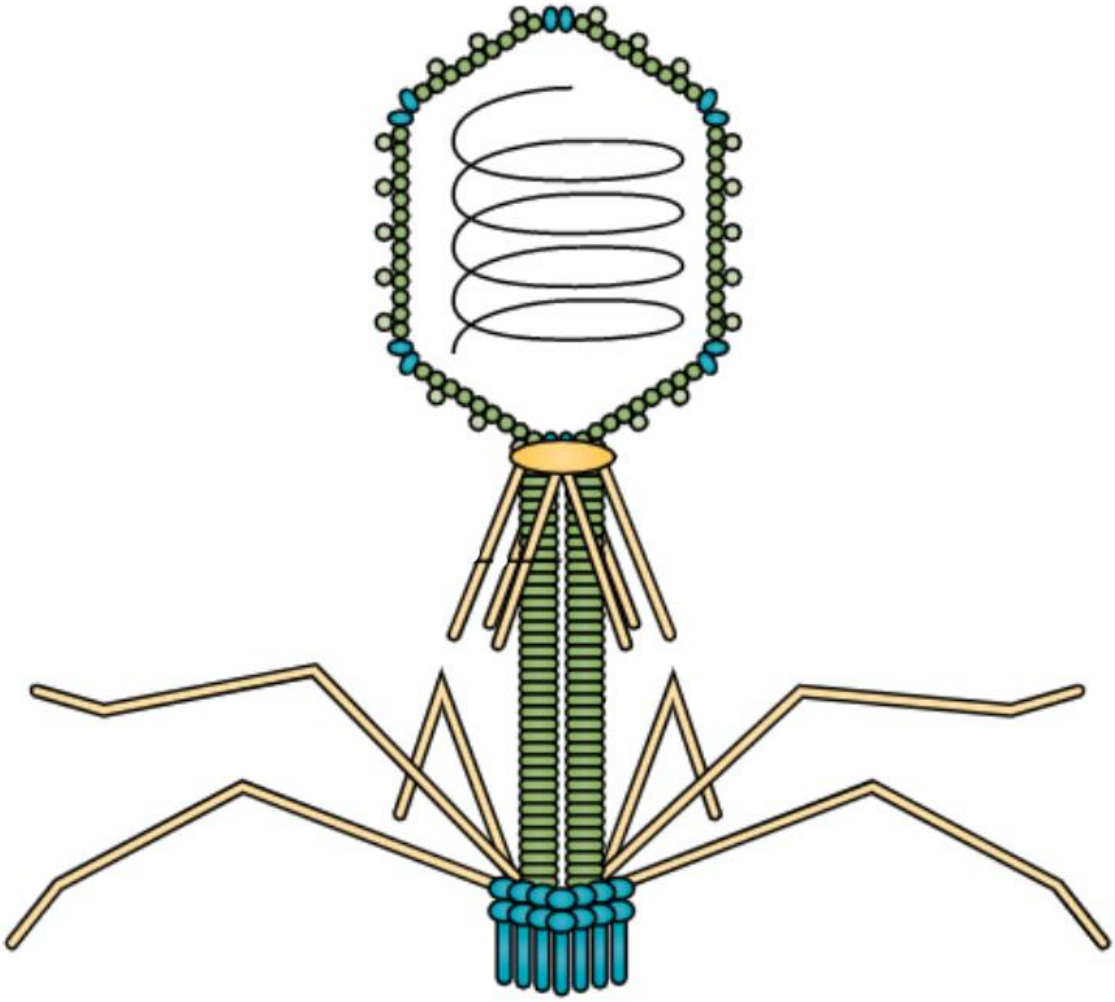


Phages in different thermal stages

A machine learning study on thermal stability and diversity of bacteriophage protein structures using AlphaFold.

Joran Schoorlemmer
Anne Kupczok
MSc Thesis
Joran.schoorlemmer@wur.nl

16-1-2023
Aalt-Jan van Dijk
Bioinformatics
1004586

Cover figure

Dion, M. B., Oechslin, F., & Moineau, S. (2020). Phage diversity, genomics and phylogeny. Nature Reviews Microbiology, 18(3), 125-138.

Index

Abstract

Phages, viruses that infect bacteria, are found in a wide range of environments and have developed advanced strategies to survive in these surroundings. To get a better understanding of their stability under thermal stress, this project performed analyses of the strategies phages use to withstand heat. Structural phage protein surfaces were classified and compared with each other. This gave insight into the wide diversity of these proteins and was used to predict thermostability using machine learning models. Using two ways of assessing thermostability, random forest models were created for proteins separated by structural class. Proteins were characterized using structural features such as the compactness or surface charge density of the protein and using sequential features retrieved using the deep learning embedding of UniRep. 75,567 structures of proteins were retrieved from the novel AlphaFold database and were checked for inaccuracies using a custom filtering pipeline, filtering out 22,843 low confidence entries and 23,454 loose structures. Model performance was found to inversely correlate with protein class diversity, indicating that within protein classes different strategies are used to withstand thermal stress. Combining different classes in one model led to lower predictive performance, confirming the high diversity between phage protein classes. The best performing model with an F1 score of 0.52 used structural features and 16S rRNA GC% estimated temperatures for the shaft class. This is far better than forced positive classification (F1=0.08) and showed the importance

of charged and turn surface residues in shaft proteins for thermostability. The use of phages in phage therapy to battle antibiotic-resistant bacteria and medicine delivery through phage design are very promising. However, problems regarding preparation and stabilization currently complicate the implementation of these phage applications. The novel characterizations of phage proteins in this project can be used to more accurately depict phages for phage therapy & design.

Introduction

Phages are viruses that infect bacteria. They are highly abundant and diverse biological entities. Phages are found in environments at many different temperatures, acidities and osmotic pressures [1]. Due to this diversity, they can be used in a plethora of situations. Phages consist of nucleic acids and proteins. They reproduce via the lytic cycle (Figure 1) or the lysogenic cycle in which they reproduce together with the host cell. The lytic cycle starts with a phage virion, the vehicle a phage uses outside of a host. A phage virion binds to a host bacterial cell and injects its genome into the cell. After injection, the phage DNA is replicated and proteins are translated. These are assembled into new virions and are released upon lysis of the cell [2]. When the phage is structured in its virion, it is open to external factors which could be a very dangerous phase for the phage. The surface proteins of the virion play a big part in its protection against these factors like extreme temperatures, low/high pH environments or organic solvents. They provide stability, which can be defined as the time a virion can remain infectious in the environment [3].

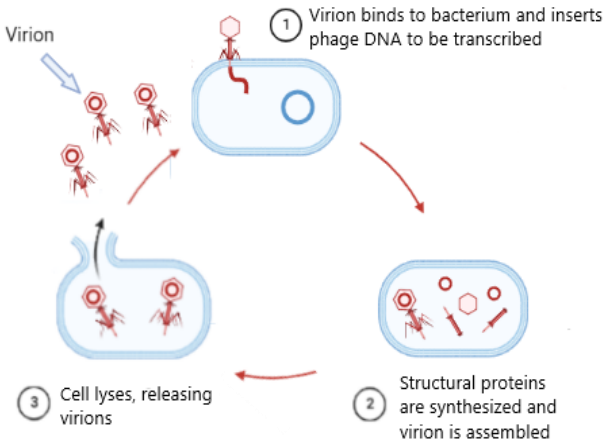


Figure 1: Lytic life cycle of a bacteriophage. 1) A phage binds to a bacterial cell. 2) The phage DNA is replicated and transcribed/translated and virion is assembled from structural proteins. 3) New virions which have been formed are released from the cell upon lysis. Created with Biorender.com

If it would be known how these proteins give rise to more stable phages, phages can be adapted or engineered to suit needs for specific high or low temperature situations. This is useful for various applications of phages. Phage therapy e.g. can be of great use to battle general bacterial infections and in non-bacterial situations with anti-inflammatory effects or by interacting with the immune system to protect human health [4]. The biggest impact however could be to use phage therapy to combat antibiotic-resistant bacteria [5]. The emergence of these bacteria is a big threat to the current healthcare system. Phage therapy could be a big step in resolving this issue. However, phage therapy still faces multiple challenges itself. These are mainly found in the preparation and stabilization of phages [6]. While some phages can be kept frozen, others show huge drops in infectivity. Some phages can be assembled quickly in high heat while others suffer from denaturation.

Currently, little is known about the structure of phage proteins in these differing temperatures. This can be explained due to the focus of research on other, non-bacterial, viruses which are found in environments with even more extreme conditions [7]. While some thermophilic phages are well known, these are more anecdotal examples. Broad studies on general properties in structural proteins of thermophilic phages are lacking [8], [9]. More information is available for other, non-bacterial viruses, which show for example a high number of disulfide bonds between coat proteins to enhance thermostability [10]. It is also known that thermophilic proteins in general show higher compactness and have more charged amino acid residues on their water-accessible surfaces [11]. Additionally, there

is some research on the survivability of phages in extreme conditions, such as different temperatures, but this does not take the 3D structure of the proteins into account. These studies mainly focus on genomic and morphological differences between phages [1].

With the availability of the novel AlphaFold [12] database, many new possibilities have opened up for in-depth analyses of protein structures. AlphaFold is a tool that can predict a protein 3D structure from an amino acid sequence with an unprecedentedly high accuracy when compared to other tools. It is among others being used to characterize structural elements, model interactions, investigate ligand binding sites [13] and opened up new possibilities in secondary structure predictions using machine learning [14]. Using this tool, over 200 million new structures have become available for use in the Protein Data Bank. This allows for characterization of phage protein structures on a large scale. This project focuses on the 3D structure of surface proteins as a basis for explaining thermal stability of phages. The surfaces are especially interesting as these are in direct contact with the environment and as such could explain more about potential interactions. It tries to answer how virion proteins differ for phages found in low/high-temperature conditions and how these proteins differ structurally. To achieve this, surfaces of virion proteins have to be characterized using structural features after which they can be compared through different protein classes and conditions.

Some structural features are known to correlate with thermal stability in general. These features will be used as a basis for this project to check for any predictive value towards thermostability in phage proteins. For example, the number of charged residues on the surface strongly increased at the expense of polar non-charged residues for thermophilic proteins [11]. As the surface charge density is highly related to the number of charged residues, it is also of interest. Other features like the secondary structure composition and packing density are found to differ between proteins found at different temperatures so can also be considered [15]. Lastly, an increase in protein compactness is also found when comparing mesophilic and thermophilic proteins [11]. Combining these features using a machine learning model, could paint a complete picture of the strategy phages use to enhance their thermostability.

Material & Methods

Data collection

358,797 sequences of phage structural proteins were retrieved from the curated database of PhANNs [16] on 12 September 2022. As these phage proteins have different functions, it can be expected that they have very different structures. This curated database has the big advantage that all proteins are already classified by 10 different functions which makes it easier to compare to each other. These classes are displayed in Figure 2. Protein structures in the same class are more similar, and they could also be impacted by temperature in a similar manner allowing for easier identification of specific strategies used per class. In the database, misannotated proteins were already manually removed by the creators of the PhANNs database. The proteins from PhANNs were retrieved from the NCBI database.

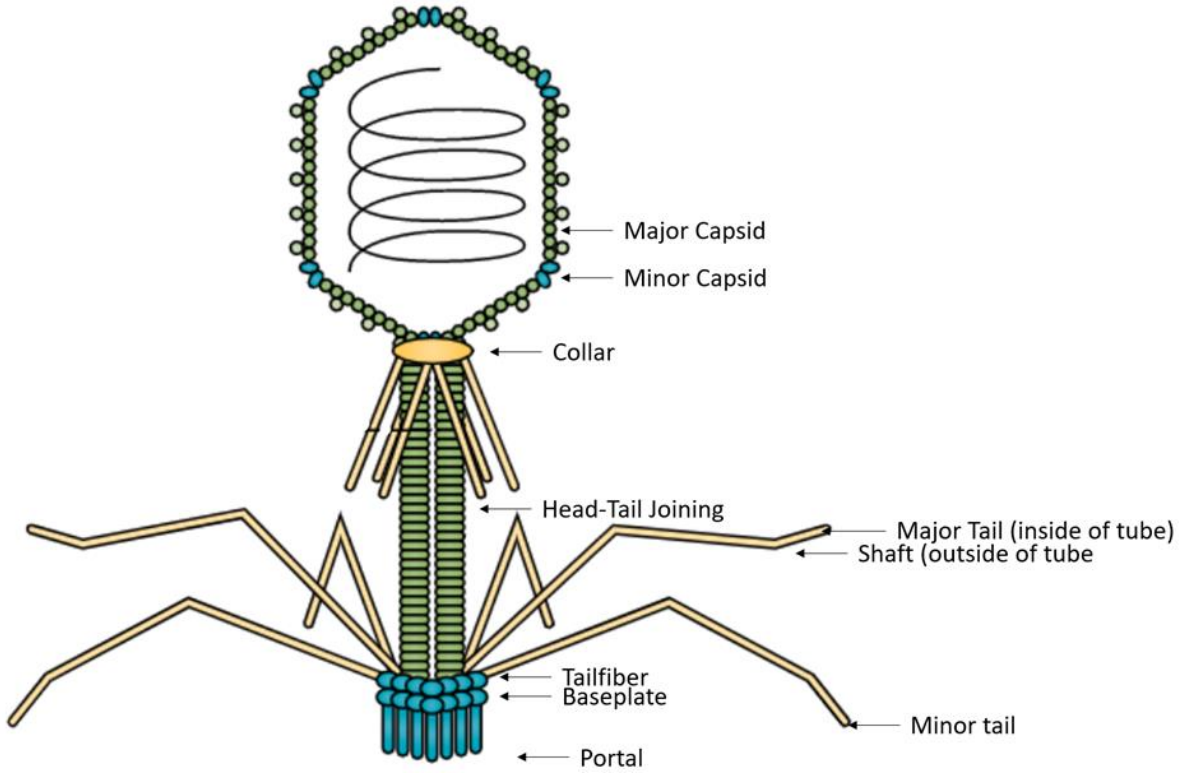


Figure 2: The 10 different structural classes used in this project for phage proteins. [16], [17]

The GenBank & RefSeq IDs from PhANNs were mapped to UniProt IDs for easier access of the Alphafold database using the UniProt online ID mapping tool. However, many of the GenBank and RefSeq entries have no stored cross reference and could thus not be mapped to UniProt. To overcome this, all GenBank and RefSeq entries were mapped to the Identical Protein Groups (IPG) database from NCBI [18]. This contains many identical proteins, of which many are cross-referenced in the UniProt database. After removing redundant entries, the new dataset was mapped to UniProt resulting in 121,695 proteins, a twofold increase compared to the original 64,279 proteins which could immediately be mapped due to the presence of a UniProt cross-reference.

