Structural Studies of pH Effects on Botulinum Toxins A & E

by

Christophe Jean Lalaurie

A thesis presented for the degree of

Doctor of Philosophy

in the

Research Department of Biochemical Engineering, University College London

September 2022
Declaration

I, Christophe Jean Lalaurie confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.
Abstract

Botulinum neurotoxins (BoNTs) are responsible for botulism, a paralytic disease which can be lethal if not treated in time. They act by entering neurons and targeting the SNARE proteins (Soluble N-ethylmaleimide-sensitive factor Attachment protein REceptor), which in turn blocks neurotransmission. However, these toxins can be repurposed for therapeutic use to treat a large number of conditions. The most studied serotypes are A and E (BoNT/A and BoNT/E, respectively), with notable differences in duration of action and domain spatial organisation. It has been shown that these toxins only exert their activity if the pH drops to 5 or lower, but it is unclear what effect the pH environment has on the toxin which drives this. Currently, the only available structural information on BoNTs is from X-ray crystallography which fixes the protein into a rigid crystal lattice. This gives limited information on its flexible regions, and no information about its dynamics and solution behaviour. To gain insight into this, molecular dynamic (MD) simulations were conducted under varying pH conditions. For BoNT/E, these simulations revealed a shift in conformational populations in solvated systems at pH ≤ 5 when compared to simulations at pH > 5, with the protein adopting a more extended conformation in the former. This was confirmed by analytical ultracentrifugation (AUC), while small-angle X-ray scattering (SAXS) validated the two major conformations observed in the MD simulations. For BoNT/A, a major conformational change was not observed, but a rare event was identified by MD (in 0.014% of frames studied) which may explain the longer onset of action compared to BoNT/E. Another key difference between the two structures of BoNT/E and BoNT/A is the large number of contacts between a conserved region termed the “switch” and the binding domain (BD) in BoNT/A, which are absent in BoNT/E.
Impact Statement

While the entry into the system and the activity of BoNTs are well understood, the exact role of the pH change they are subjected to in the vesicle is still unclear. The work presented in this PhD thesis provides insight into this characteristic, and essential, step of BoNT activity. By combining atom level resolution in MD simulations and full protein structure solution studies, we have identified some key aspects of the effects of pH likely linked to the translocation process which enables toxicity in vivo.

In BoNT/E, the protein undergoes a large-scale structural change which results in a higher SASA with key areas of the protein exposed. This is due to the electrostatic repulsion of the BD and the light chain (LC) which begins at pH 5. One of the regions which is left with higher solvent exposure is an alpha-helix also conserved in diphteria toxin, which also sees a large change in charge between pH 5.5 and pH 5. These changes occur at the same time as a smaller scale change on the translocation domain (TD), in which a lysine residue flips from buried to highly exposed at and below pH 5, in another key region which is highly conserved throughout all serotypes and in TeNT. This region in the BoNT/E simulations presented here was found to have high correlation, at pH 5 specifically, with two other regions of the protein: an unstructured segment in the TD and some residues in the BD involved in the original membrane binding event. By identifying these four key regions within the protein, and revealing their linked behaviour at pH 5 specifically, we may have identified new regions to target with antibodies as a treatment against botulism. We have also furthered our overall understanding of the full process of intoxication with BoNT/E, by exposing structural changes brought about exclusively at and below pH 5.

In BoNT/A, while a large-scale structural shift was not observed to the extent of BoNT/E, some effects due to pH were identified which may play a role in the translocation process. Starting with CD, a sudden loss of helicity was observed starting at pH 5 compared to all the conditions above pH 5, before stabilising at pH 4.5. This is in good agreement with a previous study having identified an alpha-helical region in the TD shifting to a highly exposed beta-hairpin structure. Exploring this region further in the MD simulations presented here, we identified a rare event (0.014% of frames studied, in 5 consecutive frames out of 35,001) where a large number of interactions between this region and the BD are lost compared to the other structures. This region, which is highly conserved throughout all serotypes, is significantly more buried in
BoNT/A compared to the same region in BoNT/E. The scale of the change identified in this region by the previous study, the increased interactions of this region in BoNT/A compared to BoNT/E, and the rarity of the events in which those interactions are lost may all contribute to the increased onset of action delay of BoNT/A compared to BoNT/E. In future constructs making use of BoNTs, if replacing the BD, a key feature to optimize may be the exposure of this region to ensure the fastest delivery of the payload.

Publications


Acknowledgements

In chronological order through my academic path, I would like to thank Dr. Annela Seddon of the University of Bristol for introducing the field of Biophysics as a module in the Physics undergraduate course. This optional unit motivated me to pursue further education in this area and ultimately led to where I am today. From the University of Bristol also, I would like to thank Dr. Steven Burston for running the MSc Biophysics course which provided me with strong background theoretical knowledge I would then go on to apply experimentally. Prof. Paul Race and Dr. Paul Curnow helped me hone my writing skills and experimental practice in this field through two research projects which I have since used extensively.

From UCL & IPSEN I would like to thank: Prof. Mire Zloh for his insight and assistance in the generation and analysis of the MD simulations presented in this thesis; Dr. Andrew Splevins and Dr. Teresa Barata for designing this project and giving me the opportunity to work on it, and for their assistance in identifying relevant literature to get me started; Dr. Karen Bunting and Dr. Daniel Higazi for taking over from Andrew and Teresa halfway through and running with it, and their assistance in getting my first article published; and finally my 3rd Paul supervisor, Prof. Paul Dalby for giving me this opportunity at a time of great uncertainty in my life and the constant
support throughout these 4 years during which I have learned many new skills and which I have enjoyed very much! My time at UCL was also bettered by the people of the Biochemical Engineering department who were most friendly & always willing to help. Special thanks go out to Dr Valentina Spiteri & Dr. Thomas McDonnell for their devoted assistance in SAXS and PCA analyses. This research was jointly funded by Ipsen, and the Engineering and Physical Sciences Research Council (EPSRC) Centre for Doctoral Training in Emergent Macromolecular Therapies, to whom I am also grateful and without which this research would not have been possible.

I could not have reached this point without meeting the right people at the right time, so deep thanks go out to F.S., H.L., D.M., T.P., C.F., W. & O. G., W.L., A.G. & F.P. for putting up with my antics over the years and hopefully many more to come! And last but not least J.B. & B.P. for their lifelong friendship.

Finally, a warm heartfelt thanks to my family (including the 4th & 5th Pauls in my life!) for supporting me financially for the better part of the last 10 years without which I could not have achieved this. But the biggest thanks go out to L.H., the best support I could hope for who has stuck by me for 11 years and kept me going during the hardest times, reminding me there is good in this world, and it’s worth fighting for.
UCL Research Paper Declaration Form: referencing the doctoral candidate’s own published work(s)

Please use this form to declare if parts of your thesis are already available in another format, e.g. if data, text, or figures:

- have been uploaded to a preprint server;
- are in submission to a peer-reviewed publication;
- have been published in a peer-reviewed publication, e.g. journal, textbook.

This form should be completed as many times as necessary. For instance, if you have seven thesis chapters, two of which containing material that has already been published, you would complete this form twice.

1. For a research manuscript that has already been published (if not yet published, please skip to section 2):

<table>
<thead>
<tr>
<th>a) Where was the work published? (e.g. journal name)</th>
<th>Journal of Structural Biology</th>
</tr>
</thead>
<tbody>
<tr>
<td>b) Who published the work? (e.g. Elsevier/Oxford University Press):</td>
<td>Elsevier</td>
</tr>
<tr>
<td>c) When was the work published?</td>
<td>20/06/2022</td>
</tr>
<tr>
<td>d) Was the work subject to academic peer review?</td>
<td>Yes</td>
</tr>
<tr>
<td>e) Have you retained the copyright for the work?</td>
<td>Yes</td>
</tr>
</tbody>
</table>

[If no, please seek permission from the relevant publisher and check the box next to the below statement]:

☐ I acknowledge permission of the publisher named under 1b to include in this thesis portions of the publication named as included in 1a.

2. For a research manuscript prepared for publication but that has not yet been published (if already published, please skip to section 3):

<table>
<thead>
<tr>
<th>a) Has the manuscript been uploaded to a preprint server? (e.g. medRxiv):</th>
<th>Please select.</th>
</tr>
</thead>
<tbody>
<tr>
<td>b) Where is the work intended to be published? (e.g. names of journals that you are planning to submit to)</td>
<td>Click or tap here to enter text.</td>
</tr>
</tbody>
</table>
c) List the manuscript's authors in the intended authorship order:

<table>
<thead>
<tr>
<th>Author(s)</th>
<th>Contributions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Christophe J. Lalaurie</td>
<td>Investigation, Data curation, Formal analysis, Writing – original draft, Editing.</td>
</tr>
<tr>
<td>Andrew Splevins</td>
<td>Conceptualization, Funding acquisition, Study coordination, Supervision.</td>
</tr>
<tr>
<td>Teresa S. Barata</td>
<td>Conceptualization, Funding acquisition, Study coordination, Supervision.</td>
</tr>
<tr>
<td>Karen A. Bunting</td>
<td>Conceptualization, Supervision, Validation, Writing – review &amp; editing.</td>
</tr>
<tr>
<td>Daniel R. Higazi</td>
<td>Conceptualization, Supervision, Validation, Writing – review &amp; editing.</td>
</tr>
<tr>
<td>Mire Zloh</td>
<td>Conceptualization, Supervision, Validation, Writing – review &amp; editing.</td>
</tr>
<tr>
<td>Valentina A. Spiteri</td>
<td>Resources, Data curation, Formal analysis for the SAXS data, Writing – review &amp; editing.</td>
</tr>
<tr>
<td>Stephen J. Perkins</td>
<td>Resources, Data curation, Formal analysis for the SAXS data, Writing – review &amp; editing.</td>
</tr>
<tr>
<td>Paul A. Dalby</td>
<td>Project administration, Conceptualization, Funding acquisition, Supervision, Validation, Writing – review &amp; editing.</td>
</tr>
</tbody>
</table>

3. For multi-authored work, please give a statement of contribution covering all authors (if single-author, please skip to section 4):

Christophe J. Lalaurie: Investigation, Data curation, Formal analysis, Writing – original draft, Editing. Andrew Splevins: Conceptualization, Funding acquisition, Study coordination, Supervision. Teresa S. Barata: Conceptualization, Funding acquisition, Study coordination, Supervision. Karen A. Bunting: Conceptualization, Supervision, Validation, Writing – review & editing. Daniel R. Higazi: Conceptualization, Supervision, Validation, Writing – review & editing. Mire Zloh: Conceptualization, Supervision, Validation, Writing – review & editing. Valentina A. Spiteri: Resources, Data curation, Formal analysis for the SAXS data, Writing – review & editing. Stephen J. Perkins: Resources, Data curation, Formal analysis for the SAXS data, Writing – review & editing. Paul A. Dalby: Project administration, Conceptualization, Funding acquisition, Supervision, Validation, Writing – review & editing.

4. In which chapter(s) of your thesis can this material be found?

Chapters 3 & 4

5. e-Signatures confirming that the information above is accurate (this form should be co-signed by the supervisor/ senior author unless this is not appropriate, e.g. if the paper was a single-author work):

<table>
<thead>
<tr>
<th>Candidate</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Christophe Lalaurie</td>
<td>02/09/2022</td>
</tr>
<tr>
<td>Supervisor/ Senior</td>
<td>21/09/2022</td>
</tr>
<tr>
<td>Author (where</td>
<td></td>
</tr>
<tr>
<td>appropriate):</td>
<td></td>
</tr>
<tr>
<td>Paul Dalby</td>
<td></td>
</tr>
</tbody>
</table>
# Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Declaration</td>
<td>2</td>
</tr>
<tr>
<td>Abstract</td>
<td>3</td>
</tr>
<tr>
<td>Impact Statement</td>
<td>4</td>
</tr>
<tr>
<td>Publications</td>
<td>5</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>5</td>
</tr>
<tr>
<td>Contents</td>
<td>9</td>
</tr>
<tr>
<td>List of Figures</td>
<td>14</td>
</tr>
<tr>
<td>List of Tables</td>
<td>16</td>
</tr>
<tr>
<td>List of Abbreviations</td>
<td>17</td>
</tr>
<tr>
<td>List of Units</td>
<td>19</td>
</tr>
<tr>
<td>CHAPTER 1</td>
<td>20</td>
</tr>
<tr>
<td>Introduction</td>
<td>20</td>
</tr>
<tr>
<td>1.1 Botulinum Neurotoxins (BoNTs)</td>
<td>21</td>
</tr>
<tr>
<td>1.1.1 Short history of BoNTs</td>
<td>21</td>
</tr>
<tr>
<td>1.1.2 BoNT structure</td>
<td>23</td>
</tr>
<tr>
<td>1.1.3 BoNT Activity</td>
<td>26</td>
</tr>
<tr>
<td>1.2 Protein structural studies</td>
<td>28</td>
</tr>
<tr>
<td>1.2.1 Solution vs fixed methods</td>
<td>28</td>
</tr>
<tr>
<td>1.2.2 Impact of pH environment</td>
<td>31</td>
</tr>
<tr>
<td>Chapter 2</td>
<td>33</td>
</tr>
<tr>
<td>Methods Theory</td>
<td>33</td>
</tr>
<tr>
<td>2.1 Protein expression, mutation and labelling</td>
<td>34</td>
</tr>
<tr>
<td>2.1.1 Protein expression</td>
<td>34</td>
</tr>
<tr>
<td>2.1.2 Protein mutation</td>
<td>35</td>
</tr>
<tr>
<td>2.1.3 Protein purification</td>
<td>35</td>
</tr>
<tr>
<td>2.1.4 Protein labelling</td>
<td>36</td>
</tr>
</tbody>
</table>
2.1.5 UAA incorporation

2.2 Computational methods

2.2.1 Molecular dynamic simulations

2.2.2 MD data analysis

2.2.3 Principal Component Analysis

2.2.4 Fitting of MD frames to experimental data

2.3 Solution structure methods

2.3.1 Circular dichroism (CD)

2.3.2 Analytical ultra-centrifugation (AUC)

2.3.3 Small-angle X-ray scattering (SAXS)

2.3.3 Single-molecule fluorescence resonance energy transfer (smFRET)

Chapter 3

Molecular dynamics simulations show pH mediated solution behaviour of BoNT/E

3.1 Summary

3.2 Introduction

3.3 Methods

3.3.1 Molecular dynamic simulation conditions

3.3.2 MD trajectories analysis

3.4 Results and discussion

3.4.1 Whole protein RMSD and $R_g$ analysis reveals a behaviour change in pH $\leq 5$ and pH $> 5$

3.4.2 Domain specific RMSD analysis shows that the binding domain separates from the light chain

3.4.3 Charge and salt-bridge content plays a crucial role in the activation of BoNT/E

3.4.4 SASA analysis shows low pH conditions increase BoNT/E’s solvent exposure
3.4.5 RMSF & cross-correlation analyses reveal a concerted action of multiple regions for membrane interaction at pH 5

3.4.6 PCA analysis uncovers a novel conformation exclusively observed at pH ≤ 5

3.5 Conclusions

Chapter 4

Analytical ultra-centrifugation and small-angle X-ray scattering of BoNT/E confirm the main results from the MD simulations

4.1 Summary

4.2 Introduction

4.3 Methods

4.3.1 Protein expression

4.3.2 Protein characterisation

4.3.3 AUC data acquisition

4.3.3 SAXS data acquisition

4.4 Results and discussion

4.4.1 BoNT/E Expression

4.4.2 BoNT/E Characterisation

4.4.3 AUC reveals a more extended conformation at pH ≤ 5 and a more compact conformation at pH > 5

4.4.4 SAXS confirms the extended conformation at low pH, and matches structures observed in MD simulations at the same pH values

4.5 Conclusions

Chapter 5

Biophysical studies of BoNT/A identify a potential pathway to membrane insertion

5.1 Summary

5.2 Introduction
List of Figures

**Figure 1.1:** Structural comparison of BoNTs A, E and TeNT.
**Figure 1.2:** Representation of BoNT/A1 in complex with NTNHA and HAs.

**Figure 2.1:** Schematic representation of UAA incorporation process.
**Figure 2.2:** Workflow of a typical MD simulation.
**Figure 2.3:** Illustration of the cpH simulation method.
**Figure 2.4:** Typical CD spectra of common secondary structure features.
**Figure 2.5:** Illustration of the effect of aggregation & over/under subtracting buffer in SAXS.
**Figure 2.6:** Example shape of Kratky plots based on structural features of the protein.
**Figure 2.7:** Jablonski diagram to explain fluorescence.
**Figure 2.8:** Example fluorescence overlap curves from a donor / acceptor pair of dyes.
**Figure 2.9:** Schematic representation of a smFRET setup.

**Figure 3.1:** Raw $R_g$ and RMSD data from the BoNT/E MD simulations.
**Figure 3.2:** Binned data from the $R_g$ and RMSD analysis of the MD simulations.
**Figure 3.3:** Binned RMSD data from domain specific selections of BoNT/E.
**Figure 3.4:** Charge and salt bridge distribution as a function of pH in BoNT/E.
**Figure 3.5:** Two interaction sites between the BD and the LC in BoNT/E, with a loss of positive / negative charge between pH 5.5 and pH 5.
**Figure 3.6:** SASA analysis of BoNT/E as a function of pH.
**Figure 3.7:** Cartoon representation of the TD of BoNT/E, mapped with residue SASA difference with respect to pH 7.
**Figure 3.8:** Cartoon representation of BoNT/E mapped with the residue RMSF difference with respect to pH 7.
**Figure 3.9:** Dynamic cross correlation data of three regions of interest within BoNT/E.
**Figure 3.10:** Elbow plot obtained prior to PCA analysis to determine appropriate number of clusters to use.
**Figure 3.11:** Stacked histograms representation of the PCA data from BoNT/E MD trajectories.
**Figure 3.12:** Cluster mid-points from the PCA distribution of BoNT/E MD trajectories.
**Figure 4.1:** SDS-PAGE of a purification procedure for BoNT/E.

**Figure 4.2:** Mass profile of three samples of BoNT/E obtained by MS.

**Figure 4.3:** CD spectra of BoNT/E at 6 pH conditions and compared to prediction from the PDB structure.

**Figure 4.4:** Sedimentation velocity data of BoNT/E from AUC.

**Figure 4.5:** HPLC trace of BoNT/E in different pH conditions immediately ahead of passing through the X-ray for SAXS.

**Figure 4.6:** SAXS curves (IvQ and Kratky) of BoNT/E, and the best fits from MD frames overlaid.

**Figure 4.7:** Best fit refinement from MD frames to the SAXS using backbone, full atom and hydration shell for the same frame.

**Figure 4.8:** Pair distance distribution function for the closed and open conformations of BoNT/E and their respective best fits.

**Figure 4.9:** 100 best fit frames to the open and closed conformations of BoNT/E visualised in cartoon representation.

**Figure 5.1:** Structural comparison of BoNTs A, E and TeNT.

**Figure 5.2:** Raw $R_G$ and $RMSD$ data from the BoNT/A MD simulations.

**Figure 5.3:** Binned data from the $R_G$ and $RMSD$ analysis of the MD simulations.

**Figure 5.4:** Binned $RMSD$ data from domain specific selections of BoNT/A.

**Figure 5.5:** Average SASA as a function of pH for full BoNT/A.

**Figure 5.6:** Domain specific SASA analysis of BoNT/A.

**Figure 5.7:** Charge and salt bridge distribution as a function of pH in BoNT/A.

**Figure 5.8:** Elbow plot obtained prior to PCA analysis to determine appropriate number of clusters to use.

**Figure 5.9:** Stacked histograms representation of the PCA data from BoNT/E MD trajectories.

**Figure 5.10:** Clusters viewed as a dendrogram.

**Figure 5.11:** Cluster mid-points from the PCA distribution of BoNT/A MD trajectories.

**Figure 5.12:** Interaction sites between BD and “switch” in the different cluster mid-points.

**Figure 5.13:** SDS-PAGE of a purification procedure for BoNT/A.

**Figure 5.14:** Mass profile of three samples of BoNT/A obtained by MS.
**Figure 5.15:** CD spectra of BoNT/A at 6 pH conditions and compared to prediction from the PDB structure.

**Figure 5.16:** Fraction helicity as a function of pH for BoNT/A.

**Figure 5.17:** Sedimentation velocity data of BoNT/A from AUC.

**Figure 5.18:** SAXS curves (IvQ and Kratky) of BoNT/E, and the best fits from MD frames overlaid.

**Figure 5.19:** Best fit refinement procedure for BoNT/A SAXS.

**Figure 5.20:** $R_G$ versus $R$-factor for the frames from pH 7 MD.

**Figure 5.21:** Superposition of the 100 best fit frames to the SAXS data from MD simulations.

**Figure 6.1:** Surface representation of BoNT/A & BoNT/E with the mutated residues in red; and the distances between them.

**Figure 6.2:** SDS-PAGE of an expression test for double mutant BoNT/A with UAA incorporation.

**Figure 6.3:** SDS-PAGE of the final purified fractions of BoNT/A double mutant with UAA incorporation.

**Figure 6.4:** Mass profile of a purified fraction of double mutant BoNT/A obtained by MS.

**Figure 6.5:** Mass profile of a purified fraction of a double mutant BoNT/A after labelling with two dyes, obtained by MS.

**Figure 6.6:** Western Blot of a BoNT/E double mutant expression attempt.

**List of Tables**

**Table 3.1:** Summary of PCA data for BoNT/E.

**Table 6.1:** Summary of double mutants generated for BoNTs A & E.

**Table 6.2:** Summary of PCR cycles and primers used for the mutations of BoNT/E.

**Table 6.3:** Summary of PCR cycles and primers used for the mutations of BoNT/A.
List of Abbreviations

BoNTs = Botulinum neurotoxins
TeNT = Tetanus neurotoxin
BD = Binding domain
TD = Translocation domain
LC = Light chain
HC = Heavy chain
NTNHA = Non-toxic nonhemagglutinin
GI = Gastro-intestinal
HA = Hemagglutinin
PSG = Polysialoganglioside
SV = Synaptic vesicle
SNARE = SNAP Receptor
SNAP = Soluble N-ethylmaleimide-sensitive factor attachment proteins
Cryo-EM = Cryogenic electron microscopy
CD = Circular dichroism
SAXS = Small angle X-ray scattering
smFRET = Single molecule fluorescence resonance energy transfer
NMR = Nuclear magnetic resonance
MD = Molecular dynamics
PDB = Protein data bank
SEC = Size exclusion chromatography
VMD = Visual molecular dynamics
RMSD = Root mean square deviation
RMSF = Root mean square fluctuation
RG = Radius of gyration
PCA = Principal component analysis
AUC = Analytical ultra centrifugation
PTM = Post translational modification

IPTG = Isopropyl-β-D-thiogalactoside
DNA = Deoxyribonucleic acid
PCR = Polymerase chain reaction
IMAC = Immobilized metal ion affinity chromatography
Ni-NTA = Nickel-nitrilotriacetic acid
NaCl = Sodium chloride
EDTA = Ethylenediaminetetraacetic acid
HABA = Hydroxy-azophenyl-benzoic acid
HIC = Hydrophobic interaction chromatography
IEX = Ion exchange chromatography
UAA = Unnatural amino acid
pAzF = p-azido-phenylalanine
tRNA = transfer ribonucleic acid
aaRS = acyl-tRNA synthetase
mRNA = messenger RNA
HPC = High performance computer
cpH = Constant pH
SASA = Solvent accessible surface area
MRE = Mean residue ellipticity
RPM = Rounds per minute
APD = Avalanche photo diodes
NMR = Nuclear magnetic resonance
TM = Transmembrane
MS = Mass spectrometry
OD_{600} = Optical density at 600nm
RCF = Relative centrifugal force
mTB = Modified terrific broth
MWCO = Molecular weight cut-off
QTOF LC-MS = Quadrupole time of flight liquid chromatography mass spectrometry
ESI = Electrospray ionisation
HPLC = High performance liquid chromatography
SDS-PAGE = Sodium dodecyl sulfate polyacrylamide gel electrophoresis
AF488/594 = Alexa fluor 488/594
SDM = Site-directed mutagenesis
SPAD = Single photon avalanche diode
List of Units

kDa = Kilodalton
nm = Nanometre
mg = Milligram
mL = Millilitre
μL = Microlitre
g = Gram
L = Litre
mM = Millimolar
s = Second
mdeg = Millidegrees
cm = Centimetre

ns = Nanosecond
ps = Picosecond
°C = Degree Celsius
μg = Microgram
M = Molar
mm = Millimetre
GeV = Giga electron volts
keV = Kilo electron volts
Da = Dalton
S = Svedbergs
Å = Ångstrom
CHAPTER 1

Introduction
1.1 Botulinum Neurotoxins (BoNTs)

1.1.1 Short history of BoNTs

Botulism is a condition which causes flaccid paralysis, and if left untreated it may become lethal by affecting respiratory muscles. Although it is very likely the disease has been around throughout human history (1), it was first recorded in the late-1700’s after being linked to the consumption of infected sausages in southwest Germany (2, 3). This later led to their naming in the late-1800s by a German physician, John Muller, who coined the name botulism; derived from the latin botuli, sausage (4). BoNTs are produced by anaerobic spore-forming bacteria Clostridium botulinum, a species from the Clostridium family, which also includes the Clostridium tetani species responsible for the tetanus toxin (TeNT) which is highly similar in structure and function (figure 1.1). These spores were discovered in Belgium by Emile Pierre-Marie van Ermengen, of the University of Ghent, in 1895 (5). The bacteria were isolated from infected ham, which had been linked to an outbreak in a nearby village. However, Clostridium botulinum bacteria were not observed to be pathogenic, confirming the previous observations from Germany that a toxin was responsible. While initially believed to be primarily a food-borne disease, later observations in the 1940s identified spores in wounds; and in the 70s in infants’ intestines.

Several serotypes of BoNTs have been identified over the years, the first of which was BoNT/A. Traditionally, seven serotypes are accepted and termed BoNT/A through to BoNT/G (6). Each serotype can be further split into subtypes, BoNT/A1, A2 etc. Increased accuracy of sequencing techniques and evolution will likely result in more serotypes and subtypes being discovered (7). A new serotype, BoNT/X, has already been identified as a novel variation, targeting a unique combination of proteins (8, 9).
Figure 1.1: Cartoon representation of BoNTs A (a) and E (b). (c) Overlaid cartoon representations of BoNTs A, E and TeNT. (d) Schematic representation of the domain’s relative positions in BoNTs A, E and TeNT.

The first purification of BoNTs was achieved in 1926 in California (10), where a purified BoNT/A was obtained through adsorption, elution, dialysis and evaporation. The isolation of BoNT/A marked a turning point in their history, with the advent of therapeutic uses. Building on the findings of Kerner (11, 12), potential therapeutic applications were devised. The first use was by Alan Scott to treat surgically induced strabismus in rhesus monkeys and proved extremely successful, with additional positive side-effects, notably reduction of blepharospasm (13, 14). It was observed then that a small dose of BoNT/A had beneficial effects lasting up to 8 months, without causing any negative side-effects. Scott started commercialising BoNT/A under the name “Oculinum” which was quickly bought by Allergan who then changed the name to the now internationally recognised Botox. While Botox is most famous for its cosmetic applications, the achievements of Scott inspired many other ventures in the use of BoNTs as a therapeutic protein. In the United Kingdom, doctors who had trained with Scott formed a company, later acquired by IPSEN, which specialised in using BoNTs for the treatment of dystonic conditions, under the name Dysport. Part of the
reason this is possible is the historically low occurrence of botulism, as opposed to tetanus. No botulism vaccine has ever been developed, meaning there is reduced likelihood of an immune response to the injection of this toxin compared to TeNT.

