Electronic Supplementary Material (ESI) for Biomaterials Science.

Supplementary Information for

Inverse design of viral infectivity-enhancing peptide fibrils from continuous protein-vector embeddings

Kübra Kaygisiz, Arghya Dutta, Lena Rauch-Wirth, Christopher V. Synatschke, Jan Münch, Tristan Bereau*, Tanja Weil*

K. Kaygisiz, Dr. C. V. Synatschke, Prof. Dr. T. Weil
Department Synthesis of Macromolecules
Max Planck Institute for Polymer Research
Ackermannweg 10, 55128 Mainz, Germany
Email: weil@mpip-mainz.mpg.de, synatschke@mpip-mainz.mpg.de

Dr. A. Dutta, Dr. T. Bereau
Polymer Theory
Max Planck Institute for Polymer Research
Ackermannweg 10, 55128 Mainz, Germany

Dr. A. Dutta,
Institute of Biochemistry II, Faculty of Medicine,
Goethe University,
Theodor–Stern–Kai 7, 60590 Frankfurt, Germany (*present address*)

Dr. T. Bereau
Institute for Theoretical Physics, Heidelberg University,
Philosophenweg 19, 69120 Heidelberg, Germany (*present address*)
Email: bereau@thphys.uni-heidelberg.de

L. Rauch-Wirth, Prof. Dr. J. Münch
Institute of Molecular Virology
Ulm University Medical Center
Meyerhofstraße 1, 89081 Ulm, Germany
Email: jan.muench@uni-ulm.de

Corresponding authors: Tristan Bereau (bereau@thphys.uni-heidelberg.de) and Tanja Weil (weil@mpip-mainz.mpg.de)

Contents

1. Regression Model: Infectivity Prediction

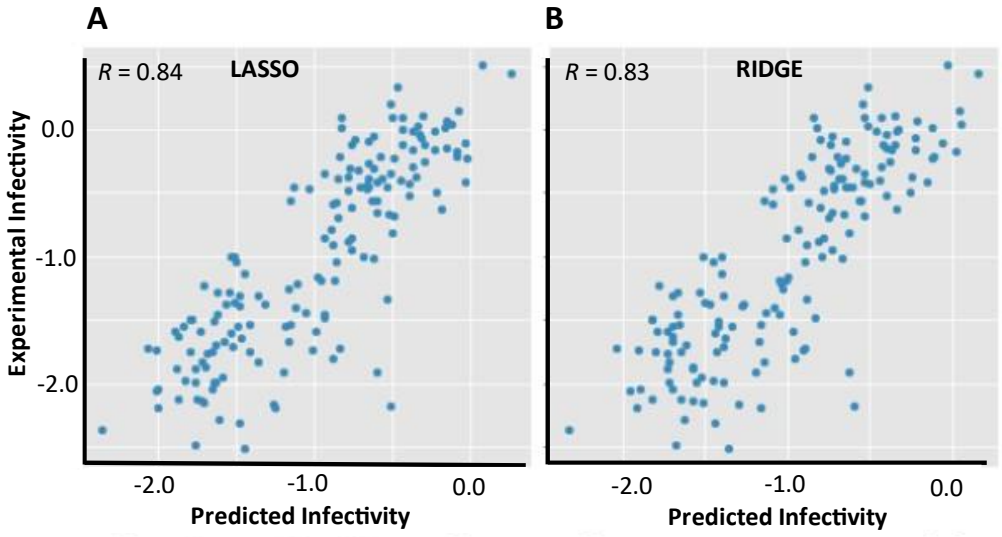


Figure S1 A LASSO and **B** RIDGE linear regression models were trained via a 5-fold cross validation.

LASSO and RIDGE regression models were trained on the training set peptides, represented as 100-d numerical vectors using continuous vector representations; the models perform similarly (**Figure S1**). Since LASSO regression applies regularization by minimizing the number of non-zero coefficients, the resulting model contains only the relevant parameters. This results in a simpler model with fewer parameters. For example, for our model only 21 vectors have a non-zero coefficient (**Eqn. 1**). Interestingly, while RIDGE regression equation contains all the vector components, it offers slightly poorer correlation as shown in (**Figure S1**).

(**Eqn. 1**) Log Infect. Rel EF-C = -2.004 + (-0.496) * vec40 + (-0.467) * vec4 + (-0.374) * vec30 + (-0.276) * vec68 + (-0.239) * vec8 + (-0.212) * vec20 + (0.179) * vec55 + (-0.17) * vec88 + (0.153) * vec16 + (0.131) * vec59 + (0.129) * vec57 + (0.099) * vec100 + (-0.095) * vec43 + (0.06) * vec22 + (0.057) * vec46 + (-0.05) * vec99 + (-0.03) * vec71 + (0.024) * vec60 + (-0.009) * vec61 + (-0.009) * vec54 + (-0.002) * vec52

Mean squared error = 0.179
Mean absolute error = 0.328
R^2= 0.695
Pearson R = 0.837

All code and data used in ML training are openly available at https://gitlab.com/arghyadutta/seq-to-infect.

2. Aggregation Prediction

Aggregation was found as a necessary property for infectivity enhancement of peptides previously by us and others.[1–4] Therefore, we applied the open accessible protein-aggregation tools Tango,[5] APPNN,[6] Waltz,[7] PATH,[8] Aggrescan[9] and PASTA 2.0[10] to preselect promising *de novo* created peptides.

Aggrescan is based on statistical analysis of the aggregation-propensity value for each amino acid residue in the sequence and a subsequent aggregation prediction by hot spot regions, identified from the peptide aggregation profile. Here, we consider a sequence as amyloidogenic if there is at least one predicted hotspot.

Waltz applies statistical analysis of a sequence and was originally developed by position specific matrix for 1089 short 6-mer peptides sequences, which were experimentally determined for fibril formation.[7] Here, we considered a sequence as amyloidogenic if at least one amyloidogenic region was detected upon entry of following parameters: threshold custom 0-100 and pH 7.0.

Tango is designed to predict aggregating regions in unfolded polypeptide chains. statistical mechanics algorithm. The method is benchmarked against experimentally observed 179 peptides.[5] Here, we applied following input parameters to determine the β-sheet aggregation tendency (aggregation parameter): pH 7.4, 298 K, ionic strength 0.1724. We select a threshold above 5.0% over 5 residues to identify hotspots for aggregation as suggested by the authors to determine amyloidogenic sequences.[5]

PATH is a structure-based method for predicting amyloidogenicity by threading and machine learning. Here, we considered a peptide as aggregating if at least one amyloidogenic region was calculated.

PASTA 2.0 is based on energetic functions which were determined experimentally from protein structures interactions potential and H-bond formation between all non-consecutive residues for parallel and anti-parallel β-pairing. A sequence is considered amyloidogenic if the pasta energy for the lowest predicted pairing is lower or equal to the threshold stated by the authors (-4.0).[10] The parameters for the prediction was threshold custom, top pairing energy 20, energy threshold -2 PEU, large scale true, protein-protein analysis: false.

APPNN applies a neural network machine learning approach based on the analysis of seven physicochemical and biochemical features such as β-sheet frequency, hydrophobic moment, helix termination parameters or isoelectric point. A sequence was considered amyloidogenic if at least one of these six amino acid windows was classified amyloidogenic.

Except for Waltz, these prediction tools were developed based on a polypeptide and protein aggregation and not on short self-assembling peptides. To find the best performing tool for our self-assembling peptide library, we applied the experimental data on self-assembly by electron microscopy[4] (**Table S1**) as a dataset to evaluate the accuracy and reliability of each tool for self-assembly with the accuracy and receiver operating characteristic (ROC) value. The accuracy was calculated from the confusion matrix according to **Eqn.2**.

(**Eqn.2**) $$accuracy = \frac{\sum true\ positive + \sum true\ negative}{\sum total\ population}$$

The ROC value for the prediction (**Figure S2**) was calculated with 10-fold stratified cross-validation and the experimental fibril formation as target value and a logistic regression learner and LASSO regularization model (17 strength) with the data-mining software orange3.[11]

The prediction tools Aggrescan, APPNN and PATH performed best with an accuracy of 76%, 69% and 69%, respectively. Even though these aggregation prediction tools are trained on polypeptides and proteins, the reported accuracy for these tools match well to our self-assembling peptide library composed of short peptides. Noteworthy, combining Aggrescan, APPNN and PATH increase the performance of aggregation further (**Figure S2**).

