Adaptive Evolutionary Optimization for the Discovery of Self-Assembling Peptides

Milica Injac[1] , Ena Dražić[2], Marko Njirjak[2], Gordana Rakić[3], Daniela Kalafatovic[2], Goran Mauša[2]

1 - University of Novi Sad, Faculty of Technical Science, Serbia, 2 - University of Rijeka, Faculty of Engineering, Croatia, 3 - University of Novi Sad, Faculty of Science, Serbia

https://doi.org/10.17952/38EPS.2026.P267

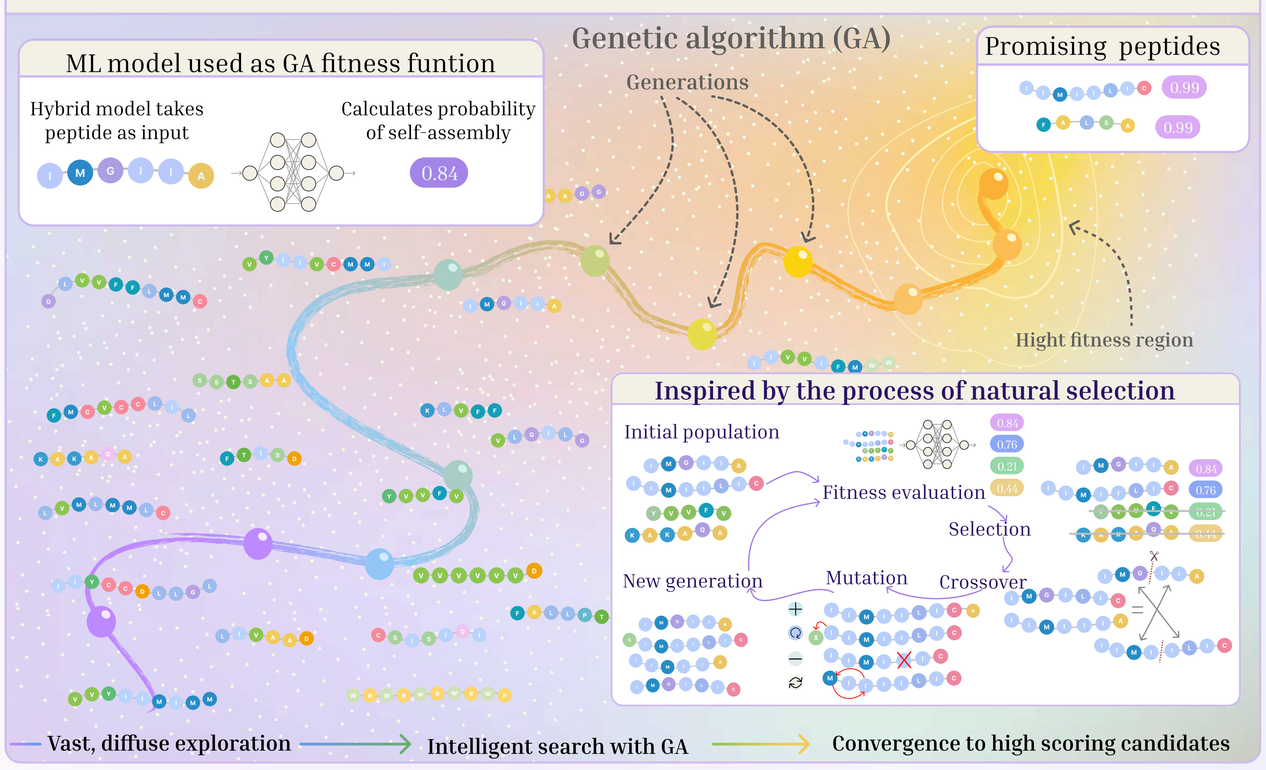
milica.injac@uns.ac.rs

RQ1 Does ML-driven GA find unique promising self-assembling peptides (high quality candidates) while exploring the global space?

RQ2 What does the GA reveal about the ML self-assembly predictor?

