Bioinformatics, 2026, **42**(5), btag188
https://doi.org/10.1093/bioinformatics/btag188
Published: 26 April 2026
Original Paper | Sequence analysis

OXFORD

Much ado about nothing: modeling amino acid replacement with predicted protein structures

Lukas Buschmann[1,2,‡], Sarah Naomi Bolz[1,‡], Ferras el-Hendi[1,2], Negin Malekian[1], Michael Schroeder[1,2,⋆]

[1]BIOTEC, CMCB, TU Dresden, Dresden 01062, Germany
[2]ScaDS.AI, TU Dresden, Dresden 01062, Germany
⋆Corresponding author. Biotec, TU Dresden, Dresden 01062, Germany. E-mail: michael.schroeder@biotec.tu-dresden.de
‡Equal contribution.
Associate Editor: Can Alkan

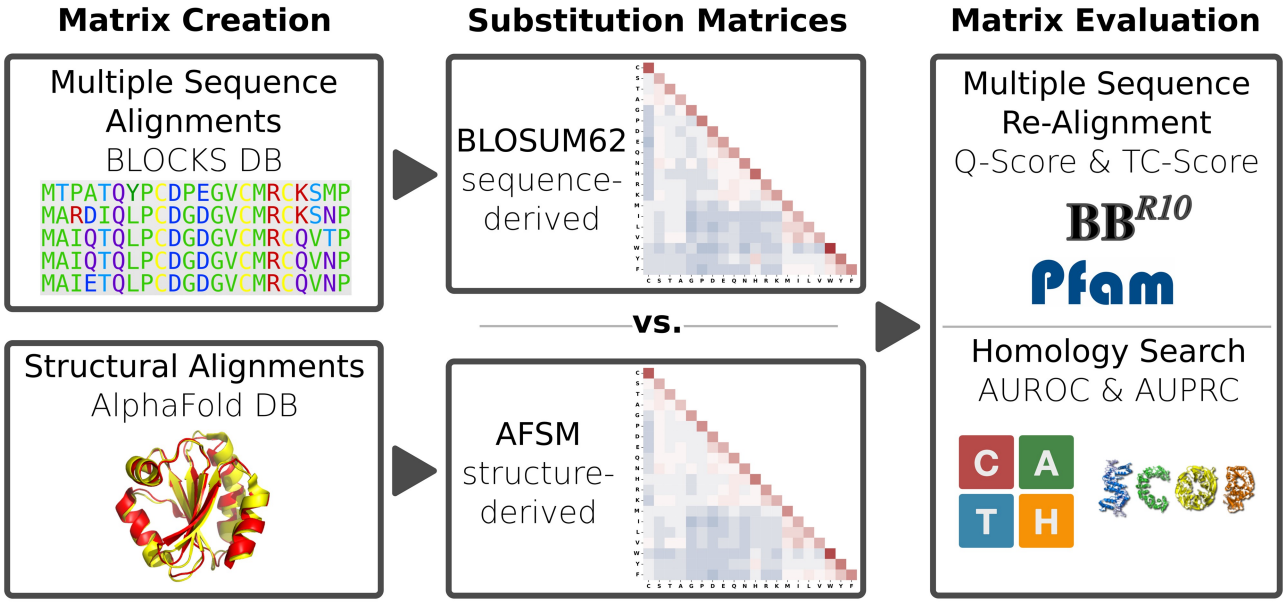
Abstract

Motivation: **Substitution matrices like BLOSUM62 model the likelihood of replacement of amino acids in evolution. Substitution matrices are used in protein sequence alignment tasks. Since the introduction of BLOSUM62 over three decades ago, many matrices have been released. Yet, to date, no effort uses large amounts of 3D structures predicted by AlphaFold.**

Results: **Here, we define AFSM, the AlphaFold Substitution Matrix derived from over 20 000 predicted 3D structures following the BLOSUM methodology. We benchmark AFSM against BLOSUM62 and 16 other matrices on five tasks in multiple sequence alignment (MSA) and protein homology search. Our analysis surprisingly reveals that all matrix families perform similarly. Only when there are few sequences in an MSA do BLOSUM62 and AFSM perform better than using no matrix. This suggests that substitution matrices were most beneficial when there was little sequence data. We corroborate this argument by showing that embeddings, which are computed from billions of sequences, perform better than substitution matrices, when sequence data is sparse. Taken together, this suggests that structural data does not improve BLOSUM62. But increased sequence data makes extrapolation with substitution matrices obsolete. Nonetheless, BLOSUM62 continues to capture chemists' intuition on amino acids by providing numerical values implicitly reflecting physicochemical properties, and it remains indispensable for sparse MSAs and direct comparison of two sequences.**

Availability and implementation: **Data is available from doi.org/10.5281/zenodo.18777546**

Graphical abstract



Received: 6 August 2025. **Revised:** 2 March 2026. **Accepted:** 25 March 2026

1 Introduction

Substitution matrices quantify the likelihood of amino acid replacements in evolution. They have been fundamental in the construction of multiple sequence alignments (MSAs) and remote homology search. They guide the alignment process by assigning scores that favor biologically meaningful matches and penalize unlikely substitutions, thereby enabling the detection of evolutionary relationships between sequences. Two widely used matrices, PAM and BLOSUM (Henikoff and Henikoff 1992), date back to 1978 and 1992, a period in which the amount of sequence data rose from hundreds to a hundred thousand. This unprecedented growth continued, with a hundred million sequences available in 2009 (www.ncbi.nlm.nih.gov/genbank/statistics). As sequence data grew, new substitution matrices were published. PAM was refined with more sequence data and an improved statistical model by Benner et al. (1994) and later by Müller and Vingron (2000) and Müller et al. (2002), resulting in the GONNET and VTML matrices. For BLOSUM, the re-implementations RBLOSUM (Styczynski et al. 2008) and CORBLOSUM (Hess et al. 2016) were published. In 2017, the PFASUM matrices (Keul et al. 2017) were calculated from the Pfam Seed alignments (Mistry et al. 2021), which consist of more than 21 000 high-quality multiple sequence alignments for Pfam protein families.

The first structural approaches emerged in the year 2000. Protein structure is more conserved than sequence (Illergard et al. 2009), which implies that it should be better suited for the construction of matrices. However, there was far less structural data available than sequence data. In 2000, when there were some ten million sequences, the protein data bank PDB held 13 583 protein structures (Berman et al. 2000). Far less, but sufficient for Prlic et al. (2000) to build two substitution matrices, SDM and HSDM. Two decades later, PDB held nearly ten times as much data, and two more approaches used structural data. Jia and Jernigan (2021) analyzed co-evolving residues in over 5000 structural families (ProtSub).

