https://doi.org/10.17952/37EPS.2024.P2275

Deep learning for Peptide Classification: From Data to Effective Models

Marko Njirjak[1], Marko Babić[2], Patrizia Jankovic[2], Erik Otović[1], Daniela Kalafatovic[2,3], Goran Mauša*[1,3]

[1] University of Rijeka, Faculty of Engineering, Rijeka, Croatia
[2] University of Rijeka, Faculty of Biotechnology and Drug Development, Rijeka, Croatia
[3] University of Rijeka, Center for Artificial Intelligence and Cybersecurity, Rijeka, Croatia

1. Overview of the proposed research pipeline

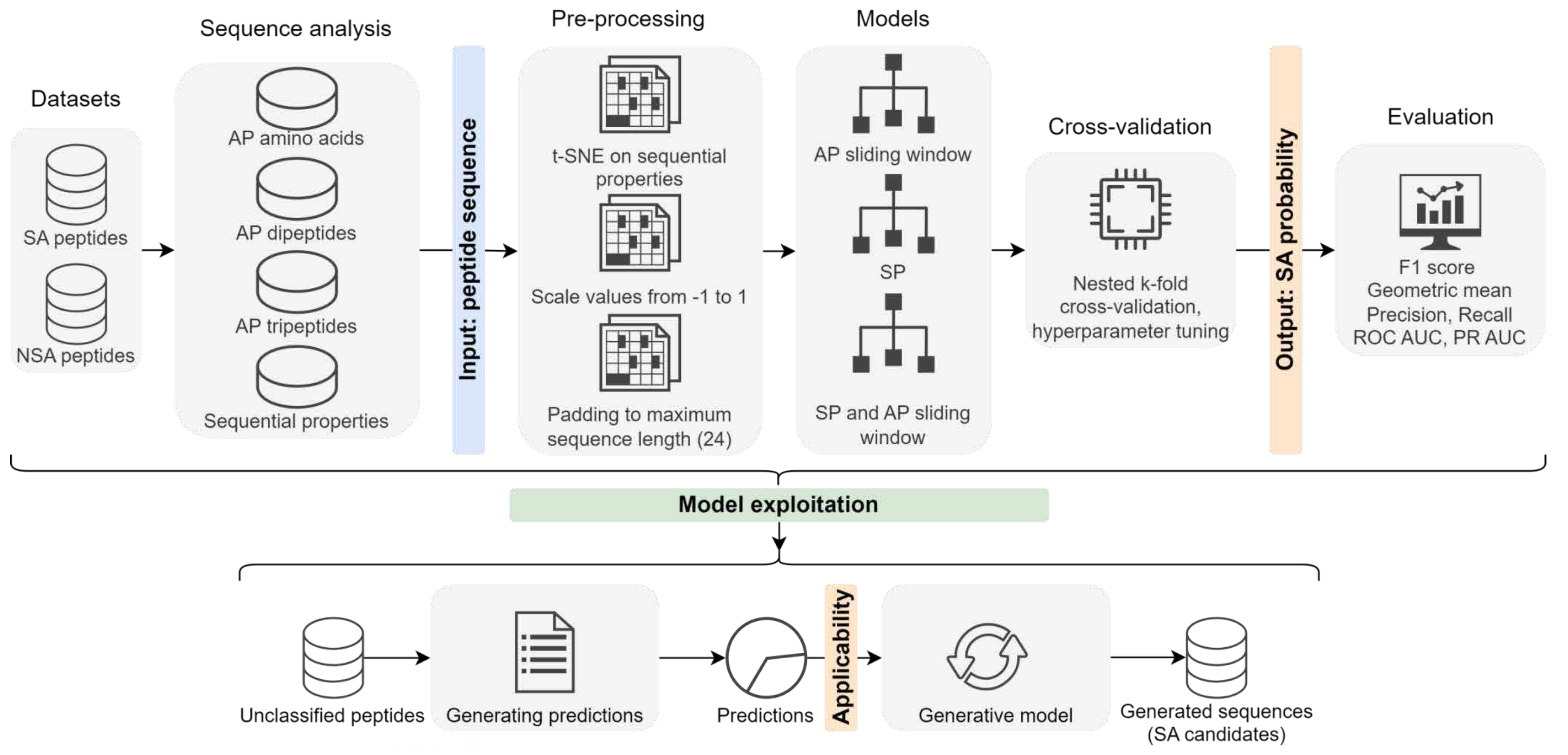
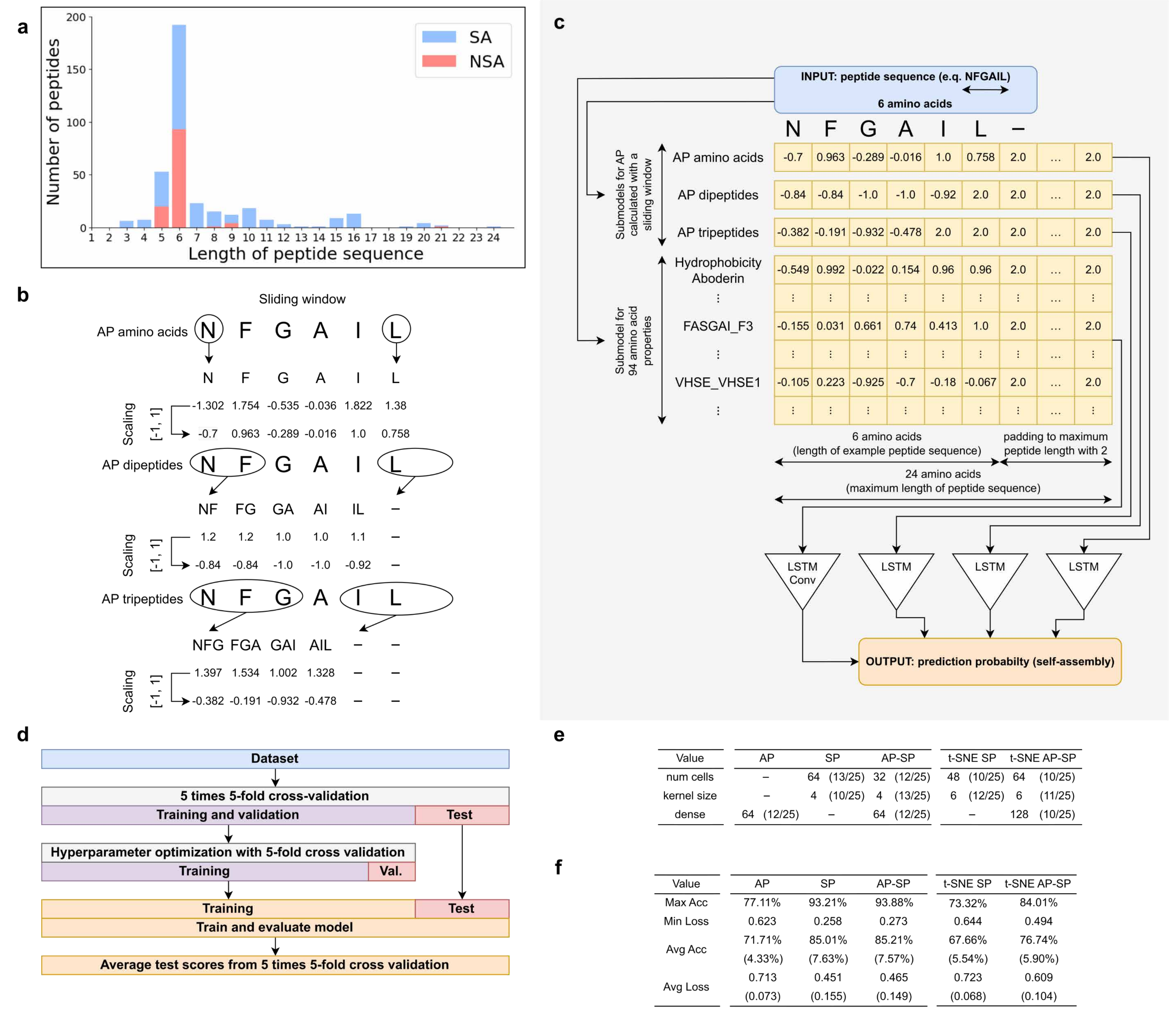


Fig. 1. The models are based on heterogeneous data obtained by applying sliding windows of varying lengths: from single amino acids to di- and tripeptides. Three models were trained on peptides with experimentally validated SA status. Data pre-processing based on t-SNE was performed for dimensionality reduction while padding shorter sequences to a maximum length of 24 residues was used to expedite training. The models were optimized and evaluated using the nested 5-fold cross-validation sampling technique before yielding the final model ready for exploitation. With the aim of discovering sequences with high self-assembly propensity, the hybrid AP-SP model was used as a guideline in a genetic algorithm-based generative approach.