Therefore, we applied Aggrescan, APPNN and PATH to predict aggregation propensity of the *de novo* predicted 3669 sequences. A sequence was considered aggregating, if at least two of Aggrescan, APPNN or PATH were positive. By applying this method 424/3669 peptides were predicted for aggregation by at least two of these tools.

As shown in **Figure S3** the aggregation tools performed with comparable accuracy for the selected 16 peptides as determined by Aggrescan 75% for the training set (**Figure S3**C) and 63% for the *de novo* predicted peptides (**Figure S3**D).

A

Aggregation prediction on training set

Tool	Accuracy	ROC
Aggrescan	76%	0.680
PATH	69%	0.696
APPNN	69%	0.619
PASTA 2.0	50%	0.512
Tango	43 %	0.437
Waltz	39%	0.487

B

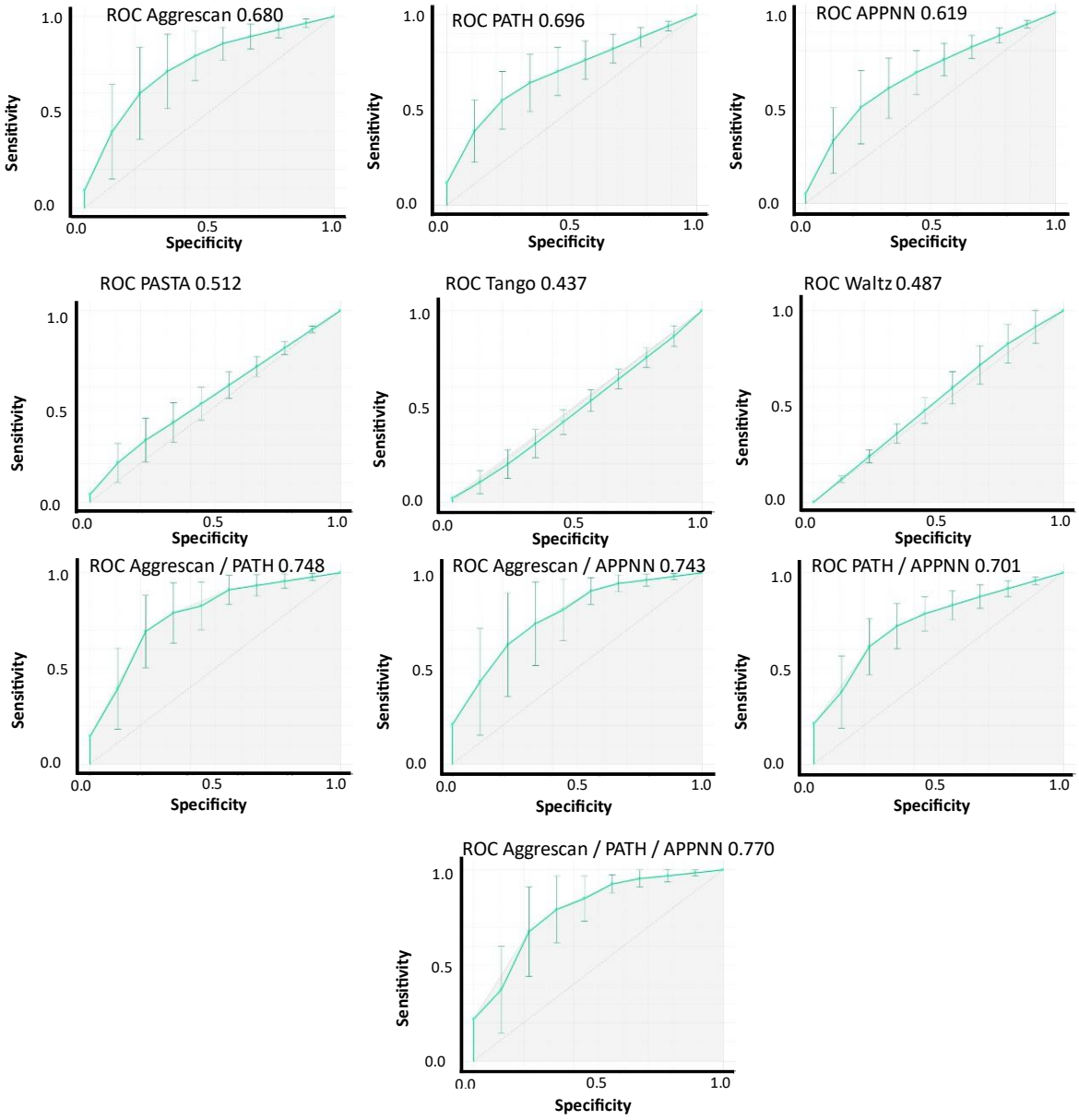
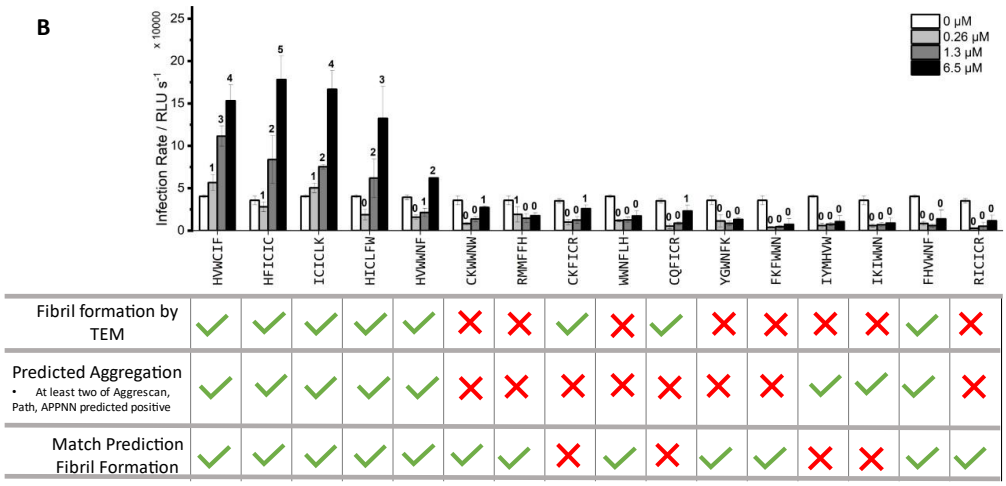


Figure S2 A Evaluation of the protein-aggregation tools Tango,[5] APPNN,[6] Waltz,[7] Path,[8] Aggrescan[9] and PASTA 2.0[10] with the training set (EF-C based library, **Table S1**). **B** ROC value for the prediction calculated with 10-fold stratified cross-validation and the experimental fibril formation as target value and a logistic regression learner and Lasso regularization model with 17 strength.

A

	Aggrescan	Path	APPNN	PASTA 2.0	Tango	Waltz	Predicted Infecticity Rel EF-C	Calculated hydrophobicity	Selection based on
HVWCIF	1	1	1	1	1	1	1.10	1.45	Highest predicted infectivity and aggregation without WWN motif
HFICIC	1	1	1	1	0	0	0.76	1.43	High N-gram similarity (0.42) with training set (HIHIQIC)
ICICLK	1	1	1	1	0	0	0.72	1.23	Cysteine rich sequence predicted for aggregation
HICLFW	1	1	1	1	0	1	0.69	1.53	Highest hydrophobicity and predicted for aggregation
HVWWNF	1	1	1	1	0	0	2.29	1.17	Highest predicted infectivity and aggregation
CKWWNW	1	0	0	0	0	0	0.86	1.12	High N-gram similarity (0.36) with training set (CKWKWQW)
RMMFFH	1	0	0	0	0	0	0.70	0.86	Low N-gram similarity (Avg 0.92, highest 0.66) with training set, not predicted for aggregation
CKFICR	1	0	0	0	0	0	0.82	0.78	High N-gram similarity (0.33) with training set (CKFQC)
WWNFLH	0	0	1	0	0	0	2.10	1.25	3rd highest infectivity prediction
CQFICR	1	0	0	0	0	0	0.96	0.91	High N-gram similarity (0.50) with training set (CQFQFQF)
YGWNFK	0	0	0	0	0	0	1.01	0.57	Low N-gram similarity (Avg 0.92, highest 0.72) with training set, not predicted for aggregation
FKFWWN	1	1	0	0	0	0	0.90	1.08	High N-gram similarity (0.56) with training set (KFKFQFNM)
IYMHVW	1	1	1	1	0	0	0.93	1.26	Low N-gram similarity (Avg 0.93, highest 0.78) with training set, predicted for aggregation
IKIWWN	1	1	1	1	0	0	0.90	1.09	High N-gram similarity (0.58) with training set (KIKIKIKIWWWW)
FHVWNF	1	1	1	1	0	0	1.09	1.10	3rd highest predicted infectivity and aggregation without WWN motif
RICICR	1	0	0	1	0	0	0.82	0.77	Moderate N-gram similarity (Avg 0.88, highest 0.66) with training set, not predicted for aggregation

