Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Discovery of Unconventional and Non-intuitive Self-assembling Peptide Materials using Experiment Driven Machine Learning

Rohit Batra

rohitbatra1989@gmail.com

IIT Madras https://orcid.org/0000-0002-1098-7035

Y Nissi Talluri

IIT Madras

Subramanian Sankaranarayanan

Argonne National Laboratory https://orcid.org/0000-0002-9708-396X

H. Christopher Fry

Argonne National Laboratory https://orcid.org/0000-0001-8343-5189

Article

Keywords: Peptide materials, machine learning, peptide informatics, automated peptide synthesis,β-sheet assembly

Posted Date: October 10th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-5100605/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

Discovery of Unconventional and Non-intuitive Self-assembling Peptide Materials using Experiment Driven Machine Learning

Y Nissi Talluri[1], Subramanian KRS Sankaranarayanan[2,3], H. Christopher Fry[2,*], and Rohit Batra[1,2,4,*]

[1]Department of Metallurgical and Materials Engineering, Indian Institute of Technology Madras, Chennai 600036, India
[2]Center for Nanoscale Materials, Argonne National Laboratory, Lemont, Illinois 60439, USA
[3]Department of Mechanical and Industrial Engineering, University of Illinois, Chicago, Illinois 60607, USA
[4]Center for Atomistic Modelling and Materials Design, IIT Madras, Chennai 600036, India
[*]Author to whom correspondence should be addressed.

September 17, 2024

Abstract

The synthesis of secondary structures such as β-sheet assembly in peptides poses significant challenges, as their formation is governed by complexity of molecular interactions, sequence-specific behavior, and influence of various environmental factors such as solvent, pH, and temperature. While rational design strategies have been attempted to predict peptide sequences that form desired nanostructures, these methods are often biased and ineffective, particularly when dealing with the complexities associated with β-sheet formation. Computational models also struggle to reliably predict β-peptide formation due to the limited understanding of the vast chemical space and the lack of comprehensive experimental data. Furthermore, existing models often fail to capture the diversity and non-intuitive nature of sequences capable of forming β-sheets, leading to a constrained discovery process. To address these issues, we developed an integrated high-throughput experimental workflow and an active learning-based AI framework to better predict the β-sheet formation propensity of peptide sequences. By focusing on sequences where ML predictions deviate from conventional design strategies, we synthesized and tested 268 pentapeptides, discovering 96 that successfully formed β-sheet assemblies. The ML models were found to surpass the commonly used β-sheet propensity table in identifying β-sheet forming sequences, especially for cases that do not contain valine, a high propensity β-sheet conformer. Furthermore, several diverse, non-intuitive and unusual β-sheet forming pentapeptides (for e.g., ILFSM, LMISI, MITIY, MISIW, and WKIYI) were identified by the ML models. Novel chemical trends were mined by analyzing the ML models and a user-friendly web-interface is provided to make these models accessible to the community. Overall, this work underscores the need for AI driven experimental exploration of diverse peptide chemistries and provides a robust strategy to overcome the biases and limitations inherent in current rational design methods to form secondary structures like β-sheets.

Keywords: Peptide materials, machine learning, peptide informatics, automated peptide synthesis, β-sheet assembly

1 Introduction

Peptide-based materials have found extensive applications in various areas of materials science, ranging from regenerative medicine and drug delivery to novel adhesives and electronic materials [1, 2, 3, 4, 5, 6, 7, 8]. A critical feature across these materials is their inherent ability to self-assemble into well-defined nanostructures, with peptide fibrils (1D) and sheets (2D) often involving the formation of β-sheets, stabilized by intermolecular hydrogen bonds between amide groups. Considerable effort has been made to understand the relationship between the peptide sequence, the assembled nanoscale structure and material function [9, 10, 11, 12, 13, 14, 15], particularly since β-sheet networks confer enhanced stability to peptides, making them resistant to enzymatic degradation [16, 17, 9, 10, 11, 12, 13, 14] Despite the wealth of data on nanoscale peptide assemblies, the principles or strategies guiding peptide design remain underexplored. Traditional rational design methods rely on basic features like hydrophobicity scales or structural propensity tables and typically employ sequence patterning of polar and non-polar amino acids. These strategies, while effective, often lead to predictable outcomes that don't fully exploit the diversity of peptide chemistry. Moreover, the resulting structures frequently emerge as byproducts of these methods rather than intentional designs, highlighting the need for new approaches that can unlock the underlying design principles of peptide assembly.

Computational methods have shown promise in peptide design, especially in areas like α-helix or helical bundle formation [18, 19], but they are limited when applied to β-sheet-forming peptides. One key limitation is the inability of the models to fully capture the complexity of the self-assembly process as well as the influence of environmental effects such as solvation, pH and temperature. This is further complicated by the small size of the available experimental dataset, which prevents effective training as well as prediction of diverse, non-crystalline structures that peptides often form. Advanced imaging techniques such as cryoTEM have expanded our understanding of peptide assemblies, but they are low-throughput, time-consuming, and costly. Consequently, the scarcity of experimental data hampers the ability to establish robust design criteria for β-sheet-forming peptides, and current models seldom predict the functional outcomes (e.g., stability or enzymatic activity) of the resulting assemblies.

Recent advancements in coarse-grain molecular dynamics (CGMD) and machine learning (ML) have demonstrated potential for accelerating peptide discovery [20, 21, 22, 23, 24, 25, 26]. Simulations by Frederix et al. explored the aggregation propensity of tripeptides, revealing promising candidates for self-assembly into hydrogels [27]. These calculations were then compared with hydrophobicity to effectively evaluate the feasibility of different candidates. Furthermore, several tripeptides were synthesized and characterized to successfully form a hydrogel in the absence of organic solvents. Similarly, Yuqin Yang et al. [28], demonstrated the use of ab initio calculations and MD simulations in improving peptide-based catalysts. Furthermore, combining CGMD with deep learning models, Jiaqi Wang et al. [29] explored over 10 trillion possible sequences efficiently in identifying peptides with specific assembly properties by testing various combinations, especially in longer chains like decapeptides and mixtures of smaller chains. However, even with these advancements, computational methods alone often struggle to predict the non-intuitive sequences that give rise to β-sheet structures due to the vast chemical space and the underlying limitations of the theoretical models used. This gap highlights the necessity of combining the experimental data with AI workflows that have the ability to directly learn from the scarce datasets.

With the rise of artificial intelligence (AI) and ML, new approaches are emerging that offer significant potential in peptide exploration and discovery [30, 31, 32]. These methods not only provide insights into novel candidate sequences but also enable higher rates of discovery by iteratively refining models based on experimental feedback [33], [34]. In prior work, we demonstrated the ability of AI/ML approaches to surpass human-designed rational methods, particularly by overcoming biases toward high β-sheet propensity amino acids like valine and phenylalanine. While our earlier work was limited by a small sample size, it underscored the promise of ML in discovering non-intuitive self-assembling peptides.

In this work, we present an active learning strategy to develop ML models that can accurately predict β-sheet assembly in a pentapeptide based on experimental feedback. Using synthesis and characterization of 268 pentapeptides, we show that the ML models developed in this work perform better than commonly used β-sheet propensity tables, especially for cases that do not contain valine. Using the developed ML models we discovered several non-intuitive, unusual as well as diverse pentapeptide sequences that result in high degree of β-sheet assembly. We also show that the likeliness of randomly finding a β-sheet forming peptide is extremely low (~ 6 %) , and the developed ML model as well as the β-sheet propensity tables have a decent accuracy (~ 70 %) for this task. The developed ML models are particularly good at this task because of their ability to find both valine containing as well as valine-absent sequences that result in β-sheet formation. Another reason for the success of the ML models is the continuous improvement provided by the active learning loops based on direct experimental feedback.