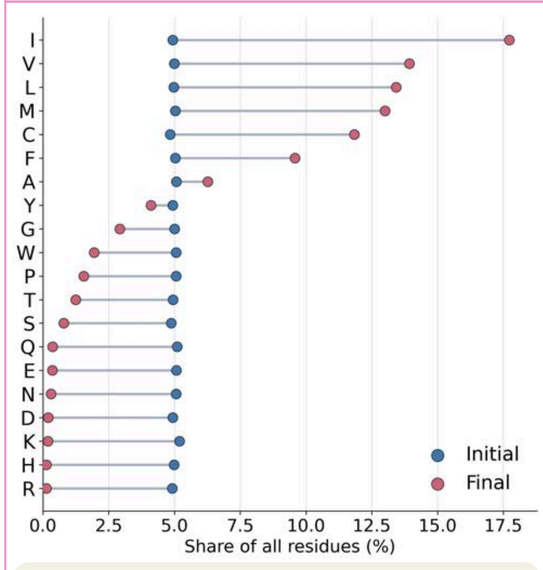
RQ3 How do GA-generated peptides compare to peptides from the experimental dataset?

Exploration of Large Peptide Space Guided by GA Driven ML



Analysis of GA-Generated Peptides Collected Over 200 Independent GA Runs

GA selection reshaped aa composition



Selection enriched I/V/L/M/C/F and strongly depleted charged residues, revealing the compositional region favored by generative model. This computational preference does not establish that individual residues cause self-assembly.

What do mutations tell us about the predictive model?

Delete Removes aa | Substitute Replaces aa | Insert Adds aa | Swap Exchanges aa

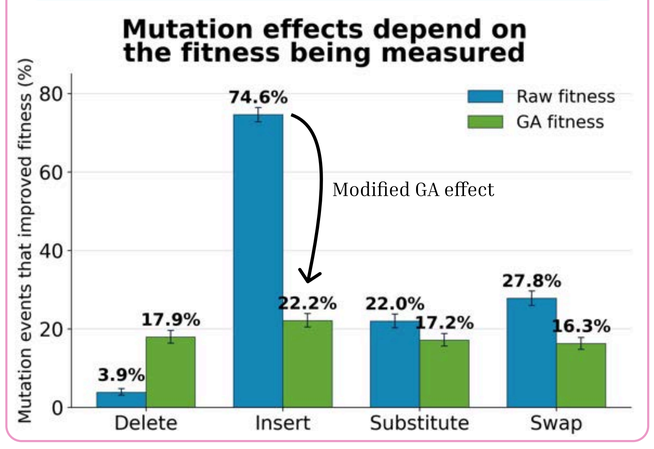
8,761 mutation events

89,6% of them happened on high-scoring peptides

GA used mutation mainly to probe local variants within the predictor's high-score region.

Length-associated preference

Adding one residue usually **increased** the predicted score, whereas **deletion** rarely did.



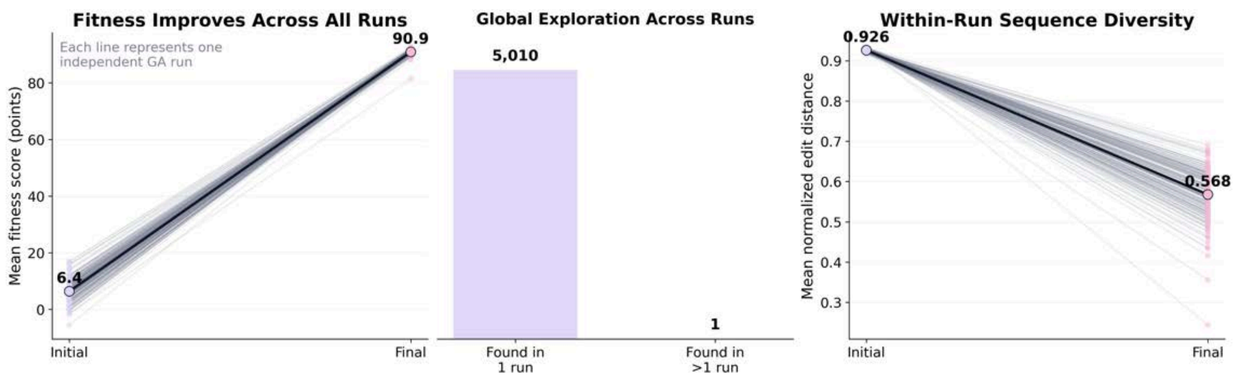
Global Exploration and Local Exploitation