Alphafold predictions for the UniProt database were retrieved from the Alphafold online database. As Alphafold is relatively new and updates are still being released, the structures were updated halfway through the project on 21 November 2022, so after the 1 November update. As not all UniProt entries have corresponding entries in the Alphafold database, the dataset currently consisted of 75,567 PDB files.

Data filtering

Alphafold provides a confidence score per amino acid residue, which can be used to filter out low quality structures. A residue with a confidence score lower than 70% was defined as a low confidence residue (LCR). While one individual LCR does not have a big influence on the overall structure, multiple together can drastically change the protein structure and its surface, which is crucial for this project. Multiple thresholds were set to create different quality classifiers for the proteins. First, 22,843 structures with over 25% LCRs were removed. Next, LCRs at the terminals of the proteins were removed as these are expected to have little influence on the total structure of the protein, while still clouding overall surface patterns. Terminals were taken as 10% of the residues of the proteins. LCRs

were removed starting from the ends until a residue with high (>70%) confidence was reached or until the 10% boundary was reached. 10,966 proteins contained no more LCRs after this step and were stored in the most stringent dataset, ready for use in this project. For the next filtering condition, no continuous stretch of 15 LCRs in the protein was allowed. This threshold was chosen due to the nature of the protein backbone forming closed loops [19]. Proteins with such a continuous stretch can be too flexible to be trusted as an accurate structure as shown in Figure 3. 16,271 proteins without such a stretch were stored in the most stringent dataset for further analyses, while the others were checked for LCRs in secondary structures. These secondary structures can have a large influence on the overall structure of the protein. Secondary structures were assigned using DSSP 2.3.0 [20] with default parameters. 2,033 proteins with less than 20% of their LCRs in secondary structures were stored in the most stringent dataset. For the remaining 23,454 proteins, the structure was cut off at the ends of the 15 LCR stretches and the biggest substructure was kept in a separate dataset. These were again checked for LCRs in secondary structures resulting in 0 proteins lost due to this step. Unless stated otherwise, the most strictly filtered dataset was used in the next steps of this project, consisting of 29,270 protein structures. Filtering steps are shown in Figure 4.

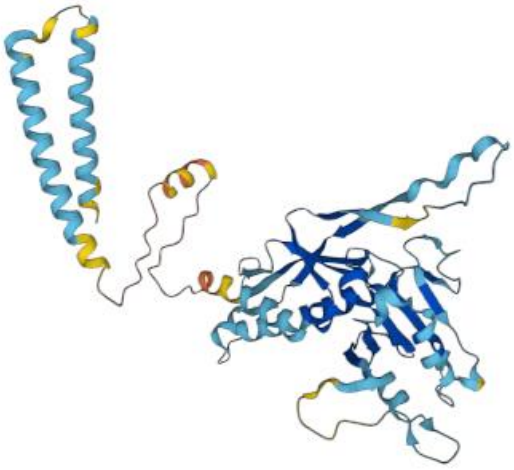


Figure 3: Alphafold prediction of UniProt entry A0A0A0G5R6. Low confidence residues are shown in yellow/red. A big stretch of LCRs can be seen in the middle of the proteins, leading to a potential untruthful structure. For this reason, the LCR stretch and accompanying smallest domain are filtered out.

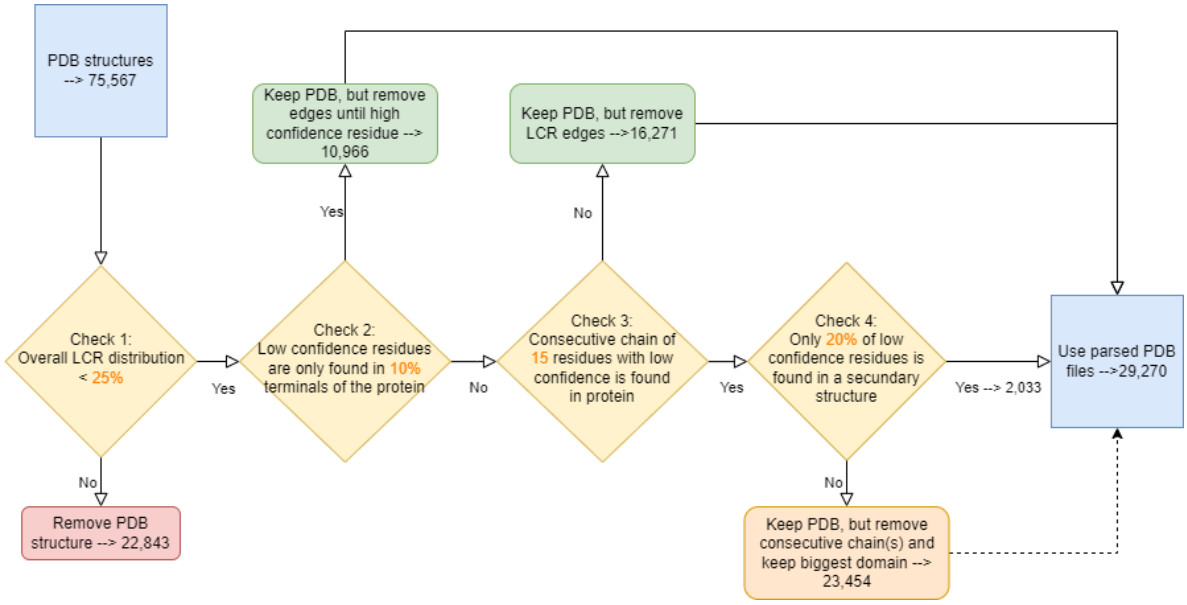


Figure 4: Filtering pipeline for curating and parsing AlphaFold structures based on confidence scores. The last step in the orange box parsed proteins into their biggest substructure, optionally available for further analysis.

Feature calculation

For characterizing the structures, features were selected and calculated using multiple different tools and custom python scripts. Surface accessibility of the amino acid residues was determined using NACCESS V2.1.1 [21] with default parameters. A residue was counted as a surface residue at a relative

solvent accessibility higher than 10.0. The residue was identified as polar, aromatic or charged based on Supplementary table A in the appendix. Compactness Z was calculated using the following equation [22].

$$Z = ASA_{Surf}/(36\pi VOL^2)^{1/3}$$

Here ASA_{Surf} is the accessible surface area (calculated with NACCESS), and VOL is the volume, calculated using ProteinVolume 1.3 [23] with a volume probe radius of 0.08 Å. Surface charge was determined using PDB2PQR 3.4.1 [24] disregarding solvent water molecules to speed up computation and APBS 1.5 [25] with default parameters. Output files from these various tools were parsed and combined using custom python scripts into one individual csv file for each protein function class. These scripts are stored on https://git.wur.nl/joran.schoorlemmer/js_bif_thesis.

Sequence-based features were retrieved to verify the findings from the structural features. This was done using the deep learning tool UniRep [26]. This tool uses amino acid embedding approaches to reveal information directly from the protein sequence and as such can be used to extract properties at the protein level. The model creating 64 dimensions was used, all 64 dimensions were extracted for each protein and were used as sequential features for further analysis in this project.

To get insight into the various features, descriptive statistics were performed. A principal component analysis (PCA) was performed per protein class to check the distribution of the proteins relating to the features. Next to this, features were compared through all ten classes using boxplots. For each feature, an ANOVA test was performed followed by a Tukey post-hoc test to check for significant differences.

Temperature prediction

For estimating the thermostability of the proteins, two approaches were used. First, the stability was determined based on empirical data of the host bacterium optimal growth temperature, T_{opt}. Alternatively, as the T_{opt} of the majority of the host bacteria is not known, the T_{opt} was predicted using the 16S rRNA stem GC content of the host bacterium.

The TEMPURA [27] database provides a concise overview of amongst others the T_{opt} of 8639 prokaryotic strains. The host organism of the phage belonging to each protein was determined with a custom python script using the package bioservices 1.10.4 [28] which uses the protein entry in UniProt. Then, the host organism was mapped incrementally to the bacteria in TEMPURA based on its taxonomic rank. First, the species level was checked for an entry in TEMPURA. If this was not present, the average value of organisms with the same genus was taken. Due to some hosts being annotated at the strain level, these had to be mapped to the species level. Navigating through the different ranks of the organisms was done using the package ete3 3.1.2 [29] which is based on the NCBI taxonomy database. Because some genera show big differences in T_{opt} for their different species, the growth temperatures were divided in two blocks. These blocks are mesophilic (<40°C) and thermophilic (>40°C)[7]. If more than 80% of the species in a genus are in a single block, all species in the genus, which did not have a strain or species value yet, are assigned to have that T_{opt}. This way, species which are not included in TEMPURA but have related species from the same genus can still be mapped. If a genus has less than 80% of its species in a single block, the genus level is regarded as none existing.

Because there is a lot of uncertainty in this approach, T_{opt} was also estimated using the 16S rRNA stem GC content of the host bacterium. Previous research showed a high correlation between GC content and the optimal growth temperature [30], [31]. 16S rRNA sequences were retrieved for the UniProt host species using NCBI Bioproject 33175 if present in this database. Using the ViennaRNA package 2.5.0a5 [32], the secondary structure of the rRNA sequence was predicted. The GC content of the stem parts of the sequence was calculated using biopython 1.79 [33]. A least squares linear regression of

the T_{opt} and stem GC content of the TEMPURA entries, shown in Supplementary figure A, was performed. Using this relation as the equation shown below, T_{opt} values for new entries without a TEMPURA entry can be estimated.

$$T_{optimal}(°C) = 1.53 * GC_{rRNA}(\%) - 65.07$$

The T_{opt} output was classified using the same temperature blocks as in the genus mapping approach to decrease noise and simplify comparing the two. The number of proteins which could be characterized as mesophilic or thermophilic are shown in Table 1. Some host organisms did not have an entry in the used 16S rRNA database, while others did not have a related entry in the TEMPURA database accounting for a loss in proteins to be interpreted. The overlapping entries which could be mapped by both approaches are also shown in Table 1. Point of notice is the low number of thermophilic proteins which could be mapped using both approaches.

Table 1: Number of proteins which could be predicted by mapping to the optimal growth temperature of their host, either using empirical data from TEMPURA or calculating using its 16S rRNA stem GC content. The overlap between the two methods is also shown.

Predicted by	Total	Mesophilic	Thermophilic
Taxonomy: species	7,655	7,028	625
Taxonomy: genus	29,892	29,511	383
Stem 16S rRNA GC%	39,623	39,008	615
Both taxonomy & 16S rRNA	28,932	27,653	134
Total proteins in dataset	52,640		

Predicting thermostability using random forest classifiers

For all ten protein classes, a random forest model was created using the randomForest package 4.7 in R. Thermophilic proteins were labeled as positive, mesophilic proteins were labeled as negative. The dataset was divided using an 80/20 split with 80% of the data being used as training data and 20% as testing data. Due to the low proportion of thermophilic proteins in the data, stratified sampling was used to ensure the presence of thermophilic proteins in the test set. The model was created using 4000 trees, the drawn samples were sampled with the same size as the number of thermophilic proteins in the training set to prevent a large overrepresentation of the mesophilic proteins. To counteract this possible bias further, the thermophilic protein type was weighted 10x as heavily as opposed to the mesophilic proteins when the majority vote is taken in a tree. Although this had little effect on model performance after stratified sampling, it was kept in the model to ensure the smallest bias possible. The importance of all structural features was retrieved from the models to explain underlying patterns. This was expressed in MeanDecreaseGini, which is a measure of the decrease in impurity upon adding an individual feature to the model.

Identity in classes

To investigate the influence of protein similarity on the predictive performance, pairwise sequence identity scores were calculated by aligning high similarity regions as in previous research [34]. The averages of these scores per class, and the highest identity score per class, were used as an indication of the similarity in a class. Due to the large number of sequences, a fast alignment tool was necessary. MMseqs2 [35] is a very fast local alignment tool enabling analysis of massive data sets. Because it is mainly used for searching for high identity homologs in large databases, some alterations had to be made to bypass the core prefiltering module of the tool. This module filters out low identity pairs to speed up computation, while these are just as important for this project. A shell function from the GitHub page of MMseqs2 was used to convert the input FASTA file into a “filtered” output file which

could be used for the alignment step. No e-value cut-off was used to make sure all sequence pairs were used.

The shaft and major capsid were analyzed in depth using the pfam_scan 1.6 tool from Pfam [36]. The composition and protein families making up the class were retrieved and checked for correlation with prediction results.

Results

Phage proteins were retrieved and separated by one of the ten structural classes according to PhANNs. Diversity in these classes was assessed and further analyses of the proteins were done separately for each class. Only proteins from the most stringent dataset were used unless stated otherwise. Two classes are not shown in the results section due to the low presence of thermophilic proteins, being the collar proteins and the minortail proteins. These contained less than 10 thermophilic proteins in both temperature prediction approaches, using the species or genus value from TEMPURA or by calculating T_{opt} using the stem GC content of the 16S rRNA sequence. All figures and tables with these two missing classes are fully shown in the Appendix.

Exploring features

To investigate the difference in the calculated structural features, a PCA was performed for each protein class. The Head-Tail Junction protein is shown in Figure 5. The other classes show similar patterns and can be found in Supplementary figure B in the Appendix.

