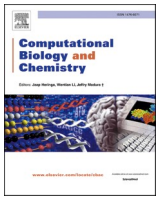
Computational Biology and Chemistry 113 (2024) 108209

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/cbac



An open source *in silico* workflow to assist in the design of fusion proteins

C.J. Lalaurie [a], C. Zhang [a], S.M. Liu [b], K.A. Bunting [b], P.A. Dalby [a,*]

[a] Department of Biochemical Engineering, University College London, London, United Kingdom
[b] IPSEN Bioinnovation, 5th Floor, The Point, 37 North Wharf Road, London W2 1AF, United Kingdom

ARTICLE INFO

Keywords:
Fusion protein
Molecular dynamics
Computational biology

ABSTRACT

Fusion proteins have the potential to become the new norm for targeted therapeutic treatments. Highly specific payload delivery can be achieved by combining custom targeting moieties, such as V_{HH} domains, with active parts of proteins that have a particular activity not naturally targeted to the intended cells. Conversely, novel drug products may make use of the highly specific targeting properties of naturally occurring proteins and combine them with custom payloads. When designing such a product, there is rarely a known structure for the final construct which makes it difficult to assess molecular behaviour that may ultimately impact therapeutic outcome. Considering the time and cost of expressing a construct, optimising the purification procedure, obtaining sufficient quantities for biophysical characterisation, and performing structural studies *in vitro*, there is an enormous benefit to conduct *in silico* studies ahead of wet lab work.

By following a repeatable, streamlined, and fast workflow of molecular dynamics assessment, it is possible to eliminate low-performing candidates from costly experimental work. There are, however, many aspects to consider when designing a novel fusion protein and it is crucial not to overlook some elements. In this work, we suggest a set of user-friendly, open-source methods which can be used to screen fusion protein candidates from the sequence alone. We used the light chain and translocation domain of botulinum toxin A (BoNT/A) fused with a selected V_{HH} domain, termed here LC-H_N-V_{HH}, as a case study for a general approach to designing, modelling, and simulating fusion proteins. Its behaviour *in silico* correlated well with initial *in vitro* work, with SEC HPLC showing multiple protein states in solution and a dynamic protein shifting between these states over time without loss of material.

1. Introduction

Despite the many functions proteins have in the body, such as enzymatic catalysis, cargo transportation, and signalling (Leader et al., 2008), their widespread use as therapeutic molecules is a relatively recent development in medicine. This was in part due to the emergence of recombinant protein expression in *E. coli* which eliminated the need to extract endogenous proteins from animal tissue. Although considerable progress has since been made in this field, there are only an order of hundreds of approved biologics in the market (Ebrahimi and Samanta, 2023) due to the challenges of purifying and storing these proteinaceous molecules without denaturation, aggregation, or degradation, which is, in part, influenced by various environmental factors such as temperature (Kellerman et al., 2022), pH (Perez and Groisman, 2007; Weeks and Sachs, 2001), and ionic strength (Möller et al., 2012). Key developments that have reduced the impact of these factors include site-directed mutations (Fryszkowska et al., 2022), PEGylation (Harris and Chess, 2003), and fusion to other proteins (Leader et al., 2008). The latter can, and has been, used to improve the half-life of a drug, such as through binding to human albumin (Nilsen et al., 2020; Andersen et al., 2014; Schelde et al., 2019), and for the creation of novel molecules with customised targeting moieties. For example, the anticancer drug, blinatumomab, is a fusion of two antibody fragments that binds CD3 T cells and CD19 on cancerous B cells, bringing them in proximity such that the T cells can secrete enzymes leading to cancer cell death (Labrijn et al., 2019; Einsele et al., 2020).

Some proteins have specific catalytic features that could be employed in cells other than their natural targets, such as the BoNT (Keith and John, 2010). This toxin is currently the most potent known (Pirazzini et al., 2017; Montal, 2010) with an estimated oral lethal dose of <1 μg/kg for humans (Cheng and Henderson, 2011; Arnon et al., 2001). It targets motor-neurons and blocks neurotransmission, resulting

* Correspondence to: University College London, Dept Biochemical Engineering, UCL, 6.07 Bernard Katz Building, Gordon Street, London WC1H 0AH, United Kingdom.
E-mail address: p.dalby@ucl.ac.uk (P.A. Dalby).

https://doi.org/10.1016/j.compbiolchem.2024.108209
Received 29 May 2024; Received in revised form 2 September 2024; Accepted 6 September 2024

in flaccid paralysis; if left untreated, this can eventually lead to respiratory failure and death. This high potency and site-specificity makes it a desirable pharmaceutical molecule due to the low amount of material required for effective treatment, as well as the near absent diffusion from the injection location. BoNT is currently used in the treatment of conditions such as cervical dystonia, blepharospasm, and glabellar lines (Cooper, 2007; Masuyer et al., 2014; Fonfria et al., 2018; Erbguth, 2004). The active toxin is comprised of a light chain (LC) and heavy chain (HC), and is structurally modular with three spatially distinct domains, each responsible for an event in the mechanism of action. The LC possesses a catalytic domain that acts as a metalloprotease when released into the neuron. The N-terminal half of the heavy chain (H_N) possesses a belt structure that wraps around the LC as well as a translocation domain (TD) that is responsible for delivering the LC from the endosome into the cytoplasm after a pH trigger. The C-terminal half of the heavy chain (H_C) possesses a binding domain (BD) responsible for the site-specific targeting of motor neurons. By replacing the BD with a new targeting moiety, it is possible to redirect the catalytic activity elsewhere to treat various secretion-based conditions.

The LC-H_N of BoNT remains stable at the endosomal pH of ~5 (Lalaurie et al., 2022; Singh and DasGupta, 1989), making it a good candidate for a fusion protein scaffold likely to retain its translocation and catalytic properties. A new targeting domain may then be designed around the LC-H_N so as not to affect the translocation mechanism while maintaining the binding sites' availability. The goal of creating fusion proteins is typically to exploit and compound the function of each component in one molecule; however, there is no guarantee that it will be functionally active due to how the protein folds. Novel inter- and intra-molecular interactions may be added unintentionally to the construct; therefore, careful consideration is required for the design of an intended fusion protein therapeutic. Flexible linker domains are frequently used to allow each component to retain a degree of mobility to permit retention of the domain function. Indeed, multiple linkers may need to be tested to find an optimal design; however, this makes successful experimental structure determination more challenging.