The potential of BoNTs as therapeutics is still being explored (15–19) with new applications being discovered, sometimes accidentally as a side effect (20). Currently approved uses, and uses under study, include the fields of ophthalmology, neurology, pain, urology, gastroenterology and psychiatry (6). Using the proteins’ modular structure (see below), many more applications are being tested by combining the targeting and binding properties of these toxins. Such a result was achieved to treat botulism in mice, by attaching an antibody to the heavy chain of the toxin (21). Currently, the vast majority of therapeutic drugs making use of BoNTs are built on BoNT/A. There is interest in attempting to use other serotypes, namely BoNT/E as it has a faster onset of action (22). However, it also has a shorter lifetime after injection. It is therefore of interest to attempt to understand where these differences come from, despite a strong sequence and structural homology between the serotypes. A better understanding of these differences and their root cause would open the door to chimeric toxins which make use of the faster onset of action of BoNT/E and the longer duration of BoNT/A for an optimised drug.

1.1.2 BoNT structure

Part of the therapeutic attractiveness of BoNTs, other than their activity (see below), is their modular structure which can be broken down into functional domains. Initially, these proteins are produced as single polypeptide chains of ~ 150 kDa and can be split into a heavy chain and a light chain (HC ~ 100 kDa, LC ~ 50 kDa). The LC is the active part of the toxin, and acts as a metalloprotease when a zinc ion is bound. This catalytic domain is held to the HC with a disulphide bond, which is reduced prior to LC release into the cytosol of the affected neuron. The HC can be further split into two domains, the translocation domain and the binding domain – referred to in this thesis as TD and BD, respectively. The TD is believed to mediate pore-formation which enables the release of the LC into the cytosol and is made up of several alpha-helices as well as a belt region which wraps around the LC. The BD domain is responsible for
the initial neuron membrane binding of the toxin as well as the uptake into the neuron using a synaptic vesicle protein.

This domain breakdown holds true for all the known BoNTs, however their spatial arrangement can vary. The crystal structures of BoNTs A (PDB ID: 3BTA), B (PDB ID: 1EPW) and E (PDB ID: 3FFZ) are known (22–24), as well as many individual domains alone or in complexes (25–28). BoNTs A (figure 1.1) and B show an extended conformation of the three domains, such that the LC and the BD are situated to the left and right of the TD. However, in BoNT/E the BD is brought into closer proximity to the LC such that it adopts a more compact conformation (figure 1.1). This difference in conformation is thought to play a part in the faster translocation of BoNT/E compared to BoNT/A (22).

BoNTs are produced in conjunction with a secondary protein, non-toxic nonhemagglutinin (NTNHA), which has a very strong structural similarity to BoNT/A but lacks the zinc binding site. When the protein is released, the two proteins form a complex (29) (figure 1.2) which is thought to protect the protein from protease-rich environments, and acidic environments such as the gastro-intestinal (GI) tract which is the most common entry path for intoxication. In this complex, BoNT/A adopts a different conformation, with the BD moved ≈ 90° such that it hangs “in front” of the TD. This conformation is induced by the NTNHA rather than the pH (30). In addition to the NTNHA protein, the BoNT and NTNHA genes are located near hemagglutination capable proteins (HA), which bind weakly to the NTNHA protein (31, 32). These are not believed to be protective, but rather enable binding to the intestinal mucus layer and hence facilitate the toxin’s entry into the bloodstream (33).
Figure 1.2: Representation of BoNT/A1 in its complex with NTNHA and HA proteins. Image taken from (6).
1.1.3 BoNT Activity

BoNTs are produced by *Clostridium Botulinum*, a gram-positive spore-forming bacteria which grows in soils around the world. Three major types of botulism exist: food-borne (the most common), infant & wound. In infant botulism, the spores are directly swallowed by the infant and the gastric environment is not yet developed enough to prevent the spores from becoming full bacteria; after which they start producing the toxin. Wound botulism is caused by spores entering an exposed wound and reproducing; eventually releasing the toxin into the bloodstream. In food borne botulism, improperly prepared food which has been contaminated by *C. Botulinum*, and hence BoNTs, is ingested. The toxin is then released and protected by a complex (see above; figure 1.2) and enters the bloodstream.

After entering the bloodstream, BoNTs target primarily motor neurons and prevent neurotransmission by cleaving a family of proteins responsible for exocytosis. It is thanks to this high-specificity and toxicity that they can also be used for therapeutic purposes, in minimal doses. When injected to a targeted location, there is little to no diffusion. The effects are not permanent, but long lasting enough to provide relief for extended periods of time. Through their targeting of motor neurons, they can be used for a very broad range of conditions, including excessive perspiration, spasticity disorders and muscle pain.

They exert their activity by first binding to a polysialoganglioside (PSG) receptor on the presynaptic membrane (34–36), and a second binding to a protein receptor which varies depending on the serotype. Some known receptors so far include synaptotagmin and glycosylated synaptic vesicle 2 (SV2). Detailed studies of the specific BoNT-PSG interactions are available but are beyond the scope of this thesis (6, 34). After binding, BoNTs are taken up into the neuron. The toxicity of BoNTs has been shown to increase when stimulating the infected neuron (37), suggesting that the internalisation is facilitated by an increased rate of exoendocytosis. The toxin is internalised in a SV before it is recycled for another cycle of neurotransmission. Once inside the vesicle, the protein is activated during the acidification step and primed to form a transmembrane channel which can release the LC into the cytosol.
Little is known of the specifics of this step, but the LC is released into the cytosol. The study of this process is challenging due to the inaccessibility of the SV lumen, and methods bypassing the intracellular delivery have to be used. This limits the information which can be gained from the study, such as the specific structural changes occurring in the protein which make the translocation possible. However, it has been used to detect that activity is only possible in the pH range 4.5 to 6 and with an increased activity below pH 6 (35). Additionally, it was found that the BoNT molecule acquires a net positive charge which pulls the protein towards the anionic surface of the membrane, generating a lipid-protein complex. A study found that neutralising negative surface charges accelerated the cytosolic delivery, further emphasising the importance of the surface charge (38). After this it is not known whether the LC undergoes a significant structural change which would allow it to enter the membrane, or whether the HC forms a transmembrane channel large enough to accommodate the LC. It is very likely that the LC release mechanism is common to all BoNTs, given their overall structural homology across all serotypes, and understanding this step could be key in developing efficient inhibitors of BoNTs following intoxication.

Following translocation of the LC, the disulphide bond linking the LC to the HC is reduced and enables the full release of the toxin’s activity. Once inside the cytosol the LC acts as a metalloprotease, requiring a zinc atom for activity (39, 40). It targets specifically the SNARE family of proteins (SNAP receptor – soluble N-ethylmaleimide-sensitive factor attachment proteins). There are three distinct SNARE proteins: VAMP/synaptobrevin, SNAP-25 and syntaxin. The main difference between the serotypes and subtypes of BoNTs are the specific proteins and peptides targeted for cleaving, details of which can be found in previous studies (41–46). These discoveries were the first time exocytosis was linked to SNARE proteins and provided insight into an additional step of the neurotransmission process.

Finally, the attractiveness of these toxins for therapeutic purposes also includes their complete reversibility. Even after an involuntary exposure to an uncontrolled dose of BoNTs, a patient will survive on the condition of being supported for respiration and given food (47). BoNTs are not cytotoxic and do not cause neuronal damage. The
duration of action varies greatly from one serotype to the other, and with the dosage used. For treatment of dystonias in humans, BoNT/A, B and E had paralytic effects lasting 4 months, 2 months and less than 4 weeks, respectively (48). By deepening our understanding of these toxins and their mechanism of action, we may further the development of novel drugs making use of BoNTs. This may be achieved by discovering key regions of the protein involved in the membrane insertion process, or by assessing the effects of different formulations on protein stability. Through thermal aggregation studies and molecular dynamics simulations, we hope to shed light on to some of these questions.

1.2 Protein structural studies

1.2.1 Solution vs fixed methods

When searching for structural features of a protein, there are two major avenues available: fixed methods or solution methods. Fixed methods include X-ray crystallography and cryogenic electron microscopy (cryo-EM), while solution methods include circular dichroism (CD), small angle X-ray scattering (SAXS), single molecule fluorescence resonance energy transfer (smFRET), nuclear magnetic resonance (NMR) and molecular dynamics (MD). In fixed methods, which contribute the majority of the entries in the PDB, the protein is first fixed in place which inherently prevents any dynamics from being observed, although dynamic regions are hinted at by the absence of coordinates. Flexible regions of the protein are often difficult to characterise with high accuracy and these methods are best suited for rigid, non-dynamic proteins. Specifically, in crystallography, the protein is locked in a crystal lattice with some proteins adopting a non-native state due to the crystal packing (49). Proteins with multiple domains, such as BoNTs, may have their relative positions displaced because of this; in turn affecting our understanding of their structure and dynamic behaviour. Moreover, it is not possible to observe the motion of the domains in different environmental conditions. In such complex proteins as BoNTs, which are believed to undergo multiple structural changes throughout their intoxication process, having a clear understanding of the motion and relative positions of the domains is crucial to understanding their activity. New methods are being developed to enhance
the power of X-ray crystallography, such as X-ray free electron lasers. In this setup, a
free electron X-ray laser generates extremely bright femtosecond scale pulses of
radiation which illuminates a flow of nanometer-sized crystals. Each individual crystal
produces a diffraction pattern, and enables the observation of biological process on
nanosecond timescales (50). Electron microscopy, and the recent improvement of it
cryo-EM, is another fixed method which can provide high resolution structures even of
dynamic, multi-domain proteins (51). Cryo-EM holds promise for the imaging of
proteins embedded in membranes, which could be applied to BoNTs. However, the
main problem remains the same: the proteins imaged by this method are fixed by
flash-freezing them, thereby occluding any information on motions within the protein.

Common to both of these methods, however, is the lengthy and highly delicate
sample preparation. The sample volumes and concentrations required are very low
which can make their handling difficult. For crystallography, many crystallising
conditions are tested in parallel until one suitable crystals are obtained. For some
proteins this never happens, and it is notoriously difficult for flexible proteins. The
quality of the structural information is highly dependent on the quality of the crystal
obtained, and collecting data from a protein can be a lengthy process. Extracting a
novel protein structure from a crystallography experiment also requires mapping the
data obtained from a protein crystal to that of a heavy-metal coated protein, adding
further difficulty to the analysis and interpretation. In cryo-EM, the proteins are placed
in a thin layer of water for vitrification, a process during which they are exposed to
conditions and surfaces unlike those of a test tube or a cell (52). These conditions may
have unknown effects on the protein to be studied, and may be destructive. The poor
understanding around this makes troubleshooting difficult during cryo-EM sample
prep. Before submitting a sample for cryo-EM it should be subjected to multiple rounds
of analysis to ensure the optimal conditions are obtained.

By contrast, solution-based methods are significantly simpler to use. In CD,
which can be used to determine the secondary structure of a protein (53, 54), the only
constraint is the interaction of buffer components with the polarised light which can
increase their signal to noise ratio. The protein can otherwise be placed in any suitable
buffer, with no modifications to it and at concentrations between 0.02 to 2 mg/mL
depending on the pathlength. In SAXS experiments, there are no constraints on the buffer or protein concentration. The protein should be in a buffer which matches perfectly the submitted blank buffer sample, which is usually achieved by overnight dialysis in a 1000 fold volume dilution (500 μL sample in 500 mL buffer). By following this condition, any buffer can be used and the protein can be subjected to many environments such as varied ionic strength, pH or temperature. Using solution methods also provides information of the protein in a close to native state, hence maintaining its dynamic behaviour. Through the use of SEC-SAXS it is also possible to separate multiple populations of the same protein in solution, if they are different enough. Therefore, if the protein is aggregation prone, these can be separated out and the monomeric protein’s data can still be collected. Finally, in smFRET experiments, the signal from individual proteins is measured and a histogram of FRET efficiencies, which can be directly linked to intra-molecular distances, is built up over time. This can also single out multiple conformations of the same protein in solution, even if it is a very rare or transient event.

MD simulations can be considered as a solution study of proteins, as they can be simulated in a box of explicit solvent. The main advantages of MD simulations are their atomic resolution, relative inexpensiveness and the possibility of running them prior to or in parallel to experimental work. The trajectories can also be saved with picosecond timesteps, which means for a simulation on the order of 100s of nanoseconds, tens of thousands of individual structures can be generated. Through the use of visualisation software such as VMD or PyMOL (55, 56) local structural changes can be visualised and interactions can be identified which can help with the subsequent interpretation of experimental data. A number of tools are also available, beyond the normal MD analysis methods such as RMSD, \( R_g \) or PCA, to directly compare MD generated structures to data obtained experimentally (57–59).

The majority of structural information currently known for BoNTs is from fixed methods (22, 25, 60, 61), and hence does not contain any information about the dynamics and relative motions of the domains. The crystals are obtained at a single pH value, often on truncated proteins, and do not explore the structural change of the full protein across the activation pH range. By applying solution-based methods to
these toxins, we hope to further our understanding of the biochemical processes driving membrane insertion across their pH activation barrier.

1.2.2 Impact of pH environment

Several environmental factors may have an impact on a protein’s function, such as temperature, ionic strength, pressure and pH (62, 63). Of particular interest for this thesis is the effect of pH on protein flexibility, as BoNTs have a strong pH dependence for their activity. From their currently known structures it is highly likely that parts of the toxin are displaced to facilitate the transition to a transmembrane channel. It has already been shown that low pH has a destabilising effect on certain proteins, such as chymotrypsin inhibitor 2 where a lower unfolding energy was required in acidic environments (62). This is partially due to the protonation of particular residues and the resulting electrostatic interactions in the protein (64). Such interactions can be identified within MD simulations and may inform on the driving forces behind any observed structural changes, and protein activity. Several studies have been conducted on individual domains of BoNTs under varying pH conditions, which have identified certain properties of the LC and the TD likely linked to the translocation mechanism (65–67). However, these studies do not account for interactions with the other two domains in the full protein.

Several other proteins have been linked to pH dependent, and specifically acidic environments, to release their payloads, such as influenza virus hemagglutinin or ebola virus GP. Both of these rely on a small scale secondary structure change which exposes a hydrophobic loop normally buried at neutral pH conditions (68). A similar structural change was observed in the BoNT/A TD in isolation, on a small region which is highly conserved between all BoNT serotypes, as well as in TeNT. TeNT also requires acidic conditions to release its payload (69, 70), in a highly similar fashion to BoNTs due to their high domain similitude, despite slight differences in their relative positions. In a previous study, TeNT was shown to adopt a compact conformation, close to that of BoNT/E, in neutral pH conditions and change to a more expanded conformation at acidic conditions (71). Due to its closeness to BoNT/E, whereby BD and LC form an interface in neutral pH conditions, a similar investigation into BoNT/E may reveal structural features key to its membrane translocation mechanism. BoNTs
also contain conserved features from the diphtheria toxin (72), another pH mediated toxin (73), in their TD. Changes in these areas (diphtheria conserved helix & “switch” region from BoNT/A) across the pH barrier could provide insight into the membrane insertion process of BoNTs, and may help to explain the differences between serotypes A and E.

By studying BoNTs A and E in a range of solution-based experiments (MD, CD, AUC, SAXS and smFRET), we hope to uncover some of the currently elusive structural features that play a role in their transition to a transmembrane channel. By observing atomic level details in the MD and directly comparing the structures generated to experimental data obtained in the same pH conditions, we hope to identify key regions of the protein for future targeted healthcare solutions. In addition to furthering our understanding of these serotypes individually, biochemical mechanisms identified may provide an answer to the differences in their kinetic properties: why does BoNT/A has a longer onset of action compared to BoNT/E?
Chapter 2

Methods Theory
In this chapter, we discuss the theory behind the methods used in this thesis. Specific protocols and conditions used for the experiments presented in this thesis are included in the relevant chapters separately.

### 2.1 Protein expression, mutation and labelling

#### 2.1.1 Protein expression

Biophysical study of proteins typically requires relatively large amounts of material to obtain a good signal to noise ratio and meaningful data. This is most commonly achieved by over-expression using an appropriate host cell organism. This can be yeast, mammalian cells, insect cells or *Escherichia coli* (*E.coli*) depending on the target protein. Mammalian cells have the advantage of adding the relevant post-translational modifications (PTMs) relative to human proteins with a good yield (up to 1 g/L), but are typically harder to maintain. Yeast cells have the highest yield (up to 10 g/L) but take longer to grow and express the protein. They are suitable for large-scale production in bioreactors. *E.coli* cells have the fastest turn around with rapid growth and requiring little maintenance, but they also provide the lowest yield due to the challenge of scaling up (up to 10 mg/mL). For this project, *E.coli* was chosen as the expression system due to its availability and ease of use, as well as the absence of PTMs required for the protein.

When using *E.coli* the gene of interest is cloned into a vector conferring antibiotic resistance, and a promoter mechanism to induce protein production. This ensures that cells growing post-transformation of the vector are the cells carrying the plasmid, and allows for control cultures to compare protein expression with and without induction. The most commonly used induction mechanism is through the T7 promoter which can be induced with isopropyl-β-D-thiogalactoside (IPTG). Different strains of *E.coli* exist which are optimised for either accurate vector replication ahead of DNA purification for sequencing or stock making (TOP10), or for increased yield of the protein of interest (BL21(DE3)). After choosing the appropriate host cell organism and transforming the vector, the protein can be expressed and purified for biophysical study.
2.1.2 Protein mutation

Some biophysical studies require the protein of interest to be mutated, for stability studies or for labelling purposes, for example. If this is the case, site-directed mutagenesis through polymerase chain reaction (PCR) may be used. This follows the standard PCR protocol, but the primers used carry the mutation required. Typically, the forward primer carries the mutation and the reverse primer is complementary to the wild type. This means the forward primer will modify the template DNA, and when the reverse primer reaches the same site it will also add the complementary base to the mutated one. After repeating this cycle 25-30 times, the amount of mutated DNA far outweighs the template DNA provided. The PCR product can be cleaned up by phosphorylation, which only reacts with the mutated strands, and ligated to be re-circularised. The template DNA is then removed, and the remaining product can be transformed into an appropriate cloning strain. After growth of these cells, a miniprep can be performed and the resulting DNA is sequenced to confirm the incorporation of the desired mutation.

2.1.3 Protein purification

Proteins for biophysical study need to be highly pure and free of contaminants to ensure any signal detected is wholly from the protein of interest. This can be achieved through chromatography, of which there are several types. For the highest purity, affinity tags may be added onto the protein of interest by adding the relevant sequences to the vector prior to transformation. Typical tags include his-tags, where several histidines are added to the end of the sequence, and strep-tags, where a custom sequence of residues is added which binds to streptavidin. A his-tagged protein can be purified by using immobilized metal ion affinity chromatography (IMAC); of which the most common is nickel-nitrilotriacetic acid (Ni-NTA). With this method, the column is equilibrated with low concentration imidazole (≈ 10 mM) then washed with slightly more concentrated imidazole (≈ 20 mM) to remove low-binding proteins. Then either a linear or stepped gradient elution is done with highly concentrated imidazole (≈ 250 mM). For strep-tags, the column is equilibrated with Tris, NaCl and EDTA and eluted with the same buffer supplemented with desthiobiotin. The column can be
regenerated by washing with the equilibration buffer supplemented with hydroxyazophenyl-benzoic acid (HABA).

Non affinity tag based purifications are also possible if the protein needs to be studied with minimal interference of the sequence. Some of the most common techniques are hydrophobic interaction chromatography (HIC), ion exchange chromatography (IEX) and size exclusion chromatography (SEC). In HIC the sample is applied to the column in a high-salt buffer, which allows the separation of proteins based on differences in hydrophobic surface. This is usually used in conjunction with other chromatography steps as it is not particularly specific. In IEX, proteins are separated based on their net charge at a specific pH. Using the isoelectric point (pI), where the net charge of a protein is 0, an appropriate pH value can be selected for optimal separation of the protein of interest. Two types of IEX exist depending on the charge of the protein. A positively charged protein will bind to a cation exchange resin, while a negatively charged protein will bind to an anion exchange resin. Once the appropriate pH conditions and resin are found, the protein is loaded onto the column in a low salt buffer and eluted with an increasing salt gradient. This can usually produce a higher separation than HIC, due to its tailoring to the protein of interest. SEC is a non-binding method of separation which relies on the size of the protein. A tightly packed resin is used which increases the apparent column volume for small molecules compared to larger ones. The larger the column volume, the better the separation; however this also results in increasingly dilute protein sample recuperated. A particular type of SEC column can be used for buffer exchanging from high-salt buffers to low-salt buffers, termed desalting columns. In practice, for a final product with high purity and without the use of affinity tags, a combination of these 3 methods can be used to gradually remove contaminant proteins.

2.1.4 Protein labelling

Protein labelling can be achieved by several methods, either using naturally occurring residues or artificially added ones. A common method for labelling is the reactivity of cysteine residues, which can be labelled through maleimide reactions. Using this, it is possible to label cysteines naturally present in the protein if these are solvent exposed, or mutate sites of interest into cysteines for site-specific labelling.
However, in the case of large proteins this can be difficult to achieve as there will be a large number of cysteines already present. This would then lead to non-specific labelling or require the additional mutation of cysteines to non-reactive residues, without affecting the structure of the protein. Therefore, a nascent field of particular interest for this is the development of unnatural amino acids (UAAs), which can be incorporated into proteins through specialised cellular machinery. These UAAs can be tailored for specific reactions through custom side-chains such as azide groups for click reactions with amine groups. Then, custom or commercially available dyes can be used for site-specific labelling of these UAAs with no risk of non-specific binding to any other site within the protein.

2.1.5 UAA incorporation

As these residues are, by definition, not naturally present in the culture media, they need to be supplemented in the growth medium. For this project, p-azido-phenylalanine (pAzF) was used, which can be purchased in powdered form and resuspended in 100 mM sodium hydroxide. This can then be added to the expression culture for incorporation into the protein of interest. The incorporation requires an additional plasmid co-transformed into the cells, which codes for a suppressor tRNA and an evolved amino acyl-tRNA synthetase (aaRS) specific for the UAA of interest. When the mRNA is being translated, the amber codon (TAG) is recognised by the tRNA loaded with the UAA and it is added to the nascent protein. Figure 2.1 represents the process schematically. This additional plasmid was pULTRA-CNF (74, 75) which is also induced by IPTG, and confers the additional spectinomycin antibiotic resistance.
2.2 Computational methods

2.2.1 Molecular dynamic simulations

2.2.1.1 Standard MD

Molecular dynamic (MD) simulations provide atom level insight into the dynamics of proteins. They are performed using computational approximations of forces applied to every atom in the system, and can be started from known structures in the PDB format. They can be used to obtain information about protein stability, conformational changes, ligand interactions, as well as reporting on the underlying causes for such changes. They have been increasingly used in drug design and the general study of proteins as they are cheap to run compared to experimental work. They also provide an almost limitless playing ground for testing of multiple conditions (temperature, pressure, ionic strength, pH) without using any physical material. However, they remain simulations and only provide an approximation, a lead to be
further investigated. Therefore, they are mostly used as part of a screening process to reduce the amount of experimental work downstream. A typical MD workflow is illustrated in figure 2.2.

Figure 2.2: A typical workflow for standard MD simulations. Image taken from: (77).

Two types of forces can be used, classical (Newtonian) mechanics or quantum mechanics. Due to the increased computational power required for quantum mechanics calculations this is typically reserved for smaller molecules, while proteins are studied under Newtonian mechanics conditions. Using an approximation of a force field, Newton’s equations of motion are applied to every atom in the simulated system to determine its velocity and position at any time point. The force on every atom is calculated with respect to every other atom such that the force on one atom is:

\[ F(x) = -\Delta U(x) \]  
(Eq. 2.1)

with \( x \) the coordinates of an atom and \( U(x) \) the potential energy function. The atoms’ motions are then determined using Newton’s second law:
\[ F = ma \]  
(Eq. 2.2)

with \( F \) the force on an atom, and \( m \) and \( a \) the mass and acceleration of the atom, respectively. Then the velocity can be determined from the acceleration, and from the velocity the atom positions can be updated for every timestep. A typical timestep to use is of the order of femtoseconds \((10^{-15} \text{ s})\), such that the positions and velocities of every atom are updated on a timescale shorter than the fastest movements in the protein.

Force fields define all the interactions between atoms, which includes bonded (bond lengths, bond rotations, dihedral angles, improper angles) and non-bonded interactions (van der Waals and electrostatic interactions). These parameters in the force field are estimated using known experimental and calculated quantum mechanical data. There is a wide choice of force fields available to use for MD simulations, all with slight differences in their parametrisation using known data for different sets of molecules. It is therefore important to choose a force field that is developed from a set of molecules similar to the molecule studied. The choice of solvent model is also affected by the force field as some water models are incompatible with specific force fields. Common force fields are CHARMM, AMBER, OPLS-AA (78–83), all with multiple iterations of their main parametrisations. MD simulations can be performed using a number of softwares, including NAMD, AMBER (separate from the AMBER forcefields), Gromacs, Desmond, CHARMM, as well as many other commercial software packages. When visualising trajectories obtained by Gromacs or NAMD, the principal software used is visual molecular dynamics (VMD), which can also be used for analysis.

Two types of solvation are possible when performing MD: implicit or explicit. The implicit solvent approximates the average effect of solvent molecules on the protein surface and does not generate additional atoms in the simulated system. However, it neglects the steric presence of water molecules. Explicit solvent is more reliable but also adds a significant number of additional atoms for which to solve Newton's equations, slowing the calculations down. When using explicit solvent, the
protein is centred inside a simulation box such that it should not extend beyond the box. Periodic boundary conditions are used to maintain the same quantity of solvent molecules, such that if a molecule leaves the simulation box it re-enters on the opposite side. This also applies to the protein if parts of it do leave the box, and the trajectory can be corrected such that the protein is re-centred. Multiple water models are available with minor property changes, the choice largely depends on the forcefield as some are incompatible. Common water models include TIP3P, SPC/E or TIP4P.

MD simulation time is largely governed by the available computational power and the size of the system. For a large system such as BoNT in explicit solvent, a significant time in the simulation can be considered as equilibration time, usually tens of nanoseconds. For significant conformational changes to be detected simulations should be at least hundreds of nanoseconds long. These long simulations have only been possible quite recently with advances in computational power. By using the high performance computer (HPC) Kathleen at UCL, we were able to achieve 400 ns long simulations and repeat these multiple times for each condition studied.

2.2.1.2 Constant pH MD

Standard MD can be used to simulate a system in set conditions such as temperature, pressure and ionic strength; however simulating true constant pH is more challenging as it would require re- or de-protonating the relevant side chains frequently throughout the simulation. This in turn would constantly affect the number of atoms present in the simulation and would require updating the files accordingly. Some workarounds are possible, and a script has been developed for use with NAMD. This script performs two operations side by side: one normal MD run as described above, and one constant pH (cpH) specific run. In the cpH specific step, both the conformation and the protonation states are tested, the latter being dependent on the conformation and the pKa of the residues. During this search, some trajectories will be discarded until one fits the correct requirements (figure 2.3). Then the regular MD starts from the end of the cpH specific trajectory for a short time until the next cpH testing. This results in significantly increased computational costs, as effectively two trajectories are being generated including one which samples both protonation state and conformation. The protonation state is calculated based on the conformation and the pKa of the residue,
and the chosen pH to simulate. To avoid the issue with changing numbers of atoms throughout the trajectory, this script generates dummy protons which are activated or not depending on the protonation state of the residue.

![Diagram](image.png)

**Figure 2.3:** Schematic representation of a cpH run. Regular MD is in blue, cpH specific in green. Image taken from (84).

