Drug discovery for misfolding diseases using structure-based iterative learning

Robert Imrie Horne

Department of Chemistry

University of Cambridge

This dissertation is submitted for the degree of

Doctor of Philosophy

Pembroke October 2023

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. I further state that no substantial part of my thesis has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. It does not exceed the prescribed word limit for the relevant Degree Committee.

Robert Imrie Horne

October 7th, 2023

Acknowledgements

I have Professor Michele Vendruscolo to thank for giving me all of my major life opportunities to date, including bringing me into this field with a Master's project that I hugely enjoyed, proposing my involvement with Wren (now Wavebreak) Therapeutics, connecting me with numerous exciting collaborators during my PhD and being a huge help in supporting my next steps. Thank you so much for your support and faith in me throughout this PhD and beyond. Professor Vendruscolo and Dr Johnny Habchi (Wavebreak VP) invited me to work at Wren after the Master's project, leading a drug discovery project for Parkinson's Disease, where I was able to learn most of the experimental skills that were then used in this PhD. Especial thanks go to Dr Xiaoting Yang, who has been a great friend and mentor and taught me the largest portion of what I know about protein misfolding kinetics and (alongside Dr Chen Gao) the entire portion of what I know about making dumplings. A big thank you also to Dr Sean Chia, a walking biophysical textbook, who taught me everything else and helped me with my first tentative steps back into both academia and online gaming during lockdown (thank you also for the gym workout routine). Shout outs also go to Dr Roxine Staats, soon-to-be-Dr Jo Menzies, Dr Benedetta Mannini and Dr Johnny Habchi for the scientific know how and life know how.

This PhD project was essentially trying to machine learn my former job to make it more efficient, which appealed to me a great deal as screening tens of thousands of molecules was a challenge. For laying the foundational groundwork and helping throughout I have to thank Dr Z. Faidon Brotzakis, whose work on in silico docking simulations of small molecules to protein targets formed the basis of nearly all of my projects. I would also like to thank Dr Andrea Possenti for his invaluable advice on which ML architectures to use for the task. For moral support and frequent snack breaks I have Magda Nowinska to thank, who is as awesome a friend as she is a scientist, I'm looking forward to dropping in on her Barcelona lab for coffee breaks in the future. Dr Becky Gregory also made all the experiments possible by synthesising nearly all the proteins needed, for which she will have my eternal gratitude. Ewa Andrzejewska had great patience with my impatience and comparatively reckless approach with the μFFE project, which enabled us to get some really nice results.

I have a small army of other collaborators within and without the Vendruscolo lab to thank for a lot of the results that follow, including my Part III students, Jared Wilson-Godber and Raghav Chandra, who started off generative modelling work with α-synuclein, and created a very nice pipeline for identification of specific binders, respectively. A number of MPhils have also helped me out including Mhd Hussein Murtada, who did an exemplary job creating an exploration pipeline that incorporated measures of molecule likelihood of reaching the CNS. Alice Aubert also did work on model explainability. Donghui Huo from the Beijing University of Chemical Technology also did some lovely work improving on my initial drug discovery pipeline with a generative model optimised for small data sets.

Although most of my work has been on α-synuclein, I also have team tau (Alessia Santambrogio, Dr Michael Metrick and Dillon Rinauro) to thank for reproducing a lot of the work for alpha-synuclein on the tau protein, which has been very successful so far. Across the Atlantic, Parvez Alam from the Caughey lab, Montana, is also owed many thanks for grounding the projects with physiological relevance via a diseased brain seeded aggregation assays, as are the Caughey and Ghetti groups for facilitating these experiments. Other collaborations that have led to cool results include work with the Roy group in Toronto, the Vekrellis group in Athens, the Skoulakis group in Vari, the de Simone group in Napoli, the Aigbirhio group in Addenbrookes, Cambridge, and the Keyser group in the Cavendish Laboratory, Cambridge. A huge thank you also to all the other people I've worked with in the Centre for Misfolding Diseases, including of course the all of the Vendruscolo group, the Knowles group (Ewa, Zenon, and Georg and Catherine the academic power couple) and the Bernardes group (Nai-Shu you are too good for this world).

The collaboration with the Keyser group was the best though, not because of the amazing results we got but because I got to do it with my favourite person, Sarah Sandler. Bumping into you was the luckiest encounter of my life so far, and it's been an awesome adventure ever since. I'm super glad that you immediately wanted to party after 2 weeks of Covid isolation, you are a ball of energy that counters my inertia wonderfully. Thank you for putting up with me for as long as you have. We endured the thesis writing together with our flatmate Teo, who I have to thank for the endless supply of artisanal coffees and generally lovely vibes which kept me going through this experience. A thesis should never be written alone. A thank you and an apology go to my amazing friends from undergrad days and beyond, Thomas, Siobhan, Sean, Grace, Ewan, Rhys, Kat and Kate who I've neglected through this process but who happily

haven’t yet given up on me. My longest running collab however, is with my family, without whom I wouldn’t have achieved any of this. A massive thank you to my parents, Susan and Charlie, who supported and often instigated my nerdy pursuits, and to my siblings, Caitriona and Jamie, who had to deal with an air headed younger brother and so tried to get him on their level (my 8-year-old brain didn’t absorb much of the literary analysis of Marlowe’s Dr Faustus but I’m sure it helped, ditto on the moral evaluation of the allied WW2 bombing campaign). Thank you also to my Gran, Joyce, for being as supportive of us when we were growing up as I hope we were of you when you were growing old.

Abstract

Computational methods such as machine learning hold the promise to reduce the costs and the failure rates of conventional drug discovery pipelines. This issue is pressing for neurodegenerative diseases, where the development of disease-modifying drugs has been particularly challenging. The high attrition rate of neurodegenerative drug discovery is especially acute for Parkinson's disease, where no disease-modifying drugs have yet been approved. Numerous clinical trials targeting α-synuclein aggregation, a process implicated in Parkinson's disease and other synucleinopathies, have failed, at least in part due to the challenges in identifying potent compounds in preclinical investigations. In **Chapter 2**, I describe machine learning approaches to identify small molecule inhibitors of α-synuclein aggregation to address this problem. Because the proliferation of α-synuclein aggregates takes place through autocatalytic secondary nucleation from fibril surfaces, we aim to identify compounds that bind the catalytic sites on the surface of the mature fibrillar aggregates (the end point polymers of the aggregation process). This prevents the formation of the toxic intermediate aggregate species, termed misfolded oligomers. Fibrils assume different structural polymorphs depending on the synucleinopathy, likely due to the different locations of the nervous system that these diseases occur within. Each tissue has an associated set of specific conditions which likely shape the final structure of the aggregates. Targeting these pathogenic polymorphs may help ameliorate disease progression more effectively than prior efforts. To achieve this goal, I use structure-based machine learning in an iterative manner to first identify and then progressively optimise secondary nucleation inhibitors. Training data for aggregation inhibition were obtained by an assay specifically isolating secondary nucleation, the major mechanism of toxic oligomer production. My results demonstrate that this approach leads to the facile identification of compounds which are two orders of magnitude more potent than previously reported ones.

This initial work formed the basis of subsequent efforts to both expand the chemical space explored, and explore it more effectively, through application of generative modelling linked with reinforcement learning. I also increased the molecular parameters considered during the process of inhibitor optimisation in **Chapter 3**, accounting for aspects of pharmacokinetics as well as potency. This work addressed a number of shortcomings in the initial approach including restricted chemical space and a focus on potency alone. The initial method was

reminiscent of the early stages of drug development, where large compound libraries are typically screened to identify compounds of promising potency against the chosen targets. Often, however, these compounds have a poor drug metabolism and pharmacokinetics (DMPK) profile, which are negative features that may be difficult to eliminate. To address this, the updated machine learning approach combines generative modelling and reinforcement learning to identify small molecules that perturb the kinetics of aggregation, thus reducing the production of oligomeric species, while also having high predicted blood brain barrier penetrance. This approach resulted in the identification of small molecules with good pharmacokinetic properties and potency against secondary nucleation.

Misfolded protein oligomers generated via secondary nucleation are clearly of central importance in both the diagnosis and treatment of Alzheimer's and Parkinson's diseases. All the methods described here are designed to counter their formation, yet accurate high-throughput methods to detect and quantify oligomer populations are still needed. Invariably bulk aggregation is the metric that is tracked, and the oligomer population is then inferred. In **Chapter 4** I present a novel single-molecule approach to detection and quantification of oligomeric species. The approach is based on the use of solid state nanopores and multiplexed DNA barcoding to identify and characterise oligomers from multiple samples. I study α-synuclein oligomers in the presence of several small molecule inhibitors of α-synuclein aggregation, as an illustration of the potential applicability of this method to assist the development of diagnostic and therapeutic methods for Parkinson's disease.

Finally, having created these pipelines for the development of α-synuclein aggregation inhibitors, I then sought to expand into other protein misfolding areas to demonstrate their generalisability as described in **Chapter 5**. The aggregation of tau into amyloid fibrils is associated with Alzheimer's disease and related tauopathies. Similarly to synucleinopathies, different tauopathies are characterised by the formation of distinct tau fibril polymorphs. Brain homogenates were used to seed the generation of tau fibrils. The aim here was to create fibrils that replicate the polymorph formed in Alzheimer's disease, thus mirroring the pathological aggregation mechanisms as closely as possible. Fibrils recovered from these efforts were capable of converting recombinant 0N3R tau into an Alzheimer's fibril polymorph in a kinetic assay, as verified through cryo-EM structural analysis. Using this kinetic assay, I illustrate the iterative machine learning drug discovery method for tau aggregation in Alzheimer's disease.

Table of Contents

List of Figures

List of Tables

List of publications

Some of the results in this thesis appear in the following publications:

RI Horne[*], SE Sandler[*], S Rocchetti, NS Hsu, MC Cruz, ZF Brotzakis, RC Gregory, R Novak, S Chia, G Bernardes, UF Keyser, M Vendruscolo, "Multiplexed, Digital Characterisation of Modulators of Protein Oligomer Flux via Solid-State Nanopores", *J. Am. Chem. Soc.* 2023 (**published**)

RI Horne[*], MA Metrick II[*], A Santambrogio[*], T Löhr[*], N Gallagher, ZF Brotzakis, D Rinauro, L Sakhnini, S Linse, B Caughey, M Vendruscolo, "A brain-seeded fibril amplification models the aggregation process of tau in Alzheimer's disease for drug discovery", 2023 (**in-preparation**)

RI Horne, E Andrzejewska, P Alam, ZF Brotzakis, A Aubert, M Nowinska, RC Gregory, R Staats, A Srivastava, A Possenti, S Chia, P Sormanni, B Caughey, TPJ Knowles, M Vendruscolo, "Discovery of Potent Inhibitors of α-Synuclein Aggregation Using Structure-Based Iterative Learning", *Nat. Chem. Bio.*, 2023 (**accepted**)

RI Horne[*], MH Murtada[*], D Huo[*], ZF Brotzakis, RC Gregory, A Possenti, S Chia, M Vendruscolo, "Exploration and Exploitation Approaches Based on Generative Machine Learning to Identify Small Molecule Inhibitors of α-Synuclein Secondary Nucleation", *J. Chem. Theory Comput.,* 2023 (**published**)

R Staats, ZF Brotzakis, S Chia, **RI Horne**, M Vendruscolo, "Structure-based optimization of a small molecule that inhibits the proliferation of α-synuclein aggregates", *Front. Mol. Biosci.,* 2023 (**published**)

S Chia, ZF Brotzakis, **RI Horne**, A Possenti, B Mannini, R Cataldi, M Nowinska, R Staats, S Linse, TPJ Knowles, J Habchi, M Vendruscolo, "Structure-Based Discovery of Small-Molecule Inhibitors of the Autocatalytic Proliferation of α-Synuclein Aggregates", *Molecular Pharmaceutics*, 2022 (**published**)

Not included in this thesis:

M Brezinova, ZF Brotzakis, **RI Horne**, M Vendruscolo "Structure-based discovery of small molecule inhibitors of Aβ aggregation from an ultra-large chemical library using Deep Docking", 2023 (**in-preparation**)

S Zhang, D Huo, **RI Horne**, Y Qi, SP Ojeda, A Yan, M Vendruscolo "Sequence-based drug design using transformers", 2023 (**in-preparation**)

M Kamal, J Knox, **RI Horne,** AR Burns, OS Tiwari, D Han, DL Bar-Yosef, E Gazit, M Vendruscolo, PJ Roy "A Rapid *in vivo* Predictor of Amyloid-Disrupting Small Molecules", *Nat. Commun.*, 2023 (**in-review**)

AJ Dear, X Teng, SR Ball, J Lewin, **RI Horne**, D Clow, N Harper, K Yahya, TCT Michaels, S Linse, TPJ Knowles, X Yang, SC Brewerton, J Thompson, J Habchi, Georg Meisl "Molecular mechanism of α-synuclein aggregation on lipid membranes revealed.", *Chem. Sci.*, 2023 (**in submission**)

R Chandra, **RI Horne**†, M Vendruscolo†, "Bayesian Optimisation for Selective Drug Generation", *J. Chem. Theory Comput.,* 2023 (**published**)

RI Horne*, MA Metrick II*, W Man, D Rinauro, G Meisl, M Vendruscolo "Secondary Processes Dominate the Quiescent, Spontaneous Aggregation of α-Synuclein at Physiological pH with Sodium Salts", *ACS Chem. Neurosci.*, 2023 (**published**)

RI Horne*, J Wilson-Godber*, ZF Brotzakis, RC Gregory, A Possenti, S Chia, M Vendruscolo, "Using Generative Modeling to Endow with Potency Initially Inert Compounds with Good Bioavailability and Low Toxicity", *J. Chem. Inf. Model.*, 2023 (**published**)

FS Ruggeri, J Habchi, S Chia, **RI Horne**, M Vendruscolo, TPJ Knowles, "Infrared Nanospectroscopy Reveals the Molecular Interaction Fingerprint of an Aggregation Inhibitor with Single Ab42 Oligomers", *Nat. Commun.*, 2021 (**published**)

R Staats, TCT Michaels, P Flagmeier, S Chia, **RI Horne**, J Habchi, S Linse, TPJ Knowles, CM Dobson, M Vendruscolo, "Screening of small molecules using the inhibition of oligomer formation in α-synuclein aggregation as a selection parameter", *Commun. Chem.*, 2020 (**published**)

*Authors contributed equally to this work

†Co-corresponding authors

Abbreviations

Aβ42: Amyloid Beta 1-42 protein

AD: Alzheimer's Disease

ALS: Amyotrophic Lateral Sclerosis

APD: Action Probability Distribution

APP: Amyloid Precursor Protein

αS: α-Synuclein

ATP: Adenosine Triphosphate

ATR-FTIR: Attenuated Total Reflection Fourier Transform Infrared

AUC: Area Under Curve

BAR: Best Agent Reminder

BBB: Blood Brain Barrier

BCA: Bicinchoninic Acid

BFS: Breadth First Traversal

CBD: Corticobasal Dementia

CLM: Chemical Language Model

LogP: Partition Coefficient

CNS: Central Nervous System

CP: Cell Painting

CTG: CellTiterGlo

CVD: Cerebrovascular Disease

DBCO: Dibenzocyclooctyne

DMEM: Dulbecco's Modified Eagle's Medium

DLB: Dementia with Lewy Bodies

DMPK: Drug Metabolism and Pharmacokinetics

DMPS: 1,2-dimyristoyl-sn-glycero-3-phospho-L-serine (sodium salt)

DQN: Deep Q-Network

DMSO: Dimethyl Sulfoxide

DTT: Dithiothreitol

ECD: Event Current Deficit

EDTA: Ethylenediaminetetraacetic Acid

ELISA: Enzyme-Linked Immunosorbent Assay

EM: Electron Microscopy

FCS: Fluorescence Correlation Spectroscopy

FDA: Food and Drug Administration

FRED: Fast Rigid Exhaustive Docking

GAN: Generative Adversarial Network

GGNN: Graph-Gated Neural Network

GPR: Gaussian Process Regressor

GuHCl: Guanidinium Hydrochloride

GUI: General User Interface

HEPES: 4-(2-Hydroxyethyl)-1-Piperazineethanesulfonic Acid

hiFBS: Heat Inactivated Foetal Bovine Serum

HTS: High Throughput Screening

IAPP: Islet Amyloid Polypeptide

IDP: Intrinsically Disordered Protein

IPA: Isopropanol

IPTG: Isopropyl β-D-1-Thiogalactopyranoside

ISF: In Silico Screening Fragment

JTNN: Junction Tree Neural Network

K_D: Equilibrium Dissociation Constant

KIC_{50}: Kinetic Inhibitory Constant. Concentration of inhibitor at which aggregation half time is extended by 50%

LB: Lysogeny Broth

LCMS: Liquid Chromatography Mass Spectrometry

LR: Linear Regressor

LSTM: Long Short Term Memory

MAE: Mean Absolute Error

MLP: Multi-Layer Perceptron

MNN: Message Neural Network

MPO: Multiparameter Optimisation

MSA: Multiple System Atrophy

MWCO: Molecular Weight Cut-Off

(R)MSE: (Root) Mean Squared Error

µFFE: Micro Free Flow Electrophoresis

NBS: Non-Binding Surface

NDA: New Drug Application

NIH: National Institutes of Health

PAGE: Polyacrylamide Gel Electrophoresis

PAINT: Points Accumulation for Imaging in Nanoscale Topography

PBS: Phosphate Buffered Saline

PCA: Principal Component Analysis

PCR: Polymerase Chain Reaction

PD: Parkinson's Disease

PDB: Protein Data Bank

PDMS: Polydimethylsiloxane

PEG: Polyethylene Glycol

PET: Positron Emission Tomography

PHF: Paired Helical Filaments

PiD: Pick's Disease

PK: Pharmacokinetics

PMSF: Phenylmethylsulfonyl Fluoride

PTFE: Polytetrafluoroethylene

QED: Quantitative Estimate of Drug-likeness

QTOF: Quadrupole Time of Flight

RBF: Radial Basis Function

ReLU: Rectified Linear Unit

RF: Random Forest

RFR: Random Forest Regressor

RMSE: Root Mean Square Error

ROC: Receiver Operating Characteristic

RT: Room Temperature (~298 K)

RT-QuIC: Real-Time Quaking-Induced Conversion

(Q)SAR: (Quantitative) Structure Activity Relationship

SDS: Sodium Dodecyl Sulfate

SEC: Size Exclusion Chromatography

SHAP: Shapley Additive Explanations

SMILES: Simplified Molecular-Input Line-Entry System

SPAAC: Strain Promoted Azide-Alkyne Cycloaddition

SPR: Surface Plasmon Resonance

SUV: Small Unilamellar Vesicle

TCCD: Two-Colour Coincidence Detection

TCEP: Tris (2-Carboxyethyl) Phosphine

TCSCP: Time-Correlated Single Photon Counting

TDMS: Technical Data Management Streaming

ThT: Thioflavin T

TIRF: Total Internal Reflection

Fluorescence

TPSA: Total Polar Surface Area

t-SNE: t-Distributed Stochastic Neighbour Embedding

Tris: Tris(hydroxymethyl)aminomethane

UMAP: Uniform Manifold Approximation and Projection

UV: Ultraviolet

VAE: Variational Autoencoder

WT: Wild Type

YT: Yeast Extract Tryptone

1. Introduction

"It seems that when you have cancer you are a brave battler against the disease, but when you have Alzheimer's you are an old fart. That's how people see you. It makes you feel quite alone."
– Sir Terry Pratchett

1.1. Motivation: The perception of neurodegenerative disease

Sir Terry Pratchett was an author of wonderful books who sadly succumbed to Alzheimer's disease in 2015, but not before supplying commentary on the disease with his usual mix of irreverence and insight. It is indeed curious that of 2 disease classes predominantly affecting the elderly, neurodegeneration is viewed so differently to cancer. Due to the array of tools now available to treat cancer, it is often viewed as a fight to be won, whereas neurodegeneration is still considered almost an inevitability of aging. Dementia especially is often looked upon as senility rather than a disease to be treated[1,2]. The reasons for this are likely due to the neurodegenerative onset occurring later in life on average than cancer, but also due to the very different treatment outcomes[3,4]. Although there remain many cancers that elude treatment, indeed it is possible for cancers to obtain resistance to treatment, prognoses are nonetheless significantly more optimistic on average than they have been in the past[5,6]. I emphasise that 'more optimistic' does not mean to imply that cancer is in any way a solved problem. It is a blight on a huge number of lives, which current treatments have alleviated in some cases but by no means all of them. The prognosis for neurodegeneration, however, remains largely unchanged from when the first example of it, Alzheimer's disease, was first described over a century ago[7].

This creates a circular problem. Neurodegeneration is perceived less as a treatable disease than as a product of aging and receives less attention, and so progress towards treatments of neurodegeneration is slow, further entrenching this perception. We have to make an effort to change this perception. The speed at which we find ways to effectively treat neurodegeneration will depend on our determination to discover them. This change will hopefully come about as therapies that have a meaningful impact on disease progression finally make their way into the clinic. After years of costly failures, a handful of drugs have been approved for Alzheimer's disease, though it is true that in some cases these treatments are contentious. Alzheimer's disease is a competitor for the field of scientific research that sparks the most controversy and academic debate. These nascent therapies are the only examples of disease modifying treatment methods that have produced tangible benefits for patients thus far. Their mechanism is the removal amyloid plaques formed in the neurons of patients with Alzheimer's disease[8,9]. This approach is based on the amyloid hypothesis of Alzheimer's disease, whereby select proteins aberrantly misfold and aggregate into misfolded oligomers and amyloid fibrils[10]. Amyloid fibrils are highly ordered fibrous protein aggregates with a cross-β sheet structure[11]. Aggregates accumulate and damage cellular organelles, eventually spreading throughout the brain of the suffering patient. Evidence for this theory is most strongly supported by the fact that familial mutations in the genes encoding these amyloidogenic proteins, or the proteins that interact with them, lead to accelerated onset of disease[12]. Other theories of how the disease progresses began to gain traction following repeated failures of drugs targeting amyloid, but with the successes of the latest clinical trials, the amyloid hypothesis retains its position as the predominant theory[13].

This does not mean that combating amyloid is the only possible treatment route, simply that is a promising one at present that could be augmented with other approaches in future. Protein misfolding and aggregate accumulation are common features of neurodegeneration, but phenotypes vary massively from patient to patient[14]. It seems likely that the various theories for how neurodegeneration begins are labouring a similar point, which is that stresses on the cell's maintenance pathways leads to loss of protein homeostasis and aggregate accrual[15]. These stresses can come from multiple sources. Aging is by far the most common factor, but the stresses that an individual experiences while they age are unique to the individual, and so the maintenance pathways which are most impaired varies accordingly[16,17]. Therefore, multiple treatment or preventative methods may ameliorate neurodegeneration, and what works best is likely to be patient specific. Early diagnosis is essential to keep as many of these options open

as possible and have the best possible prognosis. Great strides have been made in the area as well as in treatment. Methods of diagnosis vary from detecting low levels of the protein aggregates themselves in the cerebrospinal fluid (CSF), or biomarkers of neuronal death such as neurofilament light chain in the blood stream, or even machine learning of smart watch body data[18-20]. However, these methods are not yet widely used, and patients are generally diagnosed once the symptoms appear, when the disease has significantly progressed. This means there is significant accrual of protein aggregates, which are clearly capable of causing toxicity[21]. Inhibition of further aggregate formation and removal of aggregates therefore offers a route to combating the majority of cases.

This research space remains a place of vocal disagreement, but discussion propels development, whereas its lack often gives rise to stagnation. Significant advancements in our understanding are published at increasing frequency, from discoveries of the exact mechanism of neuronal death in Alzheimer's disease to high resolution determination of pathological structures in the brains of suffering patients[21-23]. Failures are of course still likely but can be overcome. For example, many have argued that the first example of an anti-amyloid drug to be FDA approved, aducanumab, did not show significant evidence of disease modification, especially given contradictory results of 2 parallel trials[24]. Aducanumab was in any case effectively rendered defunct shortly after its approval. Medicare, the US federal agency which funds treatments for those over 65, restricted funding for this drug to limited circumstances such as clinical trials due to inconclusive clinical benefit, thus contradicting the FDA's stance[25]. This decision was made in part due to the conflicting data offered in support of aducanumab but also due to the exorbitant price tag, eventually halved in an effort to save the drug, of $56,000 per year. Nonetheless, whatever its clinical efficacy, what aducanumab did achieve is outlining the roadmap (in part by showing what not to do) that such drugs should take to clinical trials, approval, and distribution, as well as galvanising other companies to push their drug programs to completion. Subsequent drug candidates targeting amyloid have fared better than aducanumab as a result. Donanemab, the best drug candidate in this class at the time of writing, slowed clinical decline by 35% in its phase III trial for example, exhibiting a tangible benefit to patients as well as the reduction in amyloid plaques which was observed for aducanumab[26]. Lecanemab has also shown greater efficacy and lower inflammatory response compared to aducanumab[9].

These trials continue to illustrate the unique challenges of modelling complex diseases that progress over decades, where our current treatments produce relatively mild effects. Not least of these is deciding who to enrol, as progressed cases may respond poorly to a treatment aimed at saving neurons that have already been destroyed. Drug candidates tested against such patients show poor responses, which results in clinical failure of therapies that might otherwise have provided a benefit. This problem goes hand in hand with the need for better diagnostics, which will have a key role to play in future trials. There are other roadblocks for neurodegenerative disease research, the most intractable of which are the lengthy duration of disease, and the fact that we do not yet fully understand the steps that lead to disease, and so struggle to model it in a manner that is reliably predictive over a reasonable timescale. Our understanding lags behind that of other disease areas partially because the attitudes described previously are reflected in the investment directed at this area. For every 1 researcher working on dementia there are 4 working on cancer and, despite dementia becoming an issue of similar scale to cancer, its research is chronically underfunded globally by comparison[27]. Investment in neurodegeneration is a far riskier prospect, with the cost of development of an Alzheimer's disease drug estimated to be over 8 times that of a cancer drug on average[28]. Some also believe it to be premature to be considering therapeutic strategies given our current knowledge of disease progression[29]. Nonetheless the scale of the problem is growing and requires rapid solutions. It is time to turn neurodegeneration into a fight to be won.

The scale of that fight is daunting. Alzheimer's disease (AD) and Parkinson's disease (PD) are estimated to affect approximately 50 million and 10 million individuals, respectively, worldwide[27]. Alzheimer's Disease International states that the number of people living with dementia is expected to double in the next 2 decades. Dementia patients take up 1 in 4 of NHS hospital beds in the UK[30], reflective of the need for round the clock care for each person living with this disease, often required for many years. Due to the escalating severity of the issue, we cannot delay therapeutic research but instead attempt to develop treatments in parallel with developing our understanding. Initially, this may result in mildly efficacious treatments, but such offsets are crucial to prevent this problem rapidly becoming an overwhelming one as we seek better alternatives. This makes drug discovery significantly more challenging and prone to attrition, but it is a necessity in order to supply improvements to patient quality of life, and relieve the disease burden on medical services and society as soon as possible.

I entered this field due to what I've seen of these diseases, my mother having worked in an end-of-life care hospice. I was struck then by how little could be done to help people with these diseases. To be able to see treatments reach patients a decade on from that point is hugely gratifying and gives me great hope for the future. There are many more challenges to come, but they can be overcome, and I am proud to have been able to make my own small contribution towards achieving that.

In **Chapter 2** I describe a pipeline to computationally enhance methods of screening for aggregation inhibitors and use state of the art experimental approaches to uncover their inhibitory mechanisms. I use the protein α-synuclein as my test case. The computational architecture is then updated and improved in **Chapter 3**. In **Chapter 4** I focus again on the experimental part of the pipeline, with a novel DNA nanostructure based nanopore detection method for protein oligomers. Finally, I show an example of generalising this method into other protein misfolding areas in **Chapter 5** and consider future directions in **Chapter 6**. The rest of this chapter will give a brief description of protein misfolding mechanisms and methods of combating them.

1.2. Protein misfolding mechanisms in neurodegenerative disease

This PhD targets the intrinsically disordered proteins (IDPs) that are implicated as a pathological agent in neurological disease[10]. IDPs lack stable tertiary structure and as a result are prone to misfolding and aggregation when the cellular homeostasis system fails[15,31,32]. An IDP possesses solvent exposed hydrophobic areas that allow it to aggregate when its behaviour is not suitably controlled[33]. The misfolded oligomers and higher order fibrils generated during the aggregation process adopt a β-sheet conformation with one or more β-strands exposed to the solvent[34,35]. These exposed β-strands act as elongation sites for subsequent extension of existing aggregates, while catalytic sites on the surface of the aggregates act as nucleation sites for the creation of new oligomers[36]. During this process some aggregate species interfere with cellular functions, resulting in cell death[21].

These aggregate species are also likely to form because many IDPs may be present in the cell at supersaturated concentrations, making them prone to aggregation, and driving the interconversion between functional states and aberrant self-assembled multimerised states[32,37].

Under normal conditions, the protein homeostasis systems, including molecular chaperones and the ubiquitin-proteasome and endosomal-lysosomal degradation pathways[38,39], ensure the correct folding and complexing of proteins and removal of aggregates[15,40]. As the body ages and experiences stresses however, this metastable proteome becomes increasingly unstable, concomitant with these maintenance pathways becoming less efficacious[17,31]. This leads to uncontrolled protein aggregation and the accumulation of misfolded oligomers, eventually converting to highly ordered polymeric fibrils[41-43]. This behaviour has been shown to be common across many IDPs, and is implicated in numerous disease areas from AD, PD, amyloid lateral sclerosis (ALS), and Pick's disease to type 2 diabetes[43].

In the case of AD the metastable proteins are the truncated peptides of amyloid precursor protein (APP), which are known as Amyloid ß (Aß). These peptides aggregate to form amyloid plaques in the extracellular space of patient neurons[44]. Tau, a microtubule binding protein, is also found in aggregated form in neurofibrillary tangles within patient neurons[45]. Whether targeting amyloid is the best route to treatment of Alzheimer's disease remains to be seen, but the more positive recent clinical data provides support to the theory that aberrant protein misfolding and subsequent amyloid formation is indeed one of the pathological steps in disease progression. Given the observation of misfolded protein aggregates in many other neurodegenerative diseases, often exacerbated by mutations in these proteins, there is support for the idea of this being a relatively conserved mechanism across many neurodegenerative diseases. This comes with the caveat that the toxic misfolded species and aggregation mechanism may differ[10]. Recent research has suggested that the most damaging aggregate species are often those that are smaller, less ordered and more mobile, termed misfolded oligomers[21,46-48]. They have greater ability to interact with cellular organelles and membranes than the larger, higher order fibrillar species, damaging them in the process. However, fibrils are able to template aggregation from their surfaces through a process termed secondary nucleation. Secondary nucleation is considered to be the main generator of these oligomers, and so if secondary nucleation cannot be prevented completely the eventual removal of fibrils is also important[49]. This removal can be achieved by the cell's degradation pathways, if given enough breathing room by an aggregation inhibitor.

The aggregation of α–synuclein (αS) in particular is thought to be responsible for the neurodegeneration observed in PD, in which the pathological accumulation of misfolded

protein results in neuronal toxicity. There is some evidence of initial aggregation in the gut before eventually spread to the central nervous system and the substantia nigra region of the brain, at which point the patient's first motor impairments become evident[50]. The link between αS aggregation and PD is supported by genetic evidence and by observations of the accumulation of αS in hallmark inclusions known as Lewy bodies within the diseased neurons of PD patients[10,51,52]. A primary aim of current research into treatment of this disease is therefore the inhibition of αS aggregate formation.

The disordered and heterogeneous structure of αS may be an important part of its functioning under normal conditions as it is involved in synaptic vesicle transport, assisting docking of

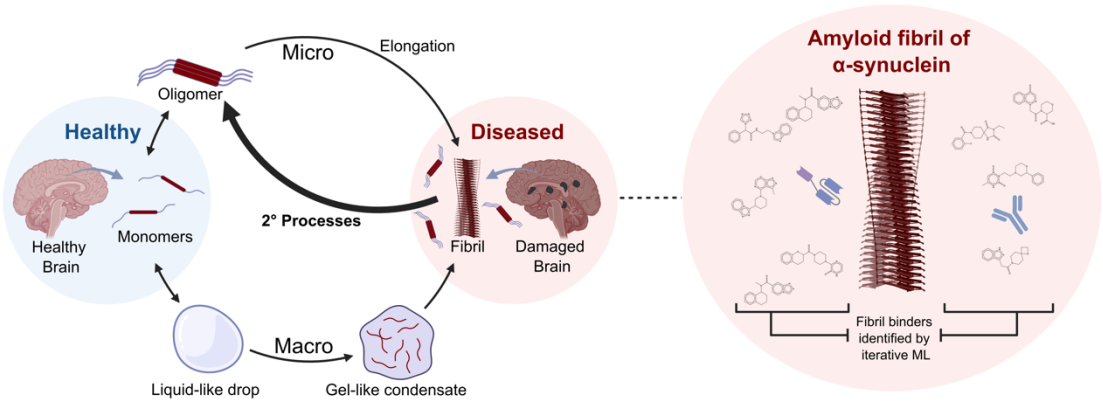


Figure 1.1. αS aggregation in Parkinson's disease and its inhibition by therapeutics. The age-related progressive impairment of the protein homeostasis system leads to the aberrant misfolding and aggregation of αS into toxic oligomeric species, which eventually convert to amyloid fibrils. The term 'amyloid' was coined to describe iodine-stained deposits in patient liver samples, that were initially thought to be carbohydrate based, before their high nitrogen content was discovered, revealing them to be proteinaceous[53]. The word is used to describe the fibrillar state of any aggregated protein. These αS fibrils are observed as the primary constituent of Lewy bodies, a hallmark structure observed in neurons of patients suffering from the disease. Fibrils can act as a catalyst for further oligomer formation via secondary processes such as secondary nucleation from catalytic sites on the fibril surface and fragmentation of the fibrils into smaller species. Secondary processes are the key generator of oligomeric species. These microscopic aggregation processes may occur in the context of macroscopic processes such as liquid liquid phase separation. This could hyper concentrate αS within the cytoplasm and organelles of the cell. This concentrated environment may drive aggregation processes for what is otherwise a relatively aggregation resistant protein[54].

these vesicles with the synaptic membrane. When bound to lipid bilayers, αS becomes enriched in amphipathic α-helical structure and promotes vesicle clustering, potentially via a double anchor mechanism[55]. Once aberrant aggregation begins however, one potential mechanism of

cellular toxicity is the perturbation of the neuronal organelle membranes such as those of the mitochondria by insertion of oligomeric forms of the protein into the phospholipid bilayer[21,46,47]. Efforts to prevent or slow this behaviour using therapeutics focus on inhibiting various steps in the aggregation process to reduce the proliferation of these toxic aggregates as much as possible[56]. The pathological relevance of these processes has led to major investment into identifying antibodies and small molecules that can inhibit those aggregation mechanisms associated with neurotoxicity[57-60]

1.3. Approaches to drug discovery for neurodegeneration

It is therefore particularly important to target αS aggregation by specifically preventing secondary aggregation processes[61]. In order to reduce the number of oligomers produced in an aggregation reaction, one approach is to target the highly ordered fibrillar aggregates, specifically the catalytic surfaces that allow oligomer formation[62]. Computational methods can contribute to these endeavours. In the following chapter, I will address the first question when beginning to look for molecular inhibitors of aggregation, which is how to identify the most promising area of the chemical space to search in, and then I will move on to enhancing the potency of the identified structures with machine learning.

