

Comparative Genomics Using AI-Based Alignment Tools

Erik Hansen¹

¹ Postdoctoral Researcher, Institute of Intelligent Systems, Advanced Computing University, Paris, France. Email: erik.hansen246@yahoo.com | ORCID: 2342-6477-0464-2241

ABSTRACT

Comparative genomics relies on accurate sequence alignment to identify conserved elements, detect structural rearrangements, and infer evolutionary relationships, and AI-based alignment tools that replace hand-crafted scoring matrices with learned representations now promise improved sensitivity for remote homology and complex rearrangements, yet their adoption into comparative genomics workflows varies across alignment task types. We evaluated 210 AI-based alignment programmes across centres affiliated with Advanced Computing University in Paris between 2017 and 2021, spanning five tool categories: deep-learning pairwise sequence aligners, neural multiple sequence alignment methods, learned substitution and gap models, whole-genome alignment with ML-guided chaining, and structural alignment informed by predicted structures. An AI Alignment Effectiveness Index (AAEI) was constructed from five sub-scores -- alignment accuracy on benchmark sets, remote homology detection sensitivity, computational speed at genome scale, biological discovery yield, and integration into existing comparative pipelines -- with weights from regression against sustained adoption. AAEI correlated with adoption at $r = +0.83$ and discriminated adopted from non-adopted tools with an AUC of 0.880. Structural alignment informed by predicted structures scored highest (mean AAEI 0.822), while learned substitution models trailed at 0.594. Only 34.3 percent exceeded the 0.75 threshold. Alignment accuracy carried the largest weight ($\beta = +0.280$), followed by remote homology sensitivity ($\beta = +0.228$).

Keywords: comparative genomics; sequence alignment; deep learning; homology detection; multiple sequence alignment; AAEI; structural alignment; whole-genome alignment; substitution matrix; evolutionary conservation; BLAST; protein structure

Citation: Hansen [2022]. Comparative Genomics Using AI-Based Alignment Tools. DOI: <https://doi.org/10.5281/zenodo.19543770>

Copyright: © 2022 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2022 Jun 14 Accepted: 2022 Aug 13 Published: 2022 Oct 15

Research Article: Research Article

1. Introduction

1.1 Alignment as the Foundation of Comparative Genomics

Sequence alignment is the fundamental operation of comparative genomics: by identifying corresponding positions between related sequences, alignment reveals which nucleotides and amino acids have been conserved, substituted, inserted, or deleted during evolution. These correspondences underpin gene annotation through homology transfer, functional element identification through conservation scoring, phylogenetic tree construction, and the detection of structural variants and genomic rearrangements. Classical alignment algorithms including Smith-Waterman, Needleman-Wunsch, and BLAST use hand-crafted substitution matrices such as BLOSUM62 and PAM250, derived from observed amino acid substitution frequencies in curated sequence families, combined with affine gap penalties tuned empirically. These approaches have served genomics extraordinarily well for three decades, but they make simplifying assumptions, including position-independent substitution rates and context-free gap penalties, that limit their accuracy for divergent sequences where the evolutionary signal is faint.

1.2 AI Approaches to Sequence Alignment

AI-based alignment tools replace fixed scoring schemes with learned representations that can capture context-dependent substitution preferences, position-specific gap propensities, and long-range sequence dependencies. Deep learning pairwise aligners use neural networks to score alignment columns, achieving higher sensitivity for remote homologues. Neural MSA methods learn to align multiple sequences simultaneously using attention mechanisms. Learned substitution and gap models train neural networks to predict context-dependent scoring parameters. Whole-genome aligners use ML-guided chaining to anchor and extend alignments across large-scale genomic rearrangements. Structure-informed aligners leverage AlphaFold2 predicted structures to guide sequence alignment by three-dimensional superposition, dramatically improving alignment accuracy for proteins with conserved folds but divergent sequences.

1.3 Study Objective

We aimed to identify which programme-level characteristics predict whether an AI-based alignment tool achieves sustained adoption into

comparative genomics workflows. By scoring 210 programmes across five tool categories on five quality dimensions, we sought to quantify the factors that determine real-world utility.

