PACT - Prediction of Amyloid Cross-interaction by Threading

Jakub W. Wojciechowski, Witold Szczurek, Małgorzta Kotulska
Department of Biomedical Engineering
Wrocław University of Science and Technology
jakub.wojciechowski@pwr.edu.pl, malgorzta.kotulska@pwr.edu.pl

Abstract

Amyloids are protein aggregates most commonly known for their role in the development of severe neurodegenerative diseases, such as Alzheimer's or Parkinson's disease. However, the unique features of such structures were also beneficially utilized by many organisms for a wide range of physiological roles e.g., biofilm formation and hormone storage. More recent studies have shown that in some cases presence of amyloid aggregates can affect aggregation kinetics of other proteins. These so-called cross-interactions may be crucial for understanding comorbidity of amyloid related diseases, including Alzheimer's disease and type II diabetes. Despite the importance of the process, our understanding of the phenomena is still very limited due to costly and time consuming experiments required to study theinteractions. To overcome this problem, we developed the PACT method. The method is based on modeling of heterogenous fibrils, formed by two amyloidogenic peptides. The resulting structural model is then assessed based on DOPE statistical potential implemented in the Modeller method, based on template based modeling. The main assumption of the method is that pairs of interacting amyloids will be energetically more favorable than negative cases. Importantly, the method can work with long protein fragments and, as a purely physicochemical model, it relies very little on training data. The method, for the first time, opens the possibility of high throughput study of amyloid interactions.
The tool is available at: https://github.com/KubaWojciechowski/PACT

***K*eywords** Amyloids · Cross-interactions · Computational prediction · Aggregation · Proteins

1 Introduction

Amyloids are unbranched protein fibers that produce a characteristic pattern in X-ray diffraction experiments. This pattern originates from the molecular structure of an amyloid core consisting of tightly packed beta strands that form the so-called steric zippers [1]. Despite significant structural similarities shared by all amyloids, their sequences are surprisingly diverse and have little homology [2]. For a long time, amyloids were thought to be hallmarks of severe diseases, such as Parkinson's and Alzheimer's diseases, type II diabetes, and many others. However, more recently a large number of functional amyloids were discovered. They can perform a variety of tasks in organisms across almost all domains of life, from bacteria and fungi to humans [3]. Due to their stability, mechanical and electrical properties, as well as their ability to self assemble, they are of great interest in nanotechnology [4]. Although both types of amyloids share remarkably similar structures, they often differ in their amino acid composition and solubility. Unlike pathogenic amyloids, many functional amyloids are enriched in polar amino acids and often they are more soluble than non-functional ones. The so-called low-complexity aromatic-rich kinked segments (LARKS) share many structural features with amyloids, yet their aggregation is reversible and often leads to the formation of hydrogels [5].

Aggregation of all types of amyloids is heavily dependent on the presence of short fragments in their sequences. These fragments, often called amyloidogenic hot-spots, are those that actually form steric zipper structure and, if exposed, they can drive aggregation of the whole protein [6]. For this reason, identification of such fragments is of great importance for understanding amyloids, and plenty of experimental methods have been proposed. Amyloids can be detected by X-ray diffraction method, however this technique is not widely used for them. In practice, biochemical

assays based on binding of Congo Red [7] and Thioflavin T [8] are more often used. Another approach involves direct observation of fibers using high resolution imaging techniques, such as electron microscopy [9], atomic force microscopy [10], and infrared spectroscopy. Especially the last method is widely used, due to its simplicity and efficiency [11]. Since different methods rely on different approaches, their results might differ in some cases. More importantly, all of them are expensive and time consuming. Experiments are hampered by difficulties in handling amyloids, including their low solubility, rapid aggregation, and need for their high-purity [12]. Currently, the use of experimental methods for identification of all amyloids in genome wide studies would be impossible. To address this problem several computational methods have been proposed based on different approaches (reviewed in [13]), starting from structural modeling [14], statistical analysis of the sequence including FoldAmyloid [15] and FishAmyloid [16], physicochemical models like PASTA 2.0 [17]. machine learning techniques such as APPNN [18] and AmyloGraph [19]. Furthermore, there are methods combining both approaches such as PATH [20] and Cordax [21]. Finally, some methods, like Aggrescan 3D [22], utilize information about protein structure. It was also shown that bioinformatic techniques are very robust and capable of identification even misannotated data despite being trained on them [23].