I focus on efforts to slow down secondary nucleation by targeting fibril aggregation templating sites with small molecules, with the eventual hope that such an approach could be used in conjunction with current antibody-based treatments that are able to clear the fibrils formed from various misfolded proteins in their respective diseases. Targeting secondary nucleation in this manner has been achieved previously[63], but here we augment it with the use of computational methods such as structure-based docking to target disease specific fibril polymorphs[64] and machine learning in order to enhance the quality of molecular matter obtained in a shorter timeframe[65]. We begin with α-synuclein in Parkinson's disease as our first test case. As therapies are beginning to be delivered for AD[66], the race is on to achieve the same outcome for PD patients[67-69].

2. Targeting Parkinson's disease with iterative learning

"The whole of life is just like watching a film. Only it's as though you always get in ten minutes after the big picture has started, and no-one will tell you the plot, so you have to work it out all yourself from the clues." – Sir Terry Pratchett

2.1. Computational approaches to drug discovery for Parkinson's disease

The above view of life could be reasonably well applied to our current understanding of Parkinson's disease. We have clues about the end point of the disease, genetic evidence for the causes, and a good idea of how the suspected players behave in various model systems, but how well this relates to what happens in a human patient is still opaque. What we do know for certain is as follows. PD is the most common neurodegenerative movement disorder, affecting 2–3% of the population over 65 years of age[52,70-73]. Mutations in the α-synuclein (αS) protein often lead to drastically accelerated onset of disease[74]. αS is a 14.46 kDa protein made of 140 amino acids. It assumes a mostly random coil structure when monomeric, excepting it's N terminus (residues 1-60) which has alpha-helical propensity, especially when interacting with lipid membranes[50]. Many of the PD promoting mutations (A30P, E46K, A53T) are located in this region, suggesting that interference in this interaction could be important to disease. This interaction is also thought to be important in the protein's function which is mediating vesicle transport in the neurons[75]. The mid-section of the protein (residues 61-95) consists of a non-amyloid component (NAC) region which is largely hydrophobic and aggregation prone. The C terminus (residues 96-140) is negatively charged giving the protein a low pI. The aggregation

of αS has been associated with the initial neurodegenerative processes underlying the disease, in which the pathological accumulation of misfolded proteins results in neuronal toxicity[51,70,71,73].

Computational methods could be expected to reduce the time and cost of traditional drug discovery pipelines targeted at these processes[76-78]. Machine learning is rapidly emerging as a powerful drug discovery strategy[79]. To explore the potential of this strategy in drug discovery programs for Parkinson's disease and other synucleinopathies, I describe here a machine learning approach to explore the chemical space to identify compounds that inhibit the aggregation of αS. My starting point is an approach that combines docking simulations with in vitro screening, which was recently employed to identify a set of compounds that bind to the fibril structures of αS, and prevent the autocatalytic proliferation of αS fibrils as a result[64]. Here, I used this initial set of compounds as input for a structure-based machine learning approach to identify chemical matter that is both efficacious and represents a significant departure from the parent structures, providing compounds that conventional similarity searches would have failed to efficiently identify.

This approach is based on the lessons learned, using chemical kinetics, about the importance of secondary nucleation in αS aggregation[62,74,80]. Because of the autocatalytic nature of this process, structure-based methods could be expected to effectively target the catalytic sites on the surface of αS aggregates[64]. As I show here, the implementation of this idea within an iterative machine learning procedure leads to the identification and optimisation of compounds with high hit/lead rates and great potency.

2.2. Results of structure based iterative active learning

2.2.1. Components of the machine learning method.

The machine learning approach used here consists of 3 main components[81]: (1) the experimental data, i.e. a readout of the potency of the compounds in an aggregation assay, (2) the variational autoencoder required to represent the compounds as latent vectors, and (3) a model for training and prediction using these vectors and the assay readouts.

For component 1, we used a chemical kinetics assay[43,57,82] that provided both the initial data for the model training and the data that were iteratively fed back into the model at each cycle of testing and prediction. This assay identifies the top compounds that inhibit the surface-catalysed secondary nucleation step in the aggregation of αS.

For component 2, we used a junction tree variational autoencoder[83], pre-trained on a set of 250,000 molecules[84] enabling accurate representation of a diverse population of molecular structures. Using this approach, simplified molecular-input line-entry system[85] (SMILES) strings representing each molecular structure were standardised using MolVS[86] and converted into latent vector representations.

For component 3, we used a random forest regressor (RFR)[87] with a Gaussian process regressor (GPR)[88] fitted to the residuals of the RFR, with both regressors using the latent vectors as training features. The residuals are the errors of the regressor, thus the second regressor is trying to correct for the errors of the first. I have gone into more depth on the technical implementation of the algorithms in the appendix and referenced each algorithm with an article explaining its function in detail. In this chapter we are approaching the relation of latent vector representation of the molecular to structures to an assay readout (quantitative structure-activity relationship, QSAR) as a regression problem. The RFR provided the highest performance compared to other combinations of multi-layer perceptrons[89] (MLPs), GPRs and linear regressors (LRs) in terms of R^2 score, mean absolute error (MAE) and root mean square error (RMSE). Performance and parameters are shown in the appendix (**Figure A.1** and **Table A.1**). Combining the RFR and GPR provided only a marginal improvement in the metrics of the RFR alone, but crucially enabled leveraging of the associated uncertainty measure of the GPR when ranking molecules during acquirement prioritisation[81]. A Gaussian process makes a prediction based on the average of many Gaussian distributions fitted to a data set, while the standard deviation of these distributions supplies the uncertainty, a value for how confident the regressor is in its prediction. Tuning the weighting applied to this uncertainty measure allowed a ranking based on both the predicted potency of the molecules and the uncertainty of that prediction. Component 3 was then trained on the 161 initial experimental data points (see **Section 2.2.2**). The best molecules predicted by the model were then tested in the same assay and the results fed back into the model in an iterative fashion (~55-65 new molecules tested at each iteration). The molecules used at each stage of the project are illustrated in **Figure A.2**, together with the structures of the most potent molecules.

2.2.2. Initial set of small molecules.

The initial set of molecules was identified via docking simulations to αS fibrils carried out previously[64] (see **Appendix A.i** for a detailed explanation of the implementation), which were expanded upon via similarity searches around molecules that performed well in the chemical kinetics assay to identify further candidates. The docking screening was carried out using the consensus strong binders predicted by AutoDock Vina[90] and Openeye's FRED[91-93] software.

2 million molecules with good central nervous system multiparameter optimisation (CNS MPO)[94,95] properties were previously docked using AutoDock Vina to target the selected binding pocket[64] (**Figure 2.1**). CNS MPO is an aggregated metric of molecular properties that predicts likelihood of a molecule passing the blood brain barrier. In that study, the binding site encompassing residues His50−Lys58 and Thr72−Val77 was selected due to its propensity to form a pocket according to the Fpocket software[92] (**Figure 2.1A**), and its mid to low solubility according to CamSol[96] (**Figure 2.1B**). Additionally, His50 is predicted to be protonated below the pH value (5.8) at which αS secondary nucleation more readily occurs[97], which may be significant for initial interactions. To increase the confidence of the calculations, the top-scoring 100,000 small molecules were selected and docked against the same αS binding site,

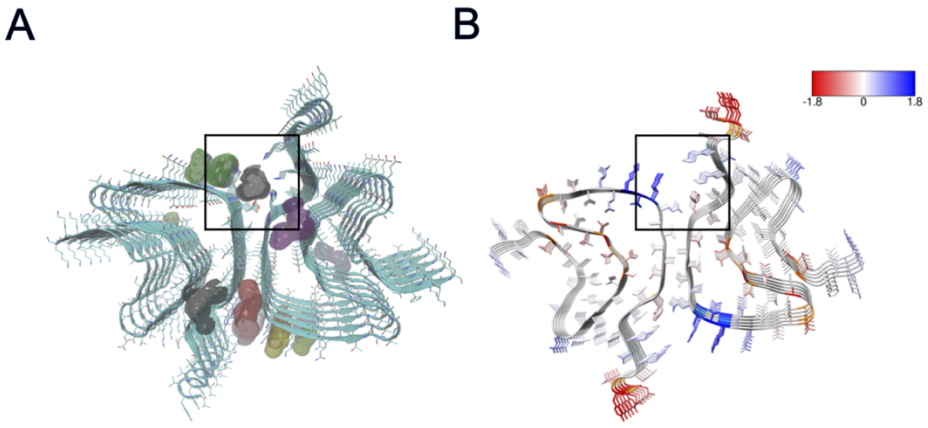


Figure 2.1. Volume and solubility based binding site prediction on polymorph 6CU7[98]. **(A)** Cavity based binding site prediction based on Fpocket[92]. **(B)** Solubility based binding site prediction based on CamSol[96]. The black box outlines the region encompassing key residues His50 and Glu57 where both cavity propensity is high and solubility is medium-low. Docking simulations and initial pocket searching were carried out by Z. Faidon Brotzakis.

using FRED[91]. The top-scoring, common 10,000 compounds in both docking protocols were selected and clustered using Tanimoto clustering[99] with a similarity cut off of 0.75, leading to a list of 79 centroids (representative molecules from each cluster). The Tanimoto similarity is a metric that compares Morgan fingerprint[100] representations (radius = 2, nbits = 2048) of 2 different molecules. These fingerprints are an older method to represent chemical structures numerically. A value of 1 for the Tanimoto similarity implies complete 2D homology between 2 structures while values closer to 0 imply little to no structural similarity. 68 compounds were available of the 79 leads identified in the in silico structure-based docking study. The first round of in vitro experiments were carried out with this set.

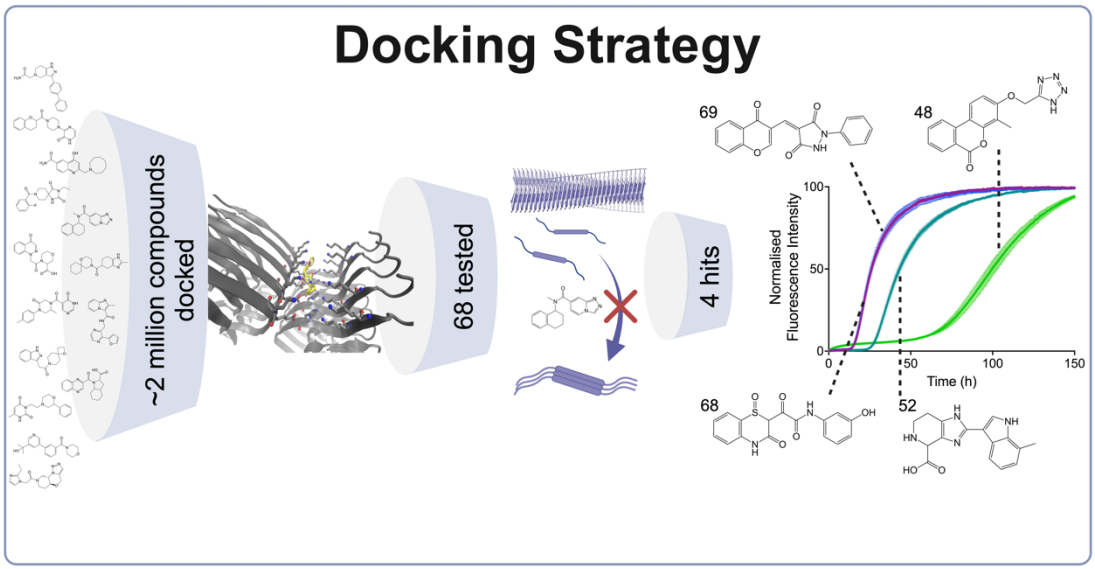


Figure 2.2. Illustration of the docking strategy for identification of leads at the start of the pipeline. 2 million compounds with good predicted blood brain barrier penetrance were computationally docked to a suspected oligomerisation site on the fibril surface. 68 of the corresponding centroids of each cluster were then obtained and experimentally tested. From 68 molecules predicted to have good binding via docking simulations, we initially identified 4 active molecules (the 'docking set') by experimental testing[64]. These 4 molecules increase the $t_{1/2}$ of αS aggregation.

Subsequent experiments to test these predicted binders in aggregation assays identified 4 active compounds[64] labelled molecule 48, 52, 68 and 69, referred to as the 'docking set', (**Figure 2.2**). We then began the process of lead generation and optimisation. Here, using the Tanimoto similarity metric between the molecules, 2 similarity searches were then carried out using these 4 structures as starting points (**Figure 2.3**). Different Tanimoto similarity thresholds were used to specify molecule subsets for testing, from initial analogue searches to the creation of a library to screen from. As such a similarity value >0.5 was used for closely related analogues, >0.4 for loosely related analogues and >0.3 for a library to screen from ('evaluation set'). While this

use of a structurally related screening library constrains the model's ability to generalise, the lack of diversity in terms of hits also makes it unlikely for the model to perform well in chemical space significantly divergent from this region. We are thus carrying out an exploitation strategy here. We remove the need for a curated screening library in subsequent work by utilising generative modelling and reinforcement learning[101], allowing for both exploitation and exploration strategies.

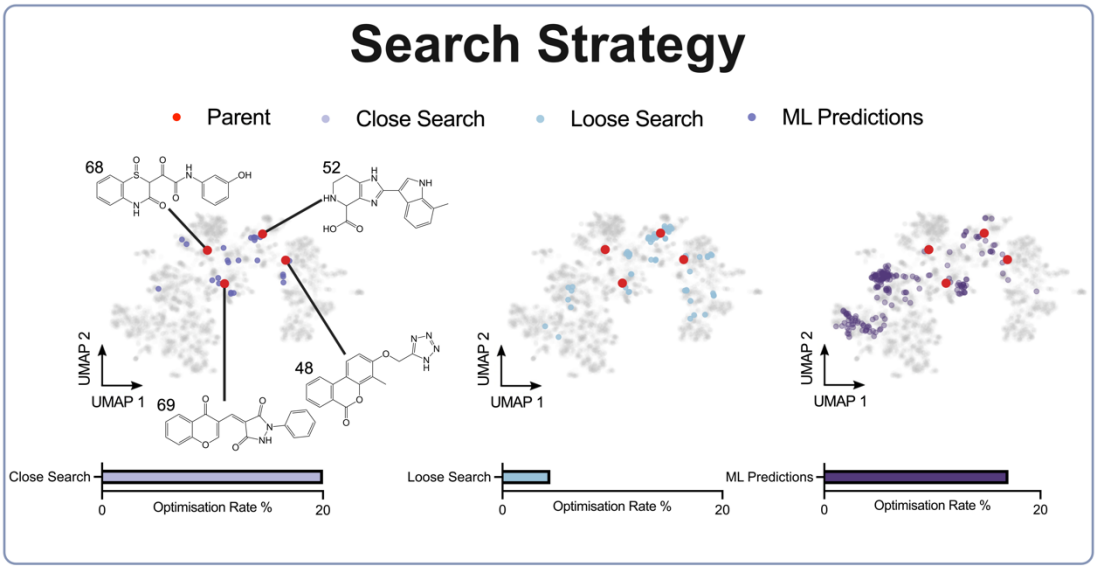


Figure 2.3. Illustration of successive search strategies around the hits identified by the docking. We initially performed a close Tanimoto similarity search around the 4 parent compounds in chemical space, represented here in two dimensions via UMAP[102]. UMAP is a dimensionality reduction technique, which is applied here to the latent vectors describing the molecular structures. We selected molecules with Tanimoto similarity cut off >0.5 (the 'close similarity docking set') followed by a loose similarity search with Tanimoto similarity cut off >0.4 (the 'loose similarity docking set'). The described machine learning method was then applied using the observed data to predict leads from a compound library derived from the ZINC database with Tanimoto similarity >0.3 to the parent structures (the 'evaluation set').

A selection of closely related molecules (Tanimoto similarity > 0.5) to the parent compounds (referred to as the 'close similarity docking set', **Figure 2.3** and **Figure A.2B**) was tested in the aggregation assay. Potent lead selection was made according to a cut off corresponding to a normalised half-time of the aggregation ($t_{1/2}$) of 2 times that of the negative control[63]. The 'optimization rate' was defined as the percentage of molecules in a set that passed this threshold. This yielded 5 new leads from 25 new molecules (**Figure A.2B**), 1 derived from molecule 48, 3 from molecule 52 and 1 from molecule 69. This step was then followed by a larger selection of compounds with a looser cut-off of structural similarity (Tanimoto similarity >0.4) to the parent compounds (referred to as the 'loose similarity docking set', **Figure 2.3**). Although new leads featured amongst this set, the optimisation rate was low (4%), and both

molecules 48 and 52, which had initially appeared the most promising of the parent structures, yielded poor results. From the 29 molecules related to molecule 48 in the loose similarity docking set, none were potent, while from the 24 molecules related to molecule 52, only 2 were potent. The functional range of molecules 48 and 52 appeared narrowly limited around the chemical space of the parent structures. Molecule 69 yielded 1 potent lead from the 16 related molecules. Overall, the optimisation rate from the loose similarity docking set was less than a quarter of that of the close similarity docking set and involved testing 3 times as many compounds. A fall in optimisation rate was expected as the similarity search has no intuition about QSAR, it simply identifies structures with similarity to the active scaffold, whether those similarities are relevant or not. It is however surprising that the drop off was so rapid, as conventional structure-activity relationship techniques would suggest there should be some activity within such close proximity to the active scaffold. This behaviour is termed an 'activity cliff', and it hampers exploration around an active scaffold[103].

These results suggest that it would be challenging to further explore the chemical space using conventional structure-activity relationship techniques without significant attrition, since the optimisation rate dramatically worsened as the similarity constraint to the parent hits was slightly loosened. To overcome this problem, the compounds resulting from these experiments were then used as input for a machine learning method for an iterative exploration of the

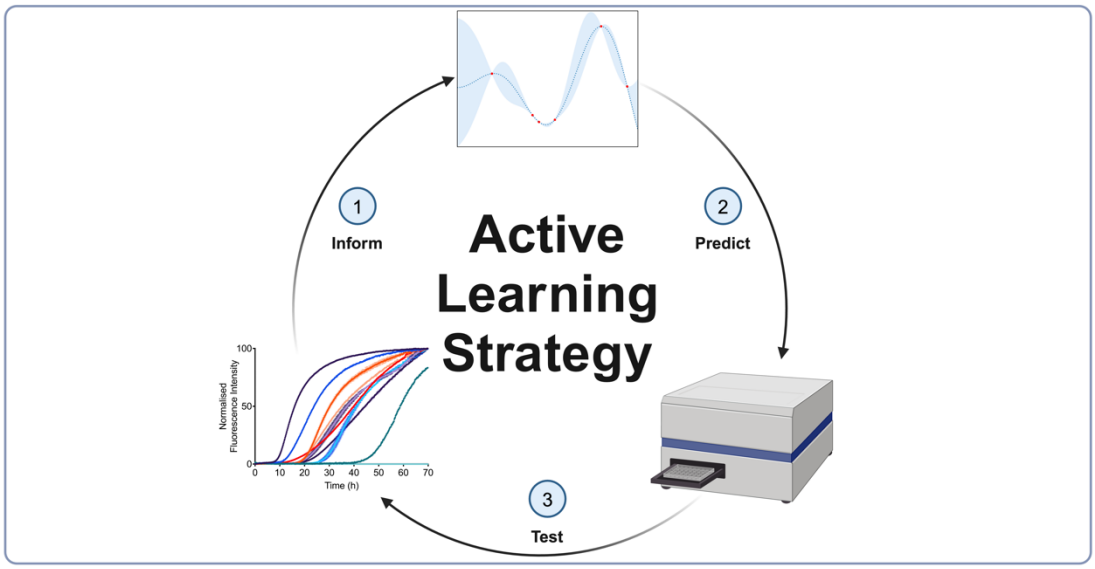


Figure 2.4. Illustration of an iterative active learning approach to inhibitor optimisation. Successive iterations of prediction and experimental testing yielded higher optimisation rates, and molecules with higher potency on average than those identified in the previous similarity searches. Validation experiments were also carried out on the leads identified.

chemical space (**Figure 2.4**). The similarity searches removed the most obvious targets of the machine learning approach, but also increased the size of the data set available for training. The training set, however, remained small by typical machine learning standards, consisting of 161 molecules. Since training sets of this size are common in early-stage research, a further aim of the work done in this chapter was to demonstrate that machine learning can be used effectively even in such data sparse scenarios.

2.2.3. Iterative application of the machine learning approach.

One of the issues with applying machine learning to a data sparse scenario is that predictions are likely to be overconfident. While this problem can be addressed to an extent by utilising Gaussian processes, a complementary strategy is to restrict the search area to a region of chemical space that is more likely to yield successful results. To this end, a structural similarity search of the 4 hit molecules in the docking set was carried out on the 'clean' and 'in stock' subset of the ZINC database, comprising ~6 million molecules. Any molecules showing a Tanimoto similarity value of >0.3 to any of the 4 structures of interest was included. This low threshold for Tanimoto similarity was intended to narrow the search space but without being overly restrictive of the available chemical landscape, yielding a data set of ~9000 compounds which comprised the prospective 'evaluation set'. The distribution of this evaluation set in terms of the predicting binding energies is shown in **Figure A.3A**.

Different machine learning models were initially trialled against the docking scores calculated for the evaluation set as a test of the project feasibility, and these models were then tuned on the much smaller aggregation data set. The best performing set up, the RFR-GPR stacked model, was then trained on the whole aggregation data set and used to predict the top set of molecules (see Machine Learning Implementation in the Appendix, and **Figures A.1, A.4** and **A.5**). For this work, the $t_{1/2}$ of a light seeding assay isolating secondary nucleation as the aggregation mechanism was used as the metric of potency to be used in machine learning, because of its robustness. For comparison, the amplification rate is more susceptible to small fluctuations in the slope of the aggregation fluorescence trace[64] (**Figure A.6**). Example traces from this assay are shown in **Figure 2.5A**, the $t_{1/2}$ being the time point at which the fluorescence reaches 50% of its maximal signal. The threshold for potency was again a $t_{1/2}$ 2-fold greater than that of the negative control under standard assay conditions (see methods **Section 2.5.5**).

The algorithm was run repeatedly from different random starting states and those molecules that appeared in the top 100 ranked molecules more than 50% of the time (64 molecules) were chosen for purchase (first iteration). In this first iteration, there was an inherent bias towards the structure of molecule 69 in the data set given the relative population sizes (**Figure A.2A**), but with the caveat that many of these structures were only loosely related to the parent (Tanimoto similarity < 0.4). Many of the potent lead molecules came from this group, suggesting chemical departures from the parent structure.

The dynamic range within the aggregation data set in terms of potency was large, in that a majority of the molecules had no effect on aggregation, while the initial docking hits and their close derivatives exhibited a relative $t_{1/2}$ of up to 4-5 times that of the negative control (limited by the length of the experimental run) at 25 μM. Molecules then found via machine learning produced a relative $t_{1/2}$ of ~4-5 at up to 8 fold lower concentration (3.12 μM, 0.3:1 molecule:protein) than that carried out in the initial screening (25 μM, 2.5:1 molecule:protein). This compares favourably with previous molecular matter tested in a less aggressive seeded aggregation assay such as the flavone derivatives, apigenin, baicalein, scutellarein, and morin which achieved relative $t_{1/2}$ of 1-2 at a stoichiometry of 0.5:1 molecule:protein in a separate study[57]. Anle-138b is another example of a well-characterised small molecule inhibitor, which will be used as a benchmark for αS aggregation inhibition throughout this thesis, and which was also taken into clinical trials[60]. The relative $t_{1/2}$ of Anle-138b is 1.22 (**Figure 2.5A**) at a ratio of 2.5:1 molecule:protein in the assay used in this work, which is significantly lower than any of the molecules discovered using the strategy employed here.

After the first iteration, the compound data were pooled together to extend the training set and a further 2 iterations were carried out, adding the resultant data to the training set at each iteration. This was followed by a fourth and final iteration trained on low dose (3.12 μM) data of all the previously obtained molecules. Example kinetic traces for a molecule from the fourth iteration are shown in **Figure 2.5A**. The molecules are labelled according to iteration number and identifier within that iteration. For example, I4.05 is the fifth potent lead (05) within iteration 4 (I4). The dose-dependent potency in the aggregation assay was investigated (**Figures 2.5A** and **A.7**) with all leads exhibiting substoichiometric potency. For comparison Anle-138b is also shown. **Figure 2.5B** shows an approximate overall rate of aggregation at different concentrations of I4.05, Anle-138b and the parent molecule. This approximate rate was taken as $1/t_{1/2}$, and fitted to a Hill slope[104]. A kinetic inhibitory constant (KIC_{50}) was then

derived. This is the concentration of molecule at which the $t_{1/2}$ is increased by 50% with respect to the control, as defined previously[63]. The KIC_{50} values for the leads were in the range of 0.5-5 μM, which compare favourably with the parent of the lead molecules (molecule 69) and Anle-138b which have KIC_{50} values of 18.2 μM and 36.4 μM (extrapolated) respectively. I4.05 had a KIC_{50} value of 0.52 μM with 95% confidence limits of 0.45 μM and 0.59 μM.

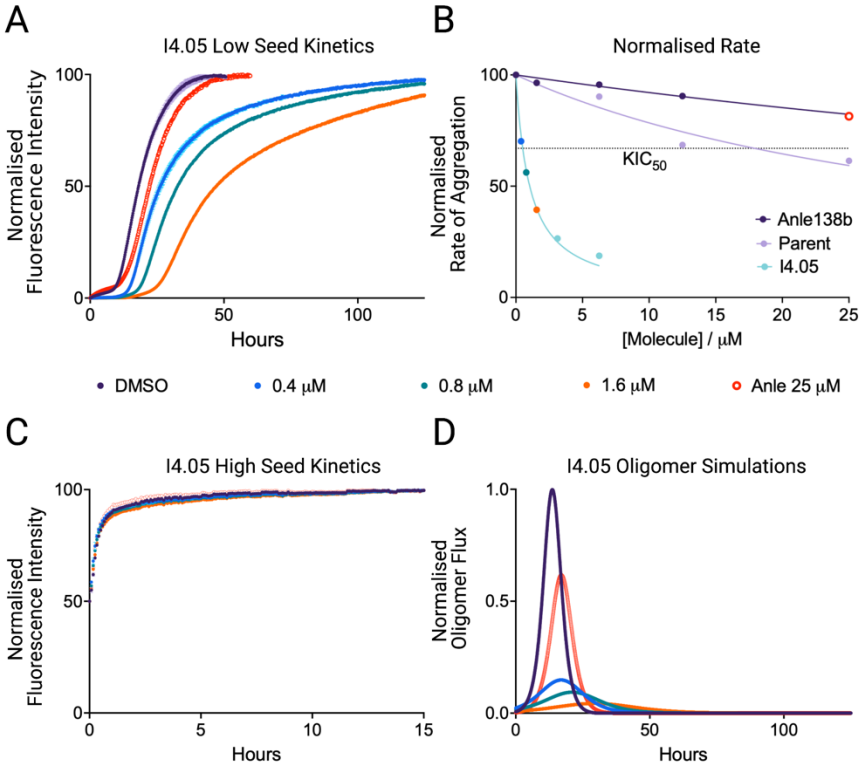


Figure 2.5. Performance comparison of a molecule from the final iteration of active learning (I4.05) vs an αS aggregation inhibitor currently in clinical trials (Anle-138b). (A) Kinetic traces of a 10 μM solution of αS with 25 nM seeds at pH 4.8, 37 °C in the presence of molecule or 1% DMSO in triplicate, with error bars denoting SD. During the initial screening, except for iteration 4, all molecules were screened at 2.5 molar equivalents (25 μM), and leads were then taken for further validation at lower concentrations: 0.4 μM (blue), 0.8 μM (teal), 1.6 μM (orange) with Anle-138b at 25 μM for comparison (red circles). The 1% DMSO negative control is shown in purple. Molecule I4.05 is shown as an example. The endpoints are normalised to the αS monomer concentration at the end of the experiment, which was detected via the Pierce™ BCA Protein Assay at t = 125 h. Furthermore, the same experiments were carried out using AlexaFluor™ 488 labelled αS yielding similar levels of inhibition as the ThT curves. **(B)** Approximate rate of reaction (taken as $1/t_{1/2}$, normalised between 0 and 100) in the presence of 3 different molecules, Anle-138b (purple), parent structure 69 (lilac) and I4.05 (blue). The KIC_{50} of I4.05 is indicated by the intersection of the fit and the horizontal dotted line. **(C)** High seeded experiments (5 μM seeds, all other conditions match **A**) were also carried out to observe any effects on the elongation rate and enable oligomer flux calculations using the secondary nucleation rate derived from **A. (D)** Oligomer flux calculations for I4.05 vs the competitor Anle-138b using the rates derived from both **A** and **C**.

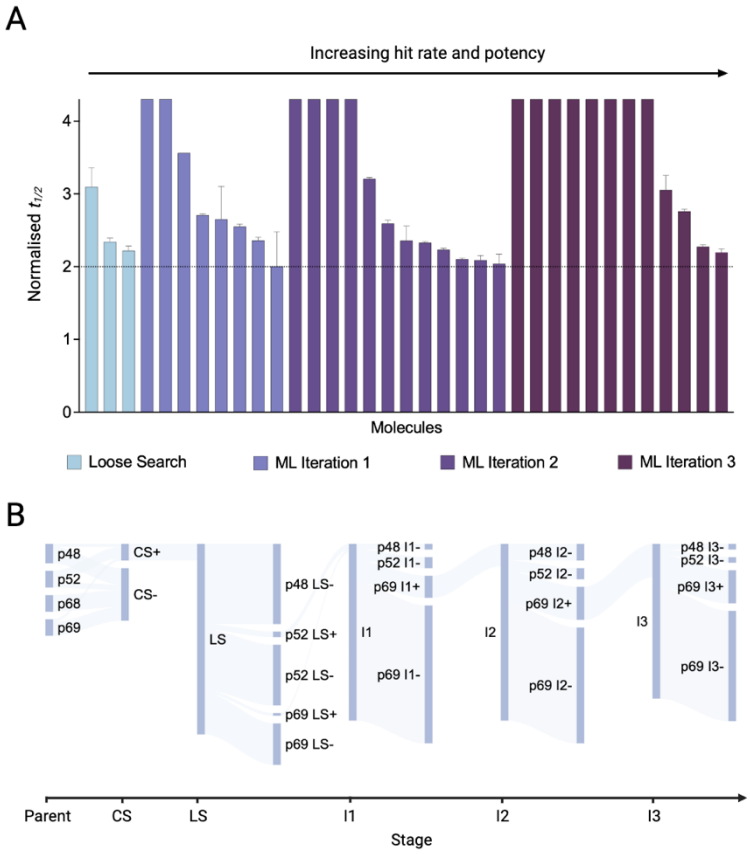


Figure 2.6. Results of the iterations of the machine learning drug discovery approach. (A) Normalised $t_{1/2}$ for the leads at 25 μM from the different stages: loose search, iteration 1, iteration 2 and iteration 3 (error bars denote SEM). The horizontal dotted line indicates the boundary for classification as a potent lead, which was normalised $t_{1/2} = 2$. For the loose search, 69 molecules were tested, while for iterations 1, 2 and 3, the number of molecules tested was 64, 64 and 56 respectively. Note that the most potent molecules exhibited complete inhibition of aggregation over the timescale observed, so the normalised $t_{1/2}$ is presented as the whole duration of the experiment. **(B)** Flow of potent leads (+) and negatives (-) in the project starting from the close search (CS), moving to the loose search (LS) and then iterations 1, 2, and 3 (I1, I2, I3). Each branch is labelled with the molecule source (e.g. p48). Attrition reached its highest point at the loose search before gradually improving with each subsequent iteration.

The elongation rate was largely unaffected in the presence of molecules at any concentration (**Figure 2.5C**). This was expected given the designed mechanism of action of the small molecule. It was also reassuring, since compounds that inhibit elongation may increase the population of oligomers[63], which are considered the most damaging of the aggregate species in vivo[46,47]. Then, using the amplification and elongation rates derived from **Figure 2.5A**, **C**, the oligomer population over time was calculated[57] (see methods **Section 2.5.7-2.5.8**). These calculations are shown in **Figure 2.5D** for I4.05 and **Figure A.7** for the rest of the leads. All leads demonstrated a dose-dependent delay and reduction of the oligomer peak. Across all metrics, I4.05 performed significantly better than Anle-138b and the parent molecule at

substoichiometric ratios, as do all of the leads obtained in previous iterations (**Figures A.7** and **A.8**).

The aggregation data from the first 3 iterations are also shown in **Figure 2.6A**. The flow of molecules derived from each parent in terms of positives and negatives over the course of the project is illustrated in **Figure 2.6B**. Of the 64 molecules from iteration 1, 8 were potent, representing an optimisation rate of 12.5%, the second iteration showed a further increase, with 11 potent molecules, representing a 17.2% optimisation rate and the third iteration, with 12 potent molecules, had an optimisation rate of 21.4%. These optimisation rates represent an order of magnitude improvement over previously reported high throughput screening (HTS) assays (<1%)[105] and, remarkably, an overall 40% improvement over the combined similarity search optimisation rates, which removed the most likely potent compounds. The potency of the machine learning leads was significantly higher on average than those identified by the similarity searches (**Figure 2.7A**), without compromising the CNS MPO scores (**Figure 2.7B**). The accumulated training data from all stages of the project for all molecules in terms of half time distribution is shown in **Figure A.3B** and **A.3C.**

Given that αS aggregation and toxicity has also been linked to membrane interactions[21,46] a parallel investigation was carried out with a lipid induced aggregation assay (**Figure A.9)** which was used as a validation of the molecules rather than for machine learning optimisation. The tested lead molecules also showed strong efficacy in this assay. A further test of these molecules in a spontaneous αS aggregation assay, without induction via pre-seeding or shaking, also exhibited strong potency[106].

