

AMP Challenge – Aggregation Score

Paulina Szymczak* Rasmus Møller-Larsen* Bruno Puczek-Szymański*
Pankhil Gawade Adam Izdebski Mathieu Seyfrid
Marcelo Der Torossian Torres Cesar de la Fuente-Nunez Ewa Szczurek

Abstract

This document describes the *in silico* metric, the aggregation score, which is used to score the 50,000 antimicrobial peptide (AMP) sequences (the “library”) submitted by each team to the AMP Challenge. Additionally, it validates the metric using control datasets and finally ranks the submissions. More in-depth documentation will be provided in the follow-up paper, including the code.

1 Overview

The objective of the aggregation score is to provide a single *in silico* metric value to compare the capabilities of generative models on the task of AMP generation. The aggregation score is a unitless value between 0 and 100 (higher is better). It is computed on 10,000 sequences across five non-overlapping folds, and the metric value is the mean across these five folds. The aggregation score considers three categories:

- *realism*: Are these sequences likely to be synthesisable?
- *activity*: Are these sequences likely to have antimicrobial properties?
- *diversity*: Do the sequences cover a diverse set of peptides?

The three categories are weighted equally.

2 Definition

In this section, the aggregation score will be defined. Recall that the goal of the aggregation score is to measure three categories – realism, activity and diversity – in a single value. These three categories are made up of several components, each of them a number in $[0, 1]$. Table 1 summarises the components. Throughout, S denotes the library being scored.

	Component	Description	Weight
<i>realism</i>	Synthesizable	Fraction of sequences that can be synthesised	1/9
	Perplexity (Hyformer)	How plausible the sequences look to a peptide language model	1/9
	Alignment Score	Similarity of the sequences to known AMPs	1/9
<i>activity</i>	Hit-rate (AMPFormer ensemble)	Fraction of sequences with MIC below $32\mu\text{M}$	1/6
	p_{amp} (OmegAMP)	Mean AMP probability under an independent classifier	1/6
<i>diversity</i>	Coverage (active)	Distinct clusters among the actives, per sequence generated	1/6
	FKEA (ESM2)	Effective number of distinct modes in ESM2’s embedding space	1/6

Table 1: The seven components of the aggregation score and their weights.

* Equal contribution.

Normalisation The components have different magnitudes and spread, so they are mapped onto a common $[0, 1]$ scale.

The activity and diversity components are normalised with a relative scheme (*rel*) against a fixed reference set R : the 1,714 experimentally confirmed broad-spectrum potent AMPs of **Broad potent**. Each statistic is computed in its natural units by the same function, applied to both S and R . Let v be its value on S and r its value on R ; the relative component is the ratio v/r squashed into $[0, 1]$:

$$\text{rel}(v) = \frac{v/r}{v/r + 1} \in [0, 1].$$

A library that matches the reference scores exactly 0.5, one twice as good scores 2/3, and one that scores nothing gets 0. The map is strictly increasing and bounded above, so the reference is a landmark rather than a ceiling: a library that outperforms natural AMPs scores above it, without the component running away. The three realism components (**Synthesizable**, **Perplexity** and **Alignment Score**) are already fractions in $[0, 1]$ by construction, and are used directly.

2.1 Realism

Synthesizable A rule-based filter flags peptides that are hard to make in the lab: it requires hydrophilic residue abundance between 30 and 70%, at most two consecutive glycines, at most one run of bulky residues, no aggregation-prone motif, a proline rate of at most 20%, at most one polymerising cysteine, and net charge in $[1.9, 10]$. The component is the fraction of S that passes. These synthesizability rules have been developed by Marcelo Der Torossian Torres based on wet-lab experience and have been used successfully as criteria for pre-screening peptide sequences.

Perplexity (Hyformer) Perplexity measures how plausible a sequence looks to a peptide language model. Sequences are scored under `hyformer_peptides_34M` [3]. Per-sequence perplexity is clipped at 20, and the component is how far the clipped average sits below that clip:

$$1 - \frac{1}{20|S|} \sum_{s \in S} \min(\text{ppl}(s), 20).$$

It detects damage that the other realism components miss: substituting a single residue in a known AMP leaves synthesizability untouched but is penalised here.