Furthermore, a novel peptide candidate selection strategy was adopted in this work that we believe led to the success of the ML models. Rather than simply selecting top scoring candidates as predicted by the ML models, which is generally done in many ML works, we focused on those regions of the peptide space where the ML models and the β-sheet propensity tables differed with each other. This led to identification, and subsequent discovery, of β-sheet forming peptides that are quite different from those suggested by traditional methods. An additional outcome of this study is the systematic development of a database of over 250 pentapeptides and their corresponding infrared spectroscopic data of the amide I region, which could be utilized to direct future research. The developed ML models are also made accessible via a web application for the community. This work takes an important step towards identifying many interesting and novel peptide chemistries that result in β-sheet assembly. These identified new peptides can be further explored in the future to identify novel peptide materials. Then, the key physicochemical property that correlates to its function can serve as experimental input into the ML models developed here to predict not only structure, but also their function.

2 Results and Discussion

2.1 Overall active learning workflow

Figure 1 captures the overall workflow adopted in this study to discover β-sheet forming pentapeptides. An active learning scheme consisting of four main steps was performed in an iterative manner. These include 1) generation/updation of database of pentapeptides synthesized and their associated experimental score denoting the degree of β-sheet formation observed, 2) machine learning model training, 3) screening of pentapeptides from a large candidate dataset predicted to result in β-sheet assembly, and lastly, 4) high-throughput synthesis and characterization of screened pentapeptides. In the first step, a database of experimentally studied pentapeptides sequences and the associated degree of β-sheet formation was collated. In the second ML model training step, a variety of molecular descriptors and high level descriptors previously known to be critical to self-assembly in peptides

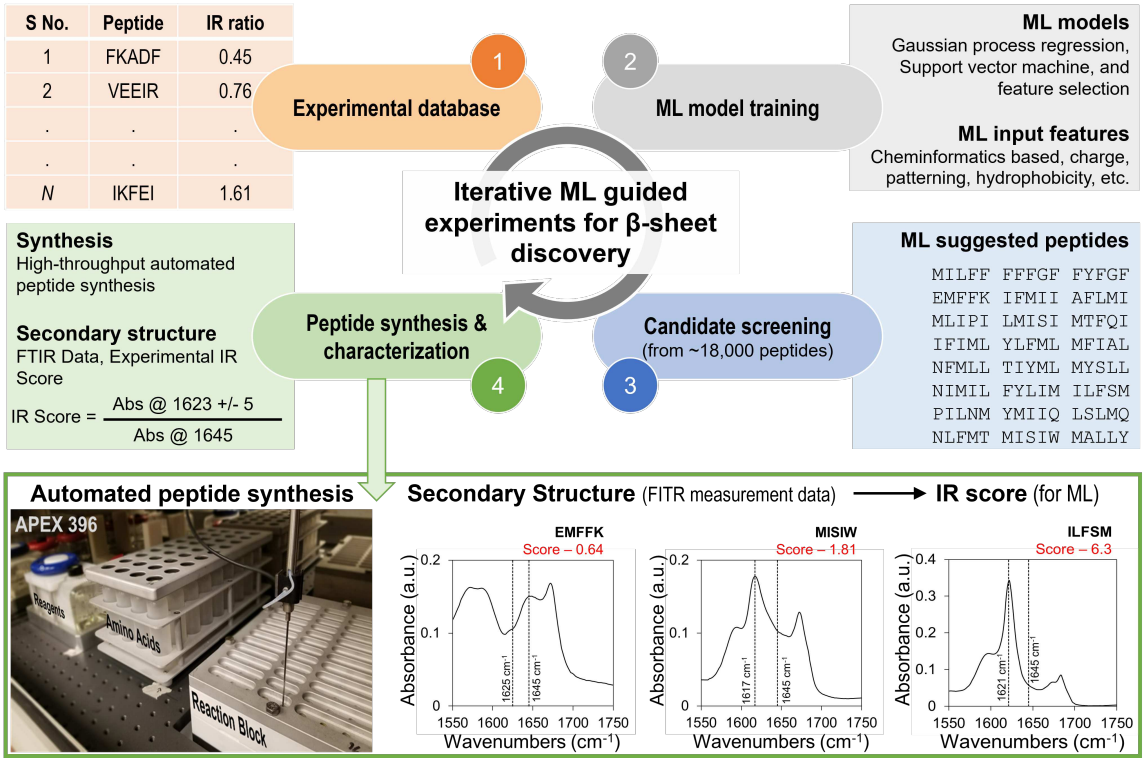
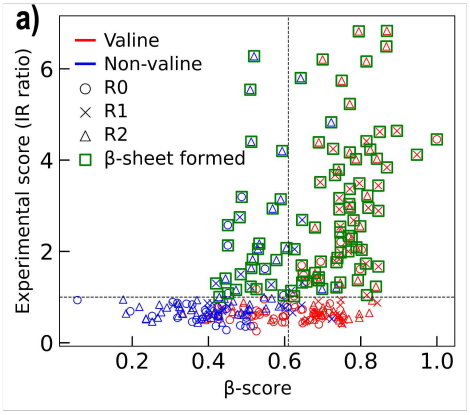


Figure 1: **The overall workflow for discovery of β-sheet forming pentapeptides.** Our active learning methodology iterates over four steps: generation/updation of experimental database, ML model training, pentapeptide candidate screening, and synthesis and characterization of screened peptide sequences. Details of each of these steps are given in the main text. The synthesis of pentapeptides was performed using an automated high-throughput multichannel peptide synthesizer. This was followed by FTIR measurements where the ratio of the absorbance at wavenumbers 1623 $\pm$ 5 cm^{-1} and 1645 cm^{-1}, titled experimental IR score, was taken as the degree of formation of β-sheets. The ML models were trained to learn this IR score given the pentapeptide sequence.

were used as input features to the ML models, which were trained to predict the degree of β-sheet assembly in a given pentapeptide. Two types of ML models, namely, support vector regression (SVR) and Gaussian process regression (GPR) were utilized (see Methods). In the candidate screening stage, the developed ML models were used to make predictions on a large and diverse database of nearly 18000 pentapeptides from our previous work [35]. Based on a screening strategy (discussed in later sections), top candidates were selected for experimental validation in the last step. An automated multichannel peptide synthesizer was used for high-throughput synthesis of pentapeptides in batches of 24. This was followed by FTIR measurements to reveal if the β-sheet assembly occurred. An experimental IR score was defined to estimate the degree of β-sheet assembly observed in a peptide and was computed using the ratio of IR signal at wavenumbers 1623 $\pm$ 5 cm^{-1} (sharp β-sheet peak) and 1645 cm^{-1} (broad random coil peak). This pentapeptide assembly data was then appended to the database, thereby, starting the next round of active learning iteration. In this work, a total of three active learning loops were performed, with around 70 new peptides studied in each loop.

2.2 Discovery of pentapeptides with β-sheet assembly



b)

Top β-sheet forming sequences			
Valine		Non-valine	
Peptide	IR score	Peptide	IR score
IIVYV	6.83	ILFSM	6.27
VTVVL	6.82	IFIML	5.79
IVVLV	6.49	YLFML	5.54
WIVTY	6.20	IFMII	4.82
LVVYV	6.16	AFLMI	4.40
WFVYV	5.74	MILFF	4.20
MVVYI	5.23	FYLIM	3.14
IVVVY	4.64	LMISI	2.94
IVVMV	4.61	MLFLL	2.75
VVVYL	4.41	MITIY	2.69

Diverse β-sheet forming sequences			
Valine		Non-valine	
Peptide	IR score	Peptide	IR score
WIVTY	6.2	YLFML	5.54
LVVYV	6.16	AFLMI	4.4
IVVMV	6.61	LMISI	2.94
VAVFY	4.39	FFFGF	2.17
SVVVM	4.24	MTFQI	1.82
FVVFY	2.96	MISIW	1.81
GVVII	2.68	MFIAL	1.63
HVVYI	2.34	PFIII	1.32
IVVQV	1.73	FRFEI	1.23
IVVVD	1.37	WKIYI	1.14

Figure 2: **Identified β-sheet forming pentapeptides** (a) The complete space of 268 pentapeptides synthesized in this work as illustrated using their β-sheet and experimental score (based on IR spectra). Beyond several valine containing sequences, many non-valine sequences were discovered that showed β-sheet assembly although their β-score was quite low. (b) Top and diverse examples of valine and non-valine sequences that resulted in high degree of β-sheet assembly.

