Biomaterials Science

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Check for updates

Cite this: *Biomater. Sci.*, 2023, **11**, 5251

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

Kübra Kaygisiz, [a] Arghya Dutta, ‡[b] Lena Rauch-Wirth,[c] Christopher V. Synatschke, [a] Jan Münch, [c] Tristan Bereau *§[b] and Tanja Weil *[a]

Amyloid-like nanofibers from self-assembling peptides can promote viral gene transfer for therapeutic applications. Traditionally, new sequences are discovered either from screening large libraries or by creating derivatives of known active peptides. However, the discovery of *de novo* peptides, which are sequence-wise not related to any known active peptides, is limited by the difficulty to rationally predict structure–activity relationships because their activities typically have multi-scale and multi-parameter dependencies. Here, we used a small library of 163 peptides as a training set to predict *de novo* sequences for viral infectivity enhancement using a machine learning (ML) approach based on natural language processing. Specifically, we trained an ML model using continuous vector representations of the peptides, which were previously shown to retain relevant information embedded in the sequences. We used the trained ML model to sample the sequence space of peptides with 6 amino acids to identify promising candidates. These 6-mers were then further screened for charge and aggregation propensity. The resulting 16 new 6-mers were tested and found to be active with a 25% hit rate. Strikingly, these *de novo* sequences are the shortest active peptides for infectivity enhancement reported so far and show no sequence relation to the training set. Moreover, by screening the sequence space, we discovered the first hydrophobic peptide fibrils with a moderately negative surface charge that can enhance infectivity. Hence, this ML strategy is a time- and cost-efficient way for expanding the sequence space of short functional self-assembling peptides exemplified for therapeutic viral gene delivery.

Received 8th March 2023,
Accepted 14th June 2023

DOI: 10.1039/d3bm00412k

rsc.li/biomaterials-science

Open Access Article. Published on 15 June 2023. Downloaded on 8/4/2025 8:11:42 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

1 Introduction

New peptides to increase viral transduction: a multi-parameter challenge

Short self-assembling peptides have attracted much interest as functional materials in recent years since they can be designed in a precise and cost-efficient way.[1] However, a small change in sequence can drastically change physicochemical properties on multiple length scales, which makes it challenging to rationally design *de novo* self-assembling peptide sequences with a desired bioactivity.[2]

For example, the emerging field of gene therapy requires efficient transduction of target cells by viral vectors that deliver the therapeutic gene.[3] In this regard, self-assembled peptide fibrils that increase the colocalization of viral vectors and cellular membranes and thereby enhance gene delivery are promising candidates for new or optimized gene-therapeutic applications.[4–6] However, the discovery of self-assembling peptides as enhancers of viral transduction *via* screening methods is challenging because of the complexity in predicting sequences that show the required physicochemical properties such as sequence amphiphilicity, charge, and assembly, for biological activity.[7,8]

Traditionally, new peptides with certain desired properties, *e.g.*, viral transduction/infection enhancement, have been found mainly by serendipity during screening processes,[4,9] or by nature-inspired rational design.[7,10,11] A common strategy is

[a]*Department Synthesis of Macromolecules, Max Planck Institute for Polymer Research, Ackermannweg 10, 55128 Mainz, Germany. E-mail: weil@mpip-mainz.mpg.de*
[b]*Polymer Theory, Max Planck Institute for Polymer Research, Ackermannweg 10, 55128 Mainz, Germany. E-mail: bereau@thphys.uni-heidelberg.de*
[c]*Institute of Molecular Virology, Ulm University Medical Center, Meyerhofstraße 1, 89081 Ulm, Germany*

†Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d3bm00412k

‡Present address: Institute of Biochemistry II, Faculty of Medicine, Goethe University, Theodor–Stern–Kai 7, 60590 Frankfurt, Germany.
§Present address: Institute for Theoretical Physics, Heidelberg University, Philosophenweg 19, 69120 Heidelberg, Germany.

Open Access Article. Published on 15 June 2023. Downloaded on 8/4/2025 8:11:42 PM. This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

finding recurring motifs in known active peptides to generate new sequences.[12,13] Although changing one amino acid at a time to screen derivatives of a known active compound is often a direct and efficient way to find active peptides with similar structure and study property–activity relationship, it is not a tractable way to discover new sequences. The sequence space of peptides is huge—there are 20^6 = 64 million peptides if we only consider all possible 6-mers composed of 20 canonical amino acids. Further, the size of the sequence space increases exponentially with the number of residues. Consequently, exploring the peptide space to discover new peptides by creating randomly generated peptides or derivatives of known structures and studying them in experiments quickly becomes unfeasible. In the quest to discover new peptides with certain target bioactivity, computational methods have been established for fast and inexpensive prescreening of peptide sequences.[14]

Navigating through sequence space *via* machine learning

Machine learning (ML) can be applied to bridge the limited experimental dataset with the vast compound space by analyzing the sequence–activity relationships to reverse engineer novel non-intuitive compounds.[15–17] ML models are often used to study structure–activity relationships since they can mitigate the curse of dimensionality: peptides with similar values of a target property can be far apart in sequence space but close in some higher-dimensional underlying feature space.[18] Various supervised and unsupervised ML algorithms have been used, for example, to predict liquid–liquid phase separating protein sequences[19] and to detect antimicrobial[20] or cell-penentrating[21] peptide sequences. Recently, an ML approach, in combination with Monte Carlo tree search and molecular dynamics simulations, was reported for predicting unexpected *de novo* β-sheet rich self-assembling peptides.[22] While all these reports successfully identified either bioactive or self-assembling peptides separately, they did not predict bioactive self-assembling peptides.

Inverse design of sequences *via* continuous vector embeddings

One way of training an ML model that can infer structure–activity relationships of bioactive self-assembling peptides is to use the minimal available information as input to the model—the peptide's sequence. Continuous vector representation is the method of choice to represent complex sequence order and composition of proteins and peptides for training ML algorithms. For instance, the word embedding model Word2Vec,[23,24] that was originally developed as a natural language processing tool, can be applied to extract structural concepts to encode latent material property information across various datasets.[25]

In this direction, Asgari and Mofrad recently proposed a method that can convert any protein sequence into a unique, dense, 100-dimensional numerical vector, termed ProtVec.[26] They used a method from natural language processing that employs an artificial neural network that, while attempting to determine the context in which a word is most likely to occur in a sentence, generates a continuous distributed representation of the word; further, the ProtVecs were found to accurately capture the physicochemical properties of proteins.

Here, we report a three-step approach to explore the sequence space of bioactive self-assembling peptides and discover *de novo* sequences with high bioactivity by only using sequence and activity information. First, we trained a LASSO (Least Absolute Shrinkage and Selection Operator) regression model using ProtVec representations of peptides from a relatively small library of 163 sequences with known activity values. LASSO aims at identifying a minimal subset of parameters relevant to the prediction, thereby enhancing explainability. Then, we utilized the trained model to systematically sample the sequence space of 6-mer peptides using a Monte Carlo approach.[27] Finally, we screened the sequences with highest predicted activities based on their charge and tendency to aggregate. The search yielded 16 new peptides, which were tested experimentally and found to be active with a 25% hit rate or a 50% hit rate if further predictive parameters like aggregation are included. Strikingly, the newly created peptides are very different in sequence from the ones comprising the training set and shorter than any previously reported sequence for infectivity enhancement. Taken together, our method offers a fast and computationally inexpensive way of predicting and screening potentially bioactive, self-assembling peptides for any desired target bioactivity; consequently, it can accelerate peptide screening in the early stages of research.