### 2.2.2 MD data analysis

The trajectories generated by MD simulations can provide a very broad range of information, the most common of which will be discussed here. A number of distance based values can be obtained, including the root mean square deviation ($RMSD$). This is a measure of how much the protein has deviated from a reference structure and is given for every timepoint recorded in the trajectory. Typically, there will be a sharp increase in $RMSD$ at the start of the simulation followed by a brief plateau, due to the protein in presence of explicit solvent relaxing to a lower energy state from the starting point. Then, if the simulated protein remains very close to the reference structure this value will remain within a narrow range of values. If the protein deviates from the reference structure the $RMSD$ will increase. A related value which can provide structural information is the radius of gyration ($R_g$) which reports on the apparent size.
of the protein. It is defined as the root mean square average of the distance of all scattering elements from the centre of mass of the protein. Therefore, a more extended structure would have an increased $R_g$ value with respect to a compact structure of the same protein. Finally, root mean square fluctuation ($RMSF$) measures the average displacement of individual atoms over the course of the trajectory, without the use of a reference structure. This can be used to identify dynamic or flexible regions within proteins.

Linked to the $RMSF$ calculation is the cross-correlation value, which provides a view of residues in the protein moving at the same time. This can inform, for example, of potential sites to mutate for stabilising effects without affecting an active site. It can also highlight whether multiple regions are likely to be involved in the same interaction event such as membrane binding.

The solvent accessible surface area ($SASA$) of a protein can also be calculated and monitored over time in MD trajectories, and can be used to detect structural changes linked to burying more of the protein or exposing larger patches to the solvent. This value can be obtained for both the full protein and on a residue basis, to detect specific residues potentially involved in interactions linked to solvent accessibility.

2.2.3 Principal Component Analysis

Principal component analysis (PCA) is a useful tool for the classification of large datasets and to identify patterns hidden within. It is based on the variation within the sample and can be used to cluster similar groups of data points within the data set. In the case of MD trajectories, it is used to group structures of similar conformations. This can be used within a single trajectory to identify distinct conformations; but it is perhaps better used when combining multiple trajectories and simulated conditions. By applying PCA to a single trajectory combining multiple conditions, similar structures can be grouped and the direct effect of the conditions can be observed by checking the cluster distribution as a function of the conditions tested. A package for the R language is available, Bio3D (85), which provides the tools to perform PCA and clustering on a MD trajectory. The first step is to create a three dimensional matrix from the trajectory. PCA is then applied to this matrix, and a first analysis determines the optimal number of clusters to use to account for the majority of the variance. The
PCA object is then grouped into the determined number of clusters and a two column file comprising of frame number and cluster is generated. Finally, the frame which has the lowest RMSD with respect to the other frames in the same cluster (the cluster mid-point) is output. By linking the frames to their MD conditions, the distribution of clusters per condition can be calculated and may reveal, for example, a conformation exclusive to a particular condition.

2.2.4 Fitting of MD frames to experimental data

While MD data has proven increasingly reliable over time with the development of more accurate force fields and solvent models, it is crucial to be able to compare the simulations data with real-world results. Some tools are available to help with that, and of particular interest for this project are two web-based servers which predict circular dichroism (CD) and small-angle X-ray scattering (SAXS) data from MD trajectories. The former is PDBMD2CD (59), which uses a combination of least-squares and linear fitting of reference spectra for known secondary structures to generate theoretical CD spectra of an input structure or trajectory. The latter is SASSIE (57), which comprises multiple modules including the generation of theoretical SAXS curves from input MD trajectories. These theoretical SAXS curves are then compared to the experimental by using an R-factor fit using the equation:

$$R = \frac{\sum ||I_{\text{Expt}}(Q)|| - ||I_{\text{Theor}}(Q)||}{\sum ||I_{\text{Expt}}(Q)||} \times 100$$  \hspace{1cm} (Eq. 2.3)

with $I_{\text{Expt}}(Q)$ and $I_{\text{Theor}}(Q)$ the experimental and theoretical intensities at Q, respectively. From Eq. 2.3, a perfect match of experimental and theoretical curves would lead to an R-factor of 0%; and any deviation from this would result in an increased R-factor.

2.3 Solution structure methods

2.3.1 Circular dichroism (CD)

Circular dichroism (CD) relies on the absorbance of circularly polarized light by a solution in the light path. Circularly polarized light is generated by rotating the direction of the electric field vector about its propagation direction. The signal
measured by CD is the difference between the absorbance of left- and right- circularly polarized light. In the context of bio-applications CD is used to determine the secondary structure of a protein, due to the specific CD signal produced by different secondary structure features. α-helices, β-sheets or β-turns all produce recognisable CD signals (figure 2.4) which can be deconvoluted by fitting to known CD signals and the secondary structure distribution of any signal measured can be determined. The most common used feature is that of α-helices which produce a distinct minimum at 222 nm. This can be used to determine the percentage helicity of a sample, and can be used to determine the melting temperature in a heating experiment or detect induced unfolding with increasing chaotropic agent concentration. The raw CD signal for proteins is usually converted into units of mean residue ellipticity (MRE) by dividing the molar ellipticity of the protein by the number of residues. The raw output of CD is in millidegrees (mdeg) which can be converted to molar ellipticity [θ] by the following equation:

$$[\theta] = \frac{\text{mdeg} \times M}{10 \times L \times C}$$  \hspace{1cm} (Eq. 2.4)

with [θ] the molar ellipticity, mdeg the raw CD output, M the average molecular weight in g/mol, L the pathlength in cm, C the protein concentration in mg/mL.
**Figure 2.4:** Typical CD signals of the most common secondary structural features. Image taken from: (86).
2.3.2 Analytical ultra-centrifugation (AUC)

Analytical ultra-centrifugation is a low-level biophysical study which can be used to determine protein aggregation or conformational changes in a protein in solution. The principle is straightforward: a protein sample is loaded into a sample cell alongside a slightly larger volume of matching buffer, in separate chambers. The sample cell is then spun at high RPM (typically 40K) and the protein is pushed through centrifugal force towards the bottom end of the cell. As the cycle runs, the cell passes in front of an objective which records the radial absorbance and the equipment stores all the readings in order. At the end of the run, the files can be analysed and the sedimentation velocity of the protein is determined by fitting the radial absorbance scans using the Lamm equation:

\[
\frac{\partial c}{\partial t} = D \left[ \frac{\partial^2 c}{\partial r^2} + \frac{1}{r} \frac{\partial c}{\partial r} \right] - s \omega^2 \left[ r \frac{\partial c}{\partial r} + 2c \right]
\]

(Eq 2.5)

with \( c \) the weight concentration of protein, \( D \) the translation diffusion coefficient, \( t \) the time, \( r \) the radius and \( \omega \) the rotor speed. \( D \) is defined by \( RT/N_a f \), with \( R \) the gas constant, \( T \) the temperature, \( N_a \) Avogadro’s number, and \( f \) the frictional coefficient.

The protein under these conditions is affected by a force \( M_p \omega^2 R \), where \( M_p \) is the particle mass, \( \omega \) the rotor speed and \( R \) is the distance from the centre of the rotor in radians per second. A counterforce is applied due to the mass of solvent displaced, \( M_s \omega^2 R \). Therefore the observed, or buoyant, mass of the protein is measured as \( M_b = M_p(1 - \nu p) \), with \( \nu \) the partial specific volume of the protein, and \( p \) the solvent density. Then, the sedimentation coefficient \( s \) is given by:

\[
s = \frac{M_b}{f}
\]

(Eq. 2.6)

The higher the value of \( s \), the faster the protein is sedimenting. Therefore this can be explained by either the buoyant mass increasing, or the frictional ratio decreasing. If the protein aggregates the buoyant mass will increase faster than the
frictional ratio and result in an increased sedimentation coefficient. If the protein does not aggregate and the sedimentation coefficient measured changes, this indicates a change in the frictional ratio which means a change to the protein’s conformation. It follows that an observed decrease in sedimentation coefficient translates to a more extended conformation of the same protein.

2.3.2 Small-angle X-ray scattering (SAXS)

2.3.2.1 Data acquisition

Small angle X-ray scattering (SAXS) is a higher sensitivity solution based method than AUC which can detect conformational shifts and give an indication of the resulting structure, information lacking from the AUC analysis. The sample is placed in the path of a focused X-ray beam which gets scattered by the electrons. The scattered photons are then detected and the intensity recorded as a function of the scattering angle. The recorded scattering data provides information on the rough size and shape of the molecule. A scattering equivalent of Bragg’s law ($\lambda = 2d\sin(\theta)$, $\lambda$ is the wavelength, $d$ the spacing between the planes in the atomic lattice, and $2\theta$ is the angle between the incident ray and diffraction planes) is used for SAXS measurements using the scattered vector $k_s - k_i$ ($k_s = 2\pi/\lambda$, scattered vector; $k_i = 2\pi/\lambda$, incident vector). The scattering vector magnitude is then:

$$Q = \frac{4\pi \sin(\theta)}{\lambda}$$

(Eq. 2.7)

with $2\theta$ the scattering angle; and $Q$ given in units of nm$^{-1}$.

In SAXS the process is elastic and coherent, meaning the scattered beam keeps the same energy as the incident beam but modifies the propagation direction (elastic) and multiple scattered waves can interfere to create a single wave in a specific direction (coherent). The final scattering signal recorded is the sum of the contributions from all pairs of scatterers. The final raw data output is given in intensity versus scattering vector, $I(Q)$ vs $Q$. This is plotted in reciprocal space by integration of the radial average of the scattering pattern about the position of the main beam, and can be transformed to real space as the distance distribution of the point scatterers, $P(r)$. 

48
The scattering intensity is recorded for both the protein in solution and the matching solution free of protein. The solution only scattering data is subtracted from the protein scattering data; and assuming a highly pure mono-disperse sample this scattering is equivalent to the scattering of one single particle in every direction.

The intensity typically falls very quickly with increasing Q values, and it is necessary to collect data from multiple concentrations of the same particle to then merge low Q data from low concentrations and high Q data from higher concentrations. This may not be possible if the protein is very aggregation prone, in which case a single reading at an average concentration (≈ 1-2 mg/mL) can provide sufficient information. A single reading of a SAXS measurement still records scattering data from a large number of proteins, and is the average scattering signal from a short exposure of ≈ 3 seconds during which multiple snapshots are recorded. Intensity at low Q values can be affected by inter-particle distances and are extremely sensitive to aggregation, but also provide information about the size of the protein in a non-aggregated sample. Data from the higher Q regions provide information about the intra-particle distances and correspond to higher order structural information. The intensity is measured at each Q value by the Debye equation:

\[
I(Q) = \sum_p \sum_q f_p f_q \frac{\sin(rQ)}{rQ}
\]

(Eq. 2.8)

with \( f_p \) and \( f_q \) the scattering lengths of electrons at points \( p \) and \( q \) within the protein, at a distance \( r \) from each other.

For a SAXS experiment, high energy X-rays are required due to the low probability of a scattering event between X-rays and electrons. Sources of suitable X-ray beams are relatively rare, with some synchrotron sources in Japan (INS-SOR, Tokyo), France (ESRF, Grenoble) or the U.K. (Diamond Light Source, Harwell science campus).

**2.3.2.2 Data analysis**

As mentioned previously, the raw data first needs to be processed by subtracting the buffer only scattering data from the protein scattering data. Two key aspects to look out for in this step is whether the sample is aggregated or not, and
whether the buffer matches the sample buffer. If the protein is aggregated, the intensity will have a steep upwards curve in the very low Q region, also referred to as the Guinier region. If the buffer does not match perfectly the sample buffer, the curve will slope downwards in the Guinier region. This is illustrated in figure 2.5, with a good sample for reference. If the sample is monomeric and the buffer matches perfectly then the Guinier region should be linear. This linear region is used for the Guinier analysis, which provides a value for the $R_G$ of the protein. This relies on the Guinier approximation which states that in very low Q regions, $I(Q) = I(0) \times \exp\left(-\frac{R_G^2Q^2}{3}\right)$. This can be rewritten as:

$$\ln(I(Q)) = \ln(I(0)) - \frac{R_G^2Q^2}{3}$$ (Eq. 2.9)

which when plotted as $\ln(I(Q))$ vs $Q^2$ gives a linear plot with a slope of $-R_G^2/3$. By fitting this linear curve, the $R_G$ value can be estimated. As an additional test to monitor aggregation it is recommended to check $I(0)/c$, with $c$ the concentration. If this remains constant, the curves obtained at multiple concentrations can be merged. The linear fit to the curve should only be used between values of Q. $R_G$ of 0.5 and 1.5 for the best accuracy (87). It is not unusual to measure a slightly higher $R_G$ value than expected due to the scattering of the hydration layer surrounding the protein. This is a small layer of higher density electrons in the solvent in direct contact with the protein, and is different enough from the bulk electrons to be detected in the scattering signal recorded.

By applying an inverse Fourier transform on the full Q range, the $I(Q)$ data can be converted back into real space (nm). This new curve is the distance distribution function, $P(r)$, and gives a probability distribution of all atoms in the protein by distance. This curve starts at 0, as no two pairs of atoms are at 0 nm distance, then increases to a maximum probability which represents the distance most common between two atoms. After this maximum, it falls back down to 0 beyond a maximum distance value, $D_{max}$, which corresponds to the maximum distance between two atoms. The shape of this curve holds information about the structure of the protein, particularly in multi-domain proteins such as antibodies, or BoNTs. For illustration purposes, a perfect
sphere would have a Gaussian distribution, with a maximum at $D_{\text{max}}/2$, $D_{\text{max}}$ being the diameter of the sphere. A real-space $R_G$ value can be obtained from this curve, which should agree with the Guinier approximation value. The $P(r)$ curve is defined by:

$$P(r) = \frac{1}{2\pi^2} \int_0^\infty I(Q).Q.r.\sin(Qr)\,dQ$$

(Eq 2.10)

An optimal $P(r)$ curve would use infinite $Q$ values, in practice this is constrained by $D_{\text{max}}$.

**Figure 2.5:** Example of a monomeric sample and highly / minimally aggregated sample. Example of over / under subtracted buffer, if the sample and buffer did not match. Images taken from: (88, 89).

A second commonly used transformation of the $I(Q)$ data is known as the Kratky plot (figure 2.6), which is a representation of $Q^2.I(Q)$ vs $Q$. This plot provides information on the flexibility and structure of the protein based on the shape. A globular
folded protein will display a single Gaussian-like peak; an unfolded protein would have a logarithm-like increasing slope. Anything in between those two extremes inform on flexible regions within proteins or multiple domains being responsible for the scattering.

**Figure 2.6:** Example Kratky plot of theoretical proteins. Image taken from: Stanford Synchrotron Radiation Lightsource web-page (at time of writing: https://www-ssrl.slac.stanford.edu/smb-saxs/content/data-analysis-primer).

Finally, theoretical SAXS curves can be generated from MD simulations which can then be directly compared to the experimental curves obtained using an R-factor fitting as described previously. This reinforces the value of SAXS experiments, with direct visualisation possible of matching three-dimensional structures.
2.3.3 Single-molecule fluorescence resonance energy transfer (smFRET)

Fluorescence is a physical property of certain molecules which absorb photons at a particular wavelength and re-emit new photons at an increased wavelength. By absorbing a photon they enter a higher energy state; and when relaxing back to their ground state they release photons at a higher wavelength, as represented in figure 2.7.

![Jablonski diagram representing the process of fluorescence.](image)

**Figure 2.7**: Jablonski diagram representing the process of fluorescence.
Typically, due to the diffraction limit in optical microscopy it is impossible to accurately resolve structures lower than $\lambda/2$, i.e. 200 nm when using far blue wavelengths and up to 350 nm when using far red. Proteins are significantly smaller than this, with even large proteins such as antibodies or BoNTs only reaching 10 nm. Therefore it is not possible to observe any conformational changes directly using standard microscopy. But by using the phenomenon of FRET, this becomes possible.

FRET, or fluorescence resonance energy transfer, is sometimes also referred to as Förster resonance energy transfer after Theodor Förster who developed the technique in the late 1940s. It makes use of the physical properties of fluorescent molecules and relies on the extreme distance dependence of the transfer of energy between a donor and acceptor dye. The donor is excited at an appropriate wavelength, which places it in a higher energy state. As it relaxes back to the ground state, it emits a photon at a wavelength far enough to be detected separately to the excitation beam. This photon in turn excites the acceptor dye, which undergoes the same process. A fourth wavelength photon is emitted by the acceptor and can be detected using appropriate detectors. The efficiency of the energy transfer is inversely proportional to the sixth power of the distance between the two dyes following equation 2.11 (90).

$$E(r) = \frac{R_0^6}{R_0^6 + r^6}$$

(Eq. 2.11)

With $R_0$ the Förster radius, defined as the distance where the efficiency is 50%, and $r$ the distance between the two dyes (91).

By using two dyes with a known Förster radius, FRET can be used as a highly sensitive measurement tool, with minor displacements having a large effect on the transfer efficiency. The optimal distance for FRET measurements is between 2 and 10 nm; with the efficiency above 10 nm falling to levels indistinguishable from background noise in any measurements. The efficiency can also be measured by the ratio of acceptor photons being emitted to the total emission detected by the donor and acceptor dyes (equation 2.12).
\[ E = \frac{F_a}{F_a + F_d} \]  
(Eq. 2.12)

With \( F_a \) and \( F_d \) the intensity detected in the acceptor channel and the donor channel, respectively.

By measuring the intensity of the acceptor and donor channels, a value of the transfer efficiency can be obtained and then matched to a distance by equating equations 2.12 and 2.11.

**Figure 2.8:** Example diagram of a FRET optical setup, with the excitation wavelength slightly lower than the of the donor emission; the overlap between donor emission and acceptor excitation; and the final acceptor emission wavelength.
Other transfer efficiency detection methods are possible (92) such as the recording of photon lifetimes, which are recorded over the duration of the measurement and averaged at the end. If the donor and acceptor dyes are in close proximity, the average lifetime of the donor photons will be shorter, as they are being absorbed faster by the acceptor. If the distance is larger, their lifetime will be longer. The efficiency is then defined as:

\[
E = 1 - \frac{\tau_{DA}}{\tau_D}
\]

(Eq. 2.13)

With \(\tau_{DA}\) and \(\tau_D\) the average lifetime of a photon of the donor channel in the presence and absence of the acceptor, respectively.

In smFRET, the FRET is measured within the protein itself rather than between two proteins. To achieve this, two sites are labelled with respectively a donor and an acceptor dye on the same protein, and concentrations used are low enough (10-100 pM) to only illuminate one single protein in the sample at any given time (93). The efficiency is then determined by recording the photon lifetimes and the intensity in the appropriate channels. The detectors used in a FRET setup do not natively select for wavelengths and detect photons irrespective of their source. Typically, the back-reflected light from the sample passes through a dichroic mirror which separates the excitation beam; and is then focused using a pinhole (confocal microscopy) and guided to the detectors with mirrors. By using a second dichroic mirror it is possible to maximise the reflection of the donor photons onto one detector; while allowing photons from the acceptor channel to leak through to a separate detector. To further select for the donor and acceptor photons, appropriate filters are placed ahead of the detectors to filter out any excess photons which could influence the readings and hence the transfer efficiency value. This is represented schematically in figure 2.9. The detectors are known as avalanche photodiodes (APDs), which have the capacity to detect single photons. Also mounted into the equipment are electronics which record the arrival time of every photon, which can then be used to determine the average lifetime.
The data from smFRET measurements is usually referred to as FRET bursts, due to the low background fluorescence with long photon lifetimes in the absence of a tagged protein. In contrast, when a labelled single molecule passes in the focal point, a burst of high intensity, short lifetime photons are emitted which can be differentiated from the background (94, 95).
Figure 2.9: Schematic representation of the FRET setup, with the different mirrors and filters in place to select for the acceptor and donor photons. Image taken from: (96).
The advantage of smFRET compared to other bulk-based methods is the ability to resolve each individual protein’s signal. This means even a small population of rare conformations can be detected and separated from the main population; whereas this information would be lost in SAXS due to the averaging over all scattering particles present in the sample. The use of confocal microscopy is crucial to the accurate detection of single particles, as it removes out of focus light from the beam which reaches the detectors. This means any scattered light from water molecules or proteins outside of the focus is not recorded while one protein is in the focal point.
Chapter 3

Molecular dynamics simulations show pH mediated solution behaviour of BoNT/E
3.1 Summary

Molecular dynamics are still a relatively recent tool for the study of proteins compared to more established methods such as X-ray crystallography or nuclear magnetic resonance (NMR). However, even in the early days of computation, they were able to accurately recreate known conformational changes in well-studied proteins. Nowadays, thanks to exponentially increasing computational power, they are an invaluable tool in drug design and optimisation. Continued interest in this field and collaboration with computer scientists has led to more accurate force-fields, and the ability to study larger systems on increased timescales. With high throughput testing of multiple proteins and mutations, the theoretical properties of novel compounds can be tested with minimal cost compared to experimental work. From binding studies to stabilising mutations and environmental effects, a broad range of information can be obtained to reduce the downstream experimental analysis required. This leads to fewer target compounds requiring testing and speeds up the drug design process.

In this chapter, we have generated long MD simulations of the BoNT/E protein in explicit solvent, under simulated pH conditions. The range of pH studied encompasses the critical pH 5 barrier required for activity of this protein, with 3 conditions at and below pH 5 and 3 conditions above pH 5. The results display a clear conformational shift for this protein across the threshold and may help in understanding the payload delivery across the membrane, a poorly understood step. Specifically, we show that the protein requires an increase of its apparent size by increasing the separation between the light chain (LC) and the binding domain (BD). This is made possible by the increase of positive charge on both domains. A subtle secondary structure change in the translocation domain (TD) contributes to a membrane insertion mechanism, likely in unison with two other regions of the protein. Finally, we show that a novel conformation of this protein is present exclusively in simulations at a pH lower than 5.5, which may be crucial for membrane insertion by increasing the overall surface exposure of the protein and exposing parts of the TD.
3.2 Introduction

pH-mediated activation of proteins has been observed in a number of systems, perhaps most notably in the close BoNT/E structural homologue tetanus neurotoxin (TeNT) (69). Another relevant example is the diphtheria toxin, of which some structural features are conserved in BoNT/E (72). Previous work has revealed that TeNT exists in a conformation close to that of BoNT/E in neutral pH conditions, and adopts a more extended conformation after crossing the activation pH threshold (71). The sequence and structural homology of TeNT to BoNT/E, especially with respect to the domains’ spatial arrangement, warrants an investigation of BoNT/E. Indeed, while the overall action of BoNT/E is well understood, so far the specific structural changes driving the activation have not been observed; nor have the biochemical properties driving these structural changes.

MD simulations are a relatively recent tool for the study of proteins and are increasingly accurate (97–99). As advances in computational power and development of user friendly tools accelerates, significantly longer simulation times can be achieved on large systems such as BoNTs. The current MD landscape for BoNTs, and BoNT/E in particular, is limited to short simulation times and does not scope out the range of pHs around pH 5 (100, 101). An in-depth analysis with additional pH conditions and longer times may provide insight into the mechanism of action of BoNT/E. Increased simulation times are paramount when attempting to observe significant structural changes in a large protein (102), as the system will have a longer equilibration time which should not be used for analysis. Beyond the increased simulation times, more repeat simulations can be performed in a shorter time. In MD, even more so than in experimental conditions, repeats are critical for an objective look at a system. Despite the starting point being the same in each simulation, the energy minimisation step may result in a lower energy state which will in turn affect how the protein behaves. Building a large array of structures and trajectories from the same starting point allows for statistical analysis and the observation of potential short-lived states.

Previous MD based studies identified on BoNTs have focused primarily on BoNT/A (100, 101, 103) and were limited in their range of pH simulated as well as their simulation times. Typically, for MD simulations to accurately capture full domain
movements they require to reach $10^{-11} - 10^{-8}$ s (102), which many of these studies fall short of. Other works have focused on single domains in isolation, or truncated BoNTs with just two of the three domains (66, 104). These studies also propose answers for questions relating to the catalytic activity or receptor binding of the toxin, but do not offer insight into the specific membrane insertion process prior to LC release. Beyond the particular importance of pH for these toxins, the protonation of precise residues has been shown to increase flexibility in other systems in acidic environments, such as chymotrypsin inhibitor 2 which displays a lower unfolding energy barrier (62, 64).

By performing longer simulations on a broader range of pH studies and using the full, wild-type toxin we have provided a fresh look at the biochemical properties of this toxin and some of the driving mechanisms for its transformation into a transmembrane channel.

The MD simulations presented in this chapter revealed a paradigm shift for BoNT/E in pH conditions above and below pH 5, coinciding with the activation barrier for this protein in vivo. This was first observed by an increased flexibility and apparent size of the protein through raw RMSD and $R_G$ data. The RMSD analysis was then adapted to determine inter- and intra-domain structural changes by relevant residue selections and revealed a translational movement of the BD with respect to the TD+LC domain, with very little change observed in the intra-domain data. In an attempt to explain the separation of the BD from the LC, the charge distribution as a function of pH was obtained with particular focus on the interface of the two domains. This revealed a sudden loss of positive/negative charge interaction when lowering the pH from pH 5.5 to pH 5 and may explain the movement observed.

Additionally, the overall SASA of BoNT/E increased in the lower pH conditions up to a near maximum at pH 4.5, with the steepest increase found between pH 5.5 and pH 4.5. This was in part due to a local secondary structure change of the TD, observed exclusively in pH conditions at and below pH 5. This region is conserved across BoNT serotypes and TeNT, and has been shown to be critical in membrane insertion for BoNT/A (67). Furthermore, this region was found to have high dynamic correlation with two other regions of the toxin with the highest correlation found at pH 5 specifically. Finally, using half of the total structures generated due to computational
limitations, PCA was applied to a single trajectory made up from the full set of
structures generated (6 pH conditions, 4 repeats per pH, 400 ns for each simulation,
resulting in a total of 840,048 structures after removing the 50 ns equilibration time).
This analysis reinforced the pH dependent nature of the structural changes observed,
with one cluster of conformations exclusively obtained from pH conditions at and below
pH 5. This family of conformations has an increased $R_G$ relative to the starting
structure, and leaves the TD more solvent exposed in good agreement with it being
critical for membrane insertion prior to LC release into the cytosol.

3.3 Methods

This section describes the specific protocols followed for the work presented in
this chapter. For a description of the theory behind the methods, see chapter 2.

3.3.1 Molecular dynamic simulation conditions

The starting PDB file for all of the simulations was taken from PDBID: 3FFZ
(22) which is a crystal structure obtained at pH 7.0. Simulations were performed on
the wild-type toxin to stay as close as possible to the real world; while experimental
work was performed on a mutant, non-toxic version of the toxin to remove the highly
constraining safety measures which would have been required with the wild-type toxin.
The mutations present in the toxin used experimentally are located in the buried zinc
binding region of the LC and are unlikely to impact the experimental studies described
herein.

A set of simulations was attempted using a NAMD (105) script for constant pH
(cpH) MD simulations (106). 6 pH conditions were simulated, 4, 4.5, 5, 5.5, 6 & 7, at a
temperature of 300K and pressure of 1 bar. A timestep of 2 fs was used with
coordinates saved every 10 ps. However, the calculations involved and the size of the
system resulted in very slow progress and these simulations were stopped early. The
forcefield used was CHARMM36 (78) and the simulations were performed in explicit
solvent, with a water box using the SPC/E model (107); charge inside the box was
neutralized using Na and Cl counter ions. Only one 100 ns trajectory was obtained
from these simulations for pH 4, 4.5, 5 & 7; and one trajectory each for pH 5.5 & 6 of
85.02 ns & 95.24 ns respectively. The analysis of the cpH trajectories is presented in appendix 1.