More recent studies have shown that in some cases presence of amyloid aggregates can affect the aggregation rate of other proteins [24]. Later, it was observed that interacting proteins can form heterogeneous fibers consisting of molecules of both interaction partners. Such cross-interactions were then identified between numerous proteins, including those involved in type II diabetes and neurodegenerative diseases. This shed new light on potentially new aspects regarding the origin of comorbidity of these disorders [25]. Despite the importance of this process, its mechanisms are still poorly understood, which can be attributed to a limited number of experimental data. As a result, interactions of only a few well-described proteins, such as amyloid-beta, islet amyloid polypeptide, or alpha synuclein, were extensively studied. This introduced a data bias. Despite the difficulties, it was shown that proteins with similar sequences are more likely to interact, however, many counterexamples were also shown [25]. The studies highlight the importance of the structure compatibility of amyloid cores.

Here we present a new computational tool PACT (Prediction of Amyloid Cross-interaction by Threading) for identification of potentially interacting amyloid pairs. The method is based on the molecular threading applied to the potential complex of interacting amyloids and stability assessment of the obtained molecular model.

2 Methods

2.1 Data sets

To build and test the method we have used several datasets:

- 90 amyloidogenic (amyloid) and 55 non-amyloidogenic (non-amyloid) peptides of lengths between 14 and 45 from the AmyLoad database [26].
- 119 pairs of peptides, which enhance (faster) and 73 which slow down (slower) each other aggregation from AmyloGraph database (www.amylograph.com) [27].

The first two sets (amyloid and non-amyloid) were used to test the method on cases of homoaggregation, in other words for identifying amyloid prone peptides.

For the prediction of cross-interactions, we used faster, slower and non-amyloid sets. The use of the set of non-aggregating peptides as the negative set in the interaction study was caused by the lack of a sufficient number of negative examples of non-interacting amyloid pairs. This is a common problem in the study of protein-protein interactions as negative results are rarely published, which creates a strong bias in the data [28].

2.2 Modeling

A query pair of sequences was threaded on the structure of amyloid fiber formed by Islet Amyloid Polypeptide (IAPP) [29]. In order to allow the method to deal with sequences of varying lengths, sequences shorter than the sequence of a template use only a part of a template structure. In such cases, a shorter sequence is aligned to the middle of the template sequence (Fig 1A). This choice can be justified considering that most of the currently known amyloid fragments, which are longer than a few amino acids, share a similar beta-sheet turn architecture, commonly known as the beta arch. This assumption was previously successfully applied by Ahmed and coworkers to build the ArchCandy method for amyloidogenic region prediction [30]. PACT allows sequences to be marginally longer than the template and, as a result, can be used to study cross-interactions between peptides of lengths between 14 and 45. For each of the tested pairs, 10 different models, consisting of two chains of each interacting peptide (Fig 1B) were built using Modeller 9.24 model-multichain.py procedure with default parameters [31]. Then, the model with the lowest DOPE value was

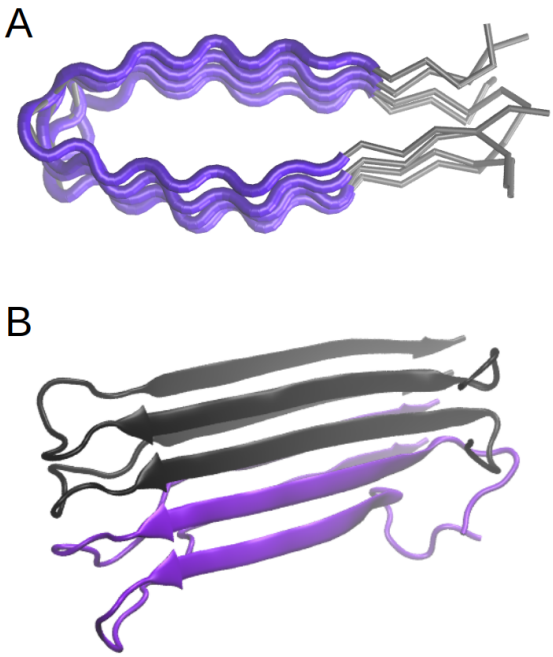

Figure 1: Schematic representation of the modeling procedure. A) In case when a query sequence is shorter than the template, only a part of it is used in modeling. B) The model of heterogenous fibril consists of two chains of each interacting peptide.