2 Results

To gain insight into sequence elements and predict *de novo* sequences, which enhance retroviral transduction, we exploited an already reported peptide library for viral transduction enhancement consisting of 163 sequences. This library was based on the self-assembling 12-mer peptide EF-C (QCKIKQIINMWQ) corresponding to residues 417–428 of the HIV envelope protein gp120 and was created by systematic peptide sequence alterations (Fig. 1A).[4,7,8] The rational design of the library results in similar sequences (Fig. 1B), which show a wide range of activity (Fig. 1C). We decided to use this peptide library as a training set to demonstrate the usefulness of ML approaches because this data set represents a common situation in the early stage of research, where one wants to explore the peptide sequence space starting from a relatively small library consisting of derivatives of a known active compound.

Exploration of sequence space

The ML approach we used is summarized in Fig. 1D. The self-assembling bioactive peptides that were used to train the ML model were previously reported by us.[8] As described in the introduction, we represented the peptide sequences as dense, 100-dimensional numerical vectors, termed as ProtVecs, using a continuous distributed representation proposed by Asgari and Mofrad.[26] One of the advantages in using continuous dis-

 This journal is © The Royal Society of Chemistry 2023

Open Access Article. Published on 15 June 2023. Downloaded on 8/4/2025 8:11:42 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

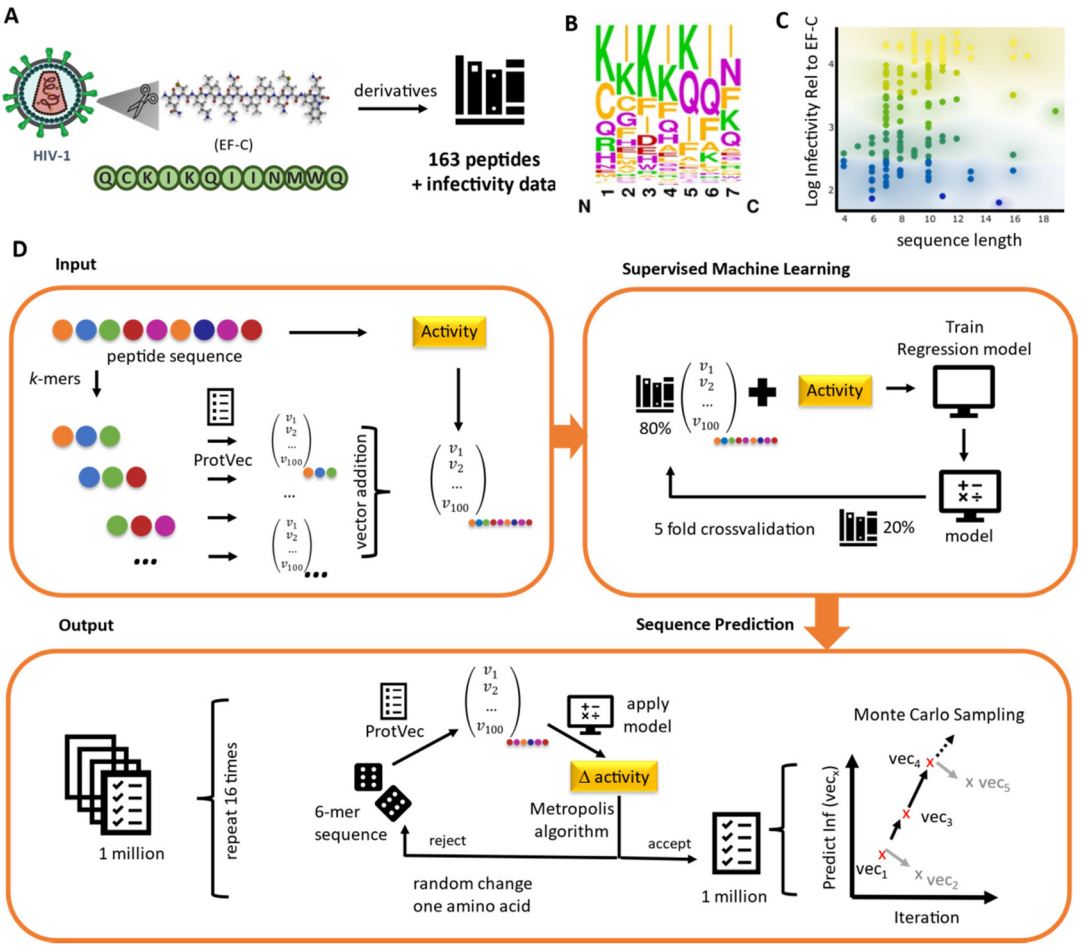

Fig. 1 Schematic overview of the workflow. (A) A peptide library[8] was created by systematic sequence derivations of the infectivity-enhancing peptide EF-C. This database contains 163 peptide sequences and their respective biological activity data (*i.e.* enhancement of HIV-1 infection, "infectivity". (B) The sequence logo plot summarizes the amino acid frequency in the library and shows that most sequences are composed of positive charged (K, lysine) and hydrophobic uncharged (F, phenylalanine or I, isoleucine) amino acids in alternating order. (C) The diagram shows the length distribution of infectivity enhancement in log scale relative to the reference peptide EF-C. Despite the sequence similarity, a wide range of activities can be found within the library. (D) Flowchart summarizing the machine-learning approach in this study. As input data, the EF-C based peptide library (A) was used. Each peptide sequence in the library was represented as a 100-dimensional vector using a continuous distributed representation, ProtVec.[26] The vector and respective activity were used in a supervised LASSO regression model to compute a linear relationship between the activity and relevant components of the 100-d ProtVec. The trained model was then used in a Monte Carlo search where 1 million 6-mers were generated, and the 6-mers that were predicted to perform better than EF-C were retained based on a Metropolis criterion (see Methods). By repeating the Monte Carlo step 16 times, each time starting with a random 6-mer, a broad portion of the 6-mer sequence space was covered.

tributed representation is that it captures underlying bioinformatic properties and information on distance relationships of neighboring amino acids and uniformly casts any arbitrary-sized peptide to a unique numerical vector with a fixed size (*e.g.*, 100-d ProtVecs), which can be readily used in numerical computations.[14,26] We constructed ProtVecs for training set peptide sequences by dividing the sequences into 3-grams (3 consecutive amino acids) and summing up the ProtVecs representing those 3-grams[26] by linear vector addition. In that way, 100-d vectors are created for each sequence, which can be assigned to an experimentally determined activity. The 100-d vectors in combination with the bioactivity data were then used to train a LASSO linear regression ML model (Fig. S1†).[28] The model (ProtVec LASSO, ESI eqn (1)†) was validated *via* 5-fold cross-validation and found to correlate well with the bioactivity (Pearson correlation coefficient R = 0.84, Fig. 2A). Though a RIDGE regression model offers an alternative with a similar Pearson correlation (R = 0.83, Fig. S1†), we used a LASSO model since, by minimizing the number of non-zero regression coefficients, it provides a simpler model with fewer parameters.