A novel construct was designed in order to develop and assess a computational workflow aimed at screening fusion proteins from the sequence alone. This construct is a retargeted BoNT where the BD is replaced with a V_{HH} domain by attaching it to the LC-H_N with a flexible linker (LC-H_N-V_{HH}). Structural models of this construct were generated by protein structure prediction algorithms from the primary sequence, as well as with docking simulations with its intended target. Using the RosettaCM algorithm (Song et al., 2013) yielded ~17,000 models, of which four within the top ten were taken for further study. Similar work has previously been performed (Ahmadi Moghaddam et al., 2023) on a ricin-based fusion protein, with 12 constructs analysed to study different linkers & active sites; however this was limited to a single repeat & a single model per construct. Here, we have used 4 models of the same construct, and repeated each three times in three pH conditions for a total of 9 simulations per model. Other computational workflows have been suggested for highly specific purposes (Gonzalez et al., 2024; Divine et al., 2021) which cannot be applied to the general study of the stability of fusion-protein based drugs.

Molecular dynamic (MD) simulations for each of these models revealed that the starting position had a significant impact on the outcomes, with some models exploring large areas of conformational space while others remained extremely stable throughout. Analysis of the trajectories also showcased the paramount importance of a detailed review of the output, with a clustering analysis identifying structures which would hinder the desired activity of the molecule exclusively in certain pH conditions. The dynamic and flexible nature of the molecule was confirmed by experimental work with HPLC SEC data which showed a fluid exchange of material from one elution peak to another without overall loss of protein content. This study highlights the importance of performing extensive *in silico* studies when designing a novel fusion protein-based product as a way to narrow down strong constructs and identify a likely optimal formulation, ahead of expensive and time-consuming practical work.

2. Methods

2.1. Homology modelling & docking

The LC-H_N-V_{HH} construct sequence was used as the target in the RosettaCM software (Song et al., 2013), and the protein databank (PDB) models 3BTA & 3EAK (Lacy and Stevens, 1998; Vincke et al., 2009) were used as the target structures for the LC-H_N and the V_{HH} domains, respectively. Docking between the V_{HH} domain and its target was simulated using ClusPro (Desta et al., 2020; Vajda et al., 2017; Kozakov et al., 2017, 2013). Finally, a model of the full molecule was built using RosettaCM by using the results of the docking simulations.

2.2. Molecular dynamics

MD simulations were performed using Gromacs 2019.3 (Van Der Spoel et al., 2005) on a high-performance computing cluster (6 nodes with two 20-core Intel Xeon Gold 6248 2.5 GHz, with hyperthreading; 192 Gb of 2933 MHz DDR4 RAM; Intel OmniPath network). The simulated temperature was set to 340 K to accelerate the process owing to increased potential energy levels, and the pressure kept at 1 bar. The protein was set in a cubic box with periodic boundary conditions and with explicit solvent. pH conditions were simulated on the starting structures using the APBS PDB2PQR server (Dolinsky et al., 2004; Jurrus et al., 2018). The ionic strength of the simulations was set to 165.7 mM to match the experimental buffers used. Three pH conditions (pH 4, 5.5, and 7.2) for four starting positions (A1, A2, A3, and A4) were run in triplicate for 300 ns each, for a total of 36 simulations.

Root mean square deviation (*RMSD*) and radius of gyration (R_G) data was collected using VMD 1.9.3 (Humphrey et al., 1996). Principal component analysis (PCA) and clustering data was obtained using the BIO3D package in R (Grant et al., 2006), using the "average" clustering algorithm. Root mean square fluctuation (*RMSF*) data was calculated using the Gromacs command "gmx rmsf". Path similarity analysis was achieved using the mdanalysis python package (Seyler et al., 2015; Gowers et al., 2016).

2.3. Protein purification and HPLC SEC

His-tagged LC-H_N-V_{HH} was purified from over-expression in *E. coli* cells using nickel affinity chromatography. HPLC SEC data was obtained using an Agilent 1260 and an Acquity UPLC Protein BEH200 SEC, with a flow rate of 0.2 mL/min in a mobile phase of 50 mM sodium phosphate, 200 mM NaCl, 200 mM L-Arginine, pH 7.0. Sample was buffer exchanged into either 50 mM sodium acetate for pH 4 and pH 5.5, or 50 mM sodium phosphate for pH 7.2; all at 165.7 mM ionic strength controlled by NaCl. The sample was maintained at 15 °C in the equipment at a concentration of 0.5 mg/mL and 20 μL fractions were taken every 2 hours.

3. Results

3.1. Homology modelling and docking

With no previously determined structure for the LC-H_N-V_{HH} construct, the starting models for the molecular dynamic simulations were generated using homology modelling from the amino acid sequence. The AlphaFold2 algorithm (Jumper et al., 2021) generated a model with a misrepresentation of the BoNT belt (Fig. 1A), whereas the RoseTTAFold software (Baek et al., 2021) did not correctly model the V_{HH} domain, predicting many loops instead of beta-sheets (Fig. 1B). The RosettaCM protocol (Song et al., 2013), however, was able to model both the V_{HH} domain and the belt correctly with its more elaborate

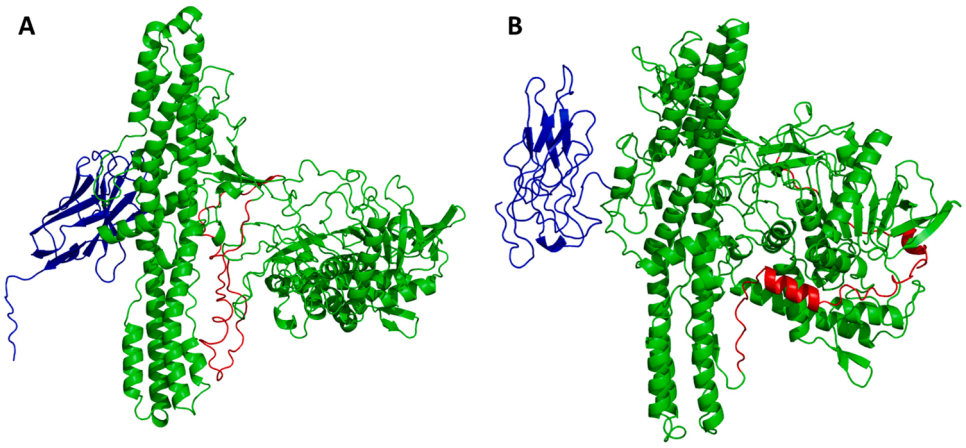


Fig. 1. A: Structure predicted by AlphaFold2. The sequence was submitted to Google Colab webserver. **B:** Structure predicted by RosettaFold. The sequence was submitted to the RosettaFold webserver. The LC and TD are coloured green, the belt region red, and the V_{HH} blue. AlphaFold2 was able to correctly predict the V_{HH} domain but not the belt conformation, whereas RosettaFold was able to predict the belt conformation but not the V_{HH}.domain.

process involving three steps: *thread*, *hybridize*, *relax*. For the first step, *thread*, the target residues (LC-H_N and V_{HH}, separately) are mapped to the template structure (PDB ID: 3BTA for the LC-H_N, PDB ID: 3EAK for the V_{HH}) (Lacy and Stevens, 1998; Vincke et al., 2009). For the second step, *hybridize*, 17,000 structures were generated which propose a final structure by combining the two distinct models (LC-H_N and V_{HH}) obtained from the *thread* step. In the final step, *relax*, each model was allowed to settle to a local potential energy minimum state. This showed that while the models retained the expected structural conformation of the two parts, the relative position of the V_{HH} and the LC-H_N varied greatly.

