Deep Mutational Scan Analysis
of
TEM 1 β-lactamase

Malte Klein, Christoph Luh, Svea Meinicke, Felix Schubert

Final Report - Datascience MoBi 2023

Topic 02 - Team 01

Supervisor: Prof. Dr. Dominik Niopek

Tutor: Benedict Wolf

Institute of Pharmacy and Molecular Biotechnology

Heidelberg University

Table of contents

1. Introduction

Bacteria have evolved many different enzymes granting antibiotic resistance as a defensive mechanism or to gain an advantage over competitors. One common class are β-Lactamases, which hydrolyze lactam rings of antibiotics, rendering them ineffective. Despite their rather small size and simple structure, the exact mechanisms and interactions inside these enzymes remain partially unknown. Even though the active site has been identified and studied intensively, the roles of many other residues are not yet clarified.

Deep Mutational Scanning (DMS) aims to assess the importance of each amino acid for enzymatic function. The activity of all possible single mutants of each amino acid position is measured, resulting in large datasets, that may hold insightful information. There are many different assays of enzymatic activity, as well as different ways to process the data leading up to a single metric for each mutant – the DMS-score.

Based on a DMS-data analysis of TEM-1 β-Lactamase, this project aims to compare two DMS measurement methods – the “growth method” and the “resistance method”. This assessment is based on three DMS-datasets originally provided by Stiffler et al., Firnberg et al. and Deng et al., and processed by Notin et al. in their ProteinGym (Notin et al., 2022). Stiffler et al. and Firnberg et al. applied a growth model, where they compared the growth of mutant cells on different Ampicillin concentrations to derive a DMS-score. Deng et al. on the other hand used a resistance model focusing on the survival of mutants on a single ampicillin concentration.

Furthermore, a goal of this project is the assessment of the mutational robustness or fragility of each residue. Robust positions can be altered without deleterious effects for the enzyme, while fragile positions significantly decrease enzymatic function upon mutation, indicating their importance in catalysis or structural stability. Different approaches of describing robustness and fragility of a position from DMS-data are compared. In order to judge the performance of these Position Effect Models, the following hypotheses are formulated:

I.) High fragility conditions higher conservation of a residue.

II.) Conservation occurs only when necessary, thus at fragile residues.

III.) Thus, high robustness conditions lower conservation.

Conservation scores are obtained from Multiple Sequence Alignment (MSA) of 100 enzyme sequences and correlations with DMS-data are calculated. Predictive accuracy of the DMS datasets and the Position Effect Models serves as a benchmark in their assessment. Additionally, consensus sequences are derived and mismatches with the wildtype TME-1 β-Lactamase sequence are investigated to evaluate the accuracy of the DMS-datasets.

2. Material & Methods

All available datasets for our enzyme of choice (TEM-1 beta-lactamase), that had been derived from measurements under selection with ampicillin have been selected from the ProteinGym for this analysis. Generally, they consist of DMS-scores for all/nearly all possible single-mutant variants of TEM-1 beta-lactamase. The scores in each dataset were derived from different measurement methods as well as different mathematical approaches that were used to summarize all measurements under one metric for each mutant. Additional to the DMS-scores, the altered sequence of the enzyme was given, as well as an abbreviation for the exchanged amino acids and amino acid positions, and a binary translation of the DMS-score, where 0 is assigned to mutants with low fitness and 1 is assigned to mutants with high fitness.

2.1 Data Cleaning & Normalization

2.1.1. Outliers

First, outliers are identified and clipped. For their identification, the datasets are plotted in a histogram, where discontinuations in the column pattern indicate outliers. Additionally, the 40 highest and lowest performing mutants of each dataset are screened for large leaps between their DMS scores. This analysis results in the clipping of a single mutant (F58N) in the Firnberg-dataset.

2.1.2. Normalization

Then, two normalization methods for each dataset are compared: a Min-Max-Normalization

$$(1) \qquad x_{Norm} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

x: DMS-Score

x_{min}/x_{max}: Minimum/Maximum DMS-score of dataset

and a Z-Normalization

$$(2) \qquad z = \frac{x - \mu}{\sigma}$$

μ: Mean DMS-score of dataset

σ: Standard deviation of DMS-scores of dataset

The goal of the normalization is to make the datasets comparable. For each dataset the mean DMS-score of all mutants corresponding to an amino acid position is calculated. To assess the normalization methods, the DMS-means for each position derived from the normalized datasets are compared. An adequate normalization method is expected to amount in similar positional patterns in a heatmap, especially between the Stiffler and the Firnberg dataset, since similar measurement methods were used.

Subsequently, the results of the visual analysis via heatmaps are confirmed with a Wilcoxon Signed Rank Test, which is typically used to calculate the probability (p-value) that two samples originate from the same population. Since all datasets were derived from measurements of the same enzyme with the same selection mechanism, the calculated p-values should be high, when the data has been normalized adequately.

2.2. Position Effect Models

2.2.1 Mean DMS-score Model

The mean DMS-scores of an amino acid position can be used to describe positional effects upon mutation. However, an attempt is made to represent position effects more precisely than by this Mean DMS-score Model.

2.2.2 Quantile Model

First, in the "Quantile Model" only the 25% highest and lowest valued mutants from the individual data sets are considered. For each position, it is determined how many percent of the mutants corresponding to this position belong to the top/bottom 25% of the respective dataset. This figure is a first approximation of the “robustness” or “fragility” of the positions. However, setting a cutoff at the top/bottom 25% is more or less arbitrary (this was also done by Leander et al., for example, but was left without justification there (Leander et al., 2022)).

2.2.3 Area Under Curve Model

Therefore, a metric is developed that summarizes the proportions of mutants of a position in the top/bottom x% for all possible x’s, serving as different cut-offs. In the first step the proportion of mutants of each position in the top/bottom 50% is determined, since the top 51% would already contain some of the below-average mutants. Then the shares of the top/bottom 49%, to 1% are determined. This results in a hypothetical curve for each position in which the proportion of mutants among the top/bottom x% of the dataset decreases as x decreases. The area under these hypothetical curves (AUC) is then calculated and can be used as a metric for robustness (when derived from the top-curve) and fragility (when derived from the bottom-curve) for each position.

2.2.4 Assessment of Accuracy of Position Effect Models

The Mean DMS-score Model, the Quantile Model and the AUC Model are then assessed for their predictive accuracy regarding conservation. Correlation scores between mean DMS-scores and conservation scores are calculated. In correspondence with hypothesis I) and III), the model of position effects that achieves the highest correlation with conservation scores (derived from multiple sequence alignment in the follwing) is regarded as the most accurate. In the same way, the different datasets are assessed in their predictive accuracy regarding position effects.