A much larger dataset was obtained from MD simulations using gromacs v5.0.4 (108, 109) with pH conditions simulated on the starting structure without adjustment throughout the trajectory. The same 6 pH conditions were tested and simulated using the AMBER03 forcefield (83) and the APBS portal (110), with 4 trajectories per pH generated, each 400 ns long. The temperature was set at 300K with 1 bar pressure, with the protein centred in a cubic box of explicit solvent (SPC/E water model) and placed at least 1 nm from the sides. The system was first energy minimised using a steepest-descent algorithm with maximum 50,000 steps, a step size of 0.01; and the process stopped once the maximum force was < 1000 kJ/mol/nm. The system was temperature equilibrated with an NVT algorithm of 100 ps, a timestep of 2 fs, using the leap-frog integrator. A 1 nm cut-off distance was set for the short range electrostatic and Van de Waals interactions. The temperature was maintained using a modified Berendsen thermostat. The pressure was equilibrated using an NPT algorithm of 1 ns, with a timestep of 2 fs and maintained with Parrinelo-Rahman coupling. Finally, for the unrestrained simulation trajectory, the timestep was kept at 2 fs with coordinates saved every 10 ps. The gromacs generated data and analysis is presented in this chapter.

### 3.3.2 MD trajectories analysis

#### 3.3.2.1 Gromacs

The solvent accessible surface area (SASA) of the protein was calculated using the GMX SASA function (111) of gromacs, which provided a global SASA measurement of the full protein as well as an average SASA value for each residue. For the SASA values, an average was calculated over the final 200 ns of the simulation for each of the 4 repeats per pH; and the average of these values was taken as the final value. Gromacs’ GMX RMSF function was used to calculate residue root mean square fluctuation (RMSF) values, and as for the SASA the average over the last 200 ns was used for each trajectory, and this value was averaged over the 4 repeats per pH. For some of the subsequent analyses (PCA, RMSD, Rd) a backbone only file was created from the main trajectory using GMX TRJCONV.

#### 3.3.2.2 VMD
VMD (112) was used on an external high performance research computer to rapidly manipulate the trajectories. The first step was to combine all 24 trajectories (6 pH conditions, 4 repeats per pH) into a single trajectory while removing the first 50 ns of each trajectory, considered as equilibration time for the size of this system. The protein in this new trajectory was aligned using the translocation domain and light chain (TD+LC) residue selection, so as to observe the relative movement of the binding domain (BD). Then 8 total trajectories were generated: 1 trajectory for each pH combining all 4 repeats; 1 trajectory combining all 24 simulations; 1 reduced trajectory combining all 24 simulations but skipping every other frame. The single trajectory per pH was used for deriving dynamic cross correlation data (using Bio3D package in R, see below); the full trajectory was used for root mean square deviation (RMSD) and radius of gyration ($R_G$) calculations; and the reduced trajectory was used for principal component analysis (PCA) (also using Bio3D in R).

The RMSD data included domain specific analyses and global protein RMSD. Scripts were written to perform these actions using the VMD tool command language (Tcl). First the reference backbone only PDB file was loaded, then the combined trajectory file. Several atom selections were created: 1 for the BD, 1 for the TD+LC, 1 for the “switch” region (details later). For each atom selection a duplicate selection was also created at the fixed frame of reference. A loop over the full trajectory then calculated the RMSD of each selection with respect to itself (intra-domain RMSD) and with respect to its starting point (inter-domain RMSD). The $R_G$ of the full protein was also calculated for each frame. All this data was written to an excel compatible file for further analysis.

### 3.3.2.3 Bio3D Package

The Bio3D package (85) in R was used to perform dynamic cross correlation analysis on the combined pH trajectories, which provided insight into areas of the protein displaying activity in a correlated manner. Using the reduced trajectory (24 trajectories, 50% of total frames) PCA was used to determine if and how the data could be clustered into groups of similar structures. For this, several common R libraries were used and some to help with the large amount of data to treat. Specifically, the
fastcluster library (113) helped to perform the hierarchical clustering on a total of 420,024 frames and the parallel library enabled parallel computing on 36 cores.

3.4 Results and discussion

3.4.1 Whole protein RMSD and $R_G$ analysis reveals a behaviour change in pH $\leq 5$ and pH $> 5$

The first level of analysis was to check the $RMSD$ and $R_G$ of the protein over time, and is presented in figure 3.1. In simulations at and above pH 5.5, the majority of the data remains at a close to constant level throughout the 400 ns and across all repeats. There are only two trajectories displaying a change in $RMSD$, however this change does not translate to a change in $R_G$. These two trajectories will be addressed further, in the PCA analysis. The $R_G$ in these conditions remains at the starting value of 3.60 nm, which corresponds to that of the PDB crystal structure (PDB ID: 3FFZ (22)). By contrast, in simulations below pH 5.5, both the $RMSD$ and $R_G$ data fluctuates significantly in several repeats. In some cases, the $RMSD$ ends on an upward slope, suggesting larger differences from the starting point could have been reached with longer simulation times. Furthermore, the change in $RMSD$ in most cases under these conditions also translates to a change in $R_G$, and specifically an increase to a common final value of 3.82 nm (+6%). The shift in behaviour occurs across the activation pH barrier of these proteins in vivo, suggesting that the change observed here may be directly related to the membrane insertion of this toxin. The behaviour change as a function of pH is made more evident by sorting the above data into bins of 0.05 nm width. This is shown in figure 3.2 and highlights the pH dependence, particularly for the increase in $R_G$. 
Figure 3.1: $R_g$ (top) and RMSD (bottom) over time for each repeat of BoNT/E MD trajectories. For future analyses, only the final 350 ns were used. This shows the variability between repeats for pH 4 to pH 5, implying increased flexibility compared to pHs 5.5 to 7.
Figure 3.2: Data from Figure 3.1 sorted into 0.05 nm width bins, and combining all trajectories for each pH. By accounting for all the repeats per pH into a single plot, the pH dependence of the $R_G$ and $RMSD$ values stands out more clearly.
3.4.2 Domain specific \textit{RMSD} analysis shows that the binding domain separates from the light chain

In order to establish what the contributing factors to the increases in \textit{RMSD} and \( R_G \) were, atom selections covering domains were used. One selection comprised the light chain and the translocation domain (TD+LC), the other the binding domain (BD). Then the \textit{RMSD} of these two selections was measured by aligning them to the reference frame (time 0) so as to measure the intra-domain \textit{RMSD}. This provided information about whether the domains were unfolding or kept their structural integrity throughout. Then, the \textit{RMSD} of both selections were measured while aligning the opposite selection. This measured the \textit{RMSD} of the selection with respect to its starting position, i.e. how far from its initial position the BD (or TD+LC) had moved while the TD+LC (or BD) remains fixed. The data was then sorted into bins and is presented in figure 3.3. The results show that each domain remains structurally sound throughout the simulations, with the intra-domain \textit{RMSD} never exceeding 0.5 nm in any pH condition. When looking at the inter-domain \textit{RMSD} however, there is a paradigm shift at and below pH 5. Under pH conditions higher than pH 5, the inter-domain \textit{RMSD} remains low, while the higher value bins are populated by simulations at and below pH 5. This indicates that the increase in global protein \textit{RMSD} and \( R_G \) is due to a translational movement of the BD with respect to the TD+LC rather than the domains losing their secondary structure. A separation of the BD and the LC would result in an increase in SASA and, more specifically, expose new areas of the TD. This is in good agreement with the pH mediated activity of this protein (and its homologues BoNT/A and TeNT), which is believed to form a trans-membrane channel with the TD (65, 114, 115), through which the LC is released into the cytosol.
Figure 3.3: Domain specific RMSD analysis, sorted into 0.05 nm width bins. (a) Intra-domain BD RMSD. (b) Intra-domain TD+LC RMSD. (c) Inter-domain RMSD with mobile BD and fixed TD+LC. (d) Inter-domain RMSD with mobile TD+LC and fixed BD. This representation shows that the domains remain stable throughout the simulations, in all pH conditions. The RMSD observed previously comes from inter-domain movements.
3.4.3 Charge and salt-bridge content plays a crucial role in the activation of BoNT/E

With a clearer picture of the contributions to the $R_{MSD}$ and $R_g$ increases in low pH simulations, the question of the underlying causes for such a behavioural shift across the pH 5 barrier arose. The main differences of the protein setup for different simulated pH conditions were protonation states of ionizable groups of amino acids resulting in different charges of the protein under each pH condition. The total charge on the protein was higher at low pHs, and lower in high pH conditions. By plotting the charge as a function of pH (figure 3.4) it became apparent that the rate of change of the charge per pH on each atom selection (BD and TD+LC) was significantly higher between pH 5.5 and pH 4.5; matching the activation barrier of this toxin. Additionally, pH 5.5 is the first pH (from high to low) where the charge on both atom selections is positive. Therefore, the subsequent increase of positive charge on both selections as the pH is lowered may also increase electrostatic repulsion between the BD and the LC. To verify this theory, the charge locations were visualised in PyMOL and two interaction sites were observed at the interface between the LC and the BD. On both these sites (figure 3.5) there is a loss of positive/negative charge interaction between pH 5.5 and pH 5, suggesting that these two sites play a role in the observed separation of the BD from the LC. With the lower pH conditions also playing a role in salt-bridge breaking, the quantity of salt-bridges present in the starting structure per pH was plotted (figure 3.4). This revealed the same pattern as the charge plot: a steep loss of salt bridges between pH 5.5 and pH 4.5; while the quantity is constant between pH 4.5 and pH 4 and between pH 5.5 and pH 7. These two factors combined are therefore believed to play a role in the activity of this toxin and explain the differences in the data presented above across the pH 5 barrier.
Figure 3.4: Charge and salt-bridge distribution as a function of pH on BoNT/E. This reveals that the steepest change in charge and loss of salt-bridges is found between pH 5.5 and pH 4.5, coinciding with the activation pH of this protein. Specifically, the charge on both atom selections is positive starting at pH 5.5, which may explain the increased flexibility observed below pH 5.5.
Figure 3.5: Interaction sites detected at the interface between LC and BD of BoNT/E. At the transition between pH 5.5 and pH 5, on both sites, a positive/negative charge interaction is lost. This is suggested to contribute to the increased flexibility of this protein below its activation pH. Protein is shown in cartoon representation, with blue and red residues depicting negative and positive charges, respectively.
3.4.4 SASA analysis shows low pH conditions increase BoNT/E’s solvent exposure

Now with a strong understanding of the underlying causes for the jump-starting of this toxin’s activity (the charge at the BD/LC interface, combined with the loss of salt bridges) a closer look at the more specific effect of the pH on the structure was required. In particular, the effect on the SASA of both the full protein and individual residues due to the pH environment was studied. Due to the nature of the activity of this protein, forming a transmembrane channel, any effect on the SASA may be revealing of a mechanism aiding the membrane insertion. Firstly, the overall SASA of the protein displayed a similar trend to the charge and salt-bridge content. That is to say, the highest rate of change of SASA as a function of pH was observed between pH 5.5 and pH 4.5 (figure 3.6a). The SASA increased as the pH lowered, in good agreement with the above RMSD and $R_G$ data. An increase in SASA also means a higher probability for areas of the protein interacting with the membrane and triggering the membrane-insertion.

Broken down by domain (figure 3.6b), the most consistent increase with lowering pH was observed in the BD, which is consistent with the separation of it from the LC. The TD showed an increase between pH 7 and pH 5.5, a drop at pH 5 and a subsequent increase down to pH 4. This reflected the pH 5 simulations where some repeats showed less movement than others. Longer simulations for each condition may have resulted in a more consistent pattern. Finally, the LC remained at a close to constant SASA due to its larger exposure on the opposite face relative to the BD.
Figure 3.6: Global and domain specific SASA of BoNT/E as a function of pH. (a) Global SASA of BoNT/E as a function of pH, shows an increase as the pH is lowered. The fastest rate of change is seen between pH 5.5 and 4.5. (b) Domain specific SASA revealing the increased exposure of the BD due to the separation from the LC. The TD average SASA fluctuated due to the variability within each repeat. The LC maintained a near constant SASA throughout.
Further breaking down this data to individual residues, some patches of higher SASA were discovered when comparing each pH’s values to pH 7. The SASA difference of the TD residues with respect to pH 7 is shown on the structure in figure 3.7 and reveals a small helical region with a consistent increase in SASA between pH 4 and pH 5; while the same region remains neutral compared to pH 7 at pH 5.5 and pH 6. After taking a closer look at this region during the simulations, it was found that the alpha helix unfolds in simulations at pH lower than 5.5, exposing a lysine residue (LYS618). This region was found to be the homologous region of BoNT/E to a region in BoNT/A previously shown to undergo a similar structural change in acidic pH (67). This structural change in BoNT/A proved to be crucial for membrane insertion, and hence for activity. This region is also highly conserved in other serotypes of BoNTs and in TeNT.
Figure 3.7: TD SASA difference with respect to pH 7, with residues coloured according to scale on the left. (a) Scale for the residue colouring, and cartoon representation of the TD of BoNT/E from two angles. (b) “Switch” region as found in the PDB structure and maintained in simulations at pH 5.5 to 7. (c) “Switch” region in an unfolded state observed exclusively in simulations at pH 4 to 5, with an increased SASA with respect to pH 7.
3.4.5 RMSF & cross-correlation analyses reveal a concerted action of multiple regions for membrane interaction at pH 5

Figure 3.8: Comparison of RMSF averages per pH, in terms of their difference with respect to pH 7. Two areas stand out that have high flexibility: the “switch” region and the end of the BD, responsible for the initial membrane binding event.

A measure of protein flexibility can also be obtained by measuring the RMSF of the protein. This gives an average per residue, and highlights dynamic areas within the protein. By measuring the RMSF over the final 200 ns of each repeat, and averaging over the four repeats, the average difference in RMSF per residue with respect to pH 7 for every pH was calculated. This is represented in figure 3.8, and highlights 2 major areas of interest. The “switch” region and the residues on the BD involved in the initial membrane binding event have the highest RMSF at pHs 4 to 5, and remains stable above pH 5.

After identifying these regions of interest, the next step was to identify whether they act independently towards the membrane insertion or if they act together, potentially with other regions of the protein. To that end, and using a combined trajectory with all 4 repeats for each pH, dynamic cross correlation matrices were
generated. Particular focus was brought to previously identified regions of interest and how they correlated with the rest of the protein. These regions were a selection of BD residues involved in the original neuron membrane binding, the “switch” region and a section of the TD identified by Kumaran et al (22) using TmPred (116, 117) as a potential membrane insertion region. The results showed that for all combinations of these three regions (“switch” vs membrane insertion, “switch” vs BD residues, BD residues vs membrane insertion) the correlation was among the highest of the full protein; and specifically, their correlation was the highest at pH 5 (figure 3.9). This may point to a concerted action of three spatially separated regions acting together at pH 5 to create the transmembrane channel, crucial for the toxin’s activity. Furthermore, these regions also displayed some of the highest RMSF values of the whole protein and in particular had higher fluctuations at pH 5 than at any other pH.
Figure 3.9: Cross correlation data of three regions of interest in BoNT/E. Cross correlation data for the “switch” region, the membrane insertion region identified by Tmpred and a region of the BD involved in membrane binding. This shows that these three regions have the highest correlation at or close to pH 5, suggesting a common
action for membrane insertion. These go from pH 4 (top left) to pH 7 (bottom right), from left to right.

3.4.6 **PCA analysis uncovers a novel conformation exclusively observed at pH $\leq 5$**

The final analysis performed on these MD trajectories was principal component analysis (PCA). The aim of this was to establish whether the structural changes observed in the simulations were merely random with similar $RMSD$ and $R_G$ values or if they could be clustered into a small number of groups with similar structures. For this method it was important that all 24 trajectories shared the same basis for the XYZ coordinates, and so the protein was aligned across all trajectories based on the TD+LC selection. This ensured that any variance found in the trajectories was due to the movement of the BD relative to the TD+LC selection. A first search algorithm was used to estimate the optimal number of clusters to use for the PCA analysis. This showed that 4 clusters were enough to capture the major conformational differences between all the frames, as represented by the flattening of the “elbow curve” (figure 3.10). Next, the 420,024 frames were separated into 4 clusters.
Figure 3.10: Elbow plot of the PCA analysis. Used to determine the optimal number of clusters to use. This suggests that 4 clusters are enough to account for most of the variance of the trajectory.
Different clustering methods were available, and six of these were tested to find the most appropriate one. These were “single”, “complete”, “average”, “McQuitty”, “median” and “centroid”. The “single” method was ruled out as this adopts a “friends of friends” approach, meaning that 100% of the data was placed into a single cluster. This is because the timestep of 20 ps (after skipping every other frame) between each frame did not produce a large enough difference of structure, and thus each frame was placed into the same cluster as the previous frame. The “median” method appeared to cluster frames based solely on the value of \( \text{RMSD} \) with respect to the starting structure, leading to some significantly different structures being clustered together due to their similar \( \text{RMSD} \) values. The “centroid” method was also excluded as it created clusters with as little as 20 frames. Given the nature of MD simulations and the 20 ps timestep between each frame, this was considered extremely unlikely to be representative of the data. Finally, the “complete”, “McQuitty” and “average” methods yielded similar results, with large clusters and robust clustering. The robustness of the clustering was tested by measuring the average \( \text{RMSD} \) value of every frame in a cluster to the cluster “mid-point”, the most representative structure of that cluster. The “average” clustering method was found to have the lowest \( \text{RMSD} \) values per cluster, meaning the structures in each cluster were closer to each other than when using any other clustering algorithm. The average \( \text{RMSD} \) of each cluster was then calculated with respect to the starting structure, and the average \( R_G \) of each cluster was also calculated. This is shown in table 3.1 and confirms that each cluster is only comprised of closely related structures with a low \( \text{RMSD} \) to their mid-points; and low standard deviations in the average \( R_G \) and the average \( \text{RMSD} \) to the starting structure. Table 3.1 also shows the proportion of each cluster found per pH.
Table 1: Summary of clustering data.

<table>
<thead>
<tr>
<th>% total frames per pH</th>
<th>Cluster 1</th>
<th>Cluster 2</th>
<th>Cluster 3</th>
<th>Cluster 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>50.03</td>
<td>49.92</td>
<td>0.05</td>
<td>0.00</td>
</tr>
<tr>
<td>4.5</td>
<td>54.80</td>
<td>41.10</td>
<td>0.86</td>
<td>3.24</td>
</tr>
<tr>
<td>5</td>
<td>52.13</td>
<td>23.57</td>
<td>23.44</td>
<td>0.86</td>
</tr>
<tr>
<td>5.5</td>
<td>75.36</td>
<td>2.00</td>
<td>0.64</td>
<td>22.00</td>
</tr>
<tr>
<td>6</td>
<td>75.14</td>
<td>0.27</td>
<td>24.59</td>
<td>0.00</td>
</tr>
<tr>
<td>7</td>
<td>97.30</td>
<td>0.00</td>
<td>2.70</td>
<td>0.00</td>
</tr>
<tr>
<td>RMSD from centroid</td>
<td>0.37 ± 0.08</td>
<td>0.53 ± 0.14</td>
<td>0.34 ± 0.16</td>
<td>0.30 ± 0.09</td>
</tr>
<tr>
<td>RMSD from reference</td>
<td>0.33 ± 0.06</td>
<td>0.63 ± 0.13</td>
<td>0.63 ± 0.13</td>
<td>0.75 ± 0.14</td>
</tr>
<tr>
<td>RG</td>
<td>3.58 ± 0.04</td>
<td>3.76 ± 0.07</td>
<td>3.59 ± 0.02</td>
<td>3.61 ± 0.03</td>
</tr>
</tbody>
</table>
Figure 3.11: PCA data breakdown shown as stacked histograms. (a) This shows how much (in percentage) each pH condition contributed to the clusters. Cluster 1 is the starting point of the simulations and is present in every condition. Cluster 2 is exclusively found in simulations at and below pH 5. N represents the total number of frames in each cluster. (b) This shows how much each cluster contributes (in percentage) to the pH conditions. In pH 7, >90% of frames are in cluster 1. In pH 4, approximately 50% of the frames are in cluster 1 and 50% in cluster 2.
For each cluster, the contribution of the different pH simulations was calculated; and for each pH, the contribution of the different clusters. This is shown in figure 3.11 and revealed that one cluster, termed cluster 2 here, was exclusively present in simulations at and below pH 5. This cluster gained in prevalence the lower the pH, climbing from 25% of all frames in pH 5 to 50% in pH 4. Looking at the cluster mid-points (figure 3.12) it can be seen that cluster 1 is the starting structure of all the simulations, and therefore explains its dominance over all the others. Clusters 3 and 4 are rare events that occurred in only one repeat of pHs 5 & 6 and pH 5.5, respectively. These may be transient states or outliers. Cluster 2, however, is a novel conformation of BoNT/E with an increased $R_G$ and the BD separated outwards from the LC, therefore increasing its SASA and that of the TD. This novel conformation may be a crucial step in the formation of the transmembrane (TM) channel required for activity of this protein.
**Figure 3.12:** Cluster mid-points shown with percentage contribution to total frames. (a) Shows the percentage of total frames each cluster represents. (b) shows the most representative structure of each cluster. Cluster 1 (black) is the starting point of each of the simulations, and dominates the structure distribution. Cluster 2 is the 2nd largest cluster, despite being exclusively present in pH conditions lower than 5.5. This indicates clusters 3 and 4 may be outliers or transient states of the protein. (c) Schematic representation of the clusters as seen from a counter clockwise 90° rotation.
3.5 Conclusions

The raw \textit{RMSD} and \textit{R}_G\ data already shows a difference between the simulations from pH 7 to pH 5.5 and those from pH 5 to pH 4, revealing a pH mediated change in behaviour of this protein in good accordance with the knowledge of its activation at pH 5. By combining all the repeats and sorting the data into bins of width 0.05 nm, the difference is made clearer still while highlighting the lack of increase in \textit{R}_G\ of the few repeats in pH > 5 where the \textit{RMSD} increased. This suggests that an increase in \textit{R}_G\ is necessary for activity, and only happens below the pH 5.5 threshold. This increase is suggested to be due to a translational movement of the BD with respect to the LC, as shown by the domain specific \textit{RMSD} analysis, in good agreement with the previous work on the structurally similar TeNT (71). The driving force of this motion is the sudden increase in charge on both domains, with two sites at the interface losing a negative/positive charge interaction which keeps the two domains together. Combined with a general loss of salt bridges, the protein is able to explore more conformations and has increased flexibility. The extended conformation results in a general increase to \textit{SASA} which contributes to the membrane interaction, while more localised structural changes may facilitate membrane insertion. Specifically, the “switch” region undergoes a loss of helicity and exposes a lysine residue, which will seek to bury itself in a membrane. This region is conserved across all serotypes of BoNT and in TeNT, and has been shown to be crucial for membrane insertion in BoNT/A (67). It is possible that several other regions of the protein aide the membrane interaction, as shown by highly correlated movement, especially at pH 5, between the “switch” and regions within the BD and in a region identified as likely to play a role in membrane interactions (22). Finally, the sum total of all frames generated can be divided into just 4 clusters of closely related conformations, with one cluster exclusively observed in simulations at and below pH 5. This conformation may be of paramount importance in the transmembrane functionality of this toxin and hence in its activity.
Chapter 4

Analytical ultra-centrifugation and small-angle X-ray scattering of BoNT/E confirm the main results from the MD simulations
4.1 Summary

Fixed imaging methods for proteins have provided invaluable insight into understanding the structure/function relationship of proteins, and comprise the vast majority of known structures to this day. However, they are inherently imperfect due to their inability to detect protein dynamics and typically have poor resolution for flexible regions within proteins. Furthermore, molecules studied in fixed methods are, by definition, fixed in a conformation. In the case of crystallography, by far the highest contributor to protein structures, this conformation is guided by the crystallisation process. It is therefore possible that if a mixed population of protein conformations exists in solution, only one will favour crystallisation and effectively remove the other population from the analysis. The crystallisation process may also coerce the protein into a non-native state which would not naturally occur in solution conditions.

To complement the data acquired through fixed methods, it is important to also observe proteins’ behaviours in a solution environment. In particular, small angle X-ray scattering (SAXS) provides a highly sensitive method to detect protein dynamics where a relatively small conformational change can result in a visible difference in the scattering properties of the protein. Combined with an increasing amount of bioinformatics tools to predict biophysical properties of proteins based on sequence and PDB structure, SAXS has become a valuable method to further our understanding of protein behaviours under solution conditions.

In this chapter, we have combined a selection of solution methods applied to BoNT/E in the same six pH conditions used for the MD simulation trajectories. Firstly, using mass spectrometry (MS) to confirm the correct expression of the endopeptidase-negative BoNT/E. Then using circular dichroism (CD) to ensure the expressed protein displayed the expected secondary structure features, by comparing the signal to a theoretical signal determined by PDBMD2CD while also monitoring the toxin’s stability in a range of buffers at the selected pH values. Finally, we subjected BoNT/E to analytical ultra-centrifugation (AUC) and SAXS to assess the validity of the MD results presented previously. Namely, whether the protein is observed in experimental conditions to behave similarly to the in silico simulations. The results of all three methods point to the same conclusion: the MD simulations are representative of
experimental observations, with no secondary structure loss observed while a novel conformation with an increased $R_G$ is detected in pH conditions lower than 5.5.

4.2 Introduction

Currently, studies on BoNTs have been focused primarily on one or two out of the three domains comprising the full protein (26, 27). Some studies also focus on identifying interaction sites by imaging individual domains in complex with potential targets (25, 28, 118); or focus on the complexes formed with the toxin prior to entering the bloodstream (119). This is in part due to the difficulty in crystallising or expressing the full length BoNT. Moreover, the majority of these studies are performed in fixed conditions such as crystallography or cryo-EM (61). With the participation of Ipsen, we had access to an endonegative mutant of BoNT/E for over-expression in E.coli cells. This facilitated the study of the full toxin, and a purification procedure is available with no affinity tags.

While the data obtained by fixed methods have helped identify parts of the mechanism of action of these toxins, the dynamics of it are still elusive. In particular, the effect of pH has not been extensively studied using fixed methods, and structures have been primarily obtained at neutral pH values. Using recent advances in computational tools in conjunction with solution based methods, we are now able to directly compare features such as secondary structure and solution conformations obtained by CD and SAXS to theoretical data obtained from MD trajectories.

Here, we present a combination of solution-based methods applied to BoNT/E to assess the structural changes occurring under a range of pH conditions. CD informs on the secondary structure changes of the protein and therefore explores BoNT/E’s stability in the buffers selected for study. Analytical ultra-centrifugation reports on the monomeric or aggregated state of the sample, while providing a first look at potential conformational changes. Small-angle X-ray scattering, the highest sensitivity method presented here, combined with bioinformatic tools delivers a direct comparison to the MD results obtained for BoNT/E. The conclusions of all three experimental tools corroborate the computational findings. No major loss of secondary structure was observed in the CD, while AUC data inferred that a more extended conformation
existed in solution in pH environments lower than 5.5. Finally, the SAXS data identified two distinct populations of BoNT/E with one observed in pH 4. The two populations found very good matches in the MD trajectories, corresponding to cluster 1 (i.e. the closed, crystal structure) and cluster 2 (i.e. a novel, extended conformation). This confirms the validity of the crystal structure in all the pH conditions observed, but also identifies and validates the novel extended conformation of BoNT/E seen exclusively in MD simulations at a pH value lower than 5.5.

4.3 Methods

This section describes the specific protocols followed for the work presented in this chapter. For a description of the theory behind the methods, see chapter 2.