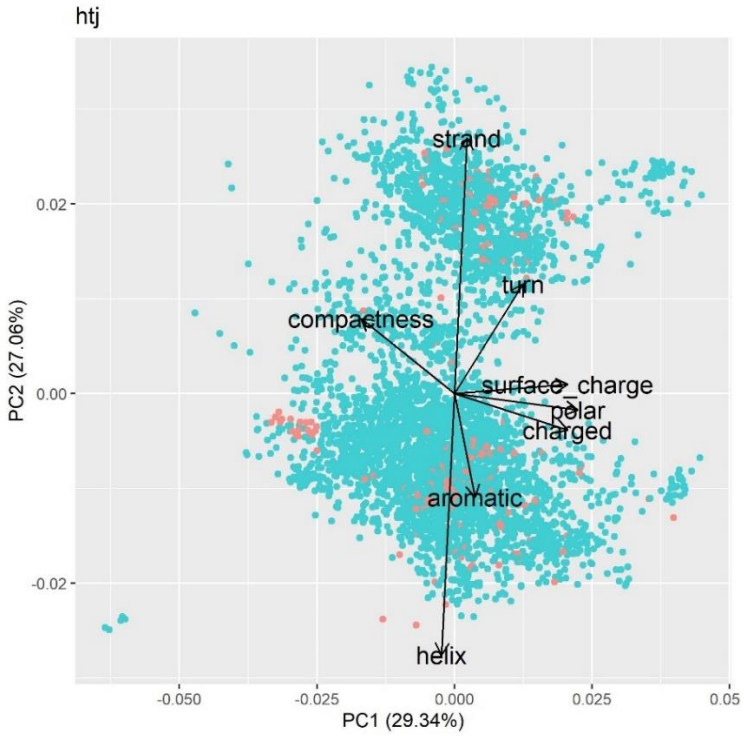


Figure 5: Biplot of the two first principal components of a PCA of the head-tail junction proteins. Mesophilic proteins are shown in blue, thermophilic proteins in red.

Some relations between the different features can be observed in Figure 5. The three charge related features (charged, surface charge and polar) group together. These show a negative correlation with compactness. The number of surface residues in a helix is negatively related to the residues in strands and turns. Proteins are widely spread out over the principal components, although some grouping is visible between the proteins correlating with the helix loading and the strand + turn loadings. No

general pattern for the thermophilic proteins can be found from this plot, although it seems that some are grouping at the left of the figure.

Having obtained an overall view of diversity between the structural features, differences between the protein classes can be investigated. The diversity in the ratio of charged residues on the surface is shown in Figure 6. The seven other features are shown in Supplementary figure C in the Appendix.

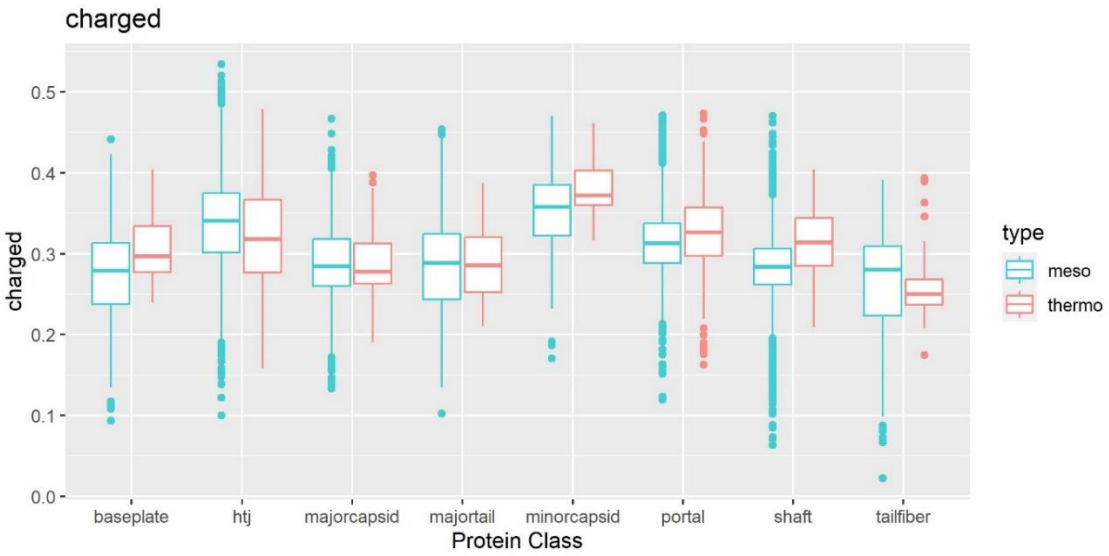


Figure 6: Distribution of the ratio of charged residues on the surface of proteins. Separated by protein class and thermostability. All differences between the classes are significant ($\alpha = 0.05$) according to a Tukey-test after an ANOVA except for the pair Majortail - Majorcapsid. The only significant differences between the thermophilic & mesophilic blocks are found in the baseplate, shaft and tailfiber class.

As can be seen in Figure 6, there are significant structural differences between the protein classes. This could also be a result of the high number of proteins in each class. The differences between the mesophilic and thermophilic proteins are less clear, which is reflected in the low number of significant differences between these groups.

Model performance in predicting thermostability

The different structural features are used to train random forest models to predict the thermostability of proteins. These are trained using both methods of T_{opt} estimations. Models were also trained using the sequence features from UniRep. Results are given in Table 2. Due to the low presence of thermophilic proteins for some protein classes, proteins from all confidence levels are used, except for proteins with a LCR content of more than 25% as these were filtered out immediately.

Table 2: F1 scores of random forest models, with the number of thermophilic proteins in the whole (training + test) dataset. A NaN score is given if the model returned zero True Positive predictions. Classes: BP (baseplate), HTJ (Head-Tail Junction), MJC (Major Capsid), MJT (Major Tail), MNC (Minor Capsid), P (Portal), S (Shaft), T (Tail fiber), All (All classes combined), All stringent (All classes combined but only the most strictly filtered structures).

T_{opt} estimation	Model	BP	HTJ	MJC	MJT	MNC	P	S	T	All	All stringent
Taxonomy	**Thermo proteins**	11	243	156	62	15	400	85	28	1,008	517
	F1 feature	0.02	0.37	0.17	0.13	0.14	0.25	0.09	0.25	0.21	0.19
	F1 sequence	*NaN*	0.35	0.21	0.09	0.13	0.18	0.12	0.35	0.20	0.13
16S rRNA	**Thermo proteins**	14	35	106	13	15	215	209	3	615	205
	F1 feature	0.05	0.05	0.12	0.10	0.30	0.14	0.52	*NaN*	0.15	0.14
	F1 sequence	0.08	0.03	0.15	0.04	0.33	0.12	0.52	*NaN*	0.17	0.13

Many models show a low F1 score, with some having no score at all. This is often the result of the size of the dataset, as there is often only one or a few thermophilic proteins in the test set. In case this single thermophilic protein is predicted incorrectly, no true positives are present resulting in the inability to calculate a F1 score. The structure-based models and the sequence-based models show very similar F1 scores for all protein classes. Multiple protein classes with a high amount of thermophilic protein entries show relatively high F1 scores, like the Head-Tail Junction class when using taxonomy based T_{opt} estimations and the Shaft class using 16S rRNA based T_{opt} estimations. However, some models with large amounts of data still show low scores. The biggest example of this is the combined classes model with 1,008 thermophilic proteins and a F1 score of 0.21 for the structural feature based, taxonomy model. The portal & major capsid protein models are also low performing in regard to their number of thermophilic proteins. The confusion matrix of the best performing structural feature-based model is shown in Table 3.

Table 3: Confusion matrix of Shaft protein predictions from a feature-based model using 16S rRNA T_{opt} data. True mesophilic proteins are shown in blue, while true thermophilic proteins are shown in red.

Ground truth\predictions	Mesophilic	Thermophilic
Mesophilic	936	52
Thermophilic	9	33

The model predicts little false negative proteins, but many false positive proteins. When comparing model performance to a baseline model which predicts everything as thermophilic, the F1 score drops drastically from 0.52 to 0.08. The importance of the different features for this model is shown in Table 4.

Table 4: Feature importance of the random forest model predicting shaft proteins using 16S rRNA T_{opt} estimations.

Feature	MeanDecreaseGini
turn	5.89
charged	5.48
polar	4.49
aromatic	3.86
helix	3.53
strand	3.07
compactness	2.64
surface_charge	2.55

All features show some impact with the highest for the ratio of charged surface residues and surface residues found in a turn. These features have the biggest impact on predicting the thermostability of a shaft protein correctly. The surface charge and compactness of the protein have a lower effect on the ability of the model to make correct predictions. Due to the high correlation in the features charged, polar and surface charge, feature importance and model performance were calculated without the first two features, charged & polar. However, both model performance and feature importance were very similar. When comparing feature importance to other high performing models in Supplementary table B, many differences can be found. Now, polar & aromatic residues and the surface charge become more important for the model. This could indicate different strategies being used per class for thermostability.

Sequence identity

To investigate the influence of protein similarity on model performance, local alignments were performed and pairwise identity scores were calculated. A summary per protein class is shown in Table 5.

Table 5: Summary of pairwise identities per protein class.

	Baseplate	HTJ	Major capsid	Major tail	Minor capsid	Portal	Shaft	Tailfiber
Min	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.12
1st Q	0.14	0.14	0.16	0.13	0.15	0.13	0.15	0.13
Median	0.15	0.17	0.18	0.14	0.16	0.15	0.16	0.14
Mean	0.20	0.24	0.21	0.19	0.24	0.18	0.24	0.16
3rd Q	0.20	0.33	0.23	0.17	0.22	0.18	0.29	0.15
Max	1	1	1	1	1	1	1	1

Similar values can be found across all classes, with low average identities for every class. However, when looking at the 3rd quantile, the shaft and head-tail junction protein classes show higher values. This indicates that the shaft and head-tail junction classes have relatively more proteins with a high similarity.

The distribution of pairwise identities between shaft proteins and between major capsid proteins are shown in Figure 7. These classes are chosen as they have a similar size and thermophilic protein composition but show very different F1 scores in Table 2. Other classes can be found in Supplementary figure D in the Appendix.

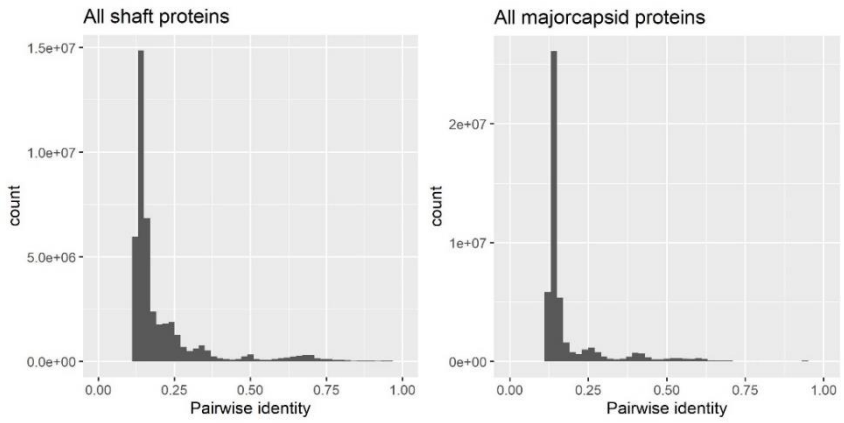


Figure 7: Distribution of pairwise identities for shaft proteins (left) and major capsid proteins (right).

No big difference in pairwise identities can be concluded from Figure 7. Many proteins show an identity score below 25%, which points to low similarity in the total dataset. To check for near duplicates in the dataset, the closest entries for each protein are also shown, in Figure 8.

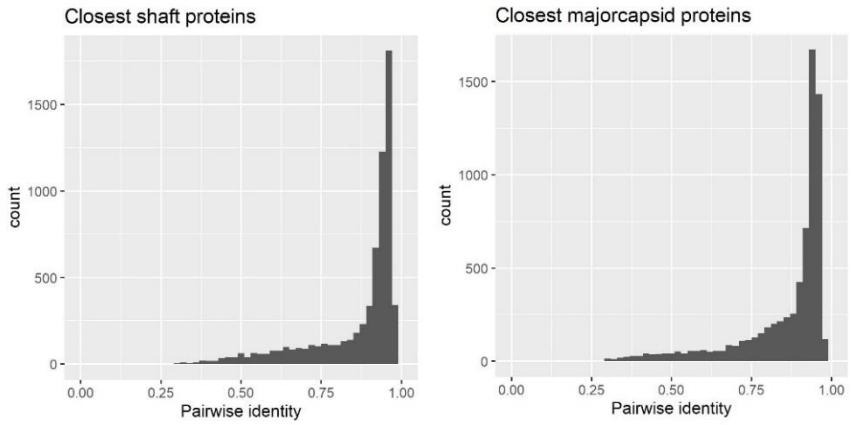


Figure 8: Distribution of pairwise identity of the most similar proteins for shaft proteins (left) and major capsid proteins (right).

Figure 8 once again highlights a similar pattern in both classes. Both have high identity scores for most proteins with fewer counts for low identity scores, indicating proteins with low similarity to every other protein in their respective class. However, upon a very close look, the shaft protein peak is a bit steeper, indicating slightly more proteins with high similarity. This is in line with the results shown in Table 5, in which it is shown that the 3rd quantile of the shaft class was greater relative to the major capsid class. The other protein classes can be seen in Supplementary figure E in the Appendix.

When comparing the composition of the classes using Pfam, a similar conclusion can be drawn. The shaft class consists of proteins containing domains originating from 15 protein families, of which only 8 have more than one individual entry in the class. The major capsid class contains 22 unique protein family domains, of which 17 have more than one individual entry amongst the class. This is an indication of a greater phylogenetic diversity in major capsid phage proteins.

Discussion

Interpretation and biologic context

Phages are diverse biological entities and can be found in almost every environment on earth. To be able to use them in phage therapy or for their anti-inflammatory effects, more in depth knowledge of their structure is necessary [6]. Their stability in different thermal conditions is of great importance for preparation and preservation of effective phage application. To assess this stability, a model was created to predict whether a phage can be classified as mesophilic (T_{opt} < 40°C) or thermophilic (T_{opt} ≥ 40°C). This way, specific phages can be analyzed and selected for previously mentioned applications. Additionally, this study provides in depth knowledge on phage protein structures and diversity in phage proteins.