Docking simulations were performed between the V_{HH} component of the models from the RosettaCM protocol and its intended target using the ClusPro software (Desta et al., 2020; Vajda et al., 2017; Kozakov et al., 2017, 2013) to identify those that were still able to bind to their target. ClusPro ranks the models based on the lowest free energy; and the highest ranked models of the V_{HH}-target complex were then combined into the full molecule by applying the RosettaCM protocol once more, this time using the LC-H_N and the bound V_{HH}-target complex. RosettaCM ranks the generated models by the *RMSD* values of the model measured relative to the template structures supplied in the *thread* step.

3.2. MD simulations reveal dynamic behaviour

From the top ten best ranked models generated with the docking simulation, four were selected at random and the bound substrate removed before use in MD simulations (Fig. 2). The aim of these simulations was to investigate the behaviour of the LC-H_N-V_{HH} in solution prior to any binding event. The simulations were performed in triplicate at three different pH conditions using the APBS PDB2PQR server

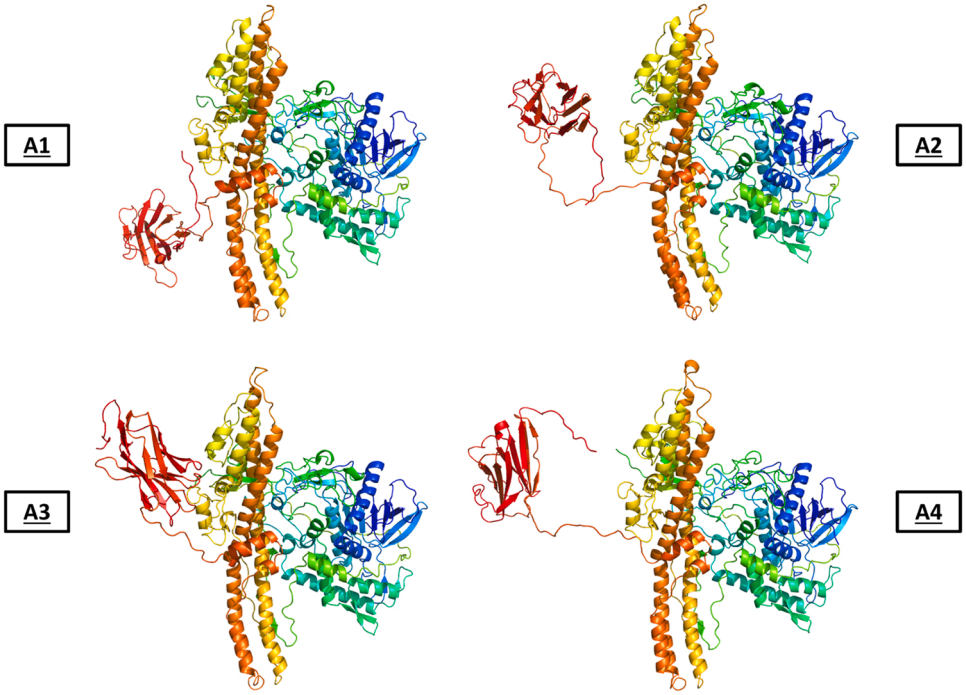


Fig. 2. Four starting models of LC-H_N-V_{HH} (A1 to A4) were selected from the 10 highest ranking models after docking simulations to intended target. This portrays the broad array of suggested placements of the V_{HH} domain relative to rest of the protein after homology modelling & docking steps. Molecules are coloured with the "chainbow" format.

(Dolinsky et al., 2004; Jurrus et al., 2018), and trajectories were analysed after 300 ns at 340 K and 1 bar pressure, to provide enough time for the model to equilibrate (Skjærven et al., 2011). The *RMSD* values of all the frames relative to their respective starting structure were between 0.5 – 3 nm for all models except A3, which remained between 0.25 – 2 nm (Fig. 3A). Model A3, therefore, appeared to be the most stable, with fewer frames having *RMSD* values greater than 1 nm (~30 % of total frames) compared to the other models (~75 % of total frames in A1 and A2; ~95 % of total frames in A4). These values imply a highly mobile structure, though they do not give insight into the nature of the changes to the molecule. When separating the distribution as a function of pH, higher *RMSD* values were observed with decreasing pH for all models except A3 (Figure S1). This indicates the construct has a preference for a neutral pH environment, leading to a more stable structure relative to the lower pH solutions.

The R_G of a protein is a measure of its compactness and can reveal conformational changes. The R_G distribution (Fig. 3B) demonstrated a broad range of values explored (3.2 nm – 3.7 nm; a 15 % increase) but with a complete overlap of data from all models, indicating that the full space available to the protein has likely been explored with these simulations. Intra-domain *RMSD* analyses showed little variation for LC-H_N (Figure S2) and V_{HH} (Fig. 4) domains for all four models, with values no greater than 0.5 and 1.2 nm, respectively, indicating high conformational stability for each domain. This suggests that the larger variation of values seen in the full protein *RMSD* data is due to a displacement of the V_{HH} domain relative to the LC-H_N complex. All together, these results indicate that this construct has a highly mobile V_{HH} domain with a broad range of LC-H_N-V_{HH} conformations explored from all starting positions except A3, which remained stable throughout. Whole protein *RMSD* and R_G values, however, can sometimes represent highly different structures, especially when studying a large molecule (>100 kDa); therefore, it is not sufficient to rely on *RMSD* or R_G analyses alone.

RMSF is a measure of the displacement of residues relative to their average position over the time period analysed, which gives an indication of the stable regions within a molecule. The *RMSF* per residue for each of the models similarly pointed to A3 as the most stable structure, with the lowest *RMSF* for all pH environments (Figure S3). Combined with the *RMSD* analysis, this highlights the impact of the starting model on the outcome of the simulations. In model A3, the protein appears to be locked in a stable position with the lowest *RMSD* and *RMSF* values of all models. This would bias the interpretation of the results if this model alone had been taken forward for analysis. By selecting multiple models from the homology modelling step, model A3 was identified as an exception, with a narrower spread of *RMSD*, and lower *RMSF* values relative to the other three models.