The strength of this ML model is that it can be applied on any other sequences made of canonical amino acids, which can be represented in a 100-d vector. With the trained ML model, we decided to screen for bioactive 6-mers. We chose peptides with only 6 amino acids because they would be shorter than any infectivity enhancing peptide from our library or the literature.

Open Access Article. Published on 15 June 2023. Downloaded on 8/4/2025 8:11:42 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

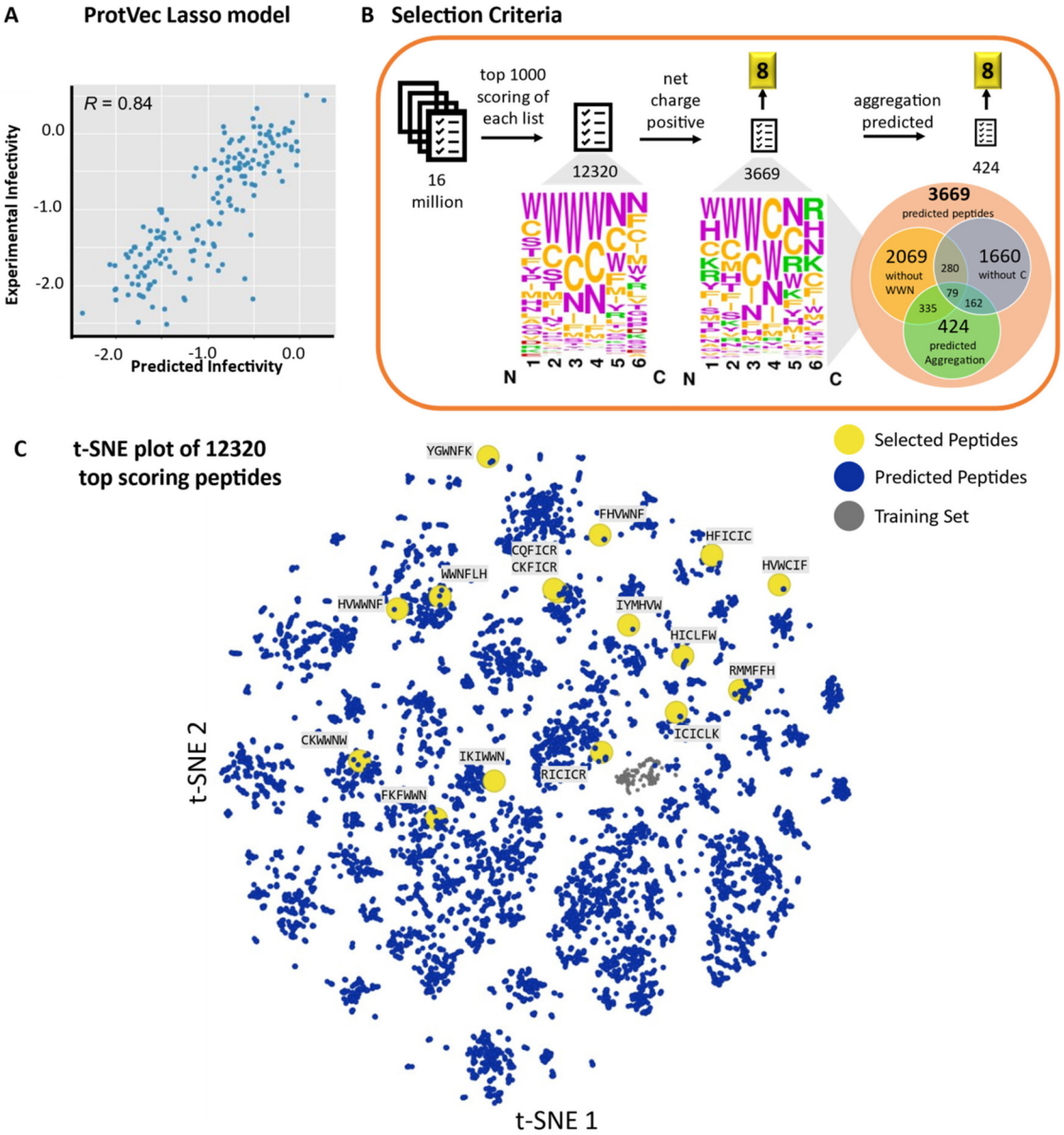

Fig. 2 (A) Performance of supervised linear regression model trained from EF-C based library. The experimentally measured data correlate highly with the predicted infectivity enhancement (Pearson R = 0.84). (B) Scheme showing the selection criteria for narrowing down putatively infectivity-enhancing peptides for experimental evaluation. From each independent prediction Monte Carlo Sampling run the best predicted 1000 sequences were selected. From this subset 3669 sequences showed a positive net charge and were further considered as promising candidates. 8 sequences predicted for aggregation and 8 sequences not predicted for aggregation were selected for experimental testing. The letter size in sequence logo plots is visualizing the amino acid frequency at the corresponding positions in the predicted 6-mer sequences. A detailed listing of amino acids abundancy for the 3669 peptides is shown in Fig. S4.† The Venn diagram is summarizing the main composition motifs of the 3669 peptides. (C) t-SNE dimension reduction plot of the predicted sequences (12 320, blue), the selected subset (16, large yellow symbols with sequence information) and training set (grey). The data points are projected from a 100-d space to a 2-d space with t-SNE.

To sample the sequence space in a time and computational wise cost-efficient way we applied a Monte Carlo model, so that not all possible 6-mer sequences (20^6 = 64 million) must be calculated. Starting from 16 initial 6-mers with randomly chosen residues, we executed 16 independent MC runs which were continued for 1 million steps, generating a peptide at each step. The generated peptides were retained based on a Metropolis criterion (Fig. 2B, see Methods; full lists of the retained peptides, along with code, data, and the trained ML models, can be found at **https://gitlab.com/arghyadutta/seq-to-infect**).

From each of the 16 MC runs, only 1000 sequences with the largest predicted infectivity values were kept, yielding 12 320 sequences after removing duplicates (Table S3†). The predicted 6-mer sequences were represented in a 2-d t-distributed stochastic neighbor embedding (t-SNE)[29] dimensionality

 This journal is © The Royal Society of Chemistry 2023

Open Access Article. Published on 15 June 2023. Downloaded on 8/4/2025 8:11:42 PM. This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

reduction plot from 100-d vector space to check their semantic varieties (Fig. 2C). The t-SNE algorithm attempts to cluster sequences that are semantically close to each other regarding their amino acid composition. As expected, the training set peptides, which have similar amino acid compositions by design, formed a cluster that is distinctly separate from the widely distributed clusters of the generated sequences.