2.3. Alignment Analysis

2.3.1 NCBI BLAST Alignment

NCBI protein BLAST is a program for aligning a query sequence with a selected data base, like the RefSeq library which is used for the TEM-1 alignment. The Refseq library is continuously updated, validated and has a consistent format which distinguishes it from other libraries.

In the alignment the matching of the sequences is maximized to the highest **identity**, which describes the proportion of matches between the query and the aligned sequence. To achieve this, gaps can be introduced in the aligned sequences (Madden et al., 2013).

Therefore, the **alignment quality** is assessed by the alignment score which is the sum of costs associated with gaps, gap extensions, identities and replacements. The definition of their scores is based on experimental data as well as on substitution matrices.

$$(3)\ S = \Sigma \text{ costs (identities, replacements)} - \Sigma \text{ penalties (no. of gaps x gap penalties)}$$

The **coverage** in the alignment is the percentage of aligned query sequence positions and another indicator to assess the alignment quality (Fassler & Cooper, 2008).

The alignment used for the analysis of the DMS data comprises 100 sequences. The identity ranges from 47.5% to 100%, while the coverage ranges from 42.66% to 100%. For analysis the alignment is exported as a fasta file and visualized in a heatmap with the consensus sequence.

2.3.2 Conservation Score Calculation

Two different strategies are used for quantifying the conservation of each residue of in the TEM-1 query sequence based on the multiple sequence alignment (MSA): the Pei & Grishin conservation, a variance-based measurement and the Shannon entropy, an entropy-based measurement (Fischer et al., 2008).

2.3.2.1 Pei & Grishin conservation

The root mean square deviation between the amino acid distribution at one position and the estimated distribution is measured. As the difference between the amino acid proportions is calculated and not their ratio, it is less sensitive to minor distributed amino acids.

$$(4) \quad \mathrm{V} = \sqrt{\Sigma \left(\frac{\mathrm{p_{ia}}}{\mathrm{N}} - \mathrm{E}\right)^2}$$

The probability of each amino acid to occur at one position is E = 1/20, as each of the 20 amino acids can be inserted. The alignment length is N = 100. Therefore, the minimum for an equal distribution (p_{ia} = 5) is V = 0. And the maximum V = 0.95, for a total conserved residue (p_{ia} = 100).

2.3.2.2 Shannon entropy

The Shannon entropy estimates the diversity within an alignment. It is calculated by the following formula, with the amino acid frequency p_{ia}.

$$(5) \qquad H = - \Sigma (p_{ia} \log_2(p_{ia}))$$

The Shannon entropy ranges from total conservation, where H = 0 , pia = 1 to maximum variance with H = 4.321, p_{ia}= 5/100. It is more sensitive than Pei & Grishin conservation and more commonly used for MSA. Thus, only Shannon entropy is used for the consensus sequence analysis, while Pei & Grishin conservation is used to verify the values calculated for the Shannon entropy.

2.3.2.3 Verification of Shannon Entropy by comparison with Pei & Grishin Conservation

For better comparability Pei & Grishin and Shannon entropy are z-normalized. As increasing values of entropy H indicate decreasing conservation and increasing variance V indicates higher conservation the sign of the z-normalized values is opposite for both methods. By plotting the z-normalized conservation scores in a bar plot, both conservation trend and strength are comparable. Exact correspondence is recognizable by an x-axis symmetry of the bars. Moreover, both methods are plotted together in a scatter plot and the correlation value is calculated.

2.3.3 Residue Categorization by DMS and Conservation

For first assessment of the four residue types, mean DMS and conservations scores of both methods are correlated and visualized in scatter plots and bar plots. That allows assessment of the connection between conservation with robustness and fragility, and categorization of the residues. In the bar plot, positive or negative trends of the bars allow an estimation of the distance between the position scores and the average. The average is represented by the x axis (DMS score & Shannon entropy = 0).

	DMS value z-normalized	Fitness effect	Shannon entropy z-normalized	Conservation	Indication
Position of Interest	Negative	Fragile	Negative	Above average	Supports Hypothesis I
Unnecessarily Conserved Position	Positive	Robust	Negative	Above average	Contradicts Hypothesis II
Random Position	Positive	Robust	Positive	Below average	Supports all Hypotheses
Destructive Position	Negative	Fragile	Positive	Below average	Contradicts Hypothesis I

Table 1: Mismatch Categories: Mismatches of Wildtype TEM-1 β-Lactamase and consensus sequences are categorized depending on the conservation score and mean DMS-score of their respective amino acid positions. The counts of the different mismatch types will be used to assess the hypotheses and accuracy of the datasets.

Position of Interest (IP): Positions with negative DMS scores (which indicates above fragility), paired with negative Shannon conservation (which indicates above average conservation). Their appearance supports hypothesis I) and are an Indication for the correctness of the DMS data.

Unnecessarily conserved Position (UP): Positions with positive DMS scores (robust), paired with negative Shannon conservation (high conservation). This type contradicts hypothesis II) and can be a sign for an error in the DMS data.

Random Positions (RP): Positions with positive DMS scores (robust), paired with positive Shannon conservation (low conservation). This type agrees with the hypothesis III), indicating DMS correctness.

Destructive Positions (DP): Positions with negative DMS scores (fragile), paired with positive Shannon conservation (low conservation). They contradict the first hypothesis and imply inaccuracy in DMS.

2.3.4 Consensus Sequence Alignments (SCC)

The consensus sequence is a theoretical sequence of RNA, DNA or amino acids representing the consensus of all aligned sequences. It can be composed by different methods. Most commonly the consensus sequence harbors the most frequent amino acids in the alignment. That allows estimation of evolutionary conservation given by the matches and mismatches of the consensus with TEM-1 sequence (Sternke et al., 2020). By consensus alignment (alignment of wildtype TEM-1 sequence with calculated consensus sequence) the extent of agreement of consensus sequence with conservation values and DMS-scores is assessed.

Therefore, the identity is calculated, and the consensus sequence is aligned by pairwise2 function from Biopython. The following parameter scores are used for calculation of the alignment score: match score = 1, mismatch score = -2, gap score = -2.5, opening gap score = -1. The self-calculated consensus sequence (SCC), composed of the most frequent amino acid at each position in the alignment is aligned and mismatches identified. Only the mismatches are visualized and categorized, since conservation does not necessarily condition fragility, but fragility does condition conservation and thus no fragile positions are to be expected in the mismatch population. Because in SCC the most frequent amino acid is inserted, errors are made at positions with similarly distributed amino acids, thereby consensus sequences were composed with Emboss as well.