The proteins were separated in ten classes, based on structural functions using the PhANNs [16] dataset. The diversity between these classes is high as is shown in Figure 6. The differences between the classes are often greater than the difference between the thermostability blocks within an individual class. This high diversity can be expected due to the different functions of these proteins [16], [17]. Different parts of the virion have different needs and interactions with their surroundings. Diversity is also reflected in the PCA, depicted in Figure 5. This is also reflected in the wide spread of these features in Supplementary figure B. Properties of proteins are found correlating with all features, although some grouping can be seen. This grouping behavior is especially clear between the helix and strand features.

Upon combining the information contained in all 10 features, random forest models were created to predict the thermostability of the phage proteins. The F1 scores shown in Table 2 differ a lot between all classes, but some interesting results can be retrieved. All high performing models contain relatively many thermophilic proteins. Differences between the two temperature estimation approaches are also showing this relation. This makes sense as the model needs a sufficient training set to accurately predict thermostability. Additionally, when few thermophilic proteins are present in the test set, the F1 score is easily skewed to low values due to false positives (mesophilic proteins predicted as thermophilic). When looking at the confusion matrix of shaft proteins in Table 3, it is clear that predicting the thermophilic proteins itself is done with a high accuracy but the F1 score is lowered by the large number of false positive mesophilic proteins. Compared to a baseline model predicting all proteins as thermophilic, model performance is greatly increased from 0.08 to 0.52 upon addition of the structural features. This indicates that these features contain and provide information on the thermostability of these proteins.

Some classes show low F1 scores, whilst having a large thermophilic presence in the dataset. This is especially true for the portal class and for the models containing all protein classes together. These contain the most thermophilic proteins but have remarkably low F1 scores. This could be due to the high diversity of these datasets. They show a low number of closely related proteins, indicated by their 3rd quantile in Table 5. This shows that the distribution of pairwise identities for shaft and head-tail junction proteins is skewed to higher values for relatively more proteins than the portal and major capsid class. The difference in diversity is also reflected in the composition of protein families in the dataset. The major capsid class contains much more domains from different families as opposed to the shaft class. This could also drive diversity and lower model performance. The models containing all classes perform also much worse, possibly due to the nature of the dataset being inherently diverse.

Little differences are found between the structural feature based and sequential based models. Both perform similarly across all protein classes and T_{opt} estimations. As only Alphafold structures have been used, which are predicted from an amino acid sequence, this could have been expected. It could even be argued that these two approaches contain the same information as it is known that a single polypeptide contains all information needed to fold into a 3D structure [37]. The difference in this project was however that when looking at specific residues, only surface residues were used. This is where the added benefit of the 3D structure is present and is why the novel AlphaFold data can be extremely impactful. As it is impossible to select for this in amino acid sequences, some differences in predictive performance were expected but are apparently not very significant if present at all.

Another interesting result is the similarity in performance of the strictly filtered dataset containing only highly confident Alphafold structures. The F1 score of the confident dataset is even slightly lower than the full dataset containing all classes, possibly due to the lower number of thermophilic proteins. Important to note is that for both datasets, the structures with low confidence residues over 25% of the total number of residues are filtered out. This ranges between 20-40% of structures being removed for each protein class. The following filtering steps have shown to be not very important for model performance in predicting thermostability. This could be due to the small difference in surface features and sequential features like mentioned before. Structures are filtered out because their predicted surface might not be representative of their actual structure. It could be that the presence of specific residues or secondary structural elements on the surface opposed to the whole protein is less impactful than initially thought. In this case, filtering becomes less important as it is less relevant if a specific residue is located on the surface or embedded in the structure.

These filtering steps could be more relevant in different projects, in which the surface of the protein is more important. These can for example be in protein-protein interaction research [38], production of

biopolymers[39] or in research on moving parts of proteins [40]. Especially for this last application, these filtering steps could be useful as it is known that Alphafold currently struggles with predicting flexible parts of proteins [41]. The filtering step which sets the most relevant restriction was related to this shortcoming of Alphafold. Circa 60% of the remaining structures at that point were filtered out due to the presence of a continuous stretch of over 15 LCRs in the middle of the structure. This could be due to a loose part of the protein, which is connecting two domains, resulting in a flexible shape-changing protein. The filtering step looking for proteins found in secondary structures had a very different impact, 92% of checked proteins were filtered out at this point. Almost all proteins apparently consisted of over 20% of LCRs in secondary structures. By raising this threshold, a more insightful filtering condition could be created as it could be too strict at 20%, allowing little flexibility.

Practical and computational limitations

Limitations of this project can be found in estimating the thermostability of a protein in the training dataset. This was done in two ways, first by retrieving optimal growing temperatures of host organisms genera from TEMPURA. Second, the T_{opt} was calculated using the 16S rRNA GC content and an empirical formula. As can be seen in Table 1, both methods already miss out on ±25% of protein entries. The two methods also show very different results on classifying a protein as thermophilic as only 134 were mapped as thermophilic by both methods. This also varies between classes, for example in the Head-Tail Junction class in Table 2. 243 proteins are classified as thermophilic using the genera mapping approach, while only 35 proteins are classified as thermophilic using the 16S rRNA approach. This discrepancy in the two approaches indicates that at least one method is not very thorough. However, this discrepancy can be seen in the shaft class as well, now with 85 thermophilic proteins using the taxonomic genera mapping approach and 209 using the 16S rRNA approach. Because this discrepancy differs per class, it can be assumed that both methods are not fully accurate. Although some correlation between the T_{opt} and species of the same genus can be expected [42], a big spread is present in many genera according to the TEMPURA database. For calculating T_{opt} using the 16S rRNA stem GC content, other complications arise. Due to computational difficulties of the ARB software needed to retrieve the actual stem regions of 16S rRNA strands, these were estimated using ViennaRNA. This could lead to inaccuracies in these stem regions and as such lead to inaccuracies in predicting T_{opt}. Although the 16S rRNA stem GC content is known to correlate with T_{opt} [27], [30], it is not precise enough for this type of analysis. To predict these more accurately, more novel tools can be used which use a combination of many genome derived features like the GC content, tRNA sequences, nucleotide or amino acid fractions and sequence length [43]. This way, using more accurate temperature values for host organisms, more accurate predictions can be made on the potential thermostability of phage proteins. A higher accuracy also allows for quantitative T_{opt} values instead of the current categorical approach. This could lead to new conclusions on phage adaptation strategies for more temperature ranges as it is known that these can differ a lot [11].

Another limitation of the project is the dataset size of certain protein classes. The baseplate, collar, minor capsid, minor tail proteins and tail fiber protein classes all contain little thermophilic proteins. This creates a problem as only a few proteins can be predicted in the test set, making verifying the quality of the model difficult. Next to this, model performance will suffer anyway due to the low number of proteins to train the model on. This size limitation could possibly be resolved by combining all classes together using a joint or ensemble learning model, ensuring the conservation of information on diversity but using the greater abundance of thermophilic proteins. This could also help resolve the large bias towards mesophilic proteins allowing for easier validation and optimization of the model. Previous research has used such an approach to predict DNA- and RNA-binding proteins, as these are relatively similar whilst being different proteins with different functions [44].

A last small limitation of this project is the use of a local alignment tool MMSeqs2 for calculating the pairwise identities. As comparing all 52,724 proteins using multiple sequence alignments or global sequence alignments was too computationally expensive, a local alignment was performed. This creates larger inaccuracies, as the alignment of small motifs is prioritized over the global alignment and as such gives lower overall identity values. While the resulting values will not differ drastically, it could result in a slightly different outcome.

Future outlook

For further research, more in depth analyses can be done on the relation of the wide diversity in protein classes to their thermostability. These classes differ widely in protein family domain compositions, which has a lot of influence on different strategies for thermostability [15]. Phages are also isolated from many different natural environments. They are retrieved from samples ranging from soil to marine and from human intestines to plant tissues. As many environmental factors next to temperature affect the efficacy and stability of phages [45], it could be interesting to include metadata of phage proteins in future work. Next to this, a similar pipeline as used in this project could be used to predict stability for different environmental factors like pH, organic matter content or salinity. When using more accurate temperature predictions, better performing models can be created. These models can be improved even further using more structural features like distributions of hydrogen bonds [15] or specific angles of aromatic features [8]. Next to this, structure embedding applications like Geometricus [46] or other sequence embedding tools like the architecture behind ProtREP [47] can add broader, more full representations of the protein. Extending this research to more broader applications than just prokaryotic phage therapy and phage design, it could be interesting to advance into the realm of eukaryotic viruses. Some research is already being done on these viruses and how they enhance thermostability [48]. Host interactions are a key part of eukaryotic virus strategies to boost virion stability, which could yield interesting different applications opposed to prokaryotic phages. Lastly, more research has to be done on the precise influence of thermostability or the lack thereof on phage infectivity to model and use phages to their fullest extent in phage therapy or their other applications.

Conclusion

This study characterized surfaces of phage protein structures by features such as the surface charge and secondary structure composition and used these protein structures to assess their diversity and predict thermostability of said proteins. This was done using the novel Alphafold protein structures, which were checked for inaccuracies using a custom filtering pipeline. For ten structural protein classes, random forest models were created to predict their thermostability. This was defined as the optimal growth temperature of their respective host organism, which was estimated for training in two ways. First by a taxonomic approach in mapping the genus wide average T_{opt} to the whole genus. Second, by calculating T_{opt} using the GC content of the 16S rRNA stem. Model performance differs greatly between the structural classes and combining all classes together in one single model hinders performance considerably. This is likely due to the increase in diversity of the dataset. As some of these classes are more diverse than others, assessing one individual strategy for thermostability is difficult. The most promising class is the shaft protein class with a highest F1 score of 0.52. The most important properties for this class were the number of residues found in turn elements and the number of charged residues on the surface. This knowledge helps in the ability to isolate or design stable phages for phage therapy. Information on their stability is crucial to ensure proper conservation and production. However, more research is required to cover all phage survival strategies to endure high temperature environments.

References

[1] A. Jurczak-Kurek *et al.*, "Biodiversity of bacteriophages: morphological and biological properties of a large group of phages isolated from urban sewage.," *Sci Rep*, vol. 6, p. 34338, Oct. 2016, doi: 10.1038/srep34338.

[2] G. P. C. Salmond and P. C. Fineran, "A century of the phage: Past, present and future," *Nature Reviews Microbiology*, vol. 13, no. 12. Nature Publishing Group, pp. 777–786, Dec. 01, 2015. doi: 10.1038/nrmicro3564.

[3] J. P. DeLong *et al.*, "Towards an integrative view of virus phenotypes," *Nature Reviews Microbiology 2021 20:2*, vol. 20, no. 2, pp. 83–94, Sep. 2021, doi: 10.1038/s41579-021-00612-w.

[4] A. P. Hynes *et al.*, "Perspectives of Phage Therapy in Non-bacterial Infections," 2019, doi: 10.3389/fmicb.2018.03306.

[5] S. A. Strathdee, G. F. Hatfull, V. K. Mutalik, and R. T. Schooley, "Phage therapy: From biological mechanisms to future directions," *Cell*, vol. 186, no. 1, pp. 17–31, Jan. 2023, doi: 10.1016/J.CELL.2022.11.017.

[6] E. Jończyk-Matysiak *et al.*, "Expert Review of Anti-infective Therapy Factors determining phage stability/activity: challenges in practical phage application Factors determining phage stability/activity: challenges in practical phage application," 2019, doi: 10.1080/14787210.2019.1646126.

[7] O. Zablocki, L. van Zyl, and M. Trindade, "Biogeography and taxonomic overview of terrestrial hot spring thermophilic phages," *Extremophiles*, vol. 22, no. 6, pp. 827–837, Nov. 2018, doi: 10.1007/s00792-018-1052-5.

[8] M. Tsuboi, J. M. Benevides, P. Bondre, and G. J. Thomas, "Structural details of the thermophilic filamentous bacteriophage PH75 determined by polarized Raman microspectroscopy," *Biochemistry*, vol. 44, no. 12, pp. 4861–4869, Mar. 2005, doi: 10.1021/bi0479306.

[9] B. Liu, F. Zhou, S. Wu, Y. Xu, and X. Zhang, "Genomic and proteomic characterization of a thermophilic Geobacillus bacteriophage GBSV1," *Res Microbiol*, vol. 160, no. 2, pp. 166–171, Mar. 2009, doi: 10.1016/J.RESMIC.2008.12.005.

[10] J. C. Caldeira and D. S. Peabody, "Stability and assembly in vitro of bacteriophage PP7 virus-like particles," *J Nanobiotechnology*, vol. 5, no. 1, pp. 1–10, Nov. 2007, doi: 10.1186/1477-3155-5-10.

[11] C. Cambillau and J. M. Claverie, "Structural and genomic correlates of hyperthermostability," *Journal of Biological Chemistry*, vol. 275, no. 42, pp. 32383–32386, Oct. 2000, doi: 10.1074/jbc.C000497200.

[12] J. Jumper *et al.*, "Highly accurate protein structure prediction with AlphaFold," *Nature 2021 596:7873*, vol. 596, no. 7873, pp. 583–589, Jul. 2021, doi: 10.1038/s41586-021-03819-2.

[13] M. Akdel *et al.*, "A structural biology community assessment of AlphaFold2 applications," *Nature Structural & Molecular Biology 2022 29:11*, vol. 29, no. 11, pp. 1056–1067, Nov. 2022, doi: 10.1038/s41594-022-00849-w.

[14] D. P. Ismi, R. Pulungan, and Afiahayati, “Deep learning for protein secondary structure prediction: Pre and post-AlphaFold,” *Comput Struct Biotechnol J*, vol. 20, pp. 6271–6286, Jan. 2022, doi: 10.1016/J.CSBJ.2022.11.012.

[15] A. Szilágyi and P. Závodszky, “Structural differences between mesophilic, moderately thermophilic and extremely thermophilic protein subunits: results of a comprehensive survey,” 2000. doi: 10.1016/S0969-2126(00)00133-7.

