

Aggrescan4D: A comprehensive tool for pH-dependent analysis and engineering of protein aggregation propensity

Mateusz Zalewski¹  | Valentin Iglesias^{2,3} | Oriol Bárcenas^{2,4} |
Salvador Ventura^{2,5}  | Sebastian Kmiecik¹ 

¹Faculty of Chemistry, Biological and Chemical Research Center, University of Warsaw, Warsaw, Poland

²Departament de Bioquímica i Biologia Molecular, Institut de Biotecnologia i de Biomedicina, Universitat Autònoma de Barcelona, Barcelona, Spain

³Clinical Research Centre, Medical University of Białystok, Białystok, Poland

⁴Institute of Advanced Chemistry of Catalonia (IQAC), CSIC, Barcelona, Spain

⁵Hospital Universitari Parc Taulí, Institut d'Investigació i Innovació Parc Taulí (I3PT-CERCA), Universitat Autònoma de Barcelona, Sabadell, Spain

Correspondence

Sebastian Kmiecik, Biological and Chemical Research Center, Faculty of Chemistry, University of Warsaw, Pasteura 1, 02-093 Warsaw, Poland.
Email: sekmi@chem.uw.edu.pl

Salvador Ventura, Institut de Biotecnologia i de Biomedicina and Departament de Bioquímica i Biologia Molecular, Universitat Autònoma de Barcelona, Barcelona, Spain.
Email: salvador.ventura@uab.cat

Funding information

Polish National Agency for Academic Exchange under the ULAM NAWA Programme, Grant/Award Number: BPN/ULM/2023/1/00189/U/00001; Spanish Ministry of Science and Innovation, Grant/Award Number: PID2022-137963OB-I00; Generalitat de Catalunya, Grant/Award Number: 2021-SGR-00635 AGA; ICREA, ICREA-Academia 2020; National Science Center, Grant/Award Number: Sheng2021/40/Q/NZ2/00078

Review Editor: Nir Ben-Tal.

Abstract

Aggrescan4D (A4D) is an advanced computational tool designed for predicting protein aggregation, leveraging structural information and the influence of pH. Building upon its predecessor, Aggrescan3D (A3D), A4D has undergone numerous enhancements aimed at assisting the improvement of protein solubility. This manuscript reviews A4D's updated functionalities and explains the fundamental principles behind its pH-dependent calculations. Additionally, it presents an antibody case study to evaluate its performance in comparison with other structure-based predictors. Notably, A4D integrates advanced protein engineering protocols with pH-dependent calculations, enhancing its utility in advising solubility-enhancing mutations. A4D considers the impact of structural flexibility on aggregation propensities, and includes a large set of precalculated predictions. These capabilities should help to open new avenues for both understanding and managing protein aggregation. A4D is accessible through a dedicated web server at <https://biocomp.chem.uw.edu.pl/a4d/>.

KEYWORDS

computational biology, monoclonal antibodies, pH-dependent aggregation, protein aggregation, protein engineering, protein solubility, protein stability, structural bioinformatics

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2024 The Author(s). *Protein Science* published by Wiley Periodicals LLC on behalf of The Protein Society.

1 | INTRODUCTION

Protein aggregation is a significant biochemical process with implications for both health and technology. It occurs when soluble proteins form insoluble aggregates, often leading to diseases such as Parkinson's, Alzheimer's, ALS, type II diabetes, and various cancers (Chiti & Dobson, 2017). From an evolutionary perspective, protein aggregation can limit the diversity of protein sequences that can be stably expressed in cells. This is because the aggregation process affects several crucial aspects, including the proper folding of proteins into their native states and the concentration at which they can exist without forming aggregates (Carija et al., 2019; de Groot & Ventura, 2010; Santos, Iglesias, & Ventura, 2020). In industry, although protein aggregation poses challenges in biotechnology, particularly in the production and purification of biotherapeutics, it also presents unique opportunities (Shukla & Thömmes, 2010). Amyloid fibers, utilized as scaffolds (Díaz-Caballero et al., 2018; Knowles & Mezzenga, 2016) and enzymatic platforms (Díaz-Caballero et al., 2018; Knowles & Mezzenga, 2016), exemplify how biological challenges can be transformed into technological advantages. Understanding these aggregation processes is essential for both mitigating their harmful effects and harnessing their potential benefits.

The growing need to understand protein aggregation across various contexts has spurred the development of sequence-based bioinformatics tools designed to pinpoint regions in proteins most susceptible to aggregation (Santos, Pujols, Pallarès, et al., 2020). Tools such as TANGO (Fernandez-Escamilla et al., 2004), Zygggregator (Tartaglia & Vendruscolo, 2008), and Aggrescan (Conchillo-Solé et al., 2007) are highly valued for their speed and cost-effectiveness, facilitating the rapid redesign of protein variants. TANGO (Fernandez-Escamilla et al., 2004) utilizes a statistical mechanics approach to predict protein aggregation by evaluating beta-sheet formation tendencies, focusing on peptide backbone energetics and side chain interactions. Zygggregator (Tartaglia & Vendruscolo, 2008) determines aggregation propensities using physicochemical properties and the context within protein sequences. Meanwhile, Aggrescan (Conchillo-Solé et al., 2007) identifies critical “hot spots” for aggregation using an aggregation propensity scale derived from *in vivo* data, making it particularly effective for detecting short sequence stretches that significantly influence protein aggregation. These tools are particularly useful for proteins in extended conformations or intrinsically disordered proteins (IDPs), where exposed aggregation prone regions (APRs) are directly related to the sequence (Belli et al., 2011). However,

sequence-based tools often misidentify APRs in folded proteins, where hydrophobic regions involved in interactions or buried within the protein are less likely to contribute to aggregation.

This limitation led to the development of structure-based predictors that utilize spatial information from experimental structures to identify aggregation-prone regions where residues are closely packed in three dimensions (Chennamsetty et al., 2009). Key advancements in this area include SolubiS (De Baets et al., 2015), CamSol (Sormanni et al., 2015), and Aggrescan3D (A3D) (Zambrano et al., 2015). These tools have improved the accuracy of predicting protein aggregation by effectively using structural data to pinpoint potential hotspots. SolubiS (De Baets et al., 2015) builds on TANGO's (Fernandez-Escamilla et al., 2004) methodology to improve protein solubility by mutating aggregation-prone regions to more stable residues, utilizing the FoldX (Schymkowitz et al., 2005) force field to assess the impact of these changes. CamSol (Sormanni et al., 2015), on the other hand, predicts protein solubility by taking into account factors such as hydrophobicity, charge, and structural context, and it provides users with protein variants engineered for increased solubility. A3D (Zambrano et al., 2015) builds on the original Aggrescan (Conchillo-Solé et al., 2007) by adding three-dimensional structural information to its sequence-based aggregation propensity analysis. This allows A3D to identify aggregation-prone regions based on both the sequence and the spatial proximity of residues, giving a more accurate depiction of potential aggregation hotspots in the three-dimensional structure of proteins.

Building on the strengths of A3D (Zambrano et al., 2015), our new in-house algorithm, referred to as Aggrescan4D (A4D) (Bárcenas et al., 2024), represents the next step in the evolution of aggregation prediction tools. A4D not only incorporates all the features of A3D but also introduces environmental pH as a variable in its calculations, offering a more nuanced view of protein aggregation under different physiological conditions. Additionally, A4D incorporates further enhancements for automated protein design and user-friendly operation.

In this overview, we explore the capabilities of A4D (Bárcenas et al., 2024), a state-of-the-art tool for predicting protein aggregation propensity, and compare its strengths and weaknesses with other aggregation prediction algorithms. Section 2 highlights the transformative advancements from A3D (Zambrano et al., 2015) to A4D (Bárcenas et al., 2024), emphasizing its key features, methodologies, and applications. Section 3 offers a concise overview of A4D's options, including its most important functionalities such as: pH calculations (Section 3.1), dynamic mode simulations (Section 3.4), and automatic

mutations (Section 3.7). Each section delves into A4D's capabilities, underscoring its significance in protein aggregation research and potential impact on biotechnology and biomedical science. Additionally, in Section 4, we describe an antibody case study evaluating the predictive accuracy of A4D, alongside other tools, providing valuable insights into its predictive accuracy. Finally, we also discuss potential future enhancements that could improve our ability to predict and manage protein aggregation.

2 | FROM Aggrescan3D TO Aggrescan4D: TRANSFORMATIVE ADVANCES IN PROTEIN AGGREGATION PREDICTION TECHNOLOGY

A3D (Zambrano et al., 2015) integrates the Aggrescan *in vivo*-derived scale (Conchillo-Solé et al., 2007) into the 3D space of protein structures by leveraging distances between amino acid residues and their overall surface exposure. This method overcomes the main limitations of sequence-based predictive approaches by disregarding solvent-protected residues that have negligible impact on protein aggregation, identifying spatially clustered residues that form STructural Aggregation Prone regions (STAPs) despite being sequentially distant, and evaluating the decay of residues' impact on STAP potency due to distance between alpha-carbon atoms (Conchillo-Solé et al., 2007; Sánchez de Groot et al., 2005). A3D's distinctive feature CABS-flex (Jamroz et al., 2014; Kurcinski et al., 2019; Kuriata et al., 2018) provides a dynamic model of protein flexibility which allows for evaluating the aggregation propensity of proteins by identifying potential STAPs that may be hidden in the static input structure (Nithin et al., 2024; Zambrano et al., 2015).