A look at the path similarity between repeats of the same model and between the different models was used as a measure of the accuracy and repeatability of MD data (Figure S4). This revealed that the trajectories of the same model and pH were always the most similar to each other, even relative to the same model at different pHs, thus confirming the repeatability of the work. Furthermore, this also revealed high similarity between repeats for models A2 and A3, highlighting the necessity to use multiple models for a full assessment of a fusion protein's behaviour. That is, had models A2 and A3 been the only selected models, only a small fraction of the available space would have been explored, leading to a biased assessment of the construct based on a minority of models.

3.3. PCA and clustering confirm consensus structure

Principal component analysis (PCA) is a tool which identifies the major axes of movement and the portions of the molecule which account for the majority of the variance. This is particularly useful for studying proteins with multiple domains because it can identify which parts of the protein are responsible for the full-protein *RMSD* values. All the data for each pH and model were combined into a single trajectory file, making it possible to compare the structures explored as a function of the starting model and pH. The full data set was grouped into six clusters of similar conformation (see elbow plot, Figure S5) using the first four principal components to account for ~75 % of the total variance (Figure S6). The average *RMSD* and standard deviation of each frame within a cluster to that cluster's mid-point frame was low, confirming the close structural similarity of all frames within each cluster (Figure S7). The distribution

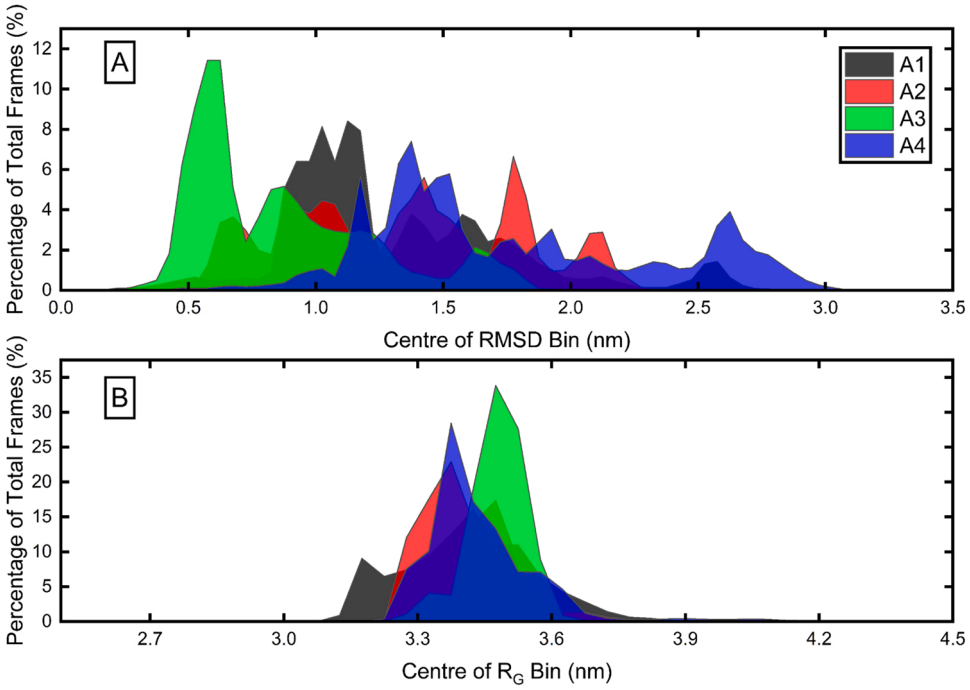


Fig. 3. A: Whole protein *RMSD* distribution after fixing the LC-H_N domains, in bins of 0.05 nm width, all pH combined, with respect to its starting position for each model. B: R_G distribution in bins of 0.05 nm width, all pH combined. This shows the large overlap of values from all starting positions, with A3 maintaining a lower *RMSD* than other starting models.