2. Neural network data pre-processing and hyperparameter optimisation



e

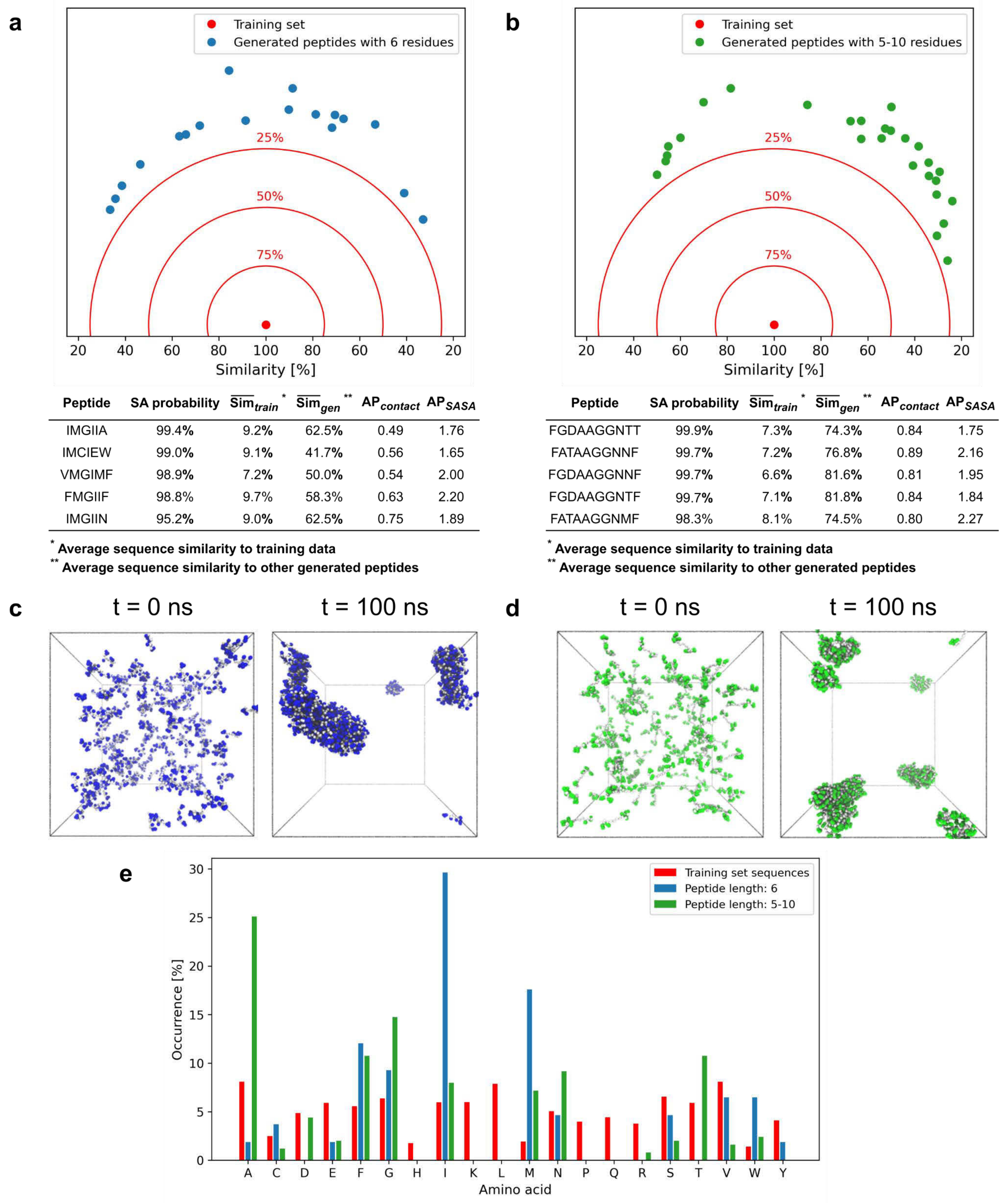
Value	AP	SP	AP-SP	t-SNE SP	t-SNE AP-SP
num cells	–	64 (13/25)	32 (12/25)	48 (10/25)	64 (10/25)
kernel size	–	4 (10/25)	4 (13/25)	6 (12/25)	6 (11/25)
dense	64 (12/25)	–	64 (12/25)	–	128 (10/25)