[16] V. A. Cantuid *et al.*, “PhANNs, a fast and accurate tool and web server to classify phage structural proteins,” 2020, doi: 10.1371/journal.pcbi.1007845.

[17] M. B. Dion, F. Oechslin, and S. Moineau, “Phage diversity, genomics and phylogeny,” *Nature Reviews Microbiology 2020 18:3*, vol. 18, no. 3, pp. 125–138, Feb. 2020, doi: 10.1038/s41579-019-0311-5.

[18] N. C. for B. I. 2004 Bethesda (MD): National Library of Medicine (US), “IPG [Internet],” Oct. 27, 2022. https://www.ncbi.nlm.nih.gov/ipg/ (accessed Oct. 27, 2022).

[19] Z. W. Tan, W.-V. Tee, E. Guarnera, and I. N. Berezovsky, “AlloMAPS 2: allosteric fingerprints of the AlphaFold and Pfam-trRosetta predicted structures for engineering and design,” *Nucleic Acids Res*, Sep. 2022, doi: 10.1093/nar/gkac828.

[20] W. Kabsch and C. Sander, “Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features,” *Biopolymers*, vol. 22, no. 12, pp. 2577–2637, 1983, doi: 10.1002/bip.360221211.

[21] S. J. Hubbard and J. M. Thornton, “NACCESS.” Department of Biochemistry and Molecular Biology, University College, London, 1993.

[22] C. J. Tsai and R. Nussinov, “Hydrophobic folding units derived from dissimilar monomer structures and their interactions,” *Protein Science*, vol. 6, no. 1, pp. 24–42, 1997, doi: 10.1002/PRO.5560060104.

[23] C. R. Chen and G. I. Makhatadze, “ProteinVolume: Calculating molecular van der Waals and void volumes in proteins,” *BMC Bioinformatics*, vol. 16, no. 1, pp. 1–6, Dec. 2015, doi: 10.1186/s12859-015-0531-2.

[24] T. J. Dolinsky *et al.*, “PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations,” *Nucleic Acids Res*, vol. 35, no. suppl_2, pp. W522–W525, Jul. 2007, doi: 10.1093/NAR/GKM276.

[25] E. Jurrus *et al.*, “Improvements to the APBS biomolecular solvation software suite,” *Protein Science*, vol. 27, no. 1, pp. 112–128, Jan. 2018, doi: 10.1002/PRO.3280.

[26] E. C. Alley, G. Khimulya, S. Biswas, M. AlQuraishi, and G. M. Church, “Unified rational protein engineering with sequence-based deep representation learning,” *Nature Methods 2019 16:12*, vol. 16, no. 12, pp. 1315–1322, Oct. 2019, doi: 10.1038/s41592-019-0598-1.

[27] Y. Sato, K. Okano, H. Kimura, and K. Honda, “TEMPURA: Database of Growth TEMPeratures of Usual and RAre Prokaryotes,” *Microbes Environ*, vol. 35, no. 3, pp. 1–3, 2020, doi: 10.1264/JSME2.ME20074.

[28] T. Cokelaer, D. Pultz, L. M. Harder, J. Serra-Musach, J. Saez-Rodriguez, and A. Valencia, "BioServices: A common Python package to access biological Web Services programmatically," *Bioinformatics*, vol. 29, no. 24, pp. 3241–3242, Dec. 2013, doi: 10.1093/bioinformatics/btt547.

[29] J. Huerta-Cepas, F. Serra, and P. Bork, "ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data," *Mol Biol Evol*, vol. 33, no. 6, pp. 1635–1638, Jun. 2016, doi: 10.1093/MOLBEV/MSW046.

[30] H. C. Wang, X. Xia, and D. Hickey, "Thermal adaptation of the small subunit ribosomal RNA gene: A comparative study," *J Mol Evol*, vol. 63, no. 1, pp. 120–126, Jul. 2006, doi: 10.1007/s00239-005-0255-4.

[31] Y. Gao and M. Wu, "Microbial genomic trait evolution is dominated by frequent and rare pulsed evolution," *Sci Adv*, vol. 8, no. 28, p. 1916, Jul. 2022, doi: 10.1126/sciadv.abn1916.

[32] R. Lorenz, I. L. Hofacker, and P. F. Stadler, "RNA folding with hard and soft constraints," *Algorithms for Molecular Biology*, vol. 11, no. 1, pp. 1–13, Apr. 2016, doi: 10.1186/S13015-016-0070-Z.

[33] P. J. A. Cock *et al.*, "Biopython: freely available Python tools for computational molecular biology and bioinformatics," vol. 25, no. 11, pp. 1422–1423, 2009, doi: 10.1093/bioinformatics/btp163.

[34] S. Daberdaku and C. Ferrari, "Exploring the potential of 3D Zernike descriptors and SVM for protein-protein interface prediction," *BMC Bioinformatics*, vol. 19, no. 1, pp. 1–23, Feb. 2018, doi: 10.1186/S12859-018-2043-3.

[35] M. Steinegger and J. Söding, "MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets," *Nature Biotechnology*, vol. 35, no. 11. Nature Publishing Group, pp. 1026–1028, Nov. 01, 2017. doi: 10.1038/nbt.3988.

[36] J. Mistry *et al.*, "Pfam: The protein families database in 2021," *Nucleic Acids Res*, vol. 49, no. D1, pp. D412–D419, Jan. 2021, doi: 10.1093/nar/gkaa913.

[37] C. B. ANFINSEN, E. HABER, M. SELA, and F. H. WHITE, "The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain.," *Proc Natl Acad Sci U S A*, vol. 47, no. 9, pp. 1309–1314, Sep. 1961, doi: 10.1073/pnas.47.9.1309.

[38] S. Jones and J. M. Thornton, "Review Principles of protein-protein interactions," 1996. [Online]. Available: https://www.pnas.org

[39] H. Chen, L. Yuan, W. Song, Z. Wu, and D. Li, "Biocompatible polymer materials: Role of protein-surface interactions," *Progress in Polymer Science (Oxford)*, vol. 33, no. 11. pp. 1059–1087, Nov. 2008. doi: 10.1016/j.progpolymsci.2008.07.006.

[40] J. W. Peng, "Exposing the moving parts of proteins with NMR spectroscopy," *Journal of Physical Chemistry Letters*, vol. 3, no. 8. pp. 1039–1051, Apr. 19, 2012. doi: 10.1021/jz3002103.

[41] A. Al-Janabi, "Has DeepMind's AlphaFold solved the protein folding problem?," *Biotechniques*, vol. 72, no. 3, pp. 73–76, Mar. 2022, doi: 10.2144/BTN-2022-0007.

[42] N. J. Russell, "Mechanisms of thermal adaptation in bacteria: blueprints for survival," 1984.

[43] D. B. Sauer and D. N. Wang, “Predicting the optimal growth temperatures of prokaryotes using only genome derived features,” *Bioinformatics*, vol. 35, no. 18, pp. 3224–3231, Sep. 2019, doi: 10.1093/BIOINFORMATICS/BTZ059.

[44] X. Du and J. Hu, “Deep Multi-Label Joint Learning for RNA and DNA-Binding Proteins Prediction,” *IEEE/ACM Trans Comput Biol Bioinform*, vol. 20, no. 1, pp. 307–320, Jan. 2023, doi: 10.1109/TCBB.2022.3150280.

[45] Y. J. Silva *et al.*, “Influence of environmental variables in the efficiency of phage therapy in aquaculture,” *Microb Biotechnol*, vol. 7, no. 5, pp. 401–413, 2014, doi: 10.1111/1751-7915.12090.

[46] J. Durairaj, M. Akdel, D. de Ridder, and A. D. J. van Dijk, “Geometricus represents protein structures as shape-mers derived from moment invariants,” *Bioinformatics*, vol. 36, pp. I718–I725, Dec. 2020, doi: 10.1093/bioinformatics/btaa839.

[47] M. L. Bileschi *et al.*, “Using deep learning to annotate the protein universe,” *Nat Biotechnol*, vol. 40, no. 6, pp. 932–937, Jun. 2022, doi: 10.1038/s41587-021-01179-w.

[48] A. K. Berger and B. A. Mainou, “Interactions between enteric bacteria and eukaryotic viruses impact the outcome of infection,” *Viruses*, vol. 10, no. 1. MDPI AG, Jan. 01, 2018. doi: 10.3390/v10010019.

[49] A. Bustamam, M. I. Sunggawa, and T. Siswantining, “Performance of multivariate mutual information and autocorrelation encoding methods for the prediction of protein-protein interactions,” *IAES International Journal of Artificial Intelligence (IJ-AI*, vol. 11, no. 2, pp. 773–786, 2022, doi: 10.11591/ijai.v11.i2.pp773-786.

Appendix

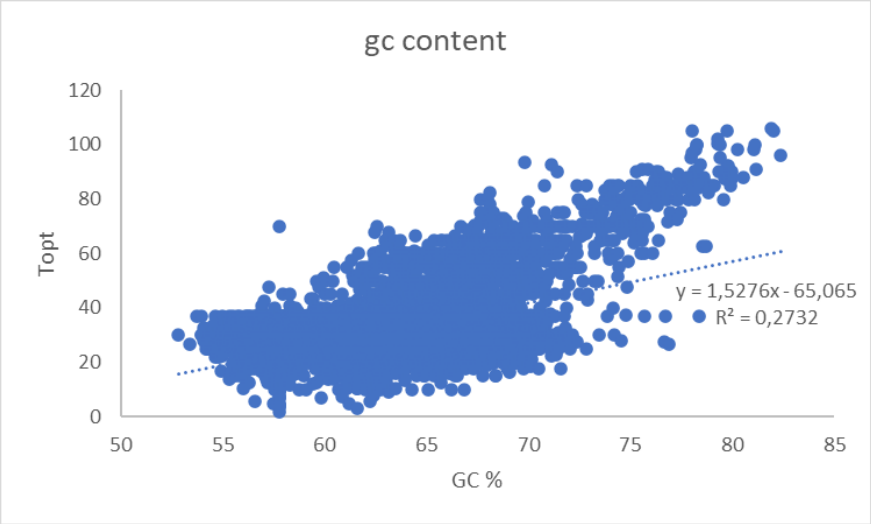
Supplementary table A: Amino acid residue types [49]

TYPE	RESIDUE
POLAR	ARG, HIS, LYS, ASP, GLU, SER, THR, ASN, GLN
CHARGED	ARG, HIS, LYS, ASP, GLU
AROMATIC	PHE, TYR, TRP, HIS

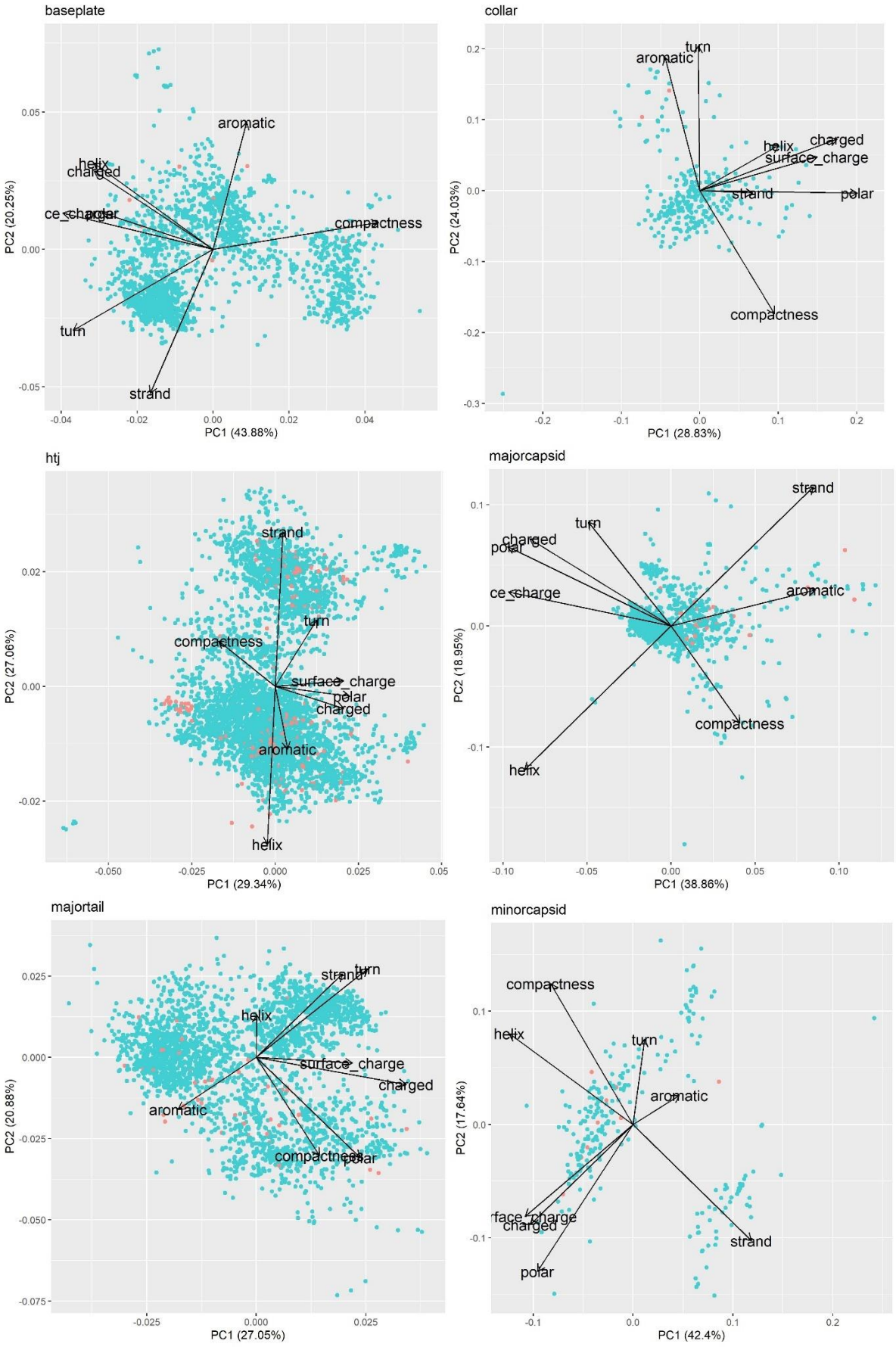
Supplementary table B: Feature importance of the random forest model predicting Head-Tail Junction proteins using taxonomic mapped T_{opt} estimations (left) and Minor capsid proteins using 16S rRNA T_{opt} estimations (right). These models are the second and third best performing models.