chosen for further analysis. Since the dataset consisted of fragments of varying lengths we proposed a normalized DOPE score (*ndope*) defined as follows:

$$ndope = \frac{DOPE}{L} \tag{1}$$

where L is an average length of sequences used to build a given model. Then, *ndope* scores were compared between amyloids and non-amyloids, as well as between pairs of amyloids interacting with non-amyloids.

The main assumption of the method is that interactions between amyloidogenic fragments that cross-interact, threaded into an amyloid fiber structure, would result in a heterogeneous aggregate which is more stable, thus energetically favorable, than in the case of a non-interacting pair. A similar assumption was successfully applied in our previous work to predict the aggregation of short amyloidogenic fragments [20]. To choose the *ndope* threshold for the classification, Receiver Operating Characteristics (ROC) curve was calculated by applying different score thresholds and recording False Positive Ratio (FPR) and True Positive Ratio (TPR). The threshold closest to the (0,1) point (representing perfect classification) was chosen.

We have also tested a variant of the method which utilized three different structural templates but it did not improve the accuracy of the method but significantly increased the computational time.

2.3 Assessment of performance and data analysis

All the data analysis was performed using Python 3.8 with Matplotlib [32], NumPy [33], Pandas [34], Scikit-Learn [35], and Seaborn [36] packages.

To test the performance of the proposed method, a dataset was randomly split into a training set and test set, which consisted of 30% of the data. Additionally, k-folds cross-validation (with k=5) was performed on the training data. Area Under ROC curve (AUC), accuracy (ACC), sensitivity (Sens), specificity (Spec) and Matthew Correlation Coefficient (MCC) were used to assess the performance of the method

3 Results

In the first step, we compared *ndope* scores obtained for models of homoagregates of amyloidogenic and non-amyloidogenic peptides from the AmyLoad database [26]. (Sup Fig 1) shows *ndope* scores for both types of peptides. The majority of models obtained for amyloidogenic peptides show much lower scores in comparison with those for non-amyloids. It allowed for calculating the threshold that could be used for the classification. Such a simple classifier, here distinguishing between amyloid and non-amyloid peptides, was able to achieve AUC of 0.76 and accuracy of 0.74, which is comparable with state-of-the-art amyloid predictors (Sup Tab 1).

We also tested if the method is capable of classifying functional amyloids, which are a major problem for most predictors due to their underrepresentation in databases of amyloids. We tested the performance of the method on imperfect repeats of CsgA protein from E. coli and S. enterica [11] (Sup Fig 2). Aggregation prone regions R1, R3, and R5 scored much lower than non-amyloidogenic regions R2 and R4 from E. coli. On this data, PACT achieved an accuracy of 0.9. Furthermore, the observed difference in *ndope* score for R4 fragments from E. coli and S. enterica corresponds very well to the difference in their aggregation propensity [11].

The results showed that the method can accurately predict aggregation prone peptides of varying lengths. Furthermore, it can be utilized to detect functional amyloids.

Then we tested the ability of the method to predict cross-interactions of amyloid peptides. The *ndope* scores were compared or models of heteroaggregates consisting of pairs of peptides whose cross-interactions resulted in faster aggregation with non-amyloidogenic pairs of peptides (Fig 2). A similar analysis was performed for pairs of peptides whose cross-interactions resulted in slower aggregation (Sup Fig 3). In both cases, models of heterologous aggregates resulting from cross-interactions showed lower values of *ndope* scores. Based on these results, a simple threshold based classifier was built.