4.3.1 Protein expression

A plasmid containing an endopeptidase-negative (i.e. non-toxic) BoNT/E DNA sequence was provided by Ipsen; with mutations in the zinc binding region of the toxin. These mutations are unlikely to affect the pH-based properties of the protein as they are buried inside the LC; hence the MD work was performed on the wild-type to stay as close as possible to the real-world. The endopeptidase-negative toxin used experimentally allowed the work to be undertaken with no cumbersome safety measures in place. The plasmid was transformed into E. coli BL21(DE3) (New England Biolabs) cells, and grown in modified terrific broth (mTB), using Kanamycin at 30 μg/mL. An overnight starter culture was used to inoculate larger cultures, which were grown at 37°C until OD600 reached 0.6. The temperature was lowered to 16°C and once cooled, 1 mM IPTG was added for induction. The induced culture was incubated for a further 21 hours, after which the cells were harvested by centrifugation at 3,196 RCF. The pellets were re-suspended with 3 mL lysis buffer (35 mM NaCl, 50 mM Tris, pH 8.0) per gram of cell paste. The cells were lysed by sonication with 10 cycles of 30 second bursts, 30 second rest. The lysed mixture was centrifuged at 19,802 RCF for 30 minutes. The supernatant was then diluted with an equal volume of 2 M ammonium sulfate, 50 mM Tris, pH 8.0; and centrifuged again at 19,802 RCF for 30 minutes. A Butyl column (hydrophobic interaction) equilibrated with 1 M ammonium sulfate, 50 mM Tris, pH 8.0 was used to capture the protein, using a 50
mM Tris, pH 8.0 elution buffer. The relevant elution fractions were pooled and desalted using a 53 mL desalting column, exchanging into 10 mM Tris, pH 8.0. The desalted sample was loaded onto a Q column (ion exchange), and eluted using 500 mM NaCl, 10 mM Tris, pH 8.0. A high purity fraction of this elution was isolated and concentrated/diluted to 0.5 mg/mL. It was then pooled with 5 μg Trypsin per mg protein at 37°C for 40 mins, to cleave the link between LC and HC (“activation” step). The activated sample was then mixed with 1.6X its volume of 5 M NaCl, 10 mM Tris, pH 8.0 and loaded onto a Butyl column equilibrated with 3 M NaCl, 10 mM Tris, pH 8.0. The final product was highly pure, activated (but still safe, endopeptidase-negative) BoNT/E and was eluted using 10 mM Tris, pH 8.0. The purity was assessed by SDS-PAGE, using 4-12% Bis-tris gels (Thermo-Fisher) and the activation was assessed by adding β-Mercapto-Ethanol in one of the lanes, which separated the band at 150 kDa into two bands at 100 and 50 kDa, respectively.

4.3.2 Protein characterisation

The expected properties of the mutant version of BoNT/E were calculated from the sequence using the ProtParam tool of ExPasy. This showed a pI of 6.07; and an expected MW prior to activation of 143,531.75 Da; 143,219.5375 Da after activation.

After purification, the protein was characterised using first order biophysical methods to ensure no extra residues were cleaved during the activation step, and the structure observed matched the expected structure of BoNT/E. The protein was prepared for MS analysis by buffer exchanging into a volatile buffer, 50 mM ammonium acetate. This was achieved using Zeba™ spin desalting columns, 40K MWCO, 2 mL. The column was first spun to remove the storage buffer, then spun 3 times with the ammonium acetate buffer and finally the sample was loaded. After the final spin, the sample was collected in the new buffer, and impurities lower than 40K molecular weight (MW) were removed. The samples were analysed on the Agilent 6510 QTOF LC-MS system in UCL Chemistry Mass Spectrometry Facility. 10 μL of each sample (at ≈ 0.2 mg/mL of protein) was injected onto a PLRP-S, 1000A, 8 μM, 150 mm x 2.1 mm column, which was maintained at 60 °C. The separation was achieved using mobile phase A (water with 0.1% formic acid) and B (acetonitrile, with 0.1% formic acid) using a gradient elution at the flow rate 0.3 mL/min. The column effluent was
continuously electrosprayed into the capillary electrospray ionisation (ESI) source of the Agilent 6510 QTOF mass spectrometer. ESI mass spectra were acquired in positive ESI mode at the m/z range 1,000–3,200 in profile mode. The raw data was converted to zero charge mass spectra using maximum entropy deconvolution algorithm within the MassHunter software version B.07.00.

In order to check the structural integrity of the final purified sample, it was loaded in a CD machine at different pH levels, and compared to the theoretical CD signal obtained via PDBMD2CD (59). The purified protein sample was buffer exchanged into six different buffers, at six different pH values, using 10 mM sodium acetate for pH 4, 4.5, 5, and 5.5, and 10 mM sodium phosphate for pH 6 and 7. All the buffers were made at 20 mM ionic strength using NaCl. The final purified sample was diluted or concentrated to ≈ 0.2 mg/mL and separated into 0.5 mL aliquots. Each aliquot was then buffer exchanged against 500 mL of the appropriate buffer in a cold-room overnight, using 0.5 mL Slide-A-Lyser G2 dialysis cassettes (Thermo-Fisher) with a 20K MWCO. After dialysis, the protein samples were loaded into a Hellma® quartz cuvette, 1 mm pathlength and placed in the CD equipment. 3 scans were recorded between 260 and 180 nm at every nm at 20°C. This was repeated for the matching buffers without protein for later subtraction. The average of the 3 scans was used as the final data, which was then converted from mdeg to molar ellipticity using the equation: \([\theta] = (\text{mdeg} \times M) / (10 \times L \times C)\); where mdeg is the raw CD output, M is the average molecular weight in g/mol, L is the pathlength in cm and C is the protein concentration in mg/mL.

4.3.3 AUC data acquisition

The sample was dialysed into 6 different pH conditions as described above. Then 450 μL of sample at 0.5 mg/mL was loaded alongside 500 μL of matching buffer in the AUC cells. The density and viscosity of the buffers were estimated using SEDNTERP (120), the partial specific volume of the protein was estimated using SLUV (121, 122). Sedimentation curves were fitted using DCDT+ by John Philo, version 2.5.1 (123). AUC data was obtained using a Beckman Coulter Optima AUC, with a spin speed of 40,000 RPM at 20°C.

4.3.3 SAXS data acquisition
Samples were buffer exchanged as described above, then concentrated to ≈ 4 mg/mL using spin concentrators immediately prior to being loaded onto a 96-well plate for auto-sampling into a SEC-HPLC (KW 403, SHODEX) with buffer-matched mobile phase to each sample. The SEC-HPLC then fed the sample directly into a temperature-controlled quartz cell capillary (20°C) with a diameter of 1.5 mm of X-ray scattering instrument B21 at the Diamond Light Source (Rutherford Appleton Laboratory, Didcot, UK), operating with a ring energy of 3 GeV, and an operational energy of 12.4 keV. This recorded 3 second snapshots for 32 minutes, while the sample was loaded and eluted off the SEC. The start of the elution, corresponding to buffer, was subtracted from the sample signal. Buffer subtraction was performed on ScÅtter 3 (BIOSIS) (124, 125); and the generated scattering curves saved for analysis. Theoretical scattering curves were generated from the MD structures using the SASCALC module on the SASSIE web-server (57) in order to find a structure which fit the experimental curve best. The raw trajectories from Gromacs were converted to .dcd format using VMD, and loaded onto the SASSIE server with a reference PDB at the relevant pH. For analysis, the best fit structure was assessed by R-factor. A hydration shell around the protein was included using the explicit solvent of the trajectories for the best-fit frame obtained by the full-atom trajectory analysis.

To facilitate comparison between all the experimental curves and the theoretical curves generated using SASSIE, the experimental data was interpolated to a fixed number of Q points using a MATLAB script. To ensure a good fit to the true data, 800 Q points were used as this was close to the actual number of Q values collected experimentally. All the intensities of every sample were then normalised by dividing by their intensity at Q=0.05. This maintained the shape of the curve, while allowing for comparison between each condition. Analysis of ln(I(Q)) vs Q curves and Kratky plots provided information about the approximate size and shape of the protein.

4.4 Results and discussion

4.4.1 BoNT/E Expression

As described in the methods, a plasmid was provided by IPSEN with the endopeptidase-negative BoNT/E sequence and kanamycin resistance. This plasmid
was transformed into BL21(DE3) *E.coli* cells for expression using the T7 promoter, induced by 1mM IPTG. After expressing for 21 hours at 16°C the cells were harvested and resuspended, then sonicated and prepared for the chromatography purification steps. Purification was performed as described in the methods, and a representative summary gel is shown in figure 4.1. A fraction from each step of the purification is shown in the lanes from left to right: from the first butyl column loading to the final purified fraction, shown non-reduced and reduced to confirm complete activation of the protein.

![Figure 4.1: SDS-PAGE of a purification procedure for BoNT/E.](image)

**4.4.2 BoNT/E Characterisation**

After purification and confirmation of activation, the protein was further investigated to confirm with certainty that no extra residues were cleaved during the activation step and that the structure was conserved and matched the expected structure. Three fractions of the final purified protein were therefore used for MS analysis, and were prepared as described in the methods. The results are shown in figure 4.2, and confirm the correct MW of BoNT/E to within 6 Da (expected from sequence: 143,219 Da; experimental: 143,225 Da). This also confirmed the high purity
of the sample, with only a small fraction of lower MW contaminants relative to the main BoNT/E mass peak. Therefore, the protein was ready for the next steps of the analysis.

![Mass spectrometry readings of the final purified fraction of BoNT/E.](image)

**Figure 4.2:** Mass spectrometry readings of the final purified fraction of BoNT/E. Reading obtained in 50 mM ammonium acetate at 0.2 mg/mL.

Before higher order analyses could be undertaken, the structural integrity of the protein needed to be confirmed. This was tested using CD under the different pH conditions wishing to be later explored, in order to ascertain the stability of the protein in the planned buffers. These were sodium acetate for pH 4, 4.5, 5 & 5.5, and sodium phosphate for pH 6 & 7. The protein samples were prepared for CD as described in the methods, and using the average of 3 scans per sample and per buffer, the data was subtracted then converted to the concentration independent unit of molar ellipticity. This provided a direct comparison of the structure in the different pH conditions, unaffected by differences in the raw signal caused by concentration effects.
The results are shown in figure 4.3, alongside a theoretical CD curve obtained from the crystal structure of BoNT/E used for starting the MD simulations (PDB ID: 3FFZ, (22)). The curves match the expected curve as predicted by PDBMD2CD for every pH studied and support two conclusions. Firstly, BoNT/E is stable in all of the buffers tested at room temperature; and secondly, its secondary structure remains unchanged in these conditions. This matches the MD simulations results, where the only change detected was the loss of a short alpha helix (the “switch” region) in pH conditions lower than 5. Such a small change on a system of this size would not cause a detectable signal modification in CD experiments. Therefore, the buffers tested here, and in the conditions tested (i.e. 20°C, atmospheric pressure) are suitable for the study of BoNT/E and are unlikely to result in protein degradation or aggregation.
Figure 4.3: Top: Comparison of theoretical CD signal obtained from the PDB structure (PDB ID: 3FFZ) and the experimental CD signal converted to molar ellipticity. The discrepancy is due to the arbitrary units of the simulated CD signal; but the overall shape of the curve is maintained with the signature minima at 222 and 208 nm,
indicating high α-helical content. **Bottom:** Comparison of molar ellipticity of BoNT/E as a function of pH. No clear correlation was identified, meaning BoNT/E does not rely on major secondary structure modifications for activity.

4.4.3 **AUC reveals a more extended conformation at pH ≤ 5 and a more compact conformation at pH > 5**

AUC is a solution method which can be used to determine aggregation of proteins over time, but can also be used to detect conformational changes in proteins. This is due to the high RPMs used to sediment proteins, which produce large drag resistances. A protein in a compact sphere-like conformation will have lower drag resistance than the same protein in a more extended conformation. Hence, the same protein will sediment faster in the compact conformation than the extended, but will display the same buoyant MW. In order to compare the sedimentation rates under different buffer conditions, the raw data is converted to the neutral unit of s(20, w) which is the sedimentation if the molecule was placed in water at 20°C. This conversion requires knowledge of the buffers’ viscosities and densities, which can be estimated or measured. In this research they have been estimated using SEDNTERP (120).

The AUC data obtained from BoNT/E is shown in figure 4.4 and confirms the trend identified by the MD previously, in which the protein explores conformations with increased apparent size ($R_g$) in lower pH values. This can be seen by the lower s(20, w) values at pH environments lower than 5.5, indicating an increase to the drag resistance. The sedimentation over time of the protein remains constant, which means BoNT/E is not aggregating under these conditions. Due to repeated equipment failures, only one repeat out of the four attempts produced reliable data. Other repeats suffered from leaks of the sample cell and obstructions in front of the AUC detector which resulted in unreliable data.
4.4.4 SAXS confirms the extended conformation at low pH, and matches structures observed in MD simulations at the same pH values

While the AUC analysis appears to confirm the MD data presented previously, it is a low sensitivity method and relies on accurate determinations of the buffer viscosities and densities and of the proteins’ partial specific volume. All of these values were estimated using software predictions and may not be totally accurate. For a more robust test of the MD data, BoNT/E was subjected to SAXS, a higher sensitivity method, and independent of user estimations. As for the AUC, the protein was placed into 6 different pH conditions using dialysis. However, due to the use of SEC-HPLC, the protein was concentrated to ≈ 4 mg/mL immediately prior to loading onto the
system. The increased concentration ensured a strong enough signal out to large Q values, because of the dilution effect of the column.

![HPLC trace](image)

**Figure 4.5:** HPLC trace of the elution prior to loading into the path of the X-ray. At pH 7 through to pH 5 BoNT/E eluted in a single peak (data for HPLC of pH 5.5 and pH 6 were not saved on the beamline server); while at pH 4.5 and pH 4 the protein displayed multiple elution peaks. The highest peak was found to be aggregated protein, and in pH 4 a distinct population was identified in the later elution.

The HPLC trace is shown in figure 4.5, and shows where the data was taken from for each pH. The flat area at the start of each trace was used for the buffer subtraction, and areas around the peaks of the protein elution were used as the protein data. Due to the high concentration (for this protein) the more acidic conditions (pH 4.5 and 4) suffered some aggregation. Despite this, regions of non-aggregated protein were found and used for analysis, marked in figure 4.5 with vertical lines. Despite a large shift in the elution time of the protein across the range of pH studied, the SAXS data revealed the protein was intact and monomeric, in the regions selected for study.
It is unclear what caused this shift, however this was not pursued further due to the absence of effect on the SAXS data. The resulting curves are shown in figure 4.6, which displays the ln(I(Q)) vs Q (top) and the Kratky (bottom) plots. The curves are overlaid with the best fit theoretical curve obtained from the SASSIE webserver using the MD simulations presented in chapter 3.

**Figure 4.6:** (a) Natural log of intensity vs Q plots for BoNT/E as a function of pH; with the best fit from MD trajectories overlaid. (b) Kratky plots of BoNT/E as a function of pH; with the best fit from MD trajectories overlaid.

Under all of the conditions tested, BoNT/E exists in solution as it is seen in the crystal structure 3FFZ. Using the Guinier approximation on the experimental curves, the $R_G$ of the structure was found to be 3.60 nm, matching that of the crystal structure. The experimental curve was compared to the theoretical curve generated on SASSIE from the MD using a backbone-only trajectory to obtain a best fit frame. This frame was then extracted with all atoms reinstated and multiple hydration shell depths, until an optimal best fit was identified (figure 4.7). The hydration shell which gave the most
accurate fit was 1.60 Å deep, using the explicit solvent of the MD. The pair distance distribution function (figure 4.8) shows a well folded structure, with a real-space $R_G$ close to the Guinier approximation and PDB values. This structure, previously only seen in fixed methods such as crystallography or cryo-EM (61), is therefore now confirmed in solution conditions, and in several pH environments.

![Graphs](image)

**Figure 4.7:** (a) Best fit refinement of the closed structure. (b) Best fit refinement of the open structure.
Figure 4.8: (a) Pair distance distribution function of the closed structure identified in the SAXS and the best fit structure obtained from the MD. (b) Pair distance distribution function of the open structure identified in the SAXS and the best fit structure obtained from the MD.

However, in pH 4, a second population was identified, on the shoulder of the aggregated protein peak. This population was not aggregated, as confirmed by the linear curve in the Guinier approximation region, and displayed a significantly different scattering curve than the other populations. Using the Guinier approximation, the $R_G$ of this structure was found to be 3.82 nm. This is in good agreement with the MD structures identified in pH conditions lower than 5.5. The second main difference with the first structure, best seen in the Kratky plot, indicates a more extended structure by the slower drop-off of the tail end of the curve. The secondary smaller peak indicates a new part of the protein is creating its own scattering profile, and is separate from the main body of the protein. By comparing this curve to the theoretical curves generated from SASSIE, a new best fit structure was identified. This structure corresponds to
that identified by PCA as “cluster 2” in chapter 3, where the BD is separated from the LC. As with the other structure, a first best fit frame was obtained with backbone atoms only. This fit was then refined with all atoms and by varying the hydration shell depth, with the best fit found to be 2.25 Å depth. The increased hydration shell depth is in good agreement with the extra charge found on the protein in low pH conditions, with water molecules further afield being affected by the protein’s surface charge. The pair distance distribution function has an increased maximum distance compared to the “closed” population, and has an increased real-space $R_G$ value close to that of the “open” PDB conformation. With a low R-factor fit to the theoretical data (5.6%), this novel structure has been shown to exist in solution conditions at pH 4, and validates the in silico results obtained using approximate pH simulations. A superposition of the top 100 best fit structure from the MD to the two populations identified in SAXS represents their differences clearly, and is shown in figure 4.9.
**Figure 4.9:** Overlaid top 100 best fit structures from the MD to the two populations identified in SAXS data for BoNT/E. In cyan are the 100 best fits to the open structure, and in orange the 100 best fits to the closed structure. In blue and red are the single best fit structures to the open and closed structures, respectively.
4.5 Conclusions

After expressing a safe-to-use variant of BoNT/E using *E. coli* as a host organism, biophysical analyses were performed to identify any systematic effects of pH on the toxin’s structure. Direct comparisons to the MD results were possible thanks to tools developed to predict biophysical properties of proteins based on PDB structures and MD trajectories (57–59). The correct expression of BoNT/E was confirmed in the first instance by SDS-PAGE, which displayed the critical split of the main protein band under reducing conditions. Further confirmation was provided by accurate MS measurements, validating the expected MW to within 6 Da. Finally, the suitability of the buffers chosen for the study of this protein was assessed by CD measurements, while also providing a first look at the effect of pH on the secondary structure of BoNT/E. The CD data confirmed the stability of the protein in the selected buffers, with no loss of secondary structure seen under any conditions, in agreement with a previous study which identified little to no secondary structure changes across the BoNT/E domains (126). It also provided a first validation of the MD data, with no effect of the pH seen on the secondary structure at the scale of the whole protein. This suggests that any structural changes required for activity of this protein are primarily focused on tertiary structure, i.e. full protein conformational changes.

AUC analysis provided more insight into the structural changes and stability of BoNT/E in the range of pH environments studied. By monitoring the sedimentation velocity at high RPMs, it was shown that the protein does not tend to aggregate at the concentration used (0.5 mg/mL) with a constant sedimentation speed over time. The first notable difference based on pH environment was seen here, after converting the sedimentation data to the common unit of s(20, w) for each pH condition. This showed two distinct populations in solution, one above pH 5 and one at and below pH 5. The higher pH population sedimented faster than the lower pH population, despite displaying the same buoyant MW. This indicated that the lower pH population was subjected to higher drag resistance, and hence in a more extended conformation. This agrees well with the MD data in which the protein explores more conformations and has an increased $R_G$ in simulations at and below pH 5.
AUC, however, is a low-sensitivity method and relies in part on user estimations of the buffer densities and viscosities, and of the protein’s partial specific volume. For a deeper look at this, SAXS experiments were performed on BoNT/E in the same buffer conditions. The first result was to confirm the currently known BoNT/E structure (PDB ID: 3FFZ) in solution conditions, which had currently only been observed in fixed methods (22, 61). The experimental curve was matched to the theoretical curve of the PDB structure with an R-factor of 1%, with 0% being a perfect match. This structure was observed in solution at every pH, however in pH 4 a second population was identified. This population had an increased $R_G$ and a unique scattering profile. By comparing this new scattering curve to the MD trajectories, a novel conformation was identified in solution with an R-factor of 5.6%. This conformation matches the family of structures identified as “cluster 2” by PCA, in which the BD is separated from the LC resulting in the increased exposure of the TD. The increased $R_G$ also agrees with the AUC data, which while incomplete still displays a more extended protein in pH ≤ 5. This novel conformation may be crucial for the activity of this toxin.
Chapter 5

Biophysical studies of BoNT/A identify a potential pathway to membrane insertion
5.1 Summary

Despite sharing a high sequence homology, BoNT/A and BoNT/E have differences in their properties and domain positioning. Notably, and of interest to the pharmaceutical industry, BoNT/A has a longer duration of action post-injection than BoNT/E. BoNT/E on the other hand has a faster onset of action than BoNT/A. Some groups have argued that this difference is due to the conformational differences between BoNT/E and the other BoNT serotypes. Where the majority of serotypes have a “butterfly” shape, with LC and BD on opposite sides of the TD, BoNT/E forms a more compact shape with interacting residues between the LC and BD. Building on the information presented in previous chapters, the same studies were applied to BoNT/A in an attempt to elucidate the structural basis for the different properties of BoNT/A and BoNT/E.

The MD results did not reveal a structural change on the scale of that seen for BoNT/E, with the $R_G$ and $RMSD$ values staying comparatively low with respect to the starting structure. However, by clustering the data through PCA, some rare events were identified with potential links to the increased onset of action relative to BoNT/E. In these rare events, exclusive to pH 5 and 5.5, multiple interactions between the “switch” region and the BD were lost. Their rarity may explain the longer onset of action due to the statistical likelihood of BoNT/A adopting this structure in vivo. CD data revealed a transition between pH 5.5 and pH 4.5, coinciding with the activation barrier for this toxin. When lowering from pH 5.5 to pH 5, a loss of alpha helicity and an increase in beta-sheet content was observed. This continued to pH 4.5 and then stabilised. This is in good agreement with a previous study identifying an alpha helical region in the TD domain, the "switch" region, which becomes a large beta-hairpin extending out from the TD at and below pH 5. Finally, AUC and SAXS studies were in good agreement with the MD results, displaying the same properties across all pH conditions studied. This provides some insight into the differences of properties between the two serotypes of BoNT studied. While BoNT/E requires a larger conformational shift than BoNT/A for its activity, this shift is driven by biochemical properties and electrostatic repulsion starting at pH 5. This change is therefore likely to happen faster than the complex secondary structure changes happening in BoNT/A.
Specifically, the change observed in the “switch” region is significantly larger in BoNT/A than in BoNT/E.

5.2 Introduction

While all BoNT serotypes share a high sequence homology, and display functional similarities, they differ in some key properties. The specific targets within the SNARE family of proteins are different from one serotype to another, but the end result is the same: blocking of neurotransmission. The spatial arrangement of the domains, while individually highly similar, differs greatly from the majority of known serotypes to BoNT/E. BoNT/E, as discussed in previous chapters, forms a compact protein with interacting residues from the BD and the LC (22) similar to the structure of Tetanus neurotoxin (TeNT) (figure 5.1). Most other serotypes, BoNT/A included, display a more extended “flat” conformation, with the LC and BD on either side of the TD and no interaction sites between the two domains (60). This is suggested to play a role in their most important difference of relevance to the pharmaceutical sector: the onset of action. BoNT/E exerts its activity on the subjects faster than the other serotypes; but remains in the system for a shorter duration (6, 127).
5.3 Methods

This section describes the specific protocols followed for the work presented in this chapter. For a description of the theory behind the methods, see chapter 2.

5.3.1 MD simulations

MD simulations of BoNT/A were performed using the same conditions as for BoNT/E. The starting structure for these simulations was the crystal structure extracted from PDB ID: 3BTA (60); obtained at pH 7.0. The same tools and conditions described in chapter 3 were used for the generation and analysis of these trajectories. Due to the time spent exploring the BoNT/E MD landscape, only 2 repeats were generated for BoNT/A; with a 3rd repeat added later for pHs 4, 4.5 & 5 to explore a potential rare event identified by PCA and clustering.

5.3.2 Protein expression

A plasmid containing an endopeptidase-negative (i.e. non-toxic) BoNT/A DNA sequence was provided by Ipsen; and was expressed and purified following the same protocol as BoNT/E up to the activation step, and with the Tris concentration
maintained at 50 mM throughout. The activation differed by the use of endoproteinase Lys-C instead of Trypsin and the activation being achieved with a 20 hour incubation at 4°C, with a final concentration of Lys-c of 0.8 μg/mL in a protein solution at 0.5 mg/mL. The final polishing step was achieved using a phenyl hydrophobic column equilibrated with 1M ammonium sulfate, 50 mM Tris, pH 8.0 and eluting the protein with a linear gradient using 50 mM Tris, pH 8.0.

5.3.3 Protein characterisation

Expected properties of the expressed protein were calculated from the sequence using ProtParam tool from ExPasy and were as follows: expected MW prior to activation: 149,319.7 Da; post-activation: 148,198.5585 Da; with a pl of 6.10.

The correct expression of endopeptidase-negative BoNT/A was validated using the same methods as for BoNT/E: SDS-PAGE, MS, CD. The conditions for all experiments were the same as described for BoNT/E in chapter 4.

5.3.4 AUC and SAXS data acquisition

The same conditions were used for both AUC and SAXS as described in chapter 4.

5.4 Results and discussion

5.4.1 MD simulations

5.4.1.1 RMSD and $R_G$ contrasts to BoNT/E

In contrast to BoNT/E, BoNT/A did not explore a wide range of conformations and remained close to the crystal structure in every pH condition. This is reflected in the near constant RMSD and $R_G$ curves (figure 5.2) throughout every pH and every repeat. By splitting the data into bins of 0.05 nm width for both RMSD and $R_G$ the contrast to BoNT/E was clear, with the range of values explored staying constant irrespective of the pH condition simulated and staying within low RMSD bins (figure 5.3).

These results were further analysed by splitting the protein into two distinct selections: the BD as one, the TD and the LC as another (figure 5.4). The intra-domain and inter-domain RMSD were measured and this revealed some similarities to BoNT/E. The intra-domain RMSD remained very low throughout, indicating the
individual domains remain stable and do not undergo random melting or unfolding during these simulations. A change within the TD in line with that observed in BoNT/E in the “switch” region was not observed here; however, multiple interactions were identified between it and the BD (figure 5.12, see PCA analysis). These interactions would not have been present in the study performed by Lam et al (67) as they studied the TD in isolation, perhaps facilitating their observations. Using higher temperature simulations, or running these for a longer duration, it is possible a similar result would have been observed.

The inter-domain RMSD revealed some displacement of one domain with respect to the other, but on a significantly smaller scale compared to that observed in BoNT/E. Additionally, the recorded displacement is unaffected by pH with similar movements seen at every pH. Together, these results appear to indicate that BoNT/A does not undergo a large-scale spatial rearrangement of its domains under specific pH conditions, as was observed in the BoNT/E simulations.
**Figure 5.2:** 1 ns smoothed $R_G$ (a) and $RMSD$ (b) data from the MD simulations of BoNT/A.

![Graph showing smoothed data]

**Figure 5.3:** Binned $RMSD$ (a) and $R_G$ (b) data from the MD simulations of BoNT/A. This is in stark contrast to BoNT/E with no clear effect of pH on the distribution of conformations explored.
Figure 5.4: Binned data after measuring RMSD of two distinct atom selections. (a) and (b) show the intra-domain RMSD of the BD and the TD+LC, respectively. (c) shows the inter-domain RMSD with mobile BD and fixed TD+LC and (d) shows the inter-domain RMSD with mobile TD+LC and fixed BD. This showed that the domains remain stable throughout, and the RMSD observed over whole protein comes from inter-domain movement.

