

AggreProt: a web server for predicting and engineering aggregation prone regions in proteins

Joan Planas-Iglesias^{1,2,†}, Simeon Borko^{1,2,†}, Jan Swiatkowski^{1,2,†}, Matej Elias³,
Martin Havlasek^{1,2}, Ondrej Salamon^{1,2}, Ekaterina Grakova^{1,2}, Antonín Kunka^{1,2},
Tomas Martinovic^{1,2}, Jiri Damborsky^{1,2}, Jan Martinovic^{1,2,*} and David Bednar^{1,2,*}

¹Loschmidt Laboratories, Department of Experimental Biology and RECETOX, Faculty of Science, Masaryk University, Brno, Czech Republic

²International Clinical Research Center, St. Anne's University Hospital Brno, Brno, Czech Republic

³IT4Innovations, VSB – Technical University of Ostrava, 17. listopadu 2172/15, 708 00 Ostrava-Poruba, Czech Republic

*To whom correspondence should be addressed. Tel: +420 605143394; Email: davidbednar1208@gmail.com

Correspondence may also be addressed to Jan Martinovic. Tel: +420 596999598; Email: jan.martinovic@vsb.cz

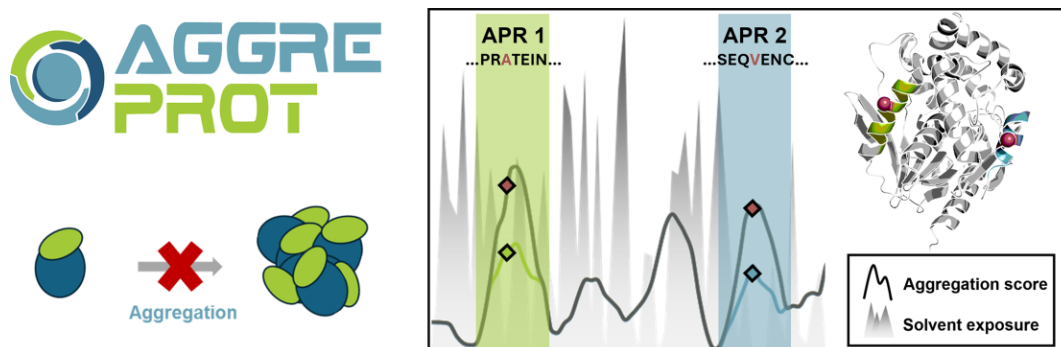
†The first three authors should be regarded as Joint First Authors.

Present address: Antonín Kunka, Protein Biophysics Group, Department of Biotechnology and Biomedicine, Technical University of Denmark, Søtofts Plads, Building 227, 2800, Kgs. Lyngby, Denmark.

Abstract

Recombinant proteins play pivotal roles in numerous applications including industrial biocatalysts or therapeutics. Despite the recent progress in computational protein structure prediction, protein solubility and reduced aggregation propensity remain challenging attributes to design. Identification of aggregation-prone regions is essential for understanding misfolding diseases or designing efficient protein-based technologies, and as such has a great socio-economic impact. Here, we introduce AggreProt, a user-friendly webserver that automatically exploits an ensemble of deep neural networks to predict aggregation-prone regions (APRs) in protein sequences. Trained on experimentally evaluated hexapeptides, AggreProt compares to or outperforms state-of-the-art algorithms on two independent benchmark datasets. The server provides per-residue aggregation profiles along with information on solvent accessibility and transmembrane propensity within an intuitive interface with interactive sequence and structure viewers for comprehensive analysis. We demonstrate AggreProt efficacy in predicting differential aggregation behaviours in proteins on several use cases, which emphasize its potential for guiding protein engineering strategies towards decreased aggregation propensity and improved solubility. The webserver is freely available and accessible at <https://loschmidt.chemi.muni.cz/aggreprot/>.

Graphical abstract



Introduction

Recombinant proteins, versatile in applications from industrial biocatalysts to therapeutics, benefit from the remarkable progress in computational prediction of protein structures (1,2). Despite accurately and rapidly predicting protein structures (3–5), new computational algorithms provide a single conformation corresponding to the global minimum of the free energy landscape without assuming anything regarding the underlying folding pathways (6,7). *In vivo*, protein folding is primarily driven by the burial of hydrophobic residues

whose exposure can lead to non-native self-association, misfolding and, ultimately, aggregation. The formation of such misfolded aggregates is triggered by various factors (8) and is associated with severe pathologies such as Alzheimer's or Parkinson's diseases (9). Additionally, binding sites, oligomerization interfaces, or other surface residues important for protein function *in vivo* can often act as potential APRs outside the native environment of proteins, e.g. during recombinant production in host organisms, or in a buffer in their purified form. Consequently, formation of inclusion bodies, low

Received: March 10, 2024. Revised: April 23, 2024. Editorial Decision: May 7, 2024. Accepted: May 13, 2024

© The Author(s) 2024. Published by Oxford University Press on behalf of Nucleic Acids Research.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

yields, and aggregation/precipitation of purified proteins are commonly encountered nuisances during protein production. Fast and accurate identification of APRs is therefore crucial for mitigating these issues and improving the efficiency of recombinant protein technologies. Among aggregates, amyloids represent a specific class characterized by highly organized two-dimensional structures. Amyloids are formed by stacked repetitive units of protein molecules stabilized by an intermolecular network of hydrogen bonds within their cross- β -sheet architecture (10) which, however, often adopt different morphologies (11,12). Nonetheless, they share a common structural kernel which is believed to be an important driver of amyloid formation and essential for its stability (13–15). These regions (APRs) are therefore perfect targets for designing mutations that decrease aggregation propensity and, consequently, improve protein solubility.

Several algorithms have been designed to address aggregation challenges. Depending on the type of input data they accept, such algorithms are classified as sequential or structural predictors. Sequential or linear predictors are typically fast, and rely either on experimentally derived scales of aggregation propensity, like in AGGRESCAN (16) or CamSol intrinsic (17), or features based on the protein amino acid composition (frequency distribution, physico-chemical properties, β -sheet propensity, hydrogen-bond optimal packing density), like in Waltz (18), Tango (19), PASTA (20), SALSA (21) or FoldAmyloid (22). Structural methods require a three-dimensional protein structure as an input, exploit structural features such as solvent accessibility or analysis of interactions, and are consequently slower. Their major advantage is the ability to identify structural APRs formed by sequentially distant residues which are impossible to detect by the linear predictors. An example of a structural predictor is AGGRESCAN 3D (23), which also considers protein dynamics in the aggregation prediction. The algorithms from both classes greatly contributed to our understanding of protein aggregation and solubility on a molecular level, and are being frequently used to identify APRs in proteins with varying degrees of success.

Over the past several years, a third generation of predictors emerged based on machine learning approaches such as support vector machine in the Budapest Amyloid Predictor (24), random forest classifiers in RF Amyloid (25) and Amylogram (26), and many others (27) including ANuPP (28), FishAmyloid (29) or CORDAX (30). The availability of computational power and advances in the ML field make the development of this class of predictors highly appealing. However, discerning molecular features important for aggregation is difficult due to the convoluted nature of the underlying calculations and therefore, the true impact of these predictors is yet to be seen based on the successful use cases, which are so far scarce. One of the main problems in engaging the prediction of APRs in proteins is the lack of training and reference data. Currently, WaltzDB (31), a database containing 1416 annotated hexapeptides entries (as of February 2024), including 515 and 901 amyloid and non-amyloid forming peptides, respectively, is the most comprehensive dataset for training aggregation predictors. Their aggregation propensity has been verified using several experimental techniques, including secondary structure determination, amyloid-specific dye binding, and electron microscopy imaging. Moreover, each hexapeptide entry includes annotations on several calculated energetic parameters from FoldX (32) and secondary structure prediction and architectural zipper class predicted by CORDAX

(30). In contrast, two other curated databases, AmyPro (33) and CPAD 2.0 (34), compile protein sequences with annotated APRs based on the literature search and are widely used for predictor benchmarking. However, we and other groups have recently contested this relatively common practice by experimentally investigating the amyloid propensity of some of the sequences and revealed many inaccuracies in the database labels (30,35). As a result, comparing protein-level predictions of aggregation propensity between different algorithms based on these databases is potentially biased, and their performance should always be validated by an experiment. Due to the scarcity of training data, the machine learning-based predictors are often hexapeptide-centred, making it difficult to get a profile view of the complete protein. Furthermore, those methods often lack contextual information such as three-dimensional structure, prediction of solvent accessibility, or transmembrane propensity, which need to be taken into account for a successful protein engineering campaign (35–37).