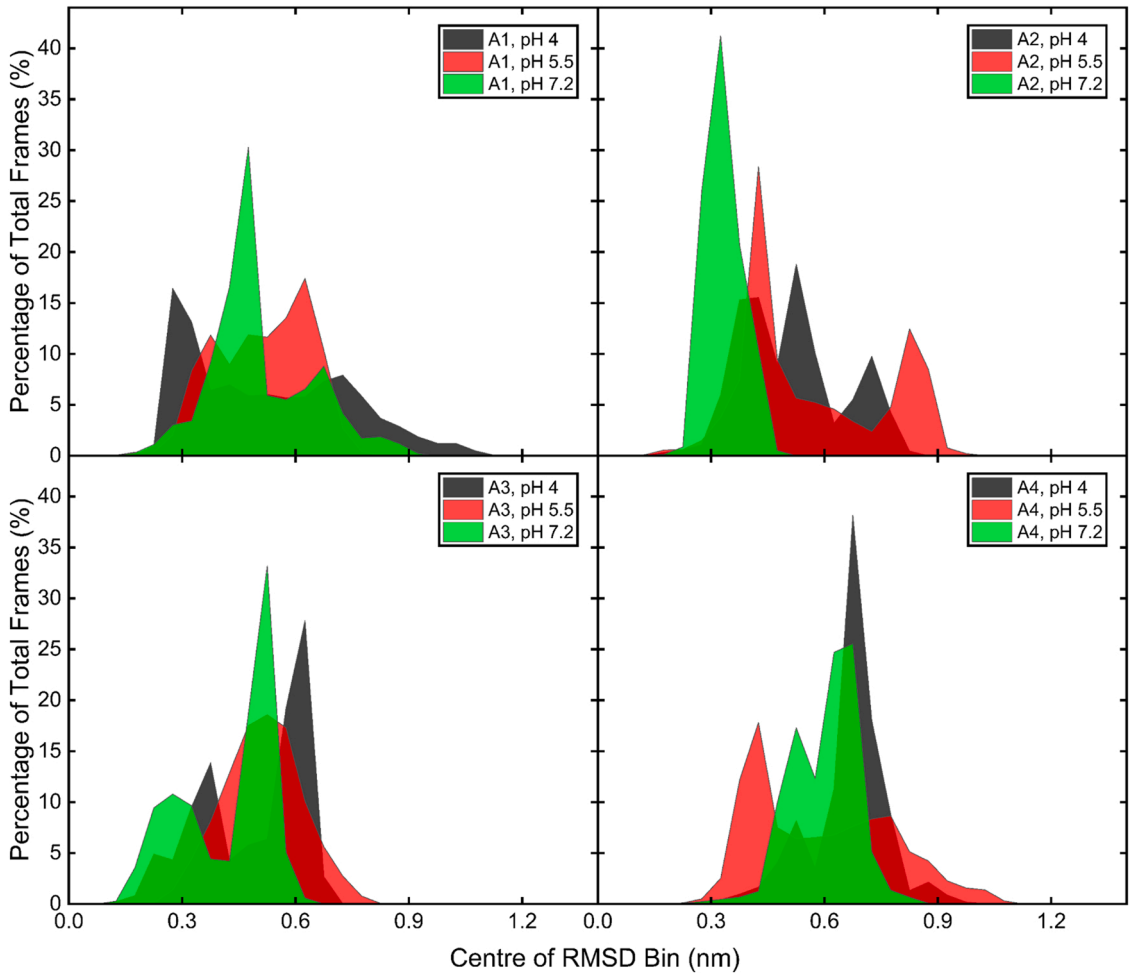


Fig. 4. Intra-domain V_{HH} *RMSD* distribution for each model and pH. The values never exceed 1.2 nm, confirming the structural integrity of the V_{HH} domain during the simulations.

of the clusters as a function of the initial model (Fig. 5) suggested high structural mobility for A1 with six clusters explored, while A3 explored just two clusters. Furthermore, the dominant cluster explored by A3 (80 % of all its frames) was very close to its starting position (Figs. 6 and 2). Interestingly, two clusters (2 and 6) were explored in all simulations despite the large disparity in the initial starting position of the V_{HH} domain. This suggests the existence of a consensus cluster (cluster 2; 46 % of all frames) that the protein explores irrespective of its initial model, making this cluster a likely representation of the dominant population in solution. This further highlights the importance of using multiple initial models to ensure most of the conformational space available to the protein has been explored. As portrayed by model A3, in some cases the protein may adopt a stable conformation which is irrelevant with respect to its viability as a potential new drug-product. For example, cluster 6 shows the active sites of the V_{HH} domain are near the TD (Fig. 6) which may result in poor binding. If model A3 had been the only model taken forward for *in silico* analysis, this construct would likely have been rejected and a separate construct implemented with e.g. a new linker, or by displacing the linkage site. However, by including multiple other models, it was possible to identify a different conformation (cluster 2), with increased availability of the binding sites, as the overall dominant and more likely solution-structure for this construct.

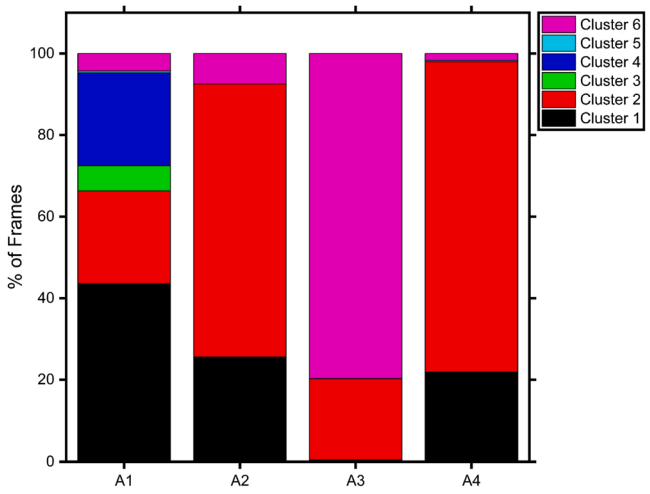


Fig. 5. Clustering of the frames, represented as a percentage of the total number of frames for each starting model. Reveals two clusters which exist in all models (2 and 6), with cluster 2 dominating the overall distribution and cluster 6 dominating in model A3.

3.4. SEC detects multiple populations in solution

The LC-H_N-V_{HH} construct was expressed, purified, and analysed by SEC HPLC in order to validate the MD data and cluster distribution. Purified protein was buffer-exchanged into each of the simulated conditions of pH and ionic strength, and loaded into the sample chamber maintained at 15 °C to avoid aggregation and evaporation. Fractions were taken every two hours and loaded onto the column for 24 hours.

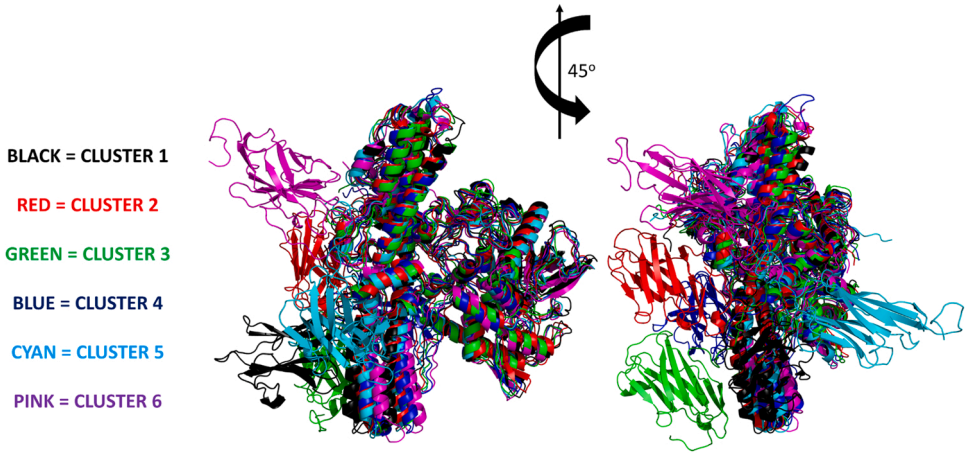


Fig. 6. Cartoon representation of the cluster mid-points, showing the high mobility of the V_{HH} domain. Cluster 6 is very close to the starting position of model A3, which is consistent with the lower *RMSD* values explored by this model and is in accordance with this cluster dominating the model's simulations.