5.4.1.2 SASA

The SASA of BoNT/A was also assessed as a function of pH, with similar results as discussed above. No clear pH-dependent trend was observed, with a relatively constant SASA across all pHs (figure 5.5). This is not unexpected, as the domain organisation is less prone to changes in SASA, even if a large domain shift had been observed. In BoNT/E, the BD separates outwards from the LC which increases the SASA of all three domains: the area of the BD and the LC which faced each other, and the area of the TD which was shielded by this interface. In BoNT/A, the three domains are in a “flat” alignment, with little room to increase or decrease the SASA. A minor loss of SASA was seen in pHs 5.5 and 6, which may be due to a minor
conformation change at these conditions. Breaking this down by domain showed a near constant area over the pH range studied (figure 5.6).

**Figure 5.5:** Average global SASA of BoNT/A as a function of pH. Average taken from the final 200 ns of each repeat. At pHs 5.5 and 6, a slight dip was observed compared to the rest. This may be due to a minor structural change.
Figure 5.6: Average SASA by domain. The area was found to be near constant throughout every pH simulated. Average taken from the final 200 ns of each repeat.

5.4.1.3 Charge and salt-bridge content

As for BoNT/E, the starting structures of each pH simulation were assessed for salt bridge content and charge distribution (figure 5.7). This contrasted again, with two sharp increases in charge from pH 7 to pH 5.5; and from pH 5 to pH 4. The slope flattens between pH 5.5 and pH 5, which may be related to the activity though it is unclear how at this stage. Similarly to BoNT/E however, the first pH with positive charge on both atom selections was observed at pH 5.5, hinting at a common electrostatic requirement for the activity of these two proteins. The headgroups on cell membranes are negatively charged; hence the positive charge on the protein would favour an interaction. The salt bridge content, however, declined with a constant rate between pH 6 and pH 4.5.
5.4.1.4 PCA

By combining all the generated MD frames into a single trajectory, using backbone only atoms and removing the first 50 ns of each repeat, the conformational space explored was separated into clusters of structural similarity. By first obtaining the elbow plot (figure 5.8) of the combined trajectory, the MD frames were split into 4 clusters of closely related structures. This was the same number of clusters as used for BoNT/E, however the distribution as a function of pH differed as well as the differences between each cluster. In BoNT/E, the 4 clusters were distinct structures with large differences in the position of the BD with respect to the TD+LC, and even the least populated cluster was made up of over 18,000 frames. In BoNT/A, two clusters make up the overwhelming majority of frames, while the two remaining clusters are made up of 5 and 1 frames (figure 5.9).

**Figure 5.7:** The charge and salt-bridge distribution as a function of pH in BoNT/A.
Figure 5.8: Elbow plot of the BoNT/A combined trajectory, indicating 4 clusters as a reasonable value to use for the grouping.
Figure 5.9: (a) pH contribution to each cluster; (b) Cluster contribution to each pH. This revealed the disappearance of cluster 1 at pH 5.5 and pH 6, while clusters 1 and 2 are otherwise evenly distributed across all other pHs. Clusters 3 and 4 are exclusive to pH 5 and 5.5, respectively, but are extremely rare events.

The robustness of the clustering was monitored by measuring the average RMSD of each cluster with respect to the cluster “mid-point” and was found to be satisfactory with values lower than 0.3 nm for every cluster. The structural differences between the clusters were minor, with only slight displacement of the BD from one cluster to the other. In cluster 2, which dominates at pHs 5.5 and 6, the BD is displaced “horizontally”, i.e. around the TD; while in clusters 1, 3 and 4 the BD maintains its alignment with the rest of the protein (figure 5.11). At pHs 5.5 and 6, the total charge on the protein is the closest to neutral, which may facilitate more movement in the “horizontal” plane. It is of interest that despite their high similarity, the two structures singled out by the clustering occur in pHs 5 and 5.5, matching the activation barrier of this protein. While less noticeable than the change observed in BoNT/E, and much
rarer, these structures may be relevant to the membrane insertion mechanism of the protein. Their relative rarity (5 consecutive frames in pH 5; 1 single frame in pH 5.5; out of 17,501 frames per trajectory) compared to the open structure of BoNT/E may explain the overall increased delay of activity post injection in BoNT/A compared to BoNT/E.

**Figure 5.10:** PCA clusters viewed as a dendrogram. Clusters 3 & 4, blue & green respectively, are the most distinct; while clusters 1 & 2, black & red respectively, are closely related and account for the majority of frames.
Figure 5.1: Cluster mid-points and their contribution to total frames generated.

Figure 5.11: Cluster mid-points and their contribution to total frames generated.

Figure 5.12: Interactions between a BD lysine and several residues in the switch region. This contact would have been lost in the Lam et al study due to the absence of the BD.
To explore this further, the individual mid-point frames were extracted as full atom structures to identify local changes including the side-chains. By monitoring the interactions between the “switch” region and the BD in each of these structures, a notable difference was identified. In clusters 1 & 2, which dominate the distribution, multiple interactions were found between the “switch” and the BD involving a total of 17 & 12 residues, respectively. In clusters 3 & 4, many of these interactions are lost with only 8 and 7 residues involved, respectively (figure 5.12). These clusters are extremely rare events, with 5 and 1 frames respectively out of 210,006 total frames analysed, and are exclusive to pHs 5 & 5.5. A further analysis on trajectories from pH 4, 4.5 & 5, with one additional repeat per pH and presenting the same pattern, is presented in appendix 2. If this distribution holds true in vivo, their relative scarcity may explain the overall delay in onset of action relative to BoNT/E. The doses used for therapeutic applications of BoNT are extremely low, meaning the probability of any one protein going through this change is also low. By contrast, in BoNT/E the “switch” region is more isolated with no contacts between it and the BD.

5.4.2 Protein expression & characterisation

Over-expression and purification of BoNT/A was assessed by the same methods as for BoNT/E, with an example SDS-PAGE gel of BoNT/A shown in figure 5.13. The mass spectrometry data for BoNT/A is shown in figure 5.14 and demonstrates the mass of the purified sample to be accurate within 5 Da of the expected MW (expected: 148,198 Da; experimental: 148,201 Da). The CD data were obtained at the six pH conditions tested in the MD simulations and are shown in figure 5.15. Contrary to BoNT/E, the pH environment had a visible effect on the protein’s secondary structure with a transition occurring between pH 5.5 and 4.5. The secondary structure appears unchanged between pH 7 and pH 5.5, a transition state is seen in pH 5, followed by an identical pattern in pH 4 and pH 4.5. The lower dip at 222 nm in the higher pH values indicates a loss of helical residues as the pH is lowered from 5.5 to 5, and a further loss beyond pH 5 (figure 5.16). This is in good agreement with previous studies having identified a significant shift in secondary structure of the “switch” region (homologous to the region identified in BoNT/E in chapter 3) which goes from an alpha helix to a beta-hairpin protruding from the TD (67). The combined
loss of alpha-helical residues and appearance of beta sheets would lead to a significant change in the CD spectrum. The signal observed in pH conditions between 5.5 and 7 matches that of the predicted signal from the 3BTA PDB structure. The signal below pH 5.5 diverges from this in the 222 nm peak. Combined with the MD data, it is possible BoNT/A relies more strongly on local secondary structure changes rather than a full-scale conformational change for its activity. Smaller changes that do not involve a full domain displacement would not be detected by SAXS, and would require a more sensitive method.

Figure 5.1: Coomassie stained SDS-PAGE of the purification process of BoNT/A. The separation of the full length protein into two bands after reduction in the final purified product confirms the successful activation of the toxin.
**Figure 5.14:** Mass profile of final purified product obtained by MS measurement.
Figure 5.15: Top: Theoretical CD spectra obtained from the 3BTA structure, and experimental data obtained at pH 7. Bottom: CD spectra corrected for sample concentration, revealing a loss of helicity for BoNT/A starting at pH 5 and remaining constant at pH 4.5 and 4.
While the level of secondary structure change observed in the lab was not seen in the MD simulations, a small change has already been noted, with many interactions broken between the “switch” and some BD residues. By extending the simulation time further, larger changes in this region may have started to happen such as a gradual shift to beta-hairpin structures.

5.4.3 AUC does not detect a significant structural change

As for BoNT/E, AUC data was collected across the same pH range in an attempt to identify a conformational shift due to the pH environment. Figure 5.17 shows the s(20, w) values in different pH conditions, that is the sedimentation rate in water at 20°C after correcting for the buffer density and viscosity. This followed the MD data
with a peak of sedimentation speed, i.e. a slightly more compact structure, at pH 5. From the PCA, cluster 1 was identified as the dominant conformation in pH 5 and the BD is brought slightly closer to the TD. This would reduce its resistance to flow and increase the average sedimentation speed recorded under these conditions. At pHs 4 and 4.5, the population is mixed between clusters 1 and 2, which results in a slightly slower sedimentation due to the population of cluster 2. At pHs 5.5 and 6, predominantly in conformations within cluster 2, the protein has the slowest sedimentation due to the increased apparent size of the protein. Overall, the change in sedimentation speed recorded for BoNT/A was on a smaller scale to that seen for BoNT/E, in good agreement with the lesser difference between the structures observed.

Figure 5.17: Sedimentation rate values, corrected for buffer density and viscosity for comparison between the different pH conditions. The data from pH 7 was discarded due to poor sample quality.
5.4.4 SAXS further validates the previous findings

As a higher sensitivity method, SAXS data was acquired for BoNT/A. The experimental data was compared to structures obtained in the MD simulations, as for BoNT/E, and is shown in figure 5.18. In a further confirmation of the previous studies, no notable difference in the scattering pattern was seen as a function of pH. The 3BTA structure therefore holds true in solution conditions, and across the pH activation range of this toxin. This confirms the increased reliance on secondary structural changes rather than a spatial re-arrangement of the domains in the activation procedure of BoNT/A. The minor changes in the BD position identified in the PCA would not be sufficient to change the scattering profile due to SAXS. In addition, SAXS scattering curves are the average of every scattering particle in solution. If multiple conformations of BoNT/A exist in solution, these would not have been resolved individually by a SAXS experiment. Figure 5.19 represents the best fit refinement process, as for BoNT/E, starting from a backbone only trajectory and adding full atoms then a hydration shell. Figure 5.20 shows the $R_g$ of MD frames from pH 7 versus the R-factor fit to the experimental data, and highlights the 100 best fits.
Figure 5.18: (a) Natural log of intensity vs Q plots for BoNT/A as a function of pH; with the best fit from MD trajectories overlaid. (b) Kratky plots of BoNT/A as a function of pH; with the best fit from MD trajectories overlaid.
Figure 5.19: Refinement procedure starting with the backbone only trajectory, then extracting single best fit frame with full atoms and hydration shell.
Figure 5.20: $R_G$ vs R-factor for the pH 7 MD simulations, with the 100 best fits in red.
Figure 5.21: Superposition of the 100 best fit frames to the BoNT/A SAXS data from the MD simulations.

5.5 Conclusions

Through extensive in silico and combined solution methods, the activation procedure of BoNT/A has been identified to be more reliant on secondary structure changes, such as the “switch” region (67), than large-scale domain movements. This is seen in the near constant sedimentation velocity across the pH range as well as the single population of conformations observed in the SAXS data. BoNT/A has previously been shown to be unchanged in pH 6.0 & 8.0 (30); but this study did not explore the effect of pH values below the activation threshold of the protein. This is in contrast to BoNT/E which has only minor secondary structure changes but a large movement of the BD relative to the LC. The sudden increase in charge below pH 5.5 at the interface which drives the separation of the two domains may partly explain the difference in the
onset of action between the two serotypes. While BoNT/A relies on complex secondary structure changes from alpha helices to beta hairpins (as seen in the CD data presented here, and in the Lam et al study (67)), BoNT/E requires smaller scale secondary structure changes and a larger domain movement driven by electrostatic repulsion. Furthermore, a sub-population of BoNT/A identified by PCA is exclusively present in pH conditions 5 & 5.5, matching the activation barrier of this toxin. These structures may be relevant to the membrane insertion due to the loss of multiple key interactions holding the “switch” region down. They represent only a fraction of the conformations explored by the protein in solution (6 frames out of 210,006 clustered), potentially explaining the overall onset of action delay until such a conformation is reached. A similar rare cluster was identified in a separate PCA study of BoNT/A with an additional repeat to the pH values lower than 5.5 (see appendix 2).

While the secondary structure change of the “switch” region was not observed in these simulations, the presence of the interaction sites identified between the BD and multiple residues within the “switch” may prevent this. These interactions would have been lost in the Lam et al study (67) due to the use of an isolated TD. Repeating these simulations at higher temperatures may impart enough energy to break these interactions further at the relevant pHs. Similar interactions are absent in BoNT/E and the “switch” region is therefore less constrained. This may also contribute to the increased onset of action in BoNT/A relative to BoNT/E.
Chapter 6

Single Molecule Fluorescent Resonance Energy Transfer
6.1 Summary

Single-molecule fluorescent resonance energy transfer (smFRET) is a highly sensitive method, inversely proportional to distance to the power 6. Due to this rapid drop-off, smFRET can be used as an extremely accurate molecular ruler. By using a pair of dyes with a known Förster radius ($R_0$), where the efficiency of transfer is 50%, precise measurements can be determined for distances lower than 10 nm. smFRET has, for example, been used to determine the walking pattern of kinesin; or to validate aggregation prone conformations of a Fab. For an efficient smFRET molecule it is necessary to ensure site-specific labelling of the protein of interest. While it is possible to use cysteines as a binding site, in the case of large systems such as BoNT serotypes there are too many natural cysteines present. An alternative is the use of unnatural amino acids (UAA), which can be incorporated into the protein of interest through the use of specialised machinery designed to respond to TAG codons. One example is p-azido-phenylalanine (pAzF), whose azide side chain can be used to react with amine-labelled fluorescent dyes using click chemistry. Here we present the initial steps in an attempt to record smFRET data for BoNT serotypes A and E in varying pH environments, using Alexafluor dyes 488 and 594 (AF488, AF594). While the expression and labelling of a BoNT/A double mutant was confirmed by MS, the majority of the sample was found to be single-labelled. A small amount of double-labelled was still present in the mixture, either with 594/594 or with 488/594. BoNT/E double mutants proved harder to express, and while expression was confirmed the amount of material generated was insufficient to perform any analysis on.
6.2 Introduction

The previous chapters revealed that BoNT/A and BoNT/E undergo some structural changes in solution at and/or below the pH barrier for their activation, pH 5. In BoNT/E this change was a major shift of the BD outwards, freeing up the TD and increasing the overall exposed surface area of the protein. The sedimentation velocity in the same pH conditions indicated an extended conformation below pH 5.5; and a strong match to a SAXS signal obtained at pH 4 was found in the MD simulations below pH 5.5. In BoNT/A two rare events were identified in PCA at pHs 5 and 5.5, which displaced the BD slightly compared to the starting point. These changes, although minor, may play a role in the membrane insertion, with the relative scarcity explaining the increased onset of action post-injection compared to BoNT/E. SAXS and AUC were not sensitive enough to accurately determine whether these changes were happening in BoNT/A, and have the disadvantage of being bulk methods. The data obtained is the average of every particle in solution and multiple populations, particularly if they are rare, cannot be accurately resolved. In order to further validate the previous results, and perhaps observe the rare populations identified by PCA at pHs 5, 5.5 & 6 for BoNT/E and 5 & 5.5 for BoNT/A, some initial steps were performed ahead of smFRET experiments.

Single molecule Förster resonance energy transfer, or smFRET, is a highly sensitive method capable of detecting conformational changes on a scale significantly lower than SAXS; and by measuring single molecules can also detect rare populations within a complex solution. smFRET relies on the non-radiative energy transfer of a donor dye to an acceptor. The FRET efficiency is inversely proportional to the distance between the dyes to the power 6. With an optimal range situated between 2 and 10 nm, this method is suited to detect domain movements within single proteins (128).

The first step toward successful FRET measurements is the accurate incorporation of donor and acceptor dyes on relevant sites within the protein (128). These sites should be solvent accessible, free of any interactions with neighbouring residues and in structurally unimportant areas of the protein. Most importantly they should be on areas of the protein suspected to move relative to one another. Finally, residues selected for labelling should have low natural flexibility to decrease the signal
to noise ratio, due to the sensitivity of this method. Measurements are performed in low sample volumes through the use of appropriate objectives which focus the excitation laser to a focal point within the sample. The protein sample should be at concentrations low enough to statistically ensure having only one protein at any time in the detection volume (10-100 pM). The sample solution is placed on a fixed microscope slide, with the protein diffusing freely within. When a labelled particle enters the detection volume, a burst of fluorescence is emitted and recorded. If FRET has occurred, a burst is detected in both the donor and acceptor channels; if the protein was only donor labelled a burst may be detected in the donor channel only which would therefore not contribute to the final FRET efficiency calculation. The emitted fluorescence is separated by wavelength using dichroic mirrors and bandpass filters along the path of the beam. By recording continuously for a given time, usually 30 minutes, a number of FRET events can be recorded and the efficiency of each burst calculated. This is then plotted as a histogram of efficiency vs frequency; and the efficiency can be directly translated to a distance between the two dyes.

FRET has been widely used with BoNTs as an activity detection assay (129–132) but not as a structural analysis method. By labelling these toxins at appropriate sites, we hoped to be able to monitor the changes observed in MD and SAXS experiments discussed in the previous chapters. For site-specific labelling, cysteines are commonly used for their reactivity. For this, either naturally present cysteines are used if they appear at the relevant sites for FRET study; or the appropriate sites are mutated to cysteines for labelling. However in large systems such as BoNTs, there are too many natural cysteines which would lead to random labelling or require mutating some of the cysteines to prevent this. This is undesirable and requires a lot of work. A relatively modern solution is the use of unnatural amino acids (UAAs) which can be incorporated into the protein sequence by the use of TAG codons and engineered plasmids. The template plasmid of the protein of interest is mutated through PCR to incorporate TAG codons at the desired sites. The mutated plasmid is then co-transformed into the expression system, here we have used *E.coli* BL21(DE3) cells, with a secondary plasmid encoding an engineered machinery which recognises TAG codons and expresses an UAA in its place.
The machinery makes use of engineered aminoacyl-tRNA synthetase (aaRS) / transfer RNA (tRNA) pairs which recognise the selected codon, here TAG. In *E. coli*, mRNA is translated to a protein through tRNA which recognise the triplet codons and attaches the relevant amino acid. The amino acid is attached with the help of the aaRS, which recognises the tRNA and binds it to the correct amino acid. For successful UAA incorporation a novel aaRS / tRNA pair needs to be used, which is orthogonal to the expression host selected. That is to say the tRNA cannot be aminoacylated by any native aaRS, and the engineered aaRS cannot aminoacylate any of the native tRNAs. The engineered pair used for this thesis was derived from the aaRS / tRNA Tyr from *Methanococcus jannaschii* (74, 75) and generated by first mutating the tRNA anticodon loop to CUA. Then a library of mutant tRNA which did not directly interact with aaRS was submitted to rounds of positive and negative selection, to ensure the tRNA did not react with any native aaRS. The final step is to engineer aaRS to selectively recognise the UAA by the same process of negative / positive selection rounds from a library of mutant aaRS.

In this chapter, the first steps towards smFRET measurements of BoNTs A and E were performed with limited results. Successful creation, co-transformation and expression of two double mutants was confirmed by SDS-PAGE, western blots and MS measurements. Multiple other double mutants were generated though there was not sufficient time to attempt expressing these. Enough material was recovered from the expression of a BoNT/A double mutant to submit samples to MS measurements and CD experiments. These confirmed the correct incorporation of both pAzF, and the structural integrity of the double mutant. The MS also later confirmed a mixture of single and double-labelled protein for the BoNT/A mutant, revealing that one of the two selected sites was less reactive than the other.

### 6.3 Methods

This section describes the specific protocols followed for the work presented in this chapter. For a description of the theory behind the methods, see chapter 2.

#### 6.3.1 Selecting labelling sites
By using the crystal structures of BoNTs A and E (PDB IDs: 3BTA and 3FFZ, respectively), appropriate residues were selected on both proteins for mutagenesis. These were chosen using PyMOL to visualise the proteins, by checking the local environment of solvent accessible residues. Residues in non-critical areas of the protein, and within FRET compatible distances, were selected by checking there were no interactions with neighbouring residues. The UAA used in this study, pAzF, is highly similar to tyrosine in its structure; however only very few tyrosines matched the criteria and only one site was identified in BoNT/A. Time limitations meant only one double mutant of each protein was expressed. The selected sites, and resulting double mutants, are summarised in table 6.1. Successful expression of the two mutants in bold was confirmed by SDS-PAGE / western blotting and mass spectrometry.

**Table 6.1:** Summary of double mutants generated for this study.

<table>
<thead>
<tr>
<th>BoNT/A</th>
<th>GLN6 – ARG1130</th>
<th>GLN6 – ASN1215</th>
<th>ASP508 – ARG1130</th>
<th>ASP508 – ASN1215</th>
<th>ASN509 – ARG1130</th>
</tr>
</thead>
<tbody>
<tr>
<td>BoNT/E</td>
<td>TYR20 – LYS1102</td>
<td>TYR20 – LYS1156</td>
<td>GLU78 – LYS1102</td>
<td>GLU78 – LYS1156</td>
<td></td>
</tr>
</tbody>
</table>

![Image showing the selected residues and distances in BoNTs A and E.](image-url)
Figure 6.1: Surface representation of BoNT/A & BoNT/E with the mutated residues in red; and the distances between them.

6.3.2 Protein expression, purification and characterisation

For each double mutant listed above, a first single mutant was created by using site-directed mutagenesis through PCR. The Q5 SDM kit from New England Biolabs was used, with the plasmid containing the endopeptidase-negative BoNT sequences as the template DNA. The PCR cycle conditions and primers used for BoNT/E and BoNT/A are summarised in 6.2a,b and 6.3a,b, respectively. The chosen sites’ codons were mutated to TAG, and the PCR result was transformed into TOP10 E.Coli cells for cloning. 3 colonies were selected for sequencing; and once confirmed a miniprep was prepared of the single-mutant plasmids. Then, the second mutation was added by the same process, using the single-mutant as the template DNA.
### Table 6.2a: Summary of PCR cycles used to mutate the BoNT/E plasmid.

<table>
<thead>
<tr>
<th>Original – Mutation</th>
<th>Initial Denaturing</th>
<th>Cycled Denaturing</th>
<th>Annealing</th>
<th>Extension</th>
<th>Final Extension</th>
</tr>
</thead>
<tbody>
<tr>
<td>TYR20 – TAG20</td>
<td>98°C, 2min30</td>
<td>98°C, 15 seconds</td>
<td>67.2°C</td>
<td>72°C, 3min24</td>
<td>72°C, 5min</td>
</tr>
<tr>
<td>GLU78 – TAG78</td>
<td></td>
<td></td>
<td>62.9°C</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LYS1102 – TAG1102</td>
<td></td>
<td></td>
<td>61.3°C</td>
<td>3min24</td>
<td></td>
</tr>
<tr>
<td>LYS1156 – TAG1156</td>
<td></td>
<td></td>
<td>61°C</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 6.2b: Primers used for preparing the mutants of BoNT/E.

<table>
<thead>
<tr>
<th>BoNT/E</th>
<th>Original – Mutation</th>
<th>Forward primer</th>
<th>Reverse primer</th>
</tr>
</thead>
<tbody>
<tr>
<td>TYR20 – TAG20</td>
<td>5’-t acc atc ctg TAG atc aaa ccg gg-3’</td>
<td>5’-cg gtc gtt aac ccg gtc g-3’</td>
<td></td>
</tr>
<tr>
<td>GLU78 – TAG78</td>
<td>5’-c cag tct gac TAG gaa aaa gac cgt ttc-3’</td>
<td>5’-ag gta gtt ccg gtc gta g-3’</td>
<td></td>
</tr>
<tr>
<td>LYS1102 – TAG1102</td>
<td>5’-ta gat aga cgc TAG gat agc ac-3’</td>
<td>5’-t gaa att gtt tgg ctt c-3’</td>
<td></td>
</tr>
<tr>
<td>LYS1156 – TAG1156</td>
<td>5’-c gtc gcg agc TAG act cat ctc t-3’</td>
<td>5’-aa att gat gta tac ctg atc gtt c-3’</td>
<td></td>
</tr>
</tbody>
</table>

### Table 6.3a: Summary of PCR cycles used to mutate the BoNT/A plasmid.

<table>
<thead>
<tr>
<th>Original – Mutation</th>
<th>Initial Denaturing</th>
<th>Cycled Denaturing</th>
<th>Annealing</th>
<th>Extension</th>
<th>Final Extension</th>
</tr>
</thead>
<tbody>
<tr>
<td>GLN6 – TAG6</td>
<td>98°C, 2min30</td>
<td>98°C, 15 seconds</td>
<td>56.2°C</td>
<td>72°C, 3min24</td>
<td>72°C, 5min</td>
</tr>
<tr>
<td>ASP508 – TAG508</td>
<td></td>
<td></td>
<td>63.1°C</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASN509 – TAG509</td>
<td></td>
<td></td>
<td>59.7°C</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARG1130 – TAG1130</td>
<td></td>
<td></td>
<td>63.1°C</td>
<td>3min24</td>
<td></td>
</tr>
<tr>
<td>ASN1215 – TAG1215</td>
<td></td>
<td></td>
<td>67.1°C</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 6.3b: Primers used for preparing the mutants of BoNT/A.

<table>
<thead>
<tr>
<th>Original – Mutation</th>
<th>Forward primer</th>
<th>Reverse primer</th>
</tr>
</thead>
<tbody>
<tr>
<td>GLN6 – TAG6</td>
<td>5'-t gtg aac aaa <strong>TAG</strong> ttc aac tat aaa g-3'</td>
<td>5'-aa cgg cat atg tat atc tc-3'</td>
</tr>
<tr>
<td>ASP508 – TAG508</td>
<td>5'-c ttc aac ttt <strong>TAG</strong> aac gag ccg g-3'</td>
<td>5'-gt cag ata ata ctg ctg aat c-3'</td>
</tr>
<tr>
<td>ASN509 – TAG509</td>
<td>5'-c aac ttt gat <strong>TAG</strong> gag ccg gaa aat atc-3'</td>
<td>5'-aa ggt cag ata ata ctg c-3'</td>
</tr>
<tr>
<td>ARG1130 – TAG1130</td>
<td>5'-t gtt ggt atc <strong>TAG</strong> ggc tat atg tat ctg-3'</td>
<td>5'-tt gtt cac atc cac ata ttt g-3'</td>
</tr>
<tr>
<td>ASN1215 – TAG1215</td>
<td>5'-g gat gtt ggt <strong>TAG</strong> ctg agc cag g-3'</td>
<td>5'-gg aat ttc cag tgc gct c-3'</td>
</tr>
</tbody>
</table>

Once the double-mutant plasmids were generated, they were co-transformed into BL21(DE3) *E.Coli* with pULTRA-CNF (74). This plasmid provided the machinery required for the incorporation of pAzF in the TAG codon locations, and conferred resistance to spectinomycin. The double antibiotic resistance, kanamycin and spectinomycin, was used as a control to ensure only cells carrying both plasmids were grown. The expression culture was the same as for the original BoNTs, with the addition of 1 mM pAzF into the media. Induction with IPTG activated both the UAA incorporation machinery and the expression of the protein of interest. The double-mutant protein was then purified following the same protocol as the wild type (described in chapters 4 & 5). The final purified fraction was tested by SDS-PAGE, MS and CD to ensure the separation into two bands after reduction, the correct incorporation of two pAzF residues, and the retention of secondary structure compared to the wild type protein.