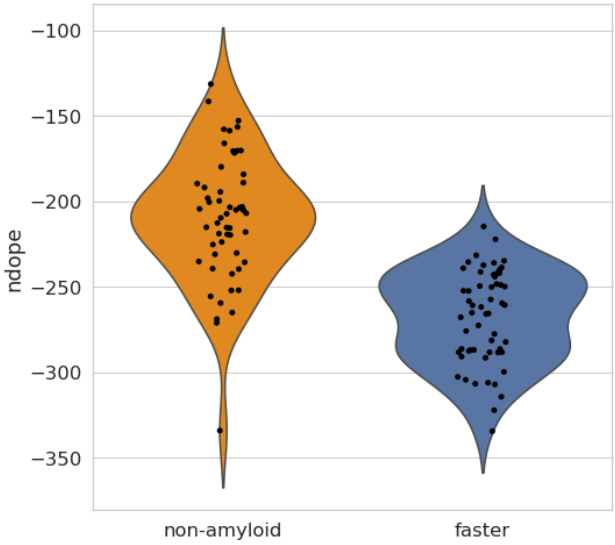

Figure 2: The score *ndope* for models of interacting identical non-amyloidogenic peptides (negative set) and interacting pairs resulting in increased aggregation rates (faster set).

To assess the performance and choose the threshold value, ROC curves were calculated for both cases; faster rate vs negative (Fig 3) and slower vs negative (Sup Fig 4) on both training and test sets. To minimize the impact of the choice of data, we performed k-folds cross-validation with k=5 on the training set and calculated several metrics describing the performance of the method (Tab 1). The same metrics were then calculated on an independent test set (Sup Tab 2). PACT performed well on both cross validation and tested on the independent data sets. It achieved an accuracy of 0.85 and 0.82 on test sets of faster vs negative and slower vs negative cases. In all cases, the results obtained on the test set were within the value of one standard deviation range away from mean values obtained on cross validation. The method performance was quite similar in both faster vs negative and slower vs negative scenarios. However, due to the smaller dataset size, larger standard deviation was obtained for slower vs negative scenario (Tab.1). The results show that the method can predict whether two peptides can cross-interact but cannot distinguish between types of interactions with regard to the rate.

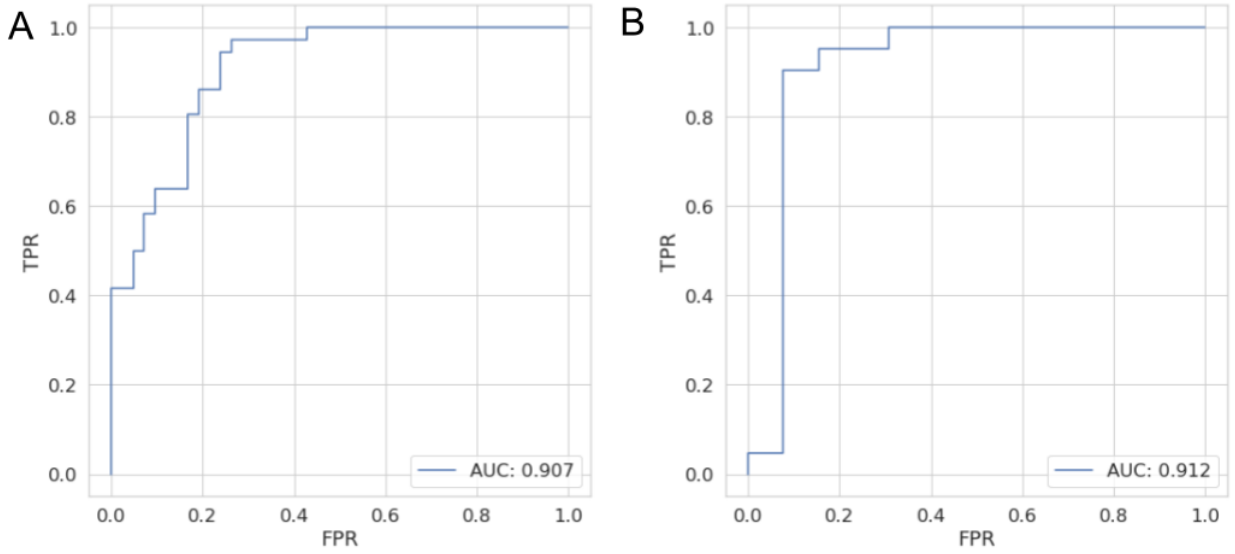

Figure 3: ROC curves for classification of non-aggregating and cross interacting pairs resulting in faster aggregation on (A) training and (B) test set.