200 independent GA runs | GA guided by Retrained prediction model | **5011** suggested peptides where **5010** are unique sequences

Within each run, the GA focused the population on a smaller set of high-fitness peptides.

Different runs produced different results: 99.9% of final sequences were not found in any other run

Within a single run, peptides converge to a place and are more similar than starting population.



RQ1 Thousands unique, high scoring candidates; runs converged locally, but explored different sequence neighborhoods. This reflects **local exploitation within each run**, while the lack of overlap between runs indicates **global exploration** of different sequence solutions. Median fitness in final populations is 90.9 which means while exploring the global space **solutions are high scoring**.

Experimental setup

4 experimental conditions x200 GA runs

Larger and more balanced dataset

[1] dataset: 256 SA, 112 NSA + [2] dataset: 91 SA, 117 NSA = Combined dataset: 334 SA, 276 NSA

368 → 610 peptides total

Retrained model accuracy: 77.1%

[1] Mauša et al., Reshaping the discovery of self-assembling peptides with generative AI guided by hybrid deep learning. Nat Mach Intell. 6. (2024). 10.1038/s42256-024-00928-1.
[2] Batra et al., Discovery of unconventional and nonintuitive self-assembling peptide materials using experiment-driven machine learning. Sci. Adv. 11. (2025). 10.1126/sciadv.adt9466

RQ2 GA search exposed predictor preferences for hydrophobicity, charge, and peptide length

Strategies to dampen length bias

	Original GA	Modified GA
Original ML model	5251	5016
Retrained ML model	3870	4270

number of peptides

Modifying GA with:
→ Fitness penalty favoring length diversity,
→ Initial populations with shorter peptides

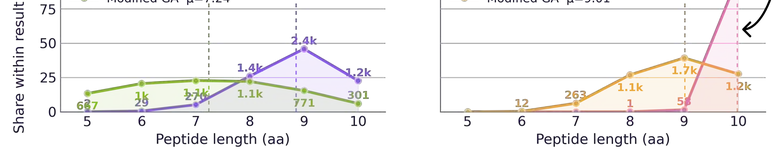
Modification of initial population strategy

Original GA: 100% peptides with 3-24 aa (3 aa – 24 aa)

Modified GA: 90% peptides with 3-6 aa (3 aa – 24 aa)

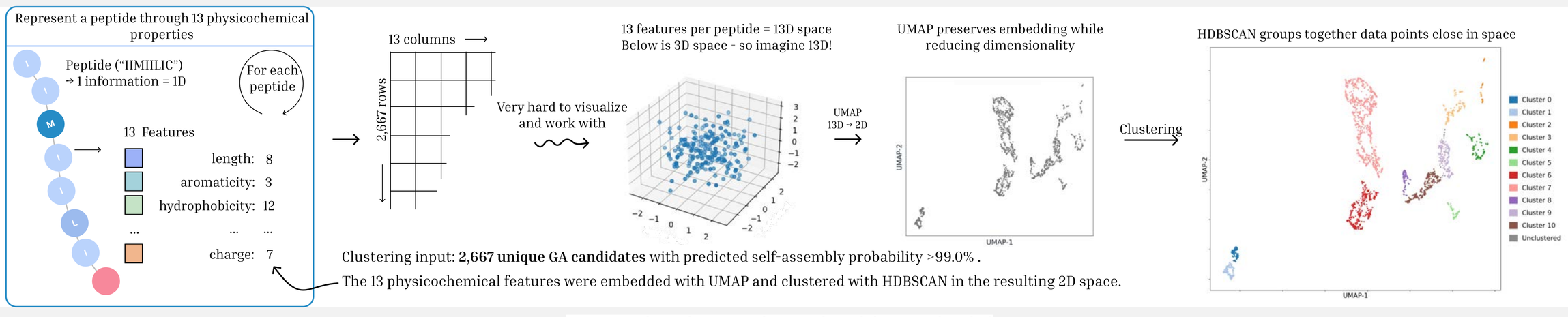
Peptide length distributions

98.5% peptides length 10



Clustering GA-Generated Peptides: Where Do Experimental Peptides Map?

