Active Learning-Driven Surrogate Modeling for the Generative Discovery of Self-Assembling Peptides

Denis Ibiši[1] · Daniela Kalafatović[1] · Goran Mauša[1]
[1]University of Rijeka, Faculty of Engineering, Rijeka, Croatia
https://doi.org/10.17952/38EPS.2026.P264
e-mail: denis.ibisi123@gmail.com

ACKNOWLEDGEMENTS

SNOOPY COST Action – CA23111
SCI-PHI – uniri-mzi-25-25
Super DeShPet 2.0 – IP-2025-02-3357

Funded by the European Union
NextGenerationEU

Croatian Science Foundation

Motivation

Self-assembling peptides are versatile building blocks that spontaneously organise through non-covalent interactions into functional nanostructures such as fibrils, micelles and hydrogels. Their assembly depends strongly on amino acid sequence and experimental conditions, while the number of possible sequences grows exponentially with peptide length.

Conventional discovery relies on experimental screening, molecular simulations and expert-designed rules, which are costly, time-consuming, or may restrict exploration to familiar sequence motifs. Machine learning (ML) can broaden the search space and reduce screening costs by prioritising promising and non-intuitive peptide candidates for simulation and experimental validation.

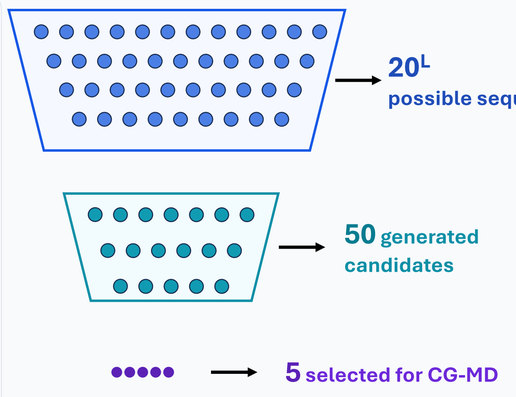
An ML surrogate model can rapidly estimate self-assembly probability, while a generative algorithm explores sequence space to propose new peptide candidates [1].

[1] Njirjak et al., Nature Machine Intelligence, 2024, 6, 1487–1500

GOAL:

Design and compare active learning / Bayesian optimisation acquisition functions for efficiently prioritising generated peptide candidates under limited labelled data and evaluation budgets

Search space challenge



AP-SP model performance

0.8686 ± 0.0253 F_1 (mean ± SD)	**0.8431 – 0.9072** F_1 (fold range)
0.9356 PR-AUC	**0.8673** ROC-AUC
0.1481 Brier score	**0.1232** ECE-10