All of these matrices predate the arrival of AlphaFold (Jumper et al. 2021; Abramson et al. 2024) and the availability of hundreds of millions of precomputed predicted protein structures (Varadi et al. 2024). Thus, in this paper, we introduce AFSM (see Fig. 1), the AlphaFold substitution matrix. We roughly follow the BLOSUM approach in construction, but derive substitutions from over 660 000 structural alignments for over 20 000 proteins. We then focus on how AFSM relates to BLOSUM62 and all of the above substitution matrices. Next, we introduce five benchmarks to evaluate the performance of AFSM. Two of the benchmarks address the construction of MSAs. The first, BAliBASE (Thompson et al. 2011), is a standard in the field. It is well curated, but comparatively small. Therefore, we consider a second MSA benchmark derived from Pfam (Mistry et al. 2021), which covers 21 000 protein families. We systematically compare the two benchmarks in terms of insertions, deletions, mismatches, matches, length of sequences, and number of sequences to assess their degree of difficulty. The MSA benchmark is complemented by benchmarks in remote homology search with varying degrees of sequence redundancy. The main result of this article is the systematic comparison of AFSM and all the above substitution matrices on these five benchmarks, revealing how they perform in identical setups.

2 Results

2.1 Construction of the AlphaFold Substitution Matrix (AFSM)

To construct the AlphaFold Substitution Matrix, we randomly selected 300 out of over 85 000 InterPro (Blum et al. 2025) protein

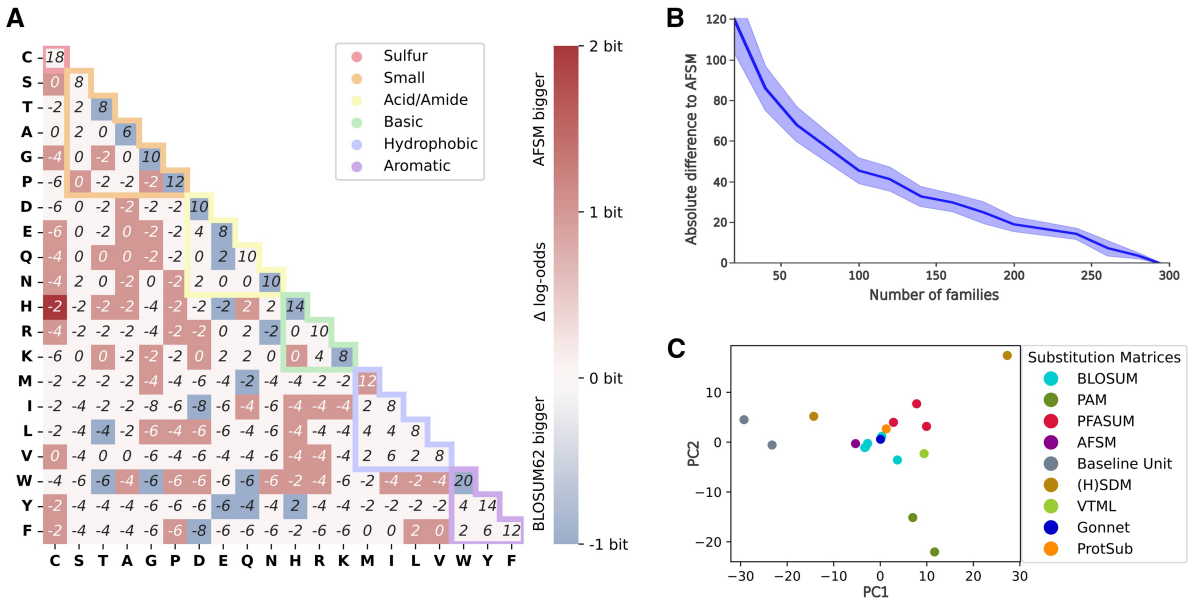


Figure 1 (A) AFSM substitution matrix. Deviation from BLOSUM62 indicated by background color. Nearly half of the entries differ slightly. (B) Convergence behavior in matrix construction. With an increasing number of families covered, AFSM construction quickly converges to the final matrix. Shown are the mean and one standard deviation across 1000 random subsets for each subset size. (C) Principal component analysis of 18 substitution matrices; from BLOSUM series: 45, 62, 80, CORBLOSUM61, RBLOSUM64; PAM series: 120, 250; PFASUM series: 31, 43, 60; (H) SDM: SDM, HSDM; Baseline Unit: 5/0, 5/-1.

families. From each family, up to 100 protein sequences were randomly sampled, resulting in a dataset comprising approximately 21 000 unique proteins. For each of these proteins, we retrieved the predicted structures from the AlphaFold database (Jumper et al. 2021; Varadi et al. 2022). For 3 of the 300 families, there were no structures at all, such that the subsequent analysis was based on 297 families only. Within each family, all possible protein pairs were generated. To mitigate redundancy and avoid overrepresentation of highly similar sequences, we filtered out pairs with high sequence identity (>62%). The remaining pairs were subjected to structural alignment. Residue-residue correspondences were then extracted from structurally aligned regions, based on the spatial proximity of residues, resulting in ungapped pairwise alignments. To account for family size variability, alignments were weighted inversely proportional to the number of pairs per family, thus reducing bias from larger families. Finally, substitution scores were computed as log-odds in half-bit units, following BLOSUM62 (Henikoff and Henikoff 1992). These scores were subsequently rounded to the nearest integer to produce the final matrix, as depicted in Fig. 1A.

To understand how the matrix depends on the amount of data, we varied the number of families. Fig. 1B suggests that further increasing the number of families would have had a limited effect on the final AFSM matrix (using 200 instead of 297 families leads to a total difference of 20 only).

2.2 Comparison of AFSM to BLOSUM62

Consider Table 1, which lists BLOSUM62, AFSM, and selected other matrices, showing meta information such as year of introduction, amount and type of data used, and similarity to BLOSUM62. Before presenting the benchmark results for all matrices in the table, let us focus on AFSM and BLOSUM62. Although AFSM uses structure instead of sequence data, and although the amount of data is over 100 times larger than for BLOSUM62, the AFSM matrix is similar to BLOSUM62, with 48% of 200 substitution scores identical. The 52% differing entries mostly show only small differences of ±1, except histidine and cysteine with a difference of 2 (see Fig. 1A). Overall, AFSM has a very strong Pearson correlation of 96% to BLOSUM62. This strong resemblance is also reflected in Fig. 1C, which depicts each substitution matrix listed in Table 1 as a high-dimensional vector, whose dimensions are reduced to two principal components (PCA). An interesting detail regarding AFSM and BLOSUM62 concerns their Evolutionary Distance (column Evo Dist in Table 1), which we captured as a ratio between match and mismatch score. More precisely formulated, the Evolutionary Distance is the absolute value of a matrix' ratio between average diagonal scores (match) over average off-diagonal scores (mismatch). BLOSUM45, 62, and 80 have distances of 5.3, 4.1, and 3.1, respectively, reflecting the degree of conservation of the underlying sequence data. With an Evolutionary Distance of 4.4, AFSM is closest to BLOSUM62 in this respect.

2.3 Benchmarks

To evaluate the performance of AFSM, BLOSUM62, and the other substitution matrices, we employ two tasks: MSA construction and protein homology search. A widely used standard benchmark in MSA construction, as noted in previous work (Blackshields et al. 2006; Edgar 2009; Long et al. 2016; Keul et al. 2017), is BAliBASE (Thompson et al. 2011), which provides curated reference MSAs designed to represent a broad range of biologically relevant alignment scenarios. Since BAliBASE R10 is comparatively small (209 MSAs), we added a subset of the Pfam Seed alignments (18 061 MSAs) as a second benchmark. The Pfam Seed dataset offers a large collection of manually curated

Table 1 Overview of substitution matrices and their metadata.