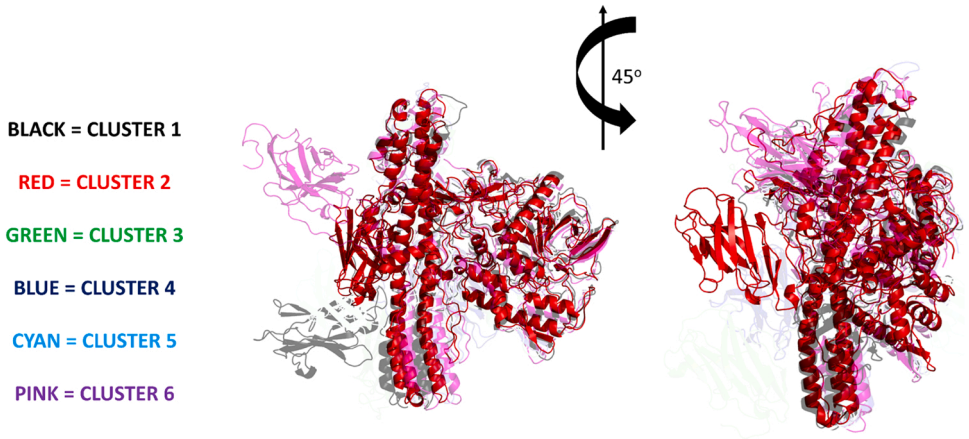


Fig. 7. View of the cluster mid-points with transparency setting inversely proportionally set to the percentage of total frames, with cluster 2 set to 0 (i.e., transparency set to 0.5 / 0 / 0.96 / 0.87 / 0.98 / 0.5, respectively for cluster 1 / 2 / 3 / 4 / 5 / 6).

The data (Fig. 8) revealed a dominant first peak followed by a smaller second peak. The latter appeared to increase in intensity and elute earlier until the 16 h mark, before gradually returning to its starting state by the 24 h mark. The sum of the two peak areas did not vary much throughout the time course – any decrease in area to the first peak was matched with a corresponding increase to the second peak; and vice-versa. This suggests a fluid exchange of material from one peak to the other, which would be consistent with the protein fluctuating between different conformations as was observed in the cluster analysis. Distribution of the clusters as a function of pH (Fig. 9) showed pH 7.2 was noticeably dominated by cluster 2; while pH 4 and pH 5.5 possessed a more even proportion of clusters 1, 3 and 6. Considering the average R_G of each cluster (Fig. 10), the average R_G in pH 4 (3.45 nm) and pH 5.5 (3.43 nm) is slightly higher than in pH 7.2 (3.42 nm), which is in good agreement with the increased proportion of the earlier elution peak at lower pH values. The presence of aggregates or fragment species has been ruled out as this would have led to significantly earlier, or later, peaks in the elution profile.

4. Conclusion

One of the first outputs of protein engineering is the generation of a novel polypeptide sequence that folds correctly into a protein with specific desirable characteristics. For our chimeric LC-H_N-V_{HH} construct, protein structure prediction algorithms and homology modelling yielded multiple potential models, each with distinct placement of the V_{HH} domain relative to the rest of the construct. While it was not feasible to perform an in-depth study of all the models generated, this study has shown that it is crucial to select a number of these at random within the highest ranked models to maximise exploration of conformational space. This is to ensure that a consensus conformation has been identified as the most likely representation of the molecule's true solution behaviour. As has been demonstrated here with model A3, some structures may be locked into a stable, low energy state with significantly reduced movement and may be a misrepresentation of the *in vivo* behaviour. However, if all models were to reach such a stable, low energy state, then it could be inferred that this conformation would likely be representative of the true structure. Following careful evaluation against the desired mechanism of action, that molecular sequence may then be down-prioritized for experimental assessment.

Utilising this assessment as the first stage in a screening funnel allows resource-intensive *in vitro* and *in vivo* assessment of molecule function and therapeutic impact to be focused on molecules with a greater chance of success. Furthermore, if such a position is only observed in particular environmental conditions, and is favourable to the activity of the protein, it is important to detect this so that the conditions can be replicated experimentally, with a view to increase confidence in assay readouts. Both situations warrant an extensive study such that the full behaviour of the protein is observed, and positive or negative outcomes can be isolated to the solution conditions in which they were detected.

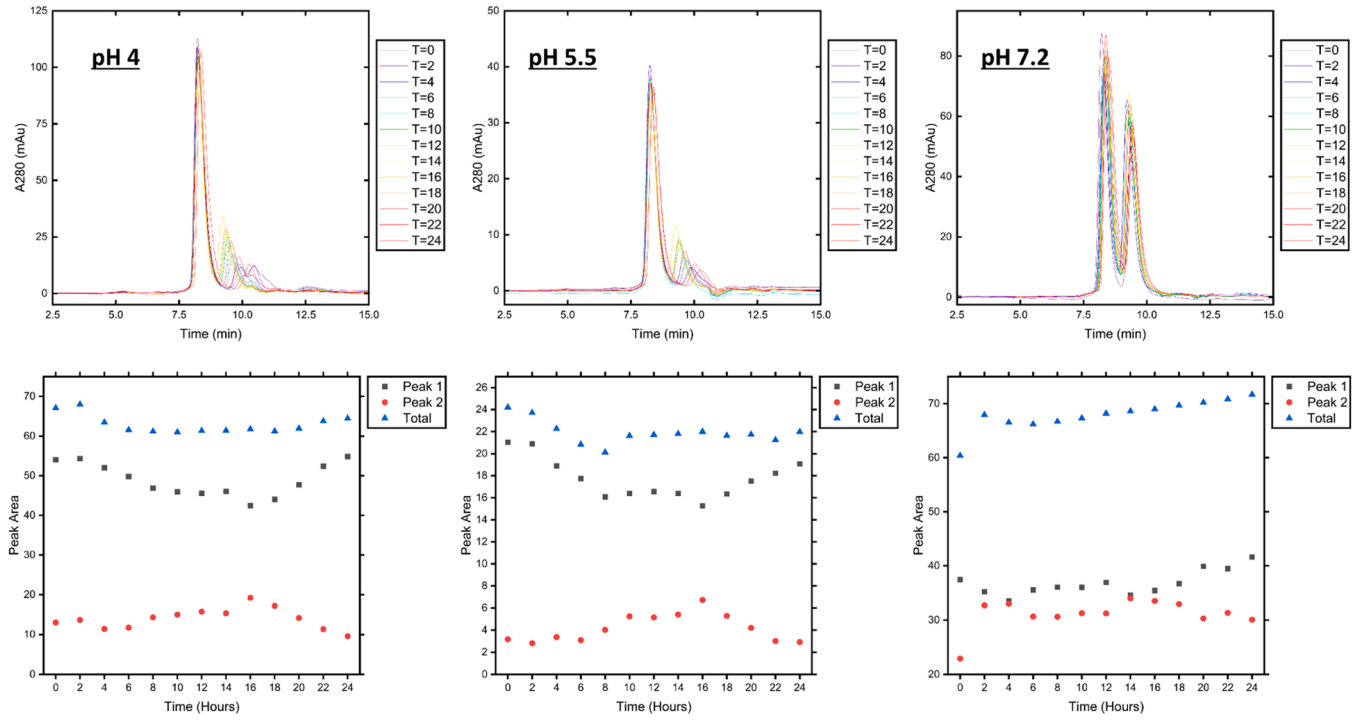


Fig. 8. HPLC SEC traces of LC-H_N-V_{HH} in the same solution conditions as used for the simulated pHs and ionic strengths (*top*). Integration of both peaks shows an inverse relationship of change in peak area over time, indicating a dynamic exchange of material from one elution peak to another, with little to no material loss (*bottom*).