B



C

Prediction accuracy **75%**
(at least two of Aggrescan, Path, APPNN predicted positive)

	Pred.	0	1	sum
actual	0	6	2	8
	1	2	6	8
sum		8	8	16

D

Prediction accuracy aggrescan only **63%**

	Pred.	0	1	sum
actual	0	2	6	8
	1	0	8	8
sum		2	14	16

Figure S3 Aggregation prediction tools applied on 16 de novo created peptides. **A** Summary of aggregation prediction results, predicted infectivity according to ProtVec *LASSO* model **Eqn. 1**, calculated hydrophobicity and comments on selection criteria. **B** Comparison of experimental and predicted aggregation. Experimental aggregation was determined by TEM fibril formation. 8 peptides were predicted for aggregation and 8 peptides were not predicted for aggregation by at least two of the tools Aggrescan, APPNN and PATH. **C** The accuracy of aggregation prediction by applying at least two prediction tools is determined 75 %. **D** Aggregation prediction accuracy for Aggrescan only calculated by confusion matrix for predicted peptides is 63 %.

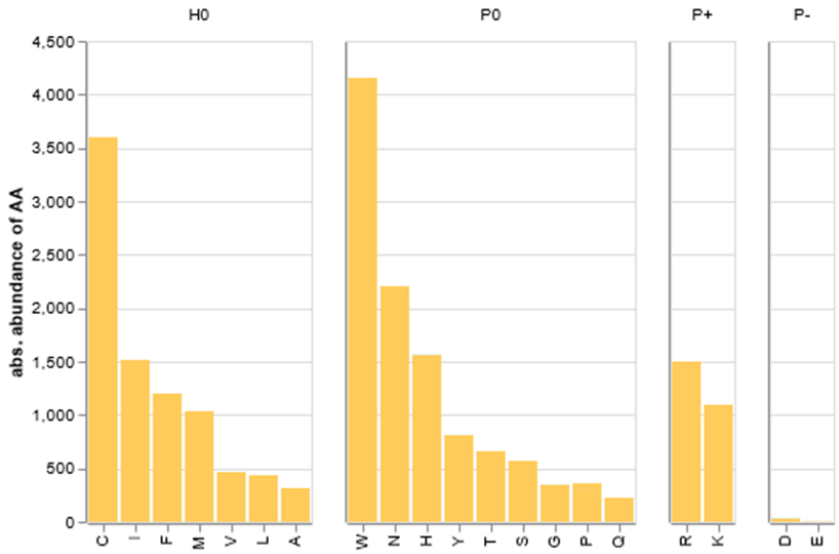


Figure S4 Absolute abundance of amino acids in net charge positive peptides (total 3669) predicted for infectivity enhancement. Cysteine (C) and Tryptophan (W) are the most prevalent amino acids.

3. N-gram similarity for predicted peptides with training set

The N-gram sequence similarity of the net charge positive peptides (total 3669) predicted for infectivity enhancement with the training set was calculated to ensure a diverse selection of peptides semantically close and far away from the training set. The N-gram similarity factor quantifies the similarity of two strings and returns 0 for the same sequence and 1 for sequences

Selected sequence	Average N-gram similarity with training set	Highest similarity	Corresponding sequence for highest similarity in the training set
HVWCIF	0.91	0.64	HIHIQIC
HFICIC	0.85	0.43	HIHIQIC
ICICLK	0.88	0.69	KIKIKIKI
HICLFW	0.88	0.64	HLHLPLL
HVWWNF	0.93	0.75	HGEHGE
CKWWNW	0.81	0.36	CKWKWQW
RMMFFH	0.93	0.67	MKFM
CKFICR	0.78	0.33	CKFQC
WWNFLH	0.94	0.80	NMWQKFKFQF
CQFICR	0.85	0.50	CKFQC
YGWNFK	0.93	0.73	KYKGAIIGNIK
FKFWWN	0.83	0.50	KFKFQFN
IYMHVW	0.93	0.79	KIKIQIW
IKIWWN	0.79	0.50	KIKIQIN
FHVWNF	0.92	0.75	KFKFQFNM
RICICR	0.88	0.67	KIKIQI

Figure S5 Overview of N-gram similarity values between the selected sequences and the training set. Average N-gram similarity describes the mean N-gram value between one selected sequence with every sequence of the training set. The highest similarity value shows the lowest value (highest similarity) for each selected sequence and the corresponding sequences from the training set. Values are colored gradually from red (0) – blue (1).

which have nothing in common. We applied the algorithm by Kondrak[12] for 2-grams with the python script shown in https://github.com/luozhouyang/python-string-similarity.git

In **Table S5** a matrix of all 3669 peptides N-gram similarity values with the training is listed. As shown in **Figure S5** the N-gram similarity values of the selected 16 peptides cover a wide range between 0.33 to 0.93 to quantify the diversity of selected sequences.

4. Evaluation of De Novo Peptide Activity with Property–Activity Model[4]

Three of the newly predicted peptides show unexpected activity despite negative Zeta-potential. To test whether these peptides follow a different mode of action a property activity model determined for the training set was applied on the *de novo* created peptides. The model established by multivariate analysis is shown in **Eqn. 3**.[4]

(**Eqn. 3**) Log Infect Rel Ef-C = -2.33462 + 0.02128*(Zeta-potential)+ 0.29879*(Log Count Rate) +0.27355*(fibril formation) + 0.26241 (Hydrophobicity) + 0.10744 (ThT-activity) + 0.00356 (β—sheet).

The peptides found from machine-learning are matching the model with a Pearson correlation coefficient of $R = 0.75$, which is comparable to the Pearson correlation coefficient of $R = 0.82$ found for the training set.[4] Noteworthy, the peptides which show at first glance unexpected behavior can be explained well with this model. According to this model the peptides HVWCIF, HFICIC, HICLFW are active due to their extraordinary high hydrophobicity and successive aggregation which outweighs the contributions of the moderately negative zeta-potential.