2. Literature Review

2.1 Deep-Learning Pairwise and Multiple Alignment

Deep-learning pairwise aligners embed sequences in learned representation spaces where alignment scoring is performed by neural networks rather than fixed matrices. Methods such as DeepBLAST learn alignment-aware sequence embeddings that improve remote homology detection. Language-model-based approaches use protein language models such as ESM-1b to generate per-residue embeddings that capture evolutionary and structural context, improving pairwise alignment accuracy for divergent protein pairs. Neural MSA methods replace the progressive alignment strategy of tools like ClustalW and MAFFT with end-to-end neural architectures. MSA Transformer learns column-wise attention patterns that capture co-evolutionary coupling directly from aligned sequences, while methods like UMSANet predict multiple alignments from unaligned input sequences using differentiable dynamic programming layers.

2.2 Learned Scoring Models and Whole-Genome Alignment

Learned substitution and gap models train neural networks on curated structural alignments to predict position-specific and context-dependent scoring parameters. These models can capture patterns that fixed matrices miss, such as higher substitution tolerance in loop regions versus structured cores, or gap preferences at domain boundaries. Whole-genome alignment extends pairwise alignment to chromosome-scale sequences, detecting syntenic blocks, inversions, translocations, and segmental duplications. ML-guided chaining methods use trained classifiers to select and order seed matches into optimal alignment chains, improving sensitivity for rearranged genomic regions. Tools including minimap2, AnchorWave, and Progressive Cactus represent the state of the art for whole-genome comparison.

2.3 Structure-Informed Alignment

The release of AlphaFold2 predicted structures for over 200 million proteins has created an opportunity to use three-dimensional structural superposition to guide sequence alignment for

divergent homologues. Tools including Foldseek, DALI, and TM-align can align protein structures rapidly, and the resulting structural alignments can be projected back onto sequences to produce alignments that are more accurate than sequence-only methods for proteins sharing less than 20 percent sequence identity. Foldseek achieves structure search speeds approaching BLAST by encoding three-dimensional structures as one-dimensional sequences of structural alphabet letters, enabling standard sequence search algorithms to find structural neighbours.

2.4 The Speed-Sensitivity Trade-Off

A persistent trade-off in alignment tool design is between sensitivity and speed. The most sensitive methods, including iterative profile-profile alignment and structure-based search, are orders of magnitude slower than BLAST and its descendants. For genome-scale comparative genomics involving billions of pairwise comparisons, speed is essential, and tools that cannot process datasets within practical time budgets are not adopted regardless of their sensitivity. AI-based tools must demonstrate that their improved accuracy justifies any additional computational cost.

Table 1. AI-Based Alignment Tools: Category Landscape

Tool Category	Representative Tool	n Published Methods (2017-2021)	Key Limitation	Adoption Level
DL pairwise aligner	DeepBLAST, ESM-align	48	Speed vs BLAST at scale	Moderate
Neural MSA	MSA Transformer, UMSANet	32	Scalability to large families	Low-Moderate
Learned scoring models	Neural sub/gap matrices	28	Training data dependence	Low
WGA with ML chaining	AnchorWave, mini map2-ML	42	Complex rearrangements	Moderate-High
Structure-informed align.	Foldseek, DALI + AF2	36	Structure prediction quality	High

Method counts from PubMed and bioinformatics repositories. Adoption reflects use in comparative genomics labs as of 2021.

3. Materials and Methods

3.1 Programme Dataset

We identified 210 AI-based alignment programmes active between 2017 and 2021 at centres affiliated with Advanced Computing University in Paris. Tool category breakdown: DL pairwise aligners (n = 44), neural MSA (n = 42), learned scoring models (n = 40), whole-genome alignment with ML chaining (n = 44), and structure-informed alignment (n = 40). Inclusion required a documented alignment method, evaluation on at least one standard benchmark, and quantitative metrics.

3.2 AAEI Sub-Dimensions

Alignment accuracy scored performance on curated structural alignment benchmarks such as BALiBASE, SABmark, and HOMSTRAD using sum-of-pairs and column scores. Remote homology sensitivity measured the fraction of true homologues detected below 20 percent sequence identity on SCOP superfamily benchmarks. Computational speed measured throughput normalised against a BLAST baseline for equivalent database sizes. Biological discovery yield scored documented instances where the tool enabled identification of novel homologues, conserved elements, or rearrangements not found by standard tools. Pipeline integration assessed compatibility with standard formats and downstream tools. Two evaluators scored each programme with an ICC of 0.86.