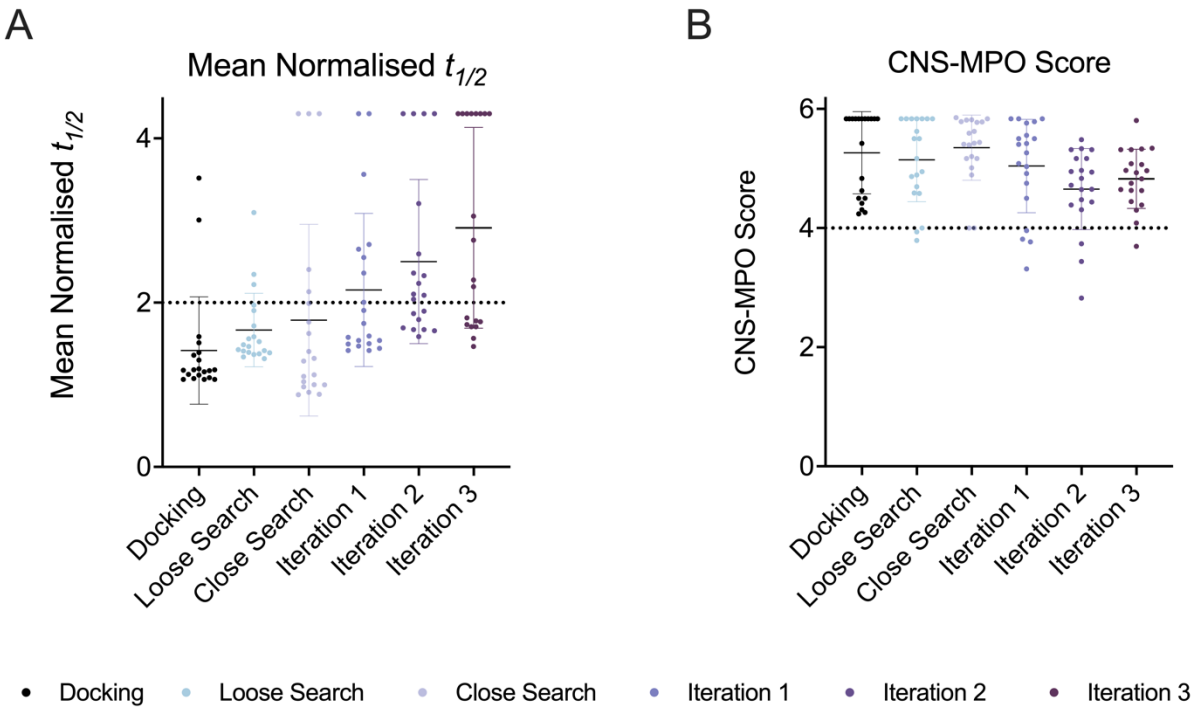


Figure 2.7. Average $t_{1/2}$ of aggregation and CNS-MPO scores for the top 20 molecules at each stage. **(A)** The stages are the initial docking simulation (68 molecules tested), loose search (69 molecules tested), close search (25 molecules tested), iteration 1 (64 molecules tested), iteration 2 (64 molecules tested) and iteration 3 (56 molecules tested). Molecules were tested at a concentration of 25 µM during screening. Molecules that completely prevented aggregation were assigned a $t_{1/2}$ value equal to the length of the experiment. **(B)** A common cut off for CNS-MPO score is 4, as indicated by the horizontal dotted line.

2.2.4. Analysis of the chemical space explored by machine learning.

The chemical space explored by the machine learning approach was inspected via dimensionality reduction techniques, including PCA, t-SNE[107] and UMAP[102] to investigate how the model was prioritising molecules (**Figure A.10 and Figure 2.8**). These methods all attempt to reduce dimensionality of the data either for clustering or visualisation purposes (reducing to two dimensions), while retaining as much information about the data distribution as possible. In this case they are used to visualise the chemical space as represented by the junction tree variational autoencoder latent vectors of the molecules. The relative positioning of the training points and the parents within the chemical space is shown in **Figure 2.8A**. The stacked RFR-GPR model assigned low uncertainty to areas of the chemical space proximal to the observed data, and the corresponding acquirement priority mirrored this when trained on the aggregation data (**Figure 2.8B-D**). This figure also illustrates how the uncertainty weighting could be altered during the ranking, depending on how conservative a prediction

was required. A drawback to a high uncertainty penalty was that the model remained in the chemical space it was confident in, while a lower uncertainty penalty ensured reasonable confidence of potent molecule acquirement while still exploring the chemical space.

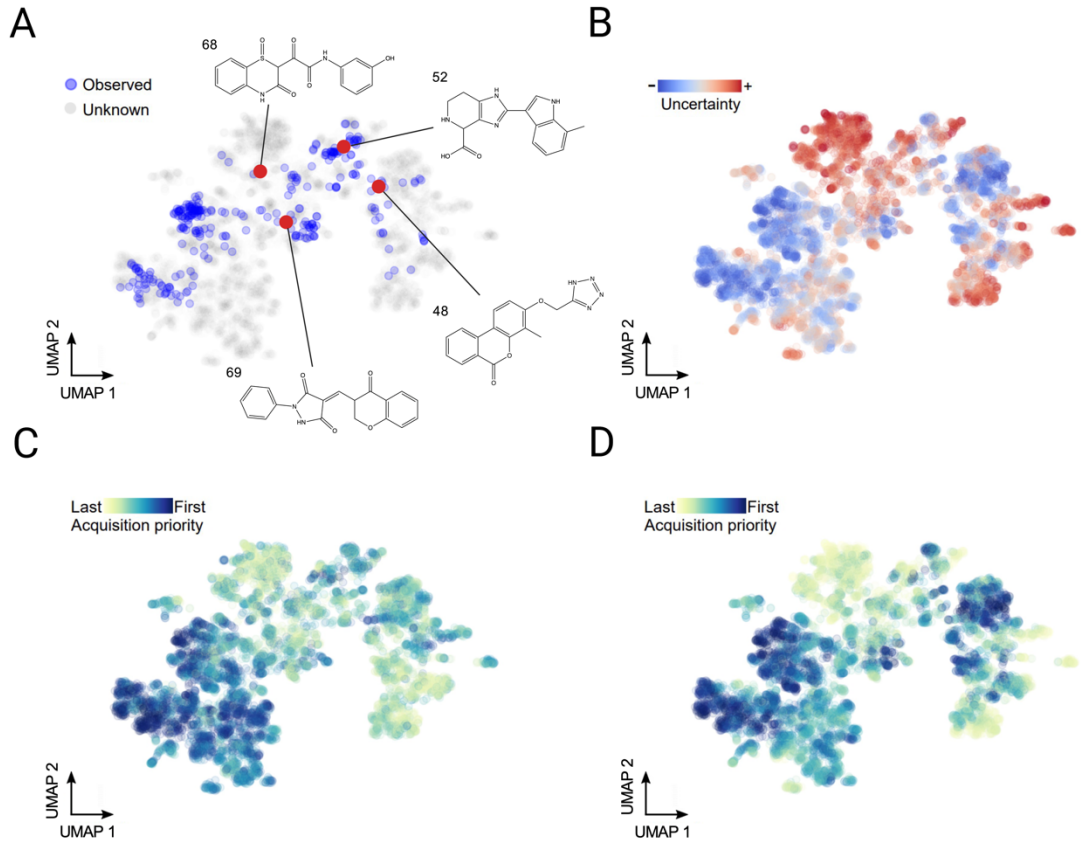


Figure 2.8. UMAP visualisation of the compound feature space using uncertainty. (A) The visualisation indicates the molecules in the chemical space that have been tested over the course of the project (blue circles) starting from the 4 initial docking molecules (red circles) in the docking set, and the relative positioning of the parent structures in this space. **(B)** GPR assigned lower uncertainty (blue) to regions of the chemical space near to the observed data and high uncertainty (red) to areas which were further away. **(C)** Acquirement ranking with a low uncertainty penalty. The lower uncertainty compounds were prioritised (dark blue) during acquirement ranking. **(D)** Acquirement ranking with a high uncertainty penalty.

The changes in similarity of the leads to the parent structures are shown in **Figure A.11**. The similarity of the molecules to their parent structure dropped for all structures at successive stages of the investigation, reaching its lowest point at the iterations of the machine learning approach. The more potent leads mostly retained the central ring and benzene substituent of molecule 69 albeit with the addition of polar groups to the benzene ring, but featured significant alterations to the rest of the scaffold. For example, from iteration 1, I1.01 replaced the fused ring substructure of molecule 69 with a single substituted benzene ring, while I1.02 replaced it

with a substituted furan ring, and subsequent iterations saw more complexity introduced. These changes were reflected in the Tanimoto similarity values, which were at the lower end of what was permitted in the evaluation set, 0.3 being the cut off. It was evident from this result that parts of the substructure were important to retain for potency, which the model did effectively while also identifying alterations in the rest of the scaffold that enhanced the potency considerably beyond that of the parent.

The observation that the QSAR model converges on the structures from two areas of the UMAP space related to structure 69 was encouraging in that it suggested the models were learning useful information and not selecting at random. While we have not tested a random set of molecules due to prohibitive resource cost, we do note that if a random selection of molecules were taken from the accumulated training data from all stages of the project, its optimisation rate (11%) would be lower than that of iterations 1, 2 and 3 on average. Though performance improves with additional data the QSAR performance in terms of R^2 remains modest (**Figure A.1**), but this is in part due to sparsity of training data. We would anticipate improvement if this approach could be implemented at medium scale with correspondingly more complex QSAR models, and we have an indication of this from trials of the model set up against the docking scores of the evaluation set, where performance in terms of R^2 score is 3 fold higher for a slightly larger data set.

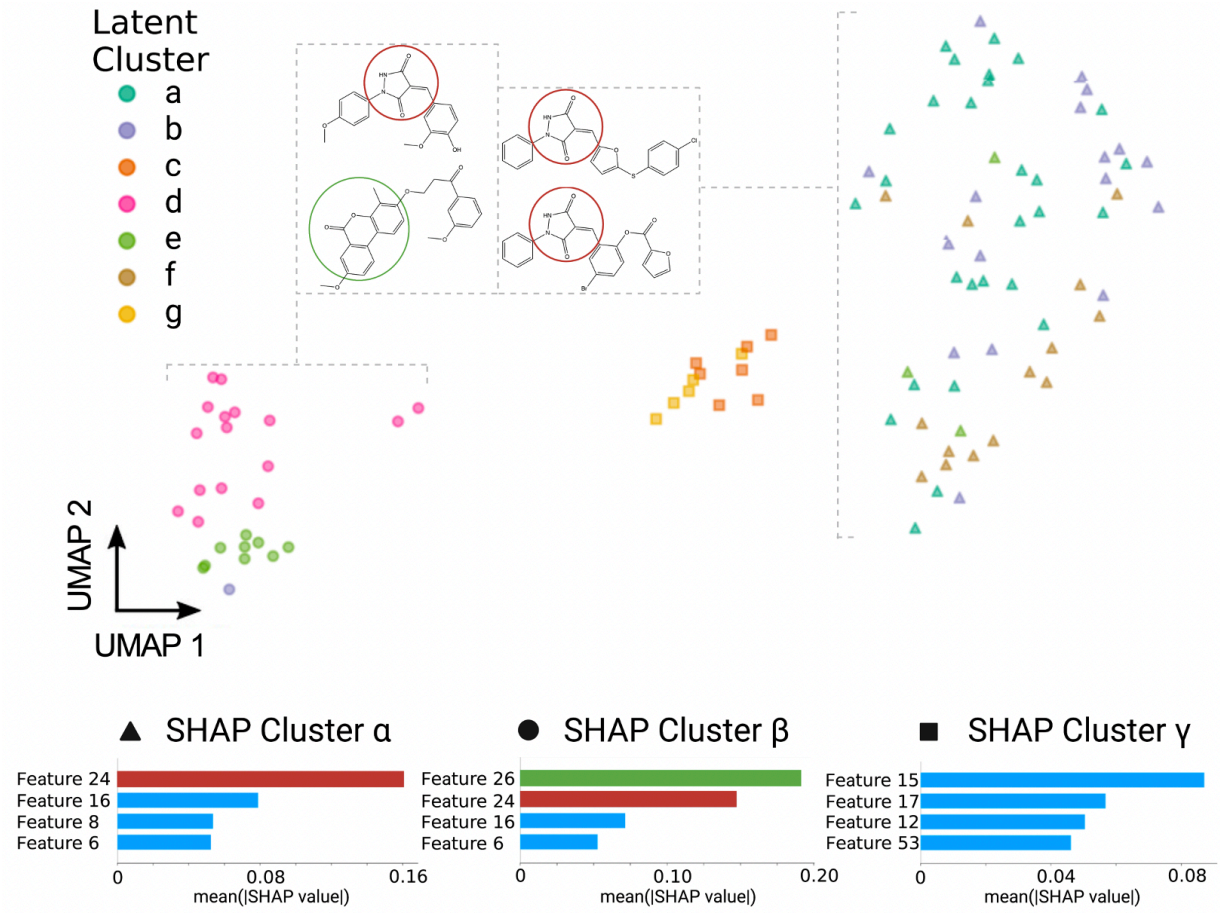


Figure 2.9. Clustering molecules based on SHAP dimensions and latent vectors. Three SHAP clusters were selected which show clear separation via UMAP. The colouring on the UMAP plot is based on the latent space clusters (a-g) and the shape of the marker is based on the SHAP value clustering (α, β, γ). Examining the plot shows that there is no separation between latent clusters c and g, which are grouped together in SHAP cluster γ. Although molecules which belonged to latent clusters a, b, and f were mostly grouped together by SHAP clustering, latent cluster e was grouped together with latent cluster d. Examination of the top dimensions of each SHAP cluster revealed that dimension 24 at least partly encodes for the key sub-structure of clusters a, b, e and f (3,5-pyrazolidinedione, highlighted in dark red), while dimension 26 at least partly encodes for the key sub-structure of cluster d (the oxygen-rich chromenone fused ring system, highlighted in dark green), and dimensions 15, 17, 12 at least partly encode for the key sub-structure of clusters c and g (carboxylic acid bearing aromatic group). Computational work was carried out by Alice Aubert under my supervision.

Next, an investigation was carried out to identify what structural information the latent vectors were encoding. Variational autoencoders are generally not built to ensure that their latent space dimensions are human interpretable, making this a challenge. The decoding of a variational autoencoder is also not deterministic, preventing facile analysis of the feature space based on single perturbation approaches of the input features and observing changes to decoded structures. Instead, hierarchical clustering was carried out on the latent vectors, followed by Shapley Additive explanation (SHAP)[108] clustering for comparison (**Figure 2.9**). While the

former differentiated groups based on large changes in any dimension, clustering based on SHAP dimensions ensured that clusters were created based only on features relevant to the prediction problem at hand. Latent space dimensions that have a large range of values had a large effect on the latent space clustering, regardless of whether these dimensions were important predictors of molecular potency. Using SHAP values, on the other hand, meant that latent space dimensions which had little effect on the model prediction were mapped to values close to zero, and therefore had a much smaller influence on the clustering. This resulted in clusters which were relevant to the prediction task. This strategy was suggested by the authors of SHAP and was recently used in the context of identifying subgroups of Covid-19 symptoms[109].

In **Figure 2.9** I show a UMAP representation of the tested molecules, with the latent vector clustering indicated by colour and the SHAP clustering indicated by shape. From the UMAP representation, it is notable that the SHAP clustering identified clusters more effectively than the hierarchical clustering. The SHAP values for each feature show the importance of that feature in the interpretation of potency, and this in turn could be used to identify which substructures within the molecules are relevant for potency by observing the structures that recurred in each cluster. For example, **Figure 2.9** shows the top dimensions of each SHAP cluster, revealing that dimension 24 at least partly encoded for the key sub-structure 3,5-pyrazolidinedione, which was present in every molecule in cluster α and a significant proportion of cluster ß, while dimension 26 at least partly encoded for the key sub-structure of cluster d, a chromenone fused ring system which was present in every other molecule in the cluster. This confirmed the hypothesis previously put forward by Jin et al.[83] that in a junction tree variational autoencoder, the latent space encoding preserved the key features of each molecule. Molecules which were clustered together shared many molecular substructures in common.

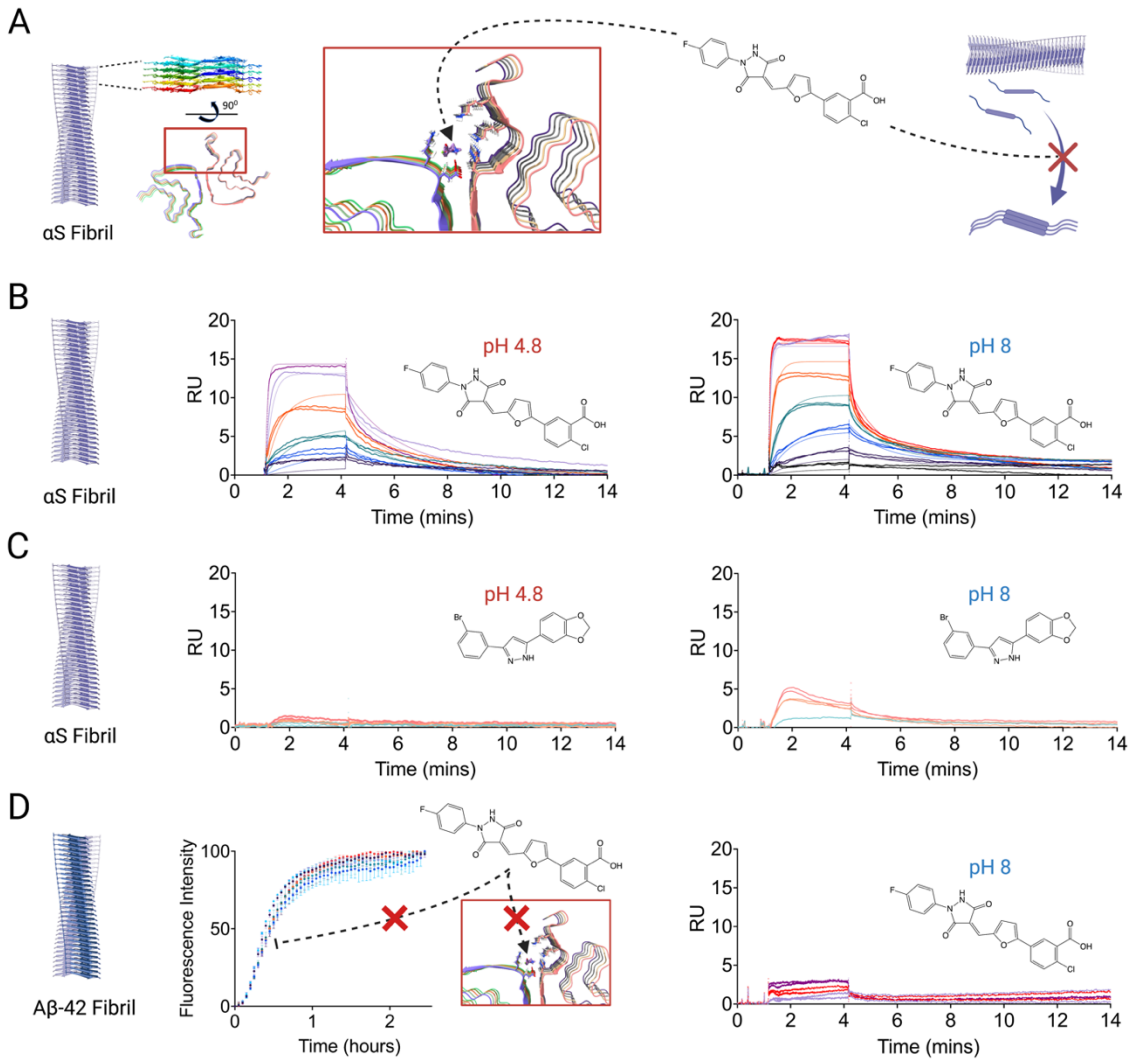


Figure 2.10. Molecule binding to αS fibrils. **(A)** A schematic representation of small molecule binding to the target binding pocket on the αS fibril, preventing secondary nucleation in the process**. (B)** SPR response curves for different concentrations of I4.05 at pH 4.8 and pH 8 binding to αS fibrils generated by a seeded assay, with the corresponding molecular structure shown. Raw data (points) and the corresponding fits (solid lines) for each molecule concentration are shown: 1.1 nM (black), 3.3 nM (purple), 11 nM (blue), 33 nM (teal), 111 nM (orange), 333 nM (red), 500 nM (magenta) and 1.1 μM (purple). Concentrations were repeated in duplicate in a pyramidal arrangement. The αS fibrils were immobilised at a concentration of 2000 pg / mm^2 on a CM5 Cytivia chip. The fits correspond to a 1:1 kinetic binding model, which yielded a K_D of 68 nM (k_a = 1.936 ± 0.007 10^5 $M^{-1}s^{-1}$, k_d = 1.315 ± 0.003 10^{-2} s^{-1}) at pH 4.8 and 13 nM at pH 8 (k_a = 5.879 ± 0.024 10^5 $M^{-1}s^{-1}$, k_d = 0.781 ± 0.002 10^{-2} s^{-1}). **(C)** SPR response curves for different concentrations of Anle-138b at pH 4.8 and pH 8 binding to αS fibrils generated by a seeded assay, with the corresponding molecular structure shown. Raw data (points) for each molecule concentration are shown: 1.1 μM (purple), 3.3 μM (light orange), 5 μM (light red). Accurate fits at pH 4.8 could not be obtained given the low dose response, but at pH 8 a 1:1 kinetic binding model yielded an approximate K_D of 8.1 μM (k_a = 0.0359 ± 0.0005 10^5 $M^{-1}s^{-1}$, k_d = 2.90 ± 0.02 10^{-2} s^{-1}). **(D)** Seeded kinetics (40 nM seed) and SPR response curves for 2 μM Aß42 in the presence of 1% DMSO or different concentrations of I4.05 (colour scheme as above). I4.05 is unable to effectively inhibit Aß42 secondary nucleation or bind to Aß42

fibrils (approximate K_D = 2.5 μM). The Aß42 fibrils were immobilised at a concentration of 2000 pg / mm^2 on a CM5 Cytivia chip.

2.3. Validation of lead molecules identified

2.3.1. Measurement of binding affinity.

A series of validation experiments were carried out on the most potent leads from the machine learning iterations. We first tested the binding to fibrils using surface plasmon resonance (SPR, see methods **Section 2.5.12**) under different buffer conditions. The results for molecule I4.05 vs Anle-138b are shown in **Figure 2.10**. The proposed mechanism of action is the binding of molecules to the fibrils thereby blocking nucleation sites for further aggregation. Support for this mechanism of action comes from the observations that the molecules function at significantly substoichiometric ratios, discounting monomer interactions, and also show negligible effect on elongation. The lack of effect on elongation suggests binding of the molecules to the fibril surface rather than to the fibril ends. Covalent interactions can also be discounted, as no mass change is observed of the αS monomer by mass spectrometry. The large effect observed in an assay that isolates secondary nucleation as the dominant mechanism implies that the molecules are specifically affecting this step, and the substoichiometry implies that the molecules must be interacting with the fibrils which are present in nM monomer equivalents at the start of the aggregation.

Proof of binding and evidence for this potential mechanism are shown by SPR in **Figure 2.10**. **Figure 2.10A** shows a schematic representation of molecule binding to the binding pocket targeted during the initial docking simulation. **Figure 2.10B** shows SPR response curves for a concentration range between 0.3 nM and 1.1 μM of I4.05, while **Figure 2.10C** shows the same experiment utilising Anle-138b from 1.1 μM to 5 μM. The binding was tested under the conditions of the light seeded assay, pH 4.8, and also at pH 8, allowing direct comparison to the seeding assay conditions of Aß42, which were tested as a control in **Figure 2.10D**. αS is highly charged at neutral pH and has a PI of 4.7[110]. It therefore requires a pH in this region to render the protein uncharged in order to aggregate on an experimentally accessible timescale under quiescent conditions, whereas Aß42 is highly aggregation prone and requires high pH to prevent it aggregating too rapidly for detection[63]. At both pH values, I4.05 exhibited binding

to αS fibrils, with kinetic fits giving K_D values of 68 nM at the lower pH and 13 nM at the higher pH. The data for Anle-138b showed no response for pH 4.8 and so no K_D could be obtained, while at pH 8 an approximate K_D of 8.1 µM was obtained. It is evident that the 2 orders of magnitude improvement in KIC_{50} of I4.05 compared to Anle-138b was matched by a similar degree of improvement in terms of binding efficacy. **Figure 2.10D** shows that I4.05 has no effect on the seeded aggregation of Aß42, nor does it bind effectively to Aß42 fibrils, which suggests that this molecule is not a promiscuous aggregation inhibitor between different amyloidogenic proteins.

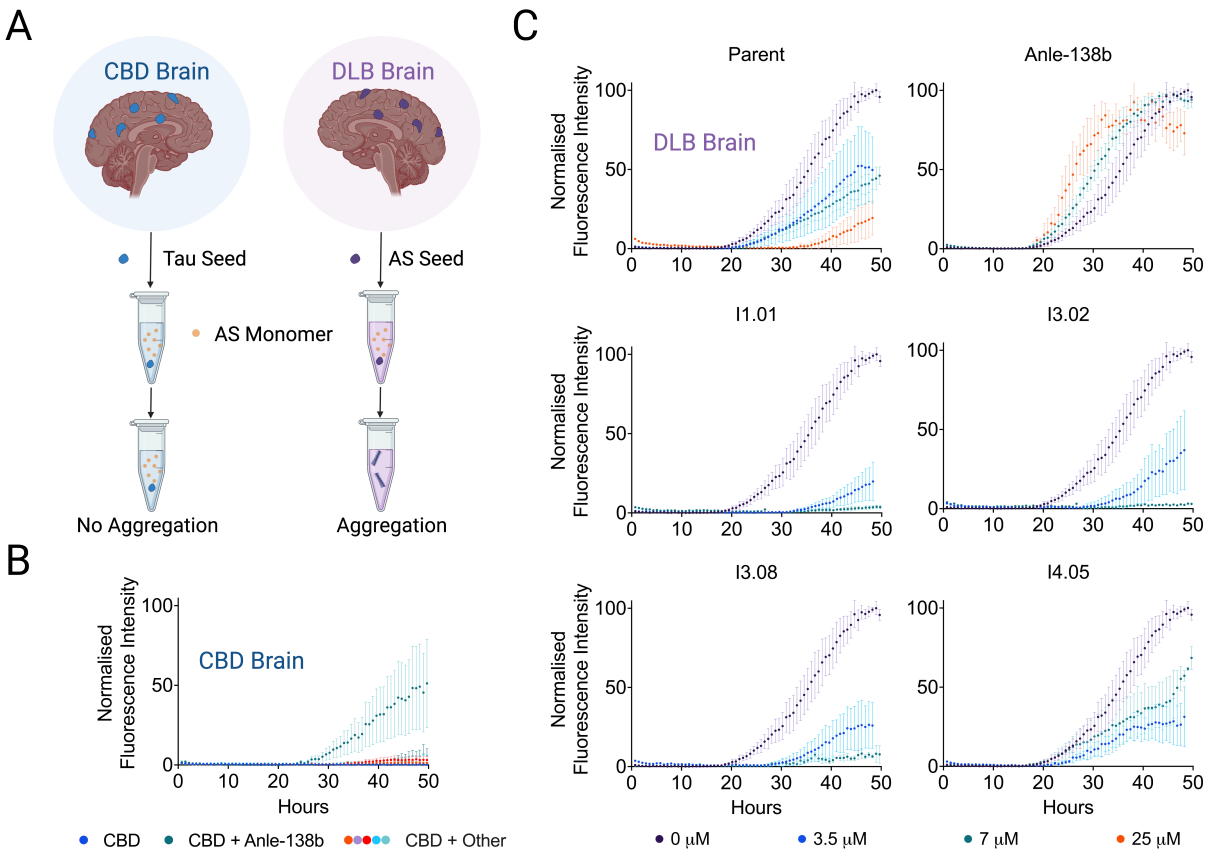


Figure 2.11. RT-QuIC brain seeding assay. (A) Schematic representation of the RT-QuIC assay, aggregates derived from the brain tissue of patients suffering with dementia with Lewy bodies (DLB) were used to induce αS aggregation. Samples from brains of patients with corticobasal degeneration (CBD) were used as a negative control. **(B)** Kinetic traces of a 7 µM solution of αS in the presence of CBD seeds (pH 8, 42°C, shaking at 400 rpm with 1 min intervals, in quadruplicate, error bars denote SD). CBD samples were 1% DMSO (blue), 7 µM Anle-138b (teal), parent (orange), I1.01 (purple), I3.02 (red), I3.08 (turquoise) and I4.05 (light blue). Anle-138b, in teal, induces aggregation under this condition. **(C)** Kinetic traces of a 7 µM solution of αS in the presence of DLB seeds. The DLB samples were 1% DMSO (purple), 3.5 µM molecule (blue), 7 µM molecule (teal) and 25 µM molecule (orange). Anle-138b again appears to accelerate rather than inhibit aggregation. Experimental work was carried out by Parvez Alam.

2.3.2. Inhibition of aggregation using brain-derived seeds.

While this result was encouraging, with the recent determination of the pathological αS fibril structure, it became clear that the recombinant in vitro fibril structure I had employed for computational and experimental work was different to that found in the brains of Parkinson's disease patients[23]. To test whether these molecules might work against patient-derived fibrils, these molecules were tested in a real-time quaking-induced conversion (RT-QuIC)[111] assay (**Figure 2.11**) that employs brain samples from patients suffering with Dementia with Lewy Bodies (DLB) to seed aggregation of αS. The dominant fibril structure identified in DLB was found to match the dominant structure observed in Parkinson's disease[23]. DLB is characterised by cognitive impairments such as hallucinations and dementia early in disease progression as well as motor impairment. Patients exhibit more aggressive spread and distribution of aggregates throughout neurons of the cerebral cortex when compared to those suffering from Parkinson's disease, which tends to only exhibit spread beyond the regions of the brain controlling motor functions in the later disease stages[50].

The RT-QuIC assay was initially introduced as a diagnostic assay[112,113], showing distinct aggregation curves in the presence of brain material derived from different pathologies[114]. In this case, I use it to test the ability for these molecules to slow the aggregation of αS induced by DLB brain material. As a negative control, samples from patients with a tauopathy (corticobasal degeneration, CBD) were also used, as these did not induce αS aggregation as only tau seeds were present rather than αS seeds (**Figure 2.11A, B**). CBD is a condition characterised by both cognitive and sensory loss as well as motor impairment (such as rigidity, which may be more pronounced on one side of the body), where aggregated tau is observed in the cortex of patient brains and elsewhere, with locations varying from patient to patient[115]. Tau is a family of closely related microtubule binding proteins (55-62 kDa). Mutations that induce tauopathies affect the alternative splicing of the protein, disrupt its interactions with microtubules, or increase its aggregation propensity. Alternative splicing of the *MAPT* gene leads to 6 tau isoforms of differing aggregation propensity, which differ by the number of N terminal inserts (0N, 1N or 2N), and microtubule binding repeats (3R, 4R). Tau aggregates in CBD are largely comprised of 4R tau, which differentiates this condition from other tauopathies, especially given 4R tau binds more strongly to microtubules and is less aggregation prone. Hyperphosphorylation of tau is also thought to be key to aggregation[45]. The conditions of this experiment were different to those initially screened, as this assay was carried

out at pH 8 and utilised shaking to induce aggregation. This is a more challenging paradigm for the molecules to function in as multiple aggregation processes occur in tandem[97]. In addition to secondary nucleation from the fibril surfaces, fragmentation of the fibrils induced via shaking results in more fibril ends for elongation, which in turn provides more fibril surface for secondary nucleation.

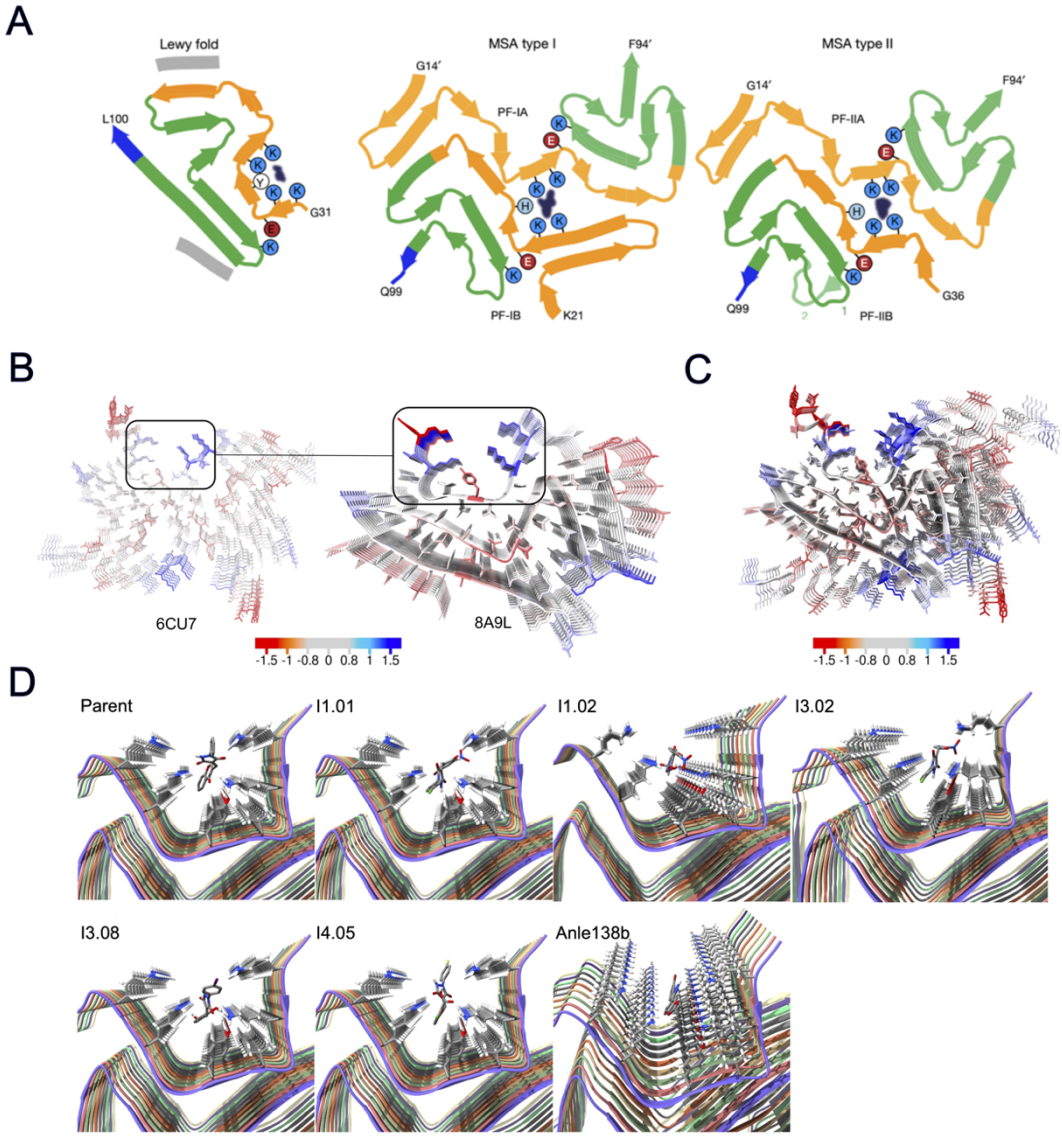


Figure 2.12. (A) Folds of the prevalent fibril polymorph in diseased brain material identified via cryo-EM in Parkinson's disease and dementia with Lewy bodies (8A9L), MSA type I and MSA type II. A common motif of 4 lysines enclosing an aromatic side chain (tyrosine in the Lewy fold and histidine in the MSA fold and 6CU7 fold) is observed in the polymorphs, with unidentified electron density in the pocket in each case (adapted from Yang, Y. *et al.*[23]). **(B)** Comparison of the cryo-EM structures of the 6CU7 (recombinant, initially targeted) and 8A9L (brain derived) with the homologous binding site indicated. **(C)** Structural overlap of the 6CU7 and 8A9L fibril structures, with the binding site in 6CU7 aligned with the similar binding site in 8A9L at the top of the diagram. The structures are coloured according to the CamSol residue solubility score[96]. **(D)** Schematics of the molecules bound in their lowest energy state within the 8A9L predicted binding site.