The first web server for A3D (Zambrano et al., 2015) was launched in 2015, marking a significant milestone in making this innovative tool accessible to the wider scientific community. Since its launch, A3D has been widely accepted in the scientific community, leading to numerous significant publications. Researchers have utilized A3D for the *in silico* redesign of antibodies and other proteins, enhancing their structural and functional properties for various biotechnological applications (Ejaz et al., 2023; Gil-Garcia et al., 2018; Xia et al., 2016). Additionally, the platform has been instrumental in understanding the aggregation mechanisms of both pathogenic and non-pathogenic proteins as well as biotherapeutics, providing insights into their behaviors and impacts (Berdyński et al., 2022; Bhandare & Ramaswamy, 2018; Gil-Garcia et al., 2020; Katina et al., 2017; Ko et al., 2023;

Leppert et al., 2023; Petrlova et al., 2022; Zakeri et al., 2023; Žerovnik, 2017). A3D has also supported studies on the interactions between antibacterial proteins and lipidic bilayers, as well as disaggregase-client recognition dynamics, aiding in the development of multi-epitope vaccines for tropical diseases and enhancing our understanding of protein evolution and its impact on aggregation (Badaczewska-Dawid et al., 2023; Carija et al., 2019; Flores-León et al., 2021; Garcia-Pardo et al., 2023; Ishwarlall et al., 2022; Pulido, Arranz-Trullén, et al., 2016; Pulido, Llamas, et al., 2016; Ruiz-Solani et al., 2023).

In 2019, the A3D 2.0 web server (Kuriata, Iglesias, Pujols, et al., 2019) was launched with major improvements, including the new CABS-flex package and the FoldX (Schymkowitz et al., 2005) force field. This version's enhanced capabilities enabled it to handle larger proteins and complexes, facilitating rational protein redesign by balancing aggregation propensity with thermodynamic stability (Castillo et al., 2010; Espargaró et al., 2008; Gil-Garcia et al., 2018; Kurcinski et al., 2019; Kuriata, Iglesias, Kurcinski, et al., 2019; Kuriata, Iglesias, Pujols, et al., 2019; Schymkowitz et al., 2005). New automated mutation tools were also introduced to streamline protein engineering processes, making them more accessible to users with varying levels of expertise in protein chemistry.

In response to growing demand for local and customizable versions, the A3D Standalone (Kuriata, Iglesias, Kurcinski, et al., 2019) version was developed, offering the full functionalities of A3D 2.0 (Kuriata, Iglesias, Pujols, et al., 2019) but with the capability for users to operate locally and customize features to suit their specific needs. This standalone version, especially useful for integrating into larger bioinformatics platforms, was well-received by researchers preferring to work within their own computational environments.

In 2024, A3D (Zambrano et al., 2015) underwent a significant update leading to its rebranding as A4D web server (Bárcenas et al., 2024), driven by the inclusion of pH calculations as the “fourth dimension.” This transformative enhancement, along with the integration of environmental factors like pH, allows for more realistic simulations of physiological conditions and provides a comprehensive understanding of how these conditions influence protein stability and aggregation (Arseni et al., 2022; Arseni et al., 2023; Garcia-Pardo & Ventura, 2024; Oeller et al., 2023; Pintado et al., 2021; Santos, Iglesias, Santos-Suárez, et al., 2020; Strohäker et al., 2019). A4D also capitalized on the AlphaFold database (Jumper et al., 2021), enhancing the accuracy of structural predictions by addressing the representation of intrinsically disordered regions (IDRs) and allowing

seamless integration of structural data from various databases (Badaczewska-Dawid et al., 2022; Badaczewska-Dawid et al., 2023; Garcia-Pardo et al., 2023; Ruff & Pappu, 2021; Varadi et al., 2024). Additionally, new automated mutation modes and expanded file compatibility have further enhanced A4D's utility in protein solubility and aggregation studies, marking a pivotal development in the field of protein aggregation prediction.

3 | A4D PIPELINE AND OPTIONS

A4D web server (Bárceñas et al., 2024) (available at <https://biocomp.chem.uw.edu.pl/a4d/>) offers users the choice between utilizing the original A3D scoring method (Zambrano et al., 2015), featuring Static mode and Dynamic mode, or adopting the A4D algorithm, which integrates pH calculations (see Figure 1).

A4D (Bárceñas et al., 2024) defaults to conducting a pH-dependent analysis of aggregation properties and structural stability assessment. It also facilitates input optimization using the FoldX (Schymkowitz et al., 2005) tool. The aggregation propensity calculations are centered around each amino acid α -Carbon atom within a default spherical radius of 10 Å. The A4D score quantifies aggregation propensity for each amino acid i using the following formula:

$$A4D \text{ score} = \text{Agg}_i (\alpha e^{\beta \text{RSA}_i}) + \sum_e [\text{Agg}_e (\alpha e^{\beta \text{RSA}_e}) (\gamma e^{-\delta \text{dist}})] \quad (1)$$

In this equation:

- Agg_i represents the aggregation propensity of the central residue i , and RSA_i denotes its relative surface area exposed to solvent. For static and dynamic mode pathways (Figure 1), the Agg_i score is based on experimental data from *in vivo* studies (Conchillo-Solé et al., 2007; Sánchez de Groot et al., 2005), whereas, in pH calculation mode, Agg_i is determined using a pH-dependent lipophilicity scale (details in Section 3.1).
- The summation includes each neighboring residue e within a 10 Å radius around i , where Agg_e and RSA_e are their respective aggregation propensities and solvent exposure.
- dist refers to the distance from the central residue i to each neighboring residue e .
- α , β , γ , and δ are numerical parameters used to adjust the influence of solvent exposure and distance on the aggregation score, reflecting residue interactions as outlined in work (Zambrano et al., 2015).

For most applications, the A4D score above 0 indicates potential aggregation, but scores higher than 0.5 are more likely to indicate residues that are critical to aggregation processes. A sphere radius of 10 Å is typically used to identify residues forming aggregation-prone regions, though a smaller radius of 5 Å may be employed for detailed analysis of specific residues within these patches (see Section 3.6).

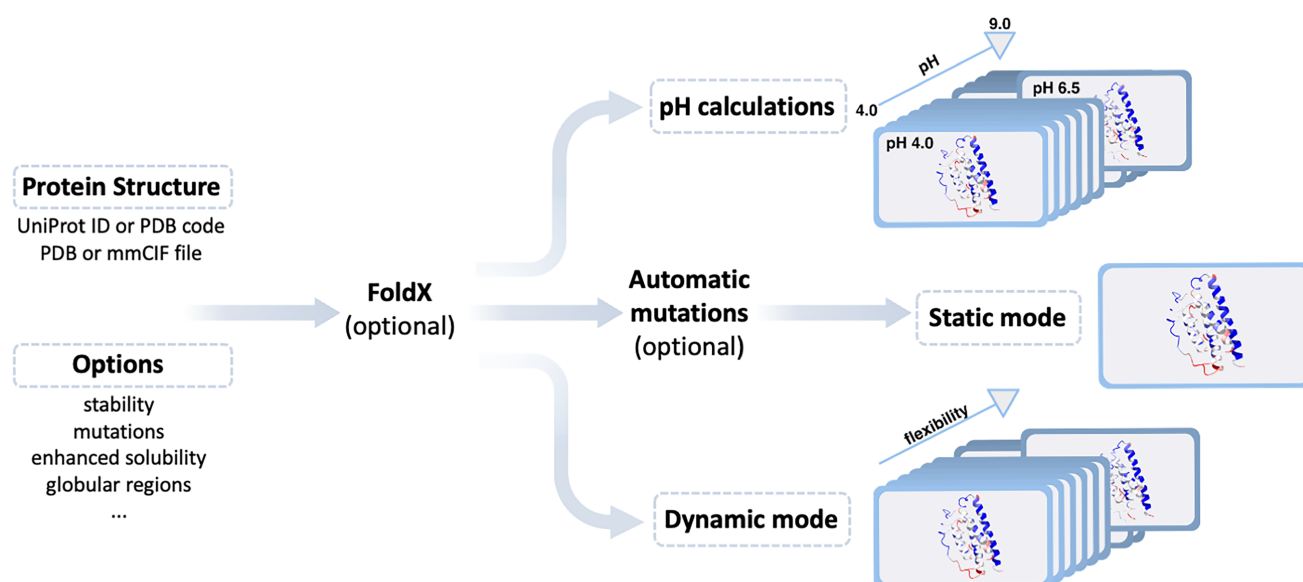


FIGURE 1 A4D pipeline.

TABLE 1 A4D advanced options list.

Name	Description	Value	Note	Section
pH calculations	The simulation will be run across a range of pH values (4–9), results for each pH value will be presented.	Yes/ No	Cannot be used alongside Dynamic mode and Improve/Enhance protein stability options.	3.1
Stability calculations	FoldX (Schymkowitz et al., 2005) is used to optimize input structure.	Yes/ No	Required for mutation calculations.	3.2
Analysis of the globular regions	Use the AlphaCutter tool (Tam & Iwasaki, 2023) to remove the disordered regions of the protein.	Yes/ No	Recommended for structures obtained from the AlphaFold database.	3.3
Dynamic mode	If Yes, A4D will take into account the flexibility of an input structure. If No, A4D will perform analysis on a static input structure.	Yes/ No	Cannot be used alongside pH calculations and Improve/Enhance protein stability options.	3.4
Mutate residues	Mutate your input protein, mutation options will be provided in the next step.	Yes/ No	–	3.5
Distance of aggregation analysis:	Select the radius (R) of the sphere used in the analysis.	5 Å/ 10 Å	–	3.6
Improve protein solubility with evolutionary conserved mutations	Automatically select aggregation prone residues and mutates them into more soluble amino acids using a conservative evolutionary approach.	Yes/ No	Cannot be used alongside pH calculations, Dynamic mode and Enhance protein solubility options.	3.7
Enhance protein solubility with charged mutations	Automatically selects aggregation prone residues and mutates them into charged amino acids.	Yes/ No	Cannot be used alongside pH calculations, Dynamic mode and Improve protein solubility options.	3.7

Note: Each option is described in the corresponding subsection of this manuscript, with the subsection number listed in the last column. Default options are highlighted in bold.

The following section provides a comprehensive explanation of A4D's (Bárcenas et al., 2024) modes and features, ensuring users have a clear understanding of its capabilities for predicting protein aggregation propensities. All available options and their default settings are presented in Table 1.

On the A4D web server, computation times vary based on protein size and the selected options. For most settings, smaller proteins (around 100 amino acids) typically require only a few minutes, while larger proteins (around 300 amino acids) usually take between 10 and 15 min. However, dynamic mode is particularly time-consuming, taking about 20 min for smaller proteins and over an hour for larger ones. Note that these times may vary if the job is queued on the A4D web server.