2.3.5 EMBOSS Consensus Alignment (ECS)

EMBOSS cons is a program which creates a consensus sequence by weights and scoring matrix values for a multiple sequence alignment. Therefore, every amino acids type at one residue is given a score, composed of the weight (defined by the alignment file), multiplied with a scoring matrix value and the residue length. The highest scored amino acid is then found and inserted if it's score reaches a certain "cut-off" (http://emboss.open-bio.org/wiki/Appdocs (10.07.23)).

2.3.5.1 PAM - point accepted substitution matrix:

This substitution matrix is based on point mutation data from 71 phylogenetic trees with 1572 mutations in total (Jia & Jarnigan, 2021). The score of the PAM matrix indicates the distance of the sequences, while increasing scores indicate greater distance (Mount, 2008).

2.3.5.2 BLOSUM -Block Substitution Matrices

BLOSUM matrices are the most common matrices used in MSA, while the BLOSUM62 matrix serves as default matrix at NCBI. They are constructed by alignment data of local alignments. (Jia & Jarnigan, 2021). The higher the number of the BLOSUM matrix the closer is the relation of the aligned sequences and therefore the higher is the identity (Pearson, 2013). The identity of TEM-1 alignment reaches from 47.5% to 100% with a huge proportion of around 90%. Therefore, BLOSUM62 and BLOSUM90 are used. For reference values consensus sequences of PAM250 (which should be similar to BLOSUM62) and PAM460 are aligned with TEM-1.

2.3.6 Optimization of BLOSUM90 Alignment

To obtain better alignment scores with BLOSUM90 consensus sequence the EMBOSS Needle pairwise alignment program is utilized for the alignment construction by scoring the positions with BLOSUM90.

3. Results

3.1. Data Cleaning & Normalization

The comparison of the normalization methods is done visually via heatmaps and with a Wilcoxon Signed Rank Test. As seen in figure 1, the patterns in the heatmaps indicate that the Z-normalization method produces more comparable data.

The Wilcoxon Signed Rank Test confirms this, as the p-values produced by the Z-normalization are significantly higher than those produced by the Min-Max-normalization approach, as seen in table 2.

Min-Max-normalization	Stiffler	Firnberg	Deng
Stiffler	1	6.89 * 10^{-45}	6.12 * 10^{-38}
Firnberg	-	1	5.98 * 10^{-19}
Deng	-	-	1
Z-normalization			
Stiffler	1	0.94	0.70
Firnberg	-	1	0.48
Deng	-	-	1

Table 2: p-Values of Wilcoxon Signed Rank Test used to compare the normalized datasets: The Z-normalization produces significantly higher p-values than the Min-Max-normalization. Since the p-value is a metric for the likelihood of the compared datasets to be derived from the same population, high p-values indicate good comparability of datasets derived from the same enzyme.

Furthermore, it becomes clear that the Deng data deviate more from the Stiffler and Firnberg data. This effect is observed for both normalization methods.

3.2. Position Effect Models

The different Position Effect Models (Mean DMS-score Model, Quantile Model and AUC Model) are assessed. In table 3 it is shown that the Mean DMS-score Model is by far the most accurate.

	Pei & Grishin Conservation	Shannon Entropy
Dataset	**Mean DMS-score Model**	
Stiffler	-0,67	0.70
Firnberg	-0.70	0.73
Deng	-0.53	0.54
Merged Model	-0.69	0.72
	Quantile Model	
Stiffler	-0.65	0.67
Firnberg	-0.65	0.68
Deng	-0.50	0.51
Merged Model	-0.66	0.68
	AUC Model	
Stiffler	0.06	-0.05
Firnberg	0.08	-0.06
Deng	0.08	-0.06
Merged Model	0.08	-0.07

Table 3: Correlation scores (r-values) of the datasets in different Position Effect Models to Pei & Grishin Conservation and Shannon Conservation scores. The Mean DMS-score Model achieves the highest correlation, closely followed by the Quantile Model. Both datasets show correlations with the expected directionality. The AUC Model achieves no significant correlation. The Firnberg-Dataset describes the level of conservation of the positions the most accurately, closely followed by the Stiffler-Dataset. The Deng-Dataset is less accurate, while still showing moderately significant r-values.

Furthermore, the Firnberg-dataset describes the conservation most accurately, closely followed by the Stiffler-Dataset. The Deng-Dataset is significantly less accurate, while still showing r-values with moderate significance. The Merged Model (mean DMS-scores from all datasets) realizes r-values closely resembling those of the Stiffler and Firnberg datasets. The Position Effect distribution described by the Merged Model through the Mean DMS-score Model is shown in figure 2.

Generally, the Shannon Entropy is predicted slightly more accurately than the Pei & Grishin Conservation.

3.3 Alignment Analysis

3.3.1 Conservation Score Calculation

The Pei & Grishin conservation (orange) and Shannon entropy (blue) after z-normalization are plotted in figure 3. The strong correspondence is visible by the x-axis symmetry of the bars. Mostly the bars of both methods are of similar length, which shows that their conservation values are close. The correlation value r = 0.991 indicates significant correlation, underlined by the datapoints resembling a straight line in the scatter plot. Therefore, both correlation methods match well.

3.3.2 Residue Categorization by DMS and Conservation

The scatter plots (figure 4) show correlation of Pei & Grishin conservation (upper row) or Shannon entropy (lower row) with all DMS models. They are almost identical but show opposite trends, resulting by the different trend-sign. Across the data sets the scatter plots are mostly consistent. Two clusters can be detected, located in the corners with high conservation and low DMS scores or low conservation and high DMS scores. The Deng distribution in the first column shows a less clear division into these groups, moreover the correlation value of Deng is the lowest (r = 0.53).

The merged model (mean values across all three datasets) (figure 5) represents the residue distribution of all three DMS data sets well, with exception of the Deng distribution at Position 180-230 (figure 6). In all data sets IP's and RP's are predominant. While UP's are rarely found in the other data sets, they are significantly present around residue 230 in the Deng distribution. Still the proportion of UP's and DP's is negligibly small.

3.3.3 SCC Alignment

The SCC alignment score is $S = 155$ and the Identity $I = 87.1\%$ which is very high. The categorization of mismatch residues in the bar plot (figure 7) shows that in all data sets RP's are predominant and account for 38-50% of all mismatches. PI's do not occur except in the Deng distribution, supporting hypothesis I). DP's (p = 2.9%) and UP's (p=5.8%) are distributed marginally. In the Deng distribution mismatches in region 200-230 differ from the other data sets.