Figure 2(a) shows results for all the 268 pentapeptide sequences that were synthesized and characterized across the three rounds of active learning iterations. 96/268 pentapeptides resulted in β-sheet formation (with experimental IR score > 1) and are highlighted in green square boxes. Further, we distinguish between sequences that do not contain any valine (termed as "non-valine") or have one or more valine amino acid (termed "valine"); this distinction will become important for analysis later. For all the sequences, we also compare their experimental IR score with a commonly utilized β-sheet propensity score, which is based on the formalism provided by Chou and Fasman where β-sheet propensity is reported as a statistical distribution of amino-acid conformers in the Protein Data Bank [36, 37]. It should be noted that β-sheet propensity score, thereafter termed as β-sheet score, is among the most common and successful metric that is often utilized by researchers for rational design of β-sheet forming peptides. However, as seen from the Figure 2(a), many pentapeptides, especially non-valine sequences, with low β-sheet score also resulted in high degree of β-sheet assembly (false negative). Further, many high β-score sequences did not result in β-sheet assembly (false positive). Both these false positive and false negative cases highlight the limitations of the β-score propensity tables, and the need of a comprehensive study to understand the underlying mechanism of self-assembly in peptides.

Before we dive into the ML models developed in this work to identify several interesting and novel peptides that result in β-sheet assembly, we give an example list of top and diverse sequences that were discovered in this work in Figure 2(b). We also distinguish between valine and non-valine sequences as valine is particularly known to induce β-sheet formation. As expected, many valine containing sequences such as, VVVVV, IIVYV, GVVII formed β-sheet because of high β-sheet propensity, patterning as well as a good trade-off between hydrophobicity and hydrophilicity. Such, valine rich sequences are often inspired by naturally occurring sequences like $(\text{VPGVG})_n$ found in elastin and employed in a multitude of biomedical applications [38]. β-turn and β-sheet structures are also found in many of the variants but it is noted that proline typically disrupts the assembly and

its mutations yield β-sheet rich structures [39, 40, 41, 42]. On the other hand, several unexpected valine containing sequences such as, WIVTY, VAVFY and HVVYI were also found to form β-sheet assembly although they consist of diverse and unpatterned amino acid sequences. Similarly, for the non-valine sequences, several non-intuitive cases (i.e. lack of polar/nonpolar patterning, lack of amino acids repeats) such as ILFSM, YLFML, AFLMI can be seen to result in β-sheet formation. Two things particularly to note is the high degree of β-sheet formation as well as the diversity of amino acids in the non-valine sequences. We note that many of the peptides identified here are unlike typical bioinspired motifs like the amyloidogenic FF or the silk inspired (GAGAGS) or the (GAGAGY) repeat unit [43, 44], highlighting the utility of the experimentally driven ML methods presented herein. Atomic force micrographs (AFM) of a select few peptides are included in Figure 3 that verify the correlation between β-sheet formation and nanoscale self-assembly (1D fibers and 2D sheets) in both the valine rich and the non-valine peptides.

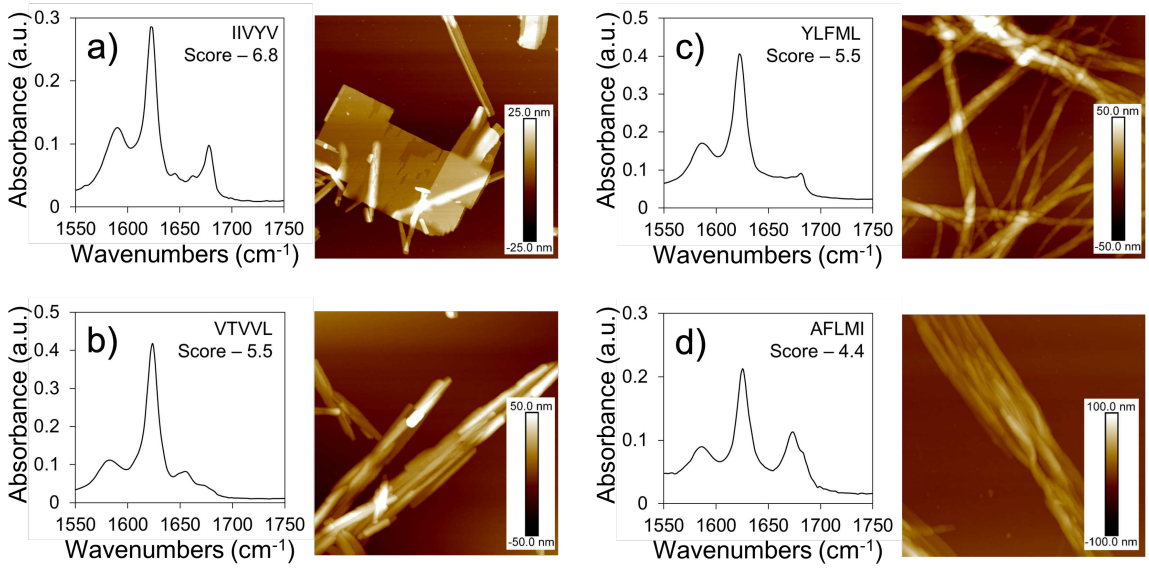


Figure 3: **Experimental validation of synthesized β-sheet forming pentapeptides.** Experimental data highlighting the correlation between infrared spectroscopy and atomic force microscopy of valine rich (a,b) and non-valine (c, d) peptides.

2.3 Details of the active learning loops

Figure 4(a) captures in detail the results obtained in each step of the active learning loops. Starting from an initial dataset of 81 pentapeptides from our previous work [35], we build ML models to directly predict the empirically observed experimental IR ratio given the peptide sequence. In this regard, peptide fingerprint is an important input to the ML models which significantly impacts the model performance. Several types of peptide fingerprints, including based on cheminformatics, β-sheet propensity tables, aggregation propensity obtained through molecular dynamics (MD) simulations, hydrophobicity scales, among others, were attempted to train the best performing ML models. See Methods section for more details on the descriptors used in this work. In terms of ML model types, GPR and SVR were attempted, mainly due to their success in the field of materials

discovery that generally involves problems with high dimensionality and low number of training data points.

Feature selection using recursive feature elimination (RFE) was also adopted and was found to considerably increase the ML model performance. Interestingly, in contrast to our previous work, the MD computed aggregation propensity (AP) feature was eliminated during the RFE stage for almost all the ML models that were trained this work (see Supporting Information). This signifies that the MD computed AP of a sequence may not be a good indicator of its empirically observed β-sheet assembly. This observation is in stark contrast to some of the recent works [29, 30] which mainly use AP values of peptides to screen some of the most promising cases. In this regard, we note that the MD simulations based ML workflows are advantageous as they can be performed entirely in silico. However, owing to the limitations of the force fields and the high computational cost of modeling such a large space of peptides those ML models should be preferred that do not rely on input features dependent on MD simulations.

We also note that two variations of each ML model type were trained, one using the full training data and the other exclusively for the non-valine sequences. Training separate ML models for the non-valine sequences was found to drastically enhance the model performance. Thus, all the predictions for the non-valine sequences are based on ML models trained on just the non-valine sequences, termed the "non-valine model", while the predictions for the valine containing sequences are based on the ML models that are trained on the full dataset, named as the "fulldata model".

Once the ML models are trained, we use them to make IR ratio predictions for several thousand pentapeptides. We note that since AP is one of the feature that was considered in this work (owing to its relevance hinted in the previous works), we restricted our candidate pentapeptide dataset to only those cases for which MD computed AP values were available, which is around 18,000 sequences with high level of diversity. Next, around 70 candidates were selected in each active learning iteration for high-throughput experimental evaluation. This selection was based on the idea of discovering novel and unusual pentapeptides that can form β-sheet assembly. Since neither the ML model nor the β-sheet propensity score is believed to be accurate in predicting β-sheet forming peptides, we specifically looked at cases wherein the ML predictions drastically differed from the commonly used β-sheet propensity tables. Broadly three categories of sequences were selected, one where both ML model and β-score were high, two where ML model prediction was high but β-score was low and three where the opposite holds. Additionally, 48 sequences were randomly selected from the complete pentapeptide space to get an idea of the baseline probability of finding β-sheet forming sequences.