Despite these challenges, and the different fibril structure present, the molecules still function well in inhibiting aggregation, and still at substoichiometric ratios (**Figure 2.11C**). There is again a clear improvement for the leads over Anle-138b, which in fact appears to accelerate aggregation in this example, and the parent molecule, although the ranking of the leads in terms of efficacy is altered compared to the screening assay. To understand these results, we note that there is a similarity in the binding pockets in the structures 6CU7 (recombinant) and 8A9L (brain derived) (**Figure 2.12**). We currently do not know whether or not this similarity is serendipitous, but binding pockets with similar features can also be observed via cryo-EM in the MSA-I and MSA-II fibril folds as well as the Lewy fold, with an unresolved species bound within the pocket[23]. Multiple system atrophy (MSA) is another synucleinopathy that differs from PD as aggregation is most pronounced within the oligodendrocytes in the former and the neurons in the latter. This difference in cellular environment likely gives rise to the altered morphology of the aggregates.

To account for differences in brain samples and also investigate potential efficacy against MSA derived brain material, we tested a single concentration of the same selection of molecules against 3 neuropathologically confirmed MSA brain samples (**Figure 2.13A, C**) and 2 further DLB brain samples (**Figure 2.13A, D**). As a further negative control, a sample with no seed was tested, to determine the degree of spontaneous nucleation in the absence of brain material (**Figure 2.13B**). Aggregation in the negative control is effectively inhibited by all the ML molecules, given αS is likely to assume the 6CU7 polymorph in this condition, and not by Anle-138b which accelerates aggregation. It is possible that Anle-138b was not fully solubilised under this condition given the heat, shaking and the relatively high concentration of Anle-138b used relative to its reported solubility[60], which may have given rise to this effect. It should also be noted that the CBD samples are the better negative control for RT-QuIC, as all brain samples contain traces of cell matrix components that may sequester αS and reduce its aggregation. The unseeded sample begins aggregation at ~40-50 h whereas CBD samples do not exhibit significant aggregation over a span of 80 h (**Figure 2.14**). Fibrils present in DLB and MSA samples are able to counteract this effect. For the DLB and MSA samples broadly similar trends were observed to those shown in **Figure 2.11**. The ML molecules did appear more efficacious against MSA samples (**Figure 2.13C**), perhaps because the MSA pocket more closely matches that of the targeted 6CU7 polymorph (4 flanking lysines around a histidine residue) compared to the 8A9L polymorph found in PD and DLB (4 flanking lysines

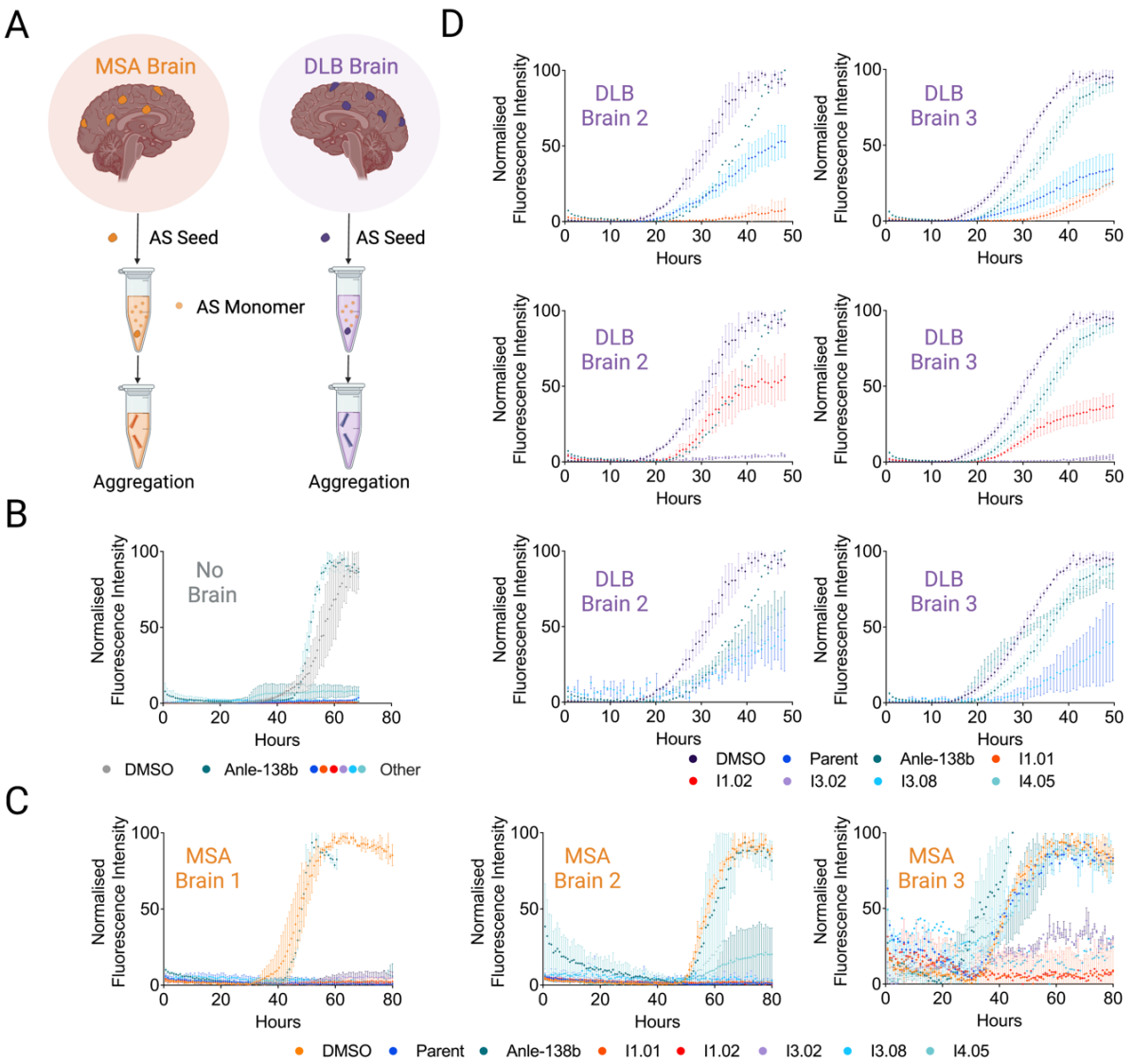


Figure 2.13. RT-QuIC brain seeding assay. (A) Schematic representation of the RT-QuIC assay, aggregates derived from the brain tissue of patients suffering with multiple system atrophy (MSA) or dementia with Lewy bodies (DLB) were used to induce αS aggregation. **(B)** Kinetic traces of a 7 µM solution of αS in the absence of brain material (pH 8, 42°C, shaking at 400 rpm with 1 min intervals, in triplicate, error bars denote SD). Unseeded samples were 1% DMSO (grey), 7 µM Anle-138b (teal), parent (blue), I1.01 (red), I3.02 (lilac), I3.08 (turquoise) and I4.05 (light blue). Anle-138b, in teal, induces aggregation under this condition. **(C)** Kinetic traces of a 7 µM solution of αS in the presence of MSA brain material. The MSA samples were 1% DMSO (light orange), 7 µM Anle-138b (teal), parent (blue), I1.01 (red), I3.02 (lilac), I3.08 (turquoise) and I4.05 (light blue). Anle-138b had no effect in samples 1 and 2 but appears to accelerate aggregation in sample 3. **(D)** Kinetic traces of a 7 µM solution of αS in the presence of DLB brain material. The DLB samples were 1% DMSO (purple), 7 µM Anle-138b (teal), parent (blue), I1.01 (red), I3.02 (lilac), I3.08 (turquoise) and I4.05 (light blue). Samples have been separated into different graphs for clarity. The DMSO and Anle-138b traces are shown on each graph, with 2 molecules from the docking or ML shown for comparison: Parent and I1.01 (top), I1.02 and I3.02 (middle), I3.08 and I4.05 (bottom). Anle-138b exerts a consistent mild inhibition for these two brain samples. Experimental work carried out by Parvez Alam.

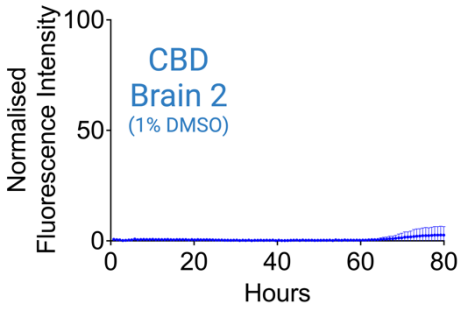


Figure 2.14. Kinetic trace of a 7 μM solution of αS in the presence of CBD brain material and 1% DMSO over a longer time course. No significant aggregation was observed over 80 h. Experimental work carried out by Parvez Alam.

around a tyrosine residue) as shown in **Figure 2.12**. The behaviour of Anle-138b was variable as, where the ML derived molecules inhibited aggregation to some extent across all examples, Anle-138b either had no effect (unseeded and MSA samples 1 and 2) or induced (CBD sample, MSA sample 3 and DLB sample 1) or mildly inhibited aggregation (DLB samples 2 and 3). No aggregation was observed in the CBD samples over the time scale observed except for Anle-138b, which accelerated aggregation under this condition.

2.3.3. Oligomer quantification by micro free-flow electrophoresis.

Having observed that molecule I3.02 was the most broadly effective in the RT-QuIC assay, an investigation was carried out to directly measure the oligomeric species formed during the reaction. This was achieved using microfluidic free-flow electrophoresis (μFFE)[116], a technique optimised using similar conditions to that used in the RT-QuIC assay, albeit at significantly higher αS concentration (100 μM). The results of this are shown in **Figure 2.15**. Aggregation time courses were tracked using AlexaFluor™ 488 labelled N122C rather than ThT. **Figure 2.15** shows a schematic of the approach, where samples were extracted from an aggregation time course, centrifuged to remove insoluble aggregates, and finally submitted to μFFE. The degree of deflection and the photon count of each particle are proportional to the size and charge of the biomolecule. The former allows the separation of monomers from oligomers and the latter gives a measure of the number and size of the oligomers at a particular time point in the presence of different inhibitors. Oligomer electrophoretic mobility (μ_o) for an oligomer comprised of n_m monomer units is proportional to oligomer charge (q_o) and inversely proportional to oligomer hydrodynamic radius (r_o) and so can be described by[116]

$$\mu_o \propto \frac{q_o}{r_o} \propto \frac{{n_m}^v}{r_o} \tag{2.1}$$

where v is a scaling exponent linking q_o with n_m. Approximating the oligomers as spherical species yields[116]

$$\mu_o \propto \frac{{n_m}^v}{r_m {n_m}^{\frac{1}{3}}} = \frac{{n_m}^{v^*}}{r_m} \tag{2.2}$$

where the oligomer electrophoretic mobility is defined only in terms of the monomer number (n_m) and hydrodynamic radius (r_m), and the scaling exponent $v^* = v - 1/3$. Samples were extracted at the $t_{1/2}$ of the negative control (1% DMSO) and the results are shown in **Figure 2.15**. Anle-138b dosing resulted in a smaller population of large aggregates, as may be expected from the slight acceleration in the aggregation observed in the fluorescence values, while I3.02 reduced both the size and the number of oligomers present in comparison to the DMSO control.

2.4. Discussion

In this chapter, we develop a machine learning approach to drug discovery for protein aggregation diseases that could improve both the optimisation rate of the in vitro assays employed and provide novel chemical matter more efficiently than conventional approaches. As of the first iterations, the optimisation rate of the approach using initial hit compounds identified via docking simulations was an over 20-fold improvement over typical HTS hit rates (~0-1%)[117]. These structures also represent discoveries that could not have been obtained by staying close in chemical space to the parent structure, as would have been dictated by similarity search approaches. There were ~4000 molecules in the test set that had Tanimoto similarity values in the same range as these leads, and all of these would potentially have had to be screened to locate these leads using similarity searches alone. This was demonstrated by the looser similarity search approach which exhibited a comparatively poor optimisation rate (4%) despite more conservative structural alterations to the parent hits. The machine learning method was therefore able to supply a degree of novelty as well as an improved optimisation rate.

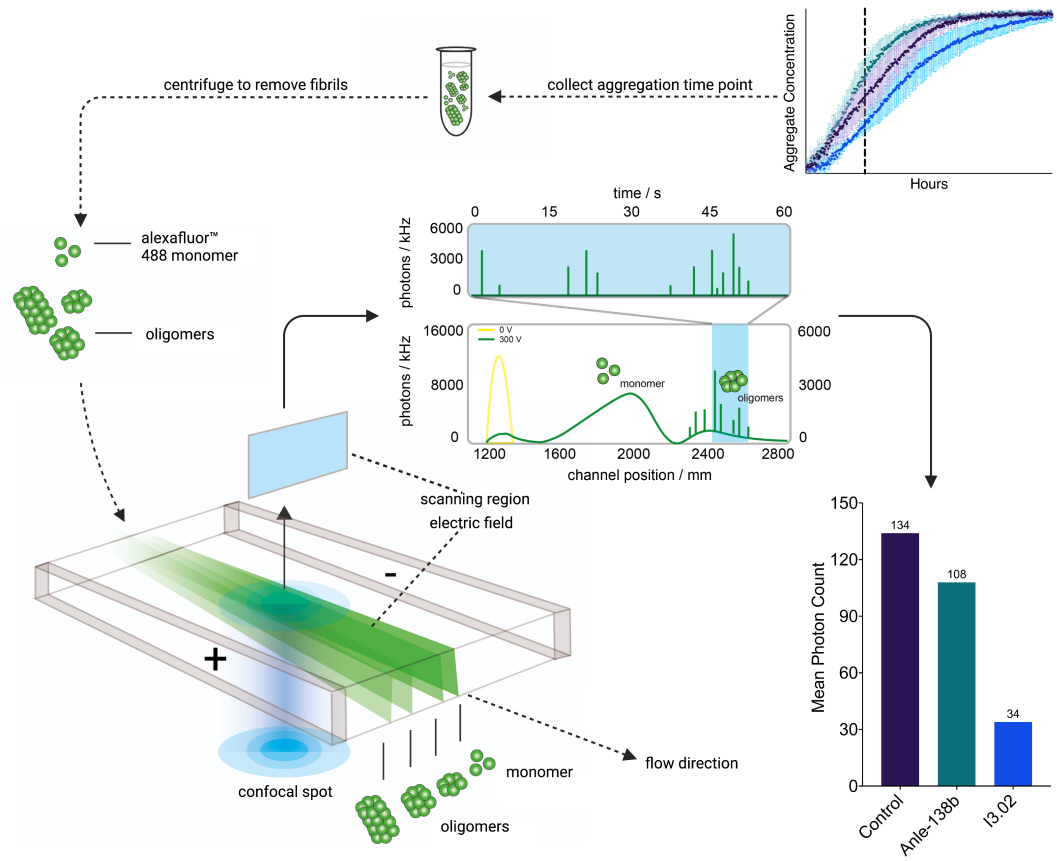


Figure 2.15. Quantification of αS oligomers using micro free-flow electrophoresis (μFFE). (Top right) αS labelled with AlexaFluor™ 488 (100 μM, pH 7.4, 37°C, cycles of 5 min shaking at 200 rpm and 1 min rest, in quadruplicate, error bars denote SD) was supplemented with 0.5 μM seed and 1% DMSO (purple) or 50 μM Anle-138b (teal) or I3.02 (blue) in 1% DMSO. Anle-138b slightly accelerates aggregation under these conditions, where fragmentation mechanisms may again play a role due to shaking, while I3.02 slows it down. Samples were extracted at 9 h from the time course of aggregation and centrifuged to remove fibrils from the mixture, leaving only αS monomers and soluble oligomeric species for analysis via μFFE. (Bottom left) Schematic representation of the μFFE approach, showing the AlexaFluor™ 488-labeled αS oligomeric mixture undergoing μFFE. The direction of fluid flow is shown by arrows. The differential deflection of the electric field allows the monomer population to be separated from the oligomer population during analysis (Middle and bottom right). Analysis of the aggregate populations detected in each sample. The number of photons emitted, proportional to particle number and size, is plotted on the y axis of the bar plot for each sample. The average number of photons emitted per particle is indicated in the inset. Experimental work carried out by Ewa Andrzejewska and I.

A limitation of this approach is the requirement to select molecules from a pre-existing library. To resolve this limitation generative modelling combined with reinforcement learning is applied in the next chapter to remove the need for a library to screen from[101,118]. A second limitation is the focus on one assay metric of interest as a learning parameter. Addressing this limitation will involve future work on multi-parameter optimisation, which is a challenging

area in rapid development[119-122]. Another topic of great interest in drug discovery approaches based on machine learning besides potency prediction is the prediction of pharmacokinetics and toxicity[123,124]. It could be possible to achieve this multi-parameter optimisation utilising multiple models in parallel and then employing a joint ranking metric, or architectures that screen for individual metrics in series, although this has primarily been demonstrated with predicted chemical properties such as clogP and quantitative estimate of drug likeness (QED) rather than experimental results[119-121]. The clogP is a calculated measure of the lipophilicity of the molecule while QED is a metric that aggregates how similar in molecular properties a particular example is to successful drugs. While CNS MPO was not optimised in this chapter it was monitored. The molecules in this work were derived from a set that passed CNS MPO criteria in the initial docking simulation, and so the CNS MPO score of the whole aggregation inhibitor set is relatively favourable with most lead molecules exceeding the common cut off value of 4[94] (**Figure 2.7B**). I explicitly attempt to optimise both CNS MPO and potency in the next chapter.

It would have been preferable to begin this approach using seeds derived from relevant pathological brain material, but this was not possible, as neither structures nor samples for these were available at the start of this study. Nonetheless, we have demonstrated that these molecules still function against disease relevant inducers, likely because of the degree of commonality between the binding sites of the fibril polymorphs. The complete loss of function against another aggregation prone protein, Aß42, does however suggest specific functionality against αS.

The identification of inhibitors of αS aggregation based on chemical kinetics approaches has advanced to the point that specific steps in the aggregation process, including primary nucleation and secondary nucleation, can be targeted in a reproducible way[43,57,82]. The mechanism targeted in this work is the surface-catalysed secondary nucleation step, which is responsible for the autocatalytic proliferation of αS fibrils. In a recent initial report, initial hit molecules identified via docking simulations were shown to bind competitively with αS monomers along specific sites on the surface of αS fibrils[64,80,125]. Specific rate measures and other aggregation metrics were derived from these experiments allowing quantitative and reliable comparisons between molecules in terms of SAR and offering metrics to optimise structures of interest[57,63]. This has been augmented with tests against diseased brain material and detailed, experimental fibril binding and oligomer flux analyses.

The results that I have presented illustrate a drug discovery approach that involves an iterative structure-based machine learning strategy to generate potent protein aggregation inhibitors. The resulting leads offered a significant improvement in potency over the parent and clinical molecules and represented a major structural departure from them. I anticipate that using machine learning approaches of the type described here could be of significant benefit to researchers working in the field of protein misfolding diseases, and indeed early-stage drug discovery research in general.

2.5. Materials and methods

2.5.1. Compounds and chemicals

Compounds were purchased from MolPort (Riga, Latvia) or Mcule (Budapest, Hungary) and prepared in DMSO to a stock of 5 mM. All chemicals used were purchased at the highest purity available.

2.5.2. Recombinant αS expression

Recombinant αS was purified based on previously described methods[74,97,126]. The plasmid pT7-7 encoding human αS was transformed into BL21 (DE3) competent cells. Following transformation, the competent cells were grown in 6L 2xYT media in the presence of ampicillin (100 μg/mL). Cells were induced with IPTG, grown overnight at 28 ºC and then harvested by centrifugation in a Beckman Avanti JXN-26 centrifuge with a JLA-8.1000 rotor at 5000 rpm (Beckman Coulter, Fullerton, CA). The cell pellet was resuspended in 10 mM Tris, pH 8.0, 1 mM EDTA, 1 mM PMSF and lysed by sonication. The cell suspension was boiled for 20 min at 85 °C and centrifuged at 18,000 rpm with a JA-25.5 rotor (Beckman Coulter). Streptomycin sulfate was added to the supernatant to a final concentration of 10 mg/mL and the mixture was stirred for 15 min at 4 ºC. After centrifugation at 18,000 rpm, the supernatant was taken with an addition of 0.36 g/mL ammonium sulfate. The solution was stirred for 30 min at 4 ºC and centrifuged again at 18,000 rpm. The pellet was resuspended in 25 mM Tris, pH 7.7, and the suspension was dialysed overnight in the same buffer. Ion-exchange chromatography was then performed using a Q Sepharose HP column of buffer A (25 mM Tris, pH 7.7) and buffer B (25

mM Tris, pH 7.7, 1.5 M NaCl). The fractions containing αS were loaded onto a HiLoad 26/600 Superdex 75 pg Size Exclusion Chromatography column, and the protein (≈ 60 ml @ 200 μM) was eluted into the required buffer. The protein concentration was determined spectrophotometrically using ε280 = 5600 M^{-1} cm^{-1}. The cysteine-containing variant (N122C) of αS was purified by the same protocol, with the addition of 3 mM DTT to all buffers.

2.5.3. Labelling of αS

αS protein was fluorophore-labelled to enable visualisation by fluorescence microscopy. In order to remove DTT, cysteine variants of αS were buffer exchanged into PBS or sodium phosphate buffer by use of P10 desalting columns packed with Sephadex G25 matrix (GE Healthcare). The protein was then incubated with an excess of AlexaFluor™ 488 dye with maleimide moieties (Thermofisher Scientific) (overnight, 4 °C on a rolling system) at a molar ratio of 1:1.5 (protein-to-dye). The labelling mixture was loaded onto a Superdex 200 16/600 (GE Healthcare) and eluted in PBS buffer at 20 °C, to separate the labelled protein from free dye. The concentration of the labelled protein was estimated by the absorbance of the fluorophores, assuming a 1:1 labelling stoichiometry (AlexaFluor™ 488: 72000 M^{-1} cm^{-1} at 495 nm).

2.5.4. αS seed fibril preparation

αS fibril seeds were produced as described previously[74,97]. Samples of αS (700 μM) were incubated in 20 mM phosphate buffer (pH 6.5) for 72 h at 40 °C and stirred at 1,500 rpm with a Teflon bar on an RCT Basic Heat Plate (IKA, Staufen, Germany). Fibrils were then diluted to 200 μM, aliquoted and flash frozen in liquid N_2, and finally stored at -80 °C. For the use of kinetic experiments, the 200 μM fibril stock was thawed, and sonicated for 15 s using a tip sonicator (Bandelin, Sonopuls HD 2070, Berlin, Germany), using 10% maximum power and a 50% cycle.

2.5.5. Measurement of αS aggregation kinetics

αS was injected into a Superdex 75 10/300 GL column (GE Healthcare) at a flow rate of 0.5 mL/min and eluted in 20 mM sodium phosphate buffer (pH 4.8) supplemented with 1 mM

EDTA. The obtained monomer was diluted in buffer to a desired concentration and supplemented with 50 μM ThT and preformed αS fibril seeds. The molecules (or DMSO alone) were then added at the desired concentration to a final DMSO concentration of 1% (v/v). Samples were prepared in low-binding Eppendorf tubes, and then pipetted into a 96-well half-area, black/clear flat bottom polystyrene NBS microplate (Corning 3881), 150 μL per well. The assay was then initiated by placing the microplate at 37 ºC under quiescent conditions in a plate reader (FLUOstar Omega, BMG Labtech, Aylesbury, UK). The ThT fluorescence was measured through the bottom of the plate with a 440 nm excitation filter and a 480 nm emission filter. After centrifugation at 5000 rpm to remove aggregates the monomer concentration was measured via the Pierce™ BCA Protein Assay Kit according to the manufacturer's protocol.

For the lipid induced assay, small unilamellar vesicles (SUVs) containing 1,2-dimyristoyl-sn-glycero-3-phospho-L-serine (DMPS), Avanti Polar Lipids Inc., Alabaster, AL, USA), were prepared from chloroform solutions of the lipids as described previously[126]. Briefly, the lipid mixture was evaporated under a stream of nitrogen gas and then dried thoroughly under vacuum to yield a thin lipid film. The dried thin film was re-hydrated by adding aqueous buffer (20 mM sodium phosphate, pH 6.5, 1 mM EDTA) at a concentration of 1 mM and heating to 40 ºC for 2 h while stirring at 1,500 rpm with a Teflon bar on an RCT Basic Heat Plate (IKA, Staufen, Germany). SUVs were obtained using several cycles of freeze-thawing followed by extrusion through membranes with 200 nm diameter pores (Avanti Polar Lipids, Inc). αS was prepared as above. Kinetic conditions were 20 μM αS, 100 μM DMPS, 50 μM ThT, 30 ºC, all other conditions remained the same as above.

Transmission electron microscopy (TEM) imaging of the fibrils produced at the end of the light seeded aggregation reaction (**Figure A.12**), was used to verify fibrils were produced

2.5.6. Determination of the αS elongation rate constant

In the presence of high concentrations of seeds (≈ μM), the aggregation of αS is dominated by the elongation of the added seeds[74,97]. Under these conditions where other microscopic processes are negligible, the aggregation kinetics for αS can be described by[57,64,74]

$$\left.\frac{dM(t)}{dt}\right|_{t=0} = 2k_{+}P(0)m(0)$$

where *M(t)* is the fibril mass concentration at time *t*, *P(0)* is the initial number of fibrils, *m(0)* is the initial monomer concentration, and k_+ is the rate of fibril elongation. In this case, by fitting a line to the early time points of the aggregation reaction as observed by ThT kinetics, $2k_+P(0)m(0)$ can be calculated for αS in the absence and presence of the compounds. Subsequently, the elongation rate in the presence of compounds is expressed as a normalised reduction as compared to the elongation rate in the absence of compounds (1% DMSO).

2.5.7. Determination of the αS amplification rate constant

In the presence of low concentrations of seeds (~ nM), the fibril mass fraction, *M(t)*, over time was described using a generalised logistic function to the normalised aggregation data[57,127]

$$\frac{M(t)}{m_{tot}} = 1 - \frac{1}{\left[1 + \frac{a}{c}e^{\kappa t}\right]^c}$$

where m_{tot} denotes the total concentration of αS monomers. The parameters a and c are defined as

$$a = \frac{\lambda^2}{2\kappa^2}$$

$$c = \sqrt{\frac{2}{n_2(n_2+1)}}$$

The parameters λ and κ represent combinations for the effective rate constants for primary and secondary nucleation, respectively, and are defined as[127]

$$\lambda = \sqrt{2k_+k_n m_{tot}^{n_c}}$$

and

$$\kappa = \sqrt{2k_+k_2 m_{tot}^{n_2+1}},$$

where k_n and k_2 denote the rate constants for primary and secondary nucleation, respectively, and n_c and n_2 denote the reaction orders of primary and secondary nucleation, respectively. In this case, n_c was fixed at 0.3 for the fitting of all data (corresponding to a reaction order of n_2 = 4), and k_2, the amplification rate, is expressed as a normalised reduction for αS in the presence of the compounds as compared to in its absence (1% DMSO).

2.5.8. Determination of the αS oligomer flux over time

The theoretical prediction of the reactive flux towards oligomers over time was calculated as[57,127]

$$\phi(t) = \frac{1}{r_+} \cdot \left[\frac{m(0)}{m(t)} \cdot \frac{d^2M}{dt^2} + \frac{1}{m(0)} \left(\frac{m(0)}{m(t)} \cdot \frac{dM(t)}{dt} \right)^2 \right]$$

where $r_+ = 2k_+m(0)$ is the apparent elongation rate constant extracted as described earlier, and $m(0)$ refers to the total concentration of monomers at the start of the reaction.

2.5.9. Recombinant Aß42 expression

The recombinant Aß42 peptide (MDAEFRHDSGY EVHHQKLVFF AEDVGSNKGA IIGLMVGGVV IA), here called Aß42, was expressed in the E. coli BL21 Gold (DE3) strain (Stratagene, CA, U.S.A.) and purified as described previously. Briefly, the purification procedure involved sonication of E. coli cells, dissolution of inclusion bodies in 8 M urea, and ion exchange in batch mode on diethylaminoethyl cellulose resin followed by lyophylisation. The lyophilised fractions were further purified using Superdex 75 HR 26/60 column (GE Healthcare, Buckinghamshire, U.K.) and eluates were analysed using SDS-PAGE for the presence of the desired peptide product. The fractions containing the recombinant peptide were combined, frozen using liquid nitrogen, and lyophilised again.

2.5.10. Aß42 aggregation kinetics and fibril preparation

Solutions of monomeric Aß42 were prepared by dissolving the lyophilised Aß42 peptide in 6 M guanidinium hydrocholoride (GuHCl). Monomeric forms were purified from potential

oligomeric species and salt using a Superdex 75 10/300 GL column (GE Healthcare) at a flowrate of 0.5 mL/min, and were eluted in 20 mM sodium phosphate buffer, pH 8 supplemented with 200 μM EDTA and 0.02% NaN_3. The centre of the peak was collected and the peptide concentration was determined from the absorbance of the integrated peak area using $\varepsilon 280 = 1490$ l mol^{-1} cm^{-1}. The obtained monomer was diluted with buffer to the desired concentration and supplemented with 20 μM thioflavin T (ThT) from a 2 mM stock. Each sample was then pipetted into multiple wells of a 96- well half-area, low-binding, clear bottom and PEG coated plate (Corning 3881), 80 μL per well, in the absence and the presence of different molar-equivalents of small molecules (1% DMSO). Assays were initiated by placing the 96-well plate at 37 ºC under quiescent conditions in a plate reader (Fluostar Omega, Fluostar Optima or Fluostar Galaxy, BMGLabtech, Offenburg, Germany). The ThT fluorescence was measured through the bottom of the plate using a 440 nm excitation filter and a 480 nm emission filter. Fibrils were extracted directly from wells and used on the day for SPR experiments.

2.5.11. Machine learning implementation and code availability

Code Availability:
Full code can be found on the GitHub repository: https://github.com/rohorne07/Iterate

Junction tree neural network variational autoencoder. The autoencoder[83] was pretrained on a library of 250,000 compounds[84], and was implemented as described previously[83] using a pip installable version (https://github.com/LiamWilbraham/jtnnencoder). Any molecules that contained substructures the autoencoder could not represent (i.e. that fell outside the substructure vocabulary of the pretrained model) were excluded.

Prediction module. All coding was carried out in Python 3. Scikit-learn[128] implementations of the Gaussian process regressor (GPR), random forest regressor (RFR), linear regressor (LR) and multi-layer perceptron (MLP) methods were tested in various combinations, and the results are shown in the supplementary section. For data handling, calculations and graph visualisation the following software and packages were used: pandas[129], seaborn[130], matplotlib[131], numpy[132], scipy[133], umap-learn[102], Multicore-TSNE[107] and GraphPad Prism 9.1.2. Cross validation and

benchmarking were also carried out for each model using scikit-learn built in functions and is described in the results section.

SHAP and latent space clustering. To compute the SHAP values, we used the SHAP python library[108]. The pretrained random-forest model was loaded, and a SHAP explainer object was created and provided with the latent representation for the top 100 highest predicted molecules. This allowed for the identification of dimensions important to the prediction of high potency molecules. The full testing set derived from the ZINC data set was also used in order to differentiate between dimensions important to distinguish high potency molecules from low potency molecules versus dimensions important to distinguish high potency molecules between themselves. This resulted in a global interpretation of the model, encompassing all data points passed to the explainer object. The resultant plots were generated using SHAP built-in plot functions. The sklearn library hierarchical clustering method was used to cluster latent vectors for comparison, with initial cluster number set to 7[134].

2.5.12. Surface plasmon resonance

All work was carried out using Biacore T200 at 25 °C. CM5 chips were activated by flowing 0.01 M NHS, 0.4 M EDC at a flow rate of 10 µL / min for 7 minutes over 2 lanes. Preformed αS or Aß42 fibrils (derived from the endpoints of low seeded aggregation reactions) at a concentration of 1 µM in sodium acetate (10 mM, pH 4.0) were injected onto a single lane in 60s bursts at 5 µL / min until a response of 2000 units was reached. Both lanes were then deactivated using a 7-minute injection of ethanolamine (1 M, pH 8.5) at 10 µL / min, and the reference lane signal was subtracted from the active lane. Different small molecule concentrations were then flowed over both lanes in a pyramidal arrangement in duplicate with blank subtraction (association time = 3 minutes, dissociation time = 10 minutes). The running buffer was sodium phosphate (20 mM, 1 mM EDTA, variable pH) with 1% DMSO. Fitting was carried out on Biacore T200 Evaluation Software, version 3.2, using a 1:1 binding model with the refractive index (RI) set to a constant value of 0 response units (RU).