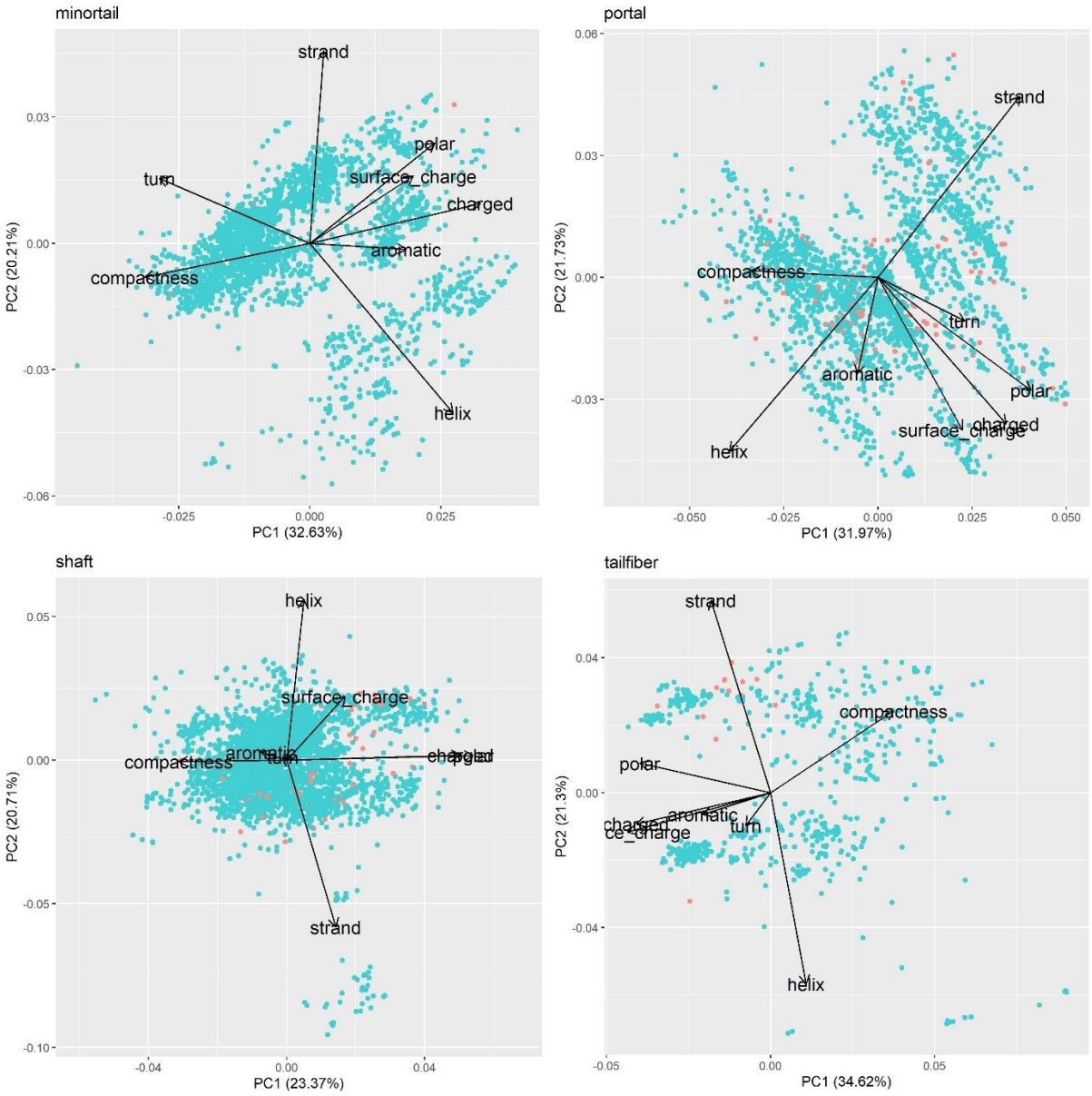
Feature	MeanDecreaseGini
polar	5,98
aromatic	5,52
compactness	5,38
turn	4,44
charged	4,40
helix	4,12
elec_energy	3,88
strand	3,09

Feature	MeanDecreaseGini
surface_charge	0,43
polar	0,42
strand	0,35
compactness	0,24
aromatic	0,22
helix	0,21
turn	0,20
charged	0,19

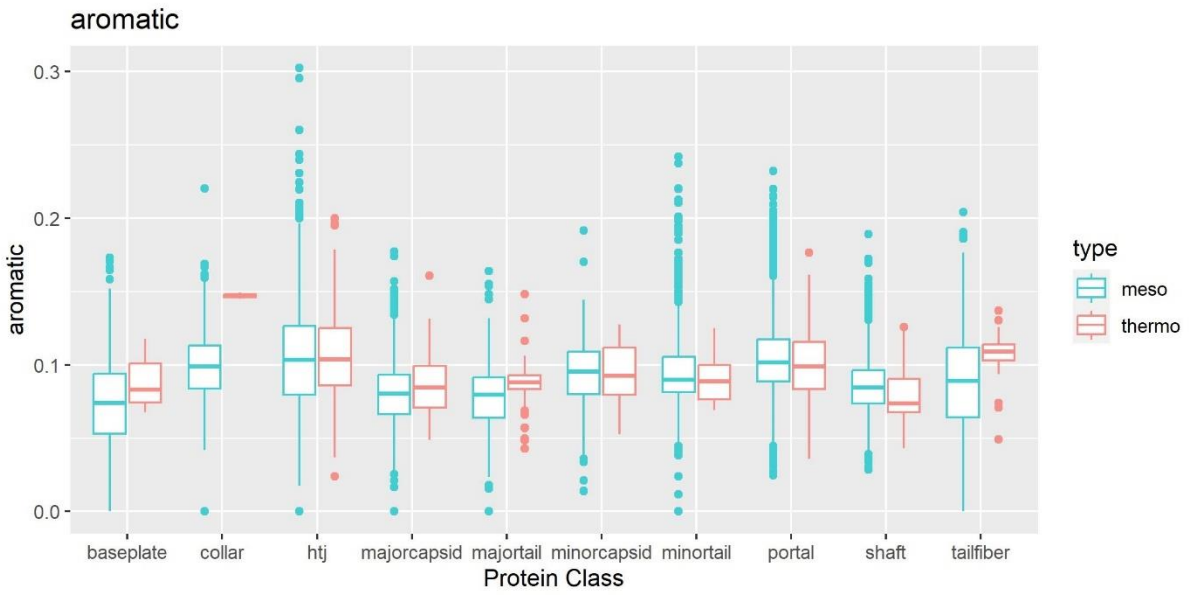


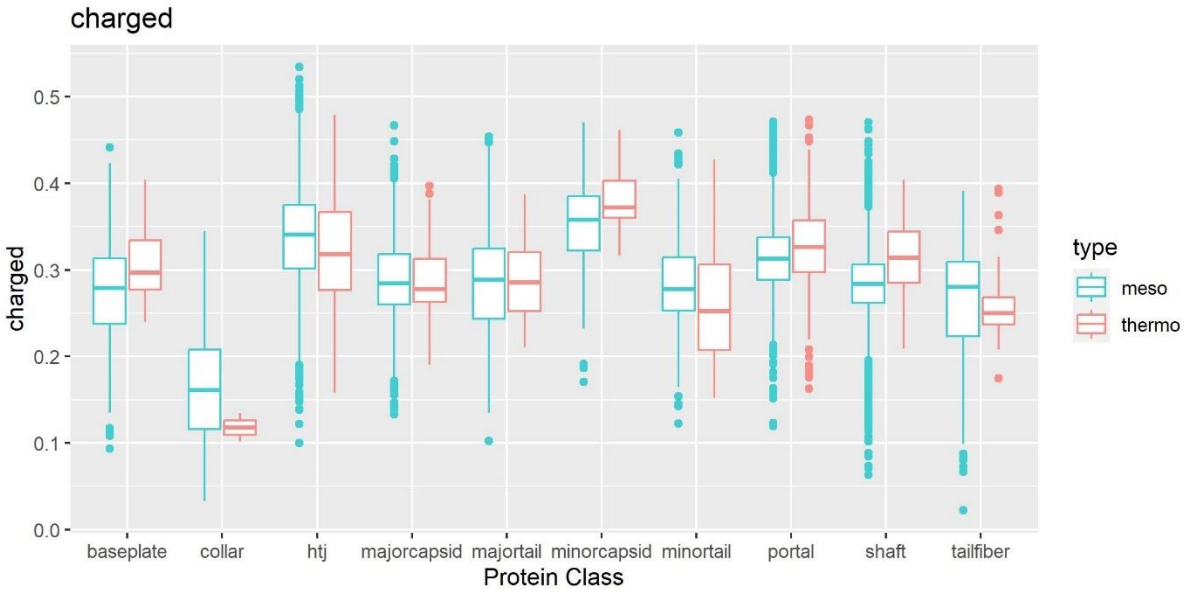
Supplementary figure A: Optimal growth temperature over stem GC content of 16S rRNA sequences of TEMPURA entries.

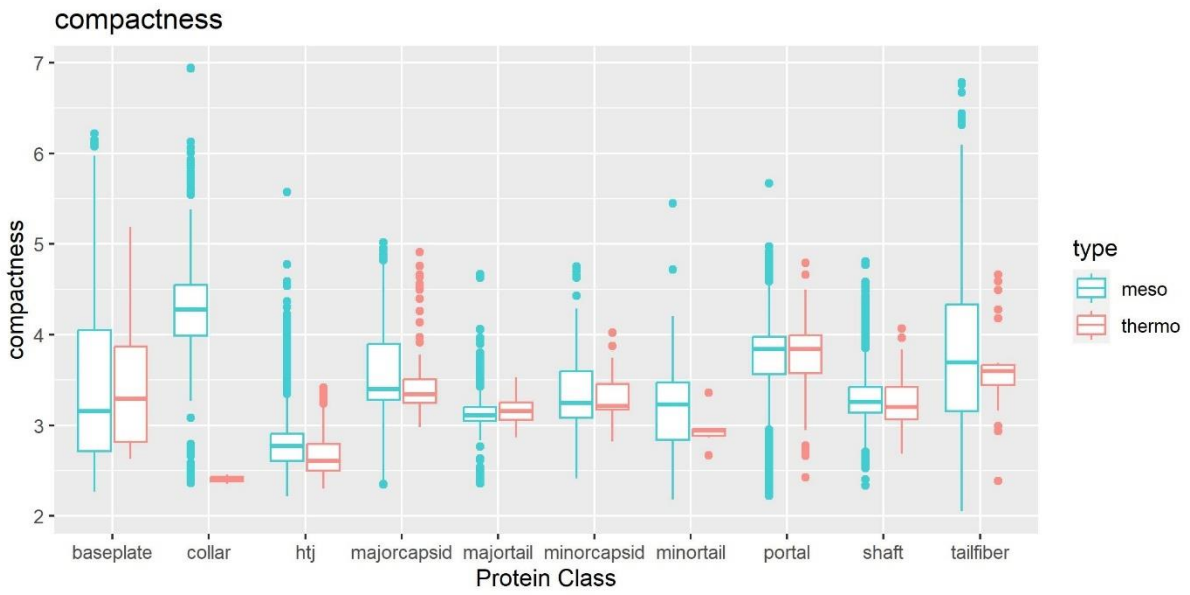

baseplate
aromatic
helix
charged
ce_charge
polar
compactness
turn
strand
PC2 (20.25%)
PC1 (43.88%)
collar
turn
aromatic
helix
charged
surface_charge
strand
polar
compactness
PC2 (24.03%)
PC1 (28.83%)
htj
strand
turn
compactness
surface_charge
polar
charged
aromatic
helix
PC2 (27.06%)
PC1 (29.34%)
majorcapsid
strand
turn
charged
polar
ce_charge
aromatic
compactness
helix
PC2 (18.95%)
PC1 (38.86%)
majortail
strand
turn
helix
surface_charge
charged
aromatic
compactness
polar
PC2 (20.88%)
PC1 (27.05%)
minorcapsid
compactness
helix
turn
aromatic
charged
strand
polar
PC2 (17.64%)
PC1 (42.4%)

