nature chemistry

Article

https://doi.org/10.1038/s41557-022-01055-3

Machine learning overcomes human bias in the discovery of self-assembling peptides

In the format provided by the authors and unedited

Supporting Information on Machine learning overcomes human bias in the discovery of self-assembling peptides

Rohit Batra, Troy D. Loeffler, Henry Chan, Srilok Srinivasan, Honggang Cui, Ivan V. Korendovych, Vikas Nanda, Liam C. Palmer, Lee A. Solomon, H. Christopher Fry, and Subramanian KRS Sankaranarayanan

May 22, 2022

1 AI-expert screened top 100 pentapeptide

The screening of 9 peptides from the proposed top 100 from the AI expert were based on selecting candidates with diverse values of parameter logP (in particular, around values of -0.5, -0.2, 0.2, 1, 2). Furthermore, sequences that broke the rational design convention (e.g., RWLDY, KWEFY and KWMDF) or employed amino acids that human experts rarely include in rationally designed peptides (proline, tyrosine, threonine, and serine) were preferred. Additionally, two sequences that conform to conventional design approaches and the human experts would have preferred to design (i.e., FFEKF and FKIDF) were also selected.

Table S1: Computational results for the top 100 pentapeptides identified by the AI-expert using the r^{penta} reward function. Aggregation propensity (AP), logP and the associated reward score with ($r^{\text{penta}}\text{w}\beta$) and without ($r^{\text{penta}}$) the β-sheet propensity factor are provided. Nine candidates that were selected for synthesis are marked with a bold font.

ID	Peptide	AP	logP	r^{penta}	$r^{\text{penta}}\text{w}\beta$
1	YPGVY	2.47	-0.59	0.441028	0.210040
2	**SYCGY**	2.42	0.17	0.427704	0.191398
3	**PTPCY**	2.54	-0.20	0.492614	0.168720
4	PCPYC	2.49	-0.47	0.455615	0.164591
5	YHSQY	2.38	-0.08	0.399644	0.162855
6	TSPCY	2.38	0.12	0.403479	0.158870
7	FKDFF	2.27	1.31	0.361773	0.158728
8	**FFEKF**	2.24	1.30	0.345121	0.155736
9	KWEFF	2.28	0.92	0.361217	0.155323
10	FKDFI	2.17	1.90	0.315894	0.153998
11	FFKEF	2.23	1.30	0.339708	0.153293
12	STPCY	2.35	0.12	0.386493	0.152181
13	IKDFF	2.15	1.90	0.305452	0.148908
14	FIKDF	2.14	1.90	0.300297	0.146395
15	CPPHY	2.50	-0.34	0.464626	0.146357
16	VPPYA	2.38	-0.39	0.393625	0.143673
17	**KWEFY**	2.20	1.92	0.332155	0.141581
18	FEKFF	2.18	1.30	0.313282	0.141368
19	**KWMDF**	2.28	1.97	0.377528	0.139213