Matrix	Pairs	Year	Str	Evo Dist	Corr
AFSM	68 423 952	2026	yes	4.4	96%
BLOSUM45	500 000	1992		5.3	96%
BLOSUM62	1 250 000	1992		4.1	100%
BLOSUM80	4 000 000	1992		3.1	98%
CORBLOSUM61		2016		4.0	98%
Gonnet	1 700 000*	1992		5.3	93%
HSDM	9947	2000	yes	5.5	89%
PAM120	471 600*	1978		2.5	85%
PAM250	471 600*	1978		3.9	84%
PFASUM31	24 000 000*	2017		6.3	95%
PFASUM43	57 000 000*	2017		5.3	95%
PFASUM60	147 000 000*	2017		4.6	97%
PROTSUB		2021	yes	4.9	96%
RBLOSUM64	1 300 000*	2008		4.0	98%
SDM	13 908	2000	yes	5.9	88%
unit matrix 5, 0					72%
unit matrix 5,-1					72%
VTML200		2002		4.2	94%

Abbreviations: Corr = Pearson correlation with BLOSUM62; Evo Dist = evolutionary distance as absolute value of mean diagonal to off-diagonal; high = large; low = small evolutionary distance; Pairs = estimated number of residue pairs used for construction (estimates marked with *); Str = structural data used; Year = year of publication.

alignments across a diverse range of protein families. This inclusion allowed us to test the generalizability of AFSM on a broader and more diverse alignment set.

Despite their different scopes, the two benchmarks exhibit broadly comparable statistical properties relevant to MSAs (see Fig. 2). However, there are notable differences in key factors that affect alignment difficulty.

One important factor is the number of sequences per MSA (see Fig. 2A). BAliBASE alignments typically contain a relatively consistent number of sequences, with a mean of 68. In contrast, Pfam displays a distribution closer to a power law, with a mean of 50 but a much lower median of 23. Notably, approximately 6% of Pfam alignments consist of only two sequences, effectively reducing them to pairwise alignments.

Sequence identity within MSAs (Fig. 2B) is another relevant factor. We calculated the average pairwise sequence identity across all alignments and observed modest differences: BAliBASE alignments averaged 31%, whereas Pfam alignments were slightly higher at 39%. Both are close to the so-called twilight zone (Rost 1999), where sequence comparison becomes difficult. The difference of 8% may be attributable to Pfam's focus on conserved domains, in contrast to BAliBASE's inclusion of full-length proteins, which often contain more variable regions. As higher sequence identity generally corresponds to easier alignment tasks, this may give a slight advantage to alignments drawn from the Pfam dataset.

Another notable distinction between the two benchmarks lies in sequence length (Fig. 2C). Pfam covers domains and BAliBASE entire proteins, resulting in longer sequences on average. In addition to sequence length and pairwise sequence identity, we also examined the overall match and mismatch rates across the two benchmark datasets. The match rate is defined as the number of matches divided by the sum of the number of matches, mismatches, and residue-gap pairs. The bivariate density plot of the mismatch and match rates (Fig. 2D) indicates that there are relatively more mismatches than matches in both MSA benchmarks. BAliBASE is closer to the origin than Pfam Seed because it has more gap-residue pairs. Pfam Seed forms a counter diagonal, which corresponds to ungapped MSAs. Overall, BAliBASE alignments exhibit very similar ratios of match to mismatch rates among all residue-residue pairs, suggesting that the underlying alignment difficulty is broadly comparable to Pfam Seed in terms of residue correspondences.

However, differences in absolute match and mismatch rates arise primarily due to varying gap rates. BAliBASE alignments tend to include more gaps, resulting in a lower total number of residue-residue pairs. In contrast, Pfam Seed alignments often contain fewer gaps, leading to a higher proportion of residue-residue pairs. This distinction is particularly evident in the density plot of Fig. 2D of match versus mismatch rates, where Pfam alignments form a sharp boundary around the line where the sum of match and mismatch rate is one, reflecting the near absence of gaps in many of these alignments.

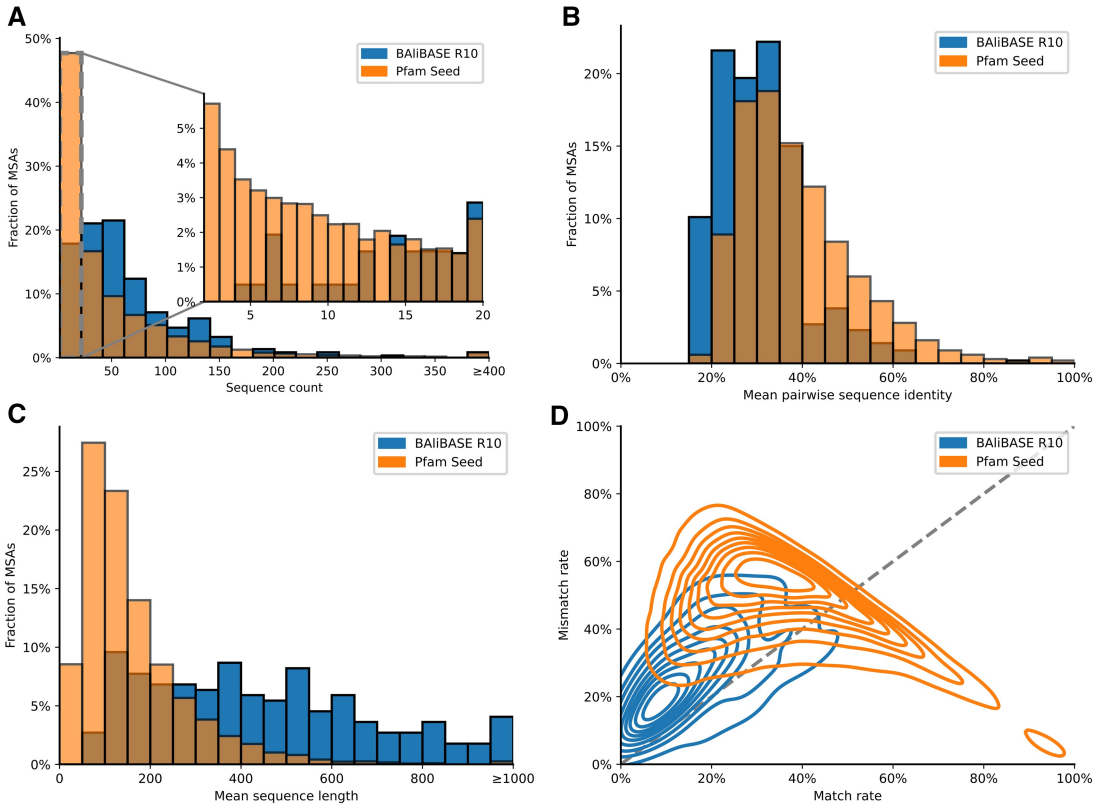


Figure 2 BAliBASE versus Pfam Seed benchmarks in comparison. (A) Distribution of the number of sequences per MSA (overview for 2 to 150 sequences and zoomed in for 2 to 20 sequences). (B) Sequence identity per MSA. (C) Mean sequence length per MSA. (D) Match versus mismatch rate. Match (x-axis) plus mismatch (y-axis) plus gap rate (not shown) equals 100%. Overall (A–D), both BAliBASE and Pfam Seed have comparable characteristics as MSA benchmarks, with Pfam Seed MSAs having slightly fewer and shorter sequences of greater sequence identity. Both generally have a higher mismatch than match rate.