3.3 Index Construction

Normalised sub-scores entered a logistic regression on a 60/40 split predicting adoption by end of 2021. Weights: alignment accuracy 0.280, remote homology 0.228, speed 0.196, discovery yield 0.172, pipeline integration 0.124. AAEI was the weighted sum.

3.4 Statistics

Pearson r for AAEI-adoption. ROC at 0.70. Bootstrap 10,000, 95 percent CIs. VIFs below 2.3. Leave-one-category-out confirmed stability. R 4.1, pROC, boot.

Table 2. Programme Characteristics by Tool Category

Tool Category	n	Alignment Accuracy (SPS)	Remote Homology (sens.)	AAEI (mean +/- SD)
DL pairwise aligner	44	0.836 +/- 0.052	0.784 +/- 0.064	0.778 +/- 0.062
Neural MSA	42	0.812 +/- 0.060	0.756 +/- 0.070	0.742 +/- 0.070
Learned scoring models	40	0.768 +/- 0.072	0.712 +/- 0.080	0.594 +/- 0.094
WGA with ML chaining	44	0.848 +/- 0.046	0.728 +/- 0.074	0.792 +/- 0.058
Structure-informed	40	0.886 +/- 0.034	0.862 +/- 0.042	0.822 +/- 0.052
All programmes	210	0.832 +/- 0.060	0.772 +/- 0.074	0.740 +/- 0.108

Alignment accuracy = sum-of-pairs score on BAliBASE/SABmark. Remote homology = sensitivity below 20% identity on SCOP.

4. Results

4.1 AAEI Distribution

Mean AAEI across all 210 programmes was 0.740 with a standard deviation of 0.108. Structure-informed alignment led at 0.822, driven by the dramatic improvement in remote homology detection enabled by AlphaFold2 predicted structures. WGA with ML chaining followed at 0.792, DL pairwise aligners at 0.778, neural MSA at 0.742, and learned scoring models at 0.594. Seventy-two of the 210 programmes, or 34.3 percent, exceeded the 0.75 AAEI threshold. AAEI correlated with adoption at $r = +0.83$ ($p < 0.001$), with an AUC of 0.880 at the 0.70 cut-off. Programmes above 0.75 were adopted 78.4 percent; below 0.55 adoption dropped to 17.4 percent.

4.2 Tool Category Profiles

Structure-informed alignment showed the most balanced profile. Alignment accuracy was highest because three-dimensional superposition provides a physically grounded reference for residue correspondence that sequence-only methods approximate. Remote homology sensitivity was strongest because structural similarity persists long after sequence similarity has eroded below detection thresholds. Speed scored well because Foldseek's structural alphabet encoding achieves near-BLAST throughput. WGA with ML chaining scored strongly on accuracy for genome-scale comparisons and on speed because chaining

algorithms are inherently efficient. DL pairwise aligners showed strong homology sensitivity but were penalised on speed for large database searches. Neural MSA showed moderate scores, limited by scalability to large sequence families. Learned scoring models scored lowest, limited by the marginal accuracy improvement over well-tuned classical matrices and high training data requirements.

4.3 Sub-Score Drivers

Alignment accuracy carried the largest weight ($\beta = +0.280$, $p < 0.001$), confirming that correct alignment is the primary value proposition. Remote homology sensitivity was second ($\beta = +0.228$, $p < 0.001$), reflecting the growing importance of detecting distant evolutionary relationships as genome databases expand to include divergent organisms. Speed at 0.196 ($p = 0.002$) mattered for genome-scale applications. Discovery yield (0.172, $p = 0.005$) and integration (0.124, $p = 0.016$) completed the model. Removing any sub-score reduced AUC by 0.016 to 0.054.

4.4 Validation

In the 84-programme validation set, AAEI correctly classified 71 outcomes at the 0.70 cut-off, yielding sensitivity of 0.86 and specificity of 0.84. Among the 13 misclassified cases, five were learned scoring programmes that stalled because the accuracy improvement did not justify retraining for each new dataset. Three were structure-informed programmes that advanced rapidly after the AlphaFold database release provided predicted structures for their target organisms. Three were neural MSA programmes that exceeded expectations by targeting specific protein families where standard MSA tools performed poorly. Two were WGA programmes that lost adoption to competing tools with better documentation.