20	PSPYV	2.34	-0.43	0.370869	0.138149
21	FEFFK	2.14	1.30	0.292911	0.132176
22	YDRFF	2.13	1.32	0.288171	0.128596
23	KEFFW	2.16	0.92	0.298071	0.128171
24	KDFYW	2.14	1.93	0.300661	0.124399
25	FKWYD	2.13	1.93	0.295544	0.122281
26	**FKIDF**	2.03	1.90	0.246485	0.120161
27	FFDFK	2.09	1.31	0.268525	0.117815
28	LKEFF	2.10	1.76	0.278522	0.116979
29	CCPYA	2.16	-0.11	0.284249	0.115121
30	CDWYY	2.05	0.11	0.236606	0.114754
31	WDPYV	2.08	0.52	0.254739	0.113040
32	KVPWY	2.07	-0.32	0.240401	0.112988
33	WVEYC	1.94	0.35	0.193019	0.111227
34	WKPYV	2.06	-0.32	0.236050	0.110943
35	**RWLDY**	2.11	1.40	0.279264	0.109960
36	KVPYF	2.01	0.06	0.218884	0.107527
37	PPNYY	2.38	-0.29	0.395577	0.107300
38	WKDMF	2.12	1.97	0.290939	0.107284
39	YKGYY	2.02	1.82	0.241072	0.106674
40	PPQYY	2.35	-0.37	0.377428	0.106623
41	YDCYW	2.00	0.11	0.215205	0.104374
42	YMEYY	1.99	0.83	0.218122	0.102790
43	**WKPYY**	2.11	-0.57	0.254996	0.099767
44	WKPYC	2.09	0.12	0.254577	0.099285
45	FELKF	2.01	1.76	0.235923	0.099088
46	VHPRY	1.99	0.89	0.218702	0.098689
47	KCPFY	2.04	0.50	0.236494	0.097258
48	PPPYT	2.44	-0.04	0.435217	0.096836
49	YSDFY	1.99	0.97	0.219472	0.096019
50	VKPWY	1.98	-0.32	0.202672	0.095256
51	YVPWK	1.98	-0.32	0.202672	0.095256
52	PGPYY	2.26	0.01	0.335962	0.092810
53	HWMEY	2.02	0.27	0.225331	0.091541
54	CDCWW	1.96	-0.58	0.192209	0.089618
55	IKPYY	1.94	0.40	0.193466	0.089236
56	MYDYY	1.92	0.84	0.189295	0.086839
57	**PPPHY**	2.46	-0.18	0.444123	0.086604
58	FVPDY	1.91	0.90	0.185823	0.086408
59	CYPDF	1.99	1.34	0.222999	0.085855
60	SWLDY	2.01	0.05	0.218780	0.085051
61	YMDYY	1.91	0.84	0.185331	0.085021
62	DVPYY	1.88	1.90	0.181661	0.083791
63	PPPCY	2.28	-0.31	0.341120	0.082295
64	KPHFY	2.01	0.63	0.224741	0.082030
65	PPPYS	2.49	0.17	0.469997	0.081074
66	CKPYY	1.92	1.50	0.194750	0.079360
67	PSPSY	2.28	0.49	0.354321	0.079279
68	PKFYY	1.95	-0.19	0.192031	0.079213
69	KPFFT	1.96	-0.23	0.195588	0.077502
70	FGYYK	1.87	0.82	0.169743	0.075748
71	FSPKF	2.01	-0.02	0.218049	0.075500
72	LKPYY	1.94	0.27	0.192303	0.073316
73	FTDYM	1.85	0.80	0.162139	0.071746

74	PSPNY	2.26	0.88	0.349655	0.070368
75	MKPYY	1.93	0.85	0.193387	0.070345
76	YKPYL	1.91	0.27	0.180592	0.068851
77	FSPDY	1.96	1.82	0.214308	0.067775
78	KWPYM	1.96	-0.53	0.192696	0.066721
79	PPPYA	2.37	0.21	0.399466	0.065912
80	PFDHF	1.92	0.47	0.186168	0.063762
81	YMPKY	1.88	0.85	0.173767	0.063208
82	PKPFY	1.99	0.66	0.216473	0.063048
83	PKSYY	1.88	1.98	0.182248	0.061737
84	PKPWY	2.01	0.28	0.221163	0.059714
85	PKPYV	1.81	1.91	0.154851	0.056714
86	PKPYF	1.93	0.66	0.191754	0.055848
87	PPPSY	2.23	0.17	0.322960	0.055711
88	PDPFF	1.97	0.50	0.206563	0.055514
89	PKPYY	1.91	1.66	0.191943	0.055184
90	PPGYY	1.96	0.01	0.197871	0.054662
91	FDPYP	1.94	1.50	0.203036	0.053805
92	GFPDF	1.86	1.51	0.170998	0.051727
93	PPPNY	2.24	0.56	0.334077	0.050112
94	PPKYY	1.86	1.66	0.172067	0.049469
95	FPPKF	1.88	-0.34	0.164363	0.048487
96	PPPYP	2.36	-0.15	0.387074	0.046933
97	DPPWY	1.91	1.12	0.187615	0.045731
98	PPPGY	2.12	0.86	0.277634	0.043033
99	PKPYM	1.83	1.70	0.160928	0.039025
100	PFEPF	1.73	0.49	0.119256	0.033541

2 Human experts proposed pentapeptides

29 Peptides were contributed by our 5 human experts and the corresponding author H.C. Fry. The peptides were selected by the corresponding author to 1) ensure equal contribution from each expert as best as possible and 2) diversity in sequence selection as best as possible. The list of the pentapeptides proposed by the human experts and the rationale for choosing/rejecting a sequences for synthesis is provided in Table S2.