3.1 | pH calculations

The possibility of including pH-dependent calculations was introduced in 2024 (Bárcenas et al., 2024), with the rebranding of A3D to A4D, marking the integration of environmental pH as the “4th dimension.” The integration of pH calculations into A4D marks a significant leap forward in protein research and

engineering. It empowers researchers to explore the impact of pH variations on protein aggregation in biotechnological contexts, aids in optimizing protein solubility across diverse environmental conditions crucial for industrial applications, and enhances our understanding of disease mechanisms, particularly in conditions where pH dysregulation contributes to protein aggregation-related pathologies.

The propensity for aggregation of pH-sensitive amino acids is influenced by changes in their charge and hydrophobicity, which vary with pH. This results in a pH-dependent alteration in their contribution to aggregation (Santos, Iglesias, Santos-Suárez, et al., 2020). SolupHred has used these factors successfully to predict how pH influences the aggregation of intrinsically disordered proteins (IDPs) (Pintado et al., 2021). The residues neighboring the pH-sensitive residues can shift the pKa values. A4D combines the pH-dependent lipophilicity scale used in SolupHred with 3D-informed pKa calculations for pH-sensitive residues available through the pKa-ANI method (Gokcan & Isayev, 2022) (Figure 2). To calculate the aggregation propensity index (Agg_i) for each amino acid at different pH values, we use these combined data and substitute them into the Equation (1). It's important to note that while there is a strong correlation between the

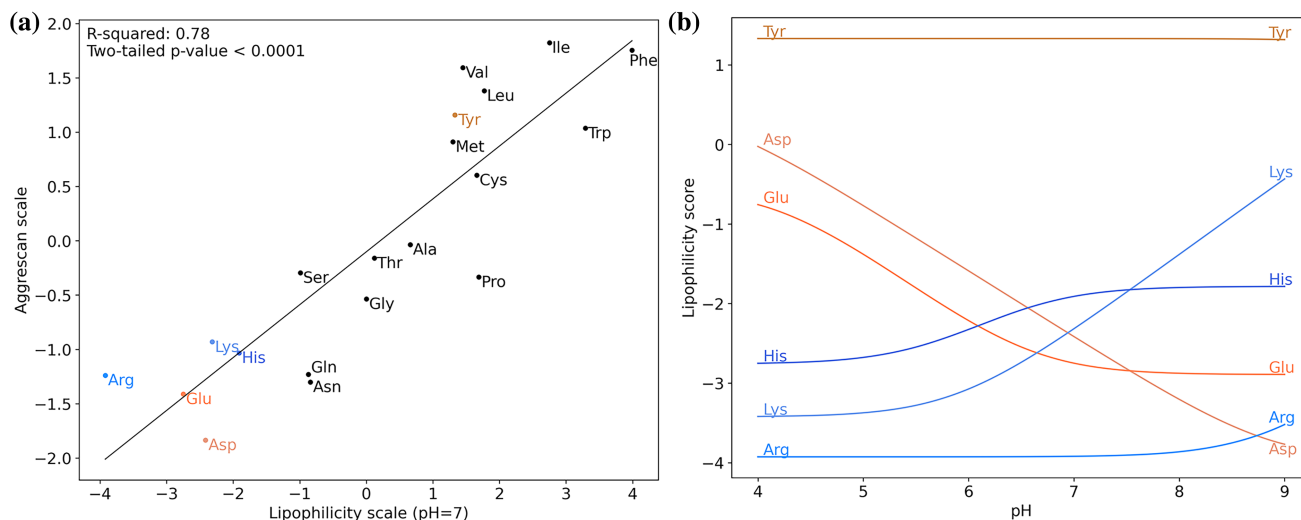


FIGURE 2 Incorporation of pH modulated changes in aggregation propensity in A4D. (a) Linear correlation between the pH-dependent lipophilicity scale used for pH calculations at neutral pH and the *in vivo* derived Aggrescan scale used for non-pH calculations. (b) pH-driven changes in the lipophilicity score. Due to limitations on satisfactory 3D pKa estimation for Arginine, its score is considered constant in the A4D assayed pH-range.

primary aggregation scale of A4D and the pH-dependent lipophilicity scale, their dynamic ranges are different. As a result, A4D reports different numerical values for the aggregation score at neutral pH compared to the “classical” A4D score (i.e., when pH calculations are not considered) (Figure 2a). Regarding pH calculations, A4D considers the pH score variation of aspartic and glutamic acids, lysine, histidine and tyrosine (see Figure 2b), while the score of other amino acids remains unchanged.

If the ‘pH calculations’ option is enabled, A4D facilitates aggregation predictions across a range of 11 pH values from 4 to 9 with increments of 0.5. Subsequently, users gain access to detailed pH-specific predictions on the result page, including structural visualization along with average and max A4D scores for each prediction. Through performance analysis, we determined that for globular proteins, the maximum value, and for IDPs, the average value, better aligns with experimental observations (Bárceñas et al., 2024). Furthermore, A4D adeptly recognizes input structures and recommends the most suitable metric for use. It is important to note that this functionality cannot be combined in a single prediction run with Dynamic mode and automatic mutations, as illustrated in Figure 1.

In our study covering A4D web server (Bárceñas et al., 2024), we examined the accuracy of A4D's pH-dependent predictions across 18 diverse proteins, including 11 globular proteins and 7 IDPs. We measured the performance by comparing the slopes of the linear regressions between the experimental data and A4D metrics. Specifically, we used the A4D max score for globular proteins and the A4D average score for IDPs, achieving 100%

agreement between the predictions and experimental results for both categories. In comparison, CamSol had an overall agreement of 78% (64% for globular proteins and 100% for disordered proteins).

The usage of A4D in predicting pH-dependent aggregation behaviors can be illustrated through the example of Human Serum Albumin (HSA), a critical blood protein involved in substance transportation. Although HSA tends to aggregate into amyloid fibrils under destabilizing conditions, such as temperature or pH changes, the precise impact of pH on its structure and the consequent exposure of the hydrophobic core—which can increase aggregation—remains complex and unpredictable. Therefore, HSA serves as a useful but challenging model for studying aggregation-related phenomena (Malik et al., 2022). Recent studies have illuminated the aggregation behavior of Human Serum Albumin (HSA) under pH alterations, particularly in the shift from pH 7 to 5. Notably, the predictions generated by A4D regarding the aggregation tendencies of HSA (PDB ID: 6YG9) are supported by experimental findings. Utilizing the pH calculations feature of A4D, it was found that at pH 5, HSA shows a heightened propensity for aggregation compared to pH 7. These findings are reflected in the scatter plot shown in Figure 3a, which illustrates a strong concordance between predicted aggregation propensity scores and experimental data, evidenced by an *R*-squared value of 0.88, indicating a robust correlation.

In Figure 3, we present an example where the correlation between experimental data and the predicted A4D score is high across the entire pH range studied. After analyzing the pH-dependent aggregation profile of

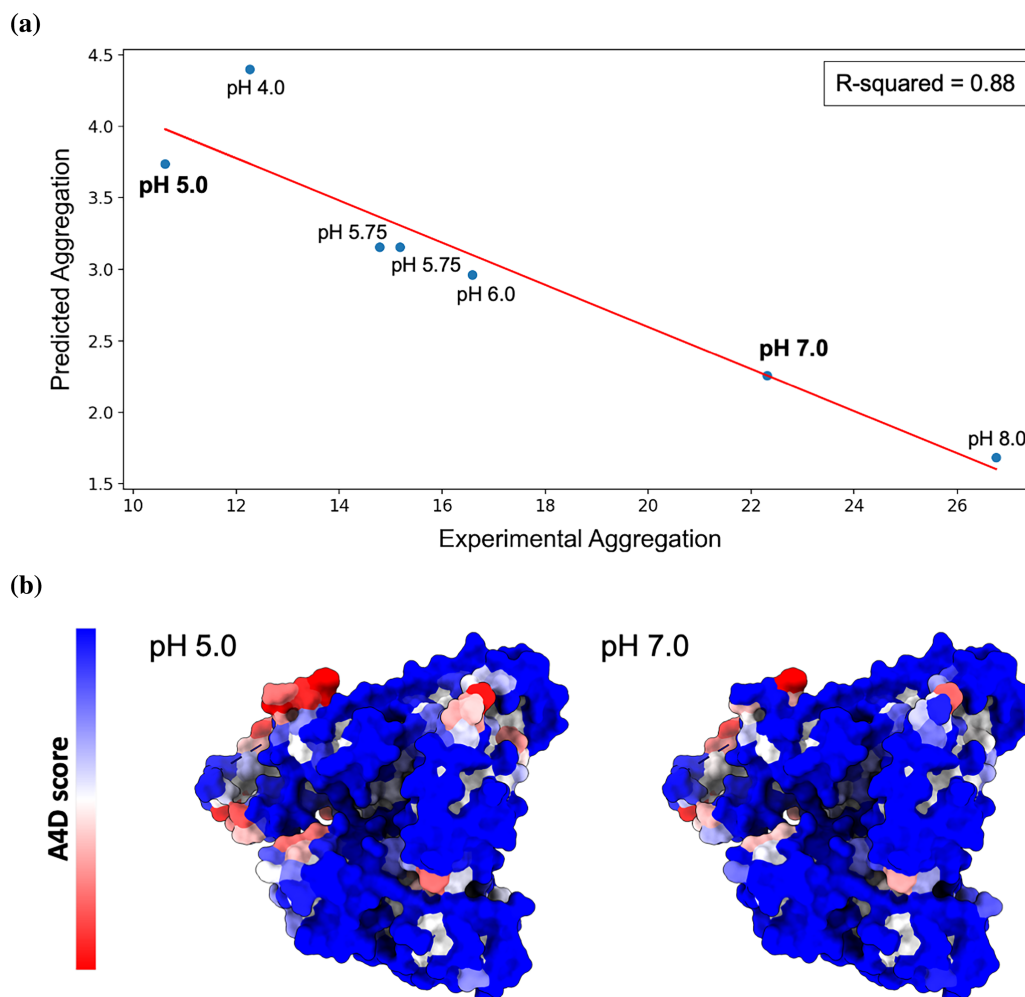


FIGURE 3 Human Serum Albumin (PDB ID: 6YG9) aggregation propensity predictions using A4D pH calculations. (a) Scatter plot illustrating the relationship between experimental data (X-axis) (Oeller et al., 2023) and predicted aggregation propensity scores (Y-axis). Each data point represents a distinct pH, with the experimental values plotted against their corresponding predicted aggregation propensity scores, while the red line represents the linear regression fit. The experimental aggregation value represents the measured aggregation propensity of proteins under specific conditions, as reported in the original A4D work (Bárcenas et al., 2024). For BSA, it is the percentage of aggregated proteins, and for lysozyme and RNase Sa, it is ThT absorbance at 595 nm. All values are normalized. The linear regression model was fitted to the experimental data to estimate the relationship between the experimental values and the predicted aggregation propensity scores. The close alignment between the points and the fitted line indicates a high level of concordance between the experimental data and the predicted aggregation propensity scores. (b) Structural representations of HSA at pH 5.0 (left) and 7.0 (right). Residues are colored by A4D scale: Blue for solubility, red for aggregation, white for neutral regions. Higher predicted propensity for aggregation at pH 5.0, compared to pH 7.0, aligns with experimental findings.