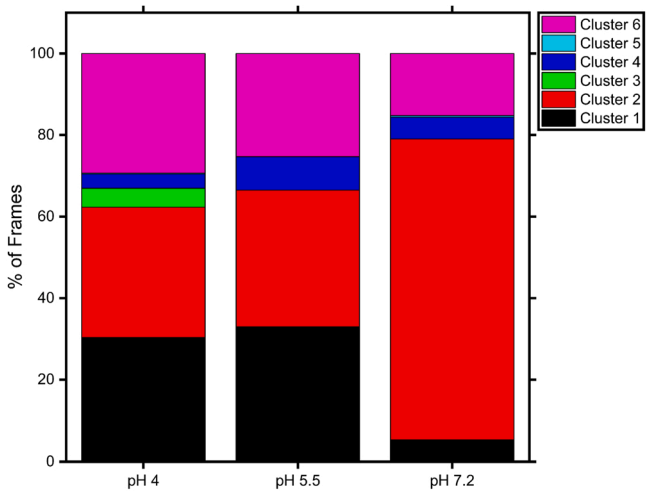


Fig. 9. Cluster distribution as a function of pH shows most frames in pH 7.2 are in cluster 2; while pH 4 and 5.5 have an almost equal proportion of frames in clusters 1, 2 and 6.

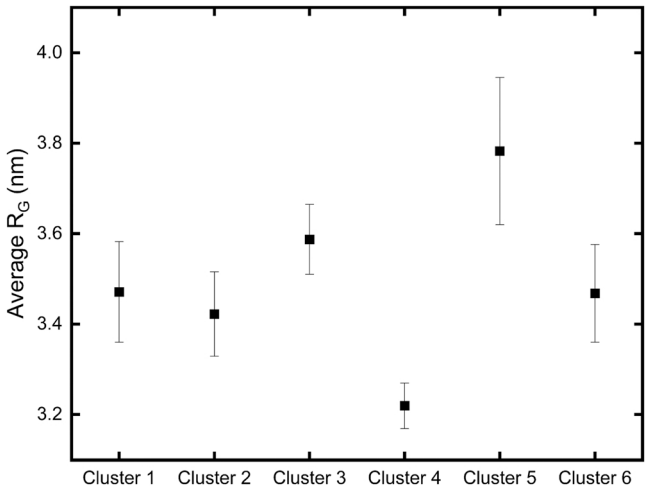


Fig. 10. Average cluster R_G, with standard deviation, showing the disparity between the most extreme clusters. This may explain the observed difference in SEC peak distribution, with pH 7.2 strongly favouring clusters with slightly lower R_G.

Furthermore, we have demonstrated the importance of analysing multiple aspects of the MD results. *RMSD* or R_G data alone would have overlooked the large array of structures hidden in similar values and resulted in false conclusions. However, when combined with a clustering analysis, we were able to resolve the large array of conformations available to the protein. From this, we identified a consensus cluster (clustering by initial model; cluster 2) while simultaneously revealing preferential conformations in particular pH environments (clustering by pH).

This study has confirmed the increasing reliability and utility of *in silico* models, by providing a detailed insight into the behaviour of our LC-H_N-V_{HH} construct using MD simulations and further corroboration with experimental work. When designing a fusion protein, there are many aspects to consider: choice of flexible linker, choice of targeting moiety, choice of payload, and impact of buffer conditions such as pH. The last point is particularly crucial for molecules whose mechanism of action will involve transition through different pH environments. All of these can be screened ahead of experimental work through MD simulations, the results of which can be used to screen out potential candidates and optimise the formulation of promising ones. In the case of the LC-H_N-V_{HH} construct, we can conclude that it would likely adopt a conformation with stable individual domains that are functionally active, and a preference for pH 7.2. Once a strong candidate has been identified, it can be expressed and taken for further characterisation

such as analytical ultra-centrifugation to monitor its aggregation properties, circular dichroism to monitor secondary structure stability, and small-angle X-ray scattering, or X-ray crystallography for a higher resolution look at its three-dimensional structure. While this workflow gives a strong indication as to the conformational stability of the constructs studied, it cannot predict how the protein will interact with other proteins. This would need to be tested experimentally on the most conformationally stable candidates in order to find the best overall construct, that being the best combination of conformational and colloidal stability. Thanks to the recent advancements in technologies for modelling proteins and generating predictions from the amino acid sequence, this work can be repeated with any number of fusion drug candidates. The nature of the computational analysis is fully transferable to any model, owing to the general nature of the *RMSD* & R_G measurements. PCA uses spatial coordinates to detect conformational similarity and can therefore equally be applied to any constructs. The scripts used here can be repurposed with residue selections for alignments relevant to the work at hand. With the ever-increasing computing power available, *in silico* analyses can give a reliable indication of how a protein behaves in solution, ultimately reducing the cost of future biologics by speeding up the design and screening process of suitable candidates.

CRediT authorship contribution statement

Christophe J. Lalaurie: Writing – original draft, Investigation, Software, Data curation, Visualization. Cheng Zhang: Investigation, Software, Methodology. Sai M. Liu: Conceptualization, Writing – review & Editing, Resources, Supervision. Karen A. Bunting: Conceptualization, Writing – review & editing, Methodology, Supervision, Funding acquisition. Paul A. Dalby: Writing – review & editing, Supervision, Funding acquisition, Project administration.

Declaration of Competing Interest

The authors declare that they have no competing interests to declare.

This research was sponsored by Ipsen Bioinnovation.

S.M. Liu was an employee of Ipsen at the time of the work & K.A. Bunting is an employee of Ipsen.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.compbiolchem.2024.108209.