Using the proposed 29 human expert sequences, we can also extract some selection trends. In the Extended Data Figure 4 it can be seen that human significantly favor F, V, and K residues and overall charge neutral pentapeptides.

Table S2: Pentapeptide sequences proposed by the human experts for self-assembly. Sequences selected for synthesis are marked with a bold font.

ID	Expert	Peptide	Comment
1	#1	**VKVFF**	This was the only contribution from human expert #1. This peptide assembled.
2	H. C. Fry	**FKFEF**	This was chosen as an example of a phenylalanine rich peptide that was electrostatically balanced. This peptide assembled.
3	H. C. Fry	FKFKF	This peptide was not chosen as it was similar to the previous peptide sequence.
4	H. C. Fry	**VKVEV**	The peptide was chosen as a high β-sheet propensity peptide that is electrostatically balanced. This peptide assembled.
5	#2	KFFFK	This peptide was not chosen due to its highly positive charge, and thus the possibility that it would not assemble.
6	#2	EFFFE	This peptide was not chosen due to its overall negative charge and similarly to KFFFK.
7	#2	**KFFFE**	This peptide was chosen as it possesses an overall neutral charge and is similar to what was proposed by the corresponding author. This peptide assembled.
8	#2	EFFFK	This peptides was not synthesized due to its similarity to the other peptides proposed by the contributor.
9	#3	LPFFD	This peptide was not synthesized because the author had already synthesized it and confirmed that it does not assemble. Therefore, we did not want to include it in our experimental group.
10	#3	**KVKVK**	This peptide was chosen as one of the two contributions that had not been synthesized previously. This peptide did NOT assemble.
11	#3	KVVVK	This peptide from human expert #3 was not synthesized as it was similar to a previous submission.
12	#4	AFAIK	This peptide was not synthesized as we believed it was a "curious" submission and is a common mnemonic for "As far as I know."
13	#4	AYLKK	This peptide was not chosen as we felt the amphipilic nature and high positive charge would not assemble into a discernible structure.
14	#4	LKLKL	This peptide was not synthesized as it resembled the peptide "VKVKV" but represented a lower β-sheet propensity than VKVKV.
15	#4	**VKVKV**	This peptide was synthesized as it resembled the peptide from human expert #3 (KVKVK) but is more hydrophobic and we believed it would assemble into a β-sheet structure due to its alternating pattern (npnpn) and employment of a high β-sheet propensity amino acid. This peptide did NOT assemble.
16	#4	LRLRL	This was peptide was not chosen as it resembled the contribution LKLKL. The same rational was used to not prepare this peptide as LKLKL.
17	#4	KKFDD	This peptide was synthesized due to its charge balanced nature between the lysine and aspartic residues along with a phenylalanine in the center that could lead to a longitudinal array of overlapping Phe residues. This peptide did NOT assemble.
18	#4	**KFAFD**	This peptide was chosen as it resembled the peptide designed by human expert #2 but included a smaller amino acid, alanine sandwiched between two phenylalanine. It followed the pattern (pnnnp). This peptide assembled.
19	#4	RRYEE	This peptide was not synthesized as we felt it was a redundant design, KKFDD, from the same contributor.
20	#4	KRYDE	This peptide was not synthesized as we felt it was a redundant design, KKFDD, from the same contributor.
21	#5	**RVSVD**	This peptide from contributor #5 was synthesized due to its employment of serine and it's patterning (pnpnp)—while the center serine represents a polar uncharged group and the Arg and Asp residues are electrostatically complementary. This peptide did NOT show assembly within the 24 hr period.
22	#5	RFNEF	This peptide was not synthesized as we felt it was a redundant design, RVSVD, from the same contributor.
23	#5	LEEAS	This peptide represents another "curious" submission from a human expert. It is a pentapeptide styling of the contributors name. We did not synthesize this.
24	#5	ANANA	Due to the alanine and the polar Asn groups we felt this peptide would be too soluble.
25	#5	ANACA	Due to the alanine and the polar Asn groups we felt this peptide would be too soluble. The cysteine group represents potential cross-linking that we wanted to avoid. This peptide was not synthesized.
26	#5	CACAA	Due to the alanine rich nature we felt this peptide would be too soluble. In addition, the cysteine residues have an unfair advantage due to potential crosss-coupling of the cysteine residues that are not comparable to our MD simulations.
27	#3	WVVHH	This peptide was not synthesized as it represented a late submission.
28	#3	WAAHH	This peptide was not synthesized as it represented a late submission.
29	#3	WGGHH	This peptide was not synthesized as it represented a late submission.
-	H.C. Fry	VVVVV*	This peptide was employed as the highest propensity β-sheet foldamer. It's logP value suggests that it is very hydrophobic but not to the point of insolubility.
-	H.C. Fry	DPDPD*	This peptide was employed as a "control" peptide which is highly polar and incredibly low β-sheet propensity.