### 6.3.3 Labelling setup

The final purified protein sample was labelled in the final elution buffer (BoNT/A: 300 mM ammonium sulfate, 35 mM Tris, pH 8.0) at a concentration of ≈ 0.3 mg/mL. A 15X molar excess of both dyes was added and incubated overnight at 20°C. The loose dye was then removed by using Zeba™ spin desalting columns (Thermo fisher scientific) with a 40K MWCO, and the protein buffer exchanged in the process. After the labelling reaction and clean-up, the sample was tested with MS to detect the MW shift and compare to the expected MW assuming double-labelling.
6.3.4 smFRET setup

The fluorescent microscope used was a picoquant microtime 200, controlled by SymPhoTime 64 software. A UPlanApo 60x/1.20W objective (Olympus) was used to focus the laser into the sample. A laser emitting at 482nm was used to excite the donor fluorophore (AF488) at 20 MHz, leading to a pulse every 50 ns. The excitation pulse was separated from the back-reflection from the sample with a 485nm dichroic mirror. The remaining photons were then focused using a 100 μm pinhole onto 2 single-photon avalanche diode (SPAD) detectors with appropriate light filters placed to select for the acceptor and donor emissions. The donor emission was selected by using a 525/50 bandpass filter, while the acceptor emission was selected using a first 585nm dichroic mirror to remove excess photons from the donor emission; then a 645/70 bandpass filter.

6.4 Results

6.4.1 Protein expression & labelling

6.4.1.1 BoNT/A

The first BoNT/A mutant selected for expression was the 6-1130 double mutant, which was expressed in BL21(DE3) in modified terrific broth supplemented with kanamycin, spectinomycin and the UAA at 1 mM concentration. Figure 6.2 confirms the correct expression of the BoNT double mutant when the media was supplemented with the pAzF, while it was not expressed in the control culture where no UAA was added in the media. Figure 6.3 is a gel of the final two steps in the purification procedure and confirm that the protein is intact (148 kDa) even after reduction prior to the activation step, and post-activation it is reduced into the characteristic LC (50 kDa) and HC (100 kDa) fragments.
Figure 6.2: SDS-PAGE of fractions collected after the IEX step of the purification, displaying the BoNT band at 148 kDa when expressed in media supplemented with pAzF (left) and lack of expression when omitting pAzF (right).
Figure 6.3: SDS-PAGE of the final two steps in the purification procedure for BoNT/A 6-1130. This confirmed the protein remains intact even after reduction prior to the activation step; then splits into two chains when reduced after activation.

Before attempting the labelling, a sample of this final purified product was submitted to MS to confirm the correct incorporation of both pAzF at the two mutated sites. Figure 6.4 represents the ESI scan elution profile of the main BoNT peak during MS analysis; and the resulting mass profiles when looking at the different regions of the peak. This confirmed the correct expression of double mutants of BoNT/A, but also expression of some single-mutants due to misincorporation of the pAzF residue. Using fractions from the same purification batch, labelling was attempted on BoNT/A. The sample was labelled using a protein concentration of 0.3 mg/mL and a 15X molar excess of both dyes and incubated at 20°C overnight, after which the loose dye was separated from the protein by using Zeba™ spin desalting columns with a 40K MWCO. This also served as a buffer exchange step into 50 mM ammonium acetate for MS. The results of the MS analysis are presented in figure 6.5.
Figure 6.4: ESI elution profile of the main BoNT peak during MS analysis (top) and resulting mass profiles (bottom). This confirms expression of double mutants, but also some single mutants due to misincorporation of the pAzF residue.
Figure 6.5: Mass profiles of different regions within the main protein peak of MS. This revealed a mixture of proteins with multiple molecular weights corresponding to single or double labelled mutants. A large portion of the sample was found to be double mutant with a single dye reacting. A smaller portion was protein with un-labelled pAzF at site 6 and misincorporation at site 1130. Finally, a minor amount of double labelled protein was confirmed, either with two acceptor dyes or with a donor/acceptor pair.

Figure 6.5 reveals a multitude of populations in the mixture post-labelling reaction with some unlabelled, some single labelled and some double labelled. The first peak, at 148,250 Da, corresponds to a single mutant with pAzF at site 6 and misincorporation at site 1130. The second and third peaks correspond to a double mutant with a single dye reacted, respectively AF594 and AF488, predominantly at site 1130. Finally, the fourth and fifth peaks correspond to double labelled double mutants, either with two acceptor dyes or an acceptor/donor pair. This showed that
site 6 was less readily available for labelling, and that AF594 was more reactive than AF488.

**6.4.1.2 BoNT/E**

The first BoNT/E mutant selected for expression was the 78-1156 double mutant, which was expressed in BL21(DE3) in modified terrific broth supplemented with kanamycin, spectinomycin and the UAA at 1 mM concentration. The expression of this double mutant was found to be significantly lower than that of the BoNT/A double mutant. The expression of the template was already slightly lower, so this was in line with expectations. No band was readily visible in SDS-PAGE fractions, so a western blot was used to confirm the presence and expression of this mutant. The result is shown in figure 6.6 and confirms the correct expression of the mutant, but in quantities too low to detect on the SDS-PAGE Coomassie stain.

Due to the very low expression levels, this was not taken any further at this stage.
1. Chemiluminescence image of western blot
2. Marker
3. Normal image of same western blot
4. Marker

**Figure 6.6:** Western blot performed on fractions after the first step of the purification. Only the relevant fraction is shown here. On the left, the chemiluminescence image overlaid on the Coomassie stained gel (right). This confirms the presence of BoNT/E just below the 150 kDa marker band, however it was too low to detect using Coomassie staining.

### 6.5 Conclusions

Presented here are the first steps towards potential work using fluorescent labelling to detect conformational changes in BoNTs A and E. The challenges here were selecting the appropriate sites on BoNT/A for labelling due to the size of the protein. The limited number of sites that were solvent exposed and free of any interactions with neighbouring residues meant the chosen residues were very distant in terms of FRET measurements. Therefore, even if the labelling had been fully optimised and a pure sample containing primarily double labelled proteins with donor
and acceptor pairs was obtained, the FRET signal would have been almost indistinguishable from the background. For BoNT/E, the challenge was the expression which was considerably reduced compared to the wild type, which itself was already lower than the expression of BoNT/A. By checking with a western blot, the expression of the protein was confirmed though the amount of material generated did not permit any analysis to be undertaken on it. Based on the experience with BoNT/A and the use of the same UAA incorporation machinery and expression conditions, it is likely the protein would have been expressed with some misincorporation too.
CHAPTER 7

Summary and Future Work
7.1 Summary

7.1.1 BoNT/E

Starting with MD simulations, a large amount of the conformational space available to BoNT/E was explored. This revealed a significant change in behaviour across the pH activation barrier of this toxin, with a highly dynamic protein at and below pH 5 and a highly stable protein above pH 5. From intra- and inter-domain RMSD analyses, this flexibility was shown to translate into a displacement of the full BD with respect to the TD+LC while maintaining its structural integrity, rather than a combined effect of many smaller motions. The underlying cause for this sudden increase in dynamic behaviour was identified as a net increase in charge on both domains, with a sharper transition between pH 5.5 and pH 4.5 than between other pHs, combined with a sudden loss of salt bridges within the protein. In particular, two interaction sites between the LC and the BD in the starting structure were lost when dropping from pH 5.5 to pH 5. A third site, found within a helical motif conserved from the diphtheria toxin, was shown to switch from negatively charged at pH > 5.5 to neutral at pH ≤ 5. The net result of this increased dynamic behaviour was an overall increase in SASA between pH 6 and pH 4.5, with the sharpest increase between pH 5.5 and 5. The increased SASA also derived from an overall increase in the \( R_G \) of the protein, with the loss of the interface between LC and BD leaving large portions of both of these and the TD exposed. A small region of interest was identified in the "switch" region, a small alpha helical area within the TD conserved between all BoNT serotypes and TeNT. This region was shown to transition from alpha helical to beta-hairpin in acidic conditions in BoNT/A, and a similar transition was observed in simulations of BoNT/E exclusively below pH 5.5. This resulted in a buried lysine becoming exposed in acidic conditions, priming the TD for membrane insertion. Finally, the "switch" region was observed to have high dynamic cross correlation, at pH 5 exclusively, with two spatially distinct regions of the protein: the region involved in the original binding event and a region identified as likely to be involved in the membrane insertion by TmPred. It is plausible these 3 regions act in tandem for the full membrane insertion process.

420,024 of the frames analysed from the MD were clustered using PCA into just 4 groups of similar structures, indicating that the trajectories were not merely
random displacements of the BD. The two dominant clusters corresponded to the starting crystal structure and a structure with an increased $R_G$, the latter being exclusively observed in pH conditions $\leq 5$. AUC and SAXS experiments under the same pH conditions validated the MD results, with a faster sedimentation in pHs $> 5$ and two distinct populations identified in SAXS. These two populations, one of which observed in pH 4, produced distinct scattering curves with both finding a strong match from the MD frames.

By generating a large population of structures from the same starting point and only modifying the simulated pH, a novel conformation of BoNT/E has been identified and validated by SAXS data. This conformation is more extended, leading to an increase to the SASA of the protein and leaving key sections exposed, such as the conserved helix from the diphtheria toxin. In conjunction to a secondary structure change in the highly conserved “switch” region, a probable combination of events has been identified as facilitating membrane insertion for BoNT/E; as well as the underlying biochemical changes happening across the pH barrier.

7.1.2 BoNT/A

Although large domain displacements on the scale of those observed in BoNT/E MD were not seen for BoNT/A, a large number of structures were generated with some subtle changes in the BD orientation. Despite no major changes to the $R_G$ or the SASA the structures were placed into 4 clusters, with 2 clusters being extremely rare events. However, these two rare events occurred in simulations at pH 5 & pH 5.5; potentially indicating a structure required for the translocation mechanism. Incidentally, their low occurrence relative to the other dominant structures may provide an answer to the increased onset of action post-injection of BoNT/A relative to BoNT/E. The difference between these four clusters were however significantly smaller than between the four clusters of BoNT/E, and any differences were too small to accurately determine in SAXS. In particular, if these were rare events within a mixed population, data from SAXS would not have been able to single out these rare events due to the averaging over every molecule in solution.

The currently known structure of BoNT/A was found to be maintained across all pHs studied, with a near-constant rate of sedimentation in AUC experiments, and
a single SAXS scattering curve observed across all pH values. This scattering curve found a strong match within the MD frames, and corresponded to the crystal structure already known. From CD experiments, a strong secondary structure change was observed across the pH activation barrier, with a stable signal between pHs 5.5 and 7 followed by a loss of helicity at pH 5 and a further loss at pH 4.5 which stayed constant at pH 4. This was in good agreement with the change identified within the “switch” region.

Although no major structural changes were observed across the pH range studied and in the conditions used for this thesis, some potential avenues for future study were identified. First, an extremely small fraction of the MD frames explored a conformation exclusively in pH 5 and 5.5, therefore potentially linking to the activity of this protein. Secondly, an interaction between a lysine on the BD and multiple residues within the “switch” region was identified and may explain the absence of a strong change in these simulations. A similar interaction is absent in BoNT/E. And finally, a strong secondary structure change was observed across the pH barrier which was not seen in BoNT/E. The rarity of the conformation explored in the MD at pH 5 & 5.5, the stronger secondary structure modifications and the interaction between the BD and the “switch” region, may combine to explain the increased onset of action of BoNT/A relative to BoNT/E.

7.2 Future Work

7.2.1 BoNT/E

To further improve on the BoNT/E information presented in this thesis, several avenues could be explored. First, taking the final structure of the MD simulations here and re-simulating the same pH from this new conformation. This would be a closer approximation to the true pH environment and may provide further insight into the mechanism for membrane insertion. Pursuing the cpH simulations for extended timescales and repeating these would be of interest, as this could validate, or invalidate, the data generated by the approximate pH simulation with no protonation updates. Exploring the importance of the switch residue identified, K618, through mutagenesis studies would add evidence to the relevance of this region. Generating
specific antibodies targeting this region could be an option for botulism treatments by blocking the membrane insertion; a method already used to treat tetanus in mice by targeting the homologous region in TeNT (133). Optimising the expression of the double mutants with UAA incorporation would enable a significantly higher resolution study of this protein, and potentially detect any short-lived states or rare events. Finally, different effects could be pursued on this protein such as performing experiments at the more biologically relevant temperature of 37°C; or monitoring the effect of multiple ionic strength / pH conditions on this protein’s stability.

7.2.2 BoNT/A

Pursuing the MD to generate additional repeats to match the full data generated for BoNT/E may be crucial in identifying critical conformations. The structures currently seen as rare events at pHs 5 & 5.5 may be more populated, potentially validating their relevance for membrane insertion. Attempting the same cpH simulations on BoNT/A may also produce more dynamic results with the protonation updated throughout the run. Mutational studies on the “switch” region would provide insight into the importance of particular residues, such as those involved in an interaction with the BD lysine. Performing all the experiments at higher temperatures and in a broader range of ionic strength buffers could be more revealing due to the additional energy input to the system.

Optimising the labelling of the double mutants would enable the acquisition of highly sensitive smFRET data, with the potential to single out rare events and small populations of distinct conformations within the same sample. This could validate the MD data if subtle changes in FRET efficiency were observed.
References


24. S. Swaminathan, S. Eswaramoorthy, Structural analysis of the catalytic and


30. T. Matsui, *et al.*, Structural basis of the pH-dependent assembly of a botulinum


33. Y. Fujinaga, *et al.*, The haemagglutinin of Clostridium botulinum type C progenitor toxin plays an essential role in binding of toxin to the epithelial cells of guinea pig small intestine, leading to the efficient absorption of the toxin. *Microbiology (N Y)* 143 (1997).


43. G. Schiavo, C. C. Shone, O. Rossetto, F. C. G. Alexander, C. Montecucco, Botulinum


50. K. Pande, *et al.*, Femtosecond structural dynamics drives the trans/cis isomerization in


58. L. Mavridis, R. W. Janes, PDB2CD: a web-based application for the generation of circular dichroism spectra from protein


65. G. Chellappan, *et al.*, Structural and functional analysis of botulinum neurotoxin subunits for pH-dependent membrane channel formation and translocation. *Biochimica et Biophysica*


R. Ratts, et al., A conserved motif in transmembrane helix 1 of diphtheria toxin


86. Y. Wei, A. A. Thyparambil, R. A. Latour, Protein helical structure determination using CD spectroscopy for solutions with strong background absorbance from 190 to 230 nm.
93. B. Schuler, H. Hofmann, Single-molecule spectroscopy of protein folding dynamics-


101. F. Wang, H. Wan, J. Hu, S. Chang, Molecular dynamics simulations of wild type and mutants of botulinum neurotoxin A
complexed with synaptic vesicle protein 2C. 


https://doi.org/10.1007/SpringerReference_28001.


122. S. J. PERKINS, Protein volumes and hydration effects: The calculations of partial specific


---

**Appendix 1 – BoNT/E Constant pH MD Analysis**

The constant pH data was first sorted into bins to assess the *RMSD* and *R_G* distribution, as with the larger data set. A similar trend was observed, with larger *R_G* values explored in the lower pH simulations.
The trajectories were then combined and analysed with PCA, through which 4 clusters were sufficient to observe the major motions of the BD with respect to the TD+LC (Elbow plot). While the crystal structure (cluster 1, black) dominated the distribution, the remaining three clusters were exclusive to one pH simulation each. The largest of these three clusters (cluster 2, red) was found in pH 4 and corresponds to the experimentally observed structure from chapter 4.
The results from just one, much shorter, simulation under constant pH conditions reflect the results from the larger dataset obtained from approximate pH simulations, most notably with the presence of the same structure observed experimentally through SAXS at pH 4. This means that for future work, simulating the starting pH with no, or minimal, adjustments throughout the trajectory may be sufficiently accurate to determine pH effects on proteins.
Appendix 2 – Low pH BoNT/A MD PCA analysis, with 3 repeats per pH

A separate PCA analysis was performed on the low pH only simulations of BoNT/A, with one extra repeat per pH (4, 4.5 & 5). The aim was to see if further rare events were identified in these, or if the same rare events were observed as presented in chapter 5. The results suggested 4 clusters were enough to represent the major structural changes throughout these trajectories, with one cluster almost exclusive to pH 5 (cluster 4) while remaining very rare (154 frames out of 315,009; 0.05%).

![Diagram showing pH contribution to cluster and cluster contribution to pH](image-url)
Taking the mid-point frames of all clusters with full atoms showed the same results as the full pH set analysis presented in chapter 5. Clusters 1 to 3, which are distributed amongst the three pH conditions, have a large amount of contacts between the “switch” and the BD; the majority of which are lost in cluster 4, close to exclusive in pH 5. This further demonstrates the importance of this region and the high dependency on pH; while potentially explaining the difference in onset of action relative to BoNT/E in which the “switch” region is free of any contacts with the rest of the protein.
Appendix 3 – Publication
Research

Elucidation of critical pH-dependent structural changes in Botulinum Neurotoxin E

Christophe J. Lalaurie a, Andrew Splevins b,d, Teresa S. Barata c,d, Karen A. Bunting d, Daniel R. Higazi e, Mire Zloh f, Valentina A. Spiteri g, Stephen J. Perkins g, Paul A. Dalby a,*

a Department of Biochemical Engineering, Bernard Katz Building, University College London, Gordon Street, London WC1H 0AH, UK
b Evox Therapeutics Ltd, Oxford Science Park, Medwar Center, Oxford, England OX4 4HG, UK
c FairJourney Biologics, 823 Rua do Campo Alegre, Porto, Porto 4150-180, Portugal
d Ipsen Bioinnovation, 102 Park Drive, Milton Park, Abingdon, Oxfordshire OX14 4RY UK
e Ipsen Biopharm Ltd., Wrexham Industrial Estate, 9 Ash Road, LL13 9UF, UK
f UCL School of Pharmacy, University College London, 29-39 Brunswick Square, London WC1N 1AX, UK
g Department of Structural and Molecular Biology, Division of Biosciences, Darwin Building, University College London, Gower Street, London WC1E 6BT, UK

A R T I C L E   I N F O

Keywords:
Botulinum Neurotoxin
Molecular dynamics
Small-angle X-ray scattering

A B S T R A C T

Botulinum Neurotoxins (BoNT) are the most potent toxins currently known. However, they also have therapeutic applications for an increasing number of motor related conditions due to their specificity, and low diffusion into the system. Although the start- and end-points for the BoNT mechanism of action are well-studied, a critical step remains poorly understood. It is theorised that BoNTs undergo a pH-triggered conformational shift, activating the neurotoxin by priming it to form a transmembrane (TM) channel. To test this hypothesis, we combined molecular dynamics (MD) simulations and small-angle X-ray scattering (SAXS), revealing a new conformation of serotype E (BoNT/E). This conformation was exclusively observed in simulations below pH 5.5, as determined by principal component analysis (PCA), and its theoretical SAXS profile matched an experimental SAXS profile obtained at pH 4. Additionally, a localised secondary structural change was observed in MD simulations below pH 5.5, in a region previously identified as instrumental for membrane insertion for serotype A (BoNT/A). These changes were found at a critical pH value for BoNTs in vivo, and may be relevant for their therapeutic use.

1. Introduction

The Botulinum NeuroToxin (BoNT) family of proteins, produced by Clostridium botulinum, are currently the most potent toxins known. With an LD₅₀ reaching as low as 1 ng/kg (Pirazzini et al., 2017) depending on the serotype, they are considered a potential bioterror threat (Rossetto et al., 2015) due to their common presence in soils around the world. There are seven widely accepted serotypes, termed A through to G (Kukreja et al., 2010); which can be divided into further subtypes (termed A1, A2 etc). A new X serotype has also been reported (Masuyer et al., 2015). More subtle variations of the known sero- and sub-types are likely to be discovered over time, with advances in deep-sequencing techniques. BoNTs are responsible for Botulism, a flaccid paralysis condition which can be lethal if not detected in time. The most common method of infection is through ingestion of contaminated meat, through the gastro-intestinal tract. BoNTs are produced with a progenitor complex, which protects them from the highly acidic conditions until they enter the bloodstream (Rummel, 2015). These proteins are highly specific to nerve terminals, and when injected locally have low diffusion into the surroundings. Their extreme potency and specificity is also of significant interest to the pharmaceutical industry, with many new medical applications discovered since the first reported use for the treatment of blepharospasm (Cooper, 2007; Fonfria et al., 2018; Jankovic, 2004), a list that is likely to expand as our understanding of these complex proteins grows.
BoNTs and their overall mechanism of action have been well studied, from entry into the system through to the target proteins and their resulting effects. The 150 kDa protein is initially produced as a single chain and is then split into a heavy (100 kDa, HC) and light (50 kDa, LC) chain. The HC contains a binding domain (BD) and a translocation domain (TD), while the LC forms a single domain and is the catalytic part of the toxin. A disulphide bond formed between the TD and the LC, maintains the protein as a single entity after a small region is cleaved between the TD and the LC, and is crucial for in vivo function.

Two structural conformations of BoNTs have been identified through crystallography. The open conformation has the LC and the BD on opposite sides of the TD (Fig. 1a), while the compact conformation brings the LC and the BD into proximity, folded around the TD (Fig. 1b). Most BoNT serotypes have been identified as being in the open conformation under physiological pH conditions. However, serotype E (BoNT/E) has been observed in the closed conformation, which is closer to the known crystal structure of Tetanus Neurotoxin (TeNT), a close homologue with similar structure and activity (Fig. 1c).

Mechanistically, the BD enables entry into the nerve cell by binding to the surface of the neuron. Upon neurotransmitter release, the BD latches onto luminal vesicle protein domains exposed on the cell surface and the toxin is internalised during re-entry into the emitting neuron. The protein is inside the vesicle as it is prepared for a new cycle of neurotransmission, during which the pH drops to \( \approx 5 \) (Fischer and Montal, 2007). At this point, the TD forms a transmembrane channel in the vesicle, and the LC enters the cytoplasm where the disulphide bond is reduced. After reduction, the LC acts as an endopeptidase, and targets proteins of the SNARE family. These proteins form the SNARE complex whose function is to fuse the vesicle, which contains the neurotransmitter, to the presynaptic membrane and facilitate neurotransmitter release into the synaptic cleft. By targeting this family of proteins, BoNTs prevent formation of the SNARE complex and hence silence neurotransmission (Montal, 2010). The modular design of BoNTs, and the functionally distinct domains, increases their potential for therapeutics, by leaving open the possibility to combine different domains for a particular purpose (Keith and John, 2010); such as Miyashita et al (Miyashita et al., 2021) achieved to treat botulism in mice by fusing antibodies to the BD and TD of a BoNT.

While this paints a clear picture of the process from start to finish, the physical effect of the pH drop on the protein structure remains poorly understood, though a conformational change is likely to take place prior to formation of the transmembrane channels. A study on TeNT (Masuyer et al., 2017), a very close homologue to BoNTs (see sequence alignment in Supplementary Fig S1a,b), revealed a more extended structure at lower pH conditions, but a more compact structure, close to that of BoNT/E, at higher pH values. This suggests that a similar structural change may occur in BoNT/E. However, previous computational studies of BoNTs have focused on BoNT/A, and were also limited to short simulation times and pH values (Chen and Deng, 2007; Chen et al., 2007; Wang et al., 2015). Despite shorter simulation times (between 60 and 200 ns), Chen et al were able to identify some small localised conformational changes in BoNT/A. In order to detect larger scale global changes such as interdomain movement, longer simulations would be required. A number of other proteins are also known to require acidic pH environments to release their payload, such as influenza virus hemagglutinin or Ebola virus GP. These proteins rely on a smaller scale secondary structure change in which a hydrophobic loop is buried at neutral pH but exposed in acidic environments (Gregory et al., 2011; Lam et al., 2018). A similar mechanism was identified in BoNT/A, and is conserved in BoNT/E and TeNT, by Lam et al (2018), and warrants analysis of this region in BoNT/E. Finally, pH is known to affect the stability of proteins in ways specific to each protein (O’Brien et al., 2012). In some cases, proteins display increased flexibility at lower pH values, such as chymotrypsin inhibitor 2, where lower forces are required to achieve unfolding. This effect is in part due to the protonation of specific residues (Tollinger et al., 2003). Identifying a similar behaviour in BoNT/E may shed light on the pH dependent activity of this protein.
To explore these hypotheses for BoNT/E and expand on the MD landscape of BoNTs, a set of 400 ns molecular dynamic (MD) simulations were conducted at several pH conditions in the range from 4 to 7. By increasing the final runtime of these MD simulations, we aimed to observe similar local structural changes but also expected larger scale (e.g. inter-domain) conformational changes to occur. In particular, these could identify the role of the BD / LC interface in BoNT/E conformational changes, and how this may affect the activity. Global root mean square deviation (RMSD, a measure of the deviation of a structure with respect to a reference structure) analyses of the trajectories, using the crystal structure of BoNT/E (PDB ID: 3FFZ) (Kumaran et al., 2009) as a point of reference, explored protein flexibility; while the radius of gyration ($R_g$, a measure of protein extension) and solvent accessible surface area (SASA) tracked any global conformational changes over time at each pH. Structural changes within each of the individual domains were also analysed in more detail. Clustering of the full trajectories by principal component analysis (PCA) determined that the protein occupied a small number of discrete conformations across the full pH range studied, and did not simply undergo a broad range of random motions. Finally, the conformations explored by the MD simulations were used to generate theoretical small angle x-ray scattering (SAXS) curves, and compared to experimental curves obtained at instrument B21 (Cowieson et al., 2020) (Diamond Light Source).

2. Results

2.1. MD simulations reveal pH dependent behaviour

MD simulations allow for the study of solution behaviour of proteins with atomic level resolution. Despite relying on a number of user-defined conditions, they have been shown to accurately mimic experimental behaviour consistently (Hollingsworth and Dror, 2018). With the current advances in computing power, they are increasingly used as a screening ahead of experimental tests (Suan Li and Khanh Mai, 2013). The trajectories presented here were first analysed using RMSD and $R_g$ calculations within Gromacs (Van Der Spoel et al., 2005), in view of identifying any local and/or global conformational changes. The $R_g$ values (Supplementary Fig S2a) showed that for pH 5 and below, most repeats explored different conformations in which an increase in $R_g$ values was observed. The two biggest changes in $R_g$ values occurred in the first pH 4 and the third pH 5 repeats. In both of these, the protein transitioned from an $R_g$ value starting at 3.60 nm and extended gradually towards a final value of ≈ 3.85 nm; an increase of 7%. The $R_g$ value for the third pH 5 repeat had not fully stabilised by the end of the simulation and so may have increased further if the simulation time were to be extended. By contrast, at pH 5.5 and above, the $R_g$ values remained at a close to constant 3.60 nm, which matches that of the crystal
structure of the closed form, obtained at pH 7 (PDB ID: 3FFZ) (Kumar et al., 2009).

RMSD analysis can potentially detect conformational changes which do not affect the \( R_G \) value. This was also calculated as a function of trajectory time (Supplementary Fig S2b). At pH 5 and below, the increase in RMSD values corresponded broadly to the increase in \( R_G \) values, indicating that the overall changes in structure were due to movements which also affected the \( R_G \) of the protein. At pH 5, 5.5 and 6 however, some trajectories had a large change in RMSD values while the \( R_G \) values remained unchanged. The first repeat at pH 6 settled at a constant RMSD value, however the second pH 5.5 repeat ended on an upward trend and may have increased further if the simulation had been extended beyond 400 ns. This pointed to a different conformational change in which the corresponding \( R_G \) value was unaffected. These three trajectories are addressed further below, when clustering the data through PCA. By contrast, at pH 7 both the RMSD and the \( R_G \) values remained largely constant throughout the four simulations.