3.3.4 ECS Alignment

The ECS Identity of all three ECS's lays in range of $I = 73.4\% - 74.1\%$, while the alignment scores are as follows; BLOSUM62 ($S = 2$), PAM250 ($S = -1$), BLOSUM90 ($S = -1.5$). There is less difference between BLOSUM90 and BLOSUM62 or PAM250 than expected. Therefore, PAM460 is calculated to obtain a reference value for a poor alignment ($S = -94.5$, $I = 61.9\%$). Compared to SCC and PAM460 the ECS lay in the middle and are a good approximation.

3.3.5 Optimization of BLOSUM90 Alignment

Compared to the previous BLOSUM90 alignment, both Identity $I = 84.4\%$ and alignment score $S = 25$ are higher and indicate a better alignment quality. The mismatch distribution of all datasets is mostly consistent. In contrast to the SCC, RP's do not predominate and make up 30-34.5% (figure 8). While RP's are consistent with our expectations, IP's which account for 23-25% of all mismatches should not appear in the mismatch group. Moreover, in all data sets except Deng the proportion of DP's lay around 7% and of UP's around 5%. While in the Deng data set UP's and DP's appear more often (DP's = 8.3%) and (UP's = 14.3%) than in the other data sets.

4. Discussion

4.1 Data Cleaning & Normalization

For the Min-Max-normalization, a similar pattern emerges for the two data sets, but the Firnberg values seem to be consistently lower than the Stiffler values. The Z-normalization on the other hand puts all datasets into the same scale, resulting in similar patterns in the heatmap.

This observation was confirmed with a Wilcoxon Signed Rank Test, resulting in p-values of $p_{\text{Firnberg-Stiffler}} = 0.94$, $p_{\text{Stiffler-Deng}} = 0.70$ and $p_{\text{Firnberg-Deng}} = 0.48$ In opposition, the p-values resulting from a Min-Max-normalization were all $p = 5.98 * 10^{-19}$ and lower. Deviations from p-values close to 100% can be explained by differences in the measurement methods, that may obscure the likeliness of the data sets, as well as different mathematical processing of the data and possible measurement inaccuracies.

4.2 Position Effect Models

The Correlation scores of the Position Effect Models show, that the Mean DMS-score Model ($r_{\text{PG,merged}} = -0.69$, $r_{\text{Shan,merged}} = 0.72$) and the Quantile Model ($r_{\text{PG,merged}} = -0.66$, $r_{\text{Shan,merged}} = 0.68$) describe position effects well. The AUC Model ($r_{\text{PG,merged}} = -0.08$, $r_{\text{Shan,merged}} = 0.07$) on the other hand shows no significant correlation and is not suited for the description of Position Effects. However, this supports the use of mean DMS-scores of positions as indicators for robustness and fragility, as is done in the subsequent analysis.

Nonetheless, cases in which the DMS scores vary strongly will allude to inaccuracies and limit the significance of our predictions with this model. In this way, mutational effects at positions with high variance within its mutant DMS-scores are especially hard to predict. Further analysis of the connection of the amino acid type, function and mutation outcome would be necessary to refine a predictive model of mutational effects on TEM-1 beta-lactamase. Additionally, Position Effects in single mutants do not necessarily translate identically to multiply mutated sequences, posing a general limitation of DMS-data.

When comparing the datasets, it is shown that significant correlation of Shannon entropy or Pei & Grishin conservation with the Stiffler and Firnberg DMS-data was found with absolute r- values ranging from r = 0.67-0.73. Whereas the r value of Deng is only moderately significant with r = 0.53. This suggests that the resistance measurement method used by Deng et al. is not as accurate as the growth measurement methods used by Stiffler et al. and Firnberg et al. To verify this hypothesis, DMS-datasets of other enzymes, where the different methods have been used, could be analyzed. The correlation of the merged model showing absolute r values ranging from 0.69 to 0.71 is a good interface between Stiffler and Firnberg. This suggests that the merged model realizes comparable accuracy, supporting its use in the alignment analysis.

4.3 Alignment analysis

By categorization of the TEM-1 residues the appearance of IP's and RP's was found to be predominant, which support hypothesis I) and II). Moreover, the small proportion of UP's in all datasets apart from the Deng dataset is consistent with hypothesis III). DP's contradict that mutations at fragile positions are depleted during evolution and their appearance is negligibly small. Based on these categorization results, all DMS data sets except Deng show no inaccuracies and can be validated. That does not hold true for the SCC and ESC mismatch distribution however. DP's and UP's are rarely distributed in both mismatch groups, ($p_{\text{SCC\_DP}} = 2.9\%$, $p_{\text{ECS\_DP}} = 7.0\%$, $p_{\text{SCC\_UP}} = 5.8\%$, $p_{\text{ECS\_UP}} = 5.0\%$) with exception of Deng distribution ($p_{\text{ECS\_DP}} = 8.3\%$, $p_{\text{UP}} = 14.7\%$). These results still agree with our hypotheses when the deviations of Deng are ignored. Inaccuracies occur in the ECS-RP's and ECS/SCC-IP's distributions. While the SCC mismatch distribution is dominated by RP's (p_{SGG_RP}= 38-50%), which is consistent with our expectation, they are not significantly dominant in the ECS mismatch group (p_{ECS_RP}= 30-34.5%,). IP's differ strongly in both mismatch groups (p_{ECS}= 23-25%, p_{SCC}= 0%). However, IP's should not mismatch from consensus, as they should be conserved. Both deviations question either the correctness of the position categorization, or the DMS data. Inaccuracies in the ESC alignment can mostly be ruled out, as the optimized ESC alignment has significantly good alignment scores. It is problematic that no gradations are made during categorization, so that scores close to mean (mean = 0) impede and limit typification of these residues. This leads to the high proportion of uncategorized residues in both mismatch groups (p_{SCC}= 35.2%, p_{ECS}= 17.8%). For further analysis these uncertain positions should be analyzed, and a more sensitive categorization should be applied. Without improvement of the categorization, it is not feasible to definitively determine whether the DMS data is inaccurate or if the error stems from the categorization process.