$*$ These cases were added later based on β-sheet scoring discussion.

3 Influence of reward function on pentapeptide screening

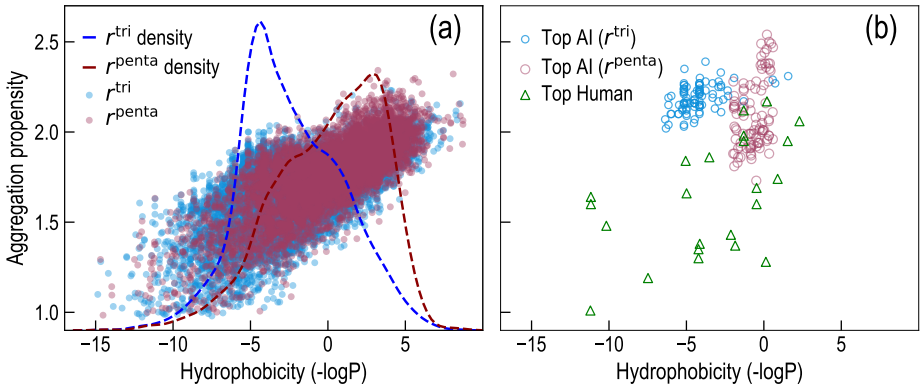

Figure S1: The dependency of the peptide search on the choice of the MCTS reward/scoring function. (a) The r^{tri}=AP^2.logP and r^{penta}=AP^2.logP$^{0.5}$ reward functions generated pentapeptide candidates in different regions of hydrophobicity and AP values during MCTS run. The probability density function of for the two reward functions is estimated using kernel density estimation. (b) Top screened pentapeptide candidates from the human experts and the AI-expert with r^{tri} and r^{penta} reward function.

The Monte Carlo tree search (MCTS) reward/scoring function r (see Eq. 1 in main text) significantly impacts the results of the peptide search. As shown in Figure S1(a), the AI-expert generated pentapeptide candidates belonging to different regions of the hydrophobicity and aggregation propensity (AP) values depending on the reward function. While the use of reward definition r^{tri} produced highly hydrophilic pentapeptides (2 < logP < 6), the r^{penta} lead to slightly hydrophobic candidates with 0 < logP < 4, although the peak in the latter case was quite broad. This is mainly because a majority of the amino acids are hydrophilic in nature and the use of r^{tri} reward function for the case of pentapeptides tips the balance between the power factor of AP (α) and logP (β) to erroneously favor highly hydrophilic peptides that are not expected to assemble. To recalibrate this balance we adjusted the logP power factor $\beta = 0.5$, which produced more sensible pentapeptide candidates. The selected top 100 candidates from the two scoring systems, along with those proposed by human experts are shown in Figure S1(b). We note that in the case of r^{penta} reward function an additional criteria of -0.6 < logP < 2 was imposed during top candidate selection to avoid highly hydrophobic candidates. The distinct region occupied by the two scoring functions is evident. The success rate of the two scoring functions also varied drastically; while only 20 % of the top scoring pentapeptides based on r^{tri} showed any aggregation, nearly 67 % of those based on r^{penta} formed aggregates. This reflects the sensitivity of the performance of the AI-expert to the choice of the reward function.