To provide the community with an easy-to-use predictor of protein aggregation and engineering tool, we have developed a sequence-based, user-friendly webserver for predicting APRs in proteins, AggreProt (<https://loschmidt.chemi.muni.cz/aggreprot/>). The server features an ensemble of five deep neural networks (DNNs) that output a single per-residue aggregation profile for up to three possible input protein sequences. Additionally, our server provides information on the solvent-accessible area and transmembrane propensity to aid users in distinguishing potential surface-exposed APRs from hydrophobic cores. The graphical interface features an interactive sequence and structure viewer, allowing the comparison of profiles from multiple proteins at once. Trained on a hexapeptides set with experimentally evaluated aggregation propensities, AggreProt outperforms or is comparable to state-of-the-art algorithms during benchmarking on training (WaltzDB) and validation (AmyPro) datasets. We successfully applied our predictor to solubilise the aggregating haloalkane dehalogenase LinB (35) and poorly soluble luciferase from *Amphiura filiformis* (in preparation). Here we illustrate how AggreProt can pinpoint differential aggregation behaviours in several proteins annotated in the recently released *in-house* database SoluProtMut^{DB} (38).

Materials and methods

Datasets

Two different datasets were used to train, validate, and test the deep neural networks (DNNs) that are at the core of AggreProt. First, WaltzDB (31) contains approximately 1400 hexapeptides, of which nearly 500 are labelled as amyloid prone and the rest (cca. 900) are not. WaltzDB was randomly split into training and testing datasets. 90% of the stratified contents of the database (keeping the label ratio) were used to train and optimise the parameters of our DNNs during their training and hyperparameter search process (WaltzDB-training). The remaining 10% was kept for testing (WaltzDB-testing). The second data set employed in testing was AmyPro (33). It consists of the sequences of 162 amyloid proteins with their amyloid-prone regions (APRs) annotated. Only 37 of such proteins do not enclose any hexapeptide present in WaltzDB (AmyPro37). Among these, 10 correspond to proteins with very long (> 50 amino acids) APRs annotated. The remaining 27 proteins (conforming AmyPro27 dataset,

Table 1. Details of network optimization

| Optimized parameter | Value range |
|--------------------------------|---|
| Number of bidirectional layers | [1,2] |
| Number of dense layers | [1,2,3] |
| Dropout values | [0–0.8], in 0.2 step range |
| Number of neurons per layer | [8–512], in powers of 2 step range |
| Batch size | [16, 32] |
| Learning rate | [10 ⁻⁴ –10 ⁻¹], in powers of 10 step range |

Supplementary File S1) were used to test the performance of our DNNs in the domain of whole protein sequences.

Metrics

The performance of the DNNs was assessed with standard metrics derived from the confusion matrix. Considering prediction and ground truth (experimental APRs), *True Positives (TP)* are defined as ground truth positives (amyloid prone) predicted as positives; *False Positives (FP)* as ground truth negatives (non-amyloid prone) predicted as positives; by opposition, ground truth positives predicted as negatives define the set of *False Negatives (FN)*; and finally, ground truth negatives predicted as negatives conform the *True Negatives (TN)*. From these definitions, the following are derived:

$$\text{Precision (Prec)} = \text{TP} / (\text{TP} + \text{FP}) \quad (1)$$

$$\text{Recall (Rec)} = \text{TP} / (\text{TP} + \text{FN}) \quad (2)$$

$$\text{Specificity (Spec)} = \text{TN} / (\text{FP} + \text{TN}) \quad (3)$$

The performance of the networks was assessed as the area under (Au) the Receiver Operating Characteristic (Recall versus 1 – specificity) and Precision/Recall curves (ROCC and PRC respectively). On ROCCs, Youden's *J* statistic (39) was used to determine the optimal threshold for prediction.

Network features and hyperparameter search

Our DNNs predict the aggregation propensity of an input hexapeptide sequence, considering the atomic feature description of each amino acid in the hexapeptide as employed in ANuPP (28). Thus, the first input layer dimension (number of neurons) corresponds to the number of atomic features, 36 per each residue in the hexapeptide. During the training procedure, the architecture and internal parameters of the networks were optimized. To this end, WaltzDB-training was used in a 5-fold cross-validation fashion. The optimised parameters and the search range are summarized in Table 1.

A total of 841 hyperparameter combinations were examined, each resulting network evaluated in terms of AuROCC. The top performing architecture, summarized in Table 2, provided five individual predictors, one for each data split in the 5-fold cross-validation procedure. The models were obtained at the training epoch where the training and validation ROC curves diverged, to avoid overfitting (see Results). The final AggreProt predictor is thus an ensemble of five DNNs (consensus values aggregated across the ensemble by average). Each of the networks is formed by an architecture composed of eight layers. All the training procedure was performed using a custom codebase developed in the scope of this work, based on the TensorFlow Framework (40).

Table 2. Architecture of AggreProt deep neural networks (DNNs)

| Layer type | Input neurons | Output neurons | Layer no. |
|----------------------|---------------|----------------|-----------|
| Input | 36*6 | 64*6 | 1 |
| Bidirectional (LSTM) | 64*6 | 64 | 2 |
| Bidirectional (LSTM) | 64 | 96 | 3 |
| Dense + dropout | 96 | 32 | 4–5 |
| Dense + dropout | 32 | 32 | 6–7 |
| Dense | 32 | 1 | 8 |

LSTM stands for long short-term memory.

Deep neural networks testing

The remaining 10% of hexapeptides not used during the training procedure (WaltzDB-testing) were used to test the performance of the resulting networks. The performance was evaluated in terms of the area under the ROC curve (AuROCC). To identify amyloid-prone regions (APRs) in the complete sequence of proteins, protein sequences were fragmented into overlapping hexapeptides using a six-amino-acid- windows shifted one amino acid at a time. Each of the resulting hexapeptides was then evaluated by the DNN ensemble, and the results were aggregated (averaged) per residue using a sliding window procedure mirroring the previous one. Protein sequences in the AmyPro27 dataset were evaluated by this procedure, and the performance of the network was assessed in terms of area under the ROC and PR curves (AuROCC and AuPRC, respectively). Finally, the Segment Overlap score (SOV), as defined by Rost and co-workers (41) and used in ANuPP (28), was also used to evaluate the correctness of the predictions in a whole sequence context. The SOV score can be calculated both for APR and non-APR segments. Other state-of-the-art predictors that base their predictions on the sole input of sequence, namely Waltz (18), Tango (19), ANuPP (28) and AGGRESCAN (16) were also evaluated using AmyPro27 and the aforementioned metrics. For the servers that limit the length of the input, several requests were made per each protein in AmyPro27 (Waltz and Tango), and their results were co-catenated.

Prediction of protein structure, transmembrane propensity, and solvent accessible surface area. Alignment of sequences and three-dimensional (3D) visualization

The input of protein structure is optional; hence, when only sequence is provided AggreProt fetches the closest sequence from AlphaFoldDB (42) using BLAST (43), provided that the *E*-value of the hit from AlphaFoldDB is at maximum 1e10⁻⁸⁰. Transmembrane (TM) propensity is predicted from sequence using TMHMM 2.0 (44). The Solvent Accessible Surface Area (SASA) is calculated only if a structure is provided or obtained from AlphaFoldDB, and the values are calculated by DSSP (45), using the SBI library (46). Sequences are aligned using the Needleman and Wunch alignment algorithm from BioJava (47). The 3D view pane implements Mol* (48).