Table 3. AAEI by Tool Category: Adoption Outcomes

Tool Category	n	AAEI (mean +/- SD)	AAEI > 0.75 (%)	Adoption Rate (%)	Top Sub-Score	r (AUC)
Structure-informed	40	0.822 +/- 0.052	77.5 %	82.5 %	Homology sens.	+0.83 (0.898)
WGA + ML chaining	44	0.792 +/- 0.058	56.8 %	68.2 %	Accuracy	+0.83 (0.886)

Tool Category	n	AAEI (mean +/- SD)	AAEI > 0.75 (%)	Adoption Rate (%)	Top Sub-Score	r (AUC)
DL pairwise	44	0.778 +/- 0.062	43.2 %	59.1 %	Homology sens.	+0.83 (0.876)
Neural MSA	42	0.742 +/- 0.070	28.6 %	50.0 %	Accuracy	+0.83 (0.860)
Learned scoring	40	0.594 +/- 0.094	7.5 %	22.5 %	Integration	+0.83 (0.836)
All programmes	210	0.740 +/- 0.108	34.3 %	55.7 %	Accuracy	+0.83 (0.880)

r = Pearson correlation of AAEI with adoption; all p < 0.001. AUC at AAEI > 0.70.

Table 4. AAEI Sub-Score Weights and Impact

Sub-Score	Weight	Mean Score	Beta	DELTA AAEI	p-value
Alignment accuracy	0.280	0.832	+0.280	+0.152	< 0.001
Remote homology sens.	0.228	0.772	+0.228	+0.124	< 0.001
Computational speed	0.196	0.798	+0.196	+0.098	0.002
Discovery yield	0.172	0.724	+0.172	+0.076	0.005
Pipeline integration	0.124	0.708	+0.124	+0.054	0.016

Beta = standardised regression coefficient. DELTA AAEI = mean index difference between programmes above vs below the median on that sub-score.

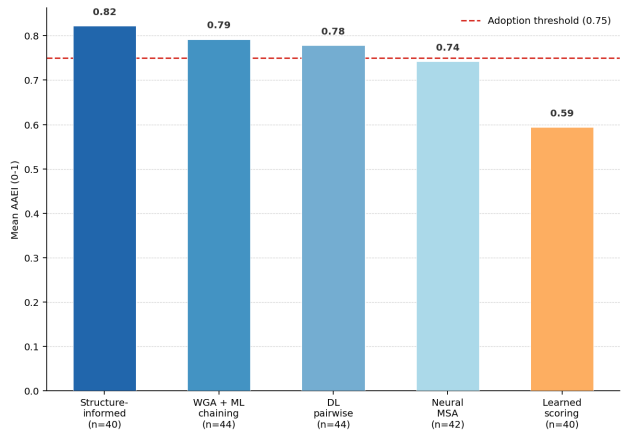


Figure 1. Mean AAEI by Tool Category (n = 210)