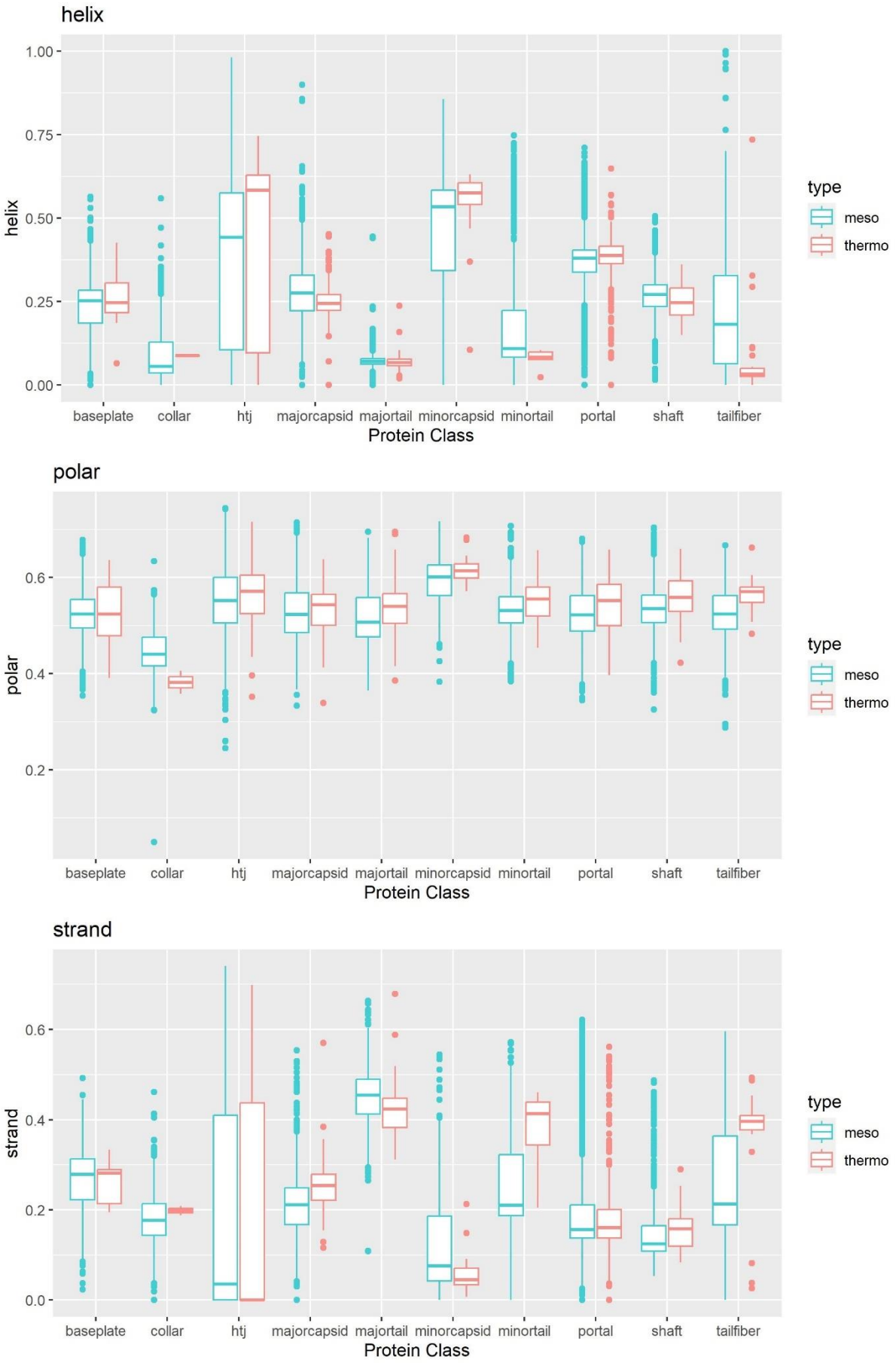


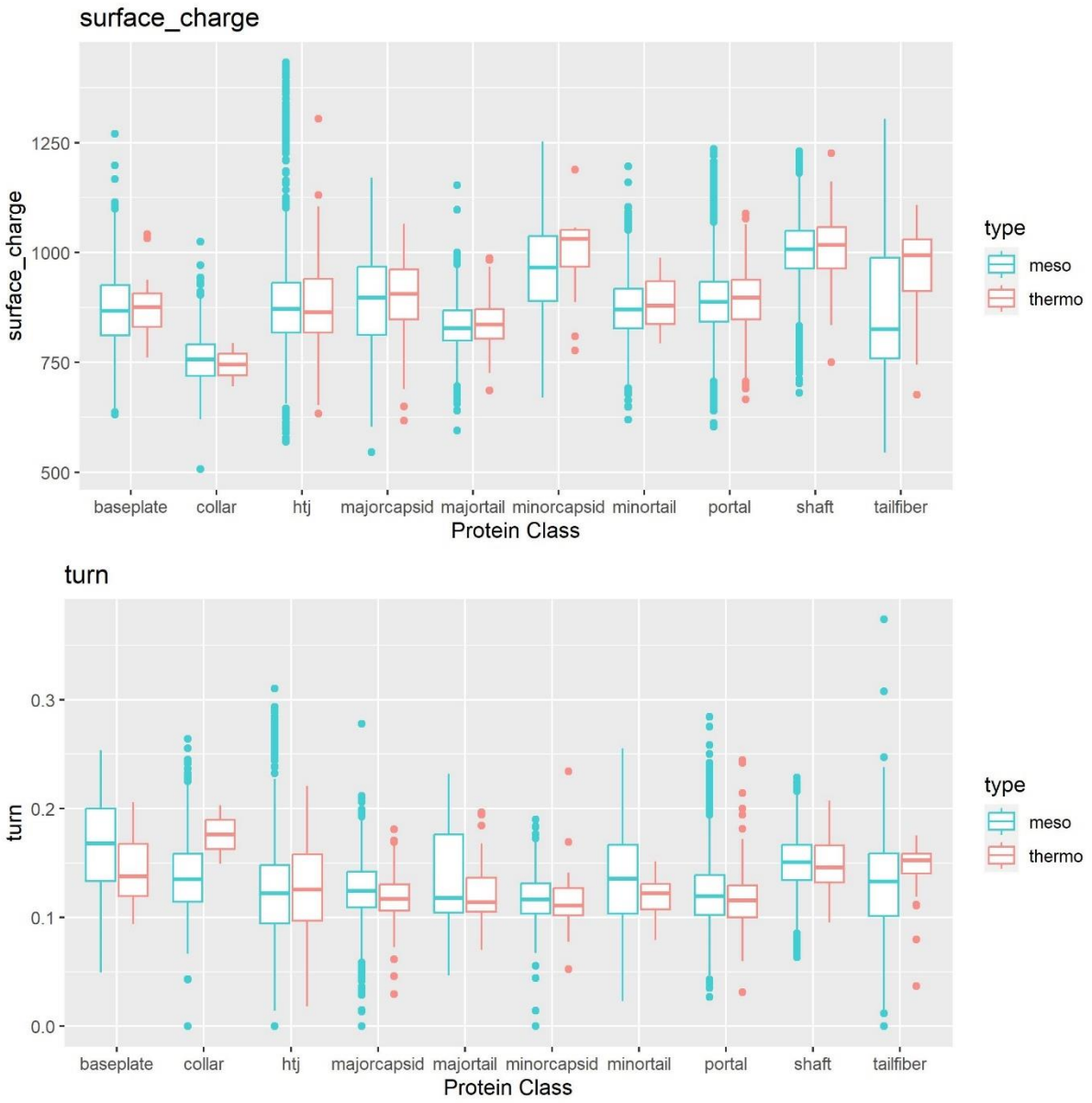
Supplementary figure B: PCA biplots of all ten protein classes.


aromatic
aromatic
0.3
0.2
0.1
0.0
baseplate
collar
htj
majorcapsid
majortail
minorcapsid
minortail
portal
shaft
tailfiber
Protein Class
type
meso
thermo

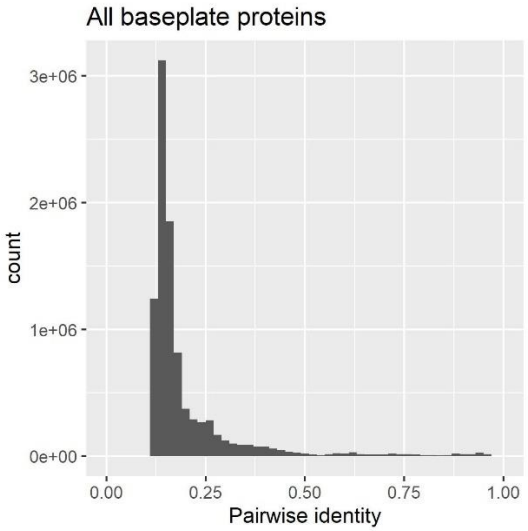

charged
charged
0.5
0.4
0.3
0.2
0.1
0.0
baseplate
collar
htj
majorcapsid
majortail
minorcapsid
minortail
portal
shaft
tailfiber
Protein Class
type
meso
thermo

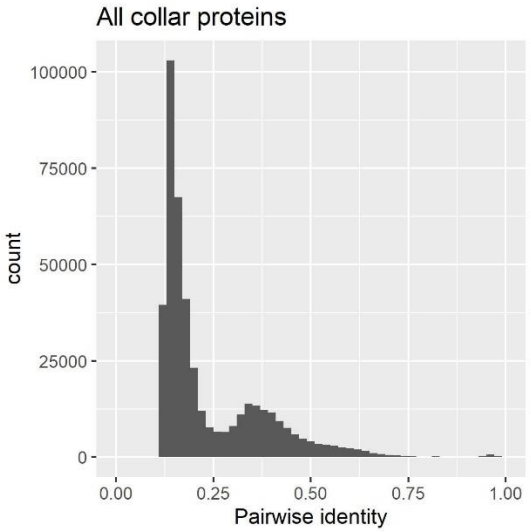

compactness
compactness
7
6
5
4
3
2
baseplate
collar
htj
majorcapsid
majortail
minorcapsid
minortail
portal
shaft
tailfiber
Protein Class
type
meso
thermo

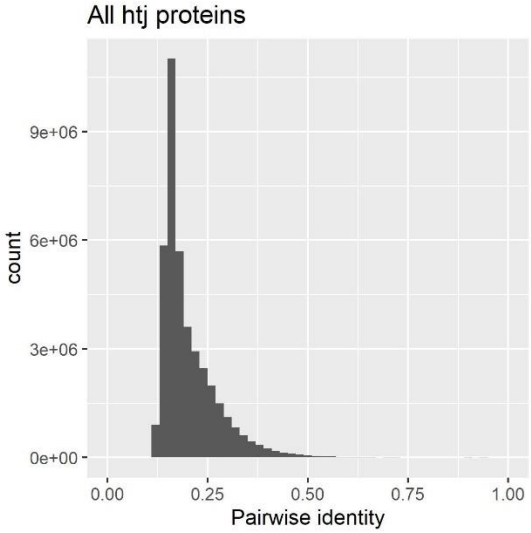

helix
1.00
0.75
0.50
0.25
0.00
helix
baseplate
collar
htj
majorcapsid
majortail
minorcapsid
minortail
portal
shaft
tailfiber
Protein Class
type
meso
thermo
polar
0.6
0.4
0.2
polar
baseplate
collar
htj
majorcapsid
majortail
minorcapsid
minortail
portal
shaft
tailfiber
Protein Class
type
meso
thermo
strand
0.6
0.4
0.2
0.0
strand
baseplate
collar
htj
majorcapsid
majortail
minorcapsid
minortail
portal
shaft
tailfiber
Protein Class
type
meso
thermo

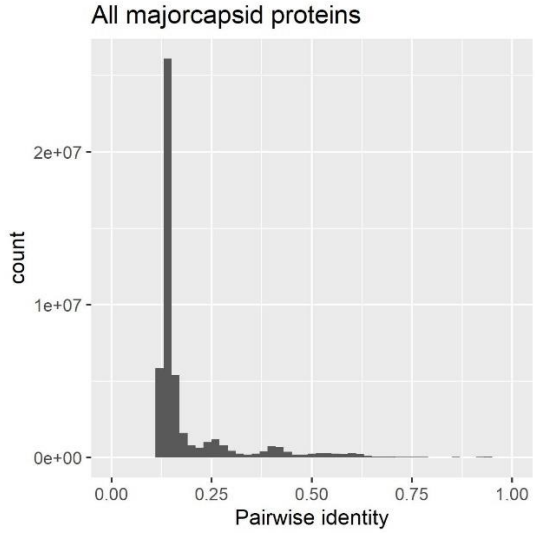


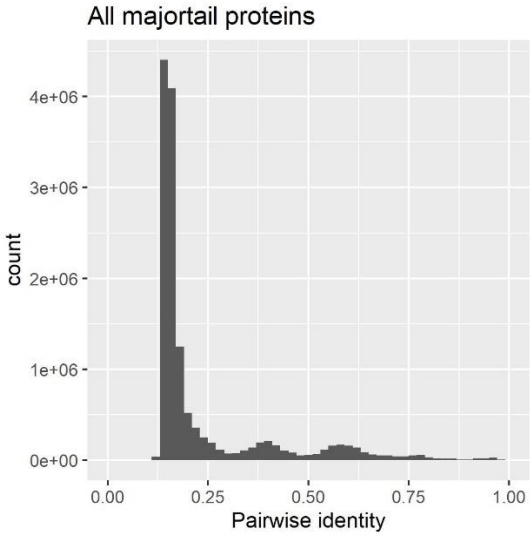
Supplementary figure C: Boxplots of all eight structural features per protein class.