Results

Performance on hexapeptides

The performance of AggreProt DNNs expectedly varied depending on the data used during the training procedure. While on the training set the AuROCC could reach 100%, it never

exceeded 90% on the validation set which provided us with the overfitting limits of the predictor. We selected the training epoch where the training and validation ROC curves diverged as the final predictive models to avoid overfitting (see Methods, [Supplementary Figure S1A](#)). The best DNN ensemble achieved an AuROCC of 88.7% ([Supplementary Figure S1B](#)) on the independent WalztDB-test set.

Performance on whole protein sequences and comparison to the state-of-the-art methods

In comparison to predicting the aggregation propensity of isolated hexapeptides, working the whole protein sequence requires further processing. First, the input sequence needs to be converted into a collection of overlapping hexapeptides. When the hexapeptide-level predictions are done, the results should be aggregated to the residue level. For validation purposes, we selected 27 sequences from AmyPro (AmyPro27) obtained after (i) filtering to remove overlaps with the training dataset (37/160), and (ii) removal of sequences containing long APRs (>50 aa, 10/37) mostly corresponding to low complexity, prion-like domains (49–51). Using AmyPro27, we compared the prediction from AggreProt to those obtained by four different sequence- or ML-based state-of-the-art methods, namely Waltz (18), Tango (19), ANuPP (28), and AGGRESAN (16). We observe a great disparity of AuROCC and AuPRC values for AggreProt performance, (see [Supplementary File S2](#)). The average performance values across the dataset reach 0.64 and 0.32 in terms of AuROCC and AuPRC, respectively. It must be noted that, if proteins with long APRs (larger than 50 amino acids long) are considered (AmyPro37) these values drop to 0.53 and 0.23, respectively. This indicates that if such long APRs correspond to relevant biological events (e.g. low complexity domains, LCDs), then the chances of detecting them by our hexapeptide-trained predictor are lower. We have recently described and experimentally verified that this drop in statistics is, at least partly, due to miss-annotations in the validation dataset (35). Our findings were in agreement with those of Louros and co-authors (30), and stress the correspondence of AggreProt predictions at hexapeptide or whole-protein levels on experimental data (35). Nonetheless, the predictor can still resolve short (between 5 and 10) residues stretches within the long-annotated APRs that often overlap with the amyloid aggregation kernels found within the cores (13–15). In comparison, other sequence-based state-of-the-art methods do not perform significantly better in our blind independent testing set. Whether on residue-level or SOV validation, AggreProt reached similar or better evaluation performance compared to other sequence-based methods (Table 3). The only notable exceptions are PASTA (20) and SALSA (21), for which their training sets significantly overlap with our validation AmyPro27 one (13 and 15 out of 33 training proteins, respectively).

We extended our comparison to an example of structure-based methods, Aggrescan3D (23), which overall achieves similar results in all metrics except in overlapping true APRs, where it stands over all other compared predictors. Finally, to test AggreProt performance on sequences with low complexity, we further analysed set of 10 proteins containing LCDs or prion-like domains (PrLDs) and compared the results with sequence regions known to form amyloids ([Supplementary Note 1, Supplementary Figure S2](#)).

Table 3. Evaluation of different sequence-based aggregation propensity predictors on AmyPro27

| Predictor | AuROCC | AuPRC | SOV APR | SOV Non-APR | Reference |
|--------------------------|-------------|-------------|-------------|-------------|------------|
| AggreProt | 0.64 | 0.32 | 54.8 | 41.7 | This study |
| Waltz | 0.57 | 0.24 | 30.5 | 56.0 | (18) |
| Tango | 0.64 | 0.33 | 52.4 | 61.5 | (19) |
| ANuPP | 0.64 | 0.27 | 54.8 | 57.3 | (28) |
| Aggrescan | 0.64 | 0.33 | 54.7 | 47.0 | (16) |
| Fold Amyloid | 0.62 | 0.30 | 50.1 | 47.7 | |
| PASTA* | 0.72 | 0.38 | 7.14 | 47.1 | (20) |
| SALSA* | 0.67 | 0.35 | 5.3 | 14.9 | |
| Camsol | 0.36 | 0.16 | 16.3 | 19.1 | |
| Aggrescan3D [§] | 0.57 | 0.26 | 71.5 | 17.8 | |

The top-performing predictors in each category are shown in boldface. AuROCC and AuPRC stand for area under the Receiver Operating Characteristic and Precision Recall curves, respectively. SOV stands for Segment Overlap score (41), which can be calculated over aggregation prone regions (APR) or non-aggregation prone regions (Non-APR). * PASTA and SALSA performance is biased since a large portion of their training sets overlaps with our AmyPro27 validation set (13 and 15 out of 33 proteins, respectively). [§]Aggrescan3D belongs to a different family of predictors, those that consider the structure of the protein as an input.

Experimental validation

A recurring problem in the protein amyloid aggregation domain is the validity of the standard datasets used (30). To face this problem, we engaged a systematic experimental validation on 37 hexapeptides and 13 mutational variants of a model haloalkane dehalogenase LinB (HLD LinB). Our experiments showed AggreProt accuracy in predicting APRs: 30 out of 37 hexapeptides were correctly predicted with 2 further cases being inconclusive. Moreover, AggreProt correctly identified APRs in HLD LinB and was used to guide the design of mutations that decreased aggregation propensity and increased yield of soluble protein (the best mutations by 100%). The ability of AggreProt and other predictors to detect APRs in LinB is shown in Figure 1. We also purposely introduced mutation into the buried APR which led to significantly compromised structural integrity of the protein, to further highlight the importance of SASA information during design (35). Finally, we employed AggreProt to detect APRs in an insoluble luciferase from *Amphiura filiformis*. Subsequent engineering of one of the detected regions leads to solubilization of the protein (in preparation). These extensive experimental validations together with the herein presented predictions on three different proteins demonstrate the usefulness of AggreProt in identifying and engineering APRs in proteins.

Web server usage

AggreProt workflow

AggreProt web server has been designed to provide the users with an easy experience when identifying APRs in their input sequence. The server combines its dedicated amyloid aggregation propensity predictor described above with transmembrane (TM) propensity and solvent accessible surface area (SASA) calculations that provide structural context for the analysed protein sequence. The overall workflow of the operations in the server is shown in Figure 2.

Data input

First, the user inputs their protein sequence(s) in FASTA format, which are quickly scanned for integrity, i.e. that headers and sequence are in place. Currently, the webserver allows to upload up to three different sequences as input for

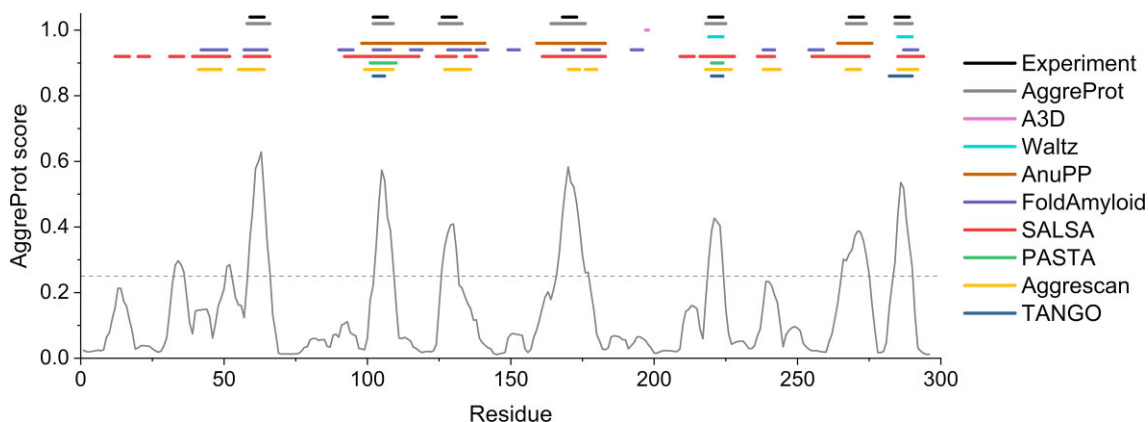


Figure 1. Experimental validation of AggreProt and comparison to other methods. APRs in LinB as previously determined in (35) (Experiment) or determined by predictors. The APR annotations were made using the default settings of each predictor according to the authors' recommendations. Even though the two peaks in AggreProt prediction around residues 35 and 50 reached the cut-off, their length was too short (less than six residues) to be considered as APRs.