5. TEM micrographs

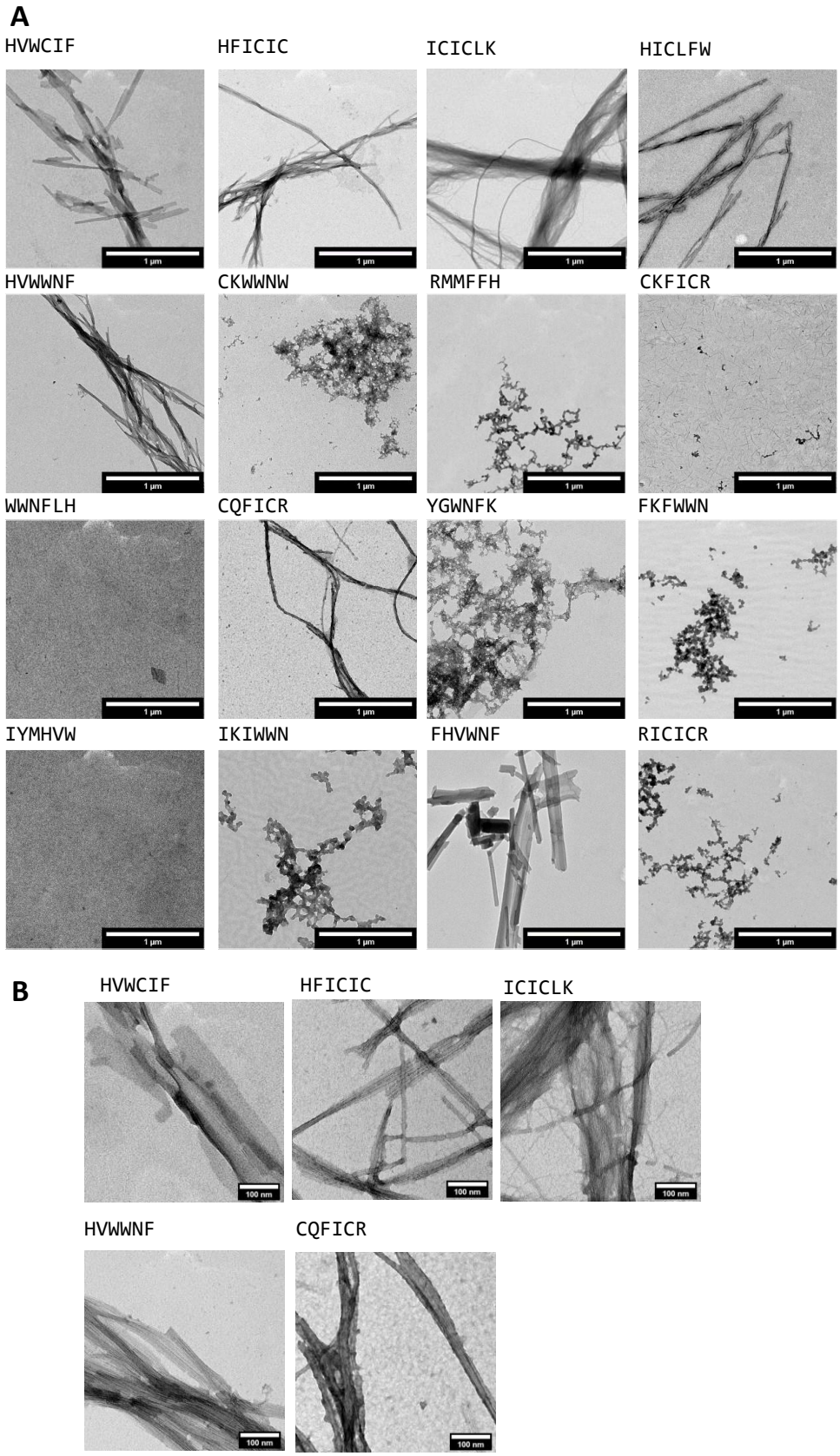


Figure S6 A TEM micrographs of *de novo* created peptides,1 mg/mL, PBS with 10% DMSO, incubated for 1d at RT, scale bar 1 µm. **B** Enlarged view on selected peptide fibrils of **A**, scale bar 100 nm.

6. Infectivity Data

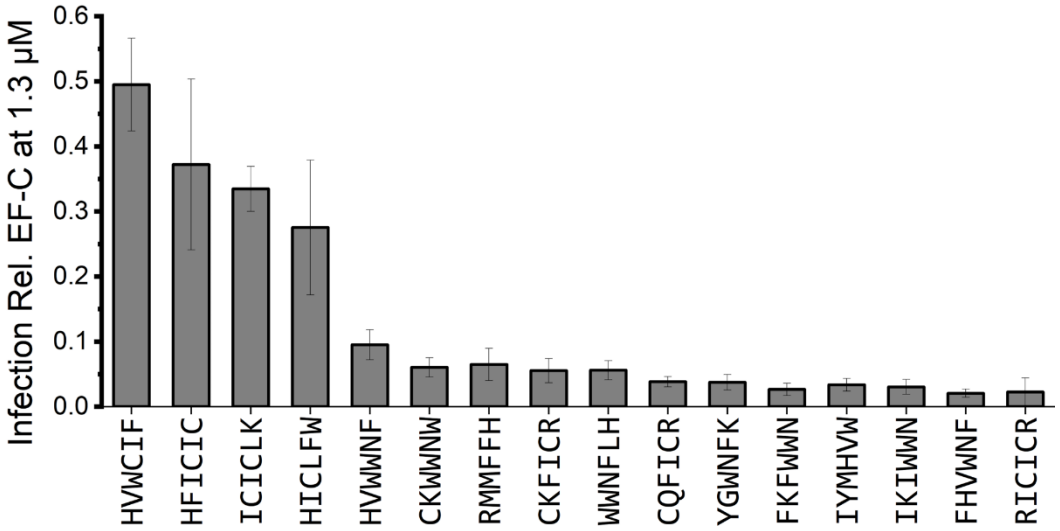


Figure S7 HIV-1 infection rates relative to EF-C (QCKIKQIINMWQ) at 1.3 µM concentration shown for the peptides from ML-prediction at 1.3 µM.

7. Cell-Viability

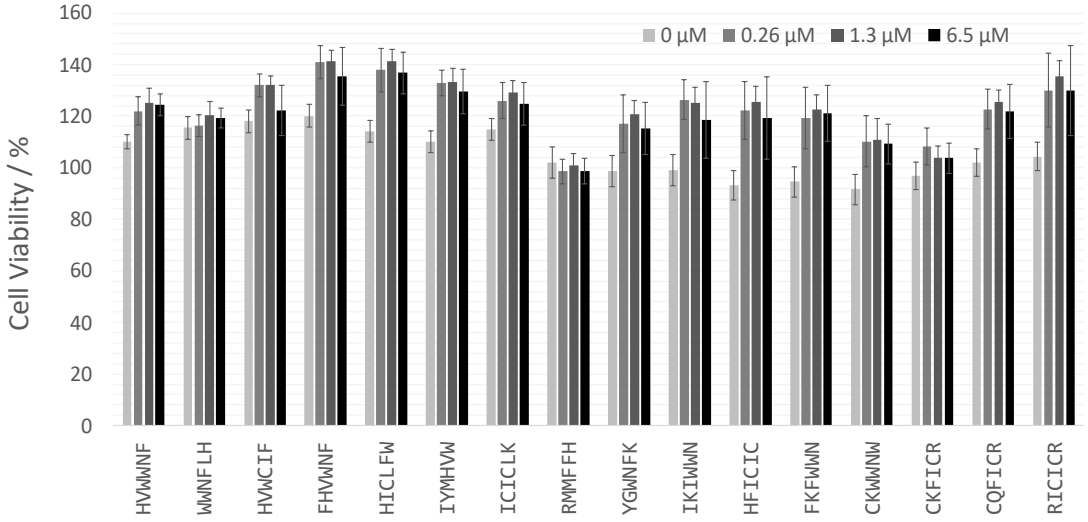


Figure S8 Cell viability normalized to 100% metabolic activity as determined by CellTiterGlo Assay of the 16 predicted peptides. Cell viability is maintained for all peptides at all tested concentrations (0.26 µM, 1.3 µM and 6.5 µM).