Designed acquisition functions

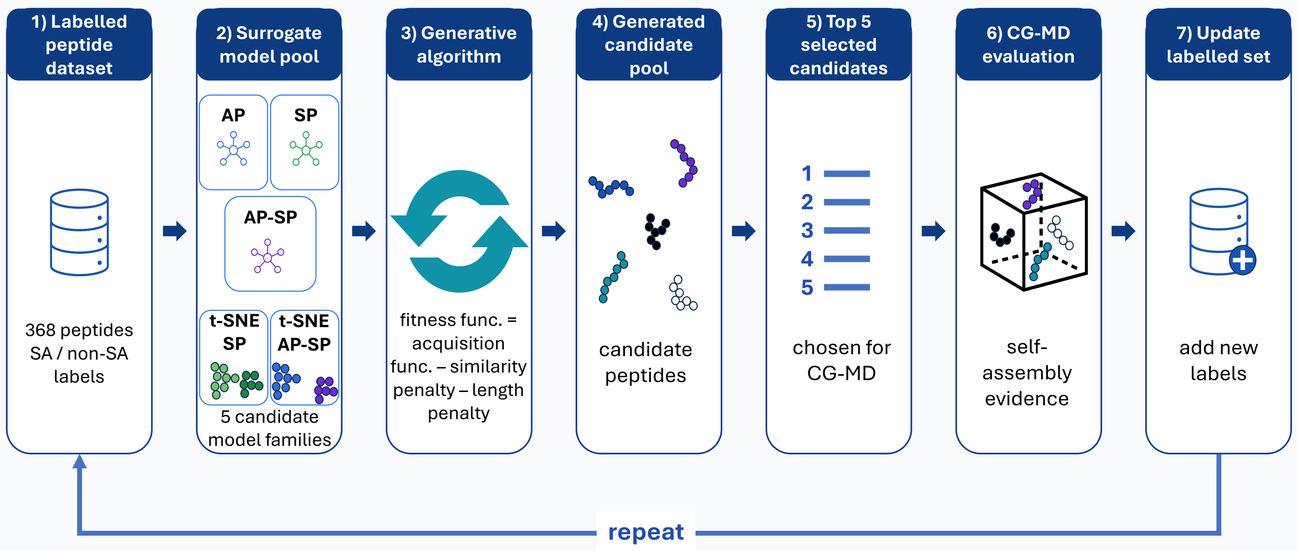
Active learning

Function	Principle(s)	Scoring method
Predictive entropy	Informativeness	Uncertainty of the mean prediction: $H(\bar{p}(x))$.
Ensemble mutual information (MI)	Informativeness	Epistemic uncertainty: $H(\bar{p}(x)) - \frac{1}{M}\sum_{m=1}^{M} H(p_m(x))$.
Family QBC	Informativeness	Vote entropy across 5 model families.
Cluster-diverse representative	Representativeness Diversity	Centroid-nearest representatives from clusters prioritised by novelty to labels.
OED logdet	Representativeness Diversity	Greedy log-determinant information gain.
Hybrid MI-diverse	Informativeness Diversity	MI selection followed by farthest-first embedding selection.
Static easy entropy	Informativeness Representativeness	Entropy ranking within the 30% most familiar candidates.
Self-paced entropy	Informativeness Representativeness	Entropy ranking within a familiar pool expanding from 30% to 100%.

