journal of Systematic and Evolutionary Microbiology* **2011**, *61*, 2827–2831.

(76) Xu, Y.; Nogi, Y.; Kato, C.; Liang, Z.; Rüger, H. J.; De Kegel, D.; Glansdorff, N. Moritella profunda sp. nov. and Moritella abyssi sp. nov., two psychropiezophilic organisms isolated from deep Atlantic sediments. *International Journal of Systematic and Evolutionary Microbiology* **2003**, *53*, 533–538.

(77) Nogi, Y.; Kato, C.; Horikoshi, K. Taxonomic studies of deep-sea barophilic Shewanella strains and description of Shewanella violacea sp. nov. *Archives of Microbiology* **1998**, *170*, 331–338.

(78) Vetriani, C.; Speck, M. D.; Ellor, S. V.; Lutz, R. A.; Starovoytor, V. Thermovibrio ammonificans sp. nov., a thermophilic, chemolithotrophic, nitrate-ammonifying bacterium from deep-sea hydrothermal vents. *International Journal of Systematic and Evolutionary Microbiology* **2004**, *54*, 175–181.

(79) Huber, R.; Stöhr, J.; Hohenhaus, S.; Rachel, R.; Burggraf, S.; Jannasch, H. W.; Stetter, K. O. Thermococcus chitonophagus sp. nov., a novel, chitin-degrading, hyperthermophilic archaeum from a deep-sea hydrothermal vent environment. *Archives of Microbiology* **1995**, *164*, 255–264.

(80) Miroshnichenko, M. L.; Kostrikina, N. A.; Chernyh, N. A.; Pimenov, N. V.; Tourova, T. P.; Antipov, A. N.; Spring, S.; Stackebrandt, E.; Bonch-Osmolovskaya, E. A. Caldithrix abyssi gen. nov., sp. nov., a nitrate-reducing, thermophilic, anaerobic bacterium isolated from a Mid-Atlantic ridge hydrothermal vent, represents a novel bacterial lineage. *International Journal of Systematic and Evolutionary Microbiology* **2003**, *53*, 323–329.

(81) Antoine, E.; Cilia, V.; Meunier, J. R.; Guezennec, J.; Lesongeur, F.; Barbier, G. Thermosipho melanesiensis sp. nov., a new thermophilic anaerobic bacterium belonging to the order Thermotogales, isolated from deep-sea hydrothermal vents in the Southwestern Pacific Ocean. *International Journal of Systematic Bacteriology* **1997**, *47*, 1118–1123.

(82) Miura, T.; Abe, F.; Inoue, A.; Usami, R.; Horikoshi, K. Purification and characterization of novel extracellular endopolygalacturonases from a deep-sea yeast, Cryptococcus sp. N6, isolated from the Japan Trench. *Biotechnology Letters* **2001**, *23*, 1735–1739.

(83) Abe, F.; Minegishi, H.; Miura, T.; Nagahama, T.; Usami, R.; Horikoshi, K. Characterization of cold- and high-pressure-active polygalacturonases from a deep-sea yeast,

Cryptococcus liquefaciens strain N6. *Bioscience, Biotechnology and Biochemistry* **2006**, *70*, 296–299.

(84) Xiao, X.; Wang, P.; Zeng, X.; Bartlett, D. H.; Wang, F. Shewanella psychrophila sp. nov. and Shewanella piezotolerans sp. nov., isolated from west Pacific deep-sea sediment. *International Journal of Systematic and Evolutionary Microbiology* **2007**, *57*, 60–65.

(85) Takai, K.; Sugai, A.; Itoh, T.; Horikoshi, K. Palaeococcus ferrophilus gen. nov., sp. nov., a barophilic, hyperthermophilic archaeon from a deep-sea hydrothermal vent chimney. *International Journal of Systematic and Evolutionary Microbiology* **2000**, *50*, 489–500.

(86) Jeanthon, C.; L'Haridon, S.; Reysenbach, A. L.; Corre, E.; Vernet, M.; Messner, P.; Sleytr, U. B.; Prieur, D. Methanococcus vulcanius sp. nov., a novel hyperthermophilic methanogen isolated from East Pacific Rise, and identification of Methanococcus sp. DSM 4213(T) as Methanococcus fervens sp. nov. *International Journal of Systematic Bacteriology* **1999**, *49*, 583–589.

(87) Rose, A. S.; Hildebrand, P. W. NGL Viewer: A web application for molecular visualization. *Nucleic Acids Research* **2015**, *43*, W576–W579.

(88) Moura, A.; Savageau, M. A.; Alves, R. Relative Amino Acid Composition Signatures of Organisms and Environments. *PLoS ONE* **2013**, *8*.

(89) Hormoz, S. Amino acid composition of proteins reduces deleterious impact of mutations. *Scientific Reports* **2013**, *3*, 1–10.

(90) Suhre, K.; Claverie, J. M. Genomic correlates of hyperthermostability, an update. *Journal of Biological Chemistry* **2003**, *278*, 17198–17202.

(91) Baek, M.; Dimaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Park, H.; Adams, C.; Glassman, C. R.; Degiovanni, A.; Pereira, J. H. Accurate prediction of

protein structures and interactions using a three-track neural network. **2021**, *8754*, 1–13.

(92) Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**,