8. FT-IR Data

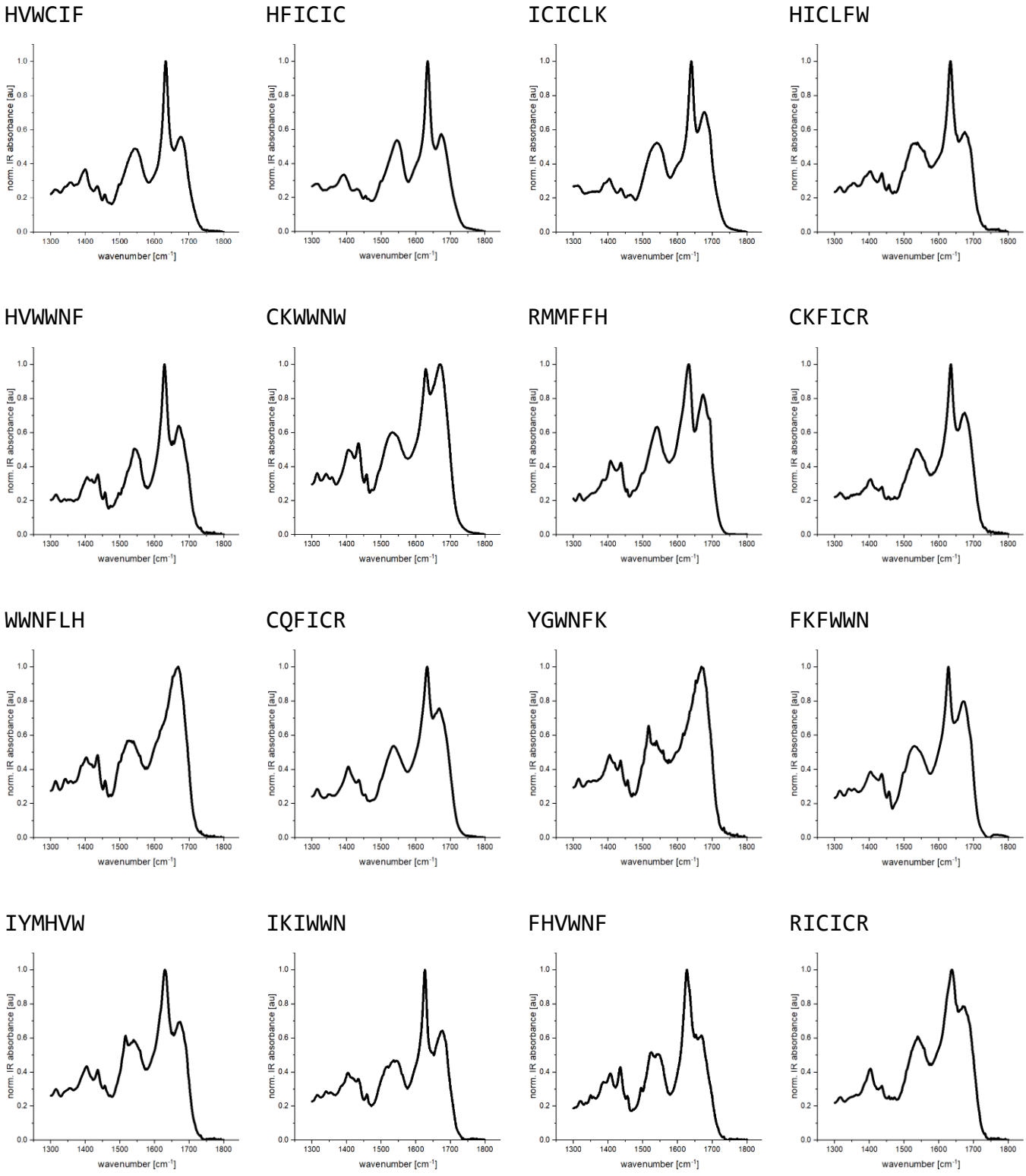


Figure S9 ATR FT-IR spectra of *de novo* created peptides.

9. Impact of Disulfide Bond Formation on Self-Assembly

Thiol groups from the side chain of cysteine can undergo disulfide bond formation with other thiol groups, which is known to influence self-assembly properties.[13]

To study the impact of disulfide bond formation of cysteine rich short peptides, we applied tris(2-carboxyethyl)phosphine (TCEP)[14] in 10 molar equivalents excess to break disulfide bonds, exemplarily studied for the peptide ICICLK. Transmission electron microscopy was performed to evaluate nanoscopic morphology, and brightfield microscopy was performed to evaluate microscopically large aggregation (Leica DMi8, 10x air objective). Surface charge and microscopic aggregation were evaluated via zeta-potential measurements.

Breaking disulfide bonds drastically changes the peptide assembly properties of ICICLK (**Figure S10**). Without disulfide bonds, no fibril formation (**Figure S10 A**) and no microscopic aggregation (**Figure S10 B, D**) can be observed, which also results in reduced surface charge (**Figure S10 C**).

Interestingly, for the original peptide EF-C in the training set, the addition of TCEP is has no visible influence on fibril formation and aggregation (**Figure S10 F**). This is likely due to the stabilizing effect of the alternating amphiphilic sequence pattern found in EF-C, which was identified earlier by us to drive assembly also without the presence of cysteine.[2,4]

Thus, we conclude that disulfide bond formation is a critical feature for self-assembly of the newly identified 6-mer peptides.

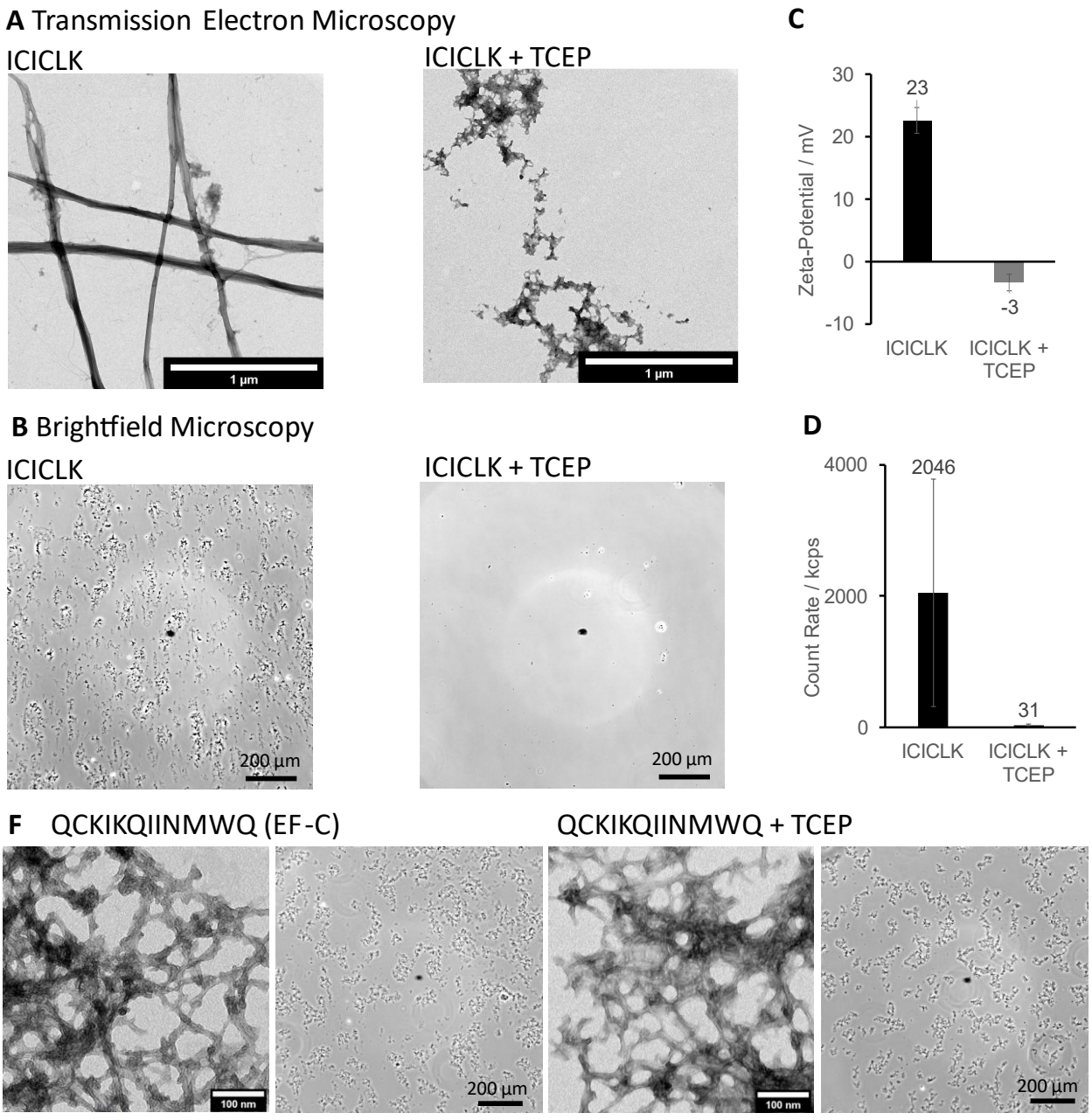


Figure S10 Effects of breaking disulfide bond on peptide self-assembly, aggregation, and surface charge of cysteine-rich peptide ICICLK. The peptide was incubated at room temperature for 1 day without and with 10 molar equivalents excess Tris(2-carboxyethyl) phosphine (TCEP). Then, the peptide was diluted from 10 mg/mL DMSO to 1 mg/mL in phosphate buffered saline, pH 7.4 (ICICLK) or in TCEP in phosphate buffered saline, pH 7.4. **A** Transmission electron microscopy micrographs depicting fibrillar (ICICLK) and non-fibrillar (ICICLK + TCEP) structures, scale bar 1 µm. **B** Brightfield microscopy images of peptide samples without (visible aggregates) and with TCEP (no visible aggregation), scale bar 200 µm. **C** Zeta-potential measurements and **D** derived count rate of scattered light of peptide samples without and with TCEP. **F** Transmission electron microscopy micrographs, scale bar 100 nm and brightfield microscopy images, scale bar 200 µm depicting fibrillar EF-C (QCKIKQIINMWQ) structures without (left) and with TCEP (right).