5. References:

(1) Deng Z, Huang W, Bakkalbasi E, Brown N, Adamski CJ, Rice K, Muzny D, Gibbs R, Palzkill T (2012). *Deep sequencing of systematic combinatorial libraries reveals β-lactamase sequence constraints at high resolution*. 424(3-4):150-67. doi: 10.1016/j.jmb.2012.09.014. Epub 2012 Sep 25. PMID: 23017428 PMCID: PMC3524589

(2) Fassler J, Cooper P (2011). *BLAST Glossary*. 2011 Jul 14. In: BLAST® Help [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2008-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK62051/

(3) Firnberg E, Labonte J, Gray J, Ostermeier M (2016). *A Comprehensive, High-Resolution Map of a Gene's Fitness Landscape*. 33(5):1378. doi: 10.1093/molbev/msw021. Epub 2016 Feb 23. PMID: 26912810 PMCID: PMC4839222

(4) Fischer JD, Mayer CE, Söding J (2008). *Prediction of protein functional residues from sequence by probability density estimation*. Bioinformatics, Volume 24, Issue 5, March 2008, Pages 613–620, https://doi.org/10.1093/bioinformatics/btm626

(5) http://emboss.open-bio.org/rel/dev/apps/embossdata.html; last accessed: 17.07.2023 7pm

(6) https://galaxy-iuc.github.io/emboss-5.0-docs/cons.html; last accessed: 10.07.2023 6pm

(7) Jia K, Jernigan RL (2021). *New amino acid substitution matrix brings sequence alignments into agreement with structure matches*. 2021 Jun;89(6):671-682. doi: 10.1002/prot.26050. Epub 2021 Feb 2. PMID: 33469973; PMCID: PMC8641535.

(8) Leander M, Liu Z, Cui Q, Raman S (2022). *Deep mutational scanning and machine learning reveal structural and molecular rules governing allosteric hotspots in homologous proteins*. 11:e79932. doi: 10/7554/eLife.79932. PMID: 36226916 PMCID: PMC9662819

(9) Madden T (2015). *The BLAST Sequence Analysis Tool*. 2013 Mar 15. In: The NCBI Handbook [Internet]. 2nd edition. Bethesda (MD): National Center for Biotechnology Information (US); 2013-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK153387/

(10)Mount DW (2008). *Comparison of the PAM and BLOSUM Amino Acid Substitution Matrices*. CSH Protoc. 2008 Jun 1:pdb.ip59. doi: 10.1101/pdb.ip59. PMID: 21356840

(11)Notin P, Dias M, Frazer J, Hurtado JM, Gomez A, Marks D, Gal Y (2022). *Proceedings of the 39th International Conference on Machine Learning*. PMLR 162:16990-17017

(12)Pearson WR (2013). *Selecting the Right Similarity-Scoring Matrix*. Curr Protoc Bioinformatics; 2013;43:3.5.1-3.5.9. doi: 10.1002/0471250953.bi0305s43. PMID: 24509512; PMCID: PMC3848038

(13)Sternke M, Tripp KW, Barrick D. (2020). *The use of consensus sequence information to engineer stability and activity in proteins*. Methods Enzymol.;643:149-179. doi: 10.1016/bs.mie.2020.06.001. Epub 2020 Jul 17. PMID: 32896279; PMCID: PMC8098710

(14)Stiffler M, Hekstra D, Ranganathan R (2015). *Evolvability as a function of purifying selection in TEM-1 β-lactamase*. 60(5):882-892. doi: 10.1016/j.cell.2015.01.035. PMID: 25723163

6. Appendix

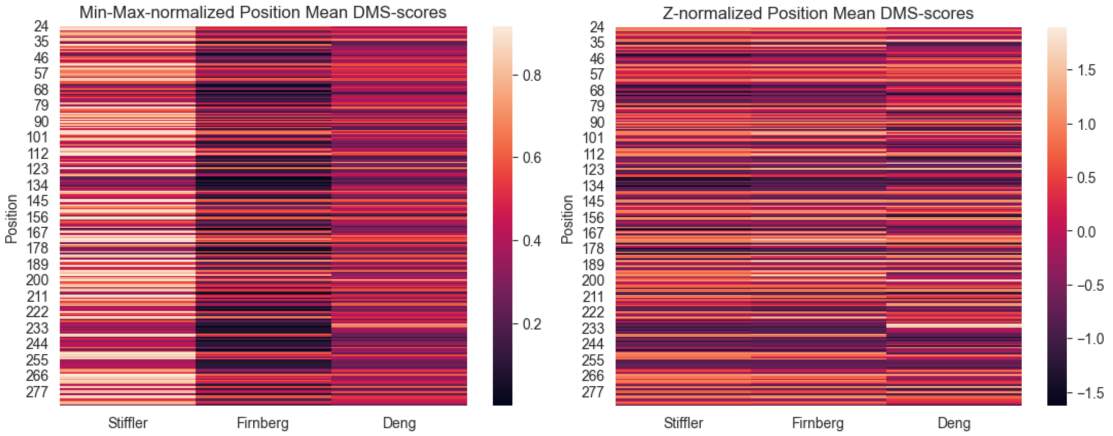

Figure 1: Visual assessment of Min-Max-normalisation and Z-normalisation of the datasets. While the Min-Max-Normalization results in similar patterns between the Stiffler and Deng datasets, the data seems to be out of scale. Stiffler-values are consistently higher than corresponding Firnberg-values. The Z-normalization on the other hand produces comparable data.