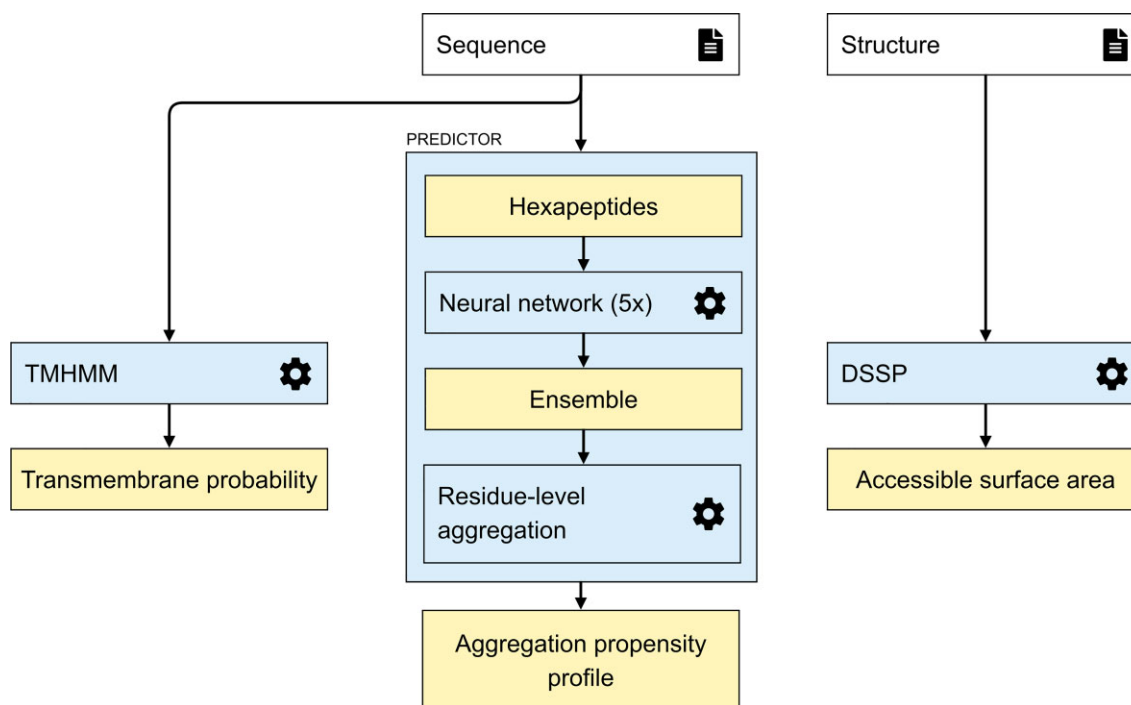


Figure 2. AggreProt schematic workflow. The different calculations on provided input types (sequence and structure) are depicted in sequential order. The input sequence is used to generate the TM and aggregation propensity profiles. The input structure is required to generate SASA profiles. Inputs (compulsory and optional) are indicated by a 'file' icon. Processes (third party software and internal processing) are indicated by a cog wheel and blue boxes. Outputs (partial and final) are indicated in yellow.

posterior comparison. Then, depending on the number of input sequences, an input box to upload a structure file associated to each of them appears dynamically. In this step the structure can be uploaded from a custom file (PDB and mmCIF formats are accepted) or can be downloaded from RSCB PDB. In case the user cannot provide an input structure, AggreProt offers the option to fetch it from AlphaFoldDB (42) either using AlaFoldDB ID, searching directly on the external database, or performing a BLAST search (43) on it (this latter option being significantly slow). Optional job title and e-mail address can be given for easy identification of the job submitted and posterior notification of results completion, respectively. However, these

two pieces of information are not required by the server, which can be used freely by all users without any login requirement.

Results output

After the job submission, the user is redirected to a page displaying a summary of the submitted job where its progression can be tracked. The job summary indicates the job ID and name (if provided), the submission time stamp, and the status of the job. Upon calculation completion, the job status becomes 'DONE', and the calculated results are made available. On the top, a profile visualization window displays the ob-

tained results in two different graphical forms: (i) as aligned profiles (central part, propensity chart) and (ii) as a glyph annotation (bottom, sequence display) (Figure 3, top panel). The profiles and the glyphs are colour coded so that the results for individual input proteins (if more than one was submitted) can be easily compared.

The aggregation propensity profile is shown in a semi-transparent solid hue, while TM propensity and SASA are indicated by dotted and dashed lines, respectively. Within this section of the visualizer (propensity chart), hovering over any sequence position, renders additional information about the protein residue and the individual prediction values for each of the propensities calculated. The corresponding glyphs are shown for each propensity series (amyloid aggregation, TM and SASA) according to their own thresholds. These glyphs help detecting continuous stretches of protein sequence that are predicted to be in a membrane or exposed. At the top of the visualization pane, the aggregation propensity threshold can be adjusted (by default it displays the threshold associated with the Youden J statistic for the predictor), and the display of the different input proteins can be adjusted independently. Increasing the threshold results in decreasing sensitivity (number of peaks detected) and increasing specificity (number of true peaks detected); decreasing the threshold achieves the opposite. The thresholds for TM (0.4) and SASA (0.25) are predetermined based on previous experience (52,53). Below, an adjustable slide ranger tool allows to zoom in on a particular section in the protein sequence(s) or to zoom out to a wider protein sequence range.

Linked to the profile visualization pane the server provides a three-dimensional (3D) representation of the input proteins (Figure 3, bottom panel) if provided by the user or otherwise fetched from the databases. The 3D view pane is interactively connected to the profile visualizer, such as when clicking an element (i.e. one amino-acid position) in one of the panes, the corresponding selection is also highlighted. In this way, from the upper visualizer the amino-acids conforming an APR (represented by a peak or by a continuous stretch of glyphs in the Aggregation series) can be selected and then highlighted for inspection in the 3D panel. The interactive behaviour also works vice-versa: selecting a region of interest from the structure will result in that region highlighted in the sequence profile visualizer, where the user can easily check what is the aggregation propensity of the selected region. At the very bottom of the results page, an executive help section indicating the user how to navigate the profile and the 3D visualizations.

Use cases

In our previous study we showed that AggreProt can correctly identify APRs that form amyloid fibrils as isolated peptides and promote aggregation within the context of the model protein HLD LinB (35). Furthermore, we show that mutations in these APRs decrease aggregation propensity and simultaneously increase solubility, providing a higher yield of the model protein. Here, we use the observed link between aggregation and solubility and extend it further to three model cases. Specifically, we validate the functionality of the presented webserver by the analysis of the deep scanning mutational data from two proteins: (i) Type III polyketide synthase and (ii) TEM β -lactamase (52,53) from SoluProtMut^{DB} (38). SoluProtMut^{DB} is a manually curated database containing effects of mutations on protein solubility. Singularly it contains

data on deep mutational scanning experiments, which can be exploited to perform a retrospective experiment: do solubilizing mutations correspond to the regions predicted by AggreProt as APRs? Are the effects of such mutations correctly estimated by AggreProt? Here, we illustrate that despite being trained on amyloid forming hexapeptides, our predictor can also detect APRs that in the correct context and with appropriately designed mutations may increase solubility. The aggregation propensities of significantly solubilized protein variants were compared to the respective wild types using AggreProt webserver. We use this reverse engineering approach to test how AggreProt can successfully guide the protein engineering strategies to reduce the aggregation and increase solubility. By exploring two of the proteins with deep mutational data, we showcase AggreProt capabilities and demonstrate the potential pitfalls and complexity of such protein solubilizing campaigns.