10. Amino acid Composition Analysis

To explore potentially common amino acid compositions between highly active peptides in the training set and the newly discovered active 6-mer peptides, we conducted a simplified, coarse-grained analysis. This analysis calculates the percentage of charged, hydrophobic, and hydrogen-bonding amino acids in peptides using the Hopp–Woods amino acid classification (**Figure S11 A, Code S1**).[15] Additionally, we coarse-grained the peptide activity into three thresholds: "high" (infectivity relative to EF-C > 0.7), "medium" (infectivity relative to EF-C > 0.1), and "low" (infectivity relative to EF-C < 0.1) active sequences.

The analysis of the training set revealed that peptides categorized as "high" and "medium" active contained a higher proportion of hydrophobic amino acids, while "low" activity peptides displayed a greater prevalence of charged amino acids (**Figure S11 B**). Remarkably, this same trend was observed in the *de novo* predicted peptides. The four active peptides, HVWCIF, HFICIC, ICICLK, and HICLFW, displayed a significantly higher content of hydrophobic amino acids compared to charged or hydrogen-bonding classified amino acids, which were predominantly found in non-active sequences (**Figure S11 C**).

It is worth noting that traditional prediction methods often rely on a predetermined set of descriptors. In contrast, the vector embedding approach employed in our study allows for the identification of underlying descriptors without any such assumptions. Therefore, a data-driven approach utilizing vector embeddings provides the flexibility to uncover latent descriptors that may have been not considered previously.

A Amino acid composition analysis

Coarse Grain

Amino Acids according to Hopp-Woods

Charged amino acids: E, D, R, K, H
Hydrogen bonding amino acids: S, T, E, D, K, R, H, N, Q, Y
Hydrophobic amino acids: W, V, I, L, F, M, C, A, G

Activity

High: Infection relative to EF-C > 0.7
Medium: Infection relative to EF-C > 0.1
Low: Infection relative to EF-C < 0.1

B Amino acid composition analysis for the **training set**

Activity category	Charged / %	Hydrogen bonding / %	Hydrophobic / %
High	24	47	53
Medium	25	48	52
Low	34	52	47

C Amino acid composition analysis for the **predicted peptides**

Sequence	Charged / %	Hydrogen bonding / %	Hydrophobic / %
HVWCIF	17	17	83
HFICIC	17	17	83
ICICLK	17	17	83
HICLFW	17	17	83
HVWWNF	17	33	67
CKWWNW	17	33	67
RMMFFH	33	33	67
CKFICR	33	33	67
WWNFLH	17	33	67
CQFICR	17	33	67
YGWNFK	17	50	50
FKFWWN	17	33	67
IYMHVW	17	33	67
IKIWWN	17	33	67
FHVWNF	17	33	67
RICICR	33	33	67

Figure S11 Comparison of the amino acid composition of the training set and the de novo predicted peptides. **A** To quantify the amino acid distribution, the amino acids and the activity were coarse-grained. The amino-acid compositions of a peptide were calculated by counting the number of each Hopp–Woods type amino acids (charged, hydrogen-bonding, or hydrophobic) in it and normalizing each count by the peptide's length (**Code S1**). **B** The peptides of the training set were categorized in high, medium, and low active sequences. High active peptides have on average a higher percentage of hydrophobic amino acids. **C** The active *de novo* predicted sequences (bold) have a higher percentage of hydrophobic amino acids compared to inactive sequences.

11. Training Set

Table S1 Binary representation of aggregation prediction results by protein-aggregation tools Tango,[5] APPNN,[6] Waltz,[7] PATH,[8] Aggrescan[9] and PASTA 2.0[10] for the training set. The experimental evaluation for fibril formation by TEM was reported previously by us.[4] Fibril formation is indicated by 1, no fibril formation is indicated by 0. The Log Infectivity enhancement relative to EF-C (QCKIKQIINMWQ) at 1.3 μM is retrieved from our previous report.[4]

sequence	Log Infect Rel. EF-C at 1.3 μM	Fibril formation by TEM	Tango	APPNN	Waltz	Path	Aggrescan	PASTA 2.0
CKIKIQIII	-0.01	1.0	0	1	1	1	1	1
CEIEIQI	-1.55	1.0	0	1	0	0	0	0
CKIKQIINM	-0.30	1.0	0	1	0	0	1	1
CKFKFQFNMWQ	0.06	1.0	0	1	0	0	1	0
CKFKFQFNM	0.09	1.0	0	0	0	0	1	0
CKFKFQF	0.09	1.0	0	0	0	0	1	0
KFKFQFNMW	-0.35	1.0	0	1	0	0	1	0
KFKFQFNM	-0.33	1.0	0	0	0	0	1	0
KFKFQFN	-1.74	1.0	0	0	0	0	1	0
CKAKAQANMWQ	-1.36	0.0	0	1	0	0	0	0
CKAKAQANM	-1.50	0.0	0	0	0	0	0	0
QCKFKQFFNMWQ	-0.61	1.0	1	1	0	0	1	0
QCKFKQFFNM	-0.37	1.0	0	0	0	0	1	0
QCKFKQFF	-1.91	1.0	0	0	0	0	0	0
CKFKQFFNMWQ	-0.21	1.0	1	1	0	0	1	0
CKFKQFFNM	-0.39	1.0	0	0	0	0	1	0
CKFKQFF	-2.17	0.0	0	0	0	0	0	0
QCKIKIQINM	-0.40	1.0	0	1	0	1	1	0
QCKAKAQANMWQ	-1.14	0.0	0	1	0	0	0	0
QCKAKAQANM	-1.83	0.0	0	0	0	0	0	0
QCKAKAQA	-1.60	0.0	0	0	0	0	0	0
QCKIKIQI	-0.95	1.0	0	0	0	0	1	0
QCKIKQIINM	-0.46	1.0	0	0	0	0	1	1
QCKIKQII	-2.19	1.0	0	1	0	0	0	0
QCKFKFQFNMWQ	-0.21	1.0	0	0	0	0	1	0
QCKFKFQFNM	-0.29	1.0	0	1	0	0	1	0
QCKFKFQF	-0.12	1.0	0	0	0	0	1	0
CRFRFQF	-0.46	1.0	0	0	0	0	0	0
HIHIQIC	-0.57	1.0	0	0	0	1	1	1
RLRLTLC	-1.29	1.0	0	1	0	1	1	0
HLHLPLL	-2.00	0.0	0	0	0	0	0	0
RGECKFKFQF	-0.48	1.0	0	0	0	0	1	0
RGEKIKIQINM	-0.61	1.0	0	1	0	1	1	0
KYKGAIIGNIK	-2.51	0.0	1	1	0	0	1	1
HGDKCHGDKC	-1.99	0.0	0	0	0	0	0	0
RPRGLLLGNLR	-1.32	0.0	0	1	0	0	1	0
KKFQKKFQ	-1.54	0.0	0	0	0	0	0	0
PPFHPPPFHP	-1.75	0.0	0	0	0	0	0	0
MDQMDQMDQMDQMDQ	-2.37	0.0	0	0	0	0	0	0
FDPFDPFDP	-1.63	0.0	0	0	0	0	0	0
TKTLTKTL	-1.67	0.0	0	0	0	0	0	0
FKFDKFKFDK	-1.33	0.0	0	0	0	0	0	0
KVKGVGK	-1.59	0.0	0	0	0	0	0	0
SISISRRI	-1.55	0.0	0	0	0	0	0	0
HRRHFRHKITKKK	-1.87	0.0	0	0	0	0	0	0
KNERIKNERI	-1.89	0.0	0	0	0	0	0	0
KIRGKFEKED	-1.24	0.0	0	0	0	0	0	0
CKFQC	-1.64	0.0	0	0	0	0	0	0
MKFM	-1.76	0.0	0	0	0	0	0	0
CKFC	-1.96	0.0	0	0	0	0	0	0
RGDKIRGDKI	-2.13	0.0	0	0	0	0	0	0