Criteria for selecting sequences

To further narrow down the list of potentially interesting sequences for experimental evaluation, we applied two selection criteria based on previously found evidence for infectivity-enhancing self-assembling peptides: charge and propensity for aggregation. Aggregation is important for bioactivity because without the amyloid structure, the unique properties in cell-adhesion[30] and retroviral transduction enhancement[8] are lost. To find the peptides, that are prone to aggregate and form fibrillar structure, we applied open source protein-aggregation tools Tango,[31] APPNN,[32] Waltz,[33] PATH,[34] Aggrescan[35] and PASTA 2.0.[36] Since these tools are developed for polypeptides and proteins, we first evaluated the reliability of each tool to predict aggregation of self-assembling short peptides with the accuracy and receiver operating characteristic (ROC) value (Fig. S2†) by using our training set (Table S1†).[8] APPNN,[32] PATH,[34] and Aggrescan[35] performed best (Fig. S2†) and were used to select aggregation prone peptides.

Out of the 12 320 predicted sequences for infectivity enhancement, 3669 peptides have a net positive charge; 424 of these 3669 peptides were predicted for aggregation by at least two of Aggrescan, PATH, and APPNN (for detailed analysis see ESI Section 2, Table S4†).

Most of these peptides contain the motif "WWN" (1600 of 3669) or the amino acid Cysteine (2009 of 3669) as visualized in the Venn diagram and in the sequence logo plot (Fig. 2B, and Fig. S4†). Interestingly, the motif "WWN" does not appear in any of the training set peptides, whereas Cysteine was shown previously by us to contribute positively to infectivity enhancement.[7,8] For experimental evaluation, peptides with a large variation in sequences were selected from different clusters in the t-SNE map in order to cover a large sequence space (Fig. 2C). Other than predicted infectivity, hydrophobicity, and aggregation propensity (Fig. S3†), we considered *N*-gram similarity scores (Fig. S5, and Table S5†) to ensure diverse selection of sequences. Finally, from the total 16 peptides selected for experimental evaluation, 8 of the peptides were predicted for aggregation, and 8 sequences were not predicted for aggregation as a control group (Fig. S3†). All these peptides strongly differ sequence wise from the training set, as visualized *via* sequence plot and *N*-gram similarity scores (Fig. 2B, and Fig. S5†).

Prediction yields highly active *de novo* peptides

16 short 6-mer peptides predicted for high biological activity (enhanced HIV-1 infection rates) were selected for experimental evaluation of their physicochemical properties and infectivity enhancement. All the tested peptides are biocompatible as tested *via* a cell-viability assay (Fig. S8†).

4 of the 16 peptides show remarkable infectivity enhancement above 10% relative to EF-C (Fig. 3A, and Fig. S7†). It is important to recognize the inherent difference in sequence length when comparing the infectivity enhancement of the newly found sequences with EF-C. The four infectivity enhancing 6-mer peptides have roughly half the molecular weight of the 12-mer EF-C. Therefore, these peptides exhibit approximately twice the infectivity enhancing efficiency in terms of mass concentration when compared to EF-C. Consequently, a direct comparison needs to consider this significant difference of sequence length. Further, these 4 peptides were predicted for aggregation, resulting in a hit rate of 50% based on the selected 8 aggregation prone peptides (Fig. S3†) or a hit rate of 25% relative to the entire selection. Interestingly, among these peptides only one peptide (ICICLK) shows a positive zeta-potential. The other 3 peptides (HVWCIF, HICLFW, HFICIC) form fibrils, colocalize with cell-membranes (Fig. 3B–E) and show infectivity enhancement despite their moderately negative zeta-potentials. We wondered whether these hit peptides show a different mode of action and applied a property–activity correlation model, which was developed with the training set (ESI section 4†).[8] The newly created peptides fit in the model well (R = 0.72, Fig. 3F), which indicates an interaction mode comparable to the training set: the peptide fibrils associate with viruses and colocalize them with cellular membranes, which facilitates the uptake and increase infection rate.[4]

3 Discussion

ML is an emerging approach to discover new active compounds from a diverse chemical space.[37] The main benefit of a ML-based approach is that it can extract underlying structure–activity relationships and use this information to predict new structures with the desired activity. With the increasing complexity of systems, it is becoming more difficult to rationally design structures from scratch, as experienced often for self-assembling structures that span nm to μm length scales and differ strongly in their biological activities. In this regard, ML is a promising approach to discover bioactive self-assembling peptides, which are difficult to rationally design *de novo*. Here, we show that a ML approach, relying on continuous vector representation for proteins, can be applied to design and predict short 6-mer peptides with diverse sequence and physicochemical properties.

Our training set contains 163 peptide sequences based on derivatives of an active compound. Small training sets are common in early stages of research, but they are rarely considered for ML approaches that aim to predict new sequence spaces since training the model is difficult.[20,27,38–42] As shown here, new peptide sequence spaces can be discovered *via* a computational approach that combines ML and MC with further screening, while still using a small training set with a wide variation in bioactivity.

Open Access Article. Published on 15 June 2023. Downloaded on 8/4/2025 8:11:42 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