Table 1: Performance of PACT on cross validation and independent test set for classification of non-aggregating and cross interacting pairs resulting in faster aggregation

	Accuracy [std]	Sensitivity [std]	Specificity [std]	F1 [std]	MCC [std]
Cross-validation	0.82 [0.06]	0.90 [0.09]	0.77 [0.08]	0.82 [0.07]	0.66 [0.12]
Test set	0.85	0.81	0.92	0.87	0.71

A serious problem with data regarding interacting amyloids, available in the literature and, therefore, in our dataset is large overrepresentation of interactions concerning amyloid beta peptide. This may cause overfitting of the method to this group of sequences. To assess its effect, we analyzed scores for interactions between different amyloid beta variants (abeta) (Sup Fig 5). Observed scores for abeta pairs are within a similar range as values for the remaining pairs, and therefore it should not have a significant effect on the performance of the method. Those pairs obtained quite similar *ndope* values, centered slightly below *ndope* value of -250, which is relatively close to identified classification threshold of -236, while the remaining interacting pairs showed even lower scores. PACT was implemented as an open source python module, available at GitHub repository: `https://github.com/KubaWojciechowski/PACT`. For users convenience we prepared docker container for the application.

4 Conclusions

High throughput identification of amyloid cross-interaction is an important step toward our understanding of this mechanism. This can allow better understanding of the principles governing the process but it can also be used to identify novel cases of amyloid interactions. Such capabilities can shed light on a possible mechanism responsible for the comorbidity of devastating disorders.

We proposed the first computational method for predicting amyloid cross-interactions. It is based on a highly interpretable and well-established physicochemical model which is not heavily dependent on training data. This feature is especially important since the available data contains a strong interest bias toward interactions of a few popular amyloids related to neurodegenerative diseases such as abeta. We have carefully studied the effect of this overrepresentation and showed that it does not affect the performance of the method. Furthermore, good performance on functional amyloids suggests that the method is very robust and can be effectively used on a wide range of sequences. In total, PACT achieved a high accuracy of 0.85 and 0.82 on the independent test sets of interactions increasing and decreasing aggregation rates. On both sets the method achieved high AUC values of 0.91 and F1 values of 0.81 and 0.80. On the other hand, since both cases were characterized by similar interaction energies, the method cannot distinguish between

enhancement and inhibition of aggregation cases. These results suggest that both processes may be driven by similar mechanisms. This issue was addressed in a recently published work by Louros and coworkers [37], who applied a somewhat similar approach to study the effect of point mutations on aggregation characteristics.

Apart from identification of amyloid cross-interactions, the proposed method is also capable of reliably predicting amyloid prone regions in proteins with comparable accuracy to the state-of-the-art techniques. Furthermore, it overcomes their major limitations regarding the identification of functional amyloids. Unlike most of the currently available amyloid predictors, it does not rely on scanning of the sequence of interest with a short sliding window.

5 Acknowledgments

We would like to thank team of AmyloGraph database developers for their effort in preparing database used in this project.

This work was partially supported by the National Science Centre, Poland, Grant 2019/35/B/NZ2/03997. Access to Wroclaw Centre for Networking and Supercomputing is greatly acknowledged. Funding was provided by Wroclawskie Centrum Sieciowo-Superkomputerowe, Politechnika Wroclawska (Grant Number 98).

References

[1] David S Eisenberg and Michael R Sawaya. Structural studies of amyloid proteins at the molecular level. *Annu. Rev. Biochem*, 86(1):69–95, 2017.

[2] Jean D Sipe, Merrill D Benson, Joel N Buxbaum, Shu-Ichi Ikeda, Giampaolo Merlini, Maria JM Saraiva, and Per Westermark. Amyloid fibril protein nomenclature: 2010 recommendations from the nomenclature committee of the international society of amyloidosis. *Amyloid*, 17(3-4):101–104, 2010.