KNDKND	-1.98	0.0	0	0	0	0	0	0
HGEHGE	-2.19	0.0	0	0	0	0	0	0
HGEHGEHGE	-1.72	0.0	0	0	0	0	0	0
CRFRVPF	-1.70	0.0	0	0	0	0	0	0
CHLHLQL	-1.01	0.0	0	1	0	0	0	0
CETMYDKILKNLSRSR	-1.72	0.0	0	0	0	0	0	0
MKFKFQF	-1.20	1.0	0	0	0	0	1	0
QCKIKQIINMWQ	0.00	1.0	1	1	0	0	1	1
QCKIKIQINMWQ	0.02	1.0	0	1	0	1	1	1
KIKIQINMWQ	-0.53	1.0	0	1	0	1	1	1
NMWQKAKAQA	-1.38	1.0	0	0	0	0	0	0
NMWQKFKFQF	-0.57	1.0	0	0	0	0	1	0
KVKVKVQV	-1.76	0.0	0	0	0	0	1	0
KIKQIINMWQ	-0.46	1.0	1	1	0	0	1	1
KAKAQANMWQ	-2.15	0.0	0	1	0	0	0	0
KFKFQFNMWQ	-0.12	1.0	0	1	0	0	1	0
KFKQFFNMWQ	-1.05	0.0	1	1	0	0	1	0
KAKQAANMWQ	-1.51	0.0	0	1	0	0	0	0
NMWQKVGTPL	-1.61	0.0	0	0	0	0	0	0
NMWQKIKQII	-1.60	0.0	0	0	0	0	0	0
NMWQKIKIQI	-0.69	1.0	0	0	0	0	1	0
KIKQIINM	-0.35	1.0	0	1	0	0	1	1
KIKQIIN	-1.40	0.0	0	1	0	0	1	0
KIKIQINMW	0.33	1.0	0	1	0	1	1	1
KIKIQINM	-0.44	1.0	0	1	0	1	1	0
KIKIQIN	-0.58	1.0	0	1	0	1	1	0
KAKAQA	-1.74	0.0	0	0	0	0	0	0
KIKIQI	-1.17	0.0	0	0	0	0	1	0
KFKFQF	-1.21	0.0	0	0	0	0	1	0
KIKIQINMWA	-0.16	1.0	0	1	0	1	1	1
EIEIEIEI	-1.75	1.0	0	1	0	0	0	0
KIKIQINMAQ	-0.67	1.0	0	1	0	1	1	0
KIKIQINAWQ	-0.27	1.0	0	1	0	1	1	0
KIKIQIAMWQ	-0.41	1.0	1	1	0	1	1	0
KIKIKIKIYYYY	0.14	0.0	1	1	1	1	1	1
HHHHEIEIEIEI	-1.53	0.0	0	1	0	0	0	0
KYKYQY	-1.99	0.0	0	0	0	0	0	0
EFEFQF	-2.49	0.0	0	0	0	0	0	0
CEFEFQF	-2.28	1.0	0	0	0	0	0	0
CKFKFQFNMW	-0.02	1.0	0	1	0	0	1	0
CKYKYQY	-1.39	0.0	0	0	0	0	0	0
KWKWQW	-1.84	0.0	0	0	0	0	0	0
CKWKWQW	-0.60	1.0	0	0	0	0	0	0
KLKQLL	-2.13	0.0	0	0	0	0	0	0
CKIKIQINMWQ	-0.16	1.0	0	1	0	1	1	1
RGDKIKIQI	-1.81	1.0	0	0	0	0	1	0
CKIKIQINM	0.20	1.0	0	1	0	1	1	0
CKIKIQI	0.01	1.0	0	0	0	0	1	0
CKIKQII	-1.38	1.0	0	0	0	0	0	0
CKIKQIINMWQ	0.09	1.0	1	1	0	0	1	1
KIKIQIRGD	-1.68	1.0	0	0	0	1	1	0
RGDKIKIQIC	-0.08	1.0	0	1	0	1	1	0
RGDKIKIQINM	-0.38	1.0	0	1	0	1	1	0
RGDKIKIQINMWQ	-0.12	1.0	0	1	0	1	1	1
KFKFEFEF	-1.55	1.0	0	0	0	0	0	0
KIKQII	-2.31	0.0	0	0	0	0	0	0
KIEIQINM	-1.46	1.0	0	1	0	1	0	0
CKIKIQIRGD	-0.48	1.0	0	0	0	1	1	0
CKIKQIIRGD	-1.31	0.0	0	0	0	0	1	0
RGDKIKIQINMC	0.03	1.0	0	1	0	1	1	0
CQFQFQF	-0.45	1.0	0	1	0	0	0	0
KIKIQII	-1.73	1.0	0	1	0	1	1	1
KFKFQFFF	-0.39	1.0	0	0	1	1	1	0
KIKIKIQI	-1.01	1.0	0	0	0	1	1	0
KAKAKAQA	-2.06	0.0	0	0	0	0	0	0
KLKLKLQL	-1.01	1.0	0	0	0	0	0	0
KFKFKFQF	-0.10	1.0	0	0	0	0	1	0
KVKVQVV	-2.04	0.0	0	1	0	0	1	1
KLKLQLL	-2.16	1.0	0	0	0	1	0	0

KFKFQFF	-0.91	1.0	0	0	0	1	1	0
KIKIQIII	-0.45	1.0	0	1	1	1	1	1
KVKVQVVV	-1.46	1.0	0	1	0	0	1	1
HHHHKAKAKAKAYYYY	-1.88	0.0	0	0	1	0	1	0
KIKIQIC	-0.22	1.0	0	1	0	1	1	0
RGDSKIKIQIC	-0.22	1.0	0	1	0	1	1	0
RGDSGGGGGGKIKIQIC	-0.06	1.0	0	1	0	1	1	0
ILKNLSRSRKIKIQIC	-0.47	1.0	0	1	0	1	1	0
KIKIKIKIWWWW	-0.23	0.0	0	1	1	1	1	1
KIKIKIKI	-1.49	0.0	0	0	0	1	1	0
CEFEFQFNMWQ	-0.85	1.0	0	1	0	1	0	0
CSISIQI	-1.25	1.0	0	1	0	0	1	0
CEIEIQINMWQ	-0.66	1.0	0	1	0	1	0	0
CEIEIQINM	-1.19	1.0	0	1	0	1	0	0
CSISIQINM	-0.70	1.0	1	1	0	1	1	0
KAKAQANM	-2.05	0.0	0	0	0	0	0	0
CKAKAQA	-1.50	0.0	0	0	0	0	0	0
HHHHKIKIQINMYYYY	0.04	1.0	1	1	1	1	1	1
KIKIKIKIWW	-0.43	0.0	0	0	0	1	1	1
HHHHKIKIKIKIWWWW	-0.42	0.0	0	1	1	1	1	1
MKIKIQINM	-0.56	1.0	0	1	0	1	1	0
RGDCKIKIQINM	-0.63	1.0	0	1	0	1	1	0
MKIKIQINMWQ	-0.38	1.0	0	1	0	1	1	1
HHHHKIKIKIKIYYYY	-0.17	0.0	1	1	1	1	1	1
RGDCKFKFQF	-0.82	1.0	0	0	0	0	1	0
KIKIQIW	-0.85	1.0	0	1	0	1	1	1
KFKFQFW	-0.79	1.0	0	0	0	1	1	0
CKFKFQFW	-0.06	1.0	0	0	0	1	1	0
EIKIQINM	-1.00	1.0	0	1	0	1	0	0
IKVAVKIKIQINM	-0.30	1.0	0	1	0	1	1	1
HHHHKAKAKAKAWWWW	-1.05	0.0	0	0	1	0	1	0
CKIW	-1.29	0.0	0	0	0	0	0	0
CKIKIQINMW	0.11	1.0	0	1	0	1	1	1
ILKNLSRSRKIKIQINMWQ	-0.49	1.0	0	1	0	1	1	1
EIEIQINM	-1.44	1.0	0	1	0	1	0	0
KIKIQINMWWQ	0.51	0.0	1	1	0	1	1	1
KIKIQINMWWWQ	0.44	0.0	1	1	1	1	1	1
KFKQFFINMWQ	-0.04	1.0	1	1	1	1	1	1
KIKIQIMWNQ	-0.25	1.0	1	1	0	1	1	1
KIKIQIMQWN	-0.15	1.0	0	1	0	1	1	0
KIKIQINMWQRGD	-1.92	0.0	0	1	0	1	1	1
EIEIQINMWQ	-0.88	0.0	0	1	0	1	0	0
KIKIKIQINMWQ	-0.10	1.0	0	1	0	1	1	1

12. Predicted Peptides Characterization

Table S2 Physicochemical characterization of the selected *de novo* peptides. Shown are the absolute infection rates (Abs. Infect) at 6.5, 1.3, 0.26 and 0 μM peptide concentration. The logarithmic infection relative to EF-C (Log Infect Rel EF-C) refers to 1.3 μM concentration. Fibril formation is determined by transmission electron microscopy (**Figure S6**) and β-sheet content is determined by ATR-FT-IR spectroscopy (**Figure S9**). Standard deviations (Std Dev) are determined by triplicate measurements.