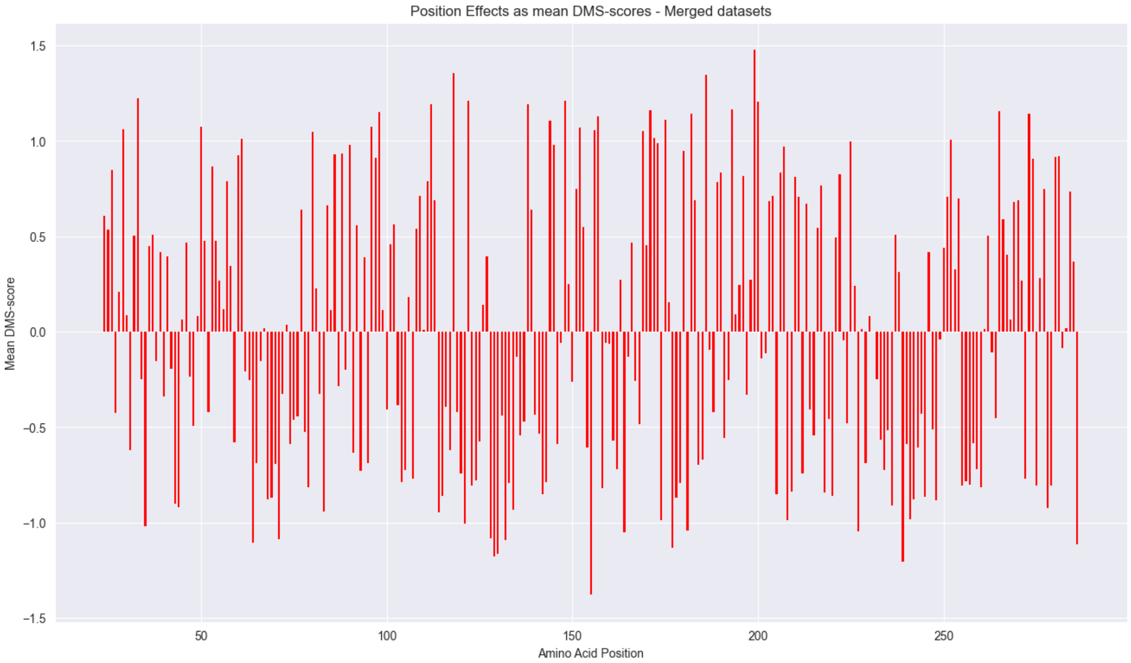

Figure 2: Position Effects of merged datasets as mean DMS-score: The mean DMS-scores of all three datasets of each position describe the robustness (high mean DMS-score) or fragility (low mean DMS-score) of each amino acid position.

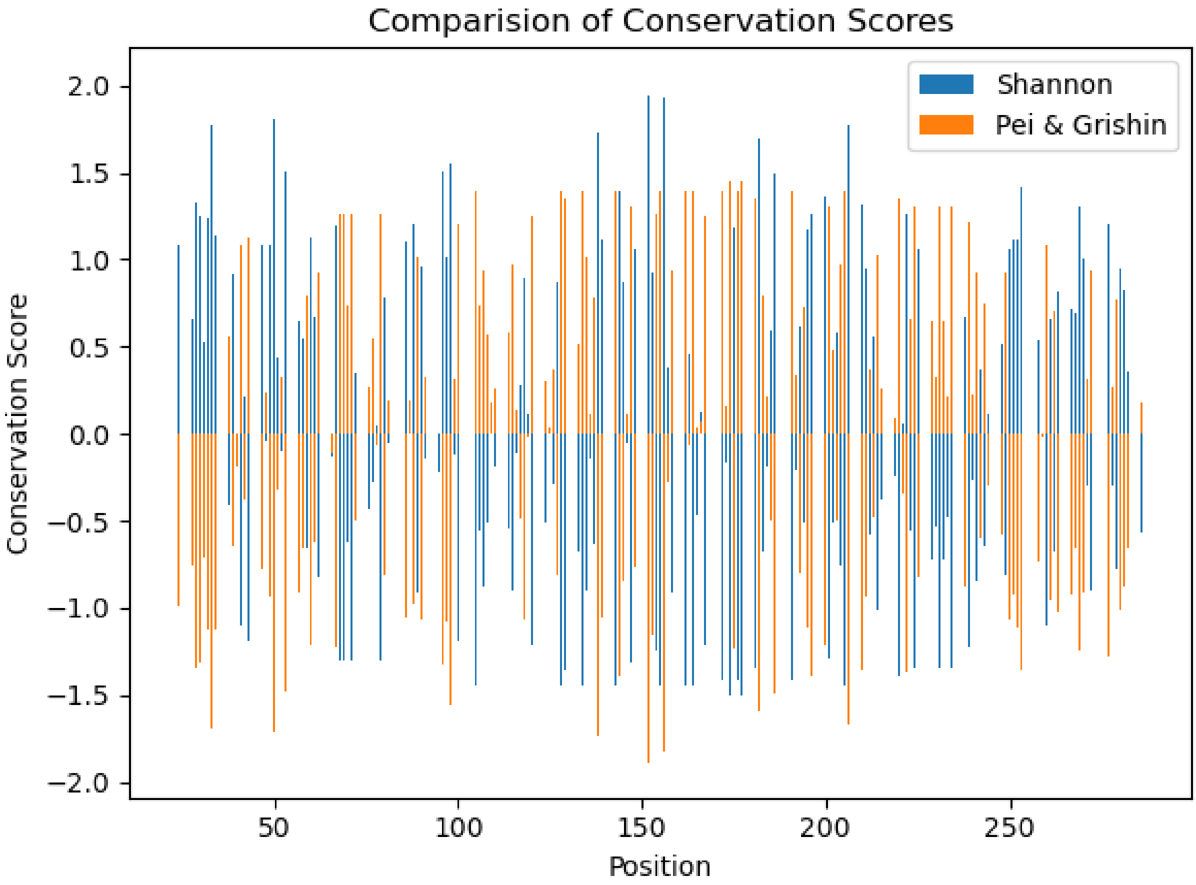

Figure 3: Comparision of z-normalized Shannon entropy (blue) and Pei & Grishin (orange) conservation for TEM-1 protein sequence. Shannon Entoropy (blue) and Pei & Grishin Conservation (orange) are compared. Negative Shannon Entropy indicates high conservation, whereas a high Pei & Grishin Conservation value corresponds to high conservation of a residue.