The sequences selected in each round were then synthesized and characterized (see Methods) to determine their actual experimental IR ratio. These results were then compared with ML model and β-score based predictions to evaluate the performance of both the methods. Finally, the obtained experimental IR ratio data was added to the peptide database and the next iteration of active learning was performed. Overall, in this work, two active learning loops were completed, wherein a total of three versions of each ML models were trained. M1 models using the initial data (R0 dataset), M2 models using the R0 dataset plus the results obtained from the first active learning loop (R1 dataset), and M3 models using the R0 and R1 dataset, plus the results obtained from the second active learning loop (R2 dataset). The M3 models are the final released versions available on the web-interface.

As mentioned above, in each round the peptides were selected based on contradiction between the ML predictions and the β-scores. The Figure 4(b) and (c) capture the selections made in rounds R1 and R2, respectively. Grey colored points mark the ML predictions from the M1 model in (b) and the M2 model in (c) for the complete candidate dataset along with the corresponding β-scores. Also, plotted in the figures are the candidates selected for next round of experimental validation (triangular markers) and those that were found to successfully result in β-sheet assembly from

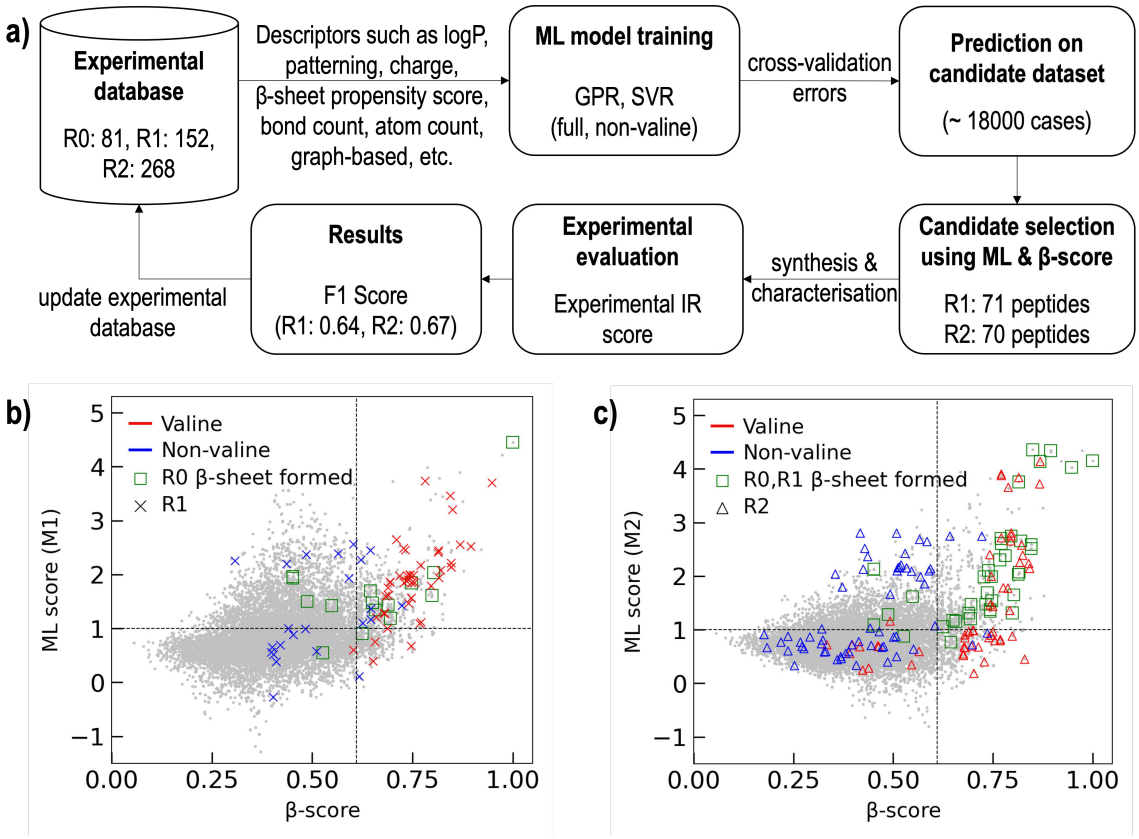


Figure 4: **Active learning scheme and the peptide selection strategy.** (a) The overall active learning scheme adopted in this work to efficiently discover interesting β-sheet forming sequences. Here, three rounds of learning iterations (R0, R1 and R2) were carried out and particular focus was given to synthesizing those interesting sequences where the ML and the β-sheet propensity models differed in their predictions. This resulted in discovery of many non-intuitive and unusual pentapeptides that showed β-sheet assembly. Selected pentapeptides in active learning loop (b) R1 and (c) R2. The colored symbols represent the selected candidates in each round, while the grey points denote the complete candidate pentapeptide dataset. Green markers represent cases for which β-sheet assembly was observed in the previous round(s). In order to improve the performance of the ML models, diverse candidates from all four different quadrants were selected.

the previous round(s) (square green colored markers). The horizontal and the vertical dotted lines respectively denote the ML predicted score and the optimal β-sheet score above which the associated model suggest β-sheet assembly (see Methods section for more details on how these threshold were identified). Interestingly, there are many candidates that lie in the top-left and the bottom-right quadrants in these figures. They represent cases where one of the ML or β-score models predicts high degree of β-sheet formation while the other suggests the opposite. On the other hand, the top-right and the bottom-left quadrants represent cases where both the ML model and β-scores agree in terms of their prediction.

The presence of green symbols in all four quadrants, particularly in 4(b), of the plots suggest that

neither of the models is able to completely capture sequences that result in β-sheet assembly. This observation particularly led us to select candidates from different quadrants of the plot. Thus, with the dual aim to select as diverse candidates as possible and to specifically provide examples to ML model where it is presently failing, we selected candidates from all the four quadrants in both the R1 and R2. Specifically, for selection from top-left and bottom-right quadrant, diverse candidates with highest predictions from one model and relatively low predictions from the other model were given preference. Similarly, diverse candidates with highest predictions from both the methods were selected from top-right quadrant. A random sequence selection resulted in candidates selected from the bottom-left quadrant. The list of candidates selected in each round is provided in Supporting Information.