What is clustering? → grouping similar data points based on defined measure of similarity. For example: HDBSCAN →Groups together data points that are tightly packed, leaving out points in sparse areas as outliers.



Clustering input: **2,667 unique GA candidates** with predicted self-assembly probability >99.0% .

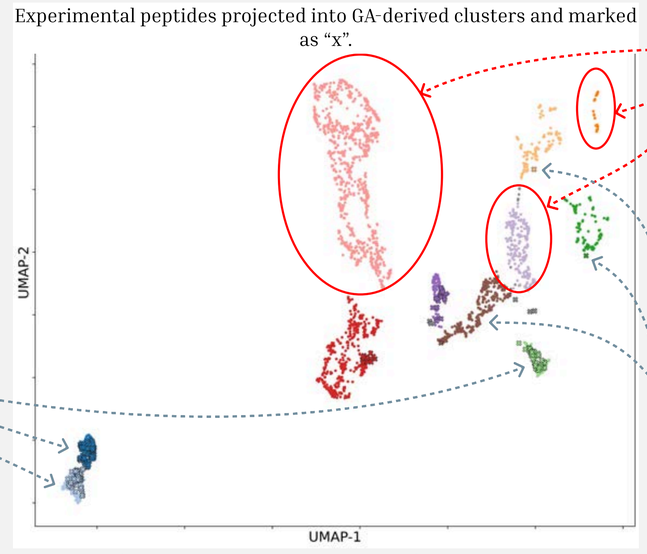
The 13 physicochemical features were embedded with UMAP and clustered with HDBSCAN in the resulting 2D space.

3.9% peptides from the dataset map to unclustered regions

0 GA candidates match sequences from the dataset

0.887 median HDBSCAN membership strength

Experimental peptides projected into GA-derived clusters and marked as "x".



3 GA-derived clusters contain no peptides from the experimental dataset

0 experimental peptides map to the largest cluster, which contains 32.4% of GA candidates

84.7% SA peptides from the dataset concentrate in those 3 clusters

3 GA clusters

13.4% generated SA peptides occupy those clusters

84.7% peptides from the dataset maps to 3 small clusters, in which 13.4% GA generated peptides is located

3 clusters contain only one experimental peptide each

RQ3 Despite the limited search space coverage by the dataset, GA explores new peptide-property regions

2,667 candidates · 11 clusters · 0.5% unclustered · largest family 32.4% · bootstrap stability: ARI 0.76 ± 0.16

TAKE HOME MESSAGE: ML-guided GA explored diverse search space neighbourhoods and identified thousands of high scoring self assembling peptide candidates, including regions not covered by the experimental dataset. The search also revealed model preferences for hydrophobicity, charge, and peptide length providing candidates for experimental testing.

Acknowledgements: This work was carried out within COST Action CA23111 – SNOOPY "Searching for Nanostructured or pOre fOrming Peptides for therapY", supported by COST (European Cooperation in Science and Technology). In addition, the authors were supported by European Union – NextGenerationEU [PepNanoDisco, grant no. uniri-mzi-25-2], Croatian Science Foundation [Super DeShPet 2.0, grant no. IP-2025-02-3357], Ministry of Science, Technological Development and Innovation (Contract No. 451-03-34/2026-03/200156) and the Faculty of Technical Sciences, University of Novi Sad through project "Scientific and Artistic Research Work of Researchers in Teaching and Associate Positions at the Faculty of Technical Sciences, University of Novi Sad 2026" (No. 01-3609/1).