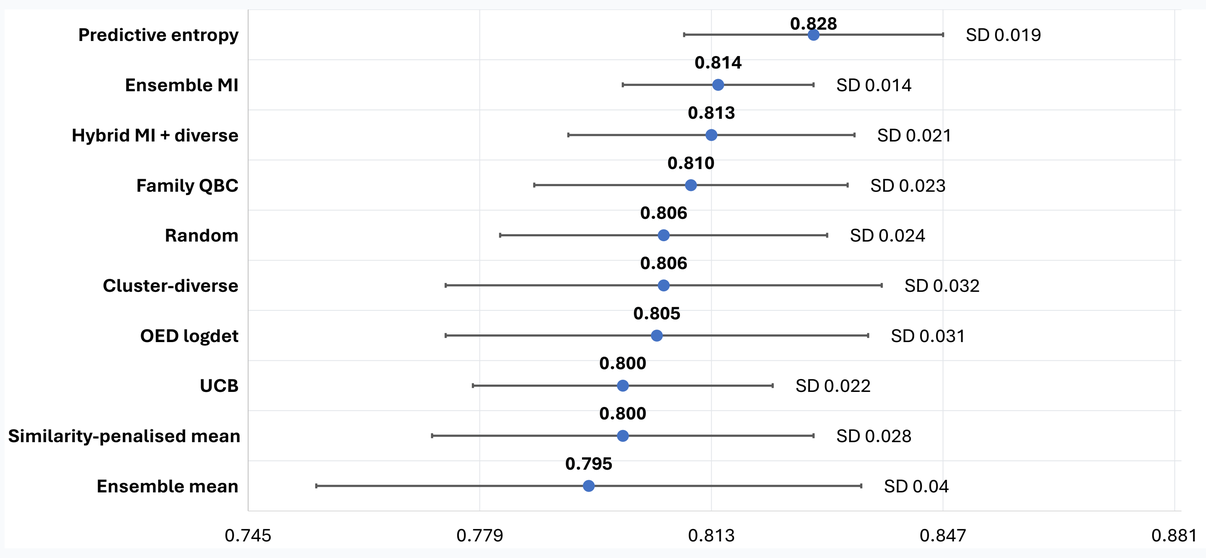
Bayesian optimisation

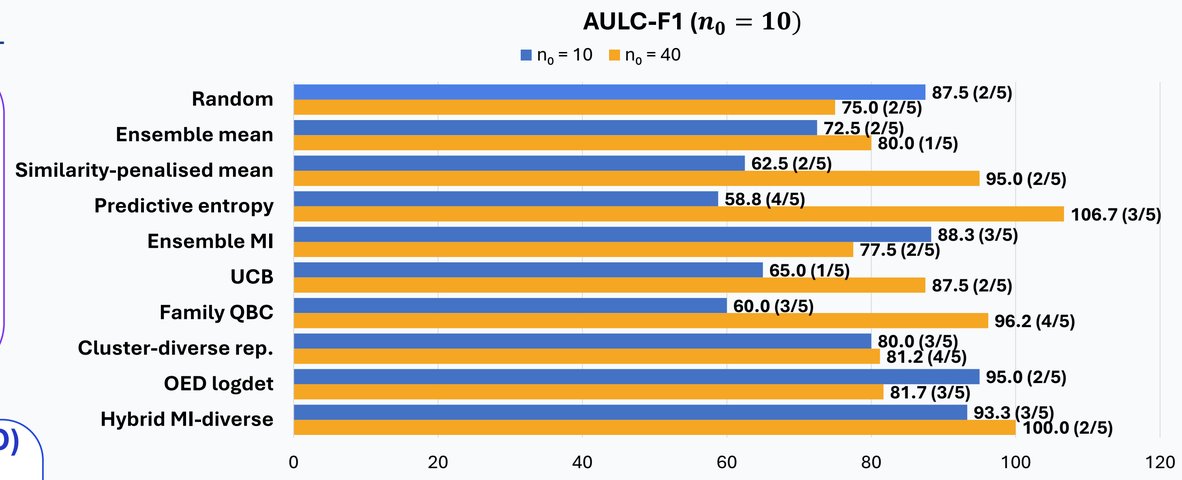
Function	Principle(s)	Scoring method
Ensemble mean / Greedy	Exploitation	Rank by mean ensemble SA probability $\bar{p}(x)$.
Similarity-penalised mean	Exploitation Exploration	$Score = \bar{p}(x) - composition\ similarity\ penalty$.
Upper Confidence Bound (UCB)	Exploitation Exploration	$Score = \bar{p}(x) + \sigma_p(x)$.
Probability of Improvement (PI)	Exploitation	Probability of improvement over incumbent f^*.
Expected Improvement (EI)	Exploitation Exploration	Expected improvement over incumbent f^*.
Maximum Entropy Search (MES)	Exploitation Exploration	Expected information gain about the optimum.

Project workflow

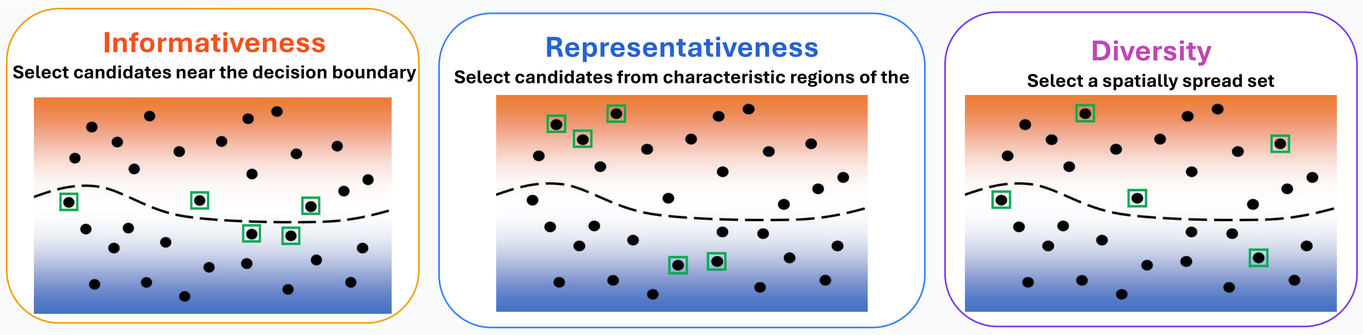


Results – AULC, Labels to F1 ≥ 0.86, Diversity gain





Candidate-selection principles: Active Learning and Bayesian Optimisation



Active learning (AL)