2.4 Iterative ML model development using active learning

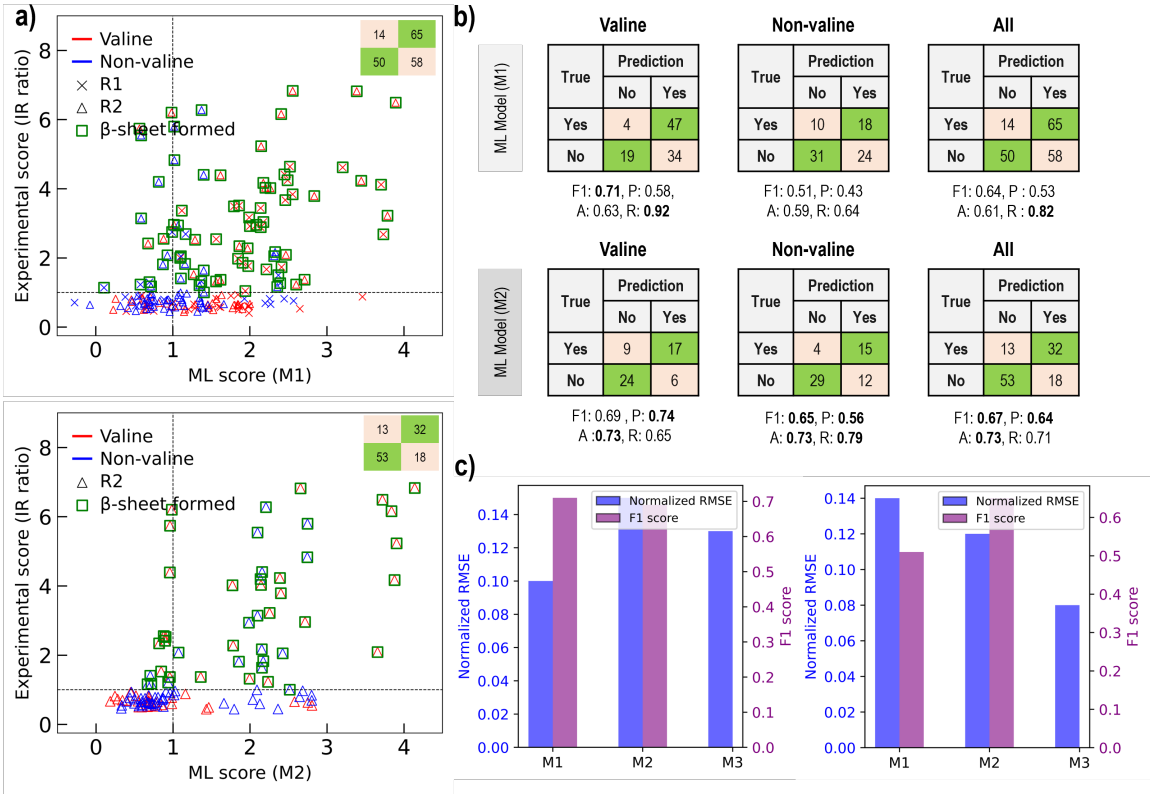


Figure 5: **Performance improvement of ML models with active learning** (a) Parity plot between the ML predicted and the experimentally measured IR ratio of different pentapeptides for the M1 and M2 models. The dotted lines denote β-sheet assembly. Confusion matrix for the each case is also included. (b) Confusion matrix for the M1 and the M2 models for only the valine containing, valine-absent and all sequences. Different error metrics in each case, such as F1-score (F1), precision (P), accuracy (A) and recall (R) are included. (c) Improvement in the normalized root mean square error (RMSE) and F1-score of the ML models with increasing active learning loops.

The active learning strategy adopted in this work allowed iterative improvement in the model performance based on direct experimental feedback. Figure 5(a) plots the ML predicted vs empirically observed IR ratios for the M1 and M2 models. Although the plots are presented for the case of SVR models, the same observations were made for the GPR models as well (see Supplementary Information). Results are shown for only those cases that were not part of the training data, i.e., R1 and R2 dataset for M1 model and R2 dataset for M2 model. Values greater than 1 (or the dotted lines) are indicative of β-sheet assembly prediction or observation. Also, cases where β-sheet assembly was observed are highlighted with green squares. Overall, the M2 model can be seen to perform better than the M1 model—the M2 model has less number of false positives, the predicted IR ratio for many of the false negatives is close to the threshold value of 1, and the coefficient of determination (R2) is higher.

To further analyze the performance of these models for different types of sequences, i.e., valine, non-valine or overall, corresponding confusion matrices are provided in Figure 5(b). Most importantly, the performance of the M2 model on the non-valine dataset is significantly better than the M1 model. Improvements in terms of several important error metrics, including F1-score, accuracy, precision and recall can be observed. Good accuracy of our M2 model on the non-valine dataset is among the major reason for identifying non-intuitive, uncommon and diverse β-sheet forming pentapeptides. One the other hand, for the valine dataset, no major change in the performance of the M1 and M2 models can be observed. This could be because both the M1 and the M2 models have reasonable performance with accuracy of 0.61 and 0.73, and F1-score of 0.64 and 0.67, respectively. Thus, addition of more training data did not result in drastic performance improvement. Figure 5(c) shows the mean five-fold cross validation errors (normalized with respect to their range) for the M1, M2 and M3 models trained separately on the full and the non-valine dataset. A consistent drop in the RMSE and an increase in the F1-score for the non-valine model is another indication of model performance improvement with active learning loops. However, for the full data model a convergence in error suggests no major change in performance. These results signifies ability of our M2 models on the valine as well as the non-valine dataset to accurately predict β-sheet assembly. It should be noted that the error results presented in Figure 5(a) and (b) are based on experiments that were performed *after* the ML model training. Thus, these results are, in a true sense, test errors of our models.

2.5 Performance comparison of ML model and rational design approaches

Next we discuss that not only the ML models improved with active learning loops, but also performed better than traditionally used β-sheet propensity tables. Figure 6 compares the performance of the trained M2 model against the β-sheet propensity scores; results are shown for SVR model here, see Supporting Information for corresponding results for GPR model. Again, cases that resulted in β-sheet assembly are highlighted using green squares. A major area of disagreement between the two methods is the top left area in Figure 6(a) which mostly contains non-valine sequences. While the ML model rightly predicted them to have high IR ratio, as also confirmed by the experiments, the β-sheet propensity score for such cases is low. We highlight that discovery of this region of the peptide space is the reason behind the success of these ML models. On the other hand, the bottom right region highlights the cases where β-sheet score is high but ML score is low. Mostly valine containing sequences can be found in this region, a few of which indeed resulted in β-sheet assembly. The false negatives predicted by the ML model in this region suggests that there is still scope of improvement in the ML models, particularly for the valine containing sequences. However, it should be noted that the ML predicted IR ratio for such false negatives is close to the threshold value of 1, and this region also contains false positives based on the β-sheet propensity score.

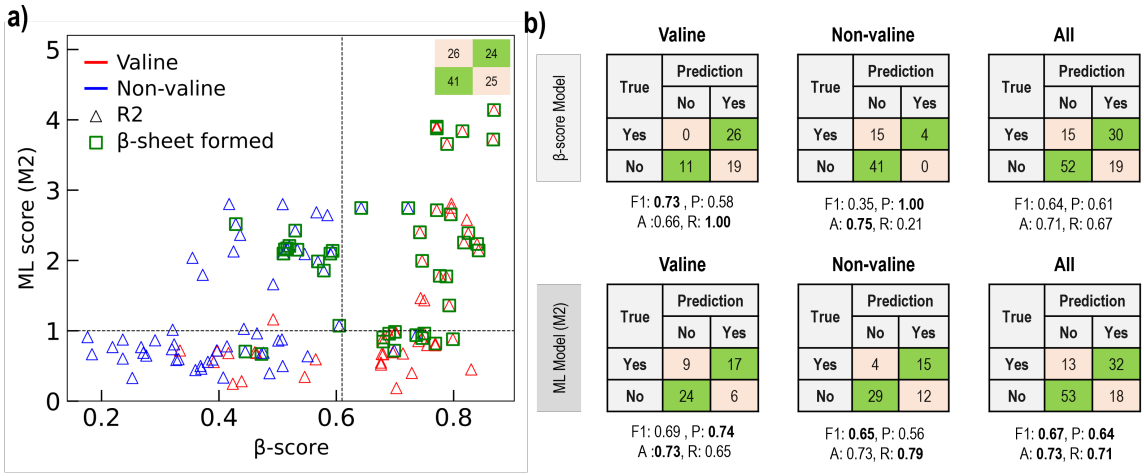


Figure 6: **Performance comparison of the β-sheet and the ML models for peptide discovery.** (a) Parity plot between the β-sheet propensity score and the ML predicted IR ratio of different pentapeptides for the M2 model.The dotted lines denote β-sheet assembly. (b) Confusion matrix for the β-sheet propensity score and the M2 model for only the valine containing, valine-absent and all sequences.

There are two other regions in the plot where both the methods agree, either both predict β-sheet assembly, or both predict no assembly. Majority of the cases lie in these two regions. Particular, many valine containing sequences lie in the top right region where both the methods suggest β-sheet assembly. This is expected since valine is empirically known to induce β-sheet formation and has the highest β-sheet propensity score, well above the other amino acids [36]. This also signifies the success of our ML model to learn that valine containing sequences should be scored high. The bottom left region is mainly constituted by the 48 randomly selected sequences from the complete space of 3.2 M pentapeptides. 41/48 randomly selected sequences were found to lie in this region, and only 3 of these 48 sequences ($\sim$ 6%) were observed to form β-sheet structures. Both these statistics highlight the high level of difficulty of finding sequences with β-sheet assembly.