f

Value	AP	SP	AP-SP	t-SNE SP	t-SNE AP-SP
Max Acc	77.11%	93.21%	93.88%	73.32%	84.01%
Min Loss	0.623	0.258	0.273	0.644	0.494
Avg Acc	71.71% (4.33%)	85.01% (7.63%)	85.21% (7.57%)	67.66% (5.54%)	76.74% (5.90%)
Avg Loss	0.713 (0.073)	0.451 (0.155)	0.465 (0.149)	0.723 (0.068)	0.609 (0.104)

Fig. 2. **a** The distribution of peptides lengths within the dataset. The sequences were labelled by their experimentally confirmed self-assembly status. **b** Schematic representation of the sliding window pre-processing procedure identifying individual amino acids, di- and tri- peptides within the sequence. **c** Structure of the input data for an example sequence NFGAIL and a hybrid AP-SP RNN model. **d** Model construction workflow diagram. **e** Hyperparameter values selected by grid-search optimisation, along with the number of occurrences. 'Num cells' represents the number of cells in a bidirectional LSTM layer, 'kernel size' is the kernel size used in the convolutional layer, while 'dense' presents the number of units in the final densely connected layer of the model. **f** Maximum accuracy, minimal loss, and average accuracy and loss during training, along with standard deviations.

4. Applicability of the proposed model: Self-assembly potential in unexplored sequence space



a

Peptide	SA probability	$\overline{\text{Sim}}_{train}$ *	$\overline{\text{Sim}}_{gen}$ **	$AP_{contact}$	AP_{SASA}
IMGIIA	99.4%	9.2%	62.5%	0.49	1.76
IMCIEW	99.0%	9.1%	41.7%	0.56	1.65
VMGIMF	98.9%	7.2%	50.0%	0.54	2.00
FMGIIF	98.8%	9.7%	58.3%	0.63	2.20
IMGIIN	95.2%	9.0%	62.5%	0.75	1.89

* Average sequence similarity to training data
** Average sequence similarity to other generated peptides