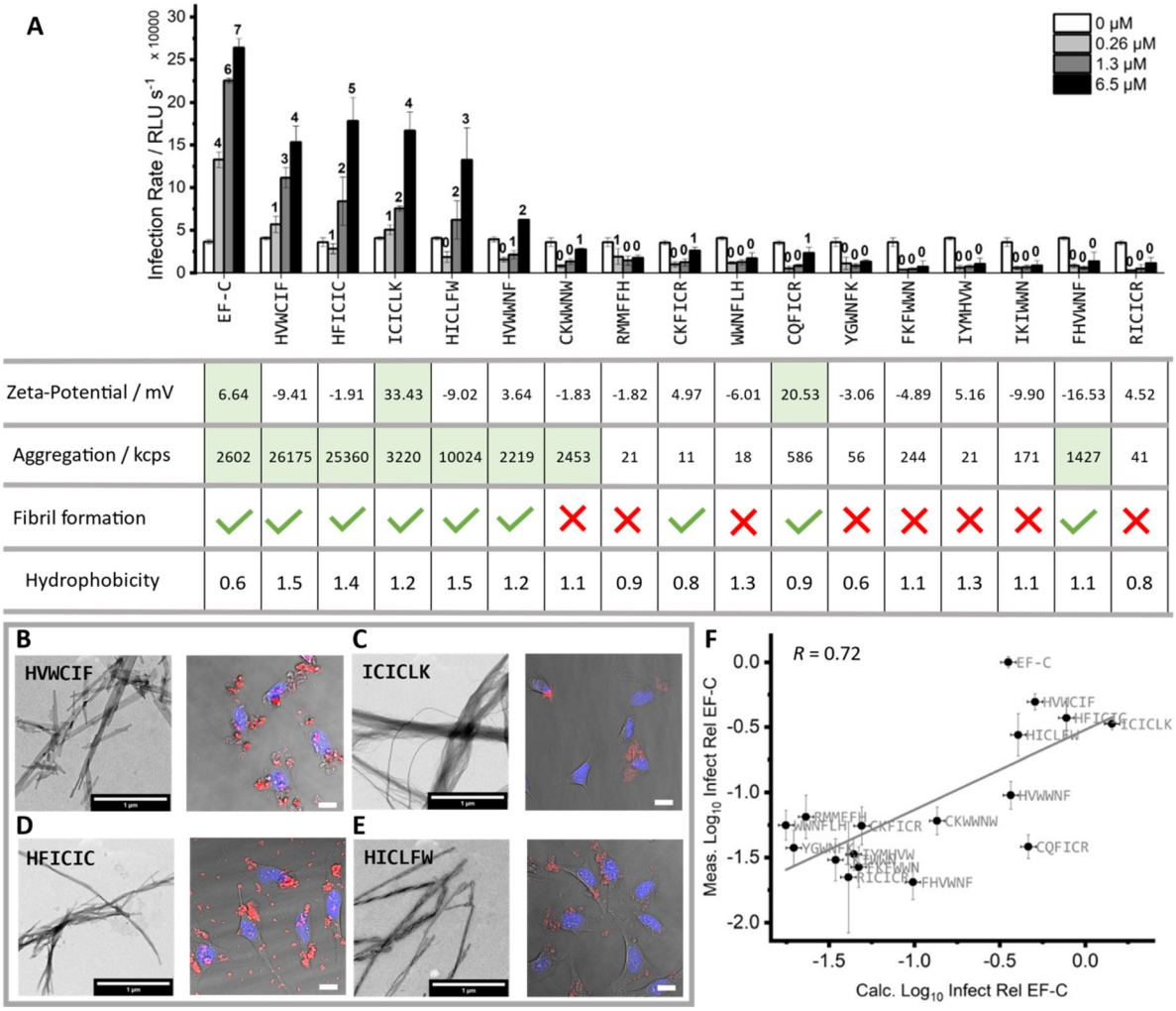

Fig. 3 (A) Summary of infectivity enhancement and physicochemical properties of *de novo* predicted peptides (Table S2†). Peptides are incubated in PBS at 1 mg mL^{-1} for 1 d at RT before characterization. Absolute HIV-1 infection rates in the presence of EF-C (QCKIKQIINMWQ) and peptides from ML-prediction at 6.5 µM, 1.3 µM, 0.26 µM and 0 µM (virus only infection). The n-fold infection rates relative to virus only control is shown for each column. The aggregation into µm-sized colloids was determined by light scattering count rate during zeta-potential measurements. The molecular aggregation into nm-sized fibrils was determined by transmission electron microscopy (TEM, Fig. S6†). Hydrophobicity was calculated according to Fauchere hydropathy scale.[50] (B–E) TEM showing fibril morphology (scale bar 1 µm) and confocal fluorescence microscopy showing cell-fibril-colocalization (scale bar 20 µm) of hit peptides (B) HVWCIF, (C) ICICLK, (D) HFICIC, (E) HICLFW. For TEM measurements the peptides were stained with 4 wt% uranylacetat during preparation. For confocal fluorescence microscopy the fibrils were stained with Proteostat and diluted to 20 µg mL^{-1} before adding to HeLa cells (40 000/1 cm^2) which nucleus was stained with Hoechst 33342. (F) Model describing property–activity correlation for the training set, see ESI Section 4,† applied on newly found peptides (Pearson correlation coefficient R = 0.72).[8] The plot shows the infectivity enhancement of peptides relative to the infectivity enhancement of EF-C in a logarithmic scale. The calculated value is determined by ESI eqn (3).†

Interestingly, most of the predicted peptides are rich in hydrophobic amino acids cysteine and tryptophan, the latter mainly from the sequence motif "WWN". Sequences which contain these hydrophobic amino acids enhance infectivity if they form fibrillar structures that are influenced by cysteine's capability to form disulfide bonds (ESI chapter 9, Fig. S10A–D†). Notably, in a previous study on the training set, we discovered a higher prevalence of cysteine in active peptides.[7,8] However, it was found that cysteine was not essential for activity.[7] In contrast to the *de novo* peptides found in this study, such as ICICLK, the non-essential role of cysteine in the training set can be attributed to the strong self-assembly tendency of peptides with amphiphilic sequence patterns. These patterns stabilize the structure even after the disulfide bonds are broken (Fig. S10F†). Therefore, our machine learning approach successfully extracted the importance of cysteine for peptide fibril formation and incorporated this information in the newly found sequences. We showed here that fibrillar structure formation can be predicted with an accuracy of ~75% through the combination of open-source aggregation prediction tools Aggrescan,[35] APPNN,[32] and PATH.[34] We found that while these algorithms were developed for polypeptides

 This journal is © The Royal Society of Chemistry 2023

Open Access Article. Published on 15 June 2023. Downloaded on 8/4/2025 8:11:42 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

and proteins, they also perform well for short self-assembling peptides.

The infectivity enhancement of peptide fibrils occurs due to improved colocalization of viruses with cell-membrane.[4] The main driving force for this interaction is believed to be electrostatic interactions;[7,9,11,43] where positively charged fibrils sequester negatively charged virions which in turn bind to the negatively charged cellular membrane. As virion attachment to the cell membrane is the major rate limiting step during viral entry, increased numbers of virions at the cell surface result in higher cell entry and infection rates. However, we here found that fibrils with a moderately negative zeta-potential can also increase infectivity. These kinds of fibrils were not included in the training set and not reported before. We hypothesize that oversimplification of the fibril–cell-membrane interaction by reducing it to solely electrostatic interactions can be misleading since the cell interaction of fibrils is regulated by an intricate balance between charge and hydrophobicity. For example, the peptide CQFICR (Fig. 3A) forms fibrils and has a positive zeta-potential but does not enhance infectivity. The low hydrophobicity for CQFICR (0.9) results in less aggregated fibrils and decreases hydrophobic cell-membrane interaction. Another example is demonstrated with the hydrophobic peptide FHVWNF (Fig. 3A), which forms aggregating fibrils with a negative zeta-potential but does not enhance infectivity due to the contribution of the strong negative zeta-potential, as supported by our property–activity model (Fig. 3F). A further example are fibrils derived from the immunoglobulin light chain that have a net negative surface charge and retain virion-binding activity but lack cell-binding and viral transduction enhancing properties.[43] More recently it has also been shown that cellular protrusions actively engage EF-C fibril/virion complexes, suggesting that not only electrostatic interactions may account for bioactivity.[44] It is important to note that meaningful comparisons of zeta-potential can only be made among peptides that either form aggregates or do not form aggregates. This is because the size of colloidal particles has an influence on the measured zeta-potential.[45]

Our method demonstrates that moderately negatively charged peptide fibrils can be active if the hydrophobicity and aggregation features are both strongly pronounced. Hydrophobic amino acids such as tryptophan, phenylalanine, and cysteine can facilitate these desired properties; the continuous vector representation of peptides successfully extracts this underlying information by processing sequence and activity information of the training set without the requirement to assume a predetermined set of relevant descriptors as often done in traditional prediction approaches.[46] The predicted sequences show a higher hydrophobicity, on average, than reported for the training set (Fig. S11†).

Taken together, our method offers a promising tool to yield diverse peptide structures, which cannot be created rationally from derivatives of active compounds or by using conventional approaches such as sequence–pattern analysis.[8,47]

Finally, all these newly found active peptides are the shortest infectivity-enhancing peptides known to us and not found in any protein databases, which makes them truly *de novo*.