Confusion matrices for the M2 model and the β-sheet propensity score as applicable to valine, non-valine and all sequences are provided in Figure 6(b). Overall, the M2 model perform better than the β-sheet propensity score in terms of accuracy, F1-score, precision and recall. However, for the valine sequences, accuracy of both the methods are somewhat similar—although the β-sheet propensity seem to have many false positives, thereby having low precision than the M2 model. Overall, the M2 models are more successful than the β-sheet propensity scores due to better prediction capability on the non-valine dataset and a better F1-score on the valine dataset. Thus, it can be concluded that the developed ML models are better at discovering β-sheet forming sequences than the β-sheet propensity tables.

2.6 New learnings from the ML models

The features retained post-RFE were further analyzed to gain new chemical insights and mine factors that promote (or demote) formation of β-sheet. The complete list of features used by different versions of the ML models is presented in Supporting Information. The analysis was done across all

versions of the ML models that were trained during the active learning process, i.e., both GPR and SVR models trained on full and non-valine data for each round of iteration. Patterning (see Figure S4) was among the most prominent feature retained post-RFE across the different ML models, which is consistent with the existing knowledge within the peptide community [45]. Other descriptors that were frequently found to be important across different ML models were related to the hydrophobicity and polarizability of the peptide, which are also well-known in the community [46, 47]. Specifically, SlogP-VSA-1 and SlogP-VSA-11, and the normalized and absolute value of logP were found to be important (see Figure S5). The former two descriptors represent the molecular surface area based hydrophobicity of the peptide, while the latter two are based on summation of individual amino acid hydrophobicity. It was observed that the SlogP-VSA-1 and SlogP-VSA-11 molecular descriptors were preferred over the high-level logP features, possibly because of the finer-level details captured by the former case. SlogP-VSA-1 aggregates the molecular surface area contributions from atoms with logP values less than -0.4, while SlogP-VSA-11 does this with logP values between 0.5 and 0.6. Accordingly, a negative correlation of SlogP-VSA-1 and a positive correlation of SlogP-VSA-11 with experimental IR score can be observed (see Supporting Information). This is expected as amino acids with neutral character (logP values near 0) are preferred. In addition to these, models also highlight that polarizability is vital for β-sheet propensity as the final models retained one of the SMR-VSA descriptors post-RFE. For instance, SMR-VSA-10 was retained in the final M3 model. SMR-VSA-10 aggregates the molecular surface area contributions from those atoms which have high polarizability (i.e., molar refractivity value greater than 4.0).

Beside the known importance of patterning, hydrophobicity and polarizability, few additional insights also emerged from the feature analysis (see Supporting Information for correlation plots between the descriptors and the IR score). For instance, it was observed that the higher the number of double bonds (fp mqns 14), the lower the degree of β-sheet propensity. Similarly, sequences containing 5-membered rings (fp mqns 35) (or amino acids H, P, and W) also exhibited low degree of self-assembly. On the contrary, several sequences with 6-membered rings (such as F, W, and Y) had a small positive correlation with the IR score and resulted in formation of β-sheet. This is already known to the community made evident by the rich history of diphenylalanine (FF) and aromatic peptides [48]. Interestingly, sulfur-containing amino acids, cysteine and methionine, were found to promote β-sheet assembly. Although, synthesizing peptides that include sulfur may not be preferred because of accompanying side reactions and purification challenges.

2.7 Graphical user interface for pentapeptide prediction

In order to make the ML models easily accessible to the community, we have developed a python based web application. The users can simply access this webpage [49] and enter pentapeptide sequence(s) to avail the M3 model based IR score predictions. The predicted results along with basic information (logP, β-sheet propensity score, patterning, etc.) can also be downloaded. We believe this app will provide an easy access to our ML models and can be used by experts to guide their research activities related to β-sheet forming peptides.

3 Conclusions

Discovery of β-sheet forming peptides is challenging owing to the associated vast chemical design space. Using machine learning methods coupled with high-throughput automated synthesis, we discovered several unusual and non-intuitive pentapeptides that resulted in β-sheet assembly. Over three rounds of active learning loops, 268 pentapeptides were synthesized, out of which 96 resulted

in β-sheets formation. Particular emphasis was put on selecting those candidates for synthesis that had contradictory predictions from the ML models and the traditional rational design approach (β-sheet propensity tables in this case) to discover non-intuitive candidates. In-line with the general understanding, several valine containing pentapeptides resulted in β-sheets formation. However, contrary to the present knowledge, several non-valine pentapeptides containing M, L, Y, S, P, W, etc. were also found to result in β-sheets. These results highlight the limitations of the currently used rational design approaches while irrefutably demonstrating the efficacy of ML when predicting "non-conventional" sequences for self-assembly. We also found iterative improvement in the performance of the ML models as more experimental training data became available with active learning loops. Further, the trained ML models performed better than existing β-sheet propensity tables, especially for the cases that do not contain valine amino acid. The ML models were explored to find factors that promote/demote β-sheet assembly. We found that patterning, polarizability and hydrophobicity are among the key factors that influence β-sheet formation. Further, absence of double bonds and 5-membered ring, and presence of 6-membered ring, cysteine and methionine were found to promote β-sheet assembly. Overall, this work presents a data-driven approach to efficiently discover novel and non-intuitive materials with desired properties which may not be possible via traditional rational design approaches.

4 Methods

4.1 Peptide Synthesis and Characterization

The peptides reported here were synthesized in batches of 20 to 24 peptides per synthesis utilizing an aapptec APEX 396 multichannel peptide synthesizer (see Supplemental Information for synthetic details). A critical step in the post-synthesis work up and isolation of the synthesized peptides was the desalting (removal of residual trifluoroacetic acid from the crude peptide) by dissolving the crude peptide in 3 mL of 10 mM hydrochloric acid and freeze drying the peptide to yield a white lyophilized powder. Each peptide was dissolved in deuterium oxide (Sigma Aldrich, 99.8% deuterated water) to achieve a 1 wt% (w/v) or 10 mg/mL solution. The peptide solution was neutralized by adding small, microlitre quantities of a 1 M ND4OD solution (0.7 mL of 14.2 M NH4OH diluted in 9.3 mL of D2O). The peptide solutions were aged for more than 24 hours before their FTIR spectra were recorded. Peptide purities were analyzed by reversed phase high performance liquid chromatography (RP-HPLC) and the identity of the peptides were confirmed by electrospray ionization mass spectrometry (ESI-MS).

4.2 β-sheet Characterization and Experimental Score Determination by Infrared Spectroscopy

The 1 wt% peptide samples were prepared in D2O in order to measure the amide 1 (1750-1550 cm-1) region in an aqueous environment. A 20 μL drop of the peptide solution was sandwiched between two calcium fluoride plates separated by a 0.050 mm gasket ring (effective path length) in a demountable liquid cell. A background spectra of D2O in the same liquid cell was used to correct the baseline. Spectra represent an average of 16 scans at a 4 cm-1 resolution. The resulting spectra were subtracted by the value at 1750 cm-1 as a zeroing function. β-sheets typically appear at 1622 cm-1 as a sharp peak. However, we observed β-sheet peaks from 1619 to 1631 cm-1. A random coil (i.e. no β-sheet) typically appears as a broad spectrum at 1645 cm-1. The experimental score was simply recorded as Abs¬1625 /Abs1645 (Note: if the maximum absorbance of the β-sheet was shifted, then the value at the Absorption max was used, if no β-sheet peak was observed then the

value at 1625 cm-1 was used. The peak at 1675 cm-1 is representative of the residual TFA salt that results as a product from peptide synthesis. As noted in the synthesis section, we always attempt to minimize this peak by desalting our crude material.

4.3 Peptide fingerprints and machine learning models

In order to train the ML models a numerical representation of pentapeptide sequences is required. In this work, we utilized a combination of two types of fingerprinting schemes, high-level and cheminformatics-based. The high-level fingerprints are inspired from our previous work [35] and include seven properties derived for a given pentapeptide sequence, namely, aggregation propensity (AP), β-sheet propensity score [36, 37], patterning, net charge, absolute charge, and hydrophobicity value with and without its absolute. The AP was computed using molecular dynamics simulations set-up discussed in our past work and is available for $\sim$18000 pentapeptides. Normalized values of these features were used for ML model development.