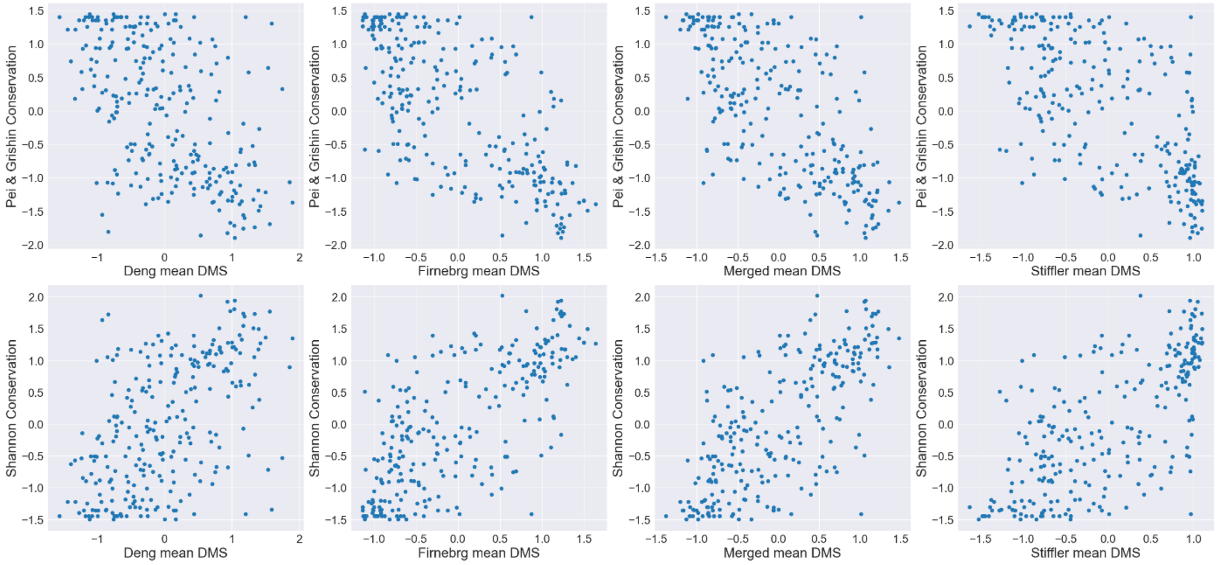

Figure 4: Correlation scatter plot of the four DMS data sets with conservation calculated by Shannon or Pei & Grishin formula. Correlation plots of Z-normalized Shannon and Pei & Grishin Conservations scores and Z-normalized mean DMS-scores for each position from each dataset, as well as the merged model are shown.

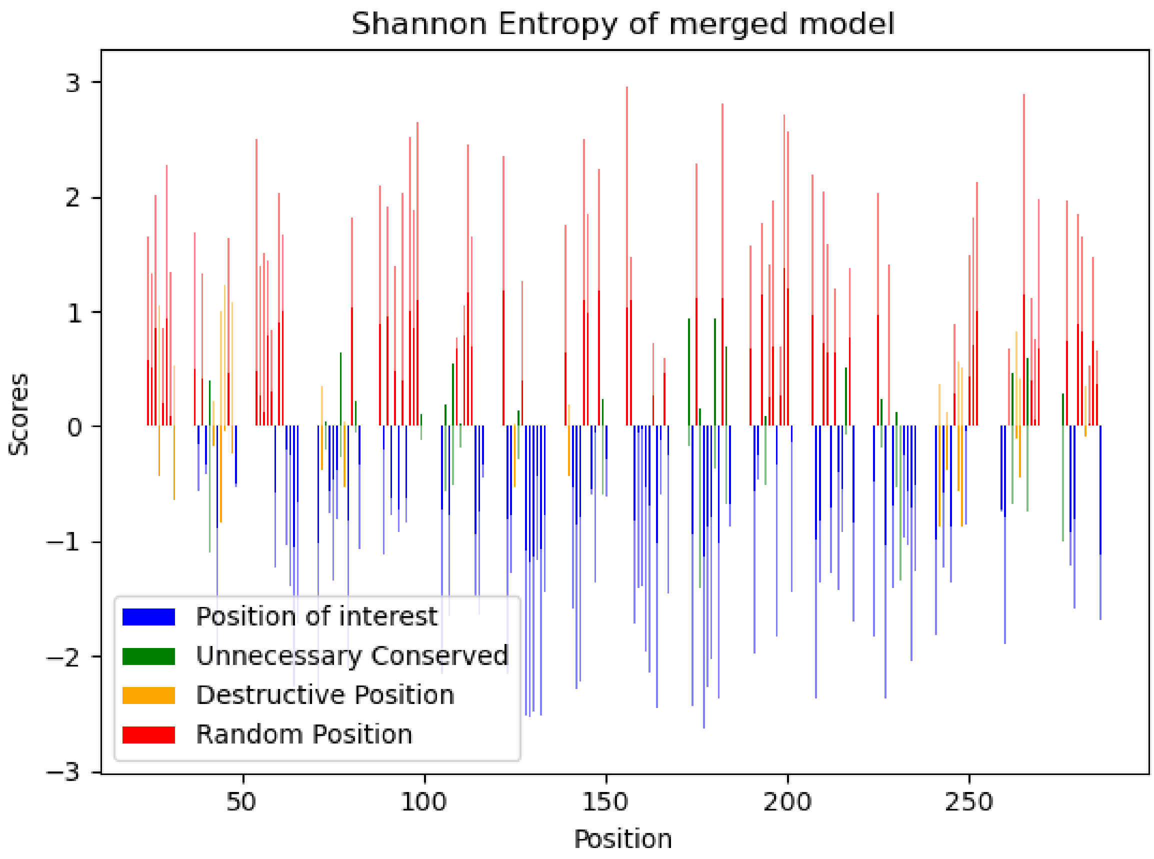

Figure 5: Categorization of TEM-1 residues with Shannon entropy (lighter color tone) and merged model DMS (darker color tone). The z-normalized Shannon Entropy and z-normalized mean DMS-scores for each position from the merged datasets are plotted. All positions are categorized into Positions of Interest (low DMS, high conservation), Unnecessarily Conserved Positions (high DMS, high conservation), Destructive Positions (low DMS, low conservation) and Random Positions (high DMS, low conservation).