Besides BAliBASE and Pfam Seed as MSA construction tasks, we evaluated the matrices' performance in three remote homology tasks. The structural domain classification databases CATH and SCOP (Pearl et al. 2003; Andreeva et al. 2020) classify domains hierarchically into more closely related families and more distantly related superfamilies. The remote homology search tasks consist of determining whether a pair of sequences belongs to the same superfamily or not. The difficulty of this task can be varied by the amount of redundancy per superfamily. Filtering each superfamily at 20% sequence identity (CATH S20 dataset) is more challenging than at 40% (SCOPe40 dataset), and 40% is more challenging than not filtering at all (CATH dataset). For details, see methods.

2.4 Benchmark results

The sequences taken from the reference MSAs from BAliBASE and Pfam Seed are realigned using the method under evaluation. Then the resulting alignments are compared against the original benchmarks. To quantify alignment accuracy, we used two widely accepted metrics (Edgar 2009; Long et al. 2016; Keul et al. 2017):

Q-score (Sum-of-Pairs Score): The Q-score evaluates how many residue pairs are correctly aligned across all sequences. It does not require entire columns to be preserved, making it a more permissive and commonly used metric in MSA evaluations.

TC-score (Total Column Score): This metric represents the percentage of correctly aligned columns relative to the reference MSA. It is a stringent measure, requiring complete agreement across all sequences in a column.

To assess the quality of the three homology search tasks, we report for each the area under the ROC curve and the area under the Precision-Recall curve.

2.5 AFSM versus BLOSUM62

Assuming that structural alignments are more accurate than sequence alignments and given that roughly half of the entries in AFSM are different from BLOSUM62, one can expect modest improvements for the evaluation metrics (TC-score and Q-score). This is, however, not the case (see Table 2). AFSM and BLOSUM62 perform nearly identically. This outcome is consistent across the BAliBASE (Q-score 71%/70%) and Pfam Seed (82%/81%) MSA construction benchmarks and the three protein homology search tasks. As expected based on prior comparisons between the benchmarks, alignments on the Pfam Seed dataset yielded slightly better scores overall. This likely reflects the higher sequence identity and lower gap content of domain-focused alignments compared to the more variable and gap-rich alignments in BAliBASE. Similarly, the results for AFSM (87%, 79%, 64%) and BLOSUM62 (87%, 80%, 64%) consistently reflect the difficulty of homology search tasks. Interestingly, gap penalties, which were optimized per matrix for the MSA construction task, were remarkably similar (see Table 3).

2.6 Choice of substitution matrices

The above results are unexpected. Despite more and other data, AFSM and BLOSUM62 perform nearly identically across five different tasks. To put this finding into perspective, we added 16 other matrices to the analysis, including classical sequence-based series PAM and BLOSUM (Henikoff and Henikoff 1992), more recent improvements such as VTML (Müller and Vingron 2000, Müller et al. 2002), RBLOSUM (Styczynski et al. 2008), CORBLOSUM (Hess et al. 2016), and PFASUM (Keul et al. 2017), as well as structure-derived matrices like SDM and HSDM (Prlic et al. 2000) and ProtSub (Jia and Jernigan 2021). Besides these

Table 2 Performance of substitution matrices across BAliBase R10, Pfam Seed, and fold recognition benchmarks.

Matrix	BAliBase R10		Pfam Seed		CATH		SCOPe40		CATH S20	
	Q	TC	Q	TC	ROC	PR	ROC	PR	ROC	PR
AFSM	71	31	82	56	87	56	79	11	64	2
BLOSUM45	72	32	83	57	87	55	81	13	65	2
BLOSUM62	70	31	81	55	87	55	80	12	64	2
BLOSUM80	64	26	77	50	88	54	82	11	69	2
CORBLOSUM61	69	30	81	54	88	56	80	12	65	2
Gonnet	73	33	83	57	86	55	79	13	63	2
HSDM	72	30	83	56	84	54	76	5	61	2
PAM120	62	24	75	48	87	53	81	11	67	1
PAM250	71	31	82	56	86	54	79	12	63	3
PFASUM31	73	32	84	58	85	53	77	7	60	1
PFASUM43	72	32	83	57	87	54	79	8	63	2
PFASUM60	71	31	83	56	87	54	79	8	63	2
PROTSUB	71	32	82	56	86	55	78	11	62	1
RBLOSUM64	69	30	81	54	88	56	81	12	65	2
SDM	72	32	83	57	87	54	79	7	66	2
unit matrix 5, 0	72	31	82	54	75	37	64	0	51	0
unit matrix 5,-1	63	25	75	46	83	49	78	6	66	1
VTML200	71	31	82	56	85	53	76	11	60	1

Abbreviations: Q = Q-score; TC = TC-score.
ROC and PR report AUC values for the corresponding tasks. All values are given in percent.

Table 3 Optimal gap penalties for substitution matrices on BAliBASE R10 and Pfam Seed benchmarks.

Matrix	BAliBASE R10				Pfam Seed			
	Q		TC		Q		TC	
	GOP	GEP	GOP	GEP	GOP	GEP	GOP	GEP
AFSM	5	0.5	7	0.5	6	0.5	7	0.5
BLOSUM45	9	0.5	8	1	7	1	9	1
BLOSUM62	6	0.5	8	0.5	7	0.5	7	1
BLOSUM80	7	0.5	9	0.5	7	0.5	7	1
CORBLOSUM61	6	0.5	7	0.5	7	0.5	7	1
Gonnet	7	0.5	9	0.5	7	0.5	7	1
HSDM	9	1	9	2	9	1.5	9	2.5
PAM120	7	0.5	7	0.5	7	0.5	7	1
PAM250	8	0.5	9	1	7	1	8	1
PFASUM31	9	0.5	9	1	8	1	9	1
PFASUM43	8	0.5	9	1	7	1	8	1
PFASUM60	9	0.5	9	1.5	8	1	9	1
PROTSUB	8	0.5	9	1	8	1	9	1
RBLOSUM64	6	0.5	7	0.5	7	0.5	7	1
SDM	4	0.5	7	0.5	4	0.5	5	0.5
unit matrix 5, 0	6	0	7	0.5	5	0.5	6	0.5
unit matrix 5,-1	5	0.5	7	0.5	6	0.5	7	0.5
VTML200	7	0.5	9	1	7	1	9	1

Abbreviations: GEP = gap extension penalty; GOP = gap opening penalty.

established matrices, we added two controls, which only distinguish matches and mismatches. To make these controls comparable to the above matrices, we assigned matches a score of 5, which is in line with the other matrices whose diagonals average to values of 4 to 6. For mismatches, we considered two scores of 0 and -1, respectively. Essentially, these two controls reflect the absence of a substitution matrix, and since they are similar to a unit matrix, we named them baseline unit matrix 5/0 and 5/-1.