4 Conclusions

Discovering new supramolecular nanostructures with a desired bioactivity is challenging due to complex multi-parameter and multi-scale dependencies. In this work, we report an inverse design approach by applying a computational pipeline that includes using a continuous vector representation-based Machine Learning model, a Monte Carlo sampling, and a final screening based on charge and aggregation propensity. A small peptide library, based on derivatives of an active peptide, was utilized to predict short 6-mer self-assembling peptides with prospects for their application as transduction enhancers in basic research and the clinics.

The strength of a continuous vector representation-based approach is that it can encode sequence and physicochemical information of a peptide into a numerical vector which can then be used to train an ML model. Monte Carlo sampling, using the trained ML model, enables us to screen a large sequence space in a time- and cost-efficient way and yield *de novo* active peptide sequences, which are structure- and property-wise very different from the training set. We envision that our data-driven method will substantially accelerate the early stages of research by screening large sequence spaces and predicting *de novo* peptides starting from a small dataset, which are unexpected by human experience and rational design.

5 Materials and methods

Materials for peptide characterization

Dimethylsulfoxide (DMSO, ACS reagent, ≥99.9%) was purchased from Honeywell, Riedel-de Haën®. Uranyl acetate was purchased from Merck. PBS was purchased from Sigma Aldrich. Proteostat® was purchased from Enzo Life Sciences. All chemicals were used as received.

Creation of peptide library and synthesis of new peptides

The data on infectivity of a peptide library based on the gp120 derivative enhancing factor C (EF-C) was reported previously by us.[8] The peptides were synthesised according to standard Fmoc solid phase peptide synthesis and were commercially obtained Phtd Peptides industrial Co. limited with purity of ≥95% as determined by liquid chromatography mass spectrometry.

Preparation of peptide fibrils

The peptides were dissolved in DMSO (c = 10 mg mL^{-1}, stored at 4 °C before usage) and added to PBS buffer to a final incubation concentration of 650 μM or 1 mg mL^{-1} as indicated in the text. Other concentrations described in the text were achieved by further dilution of the peptides.

Physicochemical characterization of peptide fibrils

TEM. 5 μL of peptide fibrils (1 mg mL^{-1}) were placed on copper grids which were placed on copper grids coated with carbon and formvar layer (300 mesh, Plano GmbH). After

Open Access Article. Published on 15 June 2023. Downloaded on 8/4/2025 8:11:42 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

10 min incubation time, the grids were staining with 4% uranyl acetate solution for 2.5 min and washed with water. Measurements were performed on a Jeol 1400 electron microscope with 120 kV acceleration voltage.

ATR FT-IR. To determine the β-sheet content of peptide fibrils ATR FT-IR spectroscopy measurements were conducted by lyophilizing 200 μL of the respective 1 mg mL^{-1} peptide solution. All spectra were recorded on a Bruker Tensor 27 spectrometer with a diamond crystal as ATR element (PIKE Miracle™, spectral resolution 2 cm^{-1}). Every sample was measured with 64 scans and processed with OriginLab software according to a previous report.[7]

ThT-assay. 4 μL of 1 mg mL^{-1} peptide fibrils solution is added to a 20 μL, 50 μM ThT-solution in PBS. The mixture is incubated for 15 min at RT. For reference PBS containing 10% DMSO (4 μL) was added instead of peptide fibril solution. The samples were placed in black UV Star® 384 microliter well-plates (Greiner bio-one). Fluorescence spectra were recorded on an Infinite® M1000 PRO microplate reader (Tecan) at e_{em} = 488 nm upon excitation at e_{ex} = 440 nm with 10 nm bandwidths and multiple reads per well (3 × 3). Measurements were repeated in triplicates and averaged with standard deviation. A peptide fibril was considered as ThT-active if the fluorescence intensity was at least twice as strong compared to the control.

Zeta-potential. For the determination of the surface charge of aggregated peptides zeta-potential measurements were conducted with a Zetasizer Nano ZS, Malvern Instruments. Unless stated otherwise the peptides were dissolved from DMSO (10 mg mL^{-1}) in PBS (pH 7.4) to concentration of 1 mg mL^{-1} and incubated for 1d at RT. Just before the measurement the 60 μL of the peptides were further diluted in 600 μL KCl (aq, 1 mM) in a 1 mL disposable folded capillary cells (DTS-1060, Zetasizer Nano series, Malvern). The zeta-potential of the peptides was derived from the electrophoretic mobility based on the Smoluchowski formula. All measurements were averaged in triplicates and reproduced at least once for each sample.

Aggregate analysis. The derived count rate of scattered light at 633 nm, 173° from the zeta-potential measurement was used as information on the light scattering intensity and turbidity of the sample as an indicator for microscopic aggregation.[8,48]

Virus-peptide interaction

Human wild-type HeLa cell line were obtained from abcam, ab260075, LOT: GR3292155-1. TZM-bl cell line and R5-tropic HIV-1 stock plasmid (pBRNL4.39-92TH014) were obtained from the National Institutes of Health AIDS Research and Reference Reagent Program as reported earlier.[49]

R5-tropic HIV-1 stocks and HIV-1 infection assays were prepared analogous to a previous report.[7] Briefly, the effect of peptide fibrils (final concentration on cells 6.5, 1.3, 0.26, 0 μM) on HIV-1 infection was studied *via* a luminescence assay for detection of β-galactosidase, which is expressed upon HIV-1 infection of TZM-bl cells. The HIV-1 infection assay was conducted in three technical replicates and reproduced at least once. Note, that *n*-fold infectivity enhancement of peptides relative to virus only infection rates are strongly dependant of initial virus concentration. To compare independent measurements with each other EF-C (QCKIKQIINMWQ) was always used as a reference peptide. EF-C is the original sequence on which the training set is based and was applied previously by us to quantify infection rates.[4,7,8]

Cell viability was determined after addition of peptides to TZM-bl cells *via* the CellTiter-Glo assay. To this end, 10 000 cells were seeded and on the next day serial diluted peptides were added. After 3 days the supernatant was removed and 100 μL CellTiter-Glo Reagent 1 : 1 diluted in PBS was added. After 10 min 50 μL was transferred to white microplate and luminescence was recorded by Orion microplate luminometer.

Confocal laser scanning microscopy studies were performed for the visualization of the cell-peptide interaction. HeLa cells were seeded one day prior to conducting the assay (40 000 per well) in an 8-well IBIDI slide. 4 μL of the preformed peptide fibrils (1 mg mL^{-1}) were diluted with 4 μL Proteostat (Enzo Life Science, 1 μL stock in 999 μL PBS) and further diluted with medium to receive a final peptide concentration of 20 μg mL^{-1}. The nucleus of the HeLa cells was stained with Hoechst 33342 (NucBlue™, Thermo Fisher Scientific). The peptide solution mixture was transferred to the HeLa cells and incubated for 30 min at 37 °C before washing three times with PBS. The interaction of fibril clusters with cells was monitored after 30 min incubation time on a Stellaris 8 confocal laser scanning microscope (Leica) equipped with a 20× air objective and laser excitation wavelength of 405 nm (Hoechst) and 561 nm (Proteostat).

Calculation of hydrophobicity and net charge

Net charge. The net charge of a peptide was calculated at pH 7.4 and the EMBOSS pKscale.