The significantly solubilizing mutations marked as manually curated from SoluProtMut^{DB} were selected for all three proteins. Different thresholds of the solubility change were used to yield a similar number of mutants for all proteins (≥ 2.0 for Type III polyketide synthase and ≥ 7.0 for TEM β -lactamase and Levoglucosan kinase). Mutations on Type III polyketide synthase showed a strikingly high correspondence with the detected peaks by AggreProt. The effect of such mutations as predicted by AggreProt agreed with the recorded experimental effect in the vast majority of cases (Figure 4, Supplementary Table S1).

For TEM β -lactamase, the solubilizing mutations were colocalized in regions with high aggregation propensity that were both exposed and buried. The solubilizing mutations found in the exposed APRs followed the same logic that we followed in our previous work (35): replacing hydrophobic residues with more polar or charged ones. On the contrary, mutations found in buried regions consisted in the increase of local hydrophobicity, and our predictor failed to recognize them as solubilizing (Figure 4, Supplementary Table S2). This observation agrees with the practice of leaving hydrophobic buried regions out of engineering campaigns (55) and designing only solvent-accessible amino acids.

Type III polyketide synthase (Uniprot ID: A0A3G4RHW3)

In total, 31 mutations (56) were selected based on the filters described above. Interestingly, 28 of those (90%) were in the regions detected as APRs by AggreProt, (Figure 4, Supplementary Table S1). The effect of these mutations was subsequently evaluated as the changes in the AggreProt profile (Figure 4, Supplementary Table S1). The peaks in APR regions were reduced in 21 cases (75%) suggesting solubility improvement, in five cases (18%) the mutations did not affect the size of the peaks in APR regions suggesting minor effects, and only in two cases (7%) the corresponding peaks were in APR regions increased indicating changes towards lower solubility. Overall, these results confirm that AggreProt can identify aggregation-prone regions with high confidence and that the effect of a large majority of substitutions introduced into these regions are correctly predicted by AggreProt.

TEM-1.1 (β -lactamase, Uniprot ID: P62593)

In total, 59 out of the 80 mutations (74%) selected based on our criteria were in the predicted APRs. The subsequent

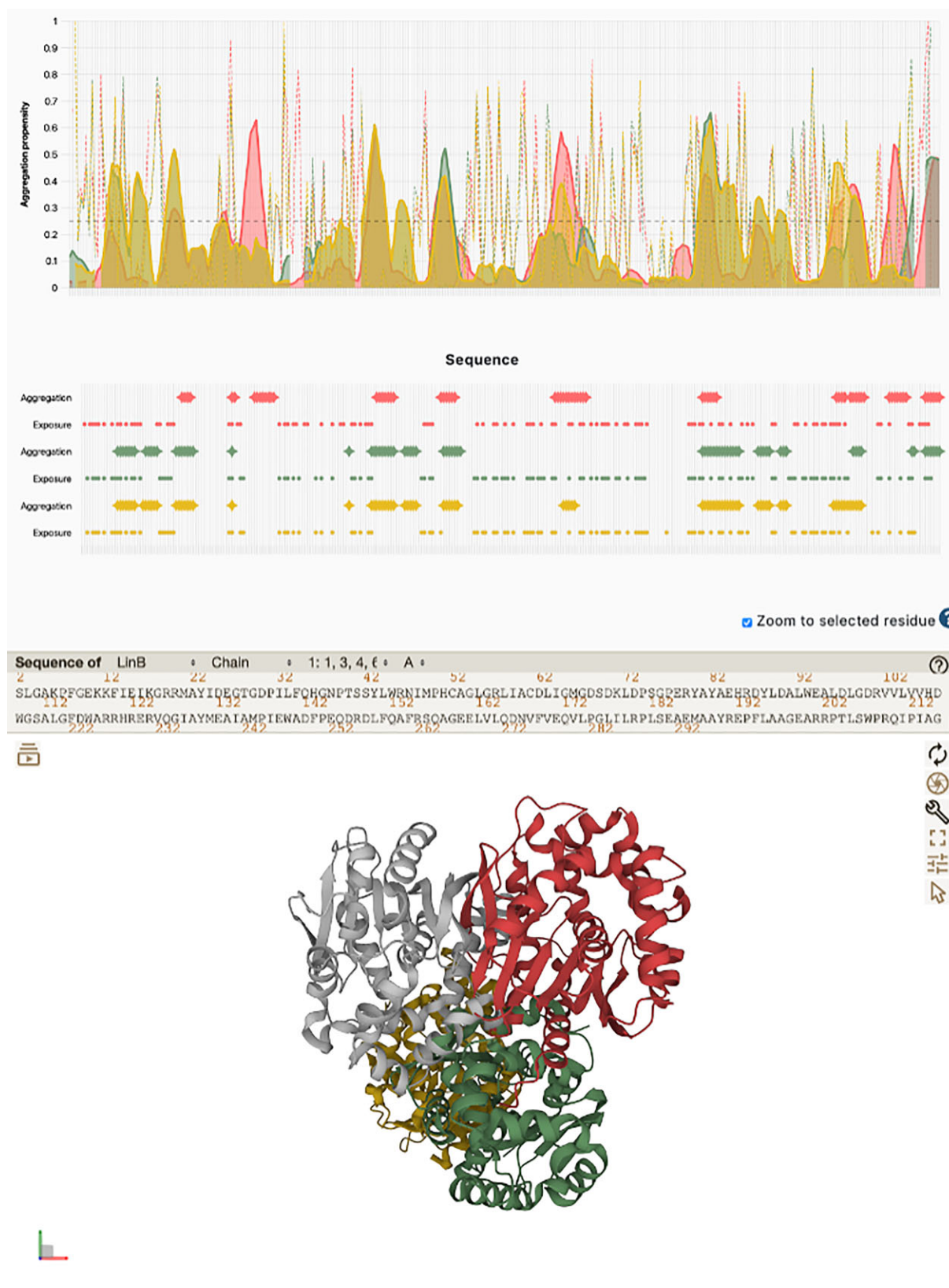


Figure 3. AggreProt results page. Propensity and 3D panes are displayed on top of each other and interactively linked. The propensity pane allows setting the cut-off threshold value for aggregation propensity and to display the results for any of the input proteins at will. The different propensity series are colour-coded identifying the input protein and differently dashed to identify each of the series (aggregation, TM, and SASA). Below the propensity chart, the protein sequences and the predictions in a glyph-encoded view are provided. At the bottom, the 3D view pane which is bidirectionally interactive with the rest of the elements of the results page. The figure displays three haloalkane dehalogenases: LinB (red), DhaA wild type (green) and DhaA115 (hyper-stable mutant, yellow) (54).