2.7 Growth of data

Table 1 lists the number of residue pairs used in the calculation of each substitution matrix. Over the decades spanning the development of these matrices, there has been an increase of roughly two orders of magnitude in the amount of data used. For example, the original PAM matrices, introduced in the late 1970s, were based on an estimated 470 000 residue pairs. In this work, we derived the AFSM from over 68 million residue pairs, reflecting both the growth in available data and advances in computational methods.

2.8 Similarity of substitution matrices

To provide an overall impression of the similarity between matrices, we represented each substitution matrix as a high-dimensional vector and applied dimensionality reduction, as shown in Fig. 1C. The resulting projection reveals that most matrices, including the BLOSUM series, AFSM, Gonnet, VTML, and PFASUM matrices, cluster closely together near the center, reflecting their high degree of similarity. In contrast, the baseline matrices form a distinct cluster separate from this main group, as do the PAM matrices. The structure-derived SDM and HSDM matrices also lie outside the central cluster. They are separated from each other as well.

2.9 Nearly all substitution matrices are similar to BLOSUM62

Despite differences in approach, whether the matrices were derived from sequence alignments or structural data, as indicated in Table 1, and substantial variation in the amount of data used for their construction, the matrices show remarkably little deviation from BLOSUM62. Table 1 includes a column reporting the correlation of each matrix with BLOSUM62. Most matrices (excluding the baseline unit matrices) exhibit a correlation above 84%, with the majority clustering even higher, at 95% or above. This highlights the overall similarity of substitution patterns captured by these matrices, regardless of their underlying methodology or data scale. However, this similarity is expected to some extent because all matrices reflect the same underlying biological principle: conservation is likely, while substitution is unlikely. We find that even a unit matrix, with identical mismatch scores 0 and match scores (1), already correlates at 72% with BLOSUM62. Note, however, despite this similarity, the matrices do differ by Evolutionary Distance, i.e., the ratio of match to mismatch score.

2.10 Parameter optimization

To ensure a fair comparison of the matrices, we individually optimized the gap opening and extension penalties for each matrix and MSA benchmark. Table 3 summarizes these settings, showing that gap opening penalties ranged from 4 to 9, while

extension penalties were consistently low, with most values at 0.5 across all configurations. Fig. 3 gives an impression of the change in Q-score of the AFSM matrix as gap parameters vary. For gap opening penalties ranging from 0 to 9 and gap extension ranging from 0 to 4, one sees a relatively consistent Q-score for most parameter combinations greater than 2 for opening and around 0.5 and 1 for extension. The optima are close to each other for both Pfam Seed and BAliBASE, but they are not identical.

2.11 Nearly all substitution matrices perform similarly to BLOSUM62

Despite differences in origin, age, and the amount of data used, the results reveal a striking consistency: matrices with higher Evolutionary Distance perform nearly identically (see Tables 1 and 2). Q-scores were consistently higher on Pfam Seed alignments (81–83%) compared to BAliBASE (69–73%), and the same trend was observed for TC-scores (54–57% for Seed; 30–33% for BAliBASE). Notably, corrected matrices derived from BLOSUM (RBLOSUM, CorBLOSUM) did not lead to significant performance gains, consistent with findings from previous studies (Styczynski et al. 2008; Hess et al. 2016). Furthermore, no clear relationship was observed between alignment performance and the number of residue pairs, year of matrix publication, or the type of data used (sequence- versus structure-based). Similarly, there were no strong performance deviations for the homology search tasks. Most of the matrices perform with a maximum 2% difference from BLOSUM62 for the ROC AUC and similarly for the PR AUC curves. All matrices consistently show that homology search is more difficult when redundancy is low. Interestingly, the baseline unit matrix 5/0 is the single matrix performing significantly worse than BLOSUM62 in the homology search tasks, despite performing equally well in MSA construction.

2.12 BLOSUM80 and PAM120 perform worse than BLOSUM62 in MSA

BLOSUM80 and PAM120 are designed to compare highly similar sequences rather than remote homologues. Consequently, one expects these two matrices to perform worse than BLOSUM62 when comparing distant sequences. And indeed, this is the case (see Table 2). There is one interesting exception: For the CATH S20 homology search task, BLOSUM80 performs 5% better than BLOSUM62. This appears contradictory, as CATH S20 is the most challenging benchmark, whose sequences are highly divergent. In fact, they are so divergent that any sequence approach and any matrix is expected to perform poorly. One reason for the seemingly contradictory result is that gap scores were transferred from the MSA evaluation of the matrices and applied to the homology benchmarks. Because CATH S20 contains substantially more divergent sequences than the MSA benchmarks, the optimal gap scores for this benchmark may differ, potentially affecting the relative performance of substitution matrices.

2.13 Highly correlated matrices do not necessarily perform similarly

We initially expected that a high correlation between substitution matrices would imply similar performance. However, our experiments show that this is not the case. For example, BLOSUM80 has the highest correlation with BLOSUM62 among the tested matrices, yet their performance on the BAliBASE MSA task differs substantially, with a Q-score difference of 6%. In contrast, PAM250 correlates less strongly with BLOSUM62 (84% correlation) but shows a much smaller Q-score difference of only 1%.

These results indicate that the correlation between matrices has little predictive value for their performance similarity. Instead, performance in the MSA task appears to be more strongly associated with the Evolutionary Distance of the matrices.

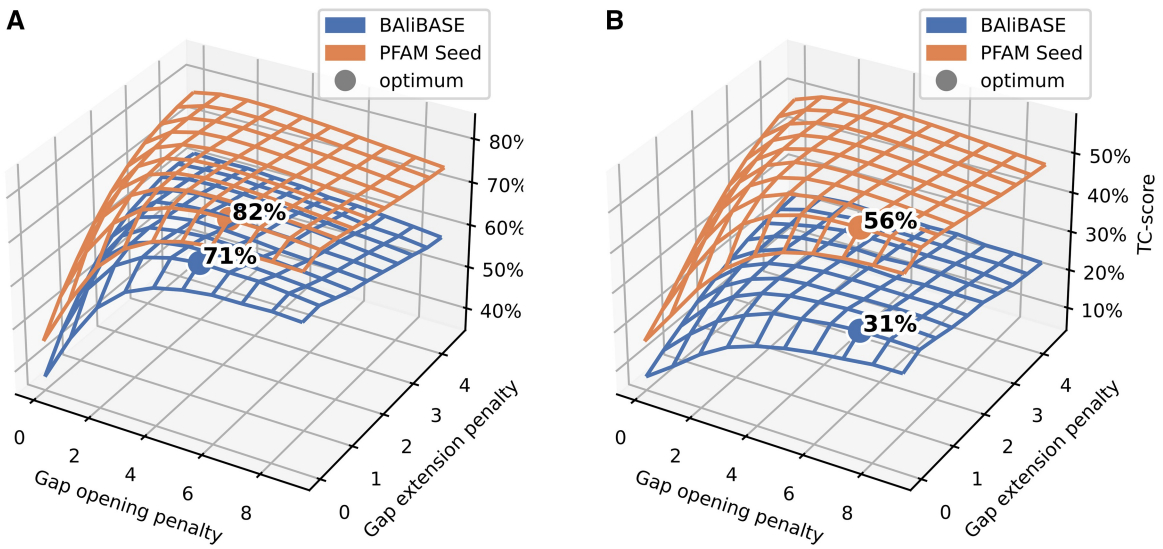


Figure 3 Optimal gap penalties and corresponding alignment scores for the AFSM matrix on BAliBASE and Pfam Seed benchmarks. (A) Q-scores and (B) TC-scores across a range of gap opening (GOP) and gap extension (GEP) penalties. For BAliBASE, the highest Q-score (71%) was achieved with GOP 5 and GEP 0.5, while the highest TC-score (31%) was obtained with GOP 7 and GEP 0.5. For Pfam Seed, the optimal Q-score was 82% with GOP 6 and GEP 0.5, and the highest TC-score was 56% with GOP 7 and GEP 0.5. Both benchmarks show similar trends and relatively stable scores across a broad range of parameter settings.