The cheminformatics-based fingerprints were generated using the RDKit python library [50], which utilized the SMILES representations [51] of pentapeptides to compute several descriptors such as molecular quantum number (MQNs), molecular surface area (VSA), among others. MQNs descriptors are inspired from quantum chemical calculations and primarily describe molecular topology, connectivity and symmetry. They include features such as counts of specific atom types (e.g., nitrogen, oxygen, sulfur), counts of specific bonds, the presence of certain functional groups, and topological indices [52]. Other VSA descriptors [53] that capture the molar refractivity (SMR-VSA) and hydrophobicity (SlogP-VSA) across the surface of a molecule were also included. Different kappa indices and chi indices [54], and the number of aromatic rings form a few other descriptors from this category. Overall, 84 cheminformatics-based descriptors were combined with seven high-level descriptors to form 91-dimensional pentapeptide fingerprint for ML model training.

In terms of ML models, both Gaussian process regression (GPR) and support vector regression (SVR) were attempted. Two versions of each the of ML model were trained, one involving the complete set of synthesized pentapeptides data (full-data model), and the other exclusively trained on those pentapeptides that do not contain any valine amino acid (non-valine model). It was observed that training a separate model for non-valine sequences resulted in enhanced model performance. The radial basis function (RBF) kernel was used to train both types of ML models. The hyperparameters of both the GPR and SVR models were optimized through ten-fold cross-validation during model training to prevent overfitting. Three hyperparameters namely, scaling coefficient, length-scale parameter and the expected noise were optimized for the GPR model. Similarly, the length-scale and the regularization factor were the hyperparameters for the SVR model. The performance of all models was assessed using root mean squared error (RMSE). Also, a recursive feature elimination (RFE) technique was used to iteratively eliminate irrelevant features one at a time. The final model was trained using the features derived from the lowest RMSE cross-validation (CV) error, with hyperparameter optimization during RFE. The RFE reduced dimensionality from 91 to 20 and 11 for GPR models and to 33 and 26 for SVR models for the full-data and the non-valine versions of the models, respectively.

Since the ML models were trained to directly predict the experimental IR score, their threshold values were taken to be 1 in Figures 4, 5 and 6. Thus, any candidate with ML score > 1 is predicted by the models to form β-sheets. However, for the β-score the optimal threshold value of 0.61 was determined by maximizing the F1-score on the collated R0, R1 and R2 datasets. This results in artificially enhancing the performance of the β-score as the threshold is determined by fitting on the data itself. Even then it was observed that the ML models performed better than the β-score in accurately predicting β-sheet assembly.

Data Availability

The experimental IR ratio for all the pentapeptides is available as Supporting Information.

Code Availability

The machine learning code used in this work are available from the authors upon reasonable request.

Competing Interests

The authors declare no competing financial or non-financial interests.

Author Contributions

H.C.F and R.B. conceived and designed the project. Y.N.T trained the ML models. All the authors contributed to the data analysis and to the preparation of the manuscript. H.C.F and R.B. supervised and directed the overall project.

References

[1] Miri Yemini, Meital Reches, Judith Rishpon, and Ehud Gazit. Novel electrochemical biosensing platform using self-assembled peptide nanotubes. *Nano Lett.*, 5(1):183–186, 2005.

[2] Tayebeh Zohrabi, Neda Habibi, Ali Zarrabi, Maryam Fanaei, and Lai Yeng Lee. Diphenylalanine peptide nanotubes self-assembled on functionalized metal surfaces for potential application in drug-eluting stent. *J. Bio. Mater. Res. A*, 104(9):2280–2290, 2016.

[3] Kai Tao, Pandeeswar Makam, Ruth Aizen, and Ehud Gazit. Self-assembling peptide semiconductors. *Science*, 358(6365), 2017.

[4] Xuehai Yan, Pengli Zhu, and Junbai Li. Self-assembly and application of diphenylalanine-based nanostructures. *Chem. Soc. Rev.*, 39(6):1877–1890, 2010.

[5] Andrei Kholkin, Nadav Amdursky, Igor Bdikin, Ehud Gazit, and Gil Rosenman. Strong piezoelectricity in bioinspired peptide nanotubes. *ACS Nano*, 4(2):610–614, 2010.

[6] Kory Jenkins, Steve Kelly, Vu Nguyen, Ying Wu, and Rusen Yang. Piezoelectric diphenylalanine peptide for greatly improved flexible nanogenerators. *Nano Energy*, 51:317–323, 2018.

[7] Xuehai Yan, Qiang He, Kewei Wang, Li Duan, Yue Cui, and Junbai Li. Transition of cationic dipeptide nanotubes into vesicles and oligonucleotide delivery. *Angew. Chem., Int. Ed.*, 119(14):2483–2486, 2007.

[8] Sarah Guerin, Syed AM Tofail, and Damien Thompson. Organic piezoelectric materials: milestones and potential. *NPG Asia Materials*, 11(1):10, 2019.

[9] Xiubo Zhao, Fang Pan, Hai Xu, Mohammed Yaseen, Honghong Shan, Charlotte AE Hauser, Shuguang Zhang, and Jian R Lu. Molecular self-assembly and applications of designer peptide amphiphiles. *Chem. Soc. Rev.*, 39(9):3480–3498, 2010.

[10] Mischa Zelzer and Rein V Ulijn. Next-generation peptide nanomaterials: molecular networks, interfaces and supramolecular functionality. *Chem. Soc. Rev.*, 39(9):3351–3357, 2010.

[11] Honggang Cui, Matthew J Webber, and Samuel I Stupp. Self-assembly of peptide amphiphiles: from molecules to nanostructures to biomaterials. *Peptide Science: Original Research on Biomolecules*, 94(1):1–18, 2010.

[12] Caroline M Rufo, Yurii S Moroz, Olesia V Moroz, Jan Stöhr, Tyler A Smith, Xiaozhen Hu, William F DeGrado, and Ivan V Korendovych. Short peptides self-assemble to produce catalytic amyloids. *Nat. Chem.*, 6(4):303–309, 2014.

[13] Fabrizio Gelain, Zhongli Luo, and Shuguang Zhang. Self-assembling peptide eak16 and rada16 nanofiber scaffold hydrogel. *Chem. Rev.*, 120(24):13434–13460, 2020.

[14] Lee A Solomon, Matthew E Sykes, Yimin A Wu, Richard D Schaller, Gary P Wiederrecht, and H Christopher Fry. Tailorable exciton transport in doped peptide–amphiphile assemblies. *ACS Nano*, 11(9):9112–9118, 2017.

[15] Shuguang Zhang. Discovery and design of self-assembling peptides. *Interface focus*, 7(6):20170028, 2017.

[16] Gang Wei, Zhiqiang Su, Nicholas P Reynolds, Paolo Arosio, Ian W Hamley, Ehud Gazit, and Raffaele Mezzenga. Self-assembling peptide and protein amyloids: from structure to tailored function in nanotechnology. *Chem. Soc. Rev.*, 46(15):4661–4708, 2017.

[17] Liam C Palmer and Samuel I Stupp. Molecular self-assembly into one-dimensional nanostructures. *Acc. Chem. Res.*, 41(12):1674–1684, 2008.

[18] Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 28(1):235–242, 2000.

[19] Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023.

[20] One-Sun Lee, Vince Cho, and George C Schatz. Modeling the self-assembly of peptide amphiphiles into fibers using coarse-grained molecular dynamics. *Nano letters*, 12(9):4907–4913, 2012.

[21] Pim WJM Frederix, Rein V Ulijn, Neil T Hunt, and Tell Tuttle. Virtual screening for dipeptide aggregation: toward predictive tools for peptide self-assembly. *J. Phys. Chem. Lett.*, 2(19):2380–2384, 2011.

[22] Yong Sun, Xing Li, Mingda Zhao, Yafang Chen, Yang Xu, Kefeng Wang, Shaoquan Bian, Qing Jiang, Yujiang Fan, and Xingdong Zhang. Bioinspired supramolecular nanofiber hydrogel through self-assembly of biphenyl-tripeptide for tissue engineering. *Bioactive Materials*, 8:396–408, 2022.