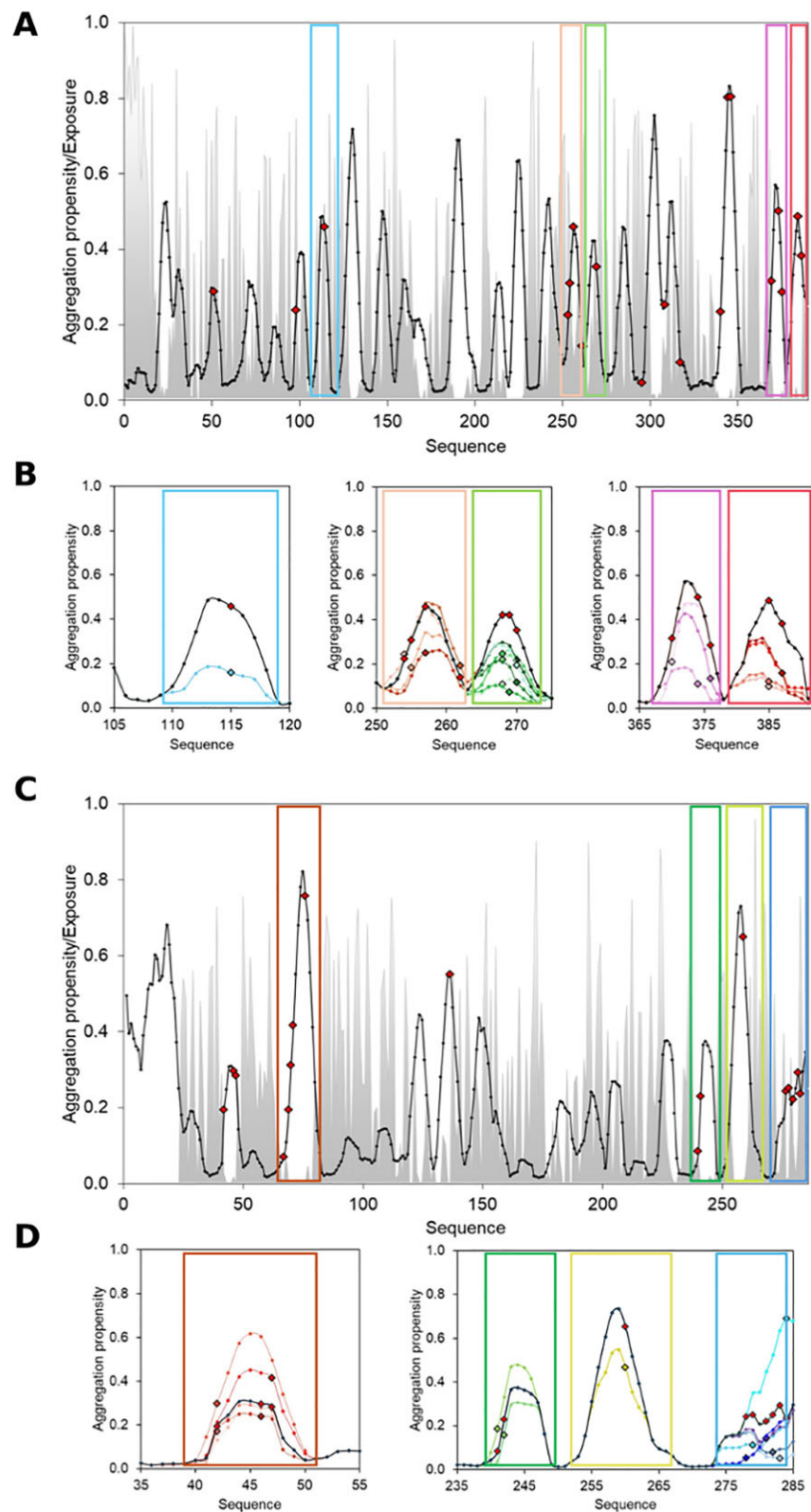


Figure 4. The overview of the predicted mutational effect of selected Type III polyketide synthase (**A, B**) and TEM-1.1 β -lactamase (**C, D**) variants. (**A, C**) The wild-type aggregation profile (solid line) is shown in the context of solubilizing mutations occurring in predicted APRs (red diamonds), other solubilizing mutations outside predicted peaks (white diamonds), predicted SASA (dotted line, greyed out area), and aggregation propensity threshold (dashed horizontal line). (**B, D**) The predicted effect of the mutation is shown for some selected exposed APRs (different colour hues); mutated residues are shown as coloured diamonds, red for the wt-profile, and matching the hue of the mutation for the mutational profiles.

analysis (Figure 4, Supplementary Table S3) showed that AggreProt predicted a reduction of the aggregation peak in 34 cases (58%), thus lowering predicted aggregation propensity, while in 25 cases (42%) the effect was neutral or negative, indicating higher predicted aggregation propensity. However, in the case of β -lactamase the solubilizing mutations corresponded to buried APRs, and the solubilizing effect estimated by AggreProt depended on the type of mutation. Mutations increasing local hydrophobicity are predicted by AggreProt to increase the aggregation propensity. While this is often the case in exposed APRs, it is not a conserved trend in buried ones. In a protein engineering campaign, the consideration of such contextual information is therefore crucial not to disrupt the hydrophobic core of the protein (55). The comprehensive presentation of aggregation-promoting residues together with their location within protein structure provides essential information for rational protein engineering using AggreProt.

Nb.b201 nanobody (PDB ID: 5vnm, chain C)

An additional use case showing the performance of AggreProt on Nb.b201 nanobody and experimentally validated solubilising mutations (57) is presented in Supplementary materials (see Supplementary Note 2 and Supplementary Figure S3).

Discussion and outlook

AggreProt web server presents the community with an easy-to-use application for predicting and engineering amyloid aggregation prone regions in proteins based on their sequence. Its performance on standard validation datasets is in par or better than the state-of-the-art sequence-based methods: Waltz (18), Tango (19), ANuPP (28), AGGRESCAN (16), PASTA (20), SALSA (21), CamSol intrinsic (17), and FoldAmyloid (22). Also, for reference, the comparison was extended to one widely-used structure-based method, Aggrescan3D (23). In comparison to those, AggreProt does not present any limitation in the input sequence length, providing a residue-level prediction for the whole protein sequence. Our tool also facilitates the interpretation of results allowing the user to compare up to three different input proteins in one single run. Moreover, it provides useful contextual information in the form of transmembrane propensity and solvent accessible surface area, whenever possible. Our previous experience indicate that this feature is crucial for successful design of aggregation-reducing mutations into proteins. AggreProt predictor has been experimentally validated: (i) at hexapeptide level, with correct predictions in 30 out of 37 cases, and (ii) in two protein solubilizing campaigns. On the first campaign, AggreProt detected 4 solvent-accessible APRs on haloalkane dehalogenase LinB; each of the APRs was re-designed resulting in a more soluble protein (35). The second, on a luciferase from *Amphiura filiformis* also detected the correct region that, upon design, resulted in increased solubility (in preparation).

We illustrate the application of our webserver in the retrospective experiment based on deep mutational scanning data on two different proteins: type III polyketide synthase and TEM β -lactamase (58). We extracted the most solubilizing mutations on each of these proteins from SoluProtMut^{DB} (38), and mapped them to the aggregation profile obtained by AggreProt. In the first example, we found a perfect overlap between the solubilizing mutations and the predicted APRs. Moreover the solubilizing effect of these mutations was reca-

pitulated by the AggreProt predictions. In the case of TEM β -lactamase, most of the solubilizing mutations corresponded to predicted APRs, but happened on buried regions of the protein. Here, the capacity of AggreProt to correctly predict the mutation behaviour depended on the nature of the mutation: we had correct (solubilizing) estimations for hydrophobic-to-hydrophilic mutations and wrong predictions otherwise. This illustrates the power of the contextual information since understanding the solvent accessibility of the amino acid to be mutated is crucial to correctly predict the effect of the mutation. We further illustrate the performance of AggreProt in an additional biomedically and biotechnologically relevant test example of a nanobody. There our predictor was able to correctly identify the mutational effect of 6 out of 8 solubilized variants.

AggreProt as presented is a powerful resource to identify APRs in proteins, based only on their sequence input. However, the final goal of reducing the protein aggregation propensity can only be achieved by engineering its sequence. We will further develop AggreProt and provide the user with a 'Design' panel, where the engineering campaign could be outlined, and its outcomes predicted. We will facilitate this task in future versions of our tool, by implementing the following strategies: First, enabling the user to fine-tune the boundaries of detected APRs and define custom ones. Correctly defining the APRs boundaries is crucial for implementing one of the trending strategies in solubilizing amyloid prone proteins developed by the Switch Lab: mutating the gatekeeper residues lying on the APR boundaries (59). Second, implementing two different mutational strategies: (i) targeting the aforementioned gatekeepers and subsisting residue each pair by any of the 5 residues arginine, lysine, aspartic acid, glutamic acid and proline (for a total of 25 variants per APR), and (ii) targeting the exposed residue closer to the peak of the APR and saturating its position (for a total of 19 variants per APR). In our vision, the user will have the ability to choose and mix-and-match among the strategies and redefine the residues to be mutated. To present the user with a final solution, first all selected mutations will be evaluated by AggreProt and their new profiles super-imposed to the original one. Then, after the user selects one mutation per APR, an ideal multi-mutation sequence will be compiled, evaluated by the predictor, and finally presented to the user for inspection.