2.14 Baseline unit matrices perform as well as BLOSUM62

A striking result of our study was the performance of the baseline unit matrices. These simple matrices consist of a single conservation score on the main diagonal and a single mismatch score for all off-diagonal entries. To match the average conservation signal of BLOSUM62, we set the conservation score to 5, which resulted in similar optimal gap penalties to those found for BLOSUM62. We tested two versions of the matrix: one where mismatches were ignored (mismatch score 0), representing a matrix suited for high evolutionary distances, and one where mismatches were mildly penalized (mismatch score −1), representing a matrix suited for low evolutionary distances like BLOSUM80 or PAM120.

Contrary to our expectations, both baseline matrices performed surprisingly well. Each performed on a par with their corresponding established matrices. Remarkably, the baseline matrix 5/0 that ignored mismatches even slightly outperformed BLOSUM62 on BAliBASE in Q-score, achieving Q-scores that were higher by about one percentage point. Similarly, baseline matrix 5/-1 performs close to BLOSUM80. Regarding the protein homology search tasks, the no matrix options generally perform slightly worse than BLOSUM62. However, the baseline unit matrix 5/-1 even outperforms BLOSUM62 on the most challenging CATH S20 task. This indicates that gap penalties play an equally important role as match/mismatch scores.

In summary, these findings underscore that the specific choice of substitution matrix has only a limited impact on MSA construction and homology search, provided the matrix is broadly appropriate for the evolutionary context.

2.15 Performance on MSA construction is dependent on MSA size

The good performance of the baseline unit matrices is surprising. One explanation may be that substitution matrices are useful when there are few sequences at a moderate distance. But as there are more and more sequences in an MSA, they impose more and more constraints arising from matches, which make the alignment task easier and not more difficult. Note, however, that besides the number of sequences, their pairwise distance also plays a role.

To test how the number of sequences influences performance, we plotted the improvement in Q- and TC-scores for BLOSUM62 and AFSM over the Baseline Unit Matrix 5/0 against the number of sequences in the MSA. Fig. 4 shows that the Q- and TC-scores for BLOSUM62 are slightly worse than using no matrix whenever there are more than around 20 sequences in the alignment. For very small numbers of sequences, BLOSUM62 improves an MSA, but for larger ones, it slightly deteriorates an MSA. This implies that BLOSUM62 was beneficial when there was little sequence data available, but its usefulness is reduced as sequence data grows. But it remains useful to this date in ordering sequences for MSA construction.

2.16 Embeddings surpass substitution matrices when sequence data is sparse

Taken together, the results of this study are intriguing. Independent of the amount of data and the type of data used, matrices perform comparably. And even the baseline unit matrices, which essentially amount to using no matrix at all, perform on a par. Does this mean that the vast amounts of sequence and structure data available today make no difference? To put this in context, we performed an additional homology search task on embeddings. Embeddings (Elnaggar et al. 2022) are numerical vectors representing sequences. They result from training a neural network on billions of sequences, thus capturing the essence of sequence space in a compact form. Embeddings can be treated as numerical vectors, which can be compared by Euclidean distance. For each of the sequences in the three homology search datasets, we obtained its embedding and selected the domain with the closest embedding (see methods), i.e. we implemented a nearest neighbor search. We then evaluated whether the query domain and its nearest neighbor are in the same superfamily. From this we computed an AUC. Similarly, we obtained for the query the domain with the highest

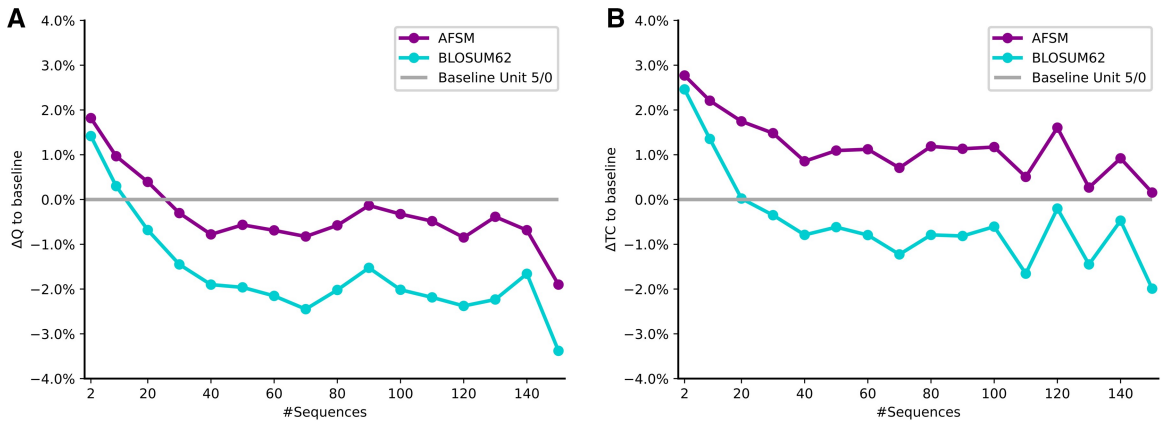


Figure 4 (A) Q-score and (B) TC-score of AFSM and BLOSUM62 for Pfam Seed compared to a baseline of using "no matrix" (Baseline Unit Matrix 5/0), in dependence of the number of sequences in an MSA. Generally, the more sequences there are, the lower the difference from the baseline. AFSM performs slightly better than BLOSUM62. Most importantly, for small amounts of sequences, AFSM and BLOSUM62 are better than no matrix; for larger amounts of sequences, they are worse.

similarity score using AFSM and BLOSUM62. And from this, we analogously computed the AUC.

We found that embeddings achieve an area under the ROC for the three homology search tasks of 99% (CATH), 94% (SCOPe40), and 90% (CATH S20). The differences nicely reflect the difficulty of the data sets resulting from increase in sparsity. In comparison, BLOSUM62 and AFSM achieve both 99%, 89%, and 78%, respectively and the baseline unit matrices 5/-1 and 5/0 scored 99%, 85%, 74% and 99%, 57%, and 62%. This is a far-reaching result. It means that when sequence data is dense, classical sequence search with or without substitution matrices performs as good as embeddings do. However, when sequence data is sparse, embeddings make a difference. Having implicitly absorbed billions of sequences, embeddings perform 12% better than the two substitution matrices on the sparse CATH S20 dataset. However, interestingly, for this dataset one of the baseline matrices performs as good as BLOSUM62 and AFSM. Overall, this means that embeddings are valuable where large evolutionary distances must be covered, i.e. when comparing e.g. protein sequences from a newly sequenced genome to model organism sequences. But, it also means as sequence data is growing and as the overall sequence density saturates, classical sequence alignment with or without substitution matrices performs comparatively.