[23] Alexander van Teijlingen, Melissa C Smith, and Tell Tuttle. Short peptide self-assembly in the martini coarse-grain force field family. *Accounts of Chemical Research*, 56(6):644–654, 2023.

[24] One-Sun Lee, Samuel I Stupp, and George C Schatz. Atomistic molecular dynamics simulations of peptide amphiphile self-assembly into cylindrical nanofibers. *Journal of the American Chemical Society*, 133(10):3677–3683, 2011.

[25] Kübra Kaygisiz, Lena Rauch-Wirth, Arghya Dutta, Xiaoqing Yu, Yuki Nagata, Tristan Bereau, Jan Münch, Christopher V Synatschke, and Tanja Weil. Data-mining unveils structure–property–activity correlation of viral infectivity enhancing self-assembling peptides. *Nature Communications*, 14(1):5121, 2023.

[26] Kübra Kaygisiz, Arghya Dutta, Lena Rauch-Wirth, Christopher V Synatschke, Jan Münch, Tristan Bereau, and Tanja Weil. Inverse design of viral infectivity-enhancing peptide fibrils from continuous protein-vector embeddings. *Biomaterials Science*, 11(15):5251–5261, 2023.

[27] 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. *Nature chemistry*, 7(1):30–37, 2015.

[28] Yuqin Yang, Xiaoyu Wang, Xialian Wu, Shuyi Guo, Haokun Yang, Junxia Lu, and Hao Dong. Computation-driven rational design of self-assembled short peptides for catalytic hydrogen production. *Journal of the American Chemical Society*, 146(19):13488–13498, 2024. PMID: 38709095.

[29] Jiaqi Wang, Zihan Liu, Shuang Zhao, Tengyan Xu, Huaimin Wang, Stan Z Li, and Wenbin Li. Deep learning empowers the discovery of self-assembling peptides with over 10 trillion sequences. *Advanced Science*, 10(31):2301544, 2023.

[30] Zihan Liu, Jiaqi Wang, Yun Luo, Shuang Zhao, Wenbin Li, and Stan Z Li. Efficient prediction of peptide self-assembly through sequential and graphical encoding. *Briefings in Bioinformatics*, 24(6):bbad409, 2023.

[31] Alexander van Teijlingen and Tell Tuttle. Beyond tripeptides two-step active machine learning for very large data sets. *Journal of Chemical Theory and Computation*, 17(5):3221–3232, 2021.

[32] Victor Daniel Aldas-Bulos and Fabien Plisson. Benchmarking protein structure predictors to assist machine learning-guided peptide discovery. *Digital Discovery*, 2(4):981–993, 2023.

[33] Tengyan Xu, Jiaqi Wang, Shuang Zhao, Dinghao Chen, Hongyue Zhang, Yu Fang, Nan Kong, Ziao Zhou, Wenbin Li, and Huaimin Wang. Accelerating the prediction and discovery of peptide hydrogels with human-in-the-loop. *Nature Communications*, 14(1), jun 2023.

[34] Silong Zhai, Yahong Tan, Chengyun Zhang, Christopher John Hipolito, Lulu Song, Cheng Zhu, Youming Zhang, Hongliang Duan, and Yizhen Yin. Pepscaf: Harnessing machine learning with in vitro selection toward de novo macrocyclic peptides against il-17c/il-17re interaction. *Journal of Medicinal Chemistry*, 66(16):11187–11200, 2023. PMID: 37480587.

[35] 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, et al. Machine learning overcomes human bias in the discovery of self-assembling peptides. *Nature chemistry*, 14(12):1427–1435, 2022.

[36] Peter Y Chou and Gerald D Fasman. Prediction of protein conformation. *Biochemistry*, 13(2):222–245, 1974.

[37] Kazuo Fujiwara, Hiromi Toda, and Masamichi Ikeguchi. Dependence of α-helical and β-sheet amino acid propensities on the overall protein fold type. *BMC structural biology*, 12:1–15, 2012.

[38] Anastasia K Varanko, Jonathan C Su, and Ashutosh Chilkoti. Elastin-like polypeptides for biomedical applications. *Annual review of biomedical engineering*, 22(1):343–369, 2020.

[39] Harald Nuhn and Harm-Anton Klok. Secondary structure formation and lcst behavior of short elastin-like peptides. *Biomacromolecules*, 9(10):2755–2763, 2008.

[40] Nan K Li, Felipe García Quiroz, Carol K Hall, Ashutosh Chilkoti, and Yaroslava G Yingling. Molecular description of the lcst behavior of an elastin-like polypeptide. *Biomacromolecules*, 15(10):3522–3530, 2014.

[41] Lisa D Muiznieks and Fred W Keeley. Proline periodicity modulates the self-assembly properties of elastin-like polypeptides. *Journal of Biological Chemistry*, 285(51):39779–39789, 2010.

[42] Pasquale Moscarelli, Federica Boraldi, Brigida Bochicchio, Antonietta Pepe, Anna Maria Salvi, and Daniela Quaglino. Structural characterization and biological properties of the amyloidogenic elastin-like peptide (vggvg) 3. *Matrix Biology*, 36:15–27, 2014.

[43] Wensi Zhang, Xiaoqing Yu, Yang Li, Zhiqiang Su, Klaus D Jandt, and Gang Wei. Protein-mimetic peptide nanofibers: Motif design, self-assembly synthesis, and sequence-specific biomedical applications. *Progress in Polymer Science*, 80:94–124, 2018.

[44] Silvia Marchesan, Attilio V Vargiu, and Katie E Styan. The phe-phe motif for peptide self-assembly in nanomedicine. *Molecules*, 20(11):19775–19788, 2015.

[45] Huayu Xiong, Brian L Buckwalter, Hong-Ming Shieh, and Michael H Hecht. Periodicity of polar and nonpolar amino acids is the major determinant of secondary structure in self-assembling oligomeric peptides. *Proceedings of the National Academy of Sciences*, 92(14):6349–6353, 1995.

[46] Shuyi Han, Sasa Cao, Yuming Wang, Jiqian Wang, Daohong Xia, Hai Xu, Xiubo Zhao, and Jian R Lu. Self-assembly of short peptide amphiphiles: the cooperative effect of hydrophobic interaction and hydrogen bonding. *Chemistry–A European Journal*, 17(46):13095–13102, 2011.

[47] WF DeGrado and JD Lear. Induction of peptide conformation at apolar water interfaces. 1. a study with model peptides of defined hydrophobic periodicity. *Journal of the American Chemical Society*, 107(25):7684–7689, 1985.

[48] Kai Tao, Haoran Wu, Lihi Adler-Abramovich, Jiahao Zhang, Xinyuan Fan, Yunxiao Wang, Yan Zhang, Syed AM Tofail, Deqing Mei, Junbai Li, et al. Aromatic short peptide architectonics: Assembly and engineering. *Progress in Materials Science*, page 101240, 2024.

[49] Machine learning based β-sheet predictor. `https://ir-score-prediction-app-bx6mcydugk2weuxcleiyhl.streamlit.app`. Accessed: 2024-09-14.

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

[51] David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988.

[52] Kong T Nguyen, Lorenz C Blum, Ruud Van Deursen, and Jean-Louis Reymond. Classification of organic molecules by molecular quantum numbers. *ChemMedChem: Chemistry Enabling Drug Discovery*, 4(11):1803–1805, 2009.

[53] Paul Labute. A widely applicable set of descriptors. *Journal of Molecular Graphics and Modelling*, 18(4-5):464–477, 2000.

[54] Lowell H Hall and Lemont B Kier. The molecular connectivity chi indexes and kappa shape indexes in structure-property modeling. *Reviews in computational chemistry*, pages 367–422, 1991.

Acknowledgements

RB acknowledges support from Center for Atomistic Modelling and Materials Design under the IOE scheme, ICSR, IIT Madras for the initiation research grants, and from Robert Bosch Centre for Data Science and AI (RBCDSAI). A part of this research was performed at the Center for Nanoscale Materials (CNM) at Argonne National Laboratory, a U.S. Department of Energy Office of Science User Facility, supported by the U.S. DOE, Office of Basic Energy Sciences, under Contract No. DE-AC02-06CH11357.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SI.pdf