2.5.13. Preparation of human brain tissue homogenates

Deidentified postmortem human brain specimens used in the RT-QuIC assay are referenced in **Table S2**. These specimens were obtained from the NIH Brain & Tissue repository-California, Human Brain & Spinal Fluid Resource Centre, VA West Los Angeles Medical Center, Los Angeles, California which is supported in part by National Institutes of Health and the US department of Veterans Affairs. Assay samples were prepared as 10% (wt/vol) brain homogenates in ice-cold phosphate-buffered saline (PBS) (pH 7.0) using 1 mm zirconia beads (BioSpec, cat#11079110z) in a Bead Mill 24 (Fisher Scientific). Subsequent dilutions of each brain homogenate (10^{-1} to 10^{-5}) for testing in the RT-QuIC assay were prepared in 1X PBS (pH 7.0).

2.5.14. αS RT-QuIC protocol

RT-QuIC assay for DLB samples were performed using the recombinant αS K23Q substrate purified using a 2-step chromatography protocol described previously (PMID: 29422107). For testing MSA samples, wild type αS (WT) recombinant substrate was purified using anion-exchange and size exclusion chromatography as described previously with minor modifications (PMID: 15939304). The WT protein expressing pET21a-αS plasmid was a gift from Michael J Fox Foundation MJFF (Addgene plasmid # 51486 ; http://n2t.net/addgene:51486 ; RRID:Addgene_51486). RT-QuIC assay was performed using black, clear bottom 96-well plates (Nalgene Nunc International) preloaded with 6 silica beads (1 mm diameter, OPS Diagnostics). Seeding was induced by addition of 2 µL of 10^{-4} (with respect to solid brain tissue) dilutions of DLB, MSA, or CBD (control) brain homogenates in quadruplicate wells containing 98 µL of the reaction buffer (40 mM phosphate buffer; pH 8.0 and 170 mM NaCl) supplemented with 6 µM (0.1 mg/ml) αS K23Q substrate (prefiltered through 100 kDa MWCO filter, Pall Corporation, Catalogue# OD100C34) and 10 µM ThT. After seeding, reaction plates were covered with a sealer film (Nalgene Nunc International) and incubated at 42 °C in a fluorescence plate reader (BMG FLUOstar Omega) with 1 min shake-rest cycles (400 rpm double orbital) for 50-90 h as indicated in the figures. ThT fluorescence ($\lambda_{excitation}$; 450 +/− 10 nm and $\lambda_{emission}$; 480 +/− 10 nm) was measured at 45 min intervals).

2.5.15. Microfluidic free-flow electrophoresis

Microfluidic device fabrication. Devices were designed using AutoCAD software (Autodesk) and photolithographic masks printed on acetate transparencies (Micro Lithography Services). Polydimethylsiloxane (PDMS) devices were produced on SU-8 moulds fabricated via photolithographic processes as described elsewhere[135,136] with UV exposure performed with custom-built LED-based apparatus[137]. Following development of the moulds, feature heights were verified by profilometer (Dektak, Bruker) and PDMS (Dow Corning, primer and base mixed in 1:10 ratio) applied and degassed before baking at 65 °C for 1.5 h. Devices were cut from the moulds and holes for tubing connection (0.75 mm) and electrode insertion (1.5 mm) were created with biopsy punches, the devices were cleaned by application of Scotch tape and sonication in IPA (5 min). After oven drying, devices were bonded to glass slides using an oxygen plasma. Before use, devices were rendered hydrophilic *via* prolonged exposure to oxygen plasma[138].

µFFE device operation. Liquid-electrode microchip free-flow electrophoresis (µFFE) devices were operated as described previously[139]. Briefly, fluids were introduced to the device by PTFE tubing, 0.012"ID x 0.030"OD (Cole-Parmer) from glass syringes (Gas Tight, Hamilton) driven by syringe pumps (Cetoni neMESYS). µFFE experiments were conducted with auxiliary buffer, electrolyte, monomer reference and sample flow rates of 1000, 200, 140 and 10 µL h^{-1}, respectively, for 15X reduction in buffer salt concentration for samples in PBS buffer.

Potentials were applied by a programmable benchtop power supply (Elektro-Automatik EA-PS 9500-06) via bent syringe tips inserted into the electrolyte outlets. Experiments were performed on a custom-built single-molecule confocal fluorescence spectroscopy setup equipped with a 488 nm wavelength laser beam (Cobolt 06-MLD 488 nm 200 mW diode laser, Cobolt). Photons were detected using a time-correlated single photon counting (TCSPC) module (TimeHarp 260 PICO, PicoQuant) with a time resolution of 25 ps.

Aggregation kinetics and sample extraction. AlexaFluor™ 488-labelled αS (100 µM) was supplemented with seed (0.5 µM) under shaking (200 rpm) at 37 °C, PBS pH 7.4 and either 1% DMSO or 50 µM molecule in 1% DMSO. Samples were extracted at the $t_{1/2}$ of the DMSO sample (9 hours). Fibrils were removed by centrifugation (21,130 rcf, 10 min, 25 °C) and the supernatant was then subjected to µFFE. For AlexaFluor™ 488-labelled oligomeric mixtures, auxiliary buffer comprised of 15X diluted PBS buffer, supplemented with 0.05% v/v Tween-20.

Using a custom-written script, single-molecule events were recorded as discrete events using a Lee filter of 4 from the acquired photon stream as fluorescence bursts with 0.05 μs of the maximum inter-photon time and containing 30 photons minimum. Using these parameters, the single-molecule bursts and their intensities were reported as a function of device position, which could be later converted to an apparent electrophoretic mobility. Oligomer bursts were distinctly characterised by a higher photon intensity detected per molecule and a higher electrophoretic mobility than monomeric protein.

2.5.16. Mass spectrometry

10 μM of preformed αS was incubated with 25 μM of molecule in 20 mM sodium phosphate buffer (pH 4.8) supplemented with 1 mM EDTA overnight under quiescent conditions at room temperature. The supernatant was removed for analysis using a Waters Xevo G2-S QTOF spectrometer (Waters Corporation, MA, USA).

2.5.17. Transmission electron microscopy

10 μM αS samples were prepared and aggregated as described in the kinetic assay, without the addition of ThT. Samples were collected from the microplate at the end of the reaction (150 hours) into low-binding Eppendorf tubes. They were then prepared on 300-mesh copper grid containing a continuous carbon support film (EM Resolutions Ltd.) and stained with 2% uranyl acetate (wt/vol) for 40s. The samples were imaged at 200kV on a Thermo Scientific (FEI) Talos F200X G2 S/TEM (Yusuf Hamied Department of Chemistry Electron Microscopy Facility). TEM images were acquired using a Ceta 16M CMOS camera.

2.6. Contributions

This chapter is substantially derived from the preprint doi: 10.1101/2021.11.10.468009 (accepted at *Nat. Chem. Bio.*). Michele Vendruscolo (M. V.) and I conceived the project, performed experiments, analysed data, and wrote the article. Specific contributions outside of this include the docking, performed by Z. Faidon Brotzakis, the RT-QuIC experiments performed by Parvez Alam and Ankit Srivastava, and the μFFE which was performed by Ewa

Andrzejewska and I. SHAP analysis was performed by Alice Aubert under my supervision as part of a summer student project. Rebecca C. Gregory produced the αS and Aß42.

3. Exploration and exploitation approaches based on generative learning

"The nice thing about artificial intelligence is that at least it's better than artificial stupidity."
– Sir Terry Pratchett

3.1. Generative modelling to expand the available chemical space

In recent years, deep learning has emerged as a powerful tool for cheminformatics[140]. With this capability, molecular generative models have emerged as promising tools for de novo molecular design. It has been shown in **Chapter 2** that computational methods can offer more efficient routes to αS aggregation inhibitors than traditional screening approaches[64,65]. A limitation of that approach was the use of pre-existing libraries to screen from, which biased the model and limited the search space. A further limitation was focussing only on the molecule potency during the machine learning task.

The work in this chapter remedies these shortcomings through the application of generative modelling approaches and multiparameter optimisation in two separate pipelines, one focussed on exploration (identifying novel and effective molecular structures) and the other on exploitation (achieving higher potency from known chemical space). The former employs an architecture derived from the GraphINVENT[141] framework for multiparameter generative modelling while the latter consists of a chemical language model (CLM) optimised for low

data regimes[142]. The GraphINVENT computational pipeline was developed as part of Mhd Hussein Murtada's Master's degree project under my supervision, while the CLM computational pipeline was developed as part of Donghui Huo's visiting studentship, also under my supervision. These pipelines are shown schematically in **Figure 3.1**.

Both pipelines feature a generative model linked to a QSAR filter. QSAR models are incorporated into generative pipelines to enable learning of the underlying relationship between the molecular structure and activity in silico[143]. Consequently, a smaller number of candidate molecules need to be tested in vitro. However, many constraints are involved in QSAR model training, such as the high dimensionality and sparsity of molecular fingerprints, in addition to the high correlation of the chemical descriptors. This makes ensemble learning models, especially Random Forest models (RFs), the most convenient and robust models for this task[144]. Moreover, one great advantage of RFs is interpretability, meaning they can be beneficial in identifying the common features of molecules with high activity levels against the target. As before, the QSAR models in this chapter predict whether a molecule can delay αS secondary nucleation. The experimental aggregation inhibition data set produced in **Chapter 2**[65] was small (453 molecules) and imbalanced, and so efforts were made to train several QSAR models to obtain acceptable accuracy.

In the initial phase of the exploration pipeline, a graph-based generative model was trained to generate drug-like molecules that could penetrate the blood-brain barrier (BBB) and reach the central nervous system (CNS). Then, the generative model was fine-tuned using reinforcement learning to generate the molecules with other desired properties including potency. For that, a scoring function was defined based on two complementary QSAR molecular activity classifiers trained on experimental aggregation data. RFs make predictions by combining the results of a set of individual decision trees that train simultaneously on subsets of the data set[145], therefore the number of predictors and their correlations do not create problems for RFs. These models were used in the reward function of a reinforcement learning model to generate new molecules with the desired activity. Using this architecture, small molecules were generated that were predicted to penetrate the BBB, and potentially delay αS aggregation.

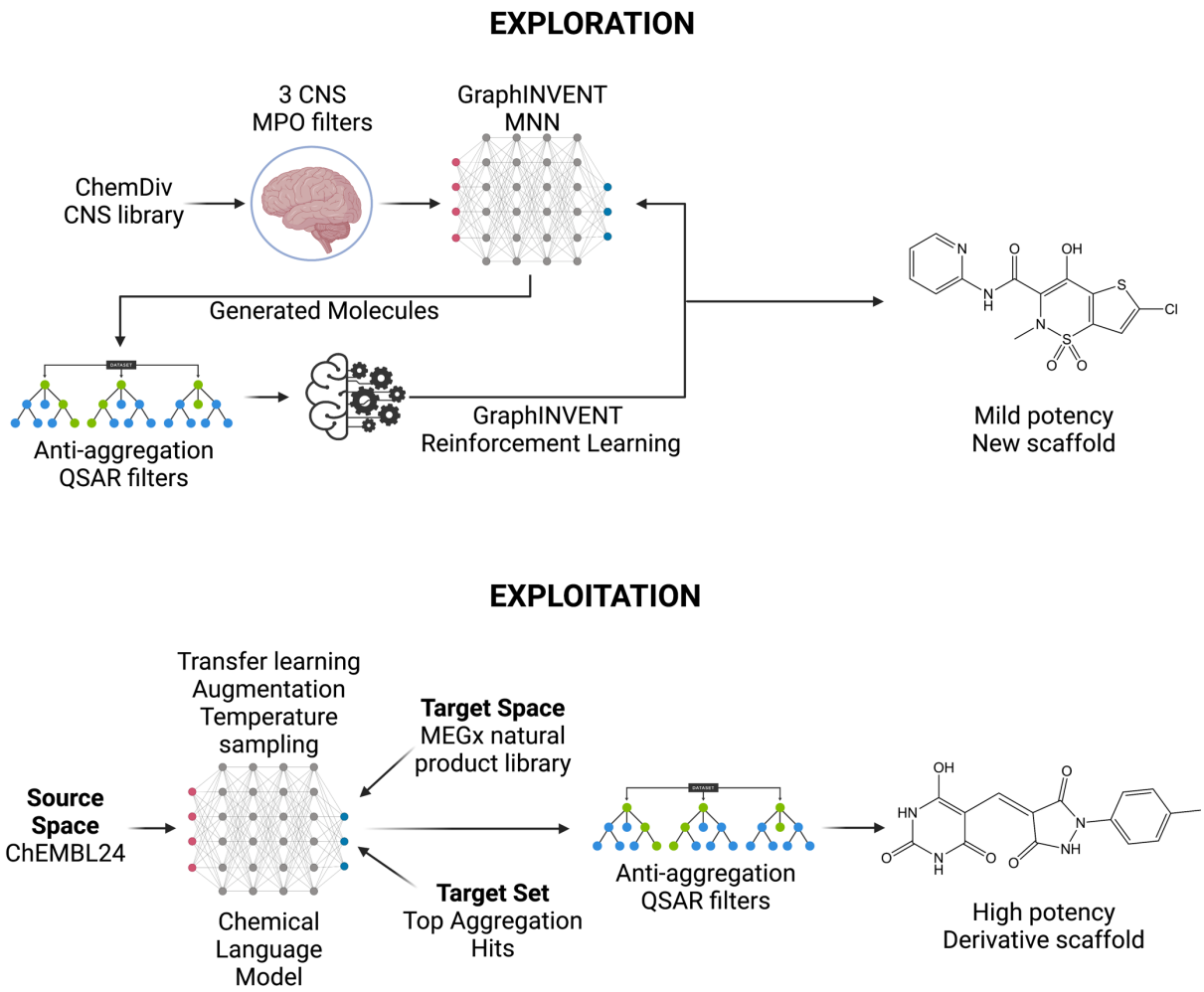


Figure 3.1. Schematic of the workflows for the exploration and exploitation pipelines. An exploration pipeline was initially pursued to complement the exploitation approach in **Chapter 2**. This has greater capacity to scaffold hop than the previous work and prioritises CNS MPO as well as potency. We then pursue a new exploitation pipeline to upgrade the method used in **Chapter 2** and replace the in silico library screening with generative modelling.

Most of the molecules, while synthetically accessible, were novel structures that were unavailable from screening libraries without custom synthesis at high expense (**Figure 3.2**). Experimental testing of these molecules was not possible for this reason, but the molecules showed strong overlap in the chemical space with the active leads found previously. As a further test an available molecule was tested from the generative model training set, which had been used in transfer learning to allow the model to create valid molecular structures. This molecule was predicted by the QSAR filters to have strong CNS properties and good anti-aggregation score. This molecule exhibited mild inhibition in the same range as existing clinical aggregation inhibitor Anle-138b[60].

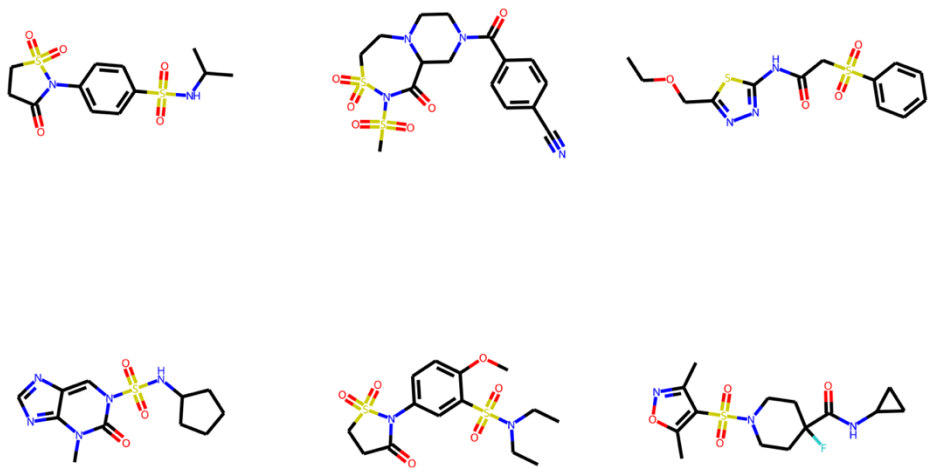


Figure 3.2. Examples of the generated molecules, which were novel and so could not be obtained without custom synthesis.

As an exploitation strategy had already been completed previously[65], in this chapter more focus is placed on the exploration approach introduced above. However, a weakness of the previous exploitation method was the use of a restricted area of the chemical space as it involved screening through a library of available compounds with a degree of similarity to the initial hits. I sought to remedy this limitation via the use of a generative chemical language model[142] (CLM), designed to function in the low data regime of this project and trained on the same aggregation set as used in the exploration pipeline. This approach employed: (i) transfer learning, (ii) temperature sampling, and (iii) data augmentation to enable the model to ably construct valid molecules with applications to the area of interest, despite very few data points. For transfer learning the model was pretrained on a synthetic compound space of bioactive molecules (ChEMBL24) to enable it to construct valid molecules with an increased likelihood of bioactivity. The model also used a library of natural products (MEGx collection, Analyticon Discovery GmbH) as a target space to optimise towards, thus indirectly optimising the pharmacokinetics of the resultant compounds via incorporation of features of a bioactive library. Rather than using a parameter such as CNS MPO as for the exploration pipeline, the aim was to imbue the generated molecules with features from a target set. Temperature sampling and data augmentation via shuffling of SMILES strings ensured the model achieved high uniqueness, validity, and novelty. As with the exploration pipeline the resultant molecules were screened for potency yielding a lead that rivalled the best molecules from the previous exploitation model in terms of potency. This molecule far outstripped the lead found via the GraphINVENT approach in terms of potency, a demonstration of the greater challenge

presented by explorative scaffold hopping compared to exploitation of known chemical space. The weaker lead could nonetheless be optimised via the exploitative approach described here, in a synergistic strategy combining the exploration and exploitation pipelines in series.

3.2. Exploration Pipeline Results

3.2.1. Creation of a library of small molecules with good CNS penetrance

To compile the training data set, the CNS drug libraries of small molecules provided by ChemDiv[146,147] were curated. In addition to these libraries, the molecules provided by the B3DB[148] data set were added. This is a benchmarking data set of BBB molecules compiled from 50 published resources and removed duplicates, creating a data set of 37,895 molecules. The data set was further filtered and assessed by a BBB permeability binary classifier[149], pre-trained on experimental brain permeability data, a CNS MPO score predictor[150], and a CNS MPO score calculator[94,95].

The CNS MPO scores are a commonly used metric for BBB penetrance in drug discovery and medicinal chemistry[94,95]. However, it is not possible to obtain the CNS MPO score of a molecule without using a machine learning predictor if pKa is included in the MPO, given that the pKa value cannot be calculated from the structure of a molecule, unlike other properties[151]. This makes the CNS MPO score prediction a regression task that highly depends on the precision of the pKa prediction. Therefore, in this project, multiple CNS MPO predictors were used to filter the initial library. A BBB permeability binary classifier (DeePred-BBB[149], using PaDEL[152] molecular descriptors as input features) and a CNS MPO score calculator not incorporating pKa prediction (GuacaMol[150]) were used alongside another CNS MPO calculator that did incorporate pKa prediction[94]. The first classified a molecule as penetrant or not using a database of experimentally tested molecules. This model had a high precision and AUC score (0.98 and 0.99, respectively) and good generalisability in the original work. AUC is ‘area under the curve’ for a receiver operating characteristic (ROC) curve of false positive rate on the x axis vs true positive rate on the y-axis plotted at different classification thresholds. It varies between 0 and 1, 1 implying a perfect true positive rate with no false positives at any classification threshold, while 0.5 would be the result for random selection. The second filter, which did not utilise pKa, calculated a probability between 0 and 1 for relevant molecular

properties of the molecule used in the CNS MPO score (molecules that achieved a probability >0.9 on average passed the filter). The third filter scored 6 calculated or predicted molecular properties (including pKa) between 0 and 1, and any molecule achieving a summed score of >4 was considered to pass[94]. The third filter was used in the previous chapter. These thresholds matched those used by the creators of these tools. Eventually, after filtering, the data set only contained the molecules that were classified to be BBB permeable by all 3 filters, removing 2260 molecules from the initial CNS ChemDiv data set. The distribution of CNS MPO scores calculated via GuacaMol for the filtered set and the structures of representative molecules within that set are shown in **Figure 3.3**, alongside structures with a range of lower CNS MPO values for comparison. This filtered data set became the training set for subsequent generative modelling.

3.2.2. Generative modelling

The GraphINVENT architecture was employed to generate molecules with desired properties. To convert the SMILES strings of the filtered CNS data set to graphs each SMILES string was turned into a node feature matrix, an adjacency tensor, and a vector *r* that resembles a step-by-step decoding route for the molecule i.e. steps to build the molecule starting from an empty graph. To obtain the vector *r*, the first step was the fragmentation of the molecular graph in a stepwise fashion using an algorithm developed in GraphINVENT. On each iteration, one edge/node was removed from the molecular graph G, and an action probability distribution (APD) was calculated for the new graph G_{n-1} until an empty graph was reached. Eventually, by aggregating APDs for all subgraphs G_n, G_{n-1}, G_{n-2}, . . ., we obtain the vector *r*

$$r = ((G_0, APD_0), (G_1, APD_1), \dots, (G_N, APD_N)) \quad (3.1)$$

The removal order of nodes and edges of the graph is determined by a breadth-first search (BFS) traversal[153].

We trained models to generate BBB-penetrant molecules and monitored the performance of the model in this respect. Therefore, three more evaluation metrics were added to GraphINVENT, using the 3 filters mentioned earlier: (1) the fraction of BBB permeable molecules, (2) the average calculated CNS MPO score, and (3) the average predicted CNS MPO score. These metrics were calculated for the novel molecules set generated by the model

while training every 2 epochs. To calculate the BBB permeable molecule fractions, the chemical descriptors of the generated molecules were computed using PaDEL software.

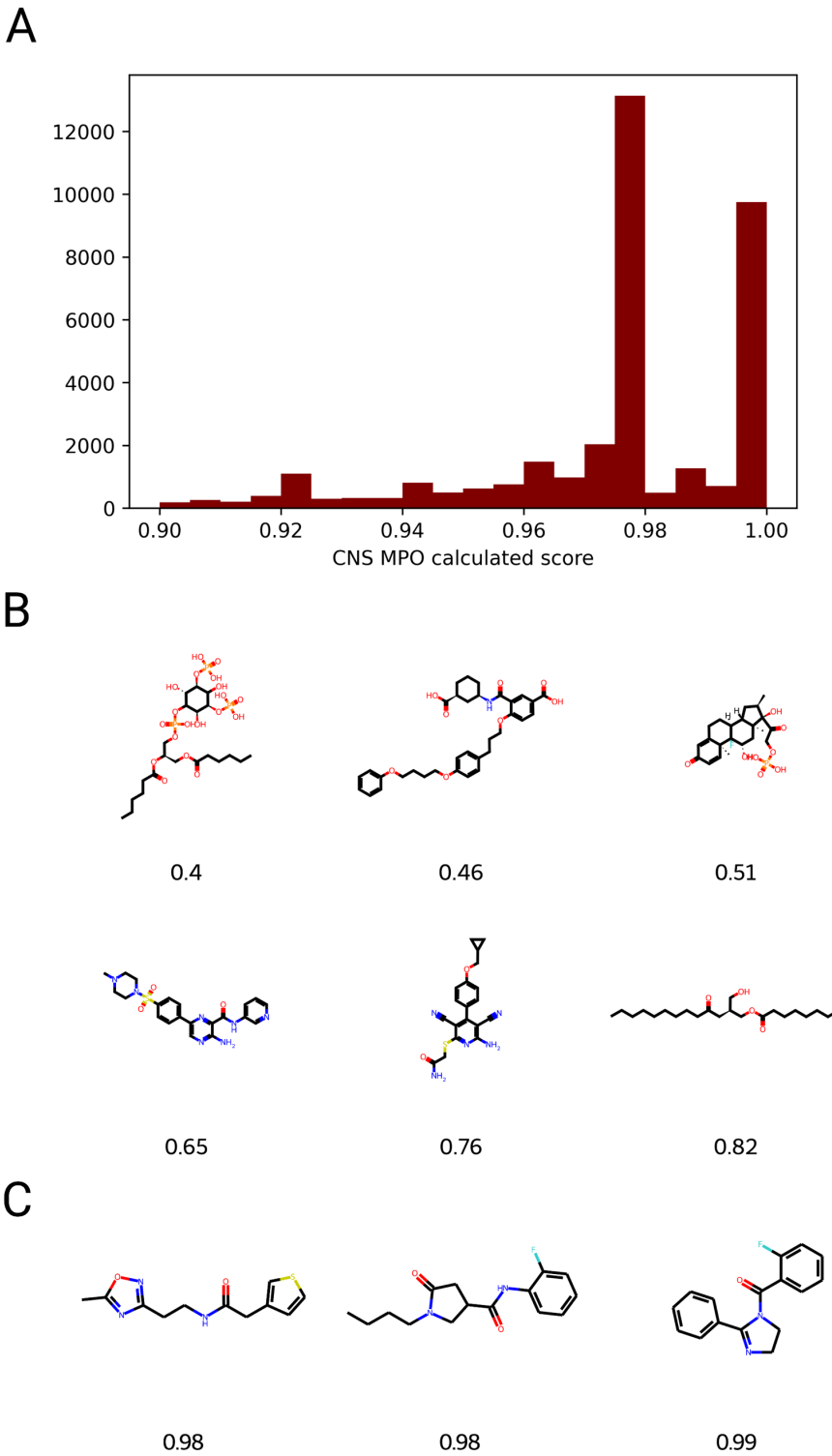


Figure 3.3. Creation of a library of small molecules with good CNS penetrance. (A) Calculated CNS MPO scores (GuacaMol) for the library subset of 35,636 molecules after filtration through the 3 different scoring methods (see text). **(B)** Randomly selected molecules spanning a range of lower CNS MPO values. (**C**) Representative molecules from the filtered set are shown. This data set was then used as the training set for molecule generation. Computational work was carried out by Mhd Hussein Murtada under my supervision.

Model (learning rate)	Metric			
	Epoch	BBB Fraction	Valid Fraction	Unique Fraction
GGNN (1×10^{-4})	72	1.0	0.92	1.0
GGNN (1×10^{-5})	66	1.0	0.94	1.0
MNN (1×10^{-4})	90	1.0	1.0	1.0
MNN (1×10^{-5})	84	0.925	0.84	1.0

Metrics are reported for the optimally performing epoch

Table 3.1. Metrics of molecules generated by MNN and GGNN at their best performing epoch for 2 different learning rates. The BBB fraction is the fraction of molecules classified as brain penetrant by DeePred-BBB. Computational work was carried out by Mhd Hussein Murtada under my supervision.

The two top performing models from the GraphINVENT package were selected, the gated graph neural network (GGNN) and the message neural network (MNN) and trained them with learning rates of 1×10^{-4} and 1×10^{-5} (4 training tasks in total). For each task, the data set was split into 80% training and 20% validation and trained the model for 100 epochs. Each time a data set passes through the model during training, and the model updates its parameters accordingly, is defined as an epoch. The MNN was found to run more efficiently given its less complex message passing and aggregation functions.

The training was done in mini batches of 50 molecules, with a block size of 1000 molecules. The loss function is the Kullback–Leibler[154] divergence which is generally used to measure the difference between probability distributions. In our case, the probability distributions to be compared are the target APD (P) and the predicted APD (Q) as

$$D_{KL}(P \parallel Q) = -\sum_{x=X} P(x)\log\left(\frac{Q(x)}{P(x)}\right) \tag{3.2}$$

An Adam optimiser was used with weight decay (L2 regulariser)[155]. Adam is the most ubiquitous method of efficient stochastic optimisation for learning of parameters during model training. The model was used to generate a batch of 100 new molecules every 2 epochs. These molecules were evaluated using the original GraphINVENT scoring metrics (**Table 3.1)** and the BBB permeability and CNS MPO metrics (**Tables 3.1** and **3.2**, respectively) implemented above. The goal was to determine the best combination of model architectures and learning rates, in addition to the epoch number in which the model performed best.

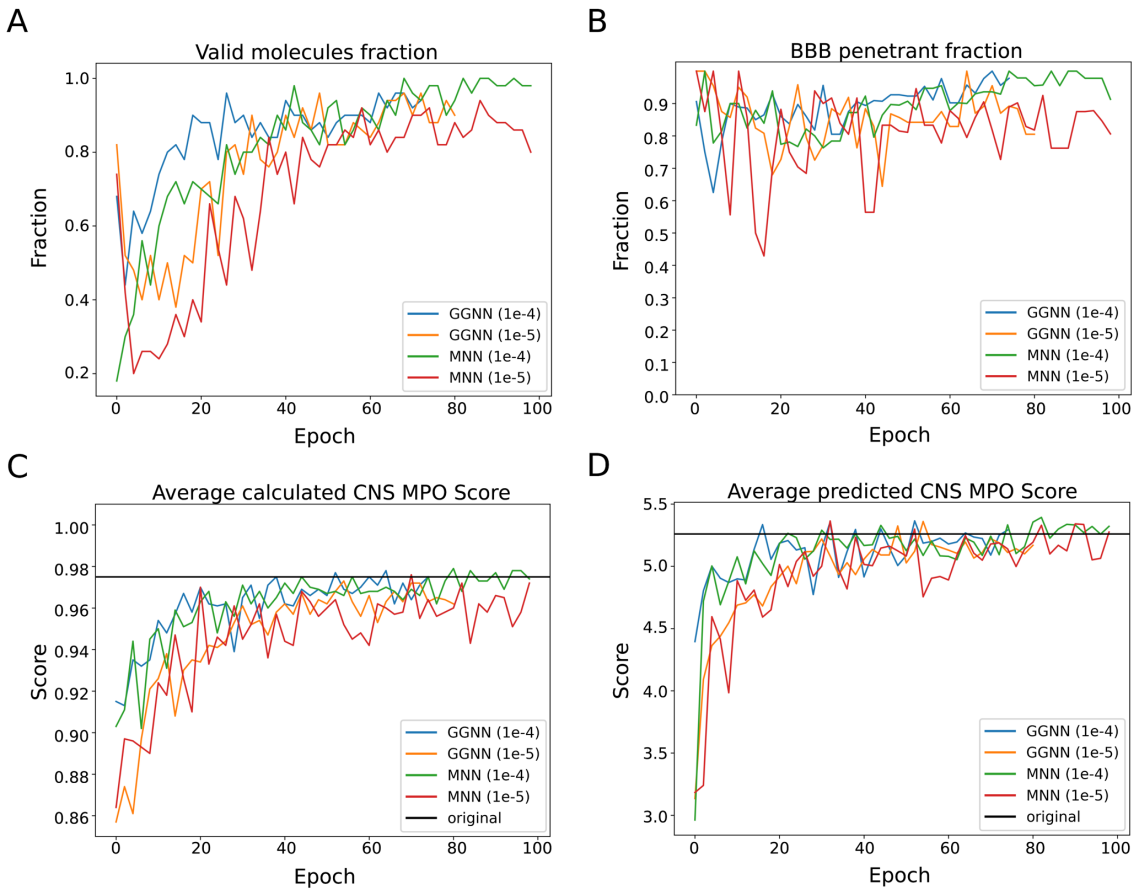


Figure 3.4. Metrics of generated small molecules during training with the GraphINVENT Gated Graph Neural Network (GGNN) and Message Neural Network (MNN) using 2 different learning rates. (A) Fraction of chemically valid molecules at each epoch. **(B)** Fraction of molecules passing the DeePred-BBB permeability classifier at each epoch. **(C)** Average calculated CNS MPO score using the GuacaMol implementation at each epoch. **(D)** Average predicted CNS MPO scores obtained using the method outlined in reference [95] at each epoch (black line indicates average of the original filtered training set).

Model (learning rate)	**Metric**				
	Epoch	Calc. CNS MPO	Change from original	Predicted CNS MPO	Change from original
GGNN (1×10^{-4})	72	0.964	-0.011	5.090	-0.170
GGNN (1×10^{-5})	66	0.963	-0.012	5.195	-0.075
MNN (1×10^{-4})	90	0.973	-0.002	5.337	+0.077
MNN (1×10^{-5})	84	0.972	-0.003	5.330	+0.070

Metrics are reported for the optimally performing epoch. Each metric has an associated 'Change from original' column, which refers to the mean change between the generated population and training set.

Table 3.2. CNS MPO average score comparison at the same epochs as in Table 1. The calculated CNS MPO score ranges between 0 and 1, 1 implying very high probability of BBB penetrance, while the predicted CNS MPO score ranges between 0 and 6. Computational work was carried out by Mhd Hussein Murtada under my supervision.

We observed that the MNN (1×10^{-4}) model outperformed the other three conditions in all metrics (**Tables 3.1** and **3.2** and **Figure 3.4**). All generated molecules were valid, unique, and BBB permeable. Moreover, the average predicted and calculated CNS MPO scores of its generated molecules were the closest to the score averages of the training data. Hence, this model was selected to be fine-tuned via reinforcement learning.

3.2.3. Reinforcement learning SMILES embedding based reward function

Having created a generator of BBB penetrant molecules, the focus moved to tailoring these molecules for potency against αS aggregation. Limitations were the size of the data set available for this task, consisting of 453 molecules, and the unbalanced nature of the data set (appendix **Figure B.1**), making the development of a high performing model challenging. In this initial proof-of-principle study, transfer learning was employed to at least in part remedy the data set size limitation. As a further measure, data were oversampled to ensure data set balance. The applied oversampling was a simple data augmentation by random duplication of the active molecules. Data were scaled afterward and split into training and testing sets (80%-20%). The metric of potency was the same as used in **Chapter 2**, the normalised half time ($t_{1/2}$) of aggregation, i.e. the time point at which 50% of the monomeric protein had converted to fibrillar aggregate, divided by the same 50% time point for the negative control. None of the active molecules in the aggregation data set were present in the generative model training set, as the aim was to identify novel structures.

Average	Metric		
	Precision	Recall	F1 Score
Micro Average	0.96	0.96	0.96
Macro Average	0.74	0.74	0.74
Weighted Average	0.96	0.96	0.96

Table 3.3. Metrics for SMILES embedding based model performance on aggregation data. Computational work was carried out by Mhd Hussein Murtada under my supervision.