3 Discussion and conclusion

BLOSUM62 was introduced over 30 years ago. Although it was corrected, recalculated on larger sequence data, complemented by different statistical approaches and by use of structural instead of sequence data, it nonetheless remains a default substitution matrix in protein BLAST searches. Why is BLOSUM62 still in use? Our results may contribute to an explanation. None of the existing approaches presented in this article, including AFSM, substantially improve MSAs or homology search in comparison to BLOSUM62. Logically, alternatives are not widely adopted. Our findings are consistent with Keul et al. (2017), Edgar (2009) and Price et al. (2005). However, one might ask whether our specific observation, that AFSM does not outperform BLOSUM62, could reflect a circularity, since AlphaFold structure predictions may themselves depend on BLOSUM substitution matrices. We consider this unlikely because AlphaFold predictions are structurally very similar to experimentally determined structures (Jumper et al. 2021). Thus, the observed similarity likely reflects the fact that both matrices capture the same underlying statistical properties of amino acid replacement, rather than methodological circularity. Most notably, our results show that the core of substitution properties is already captured at a crude level by a simple unit matrix. Consistent with this, on the MSA benchmarks used in our study, BLOSUM62 did not perform better than a simple unit matrix. In fact, only when very few sequences are used, BLOSUM62 performs slightly better than using no matrix. These results suggest that BLOSUM62 is important when data is sparse. This was the case in the early days of sequence comparison, and it remains the case today when fixing the order of sequences in the MSA construction process. The high relevance of the number of aligned sequences is further supported by our results on the use of embeddings in the three homology search tasks. Embeddings are statistical representations of sequences, incorporating the total population of all sequences. Thus, embeddings can be assumed to implicitly include knowledge of observed substitution patterns, which is made explicit in substitution matrices. The performance gap between embeddings and sequence comparison with a substitution matrix is large (12% on CATH S20) so that homology search should consider the use of embeddings. It is not obvious how embeddings can be used to construct MSAs, but Sgarbossa et al. (2023) introduces a deep neural network, which proves that language models can be adopted for this task. Another alternative may be novel representations of sequences. Edgar 2024 introduces a mega-alphabet, on which classical sequence algorithms perform as well as structure-based approaches. An intermediate approach using embedding-based protein sequence alignments is proposed by Pantolini et al. (2024). Do all these considerations mean that there is no need for BLOSUM62 and other matrices like AFSM? While substitution matrices, including AFSM, appear not to improve MSA construction and protein homology search, they also do not make results worse, and they turn the intuition of chemists and physicists into numbers. Indeed, this may be one of the main reasons for BLOSUM62's popularity. Manually assigning the physicochemical properties of amino acids numerical values is difficult (though there were attempts in the past; Grantham 1974). Moreover, in pairwise alignments or areas where sequence data is sparse, substitution matrices will still be needed to bridge the gaps between evolutionarily distant sequences.

4 Methods

4.1 Dataset acquisition

Pfam Seed release 35 was accessed from ftp.ebi.ac.uk in March 2024. After filtering out large MSAs (>100KB), 18 061 alignments with a total of 909 284 sequences were left. BAliBASE release 10 (R10) was accessed from www.lbgi.fr in February 2024. The 10 biggest alignments were removed. CATH version 4.3.0 and SCOPe version 2.07 were used. CATH was reduced by removing very short (<50AA) and very long (>250AA) sequences, resulting in 23 911 domains. Size reduction on the datasets was performed to ease computational demands. CATH, CATH S20 and SCOPe40 were cleaned as described by El-Hendi et al. (2026). InterPro release 95 was accessed from ftp.ebi.ac.uk.

4.2 AFSM calculation

To construct AFSM, 297 domain families with up to 100 domains each were randomly selected from InterPro. Not all families have 100 domains. To balance the difference in domain numbers per family, we introduce a weighting factor:

$$w_k = \frac{2}{n_k(n_k - 1)} \tag{1}$$

The weighting factor for family k, containing n_k member proteins.

AFSM scores were computed following the BLOSUM methodology (Henikoff and Henikoff 1992), with slight deviations in how substitution frequencies were acquired:

$$f_{ij} = \sum_{k=1}^{K} (o_{k,ij} w_k) \quad (2)$$

Weighted substitution frequency of amino acids i and j, where $o_{k,\,ij}$ is the number of observed substitutions between i and j in family k, and K is the total number of families.

From this point, we followed the BLOSUM methodology: substitution odds were calculated by dividing observed substitution frequencies by expected substitution frequencies, taking the base-2 logarithm of the result, multiplying by two to express scores in half-bit units, and rounding to the nearest integer.

4.3 MSA benchmarking

Benchmark MSAs were stripped of gap characters and realigned with the substitution matrices using ClustalW2 version 2.1 downloaded from clustal.org/download in March 2024. The following parameters were set for ClustalW2: -type=protein, -output=fasta, -negative, -matrix=[…], -gapopen=[…], -gapext=[…].

Gap opening and extension penalties were determined by exhaustive search (0–10 and 0–5, respectively)

Code to calculate Q- and TC-scores was obtained from drive5.com/qscore.

The MSA benchmarking pipeline was implemented using Nextflow.

4.4 Homology search benchmarking

Sequence alignment scores were obtained with parasail NW with optimal gap open and extension penalties as shown in Table 3 (BAliBASE Q-score). As parasail only accepts integers and optimal gap penalties were sometimes floating-point values, substitution matrices and gap penalties were scaled up accordingly. Similarity of all pairs of protein sequences of the CATH, CATH S20, SCOPe40 datasets were recorded separately for pairs in the same and in different superfamilies. From this AUC values for the ROC and PR-curves were calculated using scikit-learn (Version 1.6.1).

4.5 Embedding nearest neighbor search

Embeddings were obtained using ProtTransT5XLU50 v0.2.2 available under github.com/sacdallago/bio_embeddings (Elnaggar et al. 2022). To obtain the nearest neighbor of each embedding, the KNN-library FAISS (version 1.8.0) was used with Euclidean distance (L2) implemented by IndexFlatL2. After index construction, KNN searches were parallelized across queries using the joblib library (version 1.4.2) to enhance performance on a compute cluster. AUROC was computed by binarizing the labels and applying the one-versus-rest (OvR) approach from the sklearn.metrics module (version 1.3.2) with average='micro' and multi_class='ovr'.

4.6 Convergence analysis

To determine how data size influences the AFSM values, we varied the number n of families from 1 to 297 used to construct an AFSM. For each n, we repeated the analysis 1000 times. In Fig. 1B we report the absolute position-wide difference between the full and partial AFSM.

4.7 Residue pair estimation

To assess the size of the data used for construction of the substitution matrices (Table 1), we estimated the number of residues pairs as follows:

PAM1 is based on 1572 substitutions, which leads to 471 000 (=1572*300) residue pairs assuming an average sequence length of 300 (book.bionumbers.org/how-big-is-the-average-protein). For PFASUM, we counted the total number of residue pairs in the full set of Pfam Seed alignments at around 584 million. Since PFASUM applies the same clustering strategy as BLOSUM, we assume that it retains a similar fraction of residue pairs, allowing us to approximate the number of pairs used in each matrix variant based on known BLOSUM yields. For RBLOSUM64 we assume the same residue count as in BLOSUM, as both matrices are using the same dataset (BLOCKS 5) and only marginally different clustering.