Data availability

Aggreprot is freely available and accessible at <https://loschmidt.chemi.muni.cz/aggreprot/>.

Supplementary data

Supplementary Data are available at NAR Online.

Acknowledgements

This work was supported by the Ministry of Education, Youth and Sports of the Czech Republic through the e-INFRA CZ (ID:90254; LM2018140) and ELIXIR-CZ (ID:90255; LM2023055). Authors thank the RECEPTOX Research Infrastructure (No. LM2023069), Operational Programme Research, Development and Education (the CETOCOEN EXCELLENCE project No. CZ.02.1.01/0.0/0.0/17_043/0009632), and Technology

Agency of the Czech Republic (FW03010208). This research was also supported by the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no. 857560 and by the project National Institute for Neurology Research (no. LX22NPO5107 MEYS): Financed by European Union – Next Generation EU.

Funding

Horizon 2020 Framework Programme [857560]; Technology Agency of the Czech Republic [FW03010208]; Ministry of Education, Youth and Sports of the Czech Republic [CETO-COEN EXCELLENCE CZ.02.1.01/0.0/0.0/17_043/0009, ELIXIR-CZ ID:90255, LM2023055; RECETOX RI LM2023069; e-INFRA CZ ID:90254, LM2018140]; European Union – Next Generation EU [National Institute for Neurology Research LX22NPO5]. Funding for open access charge: European Union's Horizon 2020 Research and Innovation Programme [857560].

Conflict of interest statement

None declared.

References

- Wodak,S.J., Vajda,S., Lensink,M.F., Kozakov,D. and Bates,P.A. (2023) Critical assessment of methods for predicting the 3D structure of proteins and protein complexes. *Annu. Rev. Biophys.*, **52**, 183–206.
- Elofsson,A. (2023) Progress at protein structure prediction, as seen in CASP15. *Curr. Opin. Struct. Biol.*, **80**, 102594.
- Jumper,J., Evans,R., Pritzel,A., Green,T., Figurnov,M., Ronneberger,O., Tunyasuvunakool,K., Bates,R., Židek,A., Potapenko,A., *et al.* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*, **596**, 583–589.
- Baek,M., DiMaio,F., Anishchenko,I., Dauparas,J., Ovchinnikov,S., Lee,G.R., Wang,J., Cong,Q., Kinch,L.N., Schaeffer,R.D., *et al.* (2021) Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, **373**, 871–876.
- Lin,Z., Akin,H., Rao,R., Hie,B., Zhu,Z., Lu,W., Smetanin,N., Verkuil,R., Kabeli,O., Shmueli,Y., *et al.* (2023) Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, **379**, 1123–1130.
- Pinheiro,F., Santos,J. and Ventura,S. (2021) AlphaFold and the amyloid landscape. *J. Mol. Biol.*, **433**, 167059.
- Chakravarty,D. and Porter,L.L. (2022) AlphaFold2 fails to predict protein fold switching. *Protein Sci.*, **31**, e4353.
- Louros,N., Schymkowitz,J. and Rousseau,F. (2023) Mechanisms and pathology of protein misfolding and aggregation. *Nat. Rev. Mol. Cell Biol.*, **24**, 912–933.
- Soto,C. and Pritzkow,S. (2018) Protein misfolding, aggregation, and conformational strains in neurodegenerative diseases. *Nat. Neurosci.*, **21**, 1332–1340.
- Sawaya,M.R., Sambashivan,S., Nelson,R., Ivanova,M.I., Sievers,S.A., Apostol,M.I., Thompson,M.J., Balbirnie,M., Wiltzius,J.J.W., McFarlane,H.T., *et al.* (2007) Atomic structures of amyloid cross- β spines reveal varied steric zippers. *Nature*, **447**, 453–457.
- Fändrich,M., Nyström,S., Nilsson,K.P.R., Böckmann,A., LeVine,H. and Hammarström,P. (2018) Amyloid fibril polymorphism: a challenge for molecular imaging and therapy. *J. Intern. Med.*, **283**, 218–237.
- Lövestam,S., Li,D., Wagstaff,J.L., Kotecha,A., Kimanius,D., McLaughlin,S.H., Murzin,A.G., Freund,S.M.V., Goedert,M. and Scheres,S.H.W. (2024) Disease-specific tau filaments assemble via polymorphic intermediates. *Nature*, **625**, 119–125.
- Wang,H., Duo,L., Hsu,F., Xue,C., Lee,Y.K. and Guo,Z. (2020) Polymorphic A β 42 fibrils adopt similar secondary structure but differ in cross-strand side chain stacking interactions within the same β -sheet. *Sci. Rep.*, **10**, 5720.
- Sawaya,M.R., Hughes,M.P., Rodriguez,J.A., Riek,R. and Eisenberg,D.S. (2021) The expanding amyloid family: structure, stability, function, and pathogenesis. *Cell*, **184**, 4857–4873.
- van der Kant,R., Louros,N., Schymkowitz,J. and Rousseau,F. (2022) Thermodynamic analysis of amyloid fibril structures reveals a common framework for stability in amyloid polymorphs. *Structure*, **30**, 1178–1189.
- Conchillo-Solé,O., de Groot,N.S., Avilés,F.X., Vendrell,J., Daura,X. and Ventura,S. (2007) AGGRESKAN: a server for the prediction and evaluation of 'hot spots' of aggregation in polypeptides. *BMC Bioinf.*, **8**, 65.
- Sormanni,P., Aprile,F.A. and Vendruscolo,M. (2015) The CamSol method of rational design of protein mutants with enhanced solubility. *J. Mol. Biol.*, **427**, 478–490.
- Maurer-Stroh,S., Debulpaep,M., Kuemmerer,N., Lopez de la Paz,M., Martins,I.C., Reumers,J., Morris,K.L., Copland,A., Serpell,L., Serrano,L., *et al.* (2010) Exploring the sequence determinants of amyloid structure using position-specific scoring matrices. *Nat. Methods*, **7**, 237–242.
- Fernandez-Escamilla,A.-M., Rousseau,F., Schymkowitz,J. and Serrano,L. (2004) Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat. Biotechnol.*, **22**, 1302–1306.
- Walsh,I., Seno,F., Tosatto,S.C.E. and Trovato,A. (2014) PASTA 2.0: an improved server for protein aggregation prediction. *Nucleic Acids Res.*, **42**, W301–W307.
- Zibaee,S., Makin,O.S., Goedert,M. and Serpell,L.C. (2007) A simple algorithm locates β -strands in the amyloid fibril core of α -synuclein, A β , and tau using the amino acid sequence alone. *Protein Sci.*, **16**, 906–918.
- Garbuzynskiy,S.O., Lobanov,M.Y. and Galzitskaya,O.V. (2010) FoldAmyloid: a method of prediction of amyloidogenic regions from protein sequence. *Bioinformatics*, **26**, 326–332.
- Kuriata,A., Iglesias,V., Pujols,J., Kurcinski,M., Kmiecik,S. and Ventura,S. (2019) Aggrescan3D (A3D) 2.0: prediction and engineering of protein solubility. *Nucleic Acids Res.*, **47**, W300–W307.
- Keresztes,L., Szögi,E., Varga,B., Farkas,V., Perczel,A. and Grolmusz,V. (2021) The budapest amyloid predictor and its applications. *Biomolecules*, **11**, 500.
- Niu,M., Li,Y., Wang,C. and Han,K. (2018) RFamyloid: a web server for predicting amyloid proteins. *Int. J. Mol. Sci.*, **19**, 2071.
- Burdukiewicz,M., Sobczyk,P., Rödiger,S., Duda-Madej,A., Mackiewicz,P. and Kotulska,M. (2017) Amyloidogenic motifs revealed by n-gram analysis. *Sci. Rep.*, **7**, 12961.
- Navarro,S. and Ventura,S. (2022) Computational methods to predict protein aggregation. *Curr. Opin. Struct. Biol.*, **73**, 102343.
- Prabakaran,R., Rawat,P., Kumar,S. and Michael Gromiha,M. (2021) ANuPP: a versatile tool to predict aggregation nucleating regions in peptides and proteins. *J. Mol. Biol.*, **433**, 166707.
- Gasior,P. and Kotulska,M. (2014) FISH Amyloid – a new method for finding amyloidogenic segments in proteins based on site specific co-occurrence of aminoacids. *BMC Bioinf.*, **15**, 54.
- Louros,N., Orlando,G., De Vleeschouwer,M., Rousseau,F. and Schymkowitz,J. (2020) Structure-based machine-guided mapping of amyloid sequence space reveals uncharted sequence clusters with higher solubilities. *Nat. Commun.*, **11**, 3314.
- Louros,N., Konstantouleas,K., De Vleeschouwer,M., Ramakers,M., Schymkowitz,J. and Rousseau,F. (2020) WALTZ-DB 2.0: an updated database containing structural information of experimentally determined amyloid-forming peptides. *Nucleic Acids Res.*, **48**, D389–D393.
- Van Durme,J., Delgado,J., Stricher,F., Serrano,L., Schymkowitz,J. and Rousseau,F. (2011) A graphical interface for the FoldX forcefield. *Bioinformatics*, **27**, 1711–1712.