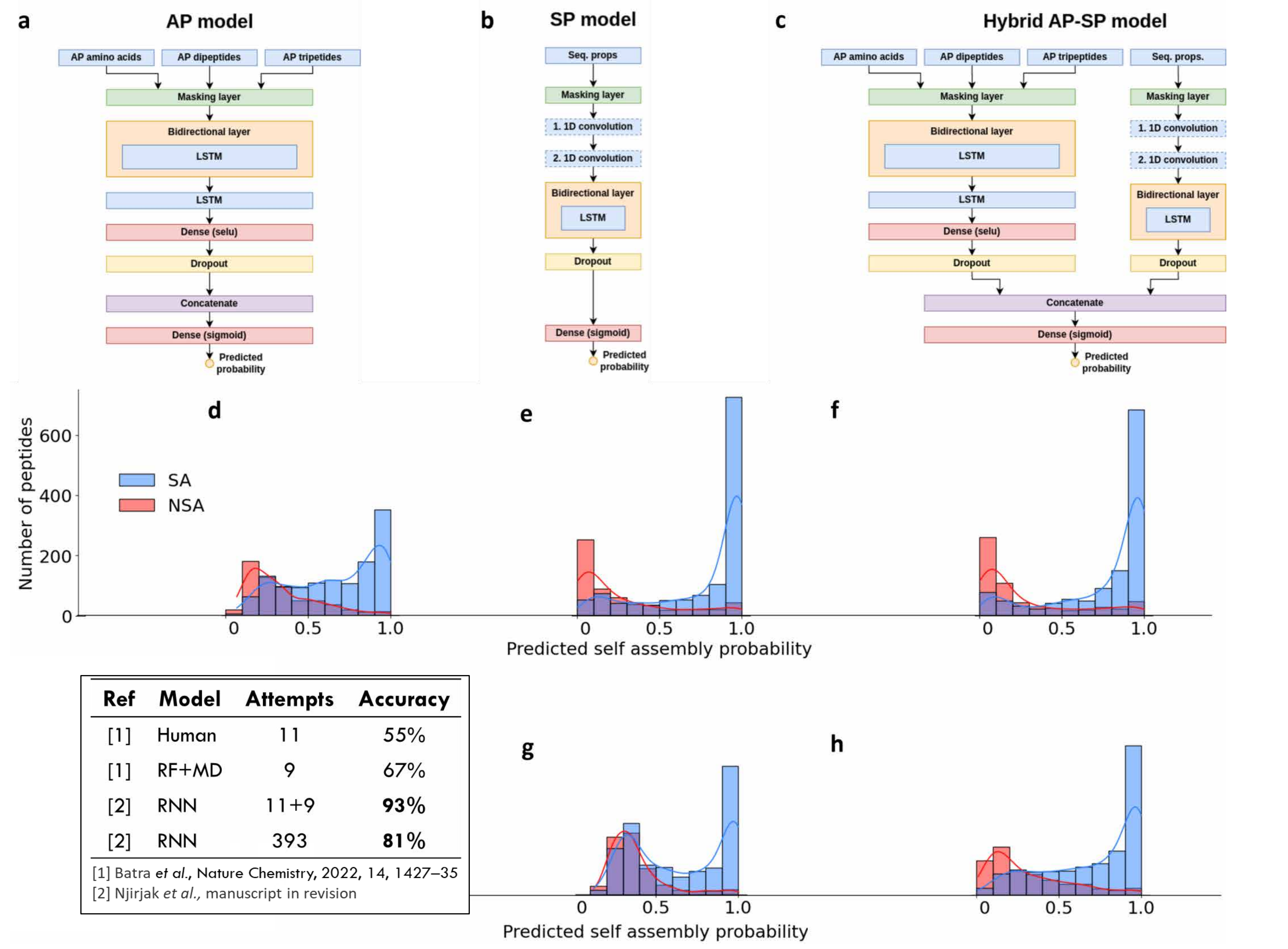
b

Peptide	SA probability	$\overline{\text{Sim}}_{train}$ *	$\overline{\text{Sim}}_{gen}$ **	$AP_{contact}$	AP_{SASA}
FGDAAGGNTT	99.9%	7.3%	74.3%	0.84	1.75
FATAAGGNNF	99.7%	7.2%	76.8%	0.89	2.16
FGDAAGGNNF	99.7%	6.6%	81.6%	0.81	1.95
FGDAAGGNTF	99.7%	7.1%	81.8%	0.84	1.84
FATAAGGNMF	98.3%	8.1%	74.5%	0.80	2.27

* Average sequence similarity to training data
** Average sequence similarity to other generated peptides

Fig. 4. **The peptides were generated using a genetic algorithm guided by the hybrid AP-SP model and validated using MD simulations.** A plot depicting the similarities between the generated peptides and the sequences used for model training, along with the accompanying table containing five generated sequences with the highest probability of self-assembly for **a** hexapeptides and **b** peptides with lengths between five and ten residues. The ability of the generated peptides to aggregate was validated using MD simulations where the initial (0 ns) and final (200 ns) frames of the simulation are shown for an example **c** hexapeptide (FMGIIF) and **d** decapeptide (FATAAGGNNF). **e** Comparison of amino acid distributions within the training dataset, generated sequences with six residues, and generated sequences with five to ten residues.

3. Architectures and respective performance assessments



Ref	Model	Attempts	Accuracy
[1]	Human	11	55%
[1]	RF+MD	9	67%
[2]	RNN	11+9	**93%**
[2]	RNN	393	**81%**

[1] Batra *et al.*, Nature Chemistry, 2022, 14, 1427–35
[2] Njirjak *et al.*, manuscript in revision

Fig. 3. Schematic representations of the architectures of the RNN models for **a** the AP model, **b** the SP model, and **c** the hybrid AP-SP model. Histograms of the self-assembly prediction probability distribution on the aggregated test folds for **d** the AP model, **e** the SP model, **f** the hybrid AP-SP model, **g** the SP model with t-SNE, and **h** the hybrid AP-SP model with t-SNE, for the experimentally confirmed SA peptides (blue) and NSA peptides (red).