Alignment score MMseqs2 [11] aligns each sequence against **Positives (all)**, a set of known AMPs (Section 4.1), and the best hit’s bitscore is normalised by that query’s self-alignment bitscore. The component is the mean of those normalised bitscores over the whole library, $\text{Alignment}(S)$. It is the only component that is sensitive to residue *order*: it separates real potent AMPs from the same peptides with their residues shuffled, which every other component rates alike.

2.2 Activity

Hit-rate (AMPFormer ensemble) MIC is predicted by the AMPFormer ensemble, which scores a peptide against 11 clinically relevant bacterial strains, seven Gram-negative and four Gram-positive. The ensemble has four members: random forests on pepCATS cross-correlation descriptors [7], on ESM2 650M embeddings [4] and on Hyformer `peptides_34M` embeddings [3], together with APEX’s own 40-model deep ensemble [13]. Members are combined on the pMIC scale, $-\log_{10}$ of the molar MIC, as mean -0.5 std, so peptides the members disagree about are scored conservatively rather than optimistically. A sequence counts as *active* when its geometric mean MIC across the eleven strains is at most $32 \mu\text{M}$:

$$\text{active}(s) = [\text{MIC}_{\text{geo}}(s) \leq 32 \mu\text{M}], \quad \text{MIC}_{\text{geo}}(s) = \left(\prod_{t=1}^{11} \text{MIC}_t(s) \right)^{1/11}.$$

Write $A(S) \subseteq S$ for the sequences that pass the gate, with $n_A = |A(S)|$. The component is the hit-rate, normalised against the reference: $\text{rel}(n_A/|S|)$.

p_{amp} (**OmegAMP**) OmegAMP [10] is an AMP classifier trained independently of the MIC surrogates. The component is its mean predicted AMP probability over S , normalised against the reference: $\text{rel}(p_{\text{OmegAMP}}(S))$.

2.3 Diversity

Coverage CD-HIT [2] clusters the active sequences $A(S)$ of Section 2.2 at 60% identity. A cluster count estimated from few actives is noisy, so it is scaled by a reliability factor $\rho(n_A) = n_A / (n_A + m)$ with $m = 20$, which is 0.5 at $n_A = 20$ and approaches 1 for large n_A . Writing $K(\cdot)$ for the cluster count, the component is

$$\text{rel}(K(A(S)) \cdot \rho(n_A) / |S|).$$

The denominator is the number of sequences *generated* ($|S|$, not n_A), so a library cannot score well by producing few but diverse actives. The reliability factor matters because the gate is strict: only a small fraction of a typical library is admitted, and some control sets admit no actives at all. Without it, a library with a single active sequence would score maximal coverage of 1 cluster per 1 sequence.

FKEA (ESM2) FKEA [8] is the Rényi kernel entropy ($\alpha = 2$) of the library in embedding space, i.e. the effective number of distinct modes it occupies. A fixed, seeded draw of 1,500 sequences is embedded with stock `esm2.t30_150M_UR50D` [4], with the kernel bandwidth set to the median pairwise distance between reference embeddings. The component is that entropy normalised against the reference, $\text{rel}(\text{FKEA}(S))$.

2.4 Aggregation

Given the individual components of the aggregation score, they are combined into a single value as follows:

$$\text{score}(S) = 100 \cdot \sum_i w_i v_i, \quad \sum_i w_i = 1, \quad v_i \in [0, 1].$$

The weights follow from a single rule: the three categories contribute equally, and the components within a category contribute equally. This gives 1/9 to each of the three realism components and 1/6 to each of the remaining four.

3 Scoring protocol

The submitted libraries contain 50,000 sequences, while the aggregation score is computed on 10,000 sequences at a time. Each library is therefore permuted once (seed 0) and split into five disjoint 10,000-sequence blocks, which are scored independently. The reported score is the mean over the blocks and the reported error is their standard deviation. The blocks are disjoint rather than resampled, since overlapping draws understate the spread. Scoring is implemented with the `seqme` library [6], which supplies the blocking as well as the embedding and entropy metrics used by the components.

4 Validation

In this section, we validate the components of the aggregation score. To evaluate the utility of the score, we defined several control datasets containing peptides with known properties. The control data were subsets of peptides from databases, such as DBAASP [9], dbAMP [14], DRAMP [5] and UniProt [1], which contain experimental measurements of peptide potency against bacteria. Additionally, we included synthetic data, which could easily fool a more naive metric. The success criterion for the score was its ability to order the control data in a series reflecting their fitness as AMPs. Additionally, the score was evaluated by its discrimination strength, i.e., absolute difference in magnitude between AMP and non-AMP control data. Control sets smaller than two blocks are scored as a whole and carry no error bar. Results are shown below.