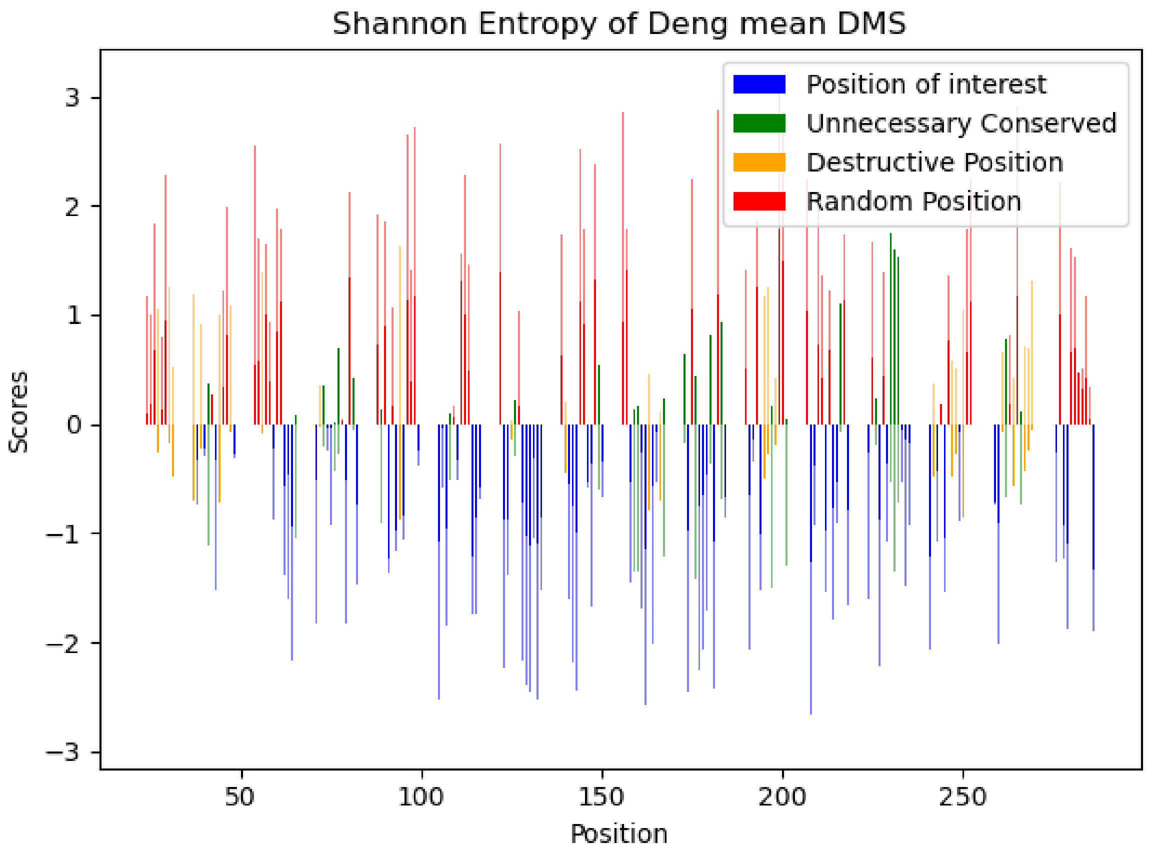

Figure 6: Categorization of TEM-1 residues with Shannon entropy (lighter color tone) and Deng DMS (darker color tone). The z-normalized Shannon Entropy and z-normalized mean DMS-scores for each position from the Deng dataset are plotted. All positions are categorized into Positions of Interest (low DMS, high conservation), Unnecessarily Conserved Positions (high DMS, high conservation), Destructive Positions (low DMS, low conservation) and Random Positions (high DMS, low conservation).

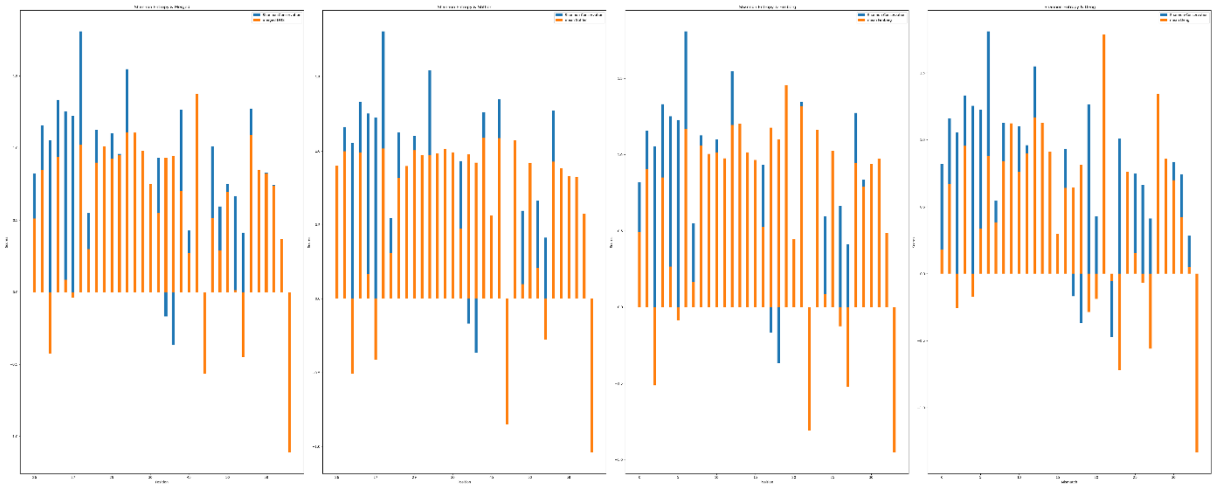

Figure 7: SCC Mismatch distribution Positions. From left to right the z-normalized DMS (blue) of following data sets is plotted, together with the z-normalized Shannon entropy (blue): Merged model, Stiffler, Firnberg, Deng.

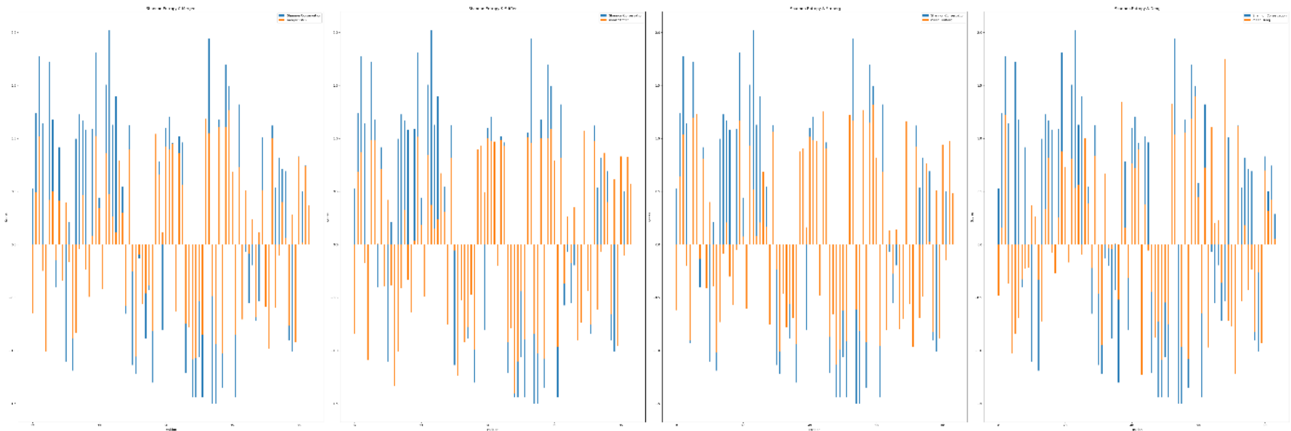

Figure 8: ESC BLOSUM90 Mismatch distribution Positions. From left to right the z-normalized DMS (blue) of following data sets is plotted, together with the z-normalized Shannon entropy (blue): Merged model, Stiffler, Firnberg, Deng.