5. Consensus motifs

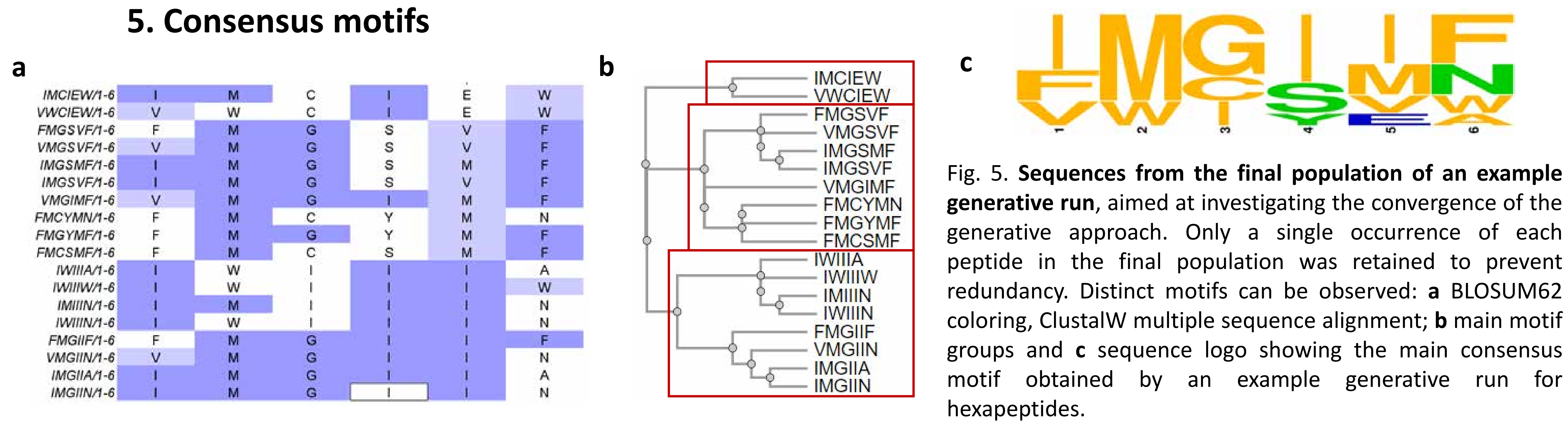


Fig. 5. **Sequences from the final population of an example generative run**, aimed at investigating the convergence of the generative approach. Only a single occurrence of each peptide in the final population was retained to prevent redundancy. Distinct motifs can be observed: **a** BLOSUM62 coloring, ClustalW multiple sequence alignment; **b** main motif groups and **c** sequence logo showing the main consensus motif obtained by an example generative run for hexapeptides.

References:

1. P. Janković, I. Šantek, A.S. Pina, D. Kalafatovic. Exploiting Peptide Self-Assembly for the Development of Minimalistic Viral Mimetics*. **Front. Chem., 2021; 9:723473***
2. G. Mauša, M. Njirjak, E. Otović, D. Kalafatovic. Configurable soft computing-based generative model: The search for catalytic peptides. ***MRS Adv., 2023;8: 1068–74.***
3. E. Otović, M. Njirjak, D. Kalafatovic, G. Mauša. Sequential Properties Representation Scheme for Recurrent Neural Network-Based Prediction of Therapeutic Peptides. ***J. Chem. Inf. Model. 2022;62(12): 2961-72.***
4. M. Negovetić, E. Otović, D. Kalafatovic, G. Mauša. Efficiently solving the curse of feature-space dimensionality for improved peptide classification. ***Digit. Discov., 2024,3, 1182-93***
5. M. Njirjak, L. Žužić, M. Babić, P. Janković, E. Otović, D. Kalafatovic, G. Mauša. AI-driven generative approach guided by hybrid deep learning. 2024; *in revision in* **Nat. Mach. Intell.**

Acknowledgements*:*

1. *Croatian Science Foundation (UIP-2019-04-7999, DOK-2021-02-3496, DOK-2020-01-4659)*
2. *University of Rijeka (uniri-23-78, uniri-23-16, UNIRI-INOVA-3-23-2)*