4.1 Control datasets

Potent sets Five category sets were extracted from DBAASP [9], each covering peptides of 8–50 standard residues: **Broad potent** (1,714 peptides with median MIC $\leq 16 \mu\text{M}$ across at least four species spanning both Gram classes), **Gram+ potent** (2,647) and **Gram- potent** (3,078), potent against at least one Gram-positive or Gram-negative strain respectively, **MDR potent** (626), potent against at least one multidrug-resistant strain, and **Selectivity potent** (296), potent overall and with a safety window $\text{HC}_{50}/\text{MIC}_{50} \geq 10$. **Positives (HQ)** is a curated set of 3,322 DBAASP [9] peptides restricted to canonical residues and to assays meeting EUCAST and CLSI conditions, counted active at MIC $\leq 32 \mu\text{g/mL}$ against at least one strain. **Positives (all)** pools 14,941 experimentally validated AMPs from dbAMP [14], AMP Scanner [12] and DRAMP [5], and **Veltri Positives** is the AMP Scanner benchmark set.

Weak sets Each category set has a weak counterpart drawn the same way but requiring MIC $\geq 64 \mu\text{M}$, the assay ceiling: **Broad weak** (660), **Gram+ weak** (948), **Gram- weak** (977) and **MDR weak** (347). The 16–64 μM grey zone is dropped, so potent and weak are cleanly separated. **Selectivity weak** (40) is the exception: those peptides are potent but hemolytic, with a safety window ≤ 1 . These sets are real AMPs that happen to be inactive, so they test whether the score reads potency rather than AMP-likeness.

Negative sets **Negatives (all)** holds 14,814 UniProt [1] sequences annotated as cytoplasmic and not as antimicrobial, antibiotic, antiviral, antifungal, effector or excreted, clustered at 40% identity with CD-HIT to remove redundancy. **Negatives (HQ)** is a harder set of 925 peptides that are classified as AMPs but experimentally confirmed inactive, at MIC $\geq 128 \mu\text{g/mL}$ against every strain tested. **Veltri Negatives** is the AMP Scanner [12] benchmark negative set.

Synthetic controls Three families of decoy probe specific failure modes. The mutated sets – **Broad potent (mutated)**, **Broad weak (mutated)** and **Gram- potent (mutated)** – substitute a single residue at a random position in each sequence of the corresponding real set, so they test whether the score detects small amounts of damage. **Shuffled** permutes the residues of the reference peptides, preserving composition and length exactly while destroying residue order. **Random (standard)** and **Random (uniform)** sample sequences at natural and at uniform amino-acid frequencies, and **UniProt** is an unfiltered background sample.