Hydrophobicity. The calculation of the hydrophobicity is based on experimental based on water–octanol partition coefficients of amino acids according to Fauchere[50] hydropathy scale. The calculation of the net charge and hydrophobicity were conducted *via* the "peptides" package in R.[51]

Details of computational pipeline

Machine learning. A peptide library consisting of 163 sequences derived from EF-C (QCKIKQIINMWQ), which was previously reported by us (Table S1†),[8] was used as the training set for the ML model. To this end, each peptide of the training set was represented as 100-dimensional numerical vectors (ProtVecs) by summing up ProtVecs of their constituent 3-grams (comprising 3 consecutive amino acid) as outlined by Asgari and Mofrad.[26] Scikit-learn was used to train and cross-validate LASSO and RIDGE regression models that connected the ProtVecs to the experimentally determined infectivity of the peptides.

Metropolis criterion. To sample the vast sequence space of the 6-mers, we used a MC method as described in the Results section. At each MC step, we generated a new peptide sequence by randomly changing one amino acid of the old peptide from the previous step and computed the infectivity of

 This journal is © The Royal Society of Chemistry 2023

Open Access Article. Published on 15 June 2023. Downloaded on 8/4/2025 8:11:42 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

the old and new peptides using the trained ProtVec LASSO model (ESI eqn (1)†). We denote the infectivity of an old and a new peptide sequence I_0 and I_n, respectively. Following the standard Metropolis protocol, we accept the Monte Carlo move under the acceptance criterion

$$\mathrm{acc}(0 \rightarrow n) = \begin{cases} 1, I_n - I_0 \geq 0 \\ \exp\left(\frac{I_n - I_o}{t}\right), I_n - I_0 < 0 \end{cases}$$

where t denotes the reduced temperature (set to 1 throughout). The Metropolis criterion is evaluated by comparing the second term to a uniformly distributed random number in the interval [0, 1).

All code and data used for ML and MC analysis are openly available at **https://gitlab.com/arghyadutta/seq-to-infect**.

Data visualization

Sequence plot. To visualize the amino acid frequency in the peptide library the sequence logo generator tool WebLogo[52] was applied. The sequence plot shows the amino acid frequency in the first seven positions starting from the N-terminus of the library or the first 6 positions starting from the N-terminus of the ML-generated peptides. All sequences of the library shorter than 7 AA (14 out of 163 peptides) were excluded for this visualization. The color code represents amino acid side chain hydrophobicity and charge categorization according to Kyte–Doolittle hydropathy scale. Green represents hydrophilic positive charge, yellow hydrophobic no charge, purple hydrophilic no charge and red hydrophilic negative charge.

t-SNE plot. For the visualization of sequence similarities in the library and the predicted peptides dimensionality reduction approach t-SNE (t-distributed stochastic neighbor embedding) was applied. To this end, the sequences were vectorized by linear vector addition of 3-mer ProtVec vector embeddings. The 100-d vector embeddings were applied as features in t-SNE, which was operated in the data analysis software Orange3 (V. 3.26.0)[53] with the following conditions: Perplexity 30, exaggeration 1, PCA components 20, normalized data. The parameters were selected empirically after testing several conditions for displaying distinct clusters.

Author contributions

K.K. designed, carried out the experiments, analyzed data, performed the formal analysis, and wrote the original draft. A.D. designed and carried out the ML prediction, analyzed data, performed the formal analysis, and revised the manuscript. L. R.W. designed and carried out the experiments related to virus transduction, analyzed data, performed the formal analysis, and revised the manuscript. T.B. conceived the idea, conceptualized, and analyzed data. C.V.S., J.M, T.B. and T.W supervised, provided the funding resources, and revised the manuscript. The final manuscript was approved by all authors.

Conflicts of interest

J.M. and L.R. hold patents for application of peptide fibrils for retroviral transduction enhancement.

Acknowledgements

The authors would like to thank Prof. Dr. Pascal Friederich for valuable discussions. Further, the authors acknowledge open-source ProtVec vector embedding script,[26,54] WebLogo,[52] Orange3[53] NumPy,[55] Pandas,[56] and Scikit-learn[57] and the freely available aggregation prediction tools Tango,[31] APPNN,[32] Waltz,[33] Path,[34] Aggrescan[35] and PASTA 2.0[36] that were used in this work.

The study was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—project number 316249678—SFB 1279 (A02, A03, A05, C01). T. B. acknowledges support from the Emmy Noether program of the Deutsche Forschungsgemeinschaft (DFG). A. D. acknowledges support by BiGmax, the Max Planck Society's Research Network on Big-Data-Driven Materials Science. Open Access funding provided by the Max Planck Society.

References

1 N. J. Sinha, M. G. Langenstein, D. J. Pochan, C. J. Kloxin and J. G. Saven, *Chem. Rev.*, 2021, **121**, 13915–13935.
2 A. L. Boyle and D. N. Woolfson, *Chem. Soc. Rev.*, 2011, **40**, 4295.
3 L. Naldini, *Nature*, 2015, **526**, 351–360.
4 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.
5 C. Meier, T. Weil, F. Kirchhoff and J. Münch, *Wiley Interdiscip. Rev.: Nanomed. Nanobiotechnol.*, 2014, **6**, 438–451.
6 K. Kaygisiz and C. V. Synatschke, *Biomater. Sci.*, 2020, **8**, 6113–6156.
7 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.
8 K. Kaygisiz, L. Rauch-Wirth, A. Dutta, X. Yu, Y. Nagata, T. Bereau, J. Münch, C. V. Synatschke and T. Weil, *ChemRxiv*, 2023, DOI: **10.26434/chemrxiv-2023-hfqxb**.
9 J. Münch, E. Rücker, L. Ständker, K. Adermann, C. Goffinet, M. Schindler, S. Wildum, R. Chinnadurai, D. Rajan, A. Specht, G. Giménez-Gallego, P. C. Sánchez,

Open Access Article. Published on 15 June 2023. Downloaded on 8/4/2025 8:11:42 PM. This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