multiple proteins (Bárcenas et al., 2024), various scenarios are possible regarding this correlation. In Figure 4, we have compiled predicted and experimental aggregation values from three proteins: Bovine Serum Albumin (BSA) (Li et al., 2016), Lysozyme (Bárcenas et al., 2024) (supplementary materials), and Ribonuclease Sa (RNase Sa) (Schmittschmitt & Scholtz, 2003). For BSA, the predicted A4D values accurately model the observed experimental aggregation. When analyzing the lysozyme data, we observe that A4D accurately models most of the pH range (pH 3.5–9.5) but fails for moderately basic pH

values (10 and above). This discrepancy is attributed to the partial unfolding of lysozyme at such pH values (Petersen et al., 2004), which is why the A4D server limits pH prediction range to values between 4.0 and 9.0. Finally, in the case of RNase Sa, the predicted and experimental values agree, showing a decrease in aggregation with an increase in pH. However, the rate of change differs. This is a known limitation in the algorithm for some protein case studies and may result from the limitations in A4D's aggregation modeling capabilities or differences in the sensitivity of the experimental value reported.

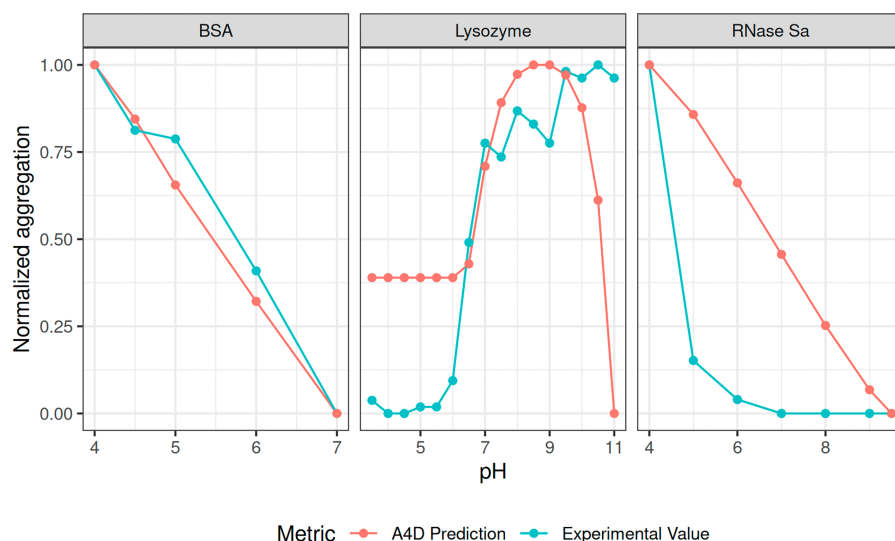


FIGURE 4 Examples of different pH-dependent normalized aggregation profiles, showcasing different behaviors in protein aggregation. Bovine Serum Albumin (BSA), Lysozyme, Ribonuclease Sa (RNase Sa); PDB IDs: 4F5S, 1LYZ chain A, 1C54 chain A, respectively. The predicted values align very well with the experimental data (BSA), fail in extreme pHs (basic pH, lysozyme), or show a different range of variability between experimental and predicted data (RNase SA).

3.2 | Stability calculations

When activated, this feature harnesses FoldX (Schymkowitz et al., 2005) to optimize input structure, consequently enhancing prediction accuracy. The setting should be toggled to “No” only for input structures that have been already optimized, in order to expedite calculations. To perform any mutation, or dynamic mode analysis, this feature has to be enabled. By default, the setting is preset to “Yes.”

3.3 | Analysis of the globular regions

Structures obtained from the AlphaFold database (Jumper et al., 2021) often present elongated, disordered regions that are not typically observed in experimentally resolved structures. These regions tend to congregate around the globular core, introducing potential surface features that could mislead the interpretation of STAPs (Ruff & Pappu, 2021) and undermine the accuracy of A4D predictions (Badaczewska-Dawid et al., 2022; Badaczewska-Dawid et al., 2023).

To address this concern, A4D introduces the “Analysis of the globular regions” feature, which employs AlphaCutter (Tam & Iwasaki, 2023) to selectively remove these disordered segments. AlphaCutter employs DSSP to pinpoint long loops and helices, followed by the calculation of a contact map. Additionally, it applies an AlphaFold pLDDT threshold to refine the domains further. By focusing on globular regions, this approach refines the structural assessment, thereby reducing the impact of artificially generated surface areas. By default, this option is deactivated.

3.4 | Dynamic mode

In some cases, analyzing static structures alone is insufficient; we must also consider their dynamics to accurately predict aggregation capabilities. Protein aggregation onset may not necessarily require global unfolding and transitioning across major energy barriers but can originate from locally unfolded states accessible under physiological conditions, such as thermal fluctuations of native states (Chiti & Dobson, 2009). A4D Dynamic mode is designed to capture and analyze this kind of structural changes, and thereby to facilitate a more comprehensive and dynamic perspective of protein aggregation tendencies. A4D Dynamic mode harnesses CABS-flex tool (Kurcinski et al., 2019; Nithin et al., 2024) to simulate protein fluctuations, allowing for the incorporation of information about protein flexibility into their aggregative properties. In this mode, an input structure, after FoldX (Schymkowitz et al., 2005) optimization, serves as a starting point for CABS-flex simulation. During this stage, a set of all-atom structures representing the predominant fluctuations observed throughout the CABS-flex conformational space sampling is produced. For each of these models, the A4D score is calculated using the same method as in the static mode, and models are ranked from highest to lowest aggregation score, including the original structure. Following the completion of calculations, a comprehensive summary of the Dynamic mode computations, including the models generated via CABS-flex and their corresponding A4D scores, is accessible in the “Dynamic mode details” tab. This summary also includes an RMSF plot, where higher RMSF values generally indicate more flexible regions, potentially exposing hydrophobic patches or aggregation-prone sequences that are otherwise buried. It's important to

note that the Dynamic mode does not support pH calculations or automatic mutations (see Figure 1). By default, Dynamic mode is set to “No.”

Below, we present an example of the bevacizumab Fab fragment (PDB ID: 1BJ1), a monoclonal antibody crucial for inhibiting vascular endothelial growth factor A (VEGF-A) in the treatment of various cancers and eye diseases. The Fab domain of bevacizumab is known for its high aggregation propensity, necessitating reduced concentrations in its therapeutic applications. While A4D static analysis of the Fab domain indicates a certain level of aggregation tendency, the incorporation of Dynamic mode unravels undisclosed aggregation-prone regions not apparent in static structural evaluation (see Figure 5). This suggests that structural fluctuations contribute to an expanded aggregation-prone area, emphasizing the significance of considering protein dynamics in aggregation predictions. The Fab fragment example, along with analogous observation for different antibodies has been comprehensively described in ref (Kuriata, Iglesias, Pujols, et al., 2019), in which the A3D 2.0 tool was used in the analysis of protein flexibility on aggregation properties.

3.5 | Mutations

A4D offers the capability to introduce mutations both before and after conducting calculations, enhancing users' control over their analyses. By default, the Mutate residues option is initially set to “No.” However, users can activate this feature to access a dedicated window where they can perform individual or multiple mutations. This process involves selecting the desired amino acid from a provided list and specifying the target amino acid for mutation. Importantly, unlike automated mutation methods, the Mutate residues option seamlessly integrates with pH calculations and Dynamic mode. Notably, enabling the Mutate residues option requires the activation of Stability calculations, leveraging the FoldX (Schymkowitz et al., 2005) tool for mutation provision.

Mutations can be also performed, through the “Aggrescan4D score” tab, accessible on the results page after completing calculations. This tab features an interactive table displaying the A4D values for each residue, along with an intuitive interface enabling users to easily select desired mutations and resubmit calculation for mutated protein.

3.6 | Distance of aggregation analysis

Aggregation propensity is determined within spherical regions centered on each residue's C α carbon. The

“Distance of aggregation analysis” option offers a choice between two distinct distance settings: 5 or 10 Å.

The selection between the 5 and 10 Å settings depends on the desired level of precision and detail in the analysis. The default 10 Å radius provides a comprehensive overview of aggregation tendencies throughout the protein structure, while opting for the 5 Å radius allows for a more refined examination, enabling researchers to delve into individual residue impacts with greater precision.

3.7 | Automated mutations to improve protein solubility

A4D offers an Automated mutation feature aimed at identifying more soluble variants of functional therapeutic proteins. This feature serves as an efficient and user-friendly method, accessible via “Enhance protein solubility with charged mutations” and “Improve protein solubility with evolutionary conserved mutations” input options. Its purpose is to offer a computational alternative to time-consuming and costly experimental approaches such as phage display.