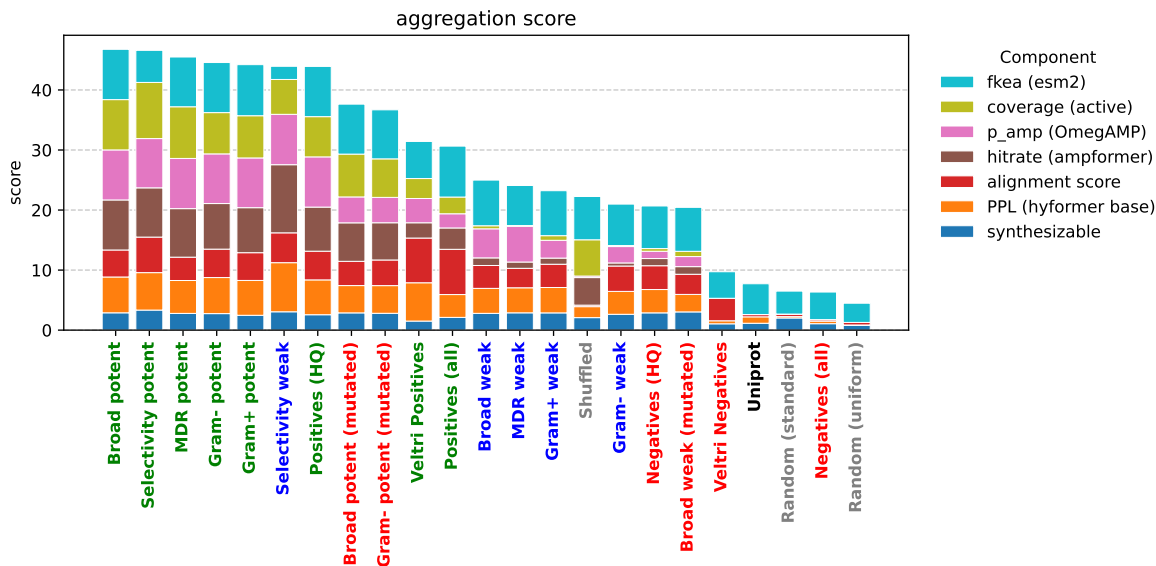


Figure 1: Aggregation score on the control sets, ordered by score and split into the weighted contribution of each component. Set labels are coloured by group: AMPs (green), weakly active AMPs (blue), non-AMPs (red), random sequences (grey) and background (black).

5 Eligibility

Every submission was reviewed by the organisers before scoring. Of the 61 submissions received, two supplied no 50,000-sequence library and could not be scored at all. The remaining 59 were scored, and 44 of those were found eligible for ranking.

The 15 scored submissions that were excluded fall into four groups. Seven made an insufficient generative contribution: an unchanged baseline model, combined with filtering, ranking or selection, which the organisers judged not to be a distinct generative contribution of the entrant’s own. Three could not be inspected, because the repository containing the model weights and inference code was inaccessible at review or was never supplied. Three violated the sequence requirements: top-100 membership in the submitted library, the novelty threshold against known antibacterials, or the minimum length of 8 residues. Two submissions had code issues: the repository contained no implementation of the generation method the submission described, so the organisers could not establish that the submitted peptides had come from the model claimed.

A further four eligible submissions came from the organising teams. They are scored on the same footing as the rest but are not ranked, which leaves 40 ranked submissions.

6 Submission results

Each submitted library was scored with the protocol of Section 3: five disjoint 10,000-sequence blocks, reported as their mean and standard deviation. Table 2 gives the 40 ranked submissions and Table 3 the four from the organising teams.

#	Team	Score	#	Team	Score
1	86886 IMI	38.86 ± 0.22	21	86992 Aishwarya967	22.11 ± 0.43
2	86974 Evgeniy Nam	36.61 ± 0.34	22	prib-lab PRIB LAB	21.37 ± 0.20
3	86831 Rishyanth Reddy	32.19 ± 0.32	23	86916 BlueWhale	21.16 ± 0.45
4	86838 Samuel Paik-Heintz	30.02 ± 0.22	24	86961 Laksh Goel	18.77 ± 0.49
5	86166 Grimmlab Team	28.76 ± 0.32	25	85630 M_K_	18.59 ± 0.49
6	86937 iMakAI	27.19 ± 0.11	26	84680 hitwhdata	17.49 ± 0.26
7	86987 Reeshabh Kotecha	27.05 ± 0.26	27	85040 ziyi Lu1	17.34 ± 0.22
8	86975 H&D: Justin Time	27.01 ± 0.27	28	86899 vnahata	16.86 ± 0.38
9	85557 Les Go	26.87 ± 0.17	29	86788 Dmitrii Magas	16.45 ± 0.52
10	86813 Qiushi Engine Team	26.81 ± 0.20	30	86683 Leontenkov Egor	15.93 ± 0.24
11	86685 Mohsen DLALA	25.67 ± 0.61	31	85381 Sim jaemin	15.83 ± 0.25
12	86825 Lux Bonteri	24.56 ± 0.28	32	86850 Team AMPsequill	15.07 ± 0.16
13	86985 j_v_v_07	24.46 ± 0.39	33	82841 Lucrezia Polloni	14.02 ± 0.22
14	84970 littlestronomer	24.11 ± 0.37	34	franco-s-lab Franco's Lab	11.22 ± 0.22
15	84912 Vadim Borisov	23.21 ± 0.20	35	86113 this-team	10.68 ± 0.31
16	86946 BeeMax	23.11 ± 0.39	36	86061 Druggable	10.28 ± 0.16
17	86158 AMP It Up	23.03 ± 0.58	37	86956 TeamPhage	9.72 ± 0.13
18	86981 Dino	23.02 ± 0.46	38	86979 Scott Jarmusch	9.37 ± 0.21
19	86768 Alireza Noroozi	22.66 ± 0.48	39	84719 DeShPet	6.36 ± 0.08
20	86844 Braxton Mensah	22.54 ± 0.52	40	86959 nicetone9	5.80 ± 0.06

Table 2: The 40 ranked submissions, as mean $\pm$ standard deviation over the five blocks.