4.8 Miscellaneous

In Table 1, Corr reports Pearson correlation (corrcoef in NumPy) and Evo Dist is mean diagonal (conservation) by mean off-diagonal (substitution).

Sequence identity in Fig. 2B was computed with parasail NW with gap open = 10, gap extend = 1, matrix=blosum62.

Match and mismatch rates in Fig. 2D were calculated per MSA in the benchmark as the proportion of identical (match) and non-identical (mismatch) residue–residue pairs, relative to the total number of residue–residue and residue–gap pairs.

Acknowledgements

Writing was assisted using ChatGPT (v 5.4) and Grammarly (v 9.63). None of the tools were used for text generation from scratch but solely for correction in spelling and grammar. All suggested changes from AI tools have been carefully checked for correctness of content. Thanks for technical support to Alexandre Mestiashvili. Special thanks to Ali Al-Fatlawi for cleaning SCOP and CATH datasets.

Author contributions

Lukas Buschmann (Conceptualization [equal], Data curation [equal], Formal analysis [equal], Investigation [equal], Methodology [equal], Resources [equal], Software [equal], Validation [equal], Visualization [equal], Writing—original draft [equal], Writing—review & editing [equal]), Sarah Bolz (Conceptualization [equal], Data curation [equal], Formal analysis [equal], Investigation [equal], Methodology [equal], Resources [equal], Software [equal], Supervision [equal], Validation [equal], Visualization [equal], Writing—original draft [equal], Writing—review & editing [equal]), Ferras El-Hendi (Data curation [equal], Formal analysis [equal], Software [equal]), Negin Malekian (Data curation [equal], Formal analysis [equal], Software [equal]), and Michael Schroeder (Conceptualization [equal], Data curation [equal], Formal analysis [equal], Investigation [equal], Methodology [equal], Supervision [equal], Validation [equal], Writing—original draft [equal], Writing—review & editing [equal])

Conflicts of interest

None declared.

Funding

Funding by BMBF projects Ebira, SNRT and ScaDS.AI is acknowledged.

References

Abramson J, Adler J, Dunger J *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* 2024;**630**:493–500.

Andreeva A, Kulesha E, Gough J, *et al.* The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures. *Nucleic Acids Res* 2020; **48**: D376-D382.

Benner SA, Cohen MA, Gonnet GH. Amino acid substitution during functionally constrained divergent evolution of protein sequences. *Protein Eng* 1994;**7**:1323–32.

Berman HM, Westbrook J, Feng Z *et al.* The protein data bank. *Nucleic Acids Res* 2000;**28**:235–42.

Blackshields G, Wallace IM, Larkin M *et al.* Analysis and comparison of benchmarks for multiple sequence alignment. *In Silico Biol* 2006;**6**:321–39.

Blum M, Andreeva A, Florentino LC, *et al.* InterPro: the protein sequence classification resource in 2025. *Nucleic Acids Res* 2025; **53**: D444-D456.

Edgar RC. Optimizing substitution matrix choice and gap parameters for sequence alignment. *BMC Bioinformatics* 2009; **10**:396.

Edgar RC. Protein structure alignment by reseek improves sensitivity to remote homologs. *Bioinformatics* 2024;**40**:btae687.

El-Hendi F, Al-Fatlawi A, Hossen MB *et al.* Structural alphabets approach performance of structural alignment in remote homology detection. bioRxiv, 2026.

Elnaggar A, Heinzinger M, Dallago C *et al.* ProtTrans: toward understanding the language of life through self-supervised learning. *IEEE Trans Pattern Anal Mach Intell* 2022;**44**:7112–27.

Grantham R. Amino acid difference formula to help explain protein evolution. *Science* 1974;**185**:862–4.

Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* 1992;**89**:10915–9.

Hess M, Keul F, Goesele M *et al.* Addressing inaccuracies in BLOSUM computation improves homology search performance. *BMC Bioinformatics* 2016;**17**:189.

Illergard K, Ardell DH, Elofsson A. Structure is three to ten times more conserved than sequence—a study of structural response in protein cores. *Proteins* 2009;**77**:499–508.

Jia K, Jernigan RL. New amino acid substitution matrix brings sequence alignments into agreement with structure matches. *Proteins* 2021;**89**:671–82.

Jumper J, Evans R, Pritzel A *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;**596**:583–9.

Keul F, Hess M, Goesele M *et al.* PFASUM: a substitution matrix from Pfam structural alignments. *BMC Bioinformatics* 2017; **18**:293.

Long H, Li M, Fu H. Determination of optimal parameters of MAFFT program based on BAliBASE3.0 database. *Springerplus* 2016;**5**:736.

Mistry J, Chuguransky S, Williams L, *et al.* Pfam: the protein families database in 2021. *Nucleic Acids Res* 2021; **49**: D412-D419.

Müller T, Spang R, Vingron M. Estimating amino acid substitution models: a comparison of Dayhoff's estimator, the resolvent approach and a maximum likelihood method. *Mol Biol Evol* 2002;**19**:8–13.

Müller T, Vingron M. Modeling amino acid replacement. *J Comput Biol* 2000;**7**:761–76.

Pantolini L, Studer G, Pereira J *et al.* Embedding-based alignment: combining protein language models with dynamic programming alignment to detect structural similarities in the twilight-zone. *Bioinformatics* 2024;**40**:btad786.

Pearl FM, Bennett CF, Bray JE *et al.* The CATH database: an extended protein family resource for structural and functional genomics. *Nucleic Acids Res* 2003;**31**:452–5.

Price GA, Crooks GE, Green RE *et al.* Statistical evaluation of pairwise protein sequence comparison with the Bayesian bootstrap. *Bioinformatics* 2005;**21**:3824–31.

Prlic A, Domingues FS, Sippl MJ. Structure-derived substitution matrices for alignment of distantly related sequences. *Protein Eng* 2000;**13**:545–50.

Rost B. Twilight zone of protein sequence alignments. *Protein Eng* 1999;**12**:85–94.

Sgarbossa D, Lupo U, Bitbol AF. Generative power of a protein language model trained on multiple sequence alignments. *eLife* 2023; **12**: e79854.

Styczynski MP, Jensen KL, Rigoutsos I *et al.* BLOSUM62 miscalculations improve search performance. *Nat Biotechnol* 2008; **26**:274–5.

Thompson JD, Linard B, Lecompte O, *et al.* A comprehensive benchmark study of multiple sequence alignment methods: current challenges and future perspectives. *PLoS One* 2011; **6**:e18093.

Varadi M, Anyango S, Deshpande M, *et al.* AlphaFold protein structure database: massively expanding the structural coverage of protein- sequence space with high-accuracy models. *Nucleic Acids Res* 2022; **50**: D439-D444.

Varadi M, Bertoni D, Magana P, *et al.* AlphaFold protein structure database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Res* 2024;**52**:D368-D375.