Under the setting “Enhance protein solubility with charged mutations,” A4D identifies aggregation-prone regions within the protein structure and virtually introduces mutations by incorporating charged amino acids at corresponding residues to impede protein self-association. Upon analysis completion, A4D generates a ranked list of suggested point mutations, accompanied by an evaluation of their solubilizing effects and energetic implications. This empowers researchers to manually select mutations that enhance solubility while considering potential destabilization. Furthermore, users can exclude specific residues from the automated mutation process to prevent mutations of critical functional fragments within the protein sequence.

The setting “Improve protein solubility with evolutionarily conserved mutations” focuses on minimizing changes in physicochemical properties that could disrupt protein folding and stability. This option ensures that suggested mutations result from a single point mutation (A $\rightarrow$ B) rather than a series of successive changes (A $\rightarrow$ C $\rightarrow$ B), thus minimizing impacts on protein structure and stability. To achieve this, A4D employs Point Accepted Mutation (PAM) matrices. These matrices store the sum of accepted substitutions in closely related sequences, with an identity threshold of 0.85 or higher. By utilizing PAM matrices, A4D ensures that the suggested mutations preserve evolutionary conservation while enhancing protein solubility. This approach optimizes protein stability and maintains biological functionality.

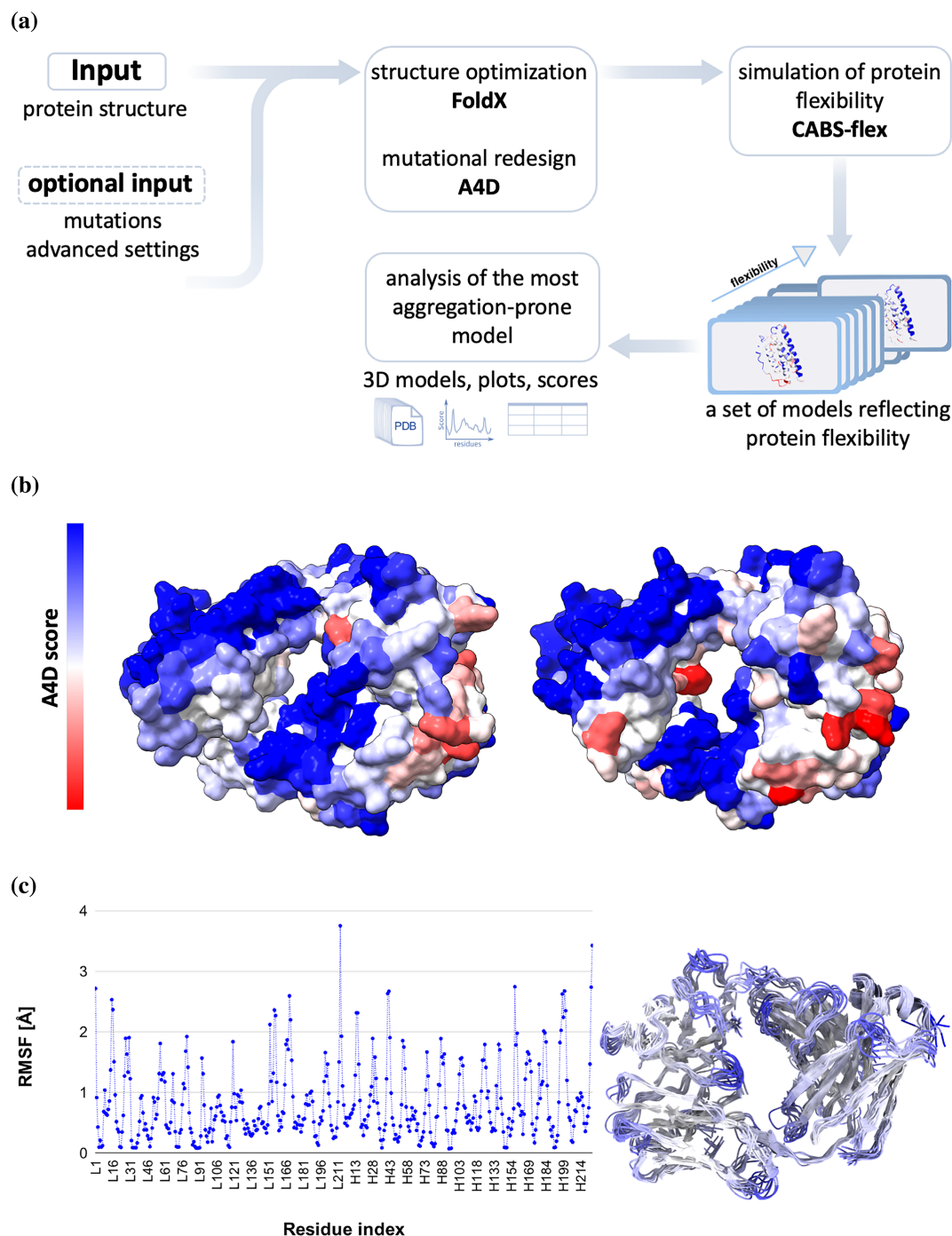


FIGURE 5 Aggrescan4D Dynamic mode. (a) Dynamic mode pipeline. (b) Bevacizumab Fab fragment (PDB ID: 1BJ1) aggregation propensity predictions using A4D Static mode (left side) and Dynamic mode (right side). Residues are colored by A4D scale: Blue for solubility, red for aggregation, white for neutral regions. Dynamic mode reveals aggregation-prone regions that are not evident in Static mode. (c) RMSF (Root Mean Square Fluctuation) plot from the CABS-flex simulation (left) and superposition of models produced by CABS-flex (right).

These functionalities are summarized in Figure 6. By default, both mutation options are disabled. It's noteworthy that automated mutations are not compatible with pH calculations and Dynamic mode (see Figure 1).

3.8 | Privacy

To prioritize privacy, users can select the “Do not show my job on the results page” checkbox, found on the main page's options panel. This feature allows users to control

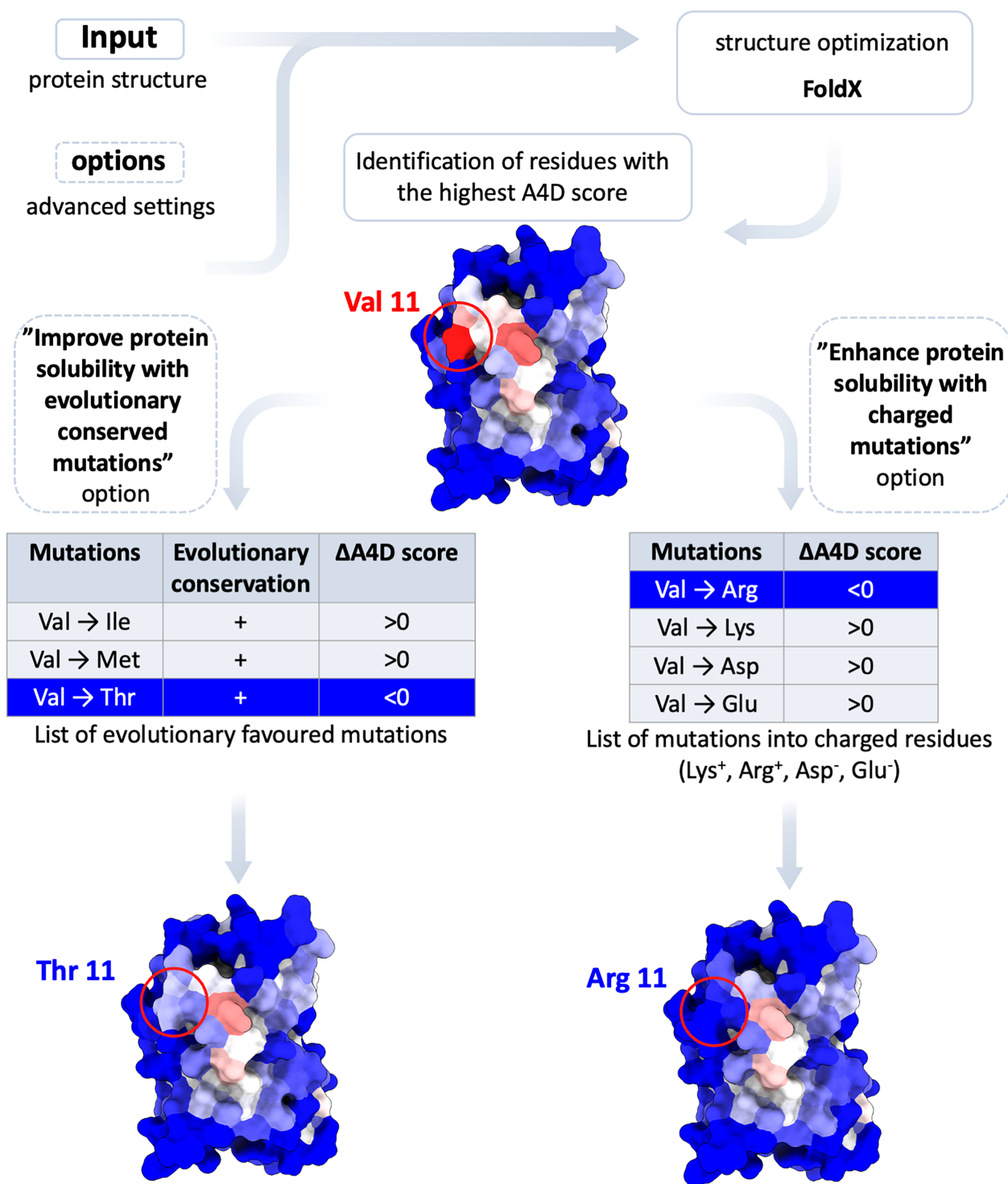


FIGURE 6 Aggrescan4D automated mutations pipeline. In the automated mutations mode, A4D identifies residues with the highest A4D score, subsequently mutating them into positively charged residues (Lysine and Arginine) or negatively charged residues (Aspartic and Glutamic acids). It then calculates A4D scores and stability increments for each variant, providing a ranked list of mutations. Mutations that do not negatively impact stability are prioritized, appearing at the top of the list, while those that affect stability are highlighted in orange (energetic effect >0.5 kcal/mol) and red (energetic effect >1.0 kcal/mol). In the "Improving protein solubility with evolutionary conserved mutations" mode, A4D performs mutations which impact less on protein stability based on protein residue based on PAM250 matrix residue distances.

the visibility of their analyses, which is especially beneficial for handling sensitive data. Furthermore, the standalone version of A3D (outlined in Methods Section 3.11)

allows users to conduct analyses locally on their own machines, granting complete control over the research process. This eliminates the need to upload sensitive data

to a remote server, ensuring maximum confidentiality for researchers.