Team	Score
86168 Piotr Wilczyński	42.08 ± 0.21
86980 Grzegorz Siudak	35.05 ± 0.23
86834 Ola Gwiazda	33.94 ± 0.09
apexmo ApexMO	26.87 ± 0.33

Table 3: Submissions from the organising teams. These are scored with the same protocol but are not ranked.

References

- [1] The UniProt Consortium. Uniprot: the universal protein knowledgebase in 2025. *Nucleic Acids Research*, 53(D1):D609–D617, 01 2025.
- [2] Limin Fu, Beifang Niu, Zhengwei Zhu, Sitao Wu, and Weizhong Li. Cd-hit: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23):3150–3152, 12 2012.
- [3] Adam Izdebski, Jan Olszewski, Pankhil Gawade, Krzysztof Koras, Serra Korkmaz, Valentin Rauscher, Jakub M. Tomczak, and Ewa Szczurek. Synergistic benefits of joint molecule generation and property prediction, 2026.
- [4] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*, 2022.
- [5] Tianyue Ma, Yanchao Liu, Bingxin Yu, Xin Sun, Huiyuan Yao, Chen Hao, Jianhui Li, Maryam Nawaz, Xun Jiang, Xingzhen Lao, and Heng Zheng. Dramp 4.0: an open-access data repository dedicated to the clinical translation of antimicrobial peptides. *Nucleic Acids Research*, 53(D1):D403–D410, 01 2025.
- [6] Rasmus Møller-Larsen, Adam Izdebski, Jan Olszewski, Pankhil Gawade, Michal Kmiciekiewicz, Wojciech Zarzecki, and Ewa Szczurek. seqme: a python library for evaluating biological sequence design, 2025.

- [7] Alex T Müller, Gisela Gabernet, Jan A Hiss, and Gisbert Schneider. modlamp: Python for antimicrobial peptides. *Bioinformatics*, 33(17):2753–2755, 09 2017.
- [8] Azim Ospanov, Jingwei Zhang, Mohammad Jalali, Xuenan Cao, Andrej Bogdanov, and Farzan Farnia. Towards a scalable reference-free evaluation of generative models, 2024.
- [9] Malak Pirtskhalava, Anthony A Amstrong, Maia Grigolava, Mindia Chubinidze, Evgenia Alimbarashvili, Boris Vishnepolsky, Andrei Gabrielian, Alex Rosenthal, Darrell E Hurt, and Michael Tartakovsky. Dbaasp v3: database of antimicrobial/cytotoxic activity and structure of peptides as a resource for development of new therapeutics. *Nucleic Acids Research*, 49(D1):D288–D297, 01 2021.
- [10] Diogo Soares, Leon Hetzel, Paulina Szymczak, Marcelo Der Torossian Torres, Johanna Sommer, Cesar de la Fuente-Nunez, Fabian Theis, Stephan Günemann, and Ewa Szczurek. Omegamp: Targeted amp discovery via biologically informed generation, 2026.
- [11] Martin Steinegger and Johannes Söding. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35(11):1026–1028, Nov 2017.
- [12] Daniel Veltri, Uday Kamath, and Amarda Shehu. Deep learning improves antimicrobial peptide recognition. *Bioinformatics*, 34(16):2740–2747, 08 2018.
- [13] Fangping Wan, Marcelo D. T. Torres, Jacqueline Peng, and Cesar de la Fuente-Nunez. Molecular de-extinction of antibiotics enabled by deep learning. *bioRxiv*, 2023.
- [14] Lantian Yao, Jiahui Guan, Peilin Xie, Chia-Ru Chung, Zhihao Zhao, Danhong Dong, Yilin Guo, Wenyang Zhang, Junyang Deng, Yuxuan Pang, Yulan Liu, Yunlu Peng, Jorng-Tzong Horng, Ying-Chih Chiang, and Tzong-Yi Lee. dbamp 3.0: updated resource of antimicrobial activity and structural annotation of peptides in the post-pandemic era. *Nucleic Acids Research*, 53(D1):D364–D376, 01 2025.