D. M. Fowler, A. Koulov, J. W. Kelly, W. Mothes, J. C. Grivel, L. Margolis, O. T. Keppler, W. G. Forssmann and F. Kirchhoff, *Cell*, 2007, **131**, 1059–1071.
10 B. Dai, D. Li, W. Xi, F. Luo, X. Zhang, M. Zou, M. Cao, J. Hu, W. Wang, G. Wei, Y. Zhang and C. Liua, *Proc. Natl. Acad. Sci. U. S. A.*, 2015, **112**, 2996–3001.
11 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.
12 P. W. J. M. Frederix, G. G. Scott, Y. M. Abul-Haija, D. Kalafatovic, C. G. Pappas, N. Javid, N. T. Hunt, R. V. Ulijn and T. Tuttle, *Nat. Chem.*, 2015, **7**, 30–37.
13 C. Wu and J. E. Shea, *Curr. Opin. Struct. Biol.*, 2011, **21**, 209–220.
14 C. Wu, R. Gao, Y. Zhang and Y. De Marinis, *BMC Bioinf.*, 2019, **20**, 456.
15 F. Strieth-Kalthoff, F. Sandfort, M. H. S. Segler and F. Glorius, *Chem. Soc. Rev.*, 2020, **49**, 6154–6168.
16 F. Sandfort, F. Strieth-Kalthoff, M. Kühnemund, C. Beecks and F. Glorius, *Chem*, 2020, **6**, 1379–1390.
17 R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams and A. Aspuru-Guzik, *ACS Cent. Sci.*, 2018, **4**, 268–276.
18 D. Ofer, N. Brandes and M. Linial, *Comput. Struct. Biotechnol. J.*, 2021, **19**, 1750–1758.
19 K. L. Saar, A. S. Morgunov, R. Qi, W. E. Arter, G. Krainer, A. A. Lee and T. P. J. Knowles, *Proc. Natl. Acad. Sci. U. S. A.*, 2021, **118**, e2019053118.
20 E. Y. Lee, B. M. Fulan, G. C. L. Wong and A. L. Ferguson, *Proc. Natl. Acad. Sci. U. S. A.*, 2016, **113**, 13588–13593.
21 E. M. López-Vidal, C. K. Schissel, S. Mohapatra, K. Bellovoda, C.-L. Wu, J. A. Wood, A. B. Malmberg, A. Loas, R. Gómez-Bombarelli and B. L. Pentelute, *JACS Au*, 2021, **1**, 2009–2020.
22 R. Batra, T. D. Loeffler, H. Chan, S. Srinivasan, H. Cui, I. V. Korendovych, V. Nanda, L. C. Palmer, L. A. Solomon, H. C. Fry and S. K. R. S. Sankaranarayanan, *Nat. Chem.*, 2022, **14**, 1427–1435.
23 T. Mikolov, I. Sutskever, K. Chen, G. Corrado and J. Dean, *Adv. Neural Inf. Process. Syst.*, 2013, 3111–3119.
24 R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu and P. Kuksa, *J. Mach. Learn. Res.*, 2011, **12**, 2493–2537.
25 V. Tshitoyan, J. Dagdelen, L. Weston, A. Dunn, Z. Rong, O. Kononova, K. A. Persson, G. Ceder and A. Jain, *Nature*, 2019, **571**, 95–98.
26 E. Asgari and M. R. K. Mofrad, *PLoS One*, 2015, **10**, e0141287.
27 C. Hoffmann, R. Menichetti, K. H. Kanekal and T. Bereau, *Phys. Rev. E*, 2019, **100**, 033302.
28 R. Tibshirani, *J. R. Stat. Soc. Ser. B*, 1996, **58**, 267–288.
29 L. Van Der Maaten and G. Hinton, *J. Mach. Learn. Res.*, 2008, **9**, 2579–2605.
30 K. Kaygisiz, A. M. Ender, J. Gačanin, L. A. Kaczmarek, D. A. Koutsouras, A. N. Nalakath, P. Winterwerber, F. J. Mayer, H. Räder, T. Marszalek, P. W. M. Blom, C. V. Synatschke and T. Weil, *Macromol. Biosci.*, 2023, **23**, 2200294.
31 A.-M. Fernandez-Escamilla, F. Rousseau, J. Schymkowitz and L. Serrano, *Nat. Biotechnol.*, 2004, **22**, 1302–1306.
32 C. Família, S. R. Dennison, A. Quintas and D. A. Phoenix, *PLoS One*, 2015, **10**, e0134679.
33 J. Beerten, J. Van Durme, R. Gallardo, E. Capriotti, L. Serpell, F. Rousseau and J. Schymkowitz, *Bioinformatics*, 2015, **31**, 1698–1700.
34 J. W. Wojciechowski and M. Kotulska, *Sci. Rep.*, 2020, **10**, 7721.
35 O. Conchillo-Solé, N. S. de Groot, F. X. Avilés, J. Vendrell, X. Daura and S. Ventura, *BMC Bioinf.*, 2007, **8**, 65.
36 I. Walsh, F. Seno, S. C. E. Tosatto and A. Trovato, *Nucleic Acids Res.*, 2014, **42**, W301–W307.
37 S. Giguère, F. Laviolette, M. Marchand, D. Tremblay, S. Moineau, X. Liang, É. Biron and J. Corbeil, *PLoS Comput. Biol.*, 2015, **11**, e1004074.
38 K. H. Kanekal and T. Bereau, *J. Chem. Phys.*, 2019, **151**, 164106.
39 A. Dutta, J. Vreeken, L. M. Ghiringhelli and T. Bereau, *J. Chem. Phys.*, 2021, **154**, 244114.
40 A. Capecchi and J.-L. Reymond, *Med. Drug Discovery*, 2021, **9**, 100081.
41 C. Rauer and T. Bereau, *J. Chem. Phys.*, 2020, **153**, 014101.
42 H. Altae-Tran, B. Ramsundar, A. S. Pappu and V. Pande, *ACS Cent. Sci.*, 2017, **3**, 283–293.
43 D. Schütz, C. Read, R. Groß, A. Röcker, S. Rode, K. Annamalai, M. Fändrich and J. Münch, *ACS Omega*, 2021, **6**, 7731–7738.
44 D. Schütz, S. Rode, C. Read, J. A. Müller, B. Glocker, K. M. J. Sparrer, O. T. Fackler, P. Walther and J. Münch, *Adv. Funct. Mater.*, 2021, **31**, 2104814.
45 Y. Chen, H. X. Gan and Y. W. Tong, *Macromolecules*, 2015, **48**, 2647–2653.
46 V. Samsoninkova, N. L. Venkatareddy, W. Wagermaier, A. Dallmann and H. G. Börner, *Soft Matter*, 2018, **14**, 1992–1995.
47 C. Schilling, T. Mack, S. Lickfett, S. Sieste, F. S. Ruggeri, T. Sneideris, A. Dutta, T. Bereau, R. Naraghi, D. Sinske, T. P. J. Knowles, C. V. Synatschke, T. Weil and B. Knöll, *Adv. Funct. Mater.*, 2019, **29**, 1809112.
48 Y. Yoshimura, Y. Lin, H. Yagi, Y. H. Lee, H. Kitayama, K. Sakurai, M. So, H. Ogi, H. Naiki and Y. Goto, *Proc. Natl. Acad. Sci. U. S. A.*, 2012, **109**, 14446–14451.
49 A. Papkalla, J. Münch, C. Otto and F. Kirchhoff, *J. Virol.*, 2002, **76**, 8455–8459.
50 J.-L. Fauchère, M. Charton, L. B. Kier, A. Verloop and V. Pliska, *Int. J. Pept. Protein Res.*, 2009, **32**, 269–278.
51 D. Osorio, P. Rondón-Villarreal and R. Torres, *R J.*, 2015, **7**, 4.
52 G. E. Crooks, G. Hon, J.-M. Chandonia and S. E. Brenner, *Genome Res.*, 2004, **14**, 1188–1190.

 This journal is © The Royal Society of Chemistry 2023

53 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.
54 Replication Data for: Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics - Harvard Dataverse, **https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/JMFHTN**, (accessed 3 February 2023).
55 C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke and T. E. Oliphant, *Nature*, 2020, **585**, 357–362.
56 W. Mckinney, in Proc. of the 9th Python in Science Conf. (SCIPY 2010), 2010, pp. 56–61.
57 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel and B. Thirion, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.

Open Access Article. Published on 15 June 2023. Downloaded on 8/4/2025 8:11:42 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.