4 Pentapeptide morphology using atomic force microscopy

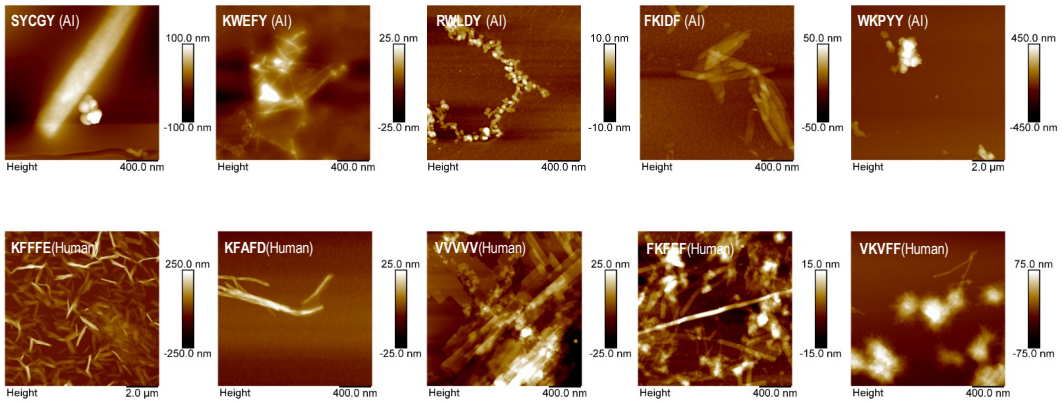

Figure S2: Atomic force microscopy (AFM) results (images are a representation of three trials yielding similar results) showcasing secondary structures in the pentapeptides suggested by the AI-expert (top) and the human experts (bottom).

5 Random forest model input features

Based on our past experience on fingerprinting organic materials [1], three hierarchical levels of features were considered, capturing different geometric and chemical information about the peptides at multiple length-scales. At the atomic scale, a count of a predefined set of motifs is included. The motifs are specified by the generic label "$A_iB_jC_k$", representing an i-fold coordinated A atom, a j-fold coordinated B atom, and a k-fold coordinated C atom, connected in the specified order. For example, N3-C3-C4 represents a three-fold coordinated N, a three-fold coordinated carbon and a four-fold coordinated carbon [2]. At a slightly larger length-scale, quantitative structure-property relationship (QSPR) descriptors, often used in chemical and biological sciences, and implemented in the RDKit Python library, were used [3, 4, 5]. Lastly, at the highest length-scale, 'morphological descriptors', such as length of the largest side-group, shortest topological distance between rings, etc. were considered. More details on the different hierarchical descriptors can be found in our previous works [6].