3.9 | A4D RESTful service

The A4D web server extends its capabilities through a programmatically accessible interface, providing researchers with the capability of integrating aggregation predictions into their computational workflows. RESTful URLs, featured by A4D web server, allow users to initiate new jobs either by uploading a custom PDB/CIF file or with specific PDB/UniProt IDs. RESTful service comprehensively supports the entire A4D web server functionality, enabling users defining project names, receiving email notifications and configuring input parameters. Upon job submission, using cURLs, users can monitor progress, access detailed information and download complete A4D output. The RESTful API can be harnessed through Python or other tools capable of POST requests.

For a comprehensive overview of A4D RESTful service options and usage examples, please refer to ref (Pujols et al., 2021). Additionally, detailed RESTful guides and examples can be found on: <http://biocomp.chem.uw.edu.pl/a4d/tutorial>.

3.10 | Database for model organisms

The introduction of AlphaFold and the accompanying AlphaFold Protein Structure Database (Varadi et al., 2024), which includes over 200 million protein structures, has made a significant impact on structural biology. This development provides access to accurate protein structures, especially for proteins that had not been structurally characterized before, enhancing our understanding and opening new avenues in scientific research. Leveraging the availability of AlphaFold's structures, we initially proposed A3D predictions for the human proteome (Badaczewska-Dawid et al., 2022) followed by predictions for yeast (Garcia-Pardo et al., 2023). Subsequently, we developed the A3D Model Organism Database (A3D-MODB) (Badaczewska-Dawid et al., 2023). This resource contains a vast collection of over 160,000 proteins from the proteomes of 12 commonly studied organisms: *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Danio rerio*, *Escherichia coli*, *Mycoplasma genitalium*, *Mus musculus*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Homo sapiens*, *Rattus norvegicus*, *Drosophila melanogaster*, and SARS-CoV-2. The primary goal of A3D-MODB is to facilitate extensive cross-species analysis related to protein aggregation. It provides

valuable predictions regarding solubility and stability, as well as confidence scores for each residue modeled by AlphaFold (pLDDT). Additionally, it offers topology predictions for membrane proteins, calculated using TOPCONS (Tsirigos et al., 2015).

Each entry in the A3D-MODB database (Badaczewska-Dawid et al., 2023) is linked to the UniProt Knowledgebase (UniProtKB) (UniProt Consortium, 2023) and species-specific databases, providing comprehensive coverage and making it an indispensable resource for researchers in various scientific fields. The A3D-MODB database is freely accessible, requiring no registration, and can be reached at <http://biocomp.chem.uw.edu.pl/A3D2/MODB>. Complementing these capabilities, the A4D web server allows users to input UniProt IDs for automatic retrieval of AlphaFold structures. It also integrates seamlessly with A3D-MODB, providing easy access to nearly 500,000 precalculated jobs. This integration simplifies research processes and enhances access to valuable structural biology data.

3.11 | A3D standalone package

A3D is also available as an independent, self-contained software application (Kuriata, Iglesias, Kurcinski, et al., 2019) designed for local use, ensuring enhanced privacy. This standalone version (Kuriata, Iglesias, Kurcinski, et al., 2019) integrates seamlessly with researchers' workflows, providing complete control over every stage of the process (the standalone version tutorial has been presented in the monograph (Kuriata et al., 2022)). Additionally, its extensional integration with the CABS-flex tool allows for interactive analysis and visualization of models generated by CABS-flex, as well as comprehensive control over CABS-flex simulations and their parameters.

Despite the robust features of A3D standalone (Kuriata, Iglesias, Kurcinski, et al., 2019), it can be easily installed via Python package managers such as pip or Conda, or as a Docker image. This installation flexibility ensures compatibility with Windows, MacOS, and Linux operating systems. The standalone version includes an integrated local server and offers a user-friendly graphical interface that mirrors the A3D web-server (Zambrano et al., 2015) interface.

It is important to note that the current standalone version of A3D operates on Python 2.7. While this is not the latest version of Python, the next iteration of the standalone A4D, is currently under development. A4D standalone will be fully compatible with modern Python releases, ensuring smooth integration and enhanced functionality. Installation guides, documentation, and

usage examples are available in the A3D repository at <https://bitbucket.org/lcbio/aggrescan3d>.

4 | CASE STUDY: EVALUATION OF AGGREGATION PREDICTION METHODS WITH MEDI1912 MONOCLONAL ANTIBODY VARIANTS

Along with A4D (Bárceñas et al., 2024), SolubiS (De Baets et al., 2015) and CamSol (Sormanni et al., 2015) are widely employed tools to understand and rationally design structural aggregation propensity. The three tools offer aggregation propensity reporters and the possibility to inform users with mutational variants predicted to have less propensity to self-association.

Recently, David Brockwell and collaborators developed an experimental pipeline for the *in vivo* development of more soluble proteoforms focusing on highly aggregating MEDI1912 monoclonal antibody which targets human nerve growth factor with picomolar affinity (Ebo et al., 2020) (Figure 7). The authors developed

18 different variants and recorded the full-length IgG aggregation propensity through retention time on high-performance size exclusion chromatography (HP-SEC), a reporter of aggregation-driven monomer loss. Although the described methodology allowed the development of novel variants, the lack of control over the mutation sites caused these less aggregating variants to impact protein functionality (Ebo et al., 2020). This dataset reporting experimental evidence for protein aggregation variants allowed us to evaluate the performance of the A4D, SolubiS, and CamSol structural aggregation prediction methods on protein mutations (antibody structure was a courtesy of the Brockwell lab at the University of Leeds). We categorized the mutants based on their aggregation propensity considered positives (9 variants) or negatives (8 variants) those that increased or decreased the retention time with respect to the midpoint. As reported in Table 2, A4D displays the highest sensitivity, accuracy, and specificity in predicting the effects of mutational variants on experimental aggregation, compared with structural-based state-of-the-art prediction algorithms.

The three tools also differ in the additional possibilities they provide. A4D like SolubiS allow users different

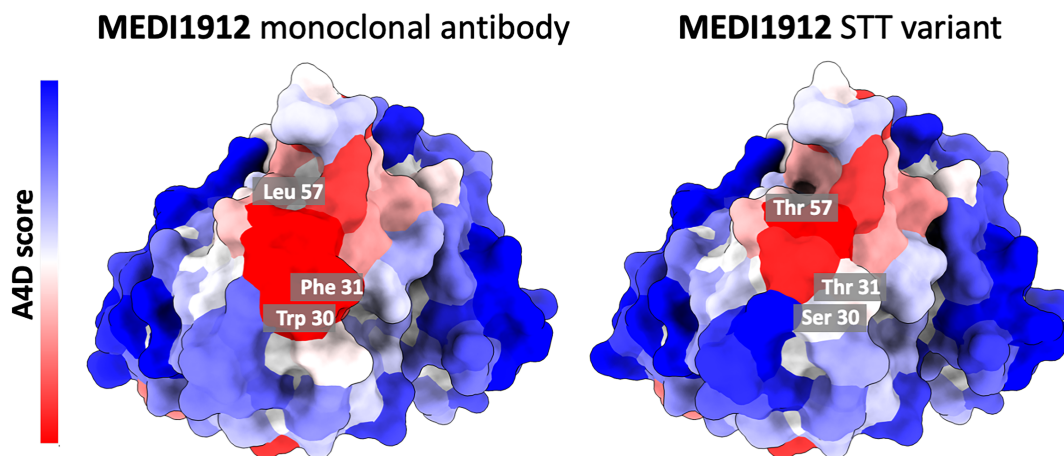


FIGURE 7 Monoclonal antibody MEDI1912 and its more soluble engineered variant STT. Residues are colored by A4D scale: Blue for solubility, red for aggregation, white for neutral regions. MEDI1912 STT differs from wild type protein with three mutations in the complementary complementarity determining region; Trp30Ser, Phe31Thr and Leu57Thr (Dobson et al., 2016).

TABLE 2 Performance of A4D, SolubiS and CamSol for a dataset of 19 mutants of the MEDI1912 monoclonal antibody as reported in ref (Ebo et al., 2020).

	Sensitivity	Specificity	Precision	Accuracy	MCC ^a	FDR ^b
A4D	0.89	0.75	0.80	0.82	0.65	0.20
CamSol	0.78	0.25	0.54	0.53	0.03	0.46
SolubiS	0.89	0.38	0.62	0.65	0.31	0.39

Note: The best performance according to each particular parameter is shown in bold. Mutational impact for A4D was assessed by the maximum score (the best reporter for experimental properties of folded proteins), for SolubiS the Tango score and for CamSol the structurally corrected score.

^aMatthews correlation coefficient.

^bFalse discovery rate.

methods of inputting structures through user-uploaded PDB/CIF files or PDB/UniProt ID without requiring registration. A4D nonetheless also allows for large-scale analysis in an automatic way through its RESTful API. Additionally, A4D allows carrying out a diversity of *in situ* analysis of the results through a potent yet user-friendly user-interface for instance allowing tagging residues in the A4D colored protein structure or calculating impact of mutations from the results page.

5 | CONCLUSIONS

Aggrescan4D (A4D) (Bárcenas et al., 2024) represents a refined evolution of Aggrescan3D (A3D) (Zambrano et al., 2015), as it considers the effect of the pH, an important environmental factor heavily impacting protein aggregation. This is important, since protein aggregation underlies the onset of human diseases, challenges drug development and protein formulation.