References

Ahmadi Moghaddam, Y., Maroufi, A., Zareei, S., Irani, M., 2023. Computational design of fusion proteins against ErbB2-amplified tumors inspired by ricin toxin. Front Mol. Biosci. 10.
Andersen, J.T., et al., 2014. Extending serum half-life of albumin by engineering neonatal Fc receptor (FcRn) binding. J. Biol. Chem. 289.
Arnon, S.S., et al., 2001. Botulinum toxin as a biological weapon: medical and public health management. JAMA 285.
Baek, M., et al., 2021. Accurate prediction of protein structures and interactions using a three-track neural network. Science (1979) 373.
Cheng, L.W., Henderson, T.D., 2011. Comparison of oral toxicological properties of botulinum neurotoxin serotypes A and B. Toxicon 58.
Cooper, G., 2007. Therapeutic uses of botulinum toxin. Ther. Uses Botulinum Toxin 5, 1–238.
Desta, I.T., Porter, K.A., Xia, B., Kozakov, D., Vajda, S., 2020. Performance and its limits in rigid body protein-protein docking. Structure 28.
Divine, R., et al., 2021. Designed proteins assemble antibodies into modular nanocages. Science (1979) 372.
Dolinsky, T.J., Nielsen, J.E., McCammon, J.A., Baker, N.A., 2004. PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. Nucleic Acids Res. https://doi.org/10.1093/nar/gkh381.
Ebrahimi, S.B., Samanta, D., 2023. Engineering protein-based therapeutics through structural and chemical design. Nat. Commun. [Preprint].
Einsele, H., et al., 2020. The BiTE (bispecific T-cell engager) platform: Development and future potential of a targeted immuno-oncology therapy across tumor types. Cancer [Preprint].
Erbguth, F.J., 2004. Historical notes on botulism, Clostridium botulinum, botulinum toxin, and the idea of the therapeutic use of the toxin. Mov. Disord. 19.
Fonfria, E., et al., 2018. The expanding therapeutic utility of botulinum neurotoxins. Toxins (Basel) 10, 1–27.
Fryszkowska, A., et al., 2022. A chemoenzymatic strategy for site-selective functionalization of native peptides and proteins. Science (1979) 376.
Gonzalez, K.J., et al., 2024. A general computational design strategy for stabilizing viral class I fusion proteins. Nat. Commun. 15.
Gowers, R., et al., 2016. MDAnalysis: a python package for the rapid analysis of molecular dynamics simulations. Proc. 15th Python Sci. Conf.
Grant, B.J., Rodrigues, A.P.C., ElSawy, K.M., McCammon, J.A., Caves, L.S.D., 2006. Bio3d: An R package for the comparative analysis of protein structures. Bioinformatics 22, 2695–2696.
Harris, J.Milton, Chess, R.B., 2003. Effect of pegylation on pharmaceuticals. Nat. Rev. Drug Discov. [Preprint].
Humphrey, W., Dalke, A., Schulten, K., 1996. VMD: Visual molecular dynamics. J. Mol. Graph. https://doi.org/10.1016/0263-7855(96)00018-5.
Jumper, J., et al., 2021. Highly accurate protein structure prediction with AlphaFold. Nature 596.
Jurrus, E., et al., 2018. Improvements to the APBS biomolecular solvation software suite. Protein Sci. 27.
Keith, F., John, C., 2010. Targeted secretion inhibitors-innovative protein therapeutics. Toxins (Basel) [Preprint].
Kellerman, M.A.W., Almeida, T., Rudd, T.R., Matejtschuk, P., Dalby, P.A., 2022. NMR reveals functionally relevant thermally induced structural changes within the native ensemble of G-CSF. Mol. Pharm. 19.
Kozakov, D., et al., 2017. The ClusPro web server for protein-protein docking. Nat. Protoc. 12.
Kozakov, D., et al., 2013. How good is automated protein docking? Protein.: Struct. Funct. Bioinforma. 81.
Labrijn, A.F., Janmaat, M.L., Reichert, J.M., Parren, P.W.H.I., 2019. Bispecific antibodies: a mechanistic review of the pipeline. Nat. Rev. Drug Discov. [Preprint].
Lacy, B., Stevens, R.C., 1998. Crystal structure of Bont A and implications for toxicity. Nat. Struct. Biol. 5, 898–902.
Lalaurie, C.J., et al., 2022. Elucidation of critical pH-dependent structural changes in Botulinum Neurotoxin E. J. Struct. Biol. 214.
Leader, B., Baca, Q.J., Golan, D.E., 2008. Protein therapeutics: a summary and pharmacological classification. Nat. Rev. Drug Discov. [Preprint].
Masuyer, G., Chaddock, J.A., Foster, K.A., Acharya, K.R., 2014. Engineered botulinum neurotoxins as new therapeutics. Annu. Rev. Pharm. Toxicol. https://doi.org/10.1146/annurev-pharmtox-011613-135935.
Möller, J., et al., 2012. The effect of ionic strength, temperature, and pressure on the interaction potential of dense protein solutions: from nonlinear pressure response to protein crystallization. Biophys. J. 102.
Montal, M., 2010. Botulinum neurotoxin: a marvel of protein design. Annu. Rev. Biochem. 79, 591–617.
Nilsen, J., et al., 2020. An intact C-terminal end of albumin is required for its long half-life in humans. Commun. Biol. 3.
Perez, J.C., Groisman, E.A., 2007. Acid pH activation of the PmrA/PmrB two-component regulatory system of Salmonella enterica. Mol. Microbiol. https://doi.org/10.1111/j.1365-2958.2006.05512.x.
Pirazzini, M., Rossetto, O., Eleopra, R., Montecucco, C., 2017. Botulinum neurotoxins: biology, pharmacology, and toxicology. Pharm. Rev. 69, 200–235.
Schelde, K.K., et al., 2019. A new class of recombinant human albumin with multiple surface thiols exhibits stable conjugation and enhanced FcRn binding and blood circulation. J. Biol. Chem. 294.
Seyler, S.L., Kumar, A., Thorpe, M.F., Beckstein, O., 2015. Path similarity analysis: a method for quantifying macromolecular pathways. PLoS Comput. Biol. 11.
Singh, B., DasGupta, B.R., 1989. Molecular topography and secondary structure comparisons of botulinum neurotoxin types A, B and E. Mol. Cell Biochem. 86, 87–95.
Skjærven, L., Reuter, N., Martinez, A., 2011. Dynamics, flexibility and ligand-induced conformational changes in biological macromolecules: a computational approach. Future Med Chem. [Preprint].
Song, Y., et al., 2013. High-resolution comparative modeling with RosettaCM. Structure 21.
Van Der Spoel, D., et al., 2005. GROMACS: fast, flexible, and free. J. Comput. Chem. [Preprint].
Vajda, S., et al., 2017. New additions to the ClusPro server motivated by CAPRI. Protein.: Struct. Funct. Bioinforma. 85.
Vincke, C., et al., 2009. General strategy to humanize a camelid single-domain antibody and identification of a universal humanized nanobody scaffold. J. Biol. Chem. 284.
Weeks, D.L., Sachs, G., 2001. Sites of PH regulation of the urea channel of Helicobacter pylori. Mol. Microbiol. https://doi.org/10.1046/j.1365-2958.2001.02466.x.