6 Score comparison with previous work

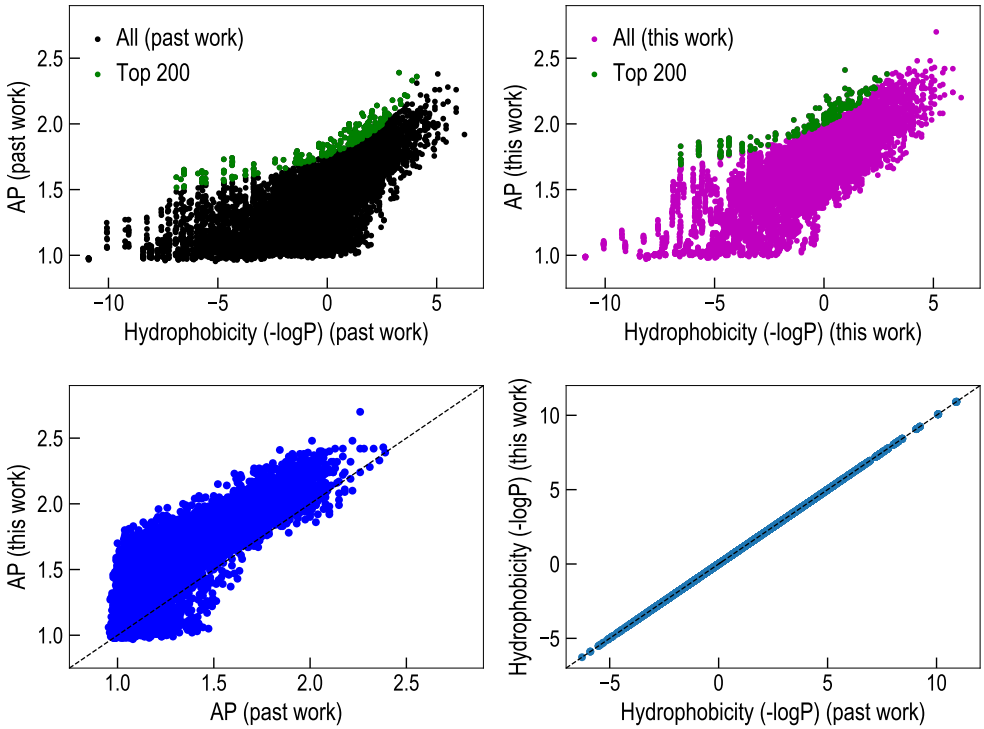

Figure S3: Comparison of aggregation propensity (AP) and hydrophobicity (-logP) of all 8000 tripeptide as computed in this work against that in the past work [7]. (Top row) The overall features of the AP vs hydrophobicity plots for all tripeptides, along with the top scoring 200 candidates, can be seen to be reproduced well. (Bottom row) While the hydrophobicity values were exactly reproduced from the past work, the AP values were consistently over predicted due to different version of the GROMACS modeling software that leads to slightly different values of the solvent accessible surface area (SASA). Statistical nature of the MD simulations also leads to slight variations in the AP values as compared to the past work. This, however, does not affect the conclusions regarding the efficiency of the AI-expert proposed in this work.

7 Mass spectrometry data

Table S3: ESI mass spectrometry data for the synthesized pentapeptides

Peptide	MS m/1	MS Calc
VVVVV	514.6	513.7
DPDPD	558.3	557.5
VKVKV	572.4	571.6
VKVEV	573.5	572.7
RVSVD	575.4	574.6
PTCPY	580.3	579.7
SYCGY	592.0	591.6
KVKVK	601.5	600.8
PPPHY	610.4	609.7
KFAFD	627.4	626.7
VKVFF	639.4	638.8
KKFDD	652.4	651.7
FKIDF	669.5	668.8
YDPKY	685.3	684.8
KDPYY	685.3	684.8
YEPYK	699.3	698.8
EPYYK	699.3	698.8
YTEYK	703.5	702.8
KDHFY	709.3	708.8
FFEKF	717.4	716.8
FKFEF	717.3	716.8
KFFFE	717.3	716.8
KFFDY	719.5	718.8
KWMDF	726.2	725.9
RWLDY	752.6	751.8
WKPYY	756.5	755.9
KWEFY	772.4	771.9

8 IR measurement of solid films

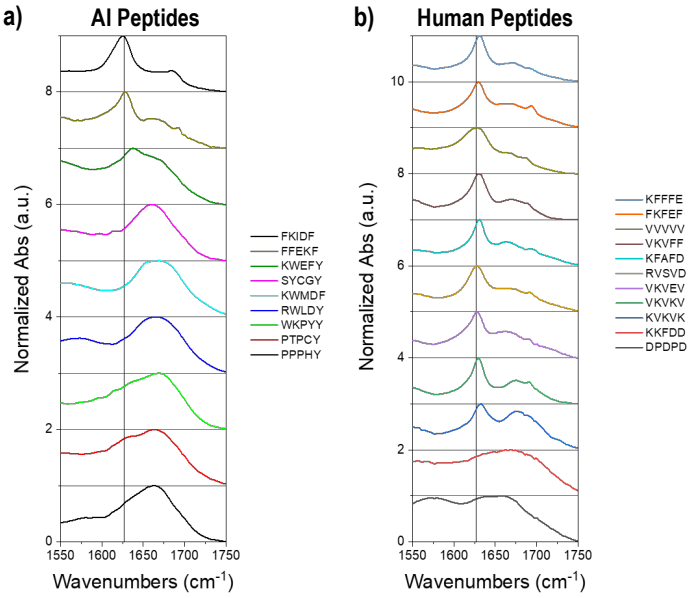

Figure S4: Spectroscopy measurements on the dried peptide films cast from diluted stock solutions (10 μL of a 0.2 wt% solution onto a CaF_2 window). The peak at 1627 cm^{-1} is representative of a β-sheet conformation [8], and can be seen in most of the human suggested candidates, indicating their bias.