A4D (Bárcenas et al., 2024) enhances the functionalities of A3D, maintaining the user-friendly interface that facilitated its wide acceptance within the protein science community. It introduces innovative features like pH-dependent aggregation prediction and an evolutionary-informed automatic mutation protocol. These enhancements aim to improve the design of protein solubility and stability without compromising structural integrity, complemented with A4D's integration with comprehensive databases like the A3D-MODB (Badaczewska-Dawid et al., 2023) and AlphaFold (Jumper et al., 2021).

Beyond its core functions, A3D's application has been extended through its integration into various bioinformatics tools, broadening its impact across different scientific areas. Its incorporation into the Protein Solubilizing Server (PSS) (Tao et al., 2022) aids in addressing the solubilization challenges of structurally complex membrane proteins, utilizing the innovative QTY code-based engineering approach to enhance their water solubility. This integration not only facilitates the design of more effective biopharmaceuticals but also supports basic research into membrane protein structures and functions. Similarly, DE-STRESS (Stam & Wood, 2021) utilizes A3D to rigorously evaluate structural models of proteins, providing researchers with critical insights that guide the engineering of proteins with optimized stability and function. This tool enables a systematic assessment of protein designs, enhancing the predictability and success rate of protein engineering efforts. We expect that A4D's versatility would be soon exploited by the bioinformatics ecosystem beyond aggregation prediction in the broader context of protein design and analysis.

Looking to the future, there are several potential enhancements to the A4D platform that could substantially improve its utility in the design of therapeutic proteins. One key advancement would involve the integration of machine learning algorithms to perform mutational analysis. These algorithms can analyze extensive datasets of known protein mutations, extracting patterns to predict their effects on protein behavior (Pang et al., 2023). Furthermore, the incorporation of empirical aggregation data into these models could refine the predictive capabilities, leveraging real-world experimental results to increase accuracy (Housmans et al., 2023). Updating the A4D database to include recent discoveries on aggregation-prone regions and resistant mutations would also be advantageous. Additionally, expanding the structural dynamics suite to include longer-term dynamics and more extensive conformational changes could enhance simulations and approach them to biologically/industrially relevant scenarios.

These prospective updates would enhance A4D's functionality and expand its application in biotechnology and biomedical research. As A4D continues to evolve, it can play an important role in uncovering new therapeutic strategies and deepening our understanding of protein aggregation-related disorders.

AUTHOR CONTRIBUTIONS

Mateusz Zalewski: Writing – original draft; investigation; visualization; conceptualization; writing – review and editing. **Valentin Iglesias:** Writing – original draft; formal analysis; investigation; visualization; conceptualization; writing – review and editing. **Oriol Bárcenas:** Writing – review and editing; writing – original draft; visualization. **Salvador Ventura:** Conceptualization; writing – review and editing. **Sebastian Kmiecik:** Conceptualization; writing – review and editing; writing – original draft; funding acquisition.

ACKNOWLEDGMENTS

The authors would like to thank Prof. David Brockwell for providing the structural model for MEDI1912 antibody. Mateusz Zalewski has been supported by the National Science Centre, Poland (Sheng 2021/40/Q/NZ2/00078). Valentin Iglesias has been supported by the Polish National Agency for Academic Exchange under the ULAM NAWA Programme (Grant agreement no. BPN/ULM/2023/1/00189/U/00001). Oriol Bárcenas has been supported by the Spanish Ministry of Science and Innovation via a doctoral grant (FPU22/03656). Salvador Ventura has been supported by the Spanish Ministry of Science and Innovation (PID2022-137963OB-I00), Generalitat de Catalunya (2021-SGR-00635 AGAUR) (CERCA Programme) and by ICREA, ICREA-Academia

2020. Sebastian Kmiecik has been supported by the National Science Centre, Poland (Sheng 2021/40/Q/NZ2/00078).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

ORCID

Mateusz Zalewski  <https://orcid.org/0000-0002-3436-8829>

Salvador Ventura  <https://orcid.org/0000-0002-9652-6351>

Sebastian Kmiecik  <https://orcid.org/0000-0001-7623-0935>

REFERENCES

- Arseni D, Chen R, Murzin AG, Peak-Chew SY, Garringer HJ, Newell KL, et al. TDP-43 forms amyloid filaments with a distinct fold in type A FTLTDP. *Nature*. 2023;620:898–903.
- Arseni D, Hasegawa M, Murzin AG, Kametani F, Arai M, Yoshida M, et al. Structure of pathological TDP-43 filaments from ALS with FTLTDP. *Nature*. 2022;601:139–43.
- Badaczewska-Dawid AE, Garcia-Pardo J, Kuriata A, Pujols J, Ventura S, Kmiecik S. A3D database: structure-based predictions of protein aggregation for the human proteome. *Bioinformatics*. 2022;38:3121–3.
- Badaczewska-Dawid AE, Kuriata A, Pintado-Grima C, Garcia-Pardo J, Burdukiewicz M, Iglesias V, et al. A3D model organism database (A3D-MODB): a database for proteome aggregation predictions in model organisms. *Nucleic Acids Res*. 2023; 52:D360–D367.
- Bárceñas O, Kuriata A, Zalewski M, Iglesias V, Pintado-Grima C, Firlik G, et al. Aggrescan4D: structure-informed analysis of pH-dependent protein aggregation. *Nucleic Acids Res*. 2024;1762: 427–443.
- Belli M, Ramazzotti M, Chiti F. Prediction of amyloid aggregation in vivo. *EMBO Rep*. 2011;12:657–63.
- Berdyński M, Misztal P, Safranow K, Andersen PM, Morita M, Filippek S, et al. SOD1 mutations associated with amyotrophic lateral sclerosis analysis of variant severity. *Sci Rep*. 2022; 12:103.
- Bhandare VV, Ramaswamy A. The proteinopathy of D169G and K263E mutants at the RNA recognition motif (RRM) domain of tar DNA-binding protein (tdp43) causing neurological disorders: a computational study. *J Biomol Struct Dyn*. 2018;36: 1075–93.
- Carija A, Pinheiro F, Iglesias V, Ventura S. Computational assessment of bacterial protein structures indicates a selection against aggregation. *Cells*. 2019;8:856.
- Castillo V, Espargaró A, Gordo V, Vendrell J, Ventura S. Deciphering the role of the thermodynamic and kinetic stabilities of SH3 domains on their aggregation inside bacteria. *Proteomics*. 2010; 10:4172–85.
- Chennamsetty N, Voynov V, Kayser V, Helk B, Trout BL. Design of therapeutic proteins with enhanced stability. *Proc Natl Acad Sci U S A*. 2009;106:11937–42.
- Chiti F, Dobson CM. Amyloid formation by globular proteins under native conditions. *Nat Chem Biol*. 2009;5:15–22.
- Chiti F, Dobson CM. Protein misfolding, amyloid formation, and human disease: a summary of Progress over the last decade. *Annu Rev Biochem*. 2017;86:27–68.
- Conchillo-Solé O, de Groot NS, Avilés FX, Vendrell J, Daura X, Ventura S. AGGRESCAN: a server for the prediction and evaluation of “hot spots” of aggregation in polypeptides. *BMC Bioinformatics*. 2007;8:65.
- De Baets G, Van Durme J, van der Kant R, Schymkowitz J, Rousseau F. Solubis: optimize your protein. *Bioinformatics*. 2015;31:2580–2.
- de Groot NS, Ventura S. Protein aggregation profile of the bacterial cytosol. *PLoS One*. 2010;5:e9383.
- Díaz-Caballero M, Navarro S, Fuentes I, Teixidor F, Ventura S. Minimalist prion-inspired polar self-assembling peptides. *ACS Nano*. 2018;12:5394–407.
- Dobson CL, Devine PWA, Phillips JJ, Higazi DR, Lloyd C, Popovic B, et al. Engineering the surface properties of a human monoclonal antibody prevents self-association and rapid clearance in vivo. *Sci Rep*. 2016;6:38644.
- Ebo JS, Saunders JC, Devine PWA, Gordon AM, Warwick AS, Schiffrin B, et al. An in vivo platform to select and evolve aggregation-resistant proteins. *Nat Commun*. 2020;11:1816.
- Ejaz S, Paracha RZ, Ejaz S, Jamal Z. Antibody designing against IIIabc junction (JIIIabc) of HCV IRES through affinity maturation; RNA-antibody docking and interaction analysis. *PLoS One*. 2023;18:e0291213.
- Espargaró A, Castillo V, de Groot NS, Ventura S. The in vivo and in vitro aggregation properties of globular proteins correlate with their conformational stability: the SH3 case. *J Mol Biol*. 2008;378:1116–31.
- Fernandez-Escamilla A-M, Rousseau F, Schymkowitz J, Serrano L. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat Biotechnol*. 2004; 22:1302–6.
- Flores-León M, Lázaro DF, Shvachiy L, Krisko A, Outeiro TF. In silico analysis of the aggregation propensity of the SARS-CoV-2 proteome: insight into possible cellular pathologies. *Biochim Biophys Acta Proteins Proteom*. 2021;1869:140693.
- Garcia-Pardo J, Badaczewska-Dawid AE, Pintado-Grima C, Iglesias V, Kuriata A, Kmiecik S, et al. A3DyDB: exploring structural aggregation propensities in the yeast proteome. *Microb Cell Fact*. 2023;22:186.
- Garcia-Pardo J, Ventura S. Cryo-EM structures of functional and pathological amyloid ribonucleoprotein assemblies. *Trends Biochem Sci*. 2024;49:119–33.
- Gil-Garcia M, Baño-Polo M, Varejão N, Jamroz M, Kuriata A, Díaz-Caballero M, et al. Combining structural aggregation propensity and stability predictions to redesign protein solubility. *Mol Pharm*. 2018;15:3846–59.
- Gil-Garcia M, Navarro S, Ventura S. Coiled-coil inspired functional inclusion bodies. *Microb Cell Fact*. 2020;19:117.
- Gokcan H, Isayev O. Prediction of protein pKa with representation learning. *Chem Sci*. 2022;13:2462–74.
- Housmans JAJ, Wu G, Schymkowitz J, Rousseau F. A guide to studying protein aggregation. *FEBS J*. 2023;290:554–83.
- Ishwarilall TZ, Adeleke VT, Maharaj L, Okpeku M, Adeniyi AA, Adeleke MA. Identification of potential candidate vaccines against *Mycobacterium ulcerans* based on the major facilitator superfamily transporter protein. *Front Immunol*. 2022;13: 1023558.