Fig. 2. RMSD analysis of MD simulations, with exclusion of the first 50 ns. (a) represents the binned data from RMSD analysis, with bin width 0.05 nm. (b) represents the binned data from the \( R_G \) analysis, with bin width 0.05 nm. (c) represents the RMSD bin counts of the BD after aligning it to the same selection of residues in the BoNT/E (PDB ID: 3FFZ). (d) represents the RMSD bin counts of the TD + LC after aligning it to the same selection of residues in the BoNT/E (PDB ID: 3FFZ). (e) represents the RMSD bin counts of the BD after fixing the BD selection, i.e. the RMSD of the BD with respect to the starting position of this selection in the PDB ID 3FFZ. (f) represents the RMSD bin count of the TD + LC after fixing the BD selection, i.e. the RMSD of the TD + LC with respect to the starting position of this selection in the PDB ID 3FFZ. The sum of all the bins presented per quarter is 100%. Data collected from 840,048 backbone-only frames.

The intra-domain RMSD values for each single domain (BD or TD + LC) remained very low throughout. This indicates that the changes in global RMSD values were not due to protein unfolding or conformational changes within any of the individual domains, and so must have instead derived from translational movements of one domain relative to the other (inter domain). An exception to this was in the first pH 6 simulation, which had a slightly higher intra-domain RMSD value for the BD, which explains the observation of an overall increase in RMSD value but with no change in \( R_G \) value. The translational motion of the domains was confirmed from the two inter-domain RMSD value distributions. However, even in this more
detailed analysis, the inter-domain RMSD values at pH 6 and 7 remained below 1 nm in most frames. This indicated that the protein was very stable at pH 7, while only a small change occurred in just one trajectory at pH 6. By contrast, at the lower pHs the inter-domain RMSD values were widely distributed up to 2.5 nm, again indicating a higher flexibility of the protein and an ability to change the relative position of its domains.

Given that the TD-membrane interactions may be driven by solvent exposure at lower pH values, for otherwise buried structure, any changes in solvent exposure resulting from the observed inter-domain conformational changes, were located and quantified. The overall SASA of the full protein was calculated for every frame. The average SASA from the last 200 ns of four repeats per pH showed a net increase when the pH was decreased from 6.0 to 4.5 (Fig. 3a), with the transition centred at between pH 5.0 and 5.5. The largest overall change came from the BD with a large portion shielded by the LC at pH 6.0, however this domain became increasingly and steadily solvent exposed as the pH decreased (Fig. 3b).

While the SASA of the LC and BD showed a clear transition at pH 5.0–5.5, the TD was found to have an increased SASA at pH 5.5 and 6 compared to pH 7. At pH 5 the SASA decreased coming closer to that of the reference structure at pH 7, however this domain became increasingly and steadily solvent exposed at this pH; with some repeats leaving the TD more exposed, while others resulted in either unchanged SASA or a more shielded domain. Fig. 4a presents the localisation of the change in SASA on the TD relative to that at pH 7, and highlights the residues most affected by the conformational change. A small section near the “lower” half (residues 808 – 823) of the TD was consistently affected, and is near the hinge region that links the TD and the BD together. Another area of importance was a short helix (residues 614 – 620) that protruded from the TD in the “upper” half (Fig. 4b,c). A homologous region was previously identified, through crystallography and functional studies, as playing a crucial role for membrane insertion in BoNT/A (Lam et al., 2018). This region saw an increase in SASA at pH 5.5 and below, and yet remained largely unchanged at pH 5.5 and 6, with respect to pH 7. This region forms several α-helices which maintained their secondary structure at pH 5.5 to 7 throughout the full trajectories. By contrast, one of the helices briefly unfolded and increased the solvent exposure of a lysine residue (K618) exclusively in pH 4 to 5 simulations.

In order to explain the observed increase in flexibility below pH 5, the charge distribution was analysed at the starting point of each pH condition. At the transition from pH 5.5 to pH 4.5 the charge on the TD + LC domain increased at the steepest rate, while the net charge switched from negative to positive at just below pH 6 (Fig. 5). This resulted in a net positive charge on both domains that increased rapidly below pH 6, and the resulting inter-domain repulsions could explain the sudden surge in flexibility and domain movement relative to each other. The number of salt bridges also decreased rapidly at below pH 5.5 which would further facilitate the loss of stability at the domain interface (Fig. 5). Of particular interest were two areas identified as interacting regions at the BD / LC interface, in which electrostatic interactions were lost between pH 5.5 and pH 5. Key salt bridges included Asp338/Lys834, where the Asp changed from negatively charged to neutral, and Lys323/ Glu1058 where the Glu shifted from negatively charged to neutral. In both cases, the Lys residues remained positively charged. Additionally, a short helix motif conserved from diphteria toxin (Ratts et al., 2005) underwent a strong electrostatic shift between pH 5.5 and pH 5 with the net loss of two negatively charged residues, Glu810 and Asp817 which became neutral upon protonation.

2.2. PCA clustering separates pH ≤ 5 and pH > 5

While the data presented above indicated a shift in the protein behaviour at below pH 5, it was not enough to confirm whether the changes were due to a concerted conformational shift between discrete populations, or whether each peak in RMSD or Re values contained broad distributions of conformations sampled through random motions. To investigate this, the trajectories were combined using only the backbone atoms (C, CA, N) and subjected to PCA to evaluate the conformational clusters present in the combined trajectories.

All 24 MD trajectories were saved as backbone-only files (C, CA, N), and then concatenated together into a single file of 70,004 frames per pH, that included only the last 350 ns of the trajectories, with a timestep of 40 ps. Most of the variance could be reasonably accounted for using...
Fig. 3. Average Solvent Accessible Surface Area of BoNT/E. (a) represents the average SASA of the full BoNT/E at each pH, with one standard deviation error bar. (b) represents the SASA broken down by domain by summing the SASA of all residues within the domain, with one standard deviation error bar. SASA values in both cases are the averaged SASA from the last 200 ns of four repeats per pH.

Fig. 4. TD of BoNT/E, coloured by SASA difference with respect to pH 7. (a) represents the location of the highest SASA change with respect to pH 7 (blue = largest negative change, red = largest positive change). (b) and (c) are zoomed visuals on the short helix region which had the largest increase in SASA at pH 4 to 5 (b) and a structure where the secondary structure is lost (c), exposing a lysine residue (white (b)/red (c)) K618. This conformation was exclusively found in pH 4 to 5 simulations. The SASA difference with respect to pH 7 was taken from the average of four repeats per pH.
only 4 clusters (Supplementary Fig S3). The PCA biplots and scree plot are presented in Supplementary Fig S4. Supplementary Fig S5 presents the contribution of each pH trajectories to the total frames per cluster (Supplementary Fig S5a) and the contribution of each cluster to the total frames per pH (Supplementary Fig S5b).

We found that at pH 5.5, 6 and 7, the BoNT MD conformations remained primarily in structural cluster 1 for the entire 350 ns. Only a single repeat at pH 5.5 explored cluster 4, and a single repeat at pH 6 explored cluster 3, each with no effect on the $R_G$ of the protein. The structures from cluster 1 were observed at every pH, but their contribution decreased from 97% at pH 7, to 75% at pH 5.5 and pH 6.0, and then again to 50–54% at pH 5.0 and below. By contrast, structures from cluster 2 contributed increasingly to the total frames per pH as it decreased from pH 5 (25%), to pH 4.5 (40%), and then pH 4.0 (50%), thus indicating overall a gradual switch from cluster 1 to cluster 2 at below pH 5.5. To expand on this, it is important to understand what each cluster represents and how robust the clustering is. To determine this, the average $R_G$ with respect to the cluster centroid was calculated as shown in Table 1, where the centroid was the structure that is closest to all the other structures within the cluster. Cluster 2 had the highest $R_MSD$ with respect to its centroid, meaning that there was more variation within that cluster than in the others. However, the average $R_MSD$ of this cluster remained below 0.6 nm, demonstrating that the structures placed into it were all closely related. The other clusters were also reliable, with an average $R_MSD$ lower than 0.4 nm with respect to their centroids. These clusters are presented as a dendrogram in Supplementary Fig S6, and as surface representations in Fig. 6b. In this representation, the LC + TD selection is aligned and shows the major movements of the BD. Clusters 1 and 2 dominate the distribution and may form two ends of a pH dependent shift required for activity. Clusters 3 and 4 are less populated and may be transient states.

Combining the clustering data with the $R_G$ and $R_MSD$ data per frame, the average $R_G$ and $R_MSD$ values per cluster were calculated (Table 1). This, in combination with the mid-point $R_MSD$ values, confirms the robustness of the clustering, with low standard deviations with respect
194

to the average. Supplementary Fig S7 shows the population weighted average RMSD and $R_g$ values per pH, and confirms the more extended state of the protein starting from pH 5 down to pH 4.

2.3. SAXS validates the structural change seen in the MD

The MD simulation analyses were validated using experimental SAXS measurements, which provide insight into the solution structure of proteins and can detect conformational differences. Samples were dialysed into six different pH buffers, concentrated, and immediately loaded onto the HPLC-SEC system which feeds directly into the path of the beam. The resulting SAXS data is presented in Fig. 7a,b. The experimental I(Q) curves for most of the samples represented a compact globular structure, with a single peak in the Kratky plot and flat tail at the higher Q-values.

$$Q = (4\pi \sin(\theta))/\lambda$$

With $2\theta$ the scattering angle, and $\lambda$ the wavelength of the beam. At pH 4 however, two elution peaks were observed (Fig. 7c), which produced two different scattering curves. The first peak matched the data from all the other conditions and implied a similarly compact structure. The second peak had a steeper low-Q region in the I(Q) vs Q plot, indicating an increased $R_g$ of the protein, and a tail in the Kratky plot which indicated that a second domain was now “visible” by the beam, with its own scattering profile.
Using the SASSIE web server (Curtis et al., 2012), theoretical SAXS curves were generated from each trajectory and matched against experimental curves obtained at the Diamond Light Source B21 beamline using SEC-SAXS. The best-fit curves were obtained by loading all 160,004 frames per pH from the MD simulations onto the SASSIE server, and using an R-factor fit to the experimental curve (Supplementary Figs S8a,b).

\[ R = \left( \sum |B(I(Q)) - Exp(I(Q))| / \sum |Exp(I(Q))| \right) \]

With \( B(I(Q)) \) the theoretical best fit intensity from the model, and \( Exp(I(Q)) \) the experimental intensity at the same Q value (Curtis et al., 2012; Press et al., 2007).

An initial screen of the backbone-only trajectory identified a best-fit structure that was then extracted with full atoms and a hydration shell included. This decreased further the R-factor between the theoretical curve and the experimental data. It was found that a 2.25 Å hydration shell best fit the open configuration, while a 1.6 Å shell best suited the closed configuration, to give final R-factors of 5.6% and 1% respectively. Fig. 8 shows a superposition of the 100 best fit structures for both conformations. The 100 best-fit conformations of the closed conformation corresponded closely to cluster 1 as identified by PCA, while the partially open conformations corresponded closely to the cluster 2 structures found predominantly within trajectories at pH 4 and pH 4.5. Both the experimentally observed structures, closed and partially open, had distinct pair distance distribution functions (P(\( r \))) with the best fit structures generating good fits in real space (Supplementary Fig S9).

3. Discussion

This joint MD and scattering study has notably clarified the conformational effect of pH on the structure of BoNT/E. Recent advances in computing power and increasingly accurate tools for MD have made simulations a key tool for major discoveries (Bai et al., 2021; Wang, 2021). In the work presented here, MD simulations have been used to extensively study the conformational space available to BoNT/E, with a total of 840,048 individual conformations generated. An overall trend was identified whereby at pH 5 and below, the protein exhibited much higher flexibility with a number of trajectories in the pH range 4 to 5 exhibiting higher \( R_G \) values, and transitioned from one conformation (cluster 1) to another (cluster 2) over time. By contrast, at pH 5.5, 6 and 7 the \( R_G \) of BoNT/E remained constant throughout all 4 repeats, and just three trajectories displayed a conformational change detected by RMSD analysis. By separating the protein into two distinct selections (BD as one and TD + LC as another), it was shown that the conformational change was not within any of the domains, but derived from a translational movement of one domain with respect to the other, at the lower pH range. Cases of higher RMSD values that did not simultaneously affect the \( R_G \) appeared at pH 5, 5.5 and 6, but were rarely populated as determined by PCA, and primarily derived from only one or two trajectories in each case. Given the extensive changes observed in the lower pH range, and the absence of any notable changes in any of the trajectories at pH 7, the simulations were terminated at a total of 400 ns, with analysis excluding the first 50 ns for equilibration. The typical timescale of domain movements in protein MD is \( 10^{-11} \) to \( 10^{-8} \) s (Skjærven et al., 2011), and falls within the timeframe presented here. The previous MD studies identified were closer to the lower limit of this range (Chen and Deng, 2007; Chen et al., 2007; Wang et al., 2015), and only focused on BoNT/A or single domains of BoNT/A (Evander Emeltan Tjoa et al., 2019). No similar computational

Fig. 7. \( \ln(I(Q)) \) vs Q and Kratky plots. (a) \( \ln(I(Q)) \) vs Q of BoNT/E at different pH conditions, with the best-fit curves to both structures. pH 4 showed two distinct populations which eluted at different times, and resulted in two different scattering profiles (shown in Fig. 7c). (b) Kratky plot displaying the difference between the two populations, with the higher tail in the red & black curves indicating a separation of two domains, compared to the lower tail in the other curves (blue, green, purple, yellow, cyan & brown). (c) HPLC trace of the elution prior to SAXS analysis. Most conditions eluted in a single peak, while pH 4.5 and pH 4 showed several. The highest peak in pH 4 & pH 4.5 was found to be aggregated protein; however in pH 4 a smaller population distinct from this peak was identified and matched a distinct conformation.
studies were identified focusing on BoNT/E. By performing more repeats for a longer duration, and on the whole protein alone in solution, we have presented here a deeper look into the conformational landscape of BoNT/E; with some results in good agreement with previous studies of BoNT/A (see below).
In order to explain the changes in behaviour at the switch to pH 5.0 and below, a transition in the charge distribution was observed, with a large drop in salt-bridge content and a large increase in positive charge on both domains at the lower pH range. It was therefore likely that the greater number of salt-bridges, combined with lower but opposing net charges on the two domains, were important for stability of the compact globular state at pH 6 and 7. Starting at pH 5.5, the salt-bridge content decreased to a point where some inter-domain movement was possible, and at pH 5 the sudden increase in positive charge on both domains increased the repulsion between them. This resulted in a much broader space explored in those simulations, with most trajectories moving in and out of more extended conformations. This transition at pH 5.5 and below matched the known activation pH required for this protein in vivo, with many studies demonstrating the absence of catalytic activity at above pH 5 (Keller et al., 2004; Pirazzini et al., 2011). It is therefore likely that a major conformational change involving the separation of the BD from the LC domain is required for activity. This change also increased the SASA of the whole protein; and specifically the TD, which transitions into a transmembrane channel through which the LC is released into the cytosol. Notably, a small region was detected with a large increase in the SASA at and below pH 5 which has homologues in all serotypes, as well as in TeNT. This small region was found to be crucial for membrane insertion in BoNT/A (Lam et al., 2018). Further analysis or mutagenesis of this region may reveal deeper insight into the structural change enabling membrane insertion in BoNT/E. Furthermore, due to the importance of this region for this protein’s toxicity, it could become a key area to target with antibodies in order to prevent the membrane interaction.

To validate the population statistics and dynamics observed from the MD simulations, the structures associated with each cluster were compared to experimental SAXS data obtained for each species separated by SEC at each pH (Fig. 7c). The high-throughput server SASSIE was used to generate theoretical SAXS curves from every structure obtained (Perkins et al., 2016; Sarachan et al., 2013). These were then compared to experimental curves obtained at Diamond Light Source, using R-factor to quantify the goodness of fit. Two scattering profiles were obtained experimentally at pH 4, separated as two peaks by SEC, indicating two
distinct conformations present in solution. The two conformations matched the scattering profiles of a compact globular protein in one case, and in the other a more extended double-domain structure. These two experimental curves were compared to the theoretical curves from the MD, and good fits were found, with R-factors of 1% and 5.6%, respectively for the closed and open conformations. The increased depth of the hydration shell around the open conformation is in good agreement with the increased surface charge associated with it. The best fit to the closed conformation came from a structure very close to the crystal structure of BoNT/E (PDB ID: 3FF2) (Kumaran et al., 2009), while the best fit for the open conformation came from cluster 2 identified by PCA.

Taken together we conclude that the BoNT/E begins its conformational journey in the compact form (i.e. within cluster 1) at physiological pH, and ends it in the open form (i.e. within cluster 2) at lower pH, some of the other structures identified in the MD are likely to be intermediate states through which the protein passes (clusters 3 and 4), before reaching a final state for creating the transmembrane channel. The separation of the two domains adjacent to the TD indicates that it is left more exposed, which may facilitate the formation of the transmembrane channel necessary for the activity of this toxin; while localised structural changes within the TD favour membrane interaction by exposing hydrophobic residues.

4. Materials and methods

4.1. Molecular dynamics simulations

The MD simulations were all started from the PDB structures (PDB ID: 3FF2) (Kumaran et al., 2009), from which a single copy of BoNT/E was extracted through text editing of the PDB. Due to the nature of X-ray crystallography, some of the more flexible regions in the protein structure were unresolved. These missing residues were rebuilt using SwissPDB Viewer (Guex and Peitsch, 1997), which infers the positions of the missing residues using residues in the same positions in homologous structures as a reference. The complete structure was uploaded to the PDB2PQR web server (Dolinsky et al., 2004) to simulate the protonation states of each ionisable group at each required input pH, and add hydrogen atoms to titrable residues using the PropKs (Olsson et al., 2011) algorithm. The file generated (.pqr) was converted back to a PDB file, and prepared for a standard MD simulation in Gromacs (Abraham et al., 2014; Van Der Spoel et al., 2005). The Amber03 (Duan et al., 2003) force field was used, as it included a topology and parameters for the Zinc ion present in the LC and has been reported to produce results in good agreement with experimental results on well-studied proteins (Hornak et al., 2006; Hu and Jiang, 2010). The protein was centred in a cubic box such that there was 2 nm between any periodic images of itself, solvated with water using the SPC/E model (Berendsen et al., 1987), and enough solvent molecules replaced with counter-ions to neutralise charge within the system. After setting up the system, it was energy-minimized using a steepest descent algorithm, temperature equilibrated at 300 K using a modified Berendsen thermostat, and pressure equilibrated at 1 bar using Parrinello-Rahman coupling. Hydrogen bonds were constrained during the temperature and pressure equilibration steps. The minimized, temperature and pressure equilibrated system was then left unrestrained and simulation trajectories were generated up to 400 ns, with a snapshot saved every 10 ps. Four repeats were obtained for each condition of pH 4, 4.5, 5, 5.5, 6 and 7.

The trajectories were generated using Gromacs 5.0.4 on an external High Power Computer (HPC) and were corrected for periodic boundary condition artefacts using “gmx trjconv” on Gromacs 2018.4, after transferring the data to a local computer. Once all the trajectories were complete, a file with backbone atoms only (C, CA, N) was generated from the 50 ns mark; and every MD trajectory (24 in total) was concatenated into a single file using VMD 1.9.3 (Humphrey et al., 1996), comprising 840,048 frames. A tool command language (Tcl) script, compiled from existing commands in VMD 1.9.3 and looped over every frame of the merged trajectory, was used to align the protein backbone to that of the BoNT/E (PDB ID: 3FF2) structure, measure the RMSD and RC of this; then align the BD alone and measure the RMSD of two selections (BD, and TD + LC) separately; then align the TD + LC, and measure the RMSD of the two selections separately. All images of the protein were generated in PyMOL (Schrodinger, LLC, 2015) or VMD 1.9.3.

4.2. Protein production and purification

A plasmid containing an endopeptidase-negative (i.e. non-toxic) BoNT/E DNA sequence was provided by Ipsen. The plasmid was transformed into E. coli BL21(DE3) (New England Biolabs) cells, and grown in modified terrific broth (mTB), using Kanamycin at 30
µg/mL. An overnight starter culture was used to inoculate larger cultures, which were grown at 37 °C until OD₀₆₀₆ reached 0.6. The temperature was lowered to 16 °C and once cooled, 1 mM IPTG was added for induction. The induced culture was incubated for a further 21 h, after which the cells were harvested by centrifugation at 3,196 RCF. The pellets were re-suspended with 3 mL lysis buffer (35 mM NaCl, 50 mM Tris, pH 8.0) per gram of cell paste. The cells were lysed by sonication with 10 cycles of 30 s bursts, 30 s rest. The lysed mixture was centrifuged at 19,802 RCF for 30 min. The supernatant was then diluted with an equal volume of 2 M ammonium sulfate, 50 mM Tris, pH 8.0; and centrifuged again at 19,802 RCF for 30 min. A Butyl column (hydrophobic interaction) equilibrated with 1 M ammonium sulfate, 50 mM Tris, pH 8.0 was used to capture the protein, using a 50 mM Tris, pH 8.0 elution buffer. The relevant elution fractions were pooled and desalted using a 53 mL desalting column, exchanging into 10 mM Tris, pH 8.0. The desalted sample was loaded onto a Q column (ion exchange), and eluted using 500 mM NaCl, 10 mM Tris, pH 8.0. A high purity fraction of this elution was isolated and concentrated/diluted to 0.5 mg/mL. It was then pooled with 5 µg Trypsin per mg protein at 37 °C for 40 mins, to cleave the link between LC and HC (“activation” step). The activated sample was then mixed with 1.6X its volume of 5 M NaCl, 10 mM Tris, pH 8.0 and loaded onto a Butyl column equilibrated with 3 M NaCl, 10 mM Tris, pH 8.0. The final product was highly pure, activated (but still safe, endopeptidase-negative) BoNT/E and was eluted using 10 mM Tris, pH 8.0. The purity was assessed by SDS-PAGE, using 4–12% Bis-tris gels (Thermo-Fisher) and the activation was assessed by adding β-Mercapto- Ethanol in one of the lanes, which separated the band at 150 kDa into two bands at 100 and 50 kDa, respectively.

4.3. SAXS

The purified protein sample was buffer exchanged into six different buffers, at the six different pH values, using 10 mM sodium acetate for pH 4, 4.5, 5, and 5.5, and 10 mM sodium phosphate for pH 6 and 7. All the buffers were made at 20 mM ionic strength using NaCl. The final purified sample was diluted or concentrated to 0.5 mg/mL and separated into 0.5 mL aliquots. Each aliquot was then buffer exchanged against 500 mL of the appropriate buffer in a coldroom overnight, using 0.5 mL Slide-A-Lyser dialysis cassettes (Thermo-Fisher) with a 20 K MWCO.

Samples were concentrated to ≈ 4 mg/mL using spin concentrators, loaded onto a 96-well plate for autosampling into a SEC-HPLC (KW 403, SHODEX) with buffer-matched mobile phase to each sample, then directly into a temperature-controlled quartz cell capillary with a diameter of 1.5 mm of X-ray scattering instrument B21 at the Diamond Light Source (Harwell Science and Innovation Campus, Didcot, UK), operating with a ring energy of 3 GeV, and an operational energy of 12.4 keV. This recorded 3 s snapshots for 32 min, while the sample was loaded and eluted off the SEC. The start of the elution, corresponding to buffer, was subtracted from the sample signal. Buffer subtraction was performed on ScÂtter 3 (BIOSIS) (Rambo, n.d.); and the generated scattering curves saved for analysis. Theoretical scattering curves were generated from the MD structures using the SASCALC module on the SASSIE web-server (Curtis et al., 2012; Watson and Curtis, 2013) in order to find a structure which fit the experimental curve best. The raw trajectories from Gromacs were converted to .dcd format using VMD, and loaded onto the SASSIE server with a reference PDB at the relevant pH. For analysis, the best fit structure was assessed by R-factor. A hydration shell around the protein was included using the explicit solvent of the trajectories for the best-fit frame obtained by the full-atom trajectory analysis.

To facilitate comparison between all the experimental curves and the theoretical curves generated using SASSIE, the experimental data was interpolated to a fixed number of Q points using a MATLAB script. To ensure a good fit to the true data, 800 Q points were used as this was close to the actual number of Q values collected experimentally. All the intensities of every sample were then normalised by dividing by their intensity at Q = 0.05 nm⁻¹. This maintained the shape of the curve, while allowing for comparison between each condition. Analysis of ln(I(Q)) vs Q curves and Kratky plots, provided information about the approximate size and shape of the protein. 4.4. Principal component analysis

Principal component analysis was used to analyse the MD results in increased detail. For this, the raw trajectories (in .xtc format) from Gromacs were converted to the .dcd format using VMD, and only backbone atoms were kept in the trajectory, so as to remove all the solvent molecules. A stride of 2 frames was used, starting from 50 ns, leaving 17,501 frames per MD trajectory,
with a frame every 20 ps. Once this reduced trajectory was generated for every pH repeat, a final file with all 24 MD trajectories was compiled by concatenating all the backbone-only files together. The PCA was performed using the Bio3D package (Grant et al., 2006) with the “fastcluster” library (Mühlner, 2013) to expand the power of the “hclust” function. A first script helped determine a sensible number of clusters to use, based on how much of the variance was covered in using varying numbers of clusters. Using the determined appropriate number of clusters, a second script performed the PCA and clustered the data accordingly. This data was sorted into bins for in-depth analysis. Six clustering algorithms were trialled (single, median, centroid, mcquitty, complete and average) (Mühlner, 2013); but only the “average” algorithm was taken forward as the best method (see Supplementary information), using 4 clusters.

Significance Statement

With the growing therapeutic potential of BoNTs, it is paramount to understand their complete mechanism of action, and their structural behaviour in formulations. Specifically, the importance of the pH drop during the vesicle acidification step, and its effect on the conformation of the toxin, remains elusive. Here we have identified a major conformational state of BoNT/E when the pH falls below 5.5, using a combination of MD simulations and SAXS. This conformational shift could be critical in facilitating the formation of the BoNT transmembrane channel, essential for function.

Funding information

This research was jointly funded by Ipsen, and the Engineering and Physical Sciences Research Council (EPSRC) Centre for Doctoral Training in Emergent Macromolecular Therapies (EP/L015218/1) for financial support of C.J.L.

Data Availability Statement

Upload of SAS data to SASDB: https://www.sasbdb.org/data/ SASDNV5/ MD data available on request.

VMD and Bio3D scripts used available at https://www.github.com/ucbecla/BoNT-E_PCA_VMD.git. CRediT authorship contribution statement

Christophe J. Lalaurie: Investigation, Data curation, Formal analysis, Writing – original draft, Editing. Andrew Splevins: Conceptualization, Funding acquisition, Study coordination, Supervision. Teresa S. Barata: Conceptualization, Funding acquisition, Study coordination, Supervision. Karen A. Bunting: Conceptualization, Supervision, Validation, Writing – review & editing. Daniel R. Higazi: Conceptualization, Supervision, Validation, Writing – review & editing. Mire Zloh: Conceptualization, Supervision, Validation, Writing – review & editing. Valentina A. Spiteri: Resources, Data curation, Formal analysis for the SAXS data, Writing – review & editing. Stephen J. Perkins: Resources, Data curation, Formal analysis for the SAXS data, Writing – review & editing. Paul A. Dalby: Project administration, Conceptualization, Funding acquisition, Supervision, Validation, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. K. A. Bunting and D. Higazi are employed by IPSEN.

Acknowledgments

C.J.L. was supported by the EPSRC Centre for Doctoral Training in Emergent Macromolecular Therapies (DPT00033549) and IPSEN Bioinnovation. S.J.P. was supported by a joint EPSRC (EP/K039121/1) and NSF (CHE01265821) grant for CCP-SAS. We thank Dr. Akash Pandya for help in setting up the MD work; Dr Katsuki Inoue for setting up the HPLC SEC-SAXS measurements on Instrument B21 at Diamond Light Source; and Dr Kevin Moore for his input on optimizing the protein purification procedures.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jsb.2022.107876.

References