33. Varadi,M., De Baets,G., Vranken,W.F., Tompa,P. and Pancsa,R. (2018) AmyPro: a database of proteins with validated amyloidogenic regions. *Nucleic Acids Res.*, **46**, D387–D392.
34. Rawat,P., Prabakaran,R., Sakthivel,R., Mary Thangakani,A., Kumar,S. and Gromiha,M.M. (2020) CPAD 2.0: a repository of curated experimental data on aggregating proteins and peptides. *Amyloid*, **27**, 128–133.
35. Cima,V., Kunka,A., Grakova,E., Planas-Iglesias,J., Havlasek,M., Subramanian,M., Beloch,M., Marek,M., Slaninova,K., Damborsky,J., *et al.* (2024) Prediction of aggregation prone regions in proteins using deep neural networks and their suppression by computational design. bioRxiv doi: <https://doi.org/10.1101/2024.03.06.583680>, 11 March 2024, preprint: not peer reviewed.
36. Marcelino,A.M.C. and Gierasch,L.M. (2008) Roles of β -turns in protein folding: from peptide models to protein engineering. *Biopolymers*, **89**, 380–391.
37. Barth,P. and Senes,A. (2016) Toward high-resolution computational design of the structure and function of helical membrane proteins. *Nat. Struct. Mol. Biol.*, **23**, 475–480.
38. Velecký,J., Hamsikova,M., Stourac,J., Musil,M., Damborsky,J., Bednar,D. and Mazurenko,S. (2022) SoluProtMutDB: a manually curated database of protein solubility changes upon mutations. *Comput. Struct. Biotechnol. J.*, **20**, 6339–6347.
39. Ruopp,M.D., Perkins,N.J., Whitcomb,B.W. and Schisterman,E.F. (2008) Youden index and optimal cut-point estimated from observations affected by a lower limit of detection. *Biometrical Journal*, **50**, 419–430.
40. Abadi,M., Agarwal,A., Barham,P., Brevdo,E., Chen,Z., Citro,C., Corrado,G.S., Davis,A., Dean,J., Devin,M., *et al.* (2015) TensorFlow: large-Scale machine learning on heterogeneous systems. Zenodo. <https://doi.org/10.5281/zenodo.4724125>.
41. Zemla,A., Venclovas,Č., Fidelis,K. and Rost,B. (1999) A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment. *Proteins Struct. Funct. Genet.*, **34**, 220–223.
42. Varadi,M., Anyango,S., Deshpande,M., Nair,S., Natassia,C., Yordanova,G., Yuan,D., Stroe,O., Wood,G., Laydon,A., *et al.* (2022) AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.*, **50**, D439–D444.
43. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
44. Krogh,A., Larsson,B., von Heijne,G. and Sonnhammer,E.L.L. (2001) Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *J. Mol. Biol.*, **305**, 567–580.
45. Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
46. Gohl,P., Bonet,J., Fornes,O., Planas-Iglesias,J., Fernandez-Fuentes,N. and Oliva,B. (2023) SBILib: a handle for protein modeling and engineering. *Bioinformatics*, **39**, btad613.
47. Lafita,A., Bliven,S., Prlić,A., Guzenko,D., Rose,P.W., Bradley,A., Pavan,P., Myers-Turnbull,D., Valasatava,Y., Heuer,M., *et al.* (2019) BioJava 5: a community driven open-source bioinformatics library. *PLoS Comput. Biol.*, **15**, e1006791.
48. Sehnal,D., Bittrich,S., Deshpande,M., Svobodová,R., Berka,K., Bazgier,V., Velankar,S., Burley,S.K., Koča,J. and Rose,A.S. (2021) Mol* Viewer: modern web app for 3D visualization and analysis of large biomolecular structures. *Nucleic Acids Res.*, **49**, W431–W437.
49. O'Rourke,T.W., Loya,T.J., Head,P.E., Horton,J.R. and Reines,D. (2015) Amyloid-like assembly of the low complexity domain of yeast Nab3. *Prion*, **9**, 34–47.
50. Wittmer,Y., Jami,K.M., Stowell,R.K., Le,T., Hung,I. and Murray,D.T. (2023) Liquid droplet aging and seeded fibril formation of the cytotoxic granule associated RNA binding protein TIA1 low complexity domain. *J. Am. Chem. Soc.*, **145**, 1580–1592.
51. Si,K., Lindquist,S. and Kandel,E.R. (2003) A neuronal isoform of the aplysia CPEB has prion-like properties. *Cell*, **115**, 879–891.
52. Cserzo,M., Eisenhaber,F., Eisenhaber,B. and Simon,I. (2004) TM or not TM: transmembrane protein prediction with low false positive rate using DAS-TMfilter. *Bioinformatics*, **20**, 136–137.
53. Schmidt,C., Macpherson,J.A., Lau,A.M., Tan,K.W., Fraternali,F. and Politis,A. (2017) Surface accessibility and dynamics of macromolecular assemblies probed by covalent labeling mass spectrometry and integrative modeling. *Anal. Chem.*, **89**, 1459–1468.
54. Markova,K., Chmelova,K., Marques,S.M., Carpentier,P., Bednar,D., Damborsky,J. and Marek,M. (2020) Decoding the intricate network of molecular interactions of a hyperstable engineered biocatalyst. *Chem. Sci.*, **11**, 11162–11178.
55. Buck,P.M., Kumar,S. and Singh,S.K. (2013) On the role of aggregation prone regions in protein evolution, stability, and enzymatic catalysis: insights from diverse analyses. *PLoS Comput. Biol.*, **9**, e1003291.
56. Wrenbeck,E.E., Bedewitz,M.A., Klesmith,J.R., Noshin,S., Barry,C.S. and Whitehead,T.A. (2019) An automated data-driven pipeline for improving heterologous enzyme expression. *ACS Synth. Biol.*, **8**, 474–481.
57. Rosace,A., Bennett,A., Oeller,M., Mortensen,M.M., Sakhnini,L., Lorenzen,N., Poulsen,C. and Sormanni,P. (2023) Automated optimisation of solubility and conformational stability of antibodies and proteins. *Nat. Commun.*, **14**, 1937.
58. Klesmith,J.R., Bacik,J.-P., Wrenbeck,E.E., Michalczyk,R. and Whitehead,T.A. (2017) Trade-offs between enzyme fitness and solubility illuminated by deep mutational scanning. *Proc. Natl. Acad. Sci. U.S.A.*, **114**, 2265–2270.
59. Houben,B., Rousseau,F. and Schymkowitz,J. (2022) Protein structure and aggregation: a marriage of necessity ruled by aggregation gatekeepers. *Trends Biochem. Sci.*, **47**, 194–205.