- Jamroz M, Kolinski A, Kmiecik S. CABS-flex predictions of protein flexibility compared with NMR ensembles. *Bioinformatics*. 2014;30:2150–4.
- Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021;596:583–9.
- Katina NS, Balobanov VA, Ilyina NB, Vasiliev VD, Marchenkov VV, Glukhov AS, et al. Sw ApoMb amyloid aggregation under nondenaturing conditions: the role of native structure stability. *Biophys J*. 2017;113:991–1001.
- Knowles TPJ, Mezzenga R. Amyloid fibrils as building blocks for natural and artificial functional materials. *Adv Mater*. 2016;28:6546–61.
- Ko SK, Björkengren G, Berner C, Winter G, Harris P, Peters GHJ. Combining molecular dynamics simulations and biophysical characterization to investigate protein-specific excipient effects on Reteplase during freeze drying. *Pharmaceutics*. 2023;15:1854.
- Kurcinski M, Oleniecki T, Ciemny MP, Kuriata A, Kolinski A, Kmiecik S. CABS-flex standalone: a simulation environment for fast modeling of protein flexibility. *Bioinformatics*. 2019;35:694–5.
- Kuriata A, Badaczewska-Dawid AE, Pujols J, Ventura S, Kmiecik S. Protocols for rational design of protein solubility and aggregation properties using Aggrescan3D standalone. In: Li MS, Kloczkowski A, Cieplak M, Kouza M, editors. *Computer simulations of aggregation of proteins and peptides*. New York, NY: Springer US; 2022. p. 17–40. https://doi.org/10.1007/978-1-0716-1546-1_2
- Kuriata A, Gierut AM, Oleniecki T, Ciemny MP, Kolinski A, Kurcinski M, et al. CABS-flex 2.0: a web server for fast simulations of flexibility of protein structures. *Nucleic Acids Res*. 2018;46:W338–43.
- Kuriata A, Iglesias V, Kurcinski M, Ventura S, Kmiecik S. Aggrescan3D standalone package for structure-based prediction of protein aggregation properties. *Bioinformatics*. 2019;35:3834–5.
- Kuriata A, Iglesias V, Pujols J, Kurcinski M, Kmiecik S, Ventura S. Aggrescan3D (A3D) 2.0: prediction and engineering of protein solubility. *Nucleic Acids Res*. 2019;47:W300–7.
- Leppert A, Chen G, Lama D, Sahin C, Railaite V, Shilkova O, et al. Liquid-liquid phase separation primes spider silk proteins for fiber formation via a conditional sticker domain. *Nano Lett*. 2023;23:5836–41.
- Li R, Wu Z, Wangb Y, Ding L, Wang Y. Role of pH-induced structural change in protein aggregation in foam fractionation of bovine serum albumin. *Biotechnol Rep (Amst)*. 2016;9:46–52.
- Malik S, Zaidi N, Siddiqi MK, Majid N, Masroor A, Salam S, et al. Mechanistic insight into inhibition of amyloid fibrillation of human serum albumin by vildagliptin. *Colloids Surf B Biointerfaces*. 2022;216:112563.
- Nithin C, Fornari RP, Pilla SP, Wroblewski K, Zalewski M, Madaj R, et al. Exploring protein functions from structural flexibility using CABS-flex modeling. *Protein Sci*. 2024;33:e5090.
- Oeller M, Kang R, Bell R, Ausserwöger H, Sormanni P, Vendruscolo M. Sequence-based prediction of pH-dependent protein solubility using CamSol. *Brief Bioinform*. 2023;24:bbad004.
- Pang KT, Yang YS, Zhang W, Ho YS, Sormanni P, Michaels TCT, et al. Understanding and controlling the molecular mechanisms of protein aggregation in mAb therapeutics. *Bio-technol Adv*. 2023;67:108192.
- Petersen SB, Jonson V, Fojan P, Wimmer R, Pedersen S. Sorbitol prevents the self-aggregation of unfolded lysozyme leading to and up to 13 degrees C stabilisation of the folded form. *J Biotechnol*. 2004;114:269–78.
- Petrlova J, Samsudin F, Bond PJ, Schmidtchen A. SARS-CoV-2 spike protein aggregation is triggered by bacterial lipopolysaccharide. *FEBS Lett*. 2022;596:2566–75.
- Pintado C, Santos J, Iglesias V, Ventura S. SolupHred: a server to predict the pH-dependent aggregation of intrinsically disordered proteins. *Bioinformatics*. 2021;37:1602–3.
- Pujols J, Iglesias V, Santos J, Kuriata A, Kmiecik S, Ventura S. A3D 2.0 update for the prediction and optimization of protein solubility. 2021.04.13.439600. 2021 <https://doi.org/10.1101/2021.04.13.439600v1>
- Pulido D, Arranz-Trullén J, Prats-Ejarque G, Velázquez D, Torrent M, Moussaoui M, et al. Insights into the antimicrobial mechanism of action of human RNase6: structural determinants for bacterial cell agglutination and membrane permeation. *Int J Mol Sci*. 2016;17:552.
- Pulido P, Llamas E, Llorente B, Ventura S, Wright LP, Rodríguez-Concepción M. Specific Hsp100 chaperones determine the fate of the first enzyme of the Plastidial isoprenoid pathway for either refolding or degradation by the stromal Clp protease in Arabidopsis. *PLoS Genet*. 2016;12:e1005824.
- Ruff KM, Pappu RV. AlphaFold and implications for intrinsically disordered proteins. *J Mol Biol*. 2021;433:167208.
- Ruiz-Solaní N, Salguero-Linares J, Armengot L, Santos J, Pallarès I, van Midden KP, et al. Arabidopsis metacaspase MC1 localizes in stress granules, clears protein aggregates, and delays senescence. *Plant Cell*. 2023;35:3325–44.
- Sánchez de Groot N, Pallarès I, Avilés FX, Vendrell J, Ventura S. Prediction of “hot spots” of aggregation in disease-linked polypeptides. *BMC Struct Biol*. 2005;5:18.
- Santos J, Iglesias V, Santos-Suárez J, Mangiagalli M, Brocca S, Pallarès I, et al. pH-dependent aggregation in intrinsically disordered proteins is determined by charge and lipophilicity. *Cells*. 2020;9:145.
- Santos J, Iglesias V, Ventura S. Computational prediction and redesign of aberrant protein oligomerization. *Prog Mol Biol Transl Sci*. 2020;169:43–83.
- Santos J, Pujols J, Pallarès I, Iglesias V, Ventura S. Computational prediction of protein aggregation: advances in proteomics, conformation-specific algorithms and biotechnological applications. *Comput Struct Biotechnol J*. 2020;18:1403–13.
- Schmittschmitt JP, Scholtz JM. The role of protein stability, solubility, and net charge in amyloid fibril formation. *Protein Sci*. 2003;12:2374–8.
- Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L. The FoldX web server: an online force field. *Nucleic Acids Res*. 2005;33:W382–8.
- Shukla AA, Thömmes J. Recent advances in large-scale production of monoclonal antibodies and related proteins. *Trends Biotechnol*. 2010;28:253–61.
- Sormanni P, Aprile FA, Vendruscolo M. The CamSol method of rational design of protein mutants with enhanced solubility. *J Mol Biol*. 2015;427:478–90.

- Stam MJ, Wood CW. DE-STRESS: a user-friendly web application for the evaluation of protein designs. *Protein Eng Des Select*. 2021;34:gzab029.
- Strohäker T, Jung BC, Liou S-H, Fernandez CO, Riedel D, Becker S, et al. Structural heterogeneity of α -synuclein fibrils amplified from patient brain extracts. *Nat Commun*. 2019;10:5535.
- Tam C, Iwasaki W. AlphaCutter: efficient removal of non-globular regions from predicted protein structures. *Proteomics*. 2023;23:e2300176.
- Tao F, Tang H, Zhang S, Li M, Xu P. Enabling QTY server for designing water-soluble α -helical transmembrane proteins. *MBio*. 2022;13:e0360421.
- Tartaglia GG, Vendruscolo M. The Zyggregator method for predicting protein aggregation propensities. *Chem Soc Rev*. 2008;37:1395–401.
- Tsirigos KD, Peters C, Shu N, Käll L, Elofsson A. The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides. *Nucleic Acids Res*. 2015;43:W401–7.
- UniProt Consortium. UniProt: the universal protein knowledgebase in 2023. *Nucleic Acids Res*. 2023;51:D523–31.
- Varadi M, Bertoni D, Magana P, Paramval U, Pidruchna I, Radhakrishnan M, et al. AlphaFold protein structure database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Res*. 2024;52:D368–75.
- Xia X, Kumru OS, Blaber SI, Middaugh CR, Li L, Ornitz DM, et al. Engineering a cysteine-free form of human fibroblast growth Factor-1 for “second generation” therapeutic application. *J Pharm Sci*. 2016;105:1444–53.
- Zakeri F, Latifi-Navid H, Soheili Z-S, Sadeghi M, Arab SS, Samiei S, et al. Design, construction and in vivo functional assessment of a hinge truncated sFLT01. *Gene Ther*. 2023;30:347–61.
- Zambrano R, Jamroz M, Szczasiuk A, Pujols J, Kmiecik S, Ventura S. AGGRESCAN3D (A3D): server for prediction of aggregation properties of protein structures. *Nucleic Acids Res*. 2015;43:W306–13.
- Žerovnik E. Putative alternative functions of human stefin B (cystatin B): binding to amyloid-beta, membranes, and copper. *J Mol Recognit*. 2017;30:e2562.

How to cite this article: Zalewski M, Iglesias V, Bárcenas O, Ventura S, Kmiecik S. Aggrescan4D: A comprehensive tool for pH-dependent analysis and engineering of protein aggregation propensity. *Protein Science*. 2024;33(10):e5180. <https://doi.org/10.1002/pro.5180>