For transfer learning a pretrained mol2vec[156] skip-gram model trained on a diversified set of 19.9 million molecules was used, so that the QSAR model would not have to learn molecular representations from scratch. The first hidden layer of the network was a frozen embedding layer initialised with the weights of the mol2vec model (these were preserved throughout training). The output of this layer was a 2D embedding vector generated based on the weights

from the base model. The next three layers were convolutional layers with a kernel size of 10 and a rectified linear unit (ReLU) activation function[157]. Between these layers, max pooling and dropout layers were added to reduce overfitting and minimise the feature space, followed by a long short-term memory (LSTM)[158] layer that greatly improved the performance, given its ability to identify trends in the data. Lastly, two dense layers with a softmax activation were added to normalise the prediction. For hyperparameters, Adam was used as an optimiser with learning rate = 1×10^{-4}, and the training loss was set to binary cross entropy. **Table 3.3** and **Figure B.2** show the metrics for the performance of this model.

We observed that the model could generalise reasonably well on the test data set. However, although an AUC score of 0.9 seemed appealing, there were many false positives in the predictions. This would be a critical issue when using this model as a reward function for reinforcement learning. The solution was to train another QSAR model that predicted molecular activity. The final reward function for reinforcement learning would then be based on the consensus of both models to increase the certainty of the prediction.

3.2.4. Reinforcement learning molecular descriptors-based reward function

Chemical descriptors were used as predictors instead of SMILES string embeddings in the second QSAR model. The idea behind this approach was that chemical descriptors are generally better able to quantify molecular properties than SMILES[153] and would reduce the classification problem and make it more explainable. Instead of learning molecular embeddings, the model would be learning measurable properties that could be compared among the molecules and associated with the output variable.

The chemical descriptors used as predictors were calculated by PaDEL software. They were the 2D and 3D physicochemical properties of the molecules, such as molecular weight, ring count and the moment of inertia (1875 descriptors in total). This meant there were more predictors than samples in the data set, meaning the model would be unable to generalise and elevating the risk that the model would learn the noise (irrelevant features) in the data. The solution was to apply feature selection with genetic algorithms, which use the principles of natural selection to identify molecular features that are most relevant to the prediction task[159,160]. Genetic algorithms are powerful in high-dimensional data sets with more features

than samples because they can handle complex, non-linear relationships between variables, whereas simple linear models such as Lasso rely on linear relationships[161]. Genetic algorithms also do not assume any distribution for the data or the errors and they can be more effective in finding an optimal set of features as they use a heuristic search method to explore the feature space. One additional advantage of this approach was that it helped to identify the common chemical properties among the active molecules.

Hence, a genetic algorithm was applied to find the best-performing subset of features when training a RF model. The features considered for selection were the most important ones identified by the trained RF model, given its ability to rank features based on the impurity (Gini impurity) of its underlying decision trees[87]. Feature importance values were calculated as the average of the impurity decrease accumulation within each decision tree of the model. A genetic algorithm mimics the process of natural selection to identify the subset of the most important features that maximise the model performance. First, an initial population of individuals was generated where each individual was a subset of features. The subsets were then scored by an RF model that predicted the target variable of interest, the anti-aggregation activity. Subsets with the highest scores were chosen to move to the next generation. Crossovers and mutations were applied so some features would switch places among the winner subsets while others would be added or removed randomly based on a mutation rate. Simple data cleaning and augmentation were applied before training the model and running the genetic algorithm to ensure data set balance.

A random grid search (with 3-fold cross-validation) was initially run for ten iterations to find the optimal hyperparameters for the RF model to ensure the best performance. After identification of the most important features, of which topological polar surface area was the most prominent, the genetic algorithm was run to select the subset of features that maximised the classification performance. **Figure 3.5** shows the ROC curve for this model and the features that were most strongly associated with the activity of the molecule according to the RFs. The hyperparameters used for RF and the genetic algorithm are shown in **Tables B.1** and **B.2**.

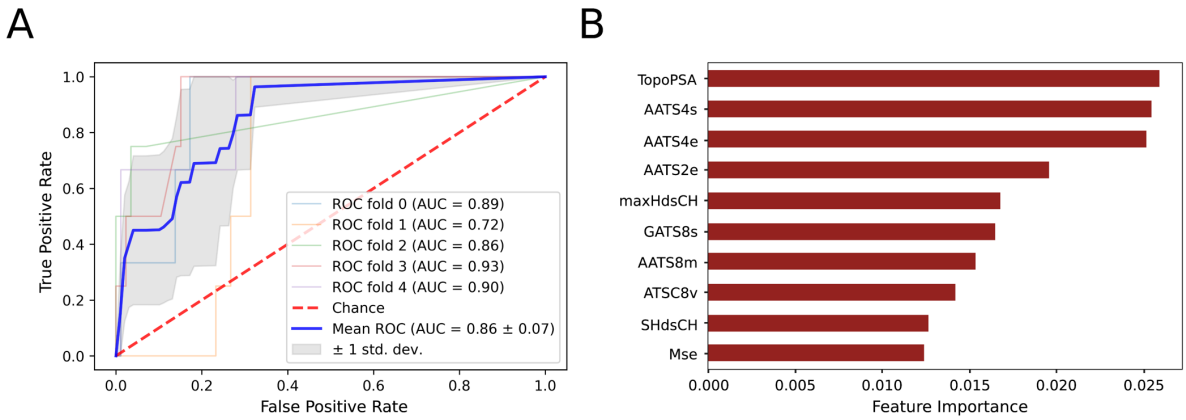


Figure 3.5. Metrics and important features in the descriptor-based RF QSAR model. (A) ROC AUC curve of the model with cross validation shown (AUC = 0.85). **(B)** Feature importance values derived from the RF QSAR model identify topological polar surface area as a key determinant. Computational work was carried out by Mhd Hussein Murtada under my supervision.

The metrics for this model vs the SMILES embedding based model are shown in **Table 3.4.** The descriptor-based model performance was an improvement, and it was better able to generalise than all previously trained models. The predictions had no false positives, and the model accuracy and average AUC scores were 0.98 and 0.85, respectively. A considerable improvement in the metric macro averages was observed for the descriptor-based model compared to the SMILES-based model which meant higher classification scores for the positive class and fewer false negatives. On the other hand, there was not a large difference in the weighted average metrics, given that both models could classify inactive molecules efficiently. Hence, both models were used in the reinforcement learning reward function, but the descriptor-based classifier was given a higher weight which was chosen based on the reinforcement learning performance.

Classifier	**Metric**		
	Precision	Recall	F1 Score
SMILES based	0.74 (0.96)	0.74 (0.96)	0.74 (0.96)
Descriptor Based	0.99 (0.98)	0.75 (0.98)	0.83 (0.97)

Macro averages are shown with weighted averages in brackets

Table 3.4. Comparison table for macro average metrics of the SMILES based model vs descriptor-based model, with weighted average shown in brackets. Computational work was carried out by Mhd Hussein Murtada under my supervision.

3.2.5. Final exploration model

A generative model was trained to produce BBB-permeable molecules and 2 QSAR classifiers were defined to filter the generated molecules based on anti-aggregation potency. The overall model architecture was then fine-tuned using reinforcement learning, an extension of the GraphINVENT package. The agents learn how to optimise the APDs of the generative model in order to maximise the QSAR reward functions. The loss function used for training was the best agent reminder loss (BAR)[162] which was responsible for the memory-awareness property of the model. This memorised the best agent with the highest score while training and was useful for reminding the new agents of the steps explored by previous agents to generate highly scoring molecules.

The fine-tuning process started by defining the prior and best agents and initializing them as the best performing MNN generative model outlined above. Then the following steps were repeated until the model converged to novel molecules with the highest scores:

- Generate a set of molecules using both priors (the current and the best).
- Score the molecules using the QSAR model.
- Compute the probabilities that the prior generative model and the current agent will assign the same actions carried out by the current agent to build a molecule.
- Compute the probabilities that the current and best agent will assign the same actions done by the best agent to build a molecule.
- Calculate the BAR loss and update the model weights to minimise it.

The prior generative model was the best performing MNN model outlined above, the hyperparameters were set as recommended in the initial paper and the learning rate was set to 1×10^{-4}. The best agent was updated every two epochs. The weights that were found to maximise the model performance after several training runs were 0.78 and 0.22 for the descriptor-based and SMILES-based models, respectively. The agents dealt with the score as a continuous value, meaning that the best agent was updated when the generated molecules gained a higher score than the last best score without any minimum thresholds for accepting the score.

After fine-tuning the model for 1000 epochs, it generated a set of novel small molecules that were predicted to be BBB permeable, druglike (high QED), and potentially able to delay the aggregation of αS. Most molecules had a CNS MPO score higher than the threshold (0.9) as calculated by GuacaMol, which meant that they had a high probability of being able to cross the BBB.

3.2.6. Investigation of generated molecules

While most of the molecules generated (**Figure 3.2**, **Figure B.3**) were not obtainable without custom synthesis, they showed an overlap (according to tSNE[107]) in the chemical space with the active molecules in the chemical inhibitor data set (**Figure 3.6A**). As a test of whether the QSAR reward functions worked appropriately, I ordered a compound (lornoxicam) within the original training set with high predicted anti-aggregation score to test experimentally in the aggregation assay used to generate the aggregation inhibition data set[65]. This was the same chemical kinetics assay used for initial screening in **Chapter 2**[43,57,82], which identifies the top compounds that significantly inhibit the surface-catalysed secondary nucleation step in the aggregation of αS. While this assay does not directly recapitulate the disease process, nor give a direct measure of oligomers, molecules previously screened through this assay showed both a prevention of aggregation seeded by diseased brain samples and also showed significant oligomer reduction, and so the assay acts as a useful screening proxy to filter potential molecules before these challenging experiments are required for validation. The potency of lornoxicam was mild in comparison to leads found previously, but was nonetheless observable and comparable to prior clinical anti-aggregation compound Anle-138b (**Figure 3.7)**. Since the original inhibitor training set only contained 4-6 distinctly different active structures, I anticipate that the performance of the model could improve as more varied training data were added.

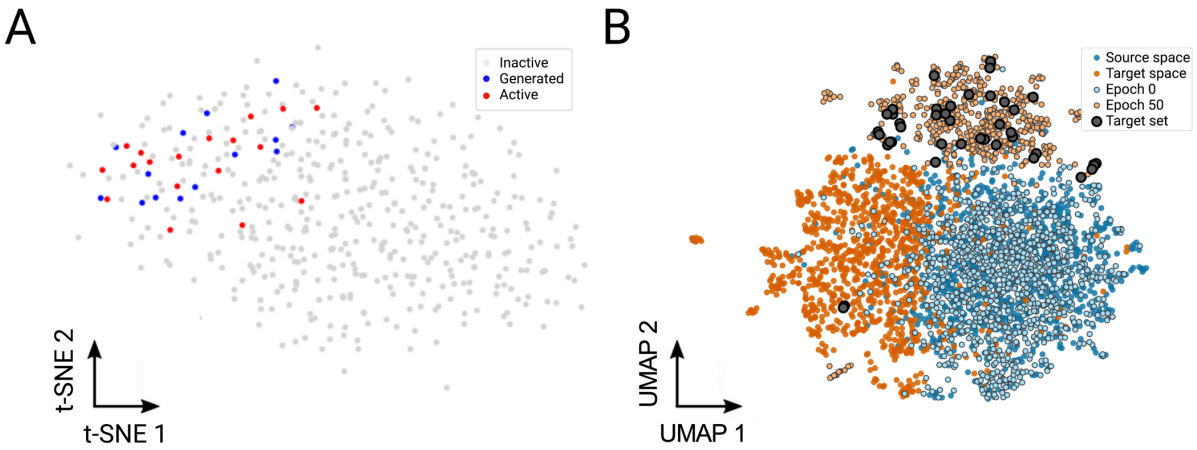


Figure 3.6. Chemical landscape of the exploration and exploitation strategies. (**A**) Comparison of the chemical space spanned by the chemical inhibitor training set and the newly generated compounds during the exploration strategy. t-SNE representation of the landscape of the chemical inhibitor training set with the original active (red) and non-active (grey) compounds, and the newly generated compounds (blue). (**B**) UMAP representation of the CLM molecule generation process. With successive iterations the generated molecules take on features similar to the target set (previously identified aggregation inhibitors) while incorporating features of a target space (natural product library). Computational work was carried out by Donghui Huo under my supervision.

3.3. Exploitation Pipeline Results

The exploitation pipeline employed the chemical language model (CLM) as previously described[142], using the bioactive library as a source space and the natural products library and the aggregation inhibitors as the target space and target set respectively. Over successive epochs, the generated molecules assume more of the features of the target space and target set, with a greater weighting assigned to the latter. Applied to the aggregation data set, a high number of training epochs were employed to ensure the resultant molecules did not deviate heavily from our selection of lead molecules, to increase the likelihood of potency. Initially, different selections of compounds were trialled from the aggregation data set, but we found that using <30 epochs and including milder potency structures as the 'target set' for the model led to a significant diversity of structures, few of which would be likely to achieve potency. This architecture could also be used in a less directed explorative approach by reducing the number of epochs and increasing the diversity of the target space, with the limitation that different parameters such as potency and CNS MPO could not be explicitly optimised for and weighted, as in the GraphINVENT pipeline.

We used 50 training epochs and only the top 20 lead structures as the target set to ensure generated molecules were close in the chemical space to the most potent structures. A UMAP representation of this process is displayed in **Figure 3.6B,** which shows molecule generation in the proximity to the area of interest around the top 20 leads. Due to lack of availability of the generated compounds, a similarity search was carried out for the first 500 generated compounds at epoch 50, which were subsequently rescreened through the QSAR model used in **Chapter 1**. 20 molecules were tested yielding 5 new leads, 1 of which (labelled CLM.1) showed a greater level of novelty compared to previously identified structures and which exhibited high potency (**Figure 3.7**). In this case a lead was classified as any molecule with a normalised half time of 1.5 or more. Historically, 1.5 was the cut-off that was used for aggregation inhibitors[63]. This was raised to 2 in **Chapter 2** to reduce the leads to be considered to a practical number. However, excluding CLM.1, milder potency was observed on average in this prototype pipeline (**Figure B.4**). The KIC_{50} value of CLM.1 (0.42 μM) was nonetheless on par with the best lead identified previously, I4.05 (0.52 μM), both of which compare very favourably with the parent hit (molecule 69) and Anle-138b which have KIC_{50} values of 18.2 μM and 36.4 μM (extrapolated) respectively. The structures of the leads derived from the CLM strategy, and their respective normalised half times are shown in **Figure B.4**.

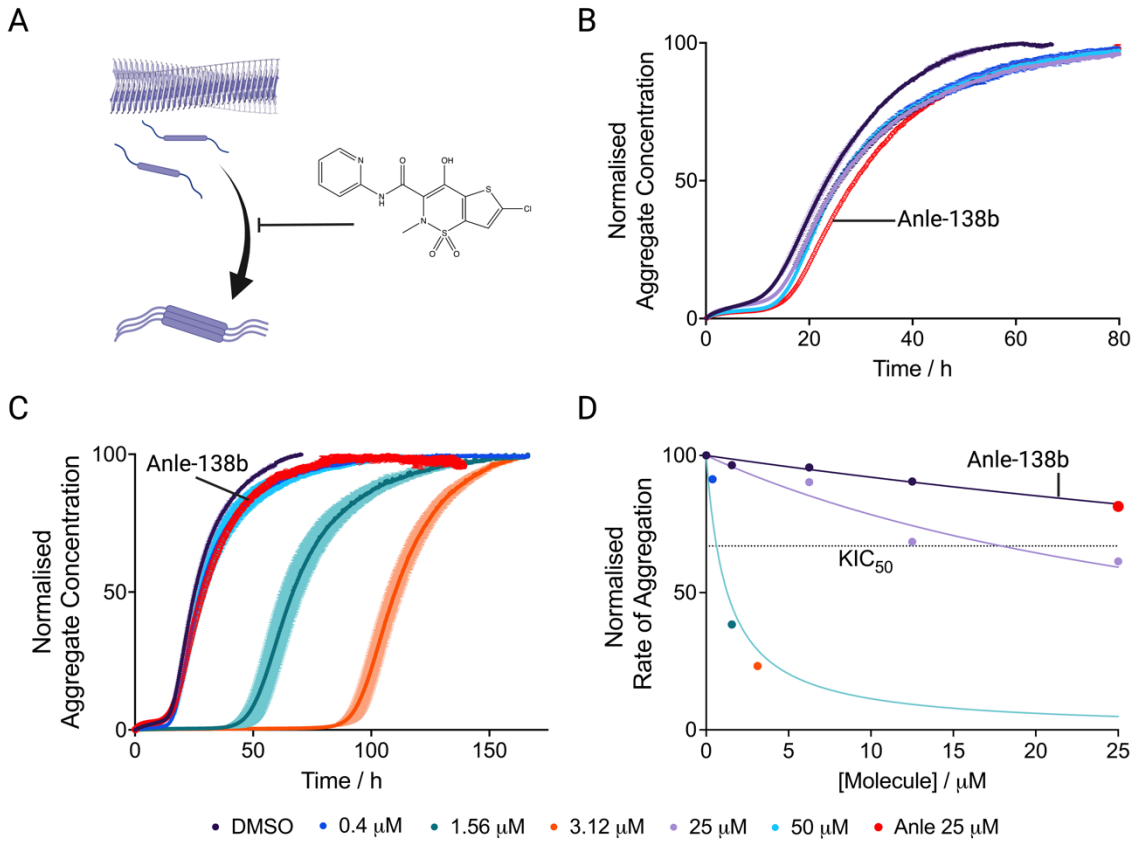


Figure 3.7. Experimental validation of compounds generated via the exploration and exploitation approaches presented in this work. (A) Schematic of the aggregation process. The dominant mechanism in oligomer formation is the nucleation of aggregates from the surfaces of existing ones (secondary nucleation). Small molecules can block this process through a proposed mechanism[65] of blocking fibril nucleation sites (lornoxicam is shown as an example). **(B)** Kinetic trace of a 10 μM solution of αS with 25 nM seeds at pH 4.8, 37°C in the presence of lornoxicam at 25 μM (lilac) and 50 μM (light blue) or in the presence of 1% DMSO (dark purple). Anle-138b (red) at 25 μM is shown as a control. **(C)** Kinetic trace of a 10 μM solution of αS with 25 nM seeds at pH 4.8, 37°C in the presence of CLM.1 at 0.4 μM (blue), 1.6 μM (teal) and 3.12 μM (orange), or in the presence of 1% DMSO (dark purple). Anle-138b (red) at 25 μM is shown as a control. The endpoints are normalised to the αS monomer concentration at the end of the experiment, which was detected via the Pierce™ BCA Protein Assay at t = 50 h. **(D)** Approximate rate of reaction (taken as $1/t_{1/2}$, and normalised between 0 and 100) in the presence of 3 different molecules, Anle-138b (purple), the parent structure of CLM.1 (lilac) and CLM.1 (blue). The KIC_{50} values of CLM.1 (0.42 μM) and its parent structure (18.2 μM) are indicated by the intersection of the fit and the horizontal dotted line.

3.4. Discussion

The objective of the machine learning approaches presented here was to demonstrate that small molecules balancing drug likeness, BBB penetrance and aggregation inhibition could be predicted, providing useful tools for therapeutic efforts against synucleinopathies. The results

illustrate the potential of generative machine learning methods to provide novel starting compounds with higher likelihood of efficacy against αS aggregation than conventional approaches. More generally, utilising exploitation and exploration pipelines in series is an effective strategy that can be applied to research projects requiring improvements in performance of small molecules and biomolecules in any assay of interest, while retaining molecular properties integral to good target engagement. Key to success in this approach is tailoring the architecture of the pipeline and the models within it for best performance, with greater emphasis placed on essential metrics. The pipelines that have been developed are concerned with the two main issues confronting research programs aimed at synucleinopathies: target engagement (blood brain barrier permeability) and potency (toxic oligomer reduction). The molecule tested for the exploration pipeline proved to be a mild inhibitor, but nonetheless marks a potential starting point for elaboration. Indeed, as shown in **Chapter 2**, the potency of an initial hit compound can be improved upon many fold if an exploitation strategy is pursued[65]. The exploitation strategy yielded a compound which made smaller departures from the previous lead compounds and yielded high potency, while addressing the restricted nature of the chemical space search approach previously employed. As such this methodology provides a strong complement to the previous work, and I anticipate that this will benefit researchers working in the field of protein misfolding diseases and drug discovery research in general.

3.5. Materials and methods

Full code can be found on the GitHub Repository: https://github.com/husseinmur/GraphINVENT-CNS

3.5.1. Compounds and chemicals

Compounds were purchased from MolPort (Riga, Latvia) or Mcule (Budapest, Hungary) and prepared in DMSO to a stock of 5 mM. All chemicals used were purchased at the highest purity available.

3.5.2. Recombinant αS expression

See **Section 2.5.2**.

3.5.3. Seed fibril preparation

See **Section 2.5.4**.

3.5.4. Measurement of aggregation kinetics

See **Section 2.5.5**.

3.5.5. Code availability

GitHub Repository: https://github.com/husseinmur/GraphINVENT-CNS.

3.6. Contributions

This chapter is substantially derived from *J. Chem. Theory Comput.* paper doi: 10.1021/acs.jctc.2c01303. Michele Vendruscolo (M. V.) and I conceived the project and wrote the article. The exploration and exploitation computational projects were executed by Mhd Hussein Murtada and Donghui Huo respectively, under my supervision as part of their master's degree/visiting student projects. I performed the laboratory experiments and analysis. Rebecca C. Gregory produced the αS.

4. Developing nanopore based oligomer detection methods

"One of the universal rules of happiness is: always be wary of any helpful item that weighs less than its operating manual." – Sir Terry Pratchett

4.1. Oligomer detection: challenges and solutions

The presence of protein oligomers is clearly associated with the onset and progression of several neurodegenerative disorders. Therapeutic efforts directed at this area have resulted in few approved drugs[163], however, in part because they are based on readouts related to fibrillar aggregates, which are the endpoint of the aggregation process. These highly ordered structures are thought to be largely inert in terms of neuronal toxicity, although they can catalyse the formation of further oligomers via secondary nucleation[57-60,62,164]. They are however large, intracellular, space occupying lesions, capable of disrupting cellular trafficking and transport and trapping important chaperones and enzymes[50]. To date, most investigations into the aggregation process rely on detection of fibrils using amyloid binding dyes, such as thioflavin T (ThT), that fluoresce strongly upon binding to fibrils. This approach, however, does not provide a direct measure of the oligomers present, the population of which vary according to the mechanism of aggregation[36,56].

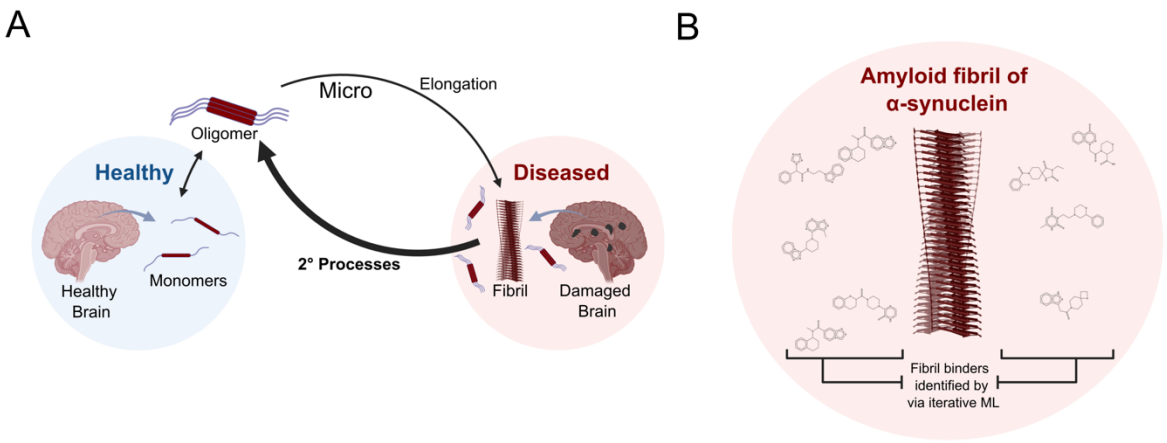


Figure 4.1. Schematic illustration of the process of αS oligomer formation in Parkinson's disease, and of its inhibition by compounds that can block secondary nucleation[65]. (A) Fibrils can catalyse oligomer formation via secondary processes such as secondary nucleation from catalytic sites on the fibril surface and fragmentation of the fibrils into smaller species. **(B)** A structure-based iterative machine learning strategy comprised of docking simulations followed by cycles of active machine learning was employed in **Chapter 2** to identify secondary nucleation inhibitors[65]. I3.08 from that work is used as a tool compound here.

The indirect nature of this measurement can cause problems, as it often means the exact inhibitory mechanism must be confirmed through further assays. This adds costs in terms of resources and time. For example, as has been argued, a promising therapeutic strategy is the blocking of secondary nucleation, which is a key accelerator of oligomer production (**Figure 4.1A**, **4.1B**)[63,165]. However, if fibril elongation were to be inhibited, this would slow the formation of endpoint fibril but increase the population of oligomers by shifting the aggregation pathway more strongly towards secondary nucleation (**Figure 4.1A**, **Figure 4.2A**)[56]. The molecule would nonetheless appear promising until tested in an elongation assay. The need to remove these false positives is a drain on any screening strategy.

Previous work has shown methods of isolating specific mechanisms of aggregation and their respective rates experimentally, and subsequently inferring the oligomer populations at a given time via fitting to an analytical model of the aggregation process[36,57]. Theoretical predictions were previously experimentally validated by taking samples during the aggregation process, tracked via ThT, and separating by size exclusion chromatography (SEC) before measuring the monomer equivalent oligomer concentration in each lyophilised sample via mass spectrometry (MS) or enzyme linked immunosorbent assays (ELISA)[36,166]. While this is a valid strategy, it is hampered by extremely low throughput and technical challenges in the implementation. Therefore, there remains a need to experimentally probe the oligomer population in a non-

disruptive and higher throughput manner to determine the size distributions of the oligomer population over time at single particle resolution[167].

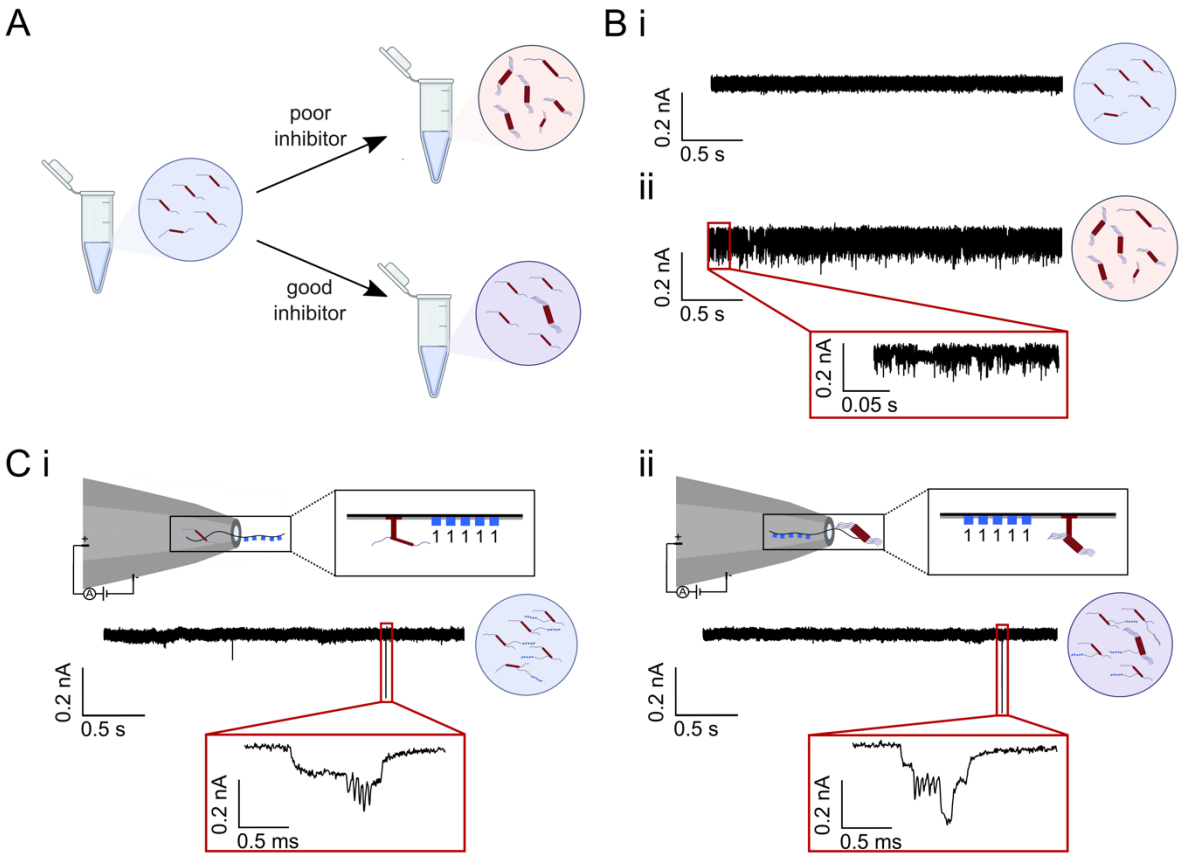


Figure 4.2. Schematic illustration of the method reported here to measure the efficacy of oligomer inhibitors, which is based on DNA nanostructure tagging of oligomers followed by detection in solid state nanopores. (A) Oligomer inhibitors have different efficacies, which have previously been challenging to establish given how difficult oligomers are to measure. **(B)** Previous approaches to oligomer measurement in nanopores have attempted to measure protein levels in the absence of any tagging methods, which is a difficult task prone to error given how challenging individual oligomer translocations are to reliably differentiate from each other and from monomer. Monomer **(i)** and heavily oligomerised **(ii)** samples are shown as examples in an uncoated pore with a diameter ~15 nm. Oligomers cannot be readily probed at a single molecule level via this approach, meaning that only bulk levels can be measured. **(C)** A novel oligomer measurement approach employing unique DNA nanostructure barcoding of each particle in a sample enables both single-molecule resolution of oligomers and multiplexing of samples, delivering improved metrics of inhibitor efficacy and increased throughput. **(i)** Monomeric protein with an attached barcode exhibits no adjacent spike, as the nanopore diameter has been tailored so that monomers do not generate a signal. **(ii)** A lightly oligomerised sample exhibits a clear spike in association with the unique barcode. The barcoded protein can enter the pore in either orientation (barcode first or protein first). Nanopore experiments were carried out by Sarah E. Sandler.

Thus far, single-molecule techniques have shown promising results in characterising oligomer distributions[167]. For example, confocal two-colour coincidence detection (TCCD)[168],

fluorescence correlation spectroscopy (FCS) measurements[169], single molecule total internal reflection fluorescence (TIRF) imaging[170], single-molecule spectrally-resolved points accumulation for imaging in nanoscale topography (sPAINT)[171], atomic force microscopy (AFM)[172], and micro free flow electrophoresis (μFFE)[116] have all allowed study of oligomer distributions under near physiological conditions. Additionally, it has been shown that using μFFE one can ascertain oligomer populations in the presence of specific secondary nucleation inhibitors[65,116]. An overview of these methods is shown in appendix **Figure C.1**. One major limitation of these approaches is that, while they are generally more efficient than the SEC/MS approach, the throughput at which inhibitors can be tested remains low.

A promising alternative towards achieving high-throughput is nanopore sensing, a single-molecule technique which relies on applying an electric field to drive molecules through a nanosize opening, allowing the measurement of changes in ionic currents relating to the size, shape, and charge of the molecule entering, or translocating, through the pore[173]. Broadly speaking, there are two types of nanopores, biological, based on pore-like proteins embedded in membranes, and solid-state, which are fabricated by creating nanosized openings in a material. Platforms containing biological nanopores are commercially available from Oxford Nanopore Technology. However, due to size, these are mostly restricted to DNA sequencing[174] or rely on protease cleavage of samples before nanopore measurements[175]. Recently, the ability to discriminate between αS variants has been accomplished using biological nanopores[176]. Using solid-state nanopores eliminates the need for fragmentation and allows the size of the nanopore to be directly tuned and optimised for detection of the analyte of interest. Auspiciously for potential high throughput applications, it has recently been demonstrated that solid-state nanopores could be manufactured at scale[174].

Previously, solid-state nanopores have proven to be a useful tool for the detection of proteins[177], as well as a way to study their conformations and interactions[178]. One of the major challenges associated with studying proteins in solid-state nanopores, however, is the rapid speed at which they translocate. This challenge can be overcome with approaches such as employing bilayer-coated solid-state nanopores[179], or by increasing the current bandwidth which increases the time resolution of the measurement[180]. In one case, αS oligomerisation was even studied in solid-state nanopores using a Tween-20 coating[181]. While these approaches are effective for studying single proteins, they are not easily adapted for multiplexed sensing. Current approaches are all based on observing single monomer or oligomer events, which can result in

ambiguous signals. Discerning individual particle translocations can be challenging and is often based on observed differences in noise profiles (**Figure 4.2B**). Additionally, these methods are low throughput as multiplexing is not possible.

Since it has been demonstrated that the combination of solid-state nanopores and digitally-encoded DNA nanostructures allows for highly multiplexed detection of single molecules[182,183], in this work, DNA nanostructures are used to study the effect of small molecule inhibitors of αS secondary nucleation in a multiplexed assay. The advantage of this approach is that every oligomer in a particular sample has a distinctive 'barcode', which clearly identifies each individual particle, and allows aggregates from different inhibitor screens to be mixed together and tested simultaneously (**Figure 4.2C**). This enables investigation of oligomer populations in more granular detail at higher throughput than was previously possible.

The small molecule inhibitors tested in this work were derived from the project described in **Chapter 2**[64,65]. In brief, inhibitors were initially identified via in silico docking to a putative catalytic site that promoted oligomer formation on the surface of αS fibrils followed by optimisation in aggregation assays via active machine learning[64,65,101]. Application of nanopore detection to quantitative protein oligomer analysis therefore offers another useful application of this technique, with the potential of high throughput analysis of a challenging target and an associated benefit to therapeutic programmes targeting these misfolded protein aggregates.