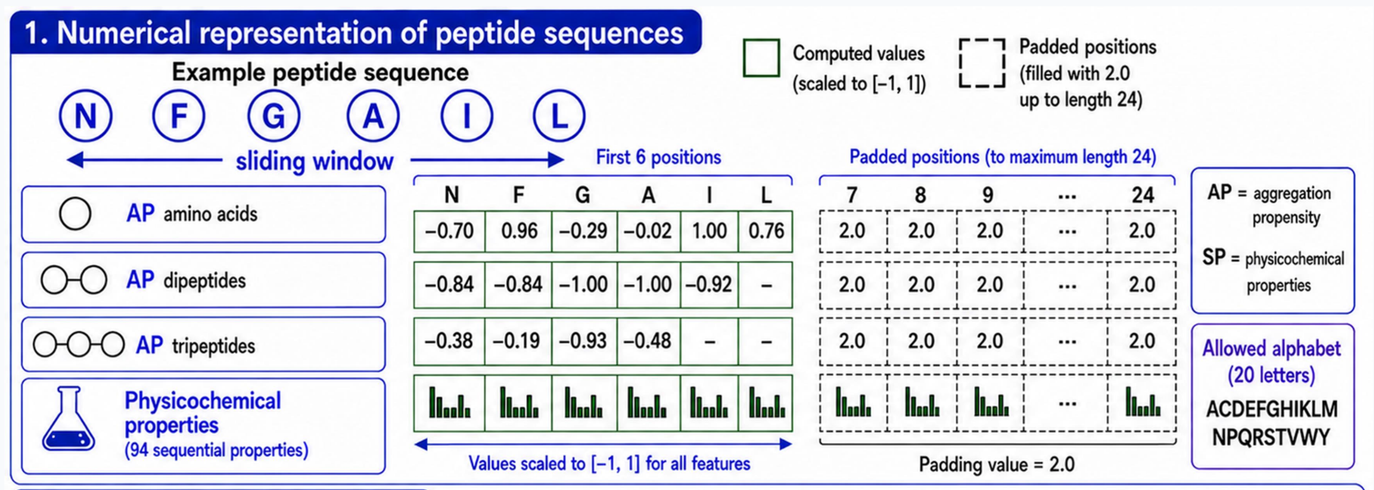
- Improve the predictive model under a limited labelling budget.
- Use informativeness, representativeness, or diversity to guide candidate selection.