	Abs Infect [RLU/s] at 6.5 μM	Std Dev.	Abs Infect [RLU/s] at 1.3 μM	Std Dev.	Abs Infect [RLU/s] at 0.26 μM	Std Dev.	Abs Infect [RLU/s] at 0 μM	Std Dev.	Infect. Rel-to EFC 1.3 μM	Std Dev.	Log Infect Rel to EFC	Std Dev.	Zeta-Pot.	Std Dev.	Count Rate / kcps	Count Rate Std	Fibril	Hydro-phobicity	ThT active	β sheet [%]
HVWCIF	2E+05	18915	111453	12016	56677	9587	40407	1946	0.55	0.08	-0.26	0.06	-9.41	0.40	26175	7913	1	1.46	1	44
HFICIC	2E+05	27604	83800	28416	28200	5615	35653	5134	0.42	0.15	-0.38	0.15	-1.91	0.22	25360	7713	1	1.43	1	54
ICICLK	2E+05	21794	75363	2873	50423	5753	40407	1946	0.37	0.04	-0.43	0.04	33.43	1.20	3220	1783	1	1.23	0	38
HICLFW	1E+05	37786	61983	22561	18680	6151	40407	1946	0.31	0.12	-0.51	0.16	-9.02	0.67	10024	3381	1	1.54	1	44
HVWWNF	62177	790.5	21423	4777	15823	2689	39163	2823	0.11	0.03	-0.97	0.10	3.64	0.23	2219	758	1	1.17	1	37
CKWWNW	27287	1426	13610	3071	8167	1749	35653	5134	0.07	0.02	-1.17	0.10	-1.83	0.82	2453	2899	0	1.12	1	26
RMMFFH	17607	3346	14603	5405	19163	8998	35653	5134	0.07	0.03	-1.14	0.16	-1.82	0.40	21	9	0	0.86	0	33
CKFICR	25873	4155	12463	3991	9960	3079	35190	2493	0.06	0.02	-1.21	0.14	4.97	0.65	11	11	1	0.78	0	37
WWNFLH	17167	6199	12610	3089	11870	1369	40407	1946	0.06	0.02	-1.20	0.11	-6.01	0.70	18	3	0	1.25	0	2
CQFICR	23253	6640	8617	1592	5157	2902	35190	2493	0.04	0.01	-1.37	0.09	20.53	0.60	586	136	1	0.91	1	33
YGWNFK	13143	2312	8433	2530	11340	7237	35653	5134	0.04	0.01	-1.38	0.14	-3.06	2.50	56	25	0	0.57	0	6
FKFWWN	7317	6784	4607	1344	3757	983	35653	5134	0.02	0.01	-1.64	0.13	-4.89	0.26	244	159	0	1.08	0	32
IYMHVW	10627	6870	7553	2037	5907	2901	40407	1946	0.04	0.01	-1.43	0.12	5.16	1.42	21	5	0	1.27	0	40
IKIWWN	8810	6012	6837	2474	6010	2049	35653	5134	0.03	0.01	-1.47	0.16	-9.90	1.26	171	57	0	1.09	0	37
FHVWNF	13743	10451	6007	2070	8390	2290	40407	1946	0.03	0.01	-1.52	0.15	-16.53	0.68	1427	931	1	1.10	0	49
RICICR	11550	6543	5030	4903	2860	163	35190	2493	0.03	0.02	-1.60	0.42	4.52	0.28	41	18	0	0.78	0	47

Table S3 contains information on top 12320 sequences from Monte Carlo ProtVec LASSO model screening with information on predicted infectivity, hydrophobicity, and net charge and is openly available at the following data repository DOI: 10.5281/zenodo.7708290

Table S4 contains information on top 3669 peptides with a net positive charge with information on aggregation prediction results from Aggrescan, APPNN, and PATH and is openly available at the following data repository DOI: 10.5281/zenodo.7708290

Table S5 contains information on N-gram similarity matrix composed of top 3669 peptides and 163 peptides from the training set and is openly available at the following data repository DOI: 10.5281/zenodo.7708290

Code S1 is a python script for calculating the amino acids composition of charged, hydrogen bonding, and hydrophobic amino acids in a peptide sequence library according to Hopp–Woods classification.[15] The code and the corresponding training set of coarse-grained peptides are openly available at the following data repository DOI: 10.5281/zenodo.8004720

13. References

1 M. Yolamanova, C. Meier, A. K. Shaytan, V. Vas, C. W. Bertoncini, F. Arnold, O. Zirafi, S. M. Usmani, J. A. Müller, D. Sauter, C. Goffinet, D. Palesch, P. Walther, N. R. Roan, H. Geiger, O. Lunov, T. Simmet, J. Bohne, H. Schrezenmeier, K. Schwarz, L. Ständker, W.-G. Forssmann, X. Salvatella, P. G. Khalatur, A. R. Khokhlov, T. P. J. Knowles, T. Weil, F. Kirchhoff and J. Münch, *Nat. Nanotechnol.*, 2013, **8**, 130–136.

2 S. Sieste, T. Mack, E. Lump, M. Hayn, D. Schütz, A. Röcker, C. Meier, K. Kaygisiz, F. Kirchhoff, T. P. J. Knowles, F. S. Ruggeri, C. V. Synatschke, J. Münch and T. Weil, *Adv. Funct. Mater.*, 2021, **31**, 2009382.

3 S. Kirti, K. Patel, S. Das, P. Shrimali, S. Samanta, R. Kumar, D. Chatterjee, D. Ghosh, A. Kumar, P. Tayalia and S. K. Maji, *ACS Biomater. Sci. Eng.*, 2019, **5**, 126–138.

4 K. Kaygisiz, L. Rauch-Wirth, A. Dutta, X. Yu, Y. Nagata, T. Bereau, J. Münch, C. V. Synatschke and T. Weil, *ChemRxiv*, 2023, https://doi.org/10.26434/chemrxiv-2023-hfqxb.

5 A.-M. Fernandez-Escamilla, F. Rousseau, J. Schymkowitz and L. Serrano, *Nat. Biotechnol.*, 2004, **22**, 1302–1306.

6 C. Família, S. R. Dennison, A. Quintas and D. A. Phoenix, *PLoS One*, 2015, **10**, e0134679.

7 J. Beerten, J. Van Durme, R. Gallardo, E. Capriotti, L. Serpell, F. Rousseau and J. Schymkowitz, *Bioinformatics*, 2015, **31**, 1698–1700.

8 J. W. Wojciechowski and M. Kotulska, *Sci. Rep.*, 2020, **10**, 7721.

9 O. Conchillo-Solé, N. S. de Groot, F. X. Avilés, J. Vendrell, X. Daura and S. Ventura, *BMC Bioinformatics*, 2007, **8**, 65.

10 I. Walsh, F. Seno, S. C. E. Tosatto and A. Trovato, *Nucleic Acids Res.*, 2014, **42**, W301–W307.

11 J. Demšar, A. Erjavec, T. Hočevar, M. Milutinovič, M. Možina, M. Toplak, L. Umek, J. Zbontar and B. Zupan, *J. Mach. Learn. Res.*, 2013, **14**, 2349–2353.

12 G. Kondrak, *SPIRE*, 2005, **3772**, 115–126.

13 D. Komáromy, M. C. A. Stuart, G. Monreal Santiago, M. Tezcan, V. V. Krasnikov and S. Otto, *J. Am. Chem. Soc.*, 2017, **139**, 6234–6241.

14 J. A. Burns, J. C. Butler, J. Moran and G. M. Whitesides, *J. Org. Chem.*, 1991, **56**, 2648–2650.

15 T. P. Hopp and K. R. Woods, *Proc. Natl. Acad. Sci. U. S. A.*, 1981, **78**, 3824–8.