4.2. Results

4.2.1. DNA nanostructure design for the capture of αS oligomer capture

A DNA nanostructure was designed that could couple to azide-tagged αS aggregates and uniquely identify them (see methods **Section 4.4.6** and **Figure 4.3**). Sarah E. Sandler designed and assembled the DNA nanostructure and carried out nanopore measurements. Using a single-stranded DNA (ssDNA) backbone as a scaffold, complementary staple DNA oligonucleotides were combined with additional oligonucleotides for detection and digitisation in a one-pot reaction and annealed. DNA dumbbells allow digitisation of the structure. Their presence creates a structured spike in the nanostructure, while their absence leaves a flat spacer region,

corresponding to either a '1' or '0'. In the proof of concept presented here, only five spike/spacer regions were used, allowing for 2^5 (32) combinations of barcodes. This design was based on previous work and was optimised to create clearly distinguishable spikes in nanopores of ~15 nm diameter[182]. However, this has the potential to be expanded with further optimisation, and it has previously been shown that it is possible to fit 56 bits onto a single DNA carrier, allowing for a library of 2^{56} ($>10^{16}$) molecules[184]. Another section of the nanostructure contained two DNA strands, one 21 base pair (bp) sequence labelled with a dibenzocyclooctyne (DBCO) tag and one which had partial complementarity to both the barcoded scaffold and the sequence containing the DBCO, connecting the DBCO tagged region to the rest of the nanostructure (**Figure 4.3A**).

The DBCO-labelled nanostructure was then combined with azide-tagged N122C αS samples for click coupling and subsequent detection (**Figure 4.3B**). The azide-tagged N122C αS monomer was prepared via reaction of the reduced cysteine thiol with the iodoacetamide moiety of iodoacetamide-PEG_3-azide. This reaction was monitored until completion via LCMS (**Figure C.2**). The monomer was isolated via SEC before use in aggregation experiments and subsequent coupling to the DNA tags.

4.2.2. Detection of stabilised oligomers via DNA nanostructures and nanopores

I chose to first test the ability of the nanopores to act as a device to detect oligomers using a stabilised oligomeric species. Stabilised αS oligomers have been extensively characterised previously[185,186]. They are typically obtained using methods such as hyper-concentration and lyophilisation, and as such have limited physiological relevance. However, they do offer a useful test case for oligomer detection methods due to their greater stability, higher concentration and larger size[186]. Stabilised oligomers were used to optimise coupling times to the DBCO-tagged DNA barcodes and also to test whether an appreciable difference could be observed in monomeric and oligomeric samples in the nanopore. Successful click coupling of the samples was confirmed via PAGE (**Figure C.3**, **Table C.1**), where monomer-bound DNA was observable.

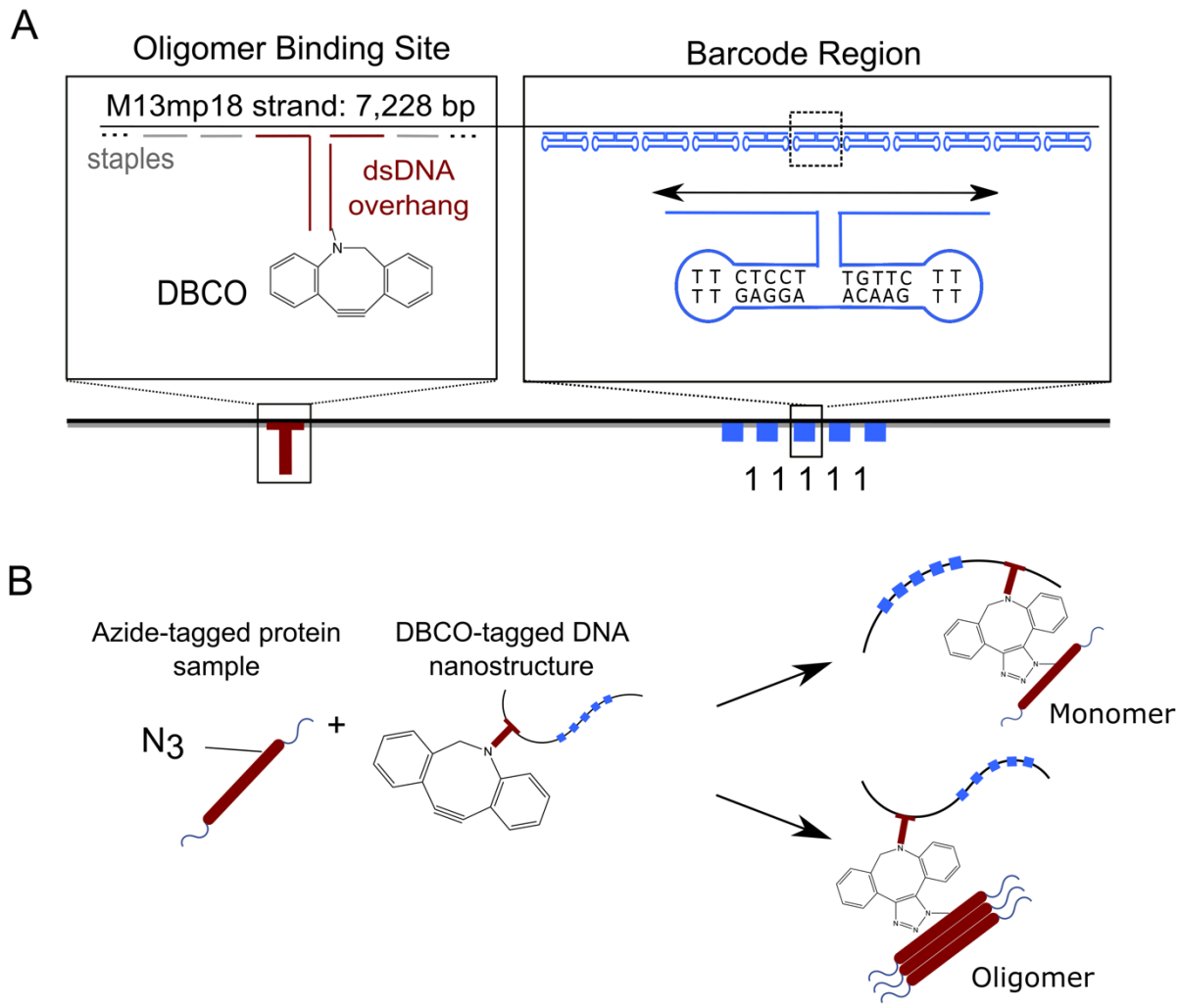


Figure 4.3. Design of a DBCO-DNA nanostructure for the capture of azide-labelled αS aggregates. (A) Schematic of the DNA nanostructure containing the DNA barcode region and a DBCO-tagged dsDNA overhang for click coupling to azide-tagged N122C-αS. DNA barcodes allow for a digital readout of the single-molecule translocations using DNA dumbbells to create distinct 1 or 0 bits. **(B)** N122C-αS is tagged with iodoacetamide-PEG$_3$-azide and then incubated with the DBCO-tagged nanostructure, allowing facile click coupling of the two components.

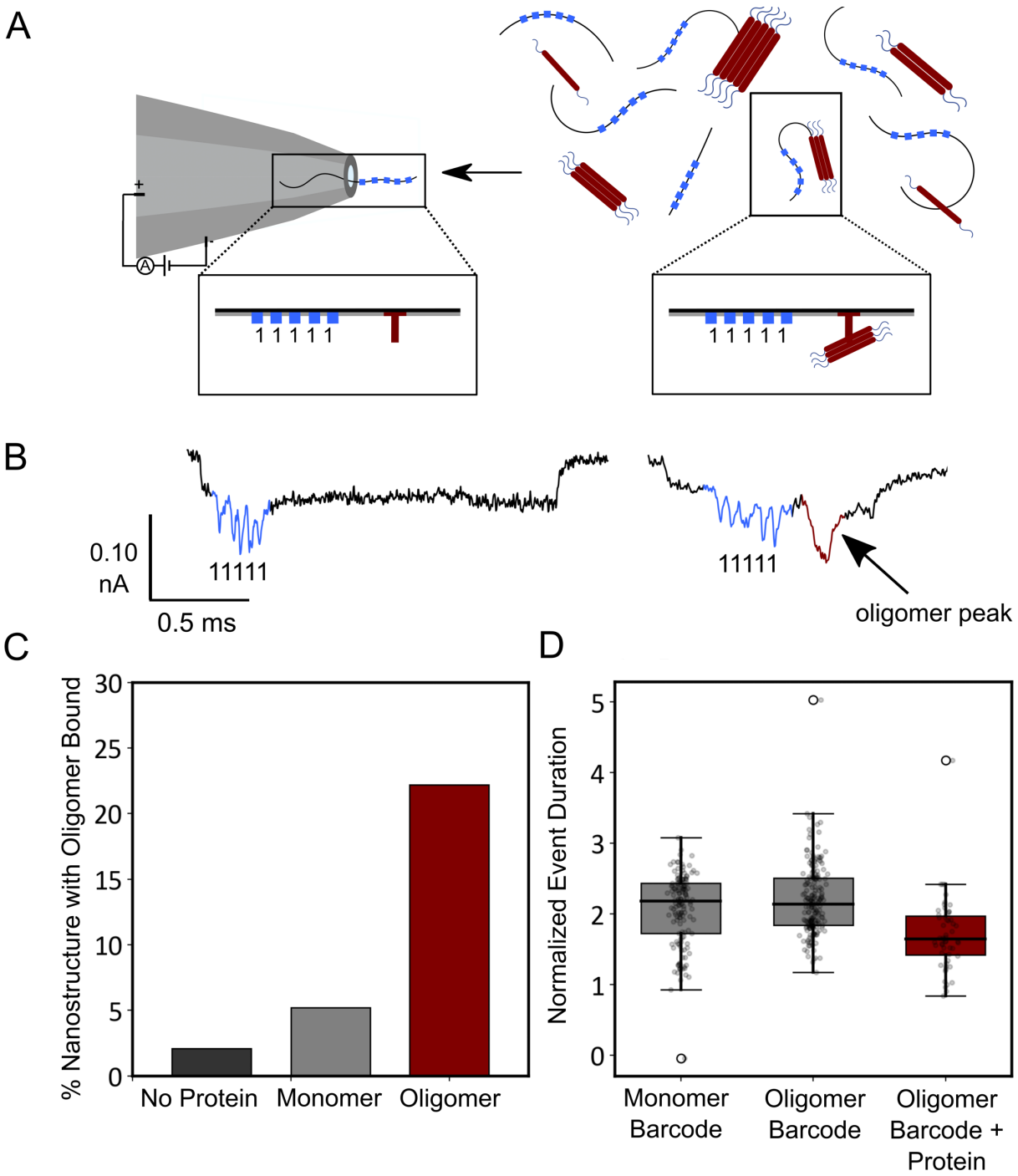


Figure 4.4. Detection of stabilised oligomers using nanopores. (A) Nanopore schematic representing the nanostructures with and without oligomers bound. **(B)** Current trace of nanopore with no protein bound (left) and with an oligomer bound (right). **(C)** Percentage of events with a spike adjacent to the bar code for control sample without protein added (N=48), monomer sample (N=154) and oligomer sample (N=248). The samples with just monomer and stabilised oligomer act as controls and show the percentage of false positives in the sample without protein. Given that is challenging to completely prevent oligomer formation in a monomeric sample the slightly higher number of peaks observed in the monomeric sample is to be expected. **(D)** Normalised event duration (normalised to pore baseline current) for samples with barcode only or with barcode and an adjacent protein spike. Nanopore experiments were carried out by Sarah E. Sandler.

Samples of the coupled DNA-protein assemblies were pushed through a nanopore using an electric current as a driving force (**Figure 4.4A**). The negatively charged nanostructure aided insertion into the pore when a current was applied. In this case, since the protein was also negatively charged at the pH used (7.4), the translocation was sped up. As the structures translocated through the nanopores, they created unique signals (**Figure 4.4B**). Monomer samples were compared against stabilised oligomer samples. Because the molecular weight of monomeric αS is ~14 kDa, and as can be seen from the low percentage of additional spikes on the nanostructure from the monomer sample in **Figure 4.4C**, we can assume it is too small to be observed via the 15 nm nanopore. In this experiment, the samples containing no protein, only monomer, or stabilised oligomers were initially tested in different pores as a control to rule out any inter-sample interactions. The lack of events observed in the monomeric sample allows us to clearly distinguish the samples with and without oligomers by their current traces, and removes the monomers as a source of additional signal as their signal is too low to be detected in a nanopore of this diameter. This demonstrates how the customisable dimensions of solid state nanopores can be utilised to focus on the subsample of interest. A significant difference in the percentage of events with proteins attached to the DNA barcodes was observed between the oligomeric and monomeric samples, demonstrating the potential utility of the approach for determining oligomer levels in a sample (**Figure 4.4C**).

It should be noted that the oligomeric samples also contained a significant proportion of monomer, which is otherwise challenging to separate entirely from the oligomer sample. Of the observed events in the oligomer sample, ~22.2% had a protein oligomer spike attached to the DNA nanostructure. The rest of the events exhibited no spike due to being bound to monomeric protein, which makes up the majority of the sample. The ability to measure with this background present is essential given the additional time cost and potential bias introduced by a need to separate oligomeric species from the bulk monomer. These events can be separated both by observing the nanopore signal generated, where little to no protein spike signifies either an uncoupled DNA nanostructure, or a nanostructure coupled to only monomer, as well as by using parameters such as event duration (**Figure 4.4D**). Because the protein is negatively charged, the event duration decreases in samples with bound proteins. These samples were measured in different pores at different times, to ensure no cross-sample contamination and reliable controls, so the duration must also be normalised to the baseline current (I_0). A normalisation was carried out as explained in the methods (**Equation 4.2**).

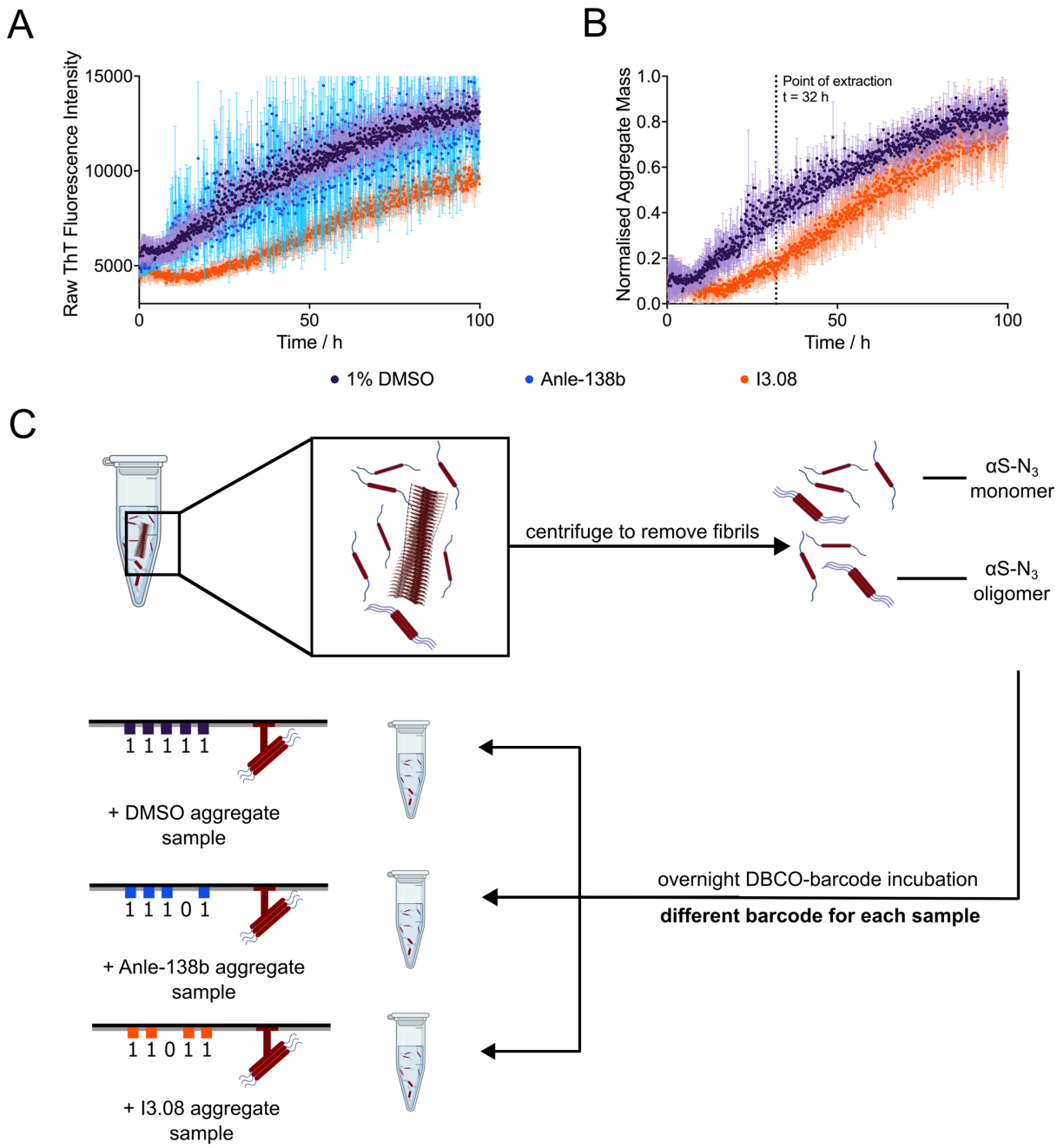


Figure 4.5. Preparation of an αS aggregation time-course in the absence and presence of inhibitor molecules, and extraction of oligomers. (A, B) Kinetic traces are shown of a 10 µM solution of azide-tagged N122C-αS supplemented with 100 nM pre-formed seeds (pH 7.4, 37 °C, shaking at 200 rpm, error bars denote SD) in the presence of 1% DMSO (purple), 25 µM Anle-138b (blue) or I3.08 (orange). The raw fluorescence (A) and normalised fluorescence (B) are shown. The endpoints were normalised to the αS monomer concentration at the end of the experiment, which was detected via the Pierce™ BCA Protein Assay at t = 100 h. The Anle-138b sample could not be suitably normalised due to the noise of the sample. **(C)** Samples were extracted at 32 h from the time course of aggregation and centrifuged to remove fibrils from the mixture, leaving only αS monomers and oligomeric species for analysis. These samples were then incubated with a unique DBCO-tagged DNA barcode overnight before analysis via solid state nanopore detection.

4.2.3. Effect of inhibitor molecules on αS oligomer production

Having optimised the conditions, I then moved on to more challenging "on time-course" samples. I carried out an aggregation beginning from monomer, under conditions designed to promote secondary nucleation[57,187]. This assay has been fully characterised for AlexaFluor-488 tagged N122C vs WT in previous works, and azide tagging did not substantially alter this behaviour[116,187,188]. Oligomer populations in this scenario are significantly lower in concentration compared to the stabilised oligomer case, and they are transient. On time-course samples of αS are stable for an unknown period, generally considered to be no more than ~36 h post extraction, compared to αS stabilised oligomers which persist for up to a week after production if left at room temperature[186,187].

The on time-course experiment was designed to better mimic the processes and species that may occur in vivo. In order to induce αS aggregation via secondary nucleation in vitro at neutral pH, a small amount of pre-formed seed was added (100 nM monomer equivalents, 1%) in the presence or absence of aggregation inhibitors of interest (**Figure 4.5A**, **B**). The aggregation process was followed using ThT fluorescence. The 3 samples of interest were a control containing only 1% DMSO, another control containing Anle-138b in 1% DMSO, and a small molecule identified previously via structure-based machine learning methods, I3.08, also in 1% DMSO. DMSO was used to dissolve the molecules before adding to the aqueous protein sample.

In **Chapter 2** I showed that I3.08 binds to the fibrils, not the monomer or oligomers, and in so doing blocks autocatalytic aggregate formation[65]. Since fibrils are removed prior to nanopore measurement by centrifugation only the oligomer and monomer population remain. The molecular mechanism of Anle-138b has not been published, and is presumably not known in detail. However, the work presented in **Chapter 1** indicates it may operate by a similar mechanism only in a milder manner. The aggregation was accelerated via shaking, which was necessary to complete the aggregation under cellular buffer conditions in an experimentally accessible time frame. As stated in **Chapter 2** this creates a more challenging paradigm for the inhibitors to function in given the increased aggregation resulting from mechanically induced fragmentation as well as secondary nucleation. Nonetheless, a significant inhibition of fibril accumulation was still observed for inhibitor I3.08, although not for the control inhibitor Anle-

138b. Samples were then taken mid-way through the time course to determine whether a reduction in oligomeric species was also observed.

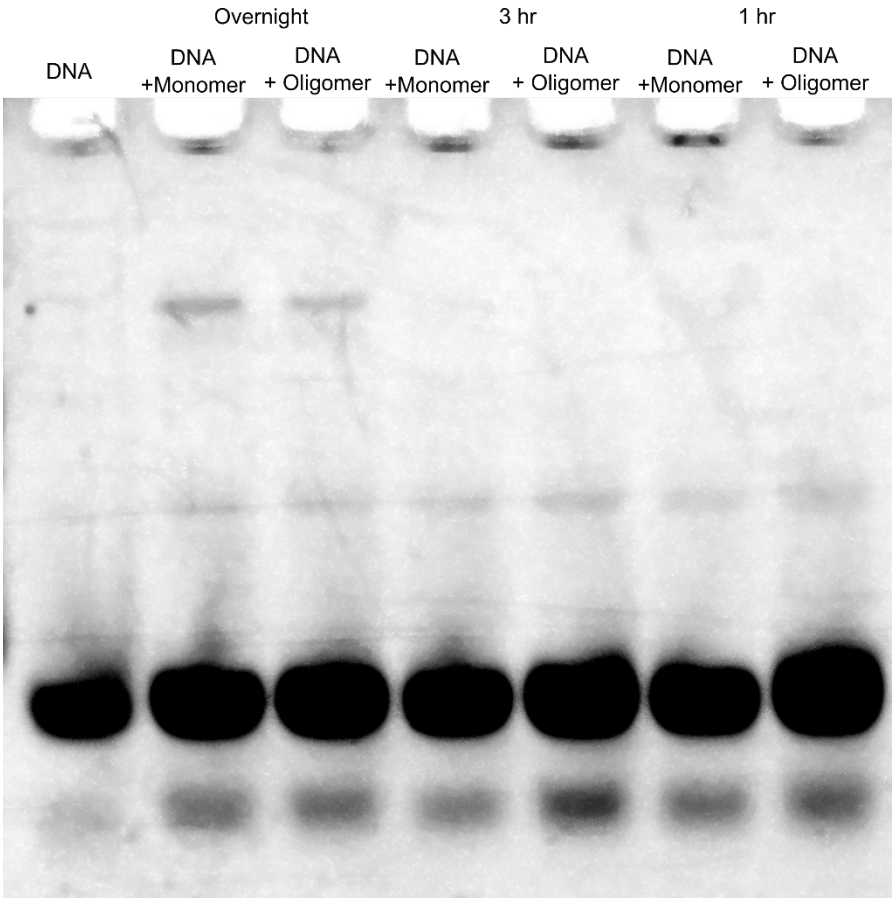


Figure 4.6. DBCO reaction over time, as observed via PAGE. In this example the 21 base pair DNA sequence alone, with DBCO attached, was reacted with the azide tagged on time course samples from an aggregation reaction (one monomeric sample and one sample extracted from the half time of aggregation, termed the oligomer sample). From the lightest (bottom) to the heaviest (top) bands there is: single stranded DNA, double stranded DNA, an impurity product of DNA synthesis, and protein tagged with DNA. The protein-DNA band is only observable for the overnight incubation. There is little observable difference in the oligomer lane vs the monomer lane for the overnight incubation, implying that only monomer is observable. This is to be expected, due to the tiny concentration of oligomer present, exacerbated by oligomer dissociation on the gel, and demonstrates the need for single molecule resolution. Gel was run by Sara Rocchetti.

Samples were extracted at 32 h into the aggregation time course and centrifuged to remove fibrils before click reaction of the azide-tagged αS with unique DBCO-tagged DNA barcodes overnight at a ratio of 1:1 (DBCO-DNA : initial monomer concentration) (**Figure 4.5C**). Each sample was labelled with a different DNA barcode; DMSO (11111), Anle-138b (11101) and I3.08 (11011). The aggregation reaction was diluted 2500-fold for this coupling, effectively quenching further aggregation. In the absence of conditions favouring phase separation[189], αS does not continue to aggregate under experimentally accessible timescales at concentrations

below 5 μM regardless of the conditions[62,97,126]. DNA-DBCO/N122C-azide reaction required at least >3 h incubation time for the reaction to proceed significantly (**Figure 4.6**). This is typical of reported strain promoted azide-alkyne click chemistry (SPAAC)[190,191]. The rate was tested by sampling 1, 3 and 12 h incubation times. No observable shift in PAGE was visible for 1 or 3 h, but an observable shift was visible for the sample incubated for 12 h overnight. These results demonstrate that we can multiplex the samples without concern for significant further coupling reactions from any residual unreacted azide/DBCO species during the nanopore measurement. Concerns over possible interchange of monomers in the sample between oligomers of different samples were addressed by the dilution at this stage, with the expectation that interactions become essentially unfeasible. Additional repeats were done using duplexed DMSO and I3.08 samples (**Figure C.4**). Similar results for samples tested in duplex and triplex support this assumption. No separation of aggregate mixtures is carried out, other than fibril removal, as this would drastically reduce throughput. Azide-tagged monomeric samples were obtained via SEC and incubated in a 1:1 ratio with DBCO tagged DNA barcodes. Oligomeric samples resulting from aggregation reactions of azide-tagged monomer were similarly incubated with a 1:1 monomer equivalent ratio of DBCO tagged DNA barcodes after fibril removal.

4.2.4. Multiplexed digital nanopore read-out of the effect of inhibitor molecules

Using the method described above, the samples were then run through the nanopore. Analysis of both the number of events containing a discernible DNA barcode and an attached protein spike, and the area of the protein spike, showed a change in oligomer distribution compared to the DMSO control (**Figure 4.7A, B, C**). The DNA barcode is the observable quantity, and so a ratio of the barcode with bound oligomer vs unbound was calculated via **Equation 4.1**. The nanopores were fabricated to be 12-15 nm, such that monomeric proteins would not be observable, while oligomeric species would be observable. The DMSO barcode was 29.8% bound to protein oligomers, the Anle-138b sample was 41.8% bound and the I3.08 sample was 14.4% bound (**Figure 4.7B**). The size distribution of the oligomers broadly matched this trend, showing decreasing oligomer mass from the DMSO sample to the Anle-138b sample, which contained a large number of small oligomers as explained below, and lastly the I3.08 sample (**Figure 4.7C**). This was calculated using **Equation 4.3**. As the samples were run simultaneously in the same pore, no normalisation was required. These results show that

compound I3.08 reduced oligomer production relative to the untreated control, and that it was a better inhibitor of oligomer production than Anle-138b.

Interestingly, first I3.08 and DMSO were tested in duplex and a similar baseline level noise (~6 pA) was maintained throughout the measurement. With the addition of Anle-138b in triplex with the other samples, the noise level increased (**Figure C.5**). This is consistent with the kinetic data (**Figure 4.5A**). The Anle-138b sample exhibits a noisy kinetic trace, consistent with increased formation of particulates, and has a correspondingly greater oligomer population. The increase in nanopore noise is most likely due to the larger oligomers present rapidly translocating through the pore at the beginning of the measurement. After 3 min, most of the larger oligomers have translocated through the pore which leads to the baseline current and noise resuming back to normal. This is also consistent with the number of events measured for Anle-138b (N=43), where fewer discernible events with Anle-138b barcode 11101, as compared to DMSO barcode 11111 (N=114) and I3.08 barcode 11011 (N=90) were observed despite all samples being added at equal concentration.

4.2.5. Comparison with a micro free flow electrophoresis (μFFE) method

For comparison, a state-of-the-art technique in protein oligomer detection is micro free flow electrophoresis (μFFE), which allows full characterisation of the oligomer distribution in physiological conditions, and was applied in **Chapter 2** to ascertain oligomer populations in the presence of a closely structurally related inhibitor to the one used here[65,116]. The μFFE requires insoluble fibrils to be removed via centrifugation, but no further separation is required, as the technique separates the monomeric fraction from the oligomeric fraction in situ using an electric field across the particle stream that deflects particles based on their electrophoretic mobility. The only disadvantage is relatively low throughput. In **Chapter 2**, molecule I3.02 induced a 37% delay in half time of aggregation compared to the negative 1% DMSO control. As a result, there was a 75% reduction in the mass of oligomers present at the half time of the negative control. The aggregation kinetics were carried out under similar conditions as used here, the primary difference in that work being the higher concentration of αS monomer and the molecule (100 μM αS, 50 μM molecule). In this work, molecule I3.08 induced a 57% delay in relative half time and, as measured by nanopore detection, the drop in oligomer events observed was 48% and the drop in oligomer mass was 22% compared to the negative DMSO

control. Anle-138b was shown to have lower effectiveness in terms of oligomer number reduction and oligomer mass reduction via both techniques, and so the ranking of effectiveness between nanopore detection and µFFE is in agreement.

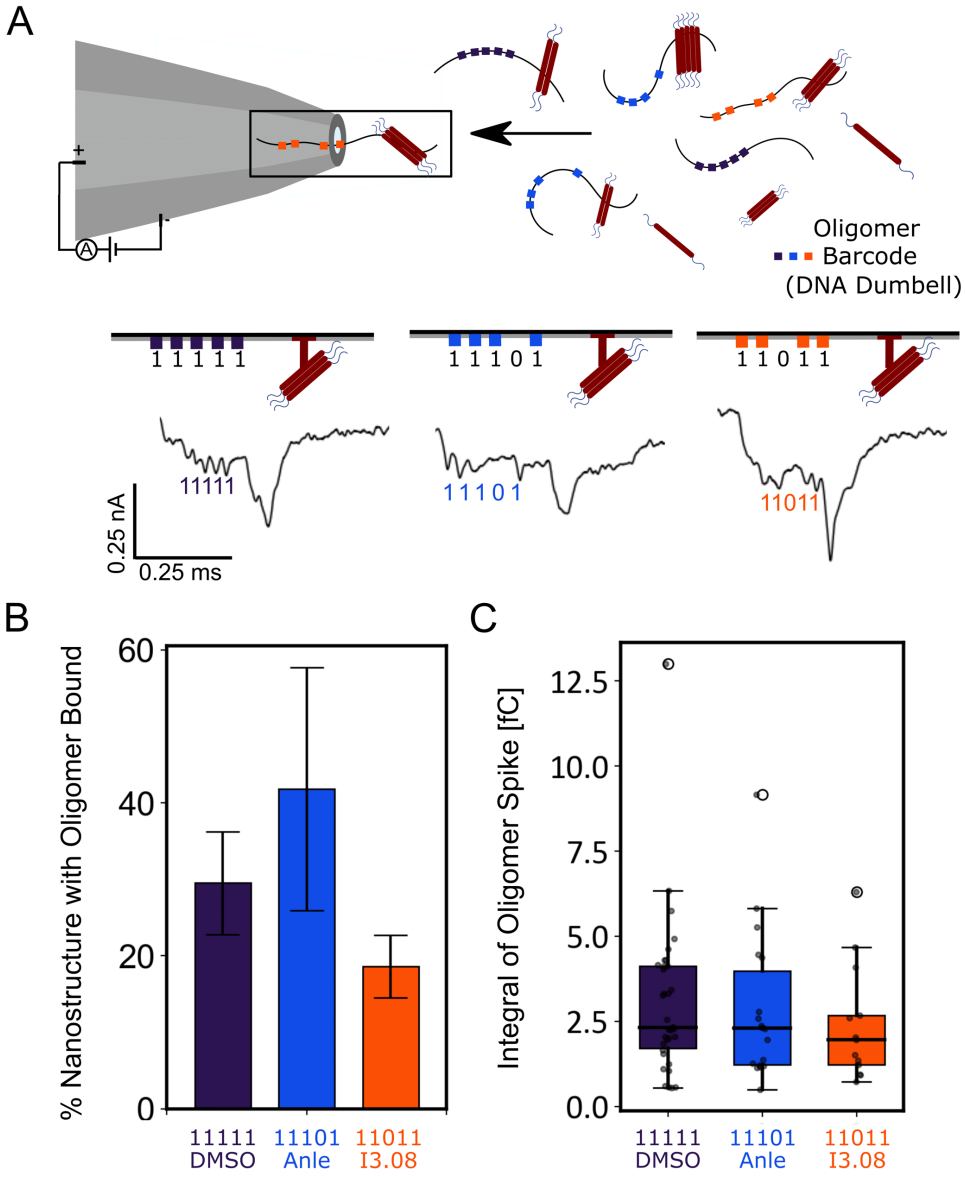


Figure 4.7. Schematic of the multiplexing pipeline and comparison of two different inhibitor molecules effects against on time-course samples. (A) Samples are tagged with a unique DNA barcode that allows identification in a multiplexed mixture, increasing the throughput. The events observed as the oligomers translocate through the nanopore can then be analysed to give an oligomer number per tag, and a relative area under the curve of each tag, proportional to oligomer size. **(B)** The fraction of events with an oligomer bound to the DNA barcode; DMSO (purple) (N=114±7), Anle-138b (blue) (N=43±16) and I3.08 (orange) (N=90±4). The standard deviation comes from repeats where the samples were combined, diluted in measurement buffer and measured for ~1 h. **(C)** Area of the current drop of the protein spike caused by bound oligomer in the DMSO (purple), Anle-138b (blue) and I3.08 (orange) samples. A larger area implies larger species are bound to the barcode on average. Nanopore experiments were carried out by Sarah E. Sandler.

The strategy here was to create a novel screening approach for aggregation inhibitors, not to fully characterise the aggregation time course, though this would represent a valid application of the technology. This has however been done multiple times previously[57,187,188] while oligomer inhibitory screening assays are scarcer, due to difficulty in applying existing methodologies with low throughput. The comparison between μFFE, one of the methods used to carry out a full time course characterisation[116,188] and then to characterise inhibitor potency[65], versus the nanopore method shown here, demonstrates that both are effective at ranking molecules in terms of inhibitory potency.

4.3. Discussion

I have reported a nanopore detection method for protein oligomer detection and analysis, with a detection limit on par with current state of the art techniques, but with significantly greater potential for throughput. To illustrate the method, I applied it to detect the inhibition of αS oligomer production by small molecules in clinical and academic development. This result was obtained with the additional benefit of multiplex capability and higher throughput.