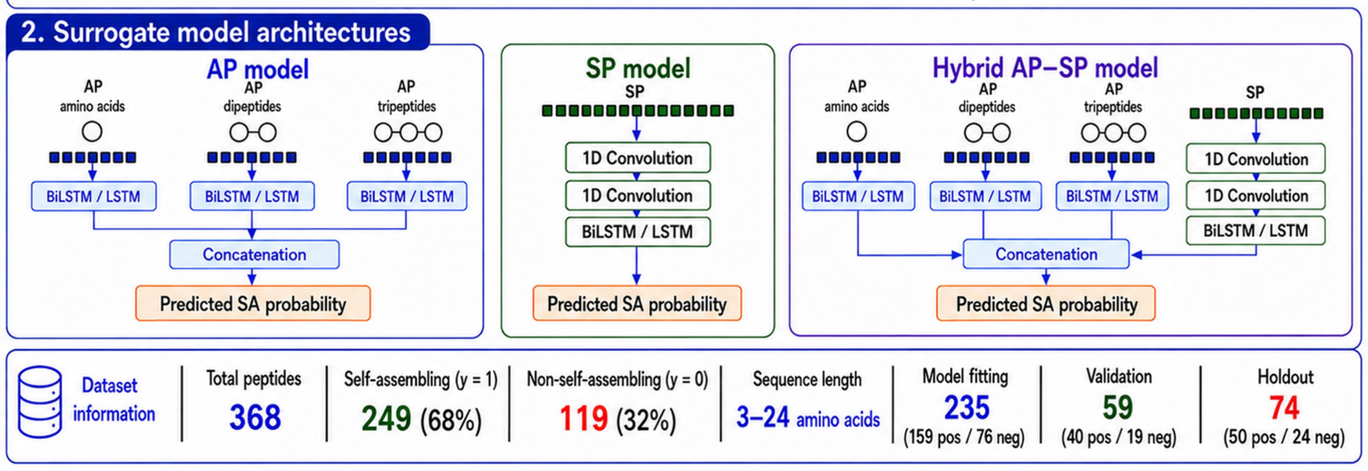
Informativeness ⟷ Exploitation
Representativeness ⟷ Exploration
Diversity ⟷ Exploration

Bayesian optimisation (BO)

- Prioritise promising candidates for expensive evaluation.
- Balance exploitation and exploration to find the best solution.

Peptide representation / Surrogate model architectures / Dataset





Dataset information	Total peptides	Self-assembling (y = 1)	Non-self-assembling (y = 0)	Sequence length	Model fitting	Validation	Holdout
	368	**249** (68%)	**119** (32%)	**3–24** amino acids	**235** (159 pos / 76 neg)	**59** (40 pos / 19 neg)	**74** (50 pos / 24 neg)

Self-paced active-learning replay

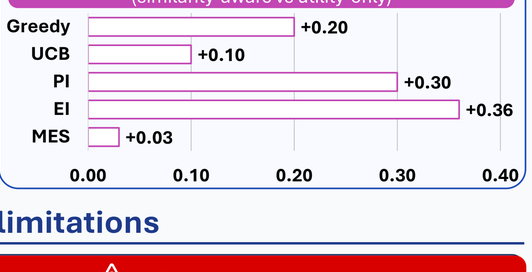
Strategy	AULC (10-235)*	Labels to F1 ≥ 0.86 **
Predictive entropy	**0.8223**	105.0 (3/3)
Random	0.8171	112.5 (2/3)
Self-paced entropy	0.8147	**71.7 (3/3)**
Static easy entropy	0.8016	102.5 (2/3)

* Normalised area under the F1 learning curve (10-235 labels); higher is better.
** Conditional mean labels revealed to first reach F1 ≥ 0.86; parentheses show folds reaching the target.

Fixed-surrogate proposal policies

Strategy	Positive	Yield	Diversity	Mean length
Random	3/5	0.600	**0.860**	**5.0 aa**
Greedy	2/5	0.400	0.480	**10.0 aa**
UCB	1/5	0.200	0.610	**10.0 aa**
PI	**5/5**	**1.000**	0.081	20.8 aa
EI	3/5	0.600	0.769	**6.2 aa**
MES	1/4	0.250	0.083	23.6 aa

Diversity gain
(similarity-aware vs utility-only)



Conclusions and limitations

☑ CONCLUSIONS

1. AULC gains were small, but some strategies reached $F1 \geq 0.86$ with fewer labels.
2. Uncertainty- and disagreement-based methods were best for model improvement.
3. Different acquisition objectives explored complementary regions of peptide space.
4. Higher utility often reduced diversity; explicit diversity control broadened coverage.
5. CG-MD simulations provided an effective way to obtain labels for iterative peptide discovery.

⚠ LIMITATIONS

1. Martini 2.2P was used; Martini 2.1P may be better for short peptides.
2. No statistical significance tests were performed.
3. CG-MD strategy selection was post hoc.
4. Fixed-surrogate evaluation used only one round with five candidates per policy.

Match the acquisition strategy to the scientific goal