9 Chronology of events in this study

Since the human experts were co-authors in this study, a natural question arises that did the human experts knowingly suggested sequences that will boost the performance of the AI-expert? In Table S4 we detail the chronology of events that led to this study. On June 17, 2020 the human experts were asked to provide pentapeptide sequences through email. No mention of co-authorship was made in that correspondence. Further, at that time, the AI sequences had not been generated and thus they were blinded to the AI-expert sequences. Care was taken to ensure that the human experts were blinded to the sequences proposed by the other groups. Only when the final comparison results were obtained after peptide synthesis and characterization, the human experts were made aware of the outcome of the comparative study and were offered co-authorship.

Table S4: Timeline of various stages involved in this study. This shows how human experts were blinded to the results received by the AI-expert, or were not incentivized in any manner to alter the outcome of this study.

Date	Action
Jan–Mar, 2020	Study conception and ideation
Mar-June 15, 2020	Validation of AI-expert for the case of tripeptides
June 17, 2020	Email sent to human experts requesting pentapeptide sequences
June 25, 2020	Responses received from all human experts
July 20, 2020	Peptides shortlisted for synthesis; synthesis begins (human experts)
Sept 25, 2020	Top 100 predictions obtained from the AI-expert
Sept 28, 2020	Peptide shortlisted for synthesis; synthesis begins (AI-expert)
Dec 4, 2020	β-sheet factor discussed, employed in ranking but not implemented in AI scoring as discussed in the manuscript
Feb 1, 2021	Experimental scoring procedure developed
March 19, 2021	Peptide synthesis and characterization ends
Dec-March, 2021	Results analysis and discussion
Feb–Mar, 2021	Manuscript preparation (by RB, HCF, SKRSS)
March 23, 2021	Manuscript sent to other authors for perusal; human experts informed about their co-authorship for the first time
May 7, 2021	Manuscript finalized and submitted for publication

References

[1] Rohit Batra, Henry Chan, Ganesh Kamath, Rampi Ramprasad, Mathew J Cherukara, and Subramanian KRS Sankaranarayanan. Screening of therapeutic agents for covid-19 using machine learning and ensemble docking studies. *J. Phys. Chem. Lett.*, 11(17):7058–7065, 2020.

[2] Tran Doan Huan, Arun Mannodi-Kanakkithodi, and Rampi Ramprasad. Accelerated materials property predictions and design using motif-based fingerprints. *Phys. Rev. B*, 92(1):014106, 2015.

[3] Chartchalerm Isarankura-Na-Ayudhya, Thanakorn Naenna, Chanin Nantasenamat, and Virapong Prachayasittikul. A practical overview of quantitative structure-activity relationship. *EXCLI Journal*, 8:74, 2009.

[4] Chanin Nantasenamat, Chartchalerm Isarankura-Na-Ayudhya, and Virapong Prachayasittikul. Advances in computational methods to predict the biological activity of compounds. *Expert Opin. Drug Discov.*, 5(7):633–654, 2010.

[5] RDKit open source toolkit for cheminformatics. http://www.rdkit.org/.

[6] Chiho Kim, Anand Chandrasekaran, Tran Doan Huan, Deya Das, and Rampi Ramprasad. Polymer genome: A data-powered polymer informatics platform for property predictions. *J. Phys. Chem. C*, 122(31):17575–17585, 2018.

[7] Pim WJM Frederix, Gary G Scott, Yousef M Abul-Haija, Daniela Kalafatovic, Charalampos G Pappas, Nadeem Javid, Neil T Hunt, Rein V Ulijn, and Tell Tuttle. Exploring the sequence space for (tri-) peptide self-assembly to design and discover new hydrogels. *Nat. Chem.*, 7(1):30, 2015.

[8] Jilie Kong and Shaoning Yu. Fourier transform infrared spectroscopic analysis of protein secondary structures. *Acta Biochim. Biophys. Sin.*, 39(8):549–559, 2007.