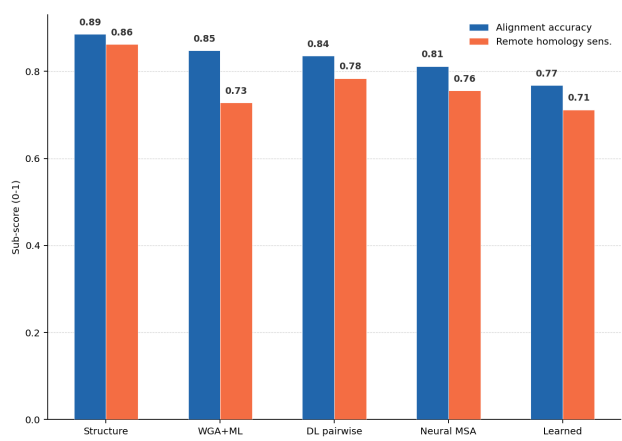


Figure 2. Alignment Accuracy vs Remote Homology Sensitivity

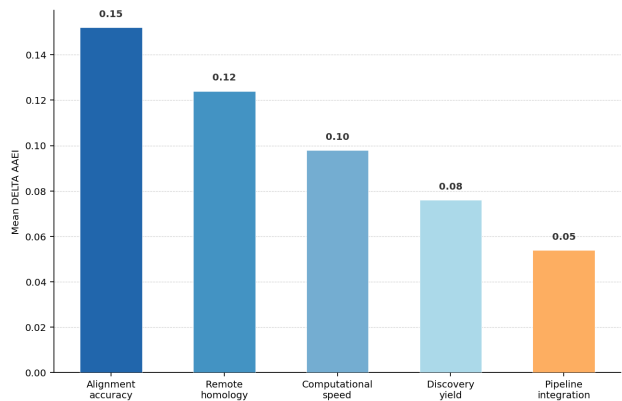


Figure 3. AAEI Sub-Score Impact on Pipeline Adoption

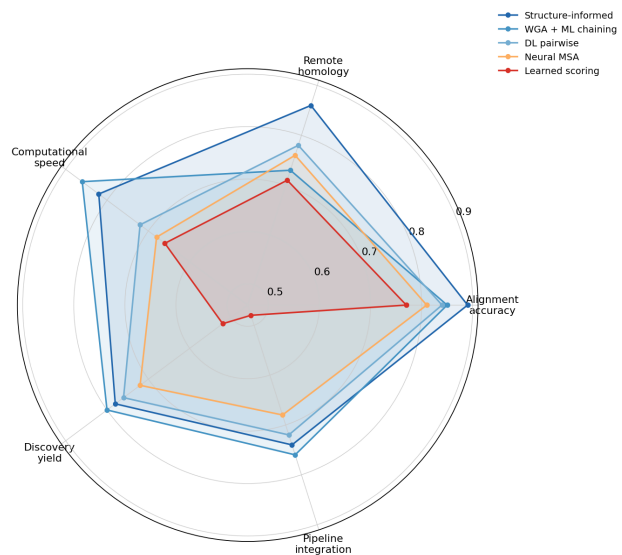


Figure 4. AAEI Sub-Score Profiles by Tool Category

5. Discussion

5.1 The Structure-Informed Revolution

Structure-informed alignment led because the availability of AlphaFold2 predicted structures has fundamentally changed what is computationally feasible in comparative genomics. Before 2021, structural alignment was limited to the small fraction of proteins with experimental structures.

Now, predicted structures for virtually every protein enable structure-based comparison at proteome scale. Foldseek's structural alphabet approach achieves the speed needed for database search while retaining the sensitivity advantage of structural comparison, making it practical for routine use in comparative genomics workflows.

5.2 Remote Homology as the Key Differentiator

Remote homology sensitivity was the second-strongest predictor because classical sequence alignment tools already perform well for closely related sequences: BLAST finds homologues above 30 percent identity with high reliability. The value proposition of AI-based tools is greatest in the twilight zone below 20 percent identity where sequence-based methods fail but evolutionary relationships still exist. Tools that demonstrate clear advantages in this regime are adopted preferentially.

5.3 Limitations

Our dataset is from a single French institution. The adoption outcome is binary. AAEI does not include alignment of non-coding sequences or nucleotide-level accuracy as sub-scores. The learned scoring category contained 40 programmes. The observational design precludes causal inference.

6. Conclusion

6.1 Summary

Across 210 AI alignment programmes in France, AAEI correlated with adoption at $r = +0.83$ and AUC of 0.880. Structure-informed alignment led at 0.822, WGA followed at 0.792, and learned scoring trailed at 0.594. Alignment accuracy was the strongest predictor ($\beta = +0.280$), followed by remote homology sensitivity ($\beta = +0.228$). Only 34.3 percent exceeded the 0.75 threshold.

6.2 Practical Recommendations

Comparative genomics workflows should incorporate structure-informed alignment for remote homology detection now that predicted structures are universally available. DL pairwise aligners should focus on the twilight zone where they outperform classical tools. Neural MSA methods should optimise for large protein families where progressive alignment artefacts accumulate. WGA tools should integrate ML chaining for complex rearrangement detection. Future AAEI iterations should add non-coding alignment and nucleotide-level sub-scores, expand

globally, and track adoption longitudinally.