[3] Ryan P McGlinchey and Jennifer C Lee. Why study functional amyloids? lessons from the repeat domain of pmel17. *Journal of molecular biology*, 430(20):3696–3706, 2018.

[4] Noémie-Manuelle Dorval Courchesne, Elizabeth P DeBenedictis, Jason Tresback, Jessica J Kim, Anna Duraj-Thatte, David Zanuy, Sinan Keten, and Neel S Joshi. Biomimetic engineering of conductive curli protein films. *Nanotechnology*, 29(45):454002, 2018.

[5] Michael P Hughes, Michael R Sawaya, David R Boyer, Lukasz Goldschmidt, Jose A Rodriguez, Duilio Cascio, Lisa Chong, Tamir Gonen, and David S Eisenberg. Atomic structures of low-complexity protein segments reveal kinked β sheets that assemble networks. *Science*, 359(6376):698–701, 2018.

[6] Manuela López de la Paz and Luis Serrano. Sequence determinants of amyloid fibril formation. *Proceedings of the National Academy of Sciences*, 101(1):87–92, 2004.

[7] Alexander J Howie and Douglas B Brewer. Optical properties of amyloid stained by congo red: history and mechanisms. *Micron*, 40(3):285–301, 2009.

[8] Liza Nielsen, Sven Frokjaer, Jens Brange, Vladimir N Uversky, and Anthony L Fink. Probing the mechanism of insulin fibril formation with insulin mutants. *Biochemistry*, 40(28):8397–8409, 2001.

[9] Tsuranobu Shirahama and Alan S Cohen. High-resolution electron microscopic analysis of the amyloid fibril. *The Journal of cell biology*, 33(3):679–708, 1967.

[10] Zhigang Wang, Chunqing Zhou, Chen Wang, Lijun Wan, Xiaohong Fang, and Chunli Bai. Afm and stm study of β-amyloid aggregation on graphite. *Ultramicroscopy*, 97(1-4):73–79, 2003.

[11] Natalia Szulc, Marlena Gąsior-Głogowska, Jakub W Wojciechowski, Monika Szefczyk, Andrzej M Żak, Michał Burdukiewicz, and Malgorzata Kotulska. Variability of amyloid propensity in imperfect repeats of csga protein of salmonella enterica and escherichia coli. *International journal of molecular sciences*, 22(10):5127, 2021.

[12] Marlena E Gąsior-Głogowska, Natalia Szulc, and Monika Szefczyk. Challenges in experimental methods. In *Computer Simulations of Aggregation of Proteins and Peptides*, pages 281–307. Springer, 2022.

[13] Małgorzata Kotulska and Jakub W Wojciechowski. Bioinformatics methods in predicting amyloid propensity of peptides and proteins. In *Computer Simulations of Aggregation of Proteins and Peptides*, pages 1–15. Springer, 2022.

[14] Michael J Thompson, Stuart A Sievers, John Karanicolas, Magdalena I Ivanova, David Baker, and David Eisenberg. The 3d profile method for identifying fibril-forming segments of proteins. *Proceedings of the National Academy of Sciences*, 103(11):4074–4078, 2006.

[15] Sergiy O Garbuzynskiy, Michail Yu Lobanov, and Oxana V Galzitskaya. Foldamyloid: a method of prediction of amyloidogenic regions from protein sequence. *Bioinformatics*, 26(3):326–332, 2010.

[16] Pawel Gasior and Malgorzata Kotulska. Fish amyloid–a new method for finding amyloidogenic segments in proteins based on site specific co-occurence of aminoacids. *BMC bioinformatics*, 15(1):1–8, 2014.

[17] Ian Walsh, Flavio Seno, Silvio CE Tosatto, and Antonio Trovato. Pasta 2.0: an improved server for protein aggregation prediction. *Nucleic acids research*, 42(W1):W301–W307, 2014.

[18] Carlos Família, Sarah R Dennison, Alexandre Quintas, and David A Phoenix. Prediction of peptide and protein propensity for amyloid formation. *PloS one*, 10(8):e0134679, 2015.