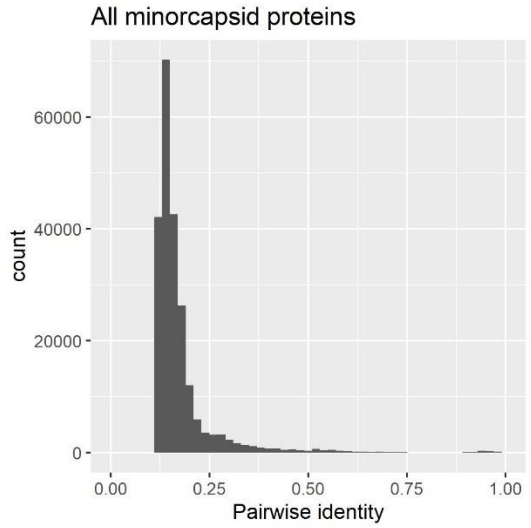

All baseplate proteins
count
3e+06
2e+06
1e+06
0e+00
0.00
0.25
0.50
0.75
1.00
Pairwise identity

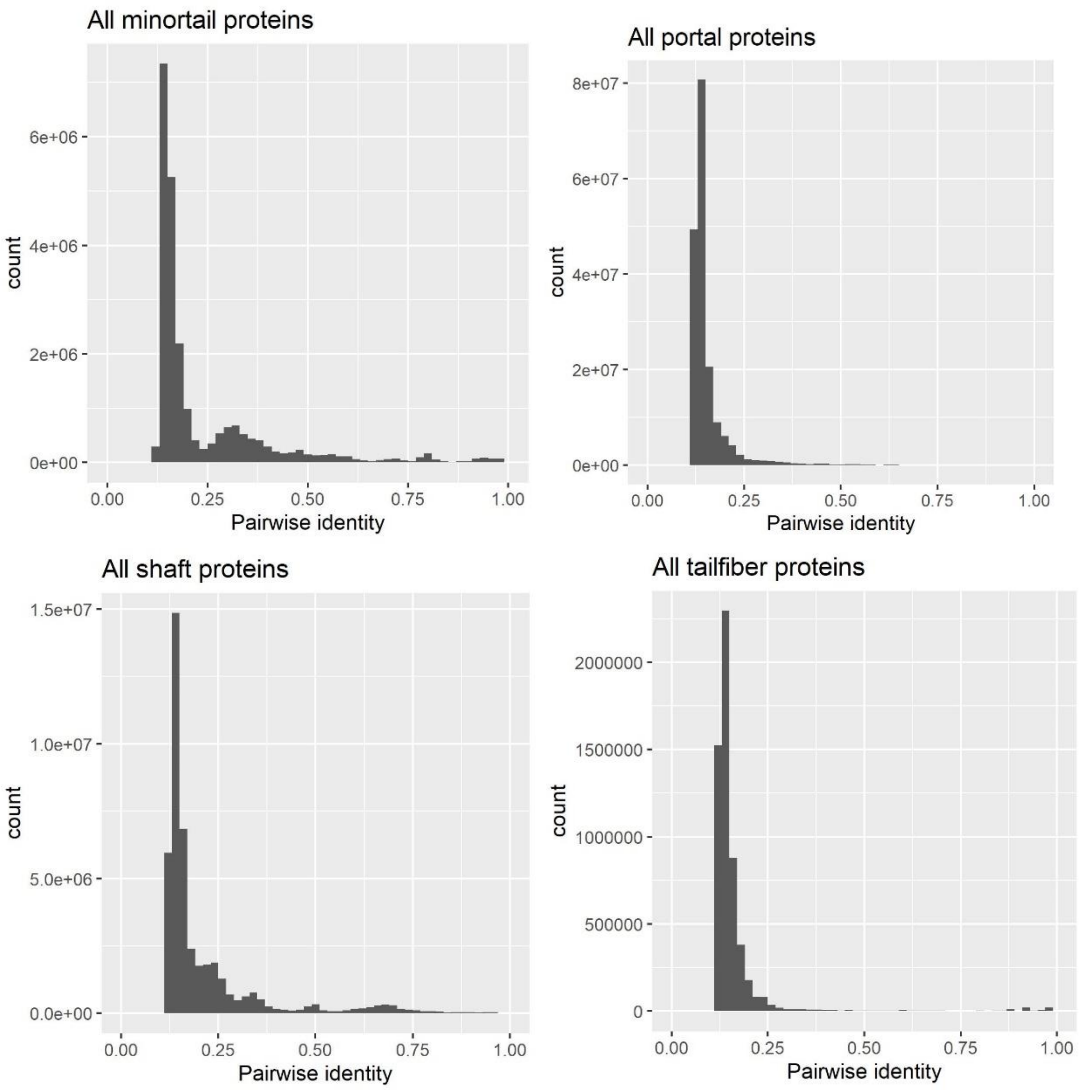

All collar proteins
count
100000
75000
50000
25000
0
0.00
0.25
0.50
0.75
1.00
Pairwise identity


All htj proteins
count
9e+06
6e+06
3e+06
0e+00
0.00
0.25
0.50
0.75
1.00
Pairwise identity

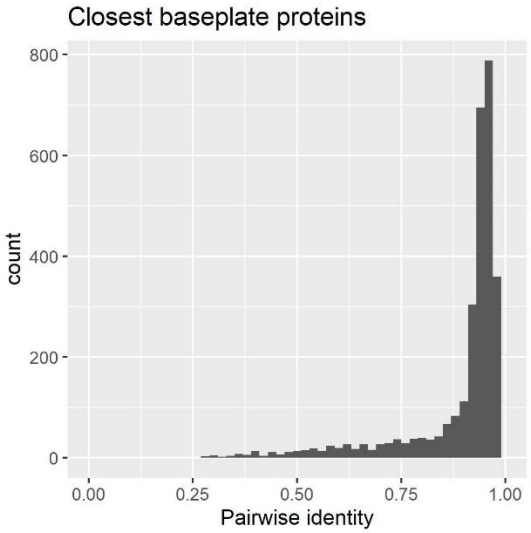

All majorcapsid proteins
count
2e+07
1e+07
0e+00
0.00
0.25
0.50
0.75
1.00
Pairwise identity

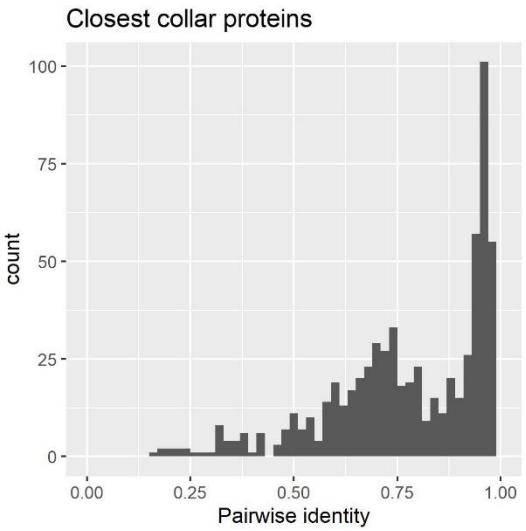

All majortail proteins
count
4e+06
3e+06
2e+06
1e+06
0e+00
0.00
0.25
0.50
0.75
1.00
Pairwise identity

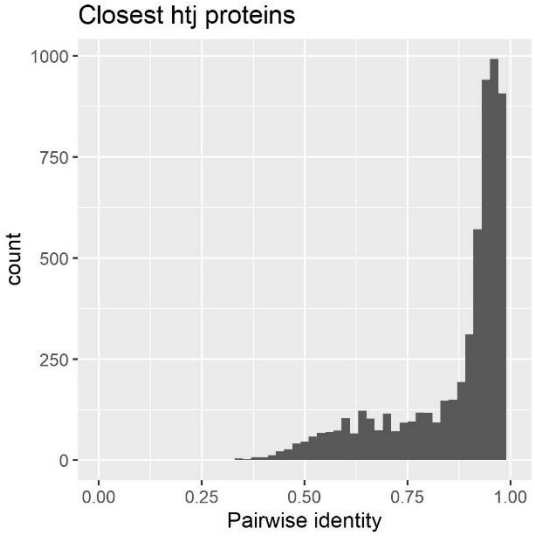

All minorcapsid proteins
count
60000
40000
20000
0
0.00
0.25
0.50
0.75
1.00
Pairwise identity

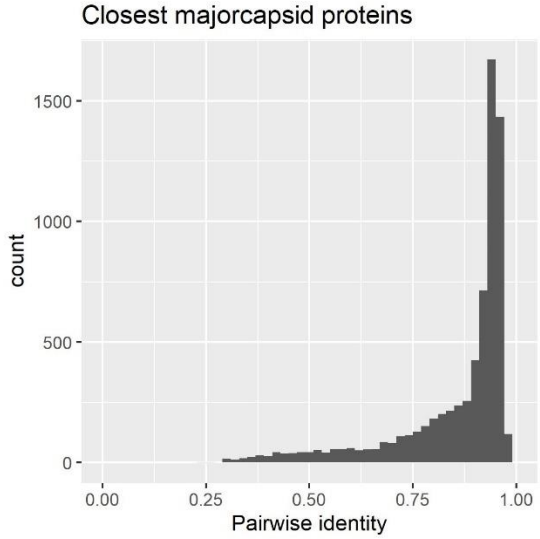


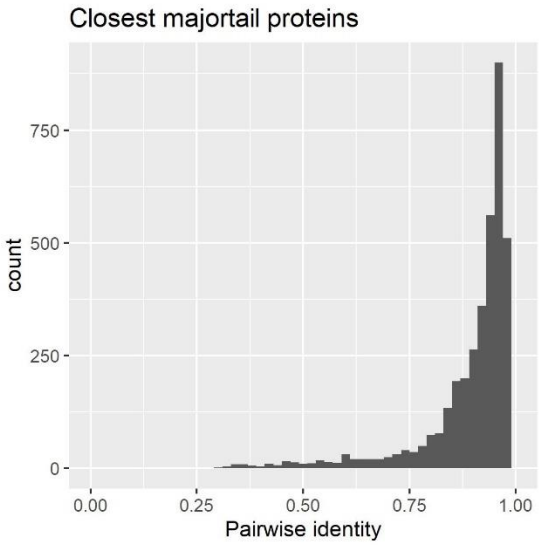
Supplementary figure D: Distribution of pairwise identities for all protein classes.

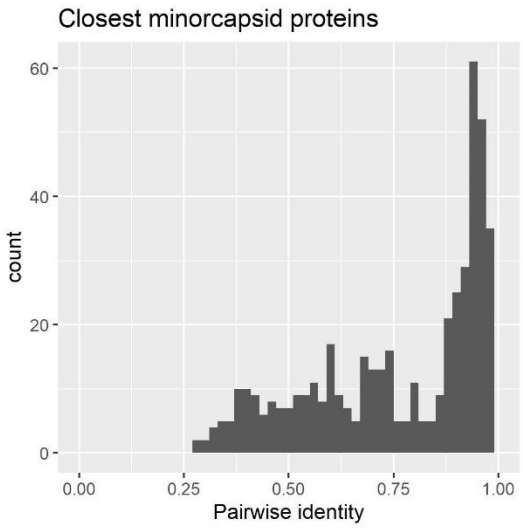

Closest baseplate proteins
count
Pairwise identity
0
200
400
600
800
0.00
0.25
0.50
0.75
1.00

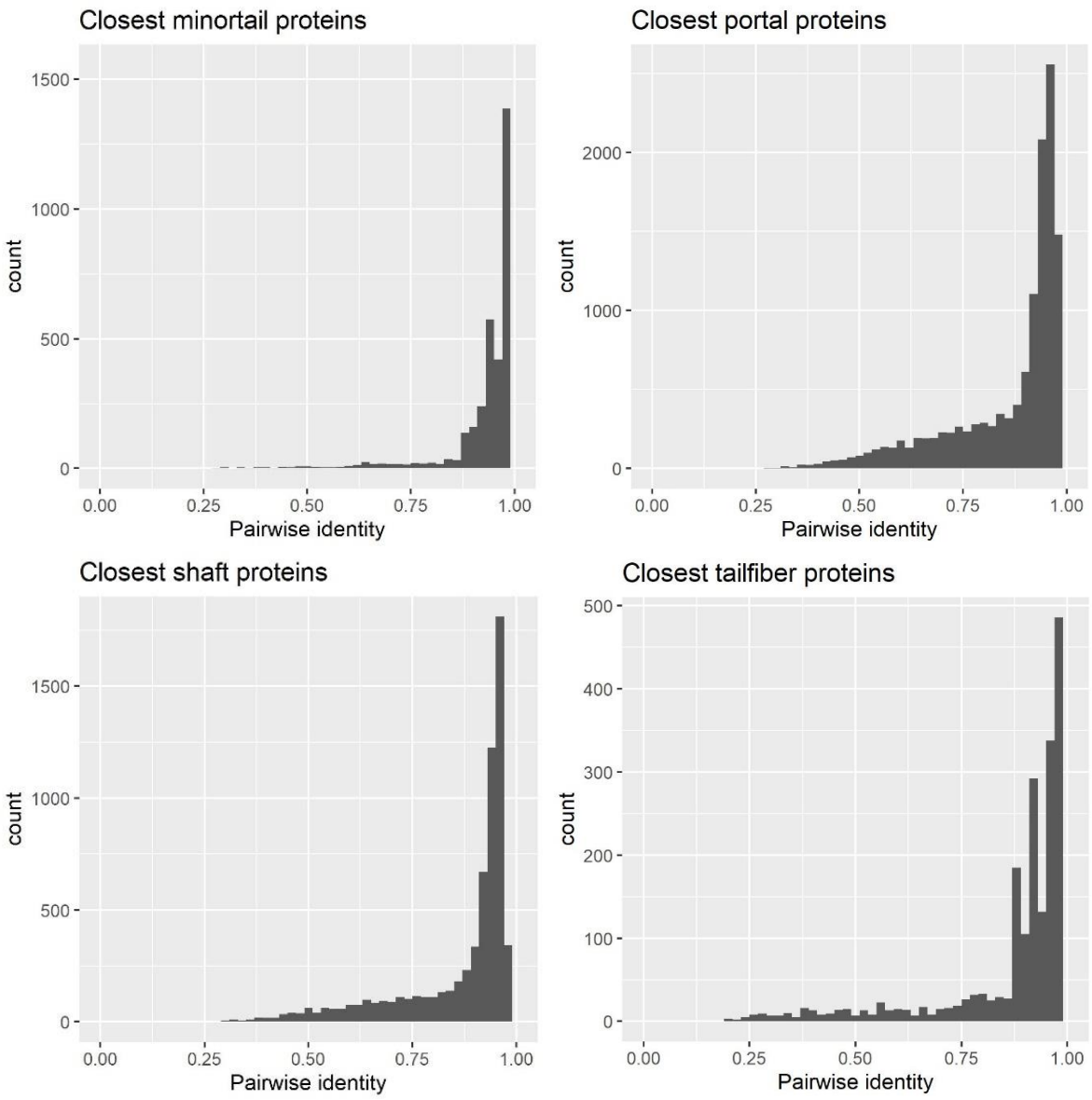

Closest collar proteins
count
Pairwise identity
0
25
50
75
100
0.00
0.25
0.50
0.75
1.00


Closest htj proteins
count
Pairwise identity
0
250
500
750
1000
0.00
0.25
0.50
0.75
1.00


Closest majorcapsid proteins
count
Pairwise identity
0
500
1000
1500
0.00
0.25
0.50
0.75
1.00


Closest majortail proteins
count
Pairwise identity
0
250
500
750
0.00
0.25
0.50
0.75
1.00


Closest minorcapsid proteins
count
Pairwise identity
0
20
40
60
0.00
0.25
0.50
0.75
1.00



Supplementary figure E: Distribution of pairwise identity of most similar proteins for all ten protein classes.