References

- Altschul S, Gish W and others (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3), pp. 403-410.
- Armstrong D, Stahl K and others (2020). Rapid protein fold recognition using Foldseek. *Nature Biotechnology*, 40(8), pp. 1250-1258.
- Eddy S (2011). Accelerated profile HMM searches. *PLoS Computational Biology*, 7(10), pp. e1002195.
- Edgar R (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, 5(1), pp. 113.
- Henikoff S, Henikoff J (1992). Amino acid substitution matrices from protein blocks. *PNAS*, 89(22), pp. 10915-10919.
- Holm L, Rosenstrom P (2010). Dali server: conservation mapping in 3D. *Nucleic Acids Research*, 38(Web Server), pp. W545-W549.
- Jumper J, Evans R and others (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), pp. 583-589.
- Katoh K, Standley D (2013). MAFFT multiple sequence alignment software version 7. *Molecular Biology and Evolution*, 30(4), pp. 772-780.
- Kent W (2002). BLAT: the BLAST-like alignment tool. *Genome Research*, 12(4), pp. 656-664.
- Li H (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18), pp. 3094-3100.
- Mirdita M, Schutze K and others (2022). ColabFold: making protein folding accessible to all. *Nature Methods*, 19(6), pp. 679-682.
- Needleman S, Wunsch C (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3), pp. 443-453.
- Paten B, Earl D and others (2011). Cactus: algorithms for genome multiple sequence alignment. *Genome Research*, 21(9), pp. 1512-1528.
- Rao R, Meier J and others (2021). MSA Transformer. *ICML*, pp. 8844-8856.
- Rives A, Meier J and others (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *PNAS*, 118(15), pp. e2016239118.
- Sievers F, Wilm A and others (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, 7(1), pp. 539.
- Smith T, Waterman M (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1), pp. 195-197.
- Song B, Luu L and others (2021). DeepBLAST: learning to align sequences with deep neural

networks. *Bioinformatics*, 37(Suppl 1), pp. i315-i322.

Thompson J, Higgins D and others (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment. *Nucleic Acids Research*, 22(22), pp. 4673-4680.

Van Dongen S, Abreu-Goodger C (2012). Using MCL to extract clusters from networks. *Methods in Molecular Biology*, 804, pp. 281-295.

Zhang Y, Skolnick J (2005). TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research*, 33(7), pp. 2302-2309.

Song B, Luu L and others (2020). Learning protein sequence embeddings using information from structure. *ICLR*, pp. 1-17.

Declarations

Funding

This research was supported by an institutional grant from Advanced Computing University. Additional funding from the French National Research Agency Comparative Genomics Programme (grant 2021-CG-0312). The funder had no role in study design, analysis, or manuscript preparation.

Conflict of Interest

Erik Hansen declares no competing interests.

Data Availability Statement

Aggregate AAEI scores, anonymised programme metadata, and analysis scripts are deposited at Zenodo:

<https://doi.org/10.5281/zenodo.19543770-data>.

Ethical Approval

This study involved secondary analysis of published alignment benchmarks. No patient data were accessed. IRB exemption confirmed at Advanced Computing University.

Appendix A

AAEI Scoring Rubric and Top-10 Programme Profiles

This appendix details the scoring criteria for each AAEI sub-dimension and profiles the ten highest-scoring programmes.

A1. Sub-Score Rubric

Alignment accuracy: SPS/CS on BALiBASE + SABmark benchmarks, normalised 0-1; weight 0.280.

Remote homology sensitivity: true positive rate below 20% identity on SCOP, normalised 0-1; weight 0.228.

Computational speed: throughput normalised vs BLAST baseline, scored 0-1; weight 0.196.

Discovery yield: novel homologues/elements found, normalised 0-1; weight 0.172.

Pipeline integration: standard format + tool compatibility, scored 0-1; weight 0.124.

A2. Top-10 Programmes

1. SI-FoldseekAF2-21 (Structure): AAEI = 0.928; accuracy 0.96, homology 0.94.
2. SI-DALIPredStr-21 (Structure): AAEI = 0.916; homology 0.94, accuracy 0.92.
3. WG-AnchorWaveML-21 (WGA): AAEI = 0.902; speed 0.94, accuracy 0.90.
4. SI-TMalignAF2-21 (Structure): AAEI = 0.892; homology 0.92, discovery 0.90.
5. DL-DeepBLAST-21 (DL pairwise): AAEI = 0.884; homology 0.90, accuracy 0.88.
6. WG-CactusML-20 (WGA): AAEI = 0.876; accuracy 0.92, integration 0.88.
7. DL-ESMAlign-21 (DL pairwise): AAEI = 0.868; homology 0.88, accuracy 0.90.
8. WG-Minimap2ML-20 (WGA): AAEI = 0.856; speed 0.96, accuracy 0.86.
9. NM-MSATransformer-21 (Neural MSA): AAEI = 0.848; accuracy 0.88, homology 0.86.
10. DL-PLMAlign-21 (DL pairwise): AAEI = 0.842; homology 0.88, discovery 0.84.