[19] Michał Burdukiewicz, Piotr Sobczyk, Stefan Rödiger, Anna Duda-Madej, Paweł Mackiewicz, and Małgorzata Kotulska. Amyloidogenic motifs revealed by n-gram analysis. *Scientific reports*, 7(1):1–10, 2017.

[20] Jakub W Wojciechowski and Małgorzata Kotulska. Path-prediction of amyloidogenicity by threading and machine learning. *Scientific Reports*, 10(1):1–9, 2020.

[21] Nikolaos Louros, Gabriele Orlando, Matthias De Vleeschouwer, Frederic Rousseau, and Joost Schymkowitz. Structure-based machine-guided mapping of amyloid sequence space reveals uncharted sequence clusters with higher solubilities. *Nature communications*, 11(1):1–13, 2020.

[22] Aleksander Kuriata, Valentin Iglesias, Jordi Pujols, Mateusz Kurcinski, Sebastian Kmiecik, and Salvador Ventura. Aggrescan3d (a3d) 2.0: prediction and engineering of protein solubility. *Nucleic acids research*, 47(W1):W300–W307, 2019.

[23] Natalia Szulc, Michał Burdukiewicz, Marlena Gąsior-Głogowska, Jakub W Wojciechowski, Jarosław Chilimoniuk, Paweł Mackiewicz, Tomas Šneideris, Vytautas Smirnovas, and Malgorzata Kotulska. Bioinformatics methods for identification of amyloidogenic peptides show robustness to misannotated training data. *Scientific reports*, 11(1):1–11, 2021.

[24] Katarzyna Lundmark, Gunilla T Westermark, Arne Olsén, and Per Westermark. Protein fibrils in nature can enhance amyloid protein a amyloidosis in mice: Cross-seeding as a disease mechanism. *Proceedings of the National Academy of Sciences*, 102(17):6098–6102, 2005.

[25] Baiping Ren, Yanxian Zhang, Mingzhen Zhang, Yonglan Liu, Dong Zhang, Xiong Gong, Zhangqi Feng, Jianxin Tang, Yung Chang, and Jie Zheng. Fundamentals of cross-seeding of amyloid proteins: an introduction. *Journal of Materials Chemistry B*, 7(46):7267–7282, 2019.

[26] Pawel P Wozniak and Malgorzata Kotulska. Amyload: website dedicated to amyloidogenic protein fragments. *Bioinformatics*, 31(20):3395–3397, 2015.

[27] Amylograph database. `www.amylograph.com`. Accessed: 2022-05-05.

[28] Alicja W Nowakowska and Malgorzata Kotulska. Topological analysis as a tool for detection of abnormalities in protein-protein interaction data. *Bioinformatics*, 2022.

[29] Sorin Luca, Wai-Ming Yau, Richard Leapman, and Robert Tycko. Peptide conformation and supramolecular organization in amylin fibrils: constraints from solid-state nmr. *Biochemistry*, 46(47):13505–13522, 2007.

[30] Abdullah B Ahmed, Nadia Znassi, Marie-Thérèse Château, and Andrey V Kajava. A structure-based approach to predict predisposition to amyloidosis. *Alzheimer's & Dementia*, 11(6):681–690, 2015.

[31] Andrej Šali and Tom L Blundell. Comparative protein modelling by satisfaction of spatial restraints. *Journal of molecular biology*, 234(3):779–815, 1993.

[32] John D Hunter. Matplotlib: A 2d graphics environment. *Computing in science & engineering*, 9(03):90–95, 2007.

[33] Charles R Harris, K Jarrod Millman, Stéfan J Van Der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J Smith, et al. Array programming with numpy. *Nature*, 585(7825):357–362, 2020.

[34] Wes McKinney et al. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56. Austin, TX, 2010.

[35] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[36] Michael L Waskom. Seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021, 2021.

[37] Nikolaos Louros, Meine Ramakers, Emiel Michiels, Katerina Konstantoulea, Chiara Morelli, Teresa Garcia, Nele Moonen, Sam D'Haeyer, Vera Goossens, Dietmar Rudolf Thal, et al. Mapping the sequence specificity of heterotypic amyloid interactions enables the identification of aggregation modifiers. *Nature communications*, 13(1):1–20, 2022.