While the nanopore system has many advantages, there are also some drawbacks. A drawback of the large dilution step required for measurement in the nanopore is the possibility that some of the oligomers may dissociate during the DBCO coupling step (12 h) due to the large dilution (2500-fold). This is a feature of most single-molecule techniques which require low concentrations in order to have clear signal to noise ratio. However, αS is a useful test case in this scenario given that its kinetics are relatively slow and its oligomers are stable[116,192] over the time scales investigated, so the measured sample is a reasonable reflection of the population present at the extraction stage. In further developments, a cross-linking step could be introduced to ensure that the protein sample extracted exactly matches the one measured. This carries the risk of cross-linking separate oligomers (potentially mitigated by appropriate dilution) and adds further processing steps, issues which I sought to avoid in the interests of throughput and preventing biasing of the oligomer population. Alternatively, if the dissociation rate in a particular case was a cause for concern, a more reactive click pair could be employed than the one used here, or the coupling could be carried out at higher concentration (followed by dilution immediately prior to measurement) to obtain coupling over a shorter time scale and

slow dissociation. A restraint on the click coupling reaction is that the sample conditions cannot be altered in terms of pH or temperature, as this would affect the oligomer distribution.

An additional concern with the nanopore measurement is the high salt concentration required for measurement, which may perturb the aggregate distribution. However, the click chemistry reaction was performed in PBS, and the samples were only mixed in the detection buffer directly before measurement. The ratio of protein bound to unbound DNA nanostructures also did not change over the time of the observation (**Figure C.6**), suggesting this is not a major issue. Again, cross-linking could remove this problem if necessary. In the interests of throughput, however, and for cases where there is a clinical trial benchmark, all that would be required is a relative measurement to compare the effect of different inhibitors. As the samples are measured under the same conditions, a ranking of effectiveness can still be obtained. For protein systems that aggregate very rapidly, the concern is more that the monomers and oligomers may further aggregate during the click reaction rather than dissociate. I anticipate that for almost all proteins the significant dilution should quench aggregation to a rate that is negligible over the time span of the coupling reaction. Finally, using nanopores as a tool to measure oligomers does have a fundamental size limit, in that particles larger than the diameter of the pore and smaller than the resolution limit will not be detected. However, with a degree of prior knowledge, the nanopore diameter can be appropriately tailored to the size distribution of interest, allowing sampling of a representative portion of the population.

The results that I have presented illustrate an approach for investigating protein assemblies that are both transient and at very low concentration. I have applied this method to the scenario of early drug discovery for Parkinson's disease and synucleinopathies in general, where αS oligomers are considered to be key to pathology. I also show comparable performance to existing single-molecule techniques, but with greater potential for throughput due to the ability to multiplex and upscale. With the introduction of artificial amino acids bearing azides into in vivo models of disease[193], this also represents a potential approach for directly quantifying oligomer populations in such models, utilising the biorthogonality of the click reaction employed here. I anticipate that this approach could be of significant benefit to researchers working in the field of protein misfolding diseases and protein multimerisation, and in early-stage drug discovery research in general.

4.4. Materials and methods

4.4.1. Compounds and chemicals

Compounds were purchased from MolPort (Riga, Latvia) or Mcule (Budapest, Hungary) and prepared in DMSO to a stock of 5 mM. All chemicals used were purchased at the highest purity available.

4.4.2. Recombinant αS expression

See **Section 2.5.2**. The cysteine-containing variant (N122C) of αS was purified by the same protocol, with the addition of 3 mM DTT to all buffers.

4.4.3. Azide labelling of αS

αS N122C protein was azide-labelled to enable click coupling to DNA tags. N122C (200 μM, PBS, pH 7.4) was incubated with TCEP-HCl (5 eq) for 1 h at RT. The reduced N122C was then desalted with a 5 mL HiTrap desalting column, (Cytiva, 29-0486-84), and eluted in PBS, pH 7.4, 10 mM EDTA and kept on ice. The extend of the reduction was then established via Ellman's method, and a sample was taken for LCMS analysis. The protein was then incubated with iodoacetamide-PEG3-azide (10 eq) for 3 h at RT, and samples were taken subjected to QTOF MS/MS analysis with a VION mass spectrometer to ascertain the progress of the reaction (**Figure C.1**). Deconvolution was conducted in UNIFI software. Upon reaction completion the reaction mixture was separated on a Superdex 75 10/300 GL column (GE Healthcare) at a flow rate of 0.5 mL/min and eluted in PBS buffer to isolate the monomeric fraction and buffer exchange into PBS. The protein concentration was determined spectrophotometrically using $\varepsilon 280 = 5600\ M^{-1}\ cm^{-1}$.

4.4.4. αS seed fibril preparation

See **Section 2.5.4**.

4.4.5. αS stabilised oligomer preparation and subsequent click coupling

αS stabilised oligomers were produced as described previously[186]. Monomeric αS was dialysed into distilled water overnight at 4 °C, using 3.5 kDa MWCO dialysis membranes. 6 mg of the dialyzed protein was aliquoted into 15 mL tubes, flash frozen in liquid nitrogen, and lyophilised for ca. 48 h at room temperature. To prepare the oligomeric samples, the 6 mg of protein was resuspended in a total of 500 µL PBS to obtain a final protein concentration of ca. 800 µM. The solution was centrifuged if necessary (1 min, 1000 g) to get rid of bubbles formed during the resuspension process. The protein solution was filtered through a 0.22 µm syringe filter and incubated in 1.5 mL tubes at 37 °C for 20–24 h under quiescent conditions. The resultant protein solution was ultracentrifuged (1 h, 288,000 g) to remove any fibrillar species that may have formed during the incubation period, and the supernatant was removed and retained. Each aliquot of supernatant was passed through four 0.5 mL 100 kDa centrifugation filters sequentially (2 min, 9300 g), in order to remove excess monomeric protein as well as the low levels of very small oligomers. To estimate the total mass concentration of the final oligomeric solution (i.e., total concentration in monomer equivalents), the absorbance was measured at 275 nm, using a molar extinction coefficient of 5600 M^{-1} cm^{-1}. This preparation results in an overall oligomeric yield of ca. 1%. Samples were then diluted to a final concentration of 88 nM monomer equivalents in PBS and incubated overnight with a final concentration 4 nM of DBCO tagged DNA nanostructure. The reason this excess was used was to attempt to ensure 1 DBCO tag per oligomer and prevent over tagging (each stabilised oligomer has a reported average monomer count of 22[186]). Subsequent on time course experiments were carried out with 1:1 labelling of the DBCO:monomer given the large excess of monomer:oligomer expected in these samples.

4.4.6. DBCO DNA nanostructures

DNA constructs with different barcoded regions plus a DBCO labelled overhang sequence were created. Each DNA construct was synthesised from pairing a linearised 7.2 kbp single-stranded (ss) M13mp18 DNA with 40 nucleotide staples complementary to the backbone in order to create a full linearised dsDNA. The backbone and staples are annealed for 45 minutes in a thermocycler. Using a 100 kDa Amicon filter, the sample is then filtered and stored in 10 mM Tris 0.5 mM $MgCl_2$ pH 8. The concentration is then measured in a nanodrop

spectrophotometer with typical yield ranging from 75-95%. The barcoded region design follows a previous work with dumbbells optimised for read out in 15 nm nanopores[182]. Each '1' bit is made of eleven simple dumbbell hairpin motifs to create the structural spikes which act as a barcode on the DNA nanostructure. This can be optimised to have fewer dumbbells per spike if needed. The exact sequences with their numbers are shown in **Table C.2** in the Supplementary Information following a previous work[182]. The overhang was created by replacing oligo No. 142 with 61 bp segment containing 40 bp to match the scaffold and a 21 bp oligo complimentary to another DNA sequence containing a DBCO label. The 21 bp dsDNA overhang is not large enough to generate a current blockade (an observable signal in the nanopore) which has been confirmed by observation. These sequences can be found in the Supplementary Information **Table C.3**.

4.4.7. Aggregation kinetics and subsequent click coupling

Azide-labelled αS N122C (10 μM) was supplemented with seed (100 nM) under shaking (200 rpm) at 37 °C, PBS pH 7.4 and either 1% DMSO or 25 μM molecule in 1% DMSO. Samples were extracted at the $t_{1/2}$ of the DMSO sample (30 hours). Fibrils were removed by centrifugation (21,130 rcf, 10 min, 25 °C). Samples were then diluted to 4 nM monomer equivalents in PBS and incubated overnight with 1 eq (relative to initial monomer concentration) of DBCO tagged DNA nanostructure.

4.4.8. Nanopore fabrication and measurement

The nanopores are made of commercially available quartz capillaries (0.2 mm ID/0.5 mm OD Sutter Instruments, CA, USA). A laser-assisted pipette puller (P-2000, Sutter Instrument, CA, USA) is used to create nanopores with diameters of 10-15 nm. 16 conical nanopores are then placed in a custom templated PDMS chip containing a communal cis reservoir and individual trans reservoirs. In order to generate the current, silver/silver-chloride (Ag/AgCl) electrodes are connected to the cis and trans reservoirs in the PDMS chip. In the baseline buffer solution for the stabilised oligomers (4 M LiCl, 1X TE, pH 8.0) and the on pathway samples (2 M LiCl, 1X TE, pH 8.0), a current-voltage curve is taken in order to estimate the nanopore size. Only one nanopore is measured at a time due to the electronics, thus the trans reservoir contains the electrode with a 500 mV bias voltage and the central cis reservoir which contains the sample

is grounded. The measurement is then run for 1-2 hours until 1500-3000 events are gathered. Typically, of these events, 30% are unfolded and are then analysed.

Current signals are collected using an Axopatch 200B patch-clamp amplifier (Molecular Devices, CA, USA). The set-up is operated in whole-cell mode with the internal filter set to 100 kHz. An 8-pole analogue low-pass Bessel filter (900CT, Frequency Devices, IL, USA) with a cutoff frequency of 50 kHz is used to reduce noise. The applied voltage is controlled through an I/O analogue-to-digital converter (DAQ-cards, PCIe-6251, National Instruments, TX, US). A LabView program records the current signal at a bandwidth of 1 MHz.

4.4.9. Nanopore data analysis

The experimental data files are stored as technical data management streaming (TDMS) files from the Labview program recording the raw traces. First, a translocation finder python script is used which identifies the events from the raw traces using user-defined thresholds (minimum 0.3 ms duration, minimum 0.1 nA current drop) and stores them in an hdf5 file. This can be found at https://gitlab.com/keyserlab/nanopyre. Next, the hdf5 file is loaded into the GUI categoriser python script, found here: https://gitlab.com/keyserlab/nanopycker. Using this, the events are printed and the user can manually sort the events time efficiently into different categories and later print events from the hdf5 file that are assigned to a specific category. In this case the categories were barcode without protein and barcode with protein. The percentage of events with oligomer bound is then calculated using

$$\%\ Oligomer\ Bound\ Events_{\ x} = \frac{N_{x\ protein}}{N_{x\ \ protein} + N_{x\ no\ protein}} \times 100 \tag{4.1}$$

Where x is the barcode. This is used in **Figure 4.4C** and **Figure 4.7B**. The duration of the events in **Figure 4.4D** is calculated using

$$Normalized\ Duration = \frac{\left(\frac{x_{right} - x_{left}}{sample\ frequency\ [Hz]}\right)}{I_0} \tag{4.2}$$

Where x_{right} is the position of the end of the event and x_{left} is the position of the end of the event. The sampling frequency is 1,000,000 Hz. I_0 is the baseline current because different pores were used for the different measurements with different baselines.

The GUI categoriser is used again on the events with protein to calculate the ECD of the protein spike using

$$g(x) = \frac{f(b) - f(a)}{b - a} \times x + f(a) - a \times \frac{f(b) - f(a)}{b - a}$$

$$Area = \sum_{a}^{b-1} \frac{1}{2} \times (x_{n+1} - x_n) \times \left[\left(g(x_{n+1}) - f(x_{n+1})\right) + \left(g(x_n) - f(x_n)\right)\right] \qquad (4.3)$$

Where $f(x)$ is the current at point x, a and b are the left and right bounds of the region of interest and $g(x)$ is the equation of the line connecting a and b.

4.4.10. Mass spectrometry

10 μM of preformed αS was incubated with 25 μM of molecule in 20 mM sodium phosphate buffer (pH 4.8) supplemented with 1 mM EDTA overnight under quiescent conditions at room temperature. The supernatant was removed for analysis using a Waters Xevo G2-S QTOF spectrometer (Waters Corporation, MA, USA).

4.5. Contributions

This chapter is substantially derived from the preprint doi: 10.1101/2023.08.09.552642 (accepted at *J. Am. Chem. Soc.*). Sarah E. Sandler (S. E. S.) and I conceived the project and wrote the article. I performed azide labelling, analysis and aggregation kinetics. S. E. S. and I performed azide-DBCO coupling reactions, and S. E. S. performed nanopore experiments. Sara Rocchetti performed gel analysis. Rebecca C. Gregory produced the N122C αS.

5. Generalising to other misfolded proteins

"There isn't one kind of dementia. There aren't a dozen kinds. There are hundreds of thousands. Each person who lives with one of these diseases will be affected in uniquely destructive ways."
– Sir Terry Pratchett

5.1. New targets: Aß-42, IAPP and tau

Having spent the vast majority of my PhD focussed on αS I hoped to demonstrate the generalisability of the developed pipelines and apply them to other misfolding protein related conditions. This work is still ongoing and will hopefully continue with other PhDs. A cross section of these projects is described in this chapter, which have reached varying degrees of progression.

5.2. Targeting secondary nucleation in Aß-42 and IAPP

5.2.1. Aß42

Amyloid ß **(**Aß) was initially the aggregating peptide implicated as the most damaging agent in Alzheimer's disease (AD) due to genetic evidence and observation of amyloid plaques in patient brains, which form in the extracellular space between neurons. These plaques consist largely of Aß peptides 40-42 amino acids in length. These truncated peptides are derived from

amyloid precursor protein (APP), a transmembrane protein of unknown function with a single membrane-spanning domain. 2 secretases, ß and γ, are responsible for cleaving the extracellular domain of APP to create the N terminal and C terminal of Aß peptides. Mutations accelerating Alzheimer's are found to increase expression of APP and so Aß production, increase the aggregation propensity of Aß, or increase the ratio of the more aggregation prone Aß42 peptide to the less aggregation prone Aß40 peptide[44]. Pathology is however now thought to be better correlated with tau aggregate distribution, in the 'trigger and bullet' hypothesis[194]. Tau undergoes aggregation to form neurofibrillary tangles within neurons, which are another hallmark of AD. These aggregates are formed of a mix of 3R and 4R tau[45]. Aß42 aggregation may trigger tau aggregation before being enhanced itself by tau aggregation in a positive feedback loop. By this logic targeting Aß42 too late in the disease may prove futile once tau aggregation becomes dominant. Removing Aß42 plaques has however shown benefit to patients as previously described. As such it seems that some of the burden on neurons in AD can be alleviated by removal of misfolded Aß42 aggregates, but that tau targeting drugs may be more effective.

Efforts were made to carry out a similar docking approach as previously described against Aß42 fibril structure 2MXU[195]. A selection of molecules were tested against 2 different sites, only 1 of which produced an effective inhibitor. This hit was obtained after screening ~120 molecules for the unsuccessful first site and ~50 molecules for the second site. Aß42 is a more aggressive aggregator than αS, capable of auto aggregating via primary nucleation in the absence of seed even at high pH where the protein is charged and so monomers should repel one another (pI = 5.31)[196]. Preventing its aggregation is therefore more challenging. I found that the inhibitor in question had few available derivatives from vendors and thus could not be easily elaborated via iterative learning in an academic context. The docking is limited here by the amount of chemical matter that can be docked (~2 million compounds). If a larger cross section of chemical space could be screened it would be possible to obtain more promising starting points.

This is something that is currently being addressed with methods such as deep docking, where the underlying calculations of the docking simulation are learned by a neural network[197]. Deep docking greatly increases the rate of the docking process and so increases the number of molecules that can be screened from 2 million to several hundred million in a similar time

frame. The PhD project of Michaela Brezinova has been focussed on optimising this computational approach. This promises to address the weak point of the pipeline established here, which is whether the initial docking produces good leads. If the correct pocket is selected, then there is a higher likelihood of obtaining a larger number of varied hits. If the wrong pocket has been selected this should become obvious more rapidly than previously, as it is easier to distinguish a successful deep docking from a failed one due to the greater discrepancy in hit rates. Initial experimental work suggests that this has succeeded, but further validation is required.

5.2.2. IAPP

Islet amyloid precursor protein (IAPP) is similarly implicated in diabetes mellitus[198,199]. Misfolded aggregates attack cell membranes, destroying the islet cells that produce insulin. How this event occurs initially is unknown but targeting it may alleviate symptoms. IAPP is rather similar to Aß42 in that it can auto-aggregate without an inducer, and also aggregates via secondary nucleation[200]. An initial effort to carry out docking against IAPP fibril structure 6Y1A[201] was also carried out, but again I could obtain relatively few molecules (~25), and only 1 showed efficacy. Again, this molecule proved challenging to explore around with iterative learning as derivatives were not available. This will hopefully also be addressed via the deep docking approach outlined above.

5.2.3. Summary

The effectiveness of this technique relies on knowledge of which binding pockets are most likely to yield effective inhibitors and, for use in academic contexts, good availability of derivatives. These factors held the approach back when targeting Aß42 and IAPP, though it may be possible to overcome these issues using new computational techniques such as deep docking. These factors were not as much of an issue for subsequent investigations into tau in the following sections, where the strategy could be employed in a similar way as implemented for αS.

5.3. Targeting pathogenic tau protein aggregation

Tau is perhaps a better target than Aß when seeking to combat AD, but its great heterogeneity in terms of isoforms and aggregate structures makes it a challenging target. A fundamental property of amyloid aggregates is their ability to promote the formation of new aggregates[164,202]. This autocatalytic process may contribute to the proliferation and spreading of the aggregates across the brain in the spatio-temporal patterns first described by Braak for AD[194,203,204]. Because this process is dependent on the presence of amyloid fibrils, their structures are likely to determine its efficiency[33]. This situation creates a significant challenge for drug discovery since, at the time of this work, over 20 different polymorph structures have been solved for tau fibrils from brain extracts by cryo-electron microscopy (cryo-EM)[205,206]. It is therefore important to develop in vitro assays of tau aggregation that generate disease-specific fibril polymorphs. This is a difficult problem because, depending on the conditions (pH, salt, temperature, cofactors) and the size and sequence of the protein isoform, in vitro studies might lead to different fibril structures from the polymorphs found in brain extracts and thus may not be disease relevant[206,207].

To reproduce in vitro the structures found in vivo, one can use brain-derived seeds to prompt recombinant monomeric tau to adopt disease-specific polymorphs[205-207]. This type of approach relies on the fact that the free energy barriers for growth are typically lower than those for nucleation, and it is therefore possible to propagate a conformer through seeding even under conditions where another more stable conformer may arise from primary nucleation. This was recently attempted to obtain αS fibrils with the morphology observed in MSA[208]. However, although individual filament halves of a mature fibril were faithfully propagated, the corresponding counter-filament in some cases adopted a novel structure not observed in reconstructions of MSA fibrils from patient brain extracts by cryo-EM[209].

An alternative method is to identify solution conditions for in vitro aggregation assays that lead to the recreation of the amyloid fibril polymorphs observed in disease. Significant progress has been recently made with tau under shaking conditions[206]. For the application of this approach to drug discovery, however, the in vitro assay should ideally also be consistent with the conditions in vivo, in order to identify candidate inhibitors with a mechanism of action expected to be clinically relevant.

Here I report a framework for addressing this problem. The three main components of this framework are: (1) a faithful propagation of strain-specific fibril polymorphs, as verified by

cryo-EM, (2) a careful kinetic analysis in quiescent reaction conditions whereby individual microscopic processes might be specifically targeted pharmacologically, using the seeds from (1) to induce aggregation and (3) the computational pipeline that has been described in **Chapter 2** to first identify and then elaborate hits. The first component is based on the brain-derived assays described above, while the second component is based on a chemical kinetics approach to study filament assembly[82]. The experimental assays for components (1) and (2) were established by Michael A. Metrick II and Alessia Santambrogio. Using this chemical kinetics formalism, several microscopic mechanisms were identified that underlie the propagation of amyloid fibrils in these conditions including secondary processes and elongation[210,211]. Preliminary data suggests that in (2) a similar fold is obtained when compared with the 503L polymorph that is dominant in AD patient brains (appendix **Figure D.1A**)[207]. The polymorph formed in (2), shown in **Figure D.1B**, has also been shown to be on the kinetic pathway to the final 503L polymorph by recently published work[212].

By using this approach, we describe the propagation of tau aggregates derived from AD brains. We then identify tau aggregation inhibitors targeted against aggregation via secondary processes, through in silico docking to binding sites on the surface of the cryo-EM structure of tau fibrils adopting an AD fold. These molecules are subsequently enhanced for tau anti-aggregation activity through the application of the iterative machine learning pipeline[65].

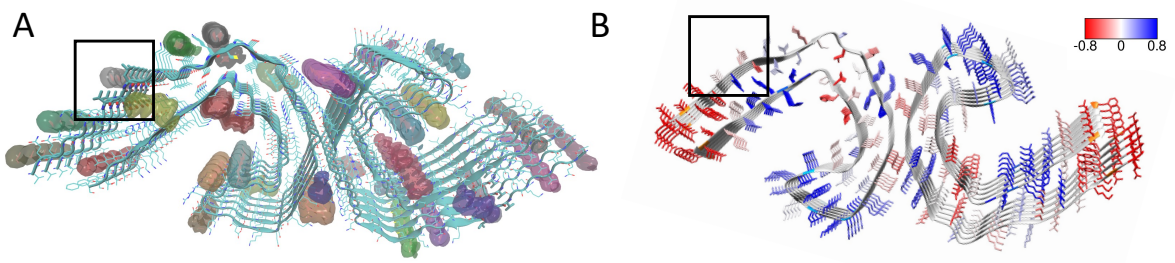


Figure 5.1. Volume and solubility based binding site prediction. (A) Cavity based binding site prediction based on Fpocket[213]. **(B)** Solubility based binding site prediction based on CamSol[96]. The black box outlines residues 313-315 where both solubility is low and cavity propensity is high. Docking simulations and initial pocket searching were carried out by Z. Faidon Brotzakis.

5.4. Results

5.4.1. Identification of a potential catalytic pocket on the tau fibril surface

Having developed a working model of the AD-derived aggregation cascade, we investigated the effects of small molecule inhibitors on the microscopic processes of the cascade. Aggregation kinetic experiments were conducted at 5 μM K12 (a C-terminally truncated recombinant 3-repeat fragment of tau substrate, with Cys mutated to Ser[114]) monomer concentration in the presence of 50 nM (monomer equivalents) first generation fibrils of K12 produced with brain-derived seeds, and with addition of 20 μM compound. A selection of small molecules was identified via docking simulations to tau fibrils, followed by experimental testing in the chemical kinetics assay described above (2) to identify potential inhibitors of aggregation. The hypothesis is again that molecules that can bind to the fibril structures could modulate secondary aggregation processes, involving formation of new misfolded aggregates from existing ones[64]. Potential binding sites on the 5O3L fibril cryo-EM structure[207] were again identified via the Fpocket and CamSol methods (see methods **Section 5.5.4** and **Figure 5.1**). The predicted binding affinity of a subset of the ZINC database passing CNS MPO criteria[94,95] to the chosen binding site was calculated using AutoDock Vina and the FRED (OpenEye Scientific Software), and the best predicted binders were obtained for experimental testing. A schematic of this approach is shown in **Figure 5.2A**.

Whether the targeted binding site is indeed the binding site of the molecules has yet to be experimentally validated. This area was targeted in part due to its reported involvement in triggering further amyloidogenesis[214], as well as identification via Fpocket and CamSol, yet the area appears comparatively featureless. There was less prior knowledge to work with here than for the αS project, and Fpocket provided a larger number of potential pockets to select from (**Figure 5.1** vs **Figure 2.1**). The pocket identification software is robust but not infallible and would benefit from a similar update as has been implemented for the docking via deep learning. That the docking hit rate was higher than expected does suggest that the docking site was chosen effectively, but this requires validation.

5.4.2. Experimental screening and initial ML optimisation

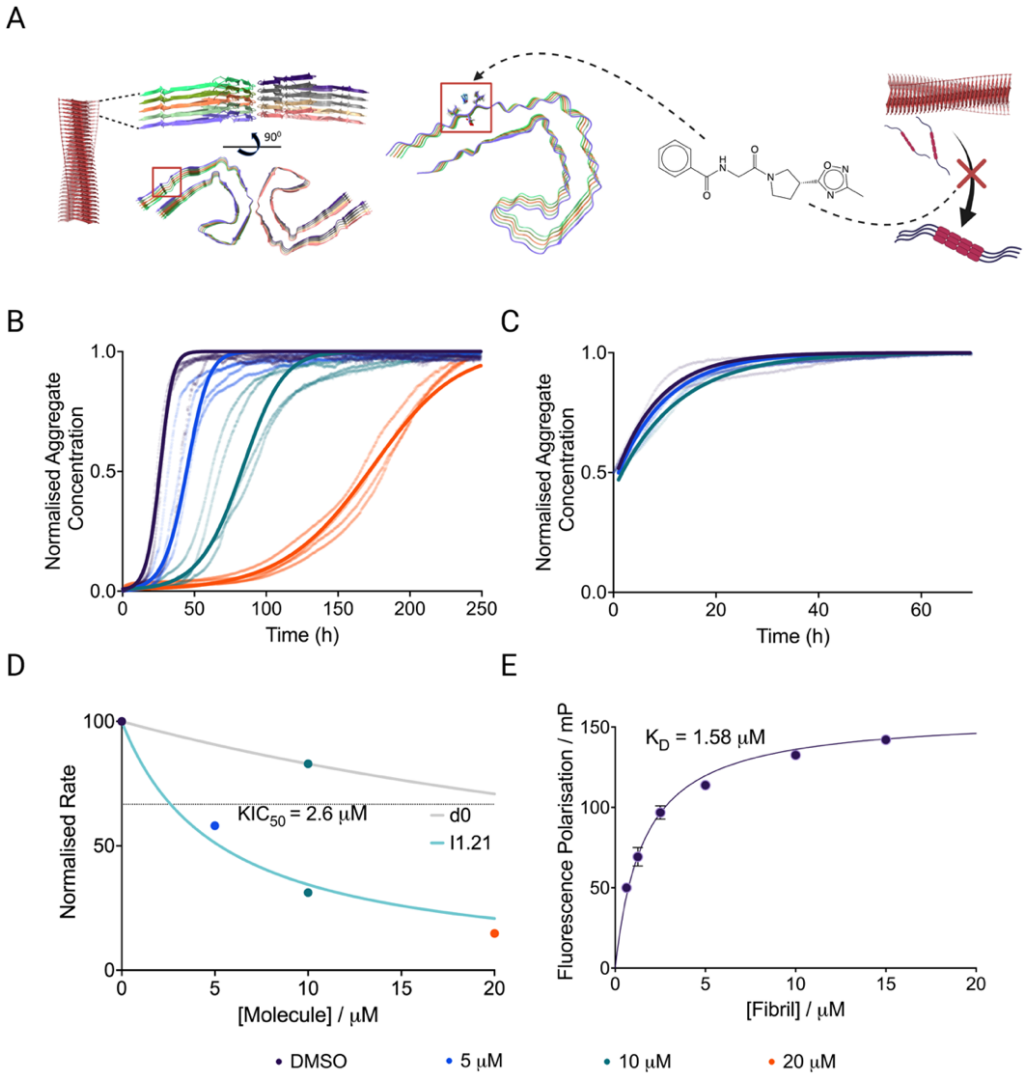


Figure 5.2. Kinetic analysis of a tau aggregation inhibitor (I.21) identified through iterative machine learning. (A) Schematic of the proposed binding mechanism, showing a K12 fibril cross section and the targeted binding site. The binding pose of I1.21 in this pocket as predicted by AutoDock Vina is also shown. This binding is thought to be responsible for the modulation of secondary pathways. **(B)** Kinetic traces for 50 nM seed, 5 μM monomer K12 aggregation reactions in the presence of 1% DMSO (purple points), and I1.21 identified via machine learning at different concentrations (coloured points). Fits of a multistep secondary nucleation model are also shown (solid lines), which describe the aggregation behaviour reasonably well. A fragmentation model also provides a similar degree of fitting. The elongation rate (k_+) for these models was derived from **(C)** showing kinetic traces for 2.5 μM seed, 5 μM monomer AD aggregate K12 reactions in the presence of 1% DMSO (purple points), and I1.21 (coloured points). The presence of molecule was not found to significantly alter elongation kinetics, so can be assumed to be mostly specific for inhibition of secondary processes. **(D)** Approximate rate of reaction (taken as $1/t_{1/2}$, normalised between 0 and 100) in the presence of 2 different molecules, the original docking hit d0 (grey), and I1.21 derived from it (light blue). The KIC_{50} of I1.21 (2.6 μM) is indicated by the intersection of the fit and the horizontal dotted line. **(E)** Change in fluorescence polarisation (in mP units) of 10 μM I1.21 with increasing concentrations of K12 fibrils (concentrations given in monomer equivalents). Error bars indicate the SD. The solid line is a fit to the points using a one-step binding curve, estimating a K_D of 1.58 ± 0.15 μM (SD) for I1.21. Kinetic aggregation experiments were run by Alessia Santambrogio.

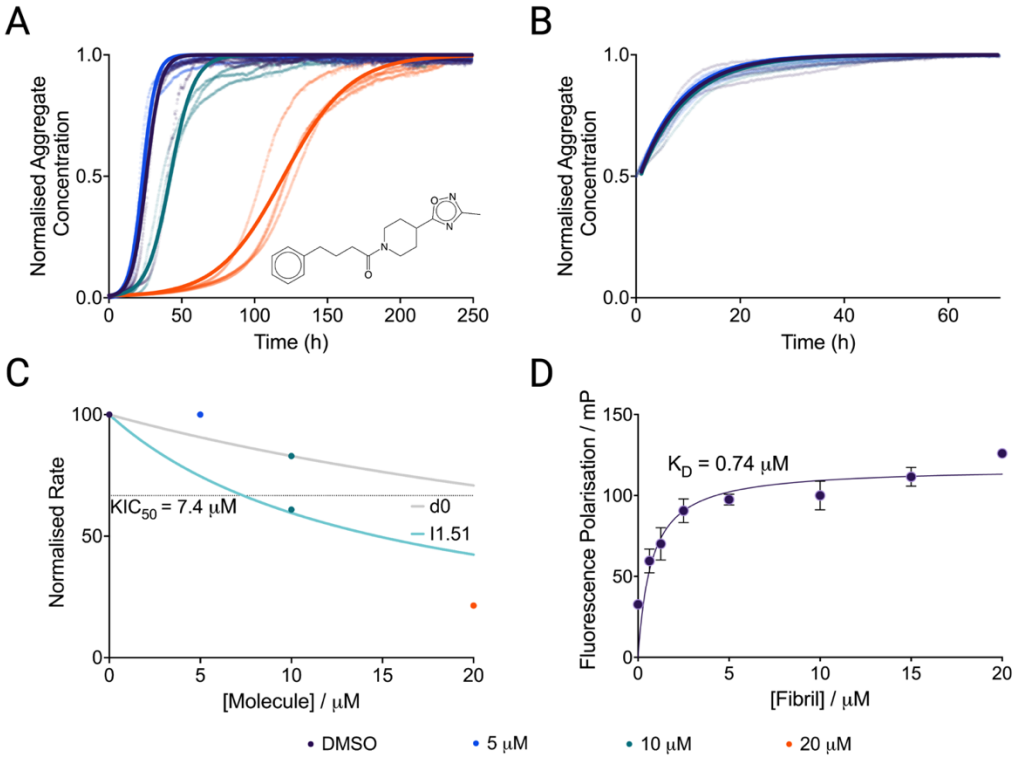


Figure 5.3. Kinetic analysis of a tau aggregation inhibitor (I.51) identified through iterative machine learning. (A) Kinetic traces for 50 nM seed, 5 μM monomer AD aggregate K12 reactions in the presence of 1% DMSO (purple points), and I1.51 identified via ML at different concentrations (coloured points). The molecular structure of I1.51 is shown. Fits of a multistep secondary nucleation model are also shown (solid lines). The elongation rate (k_+) for these models was derived from **(B)** showing kinetic traces for 2.5 μM seed, 5 μM monomer K12 aggregation reactions in the presence of 1% DMSO (purple points), and I1.51 (coloured points). The presence of molecule was not found to significantly alter elongation kinetics, so can be assumed to be specific for secondary processes. **(C)** Approximate rate of reaction (taken as $1/t_{1/2}$, normalised between 0 and 100) in the presence of 2 different molecules, the original docking hit d0 (grey), and I1.51 derived from it (light blue). The KIC_{50} of I1.51 (7.4 μM) is indicated by the intersection of the fit and the horizontal dotted line. **(D)** Change in fluorescence polarisation (in mP units) of 10 μM I1.51 with increasing concentrations of K12 fibrils (concentrations given in monomer equivalents). Error bars indicate the SD. The solid line is a fit to the points using a one-step binding curve, estimating a K_D of 0.74 μM ± 281 nM (SD) for I1.51 Kinetic aggregation experiments were run by Alessia Santambrogio.

The metric of potency in the aggregation assays was the normalised half time ($t_{1/2}$). A hit was defined as any molecule yielding a normalised half time 1.5 times greater or more than the negative control. The same threshold was used for subsequent potent lead generation. Of 102 predicted binders tested 10 were hits (9.8% hit rate). Two of these were taken forward for further elaboration (labelled d0 and d1) via a machine learning pipeline[65]. Of 32 molecules tested in a first iteration of the pipeline 6 molecules were leads (18.8% optimisation rate),

labelled I1.21, I1.51, I1.114, I1.115, I1.116 and I1.121. The labelling system for the molecules obtained via ML is the same as that used in prior chapters.

In this case the first iteration was carried out alongside a parallel iteration of tau aggregation inhibitors directed at a different aggregation assay, established by the Linse group and carried out by Dillon Rinauro, which did not utilise diseased brain derived seeds and will be mentioned briefly here[211]. More emphasis is placed on the first introduced project, given the importance of working with polymorphs relevant to disease, and all the figures in this chapter relate to the pathological fibril project. Nonetheless the availability of more closely related hit derivatives for the latter project was significantly greater. As a result, optimisation rates were higher, reaching 54% at the first iteration (33% with a relative half time cut off >2) and 62% at the second iteration (44% with a relative half time cut off >2). This demonstrates how the success of this approach in the absence of custom synthesis can be shaped by derivative availability. Interestingly, in this scenario the best leads against the former assay differed from the best leads against the pathological fibril assay, suggesting a degree of specificity for the molecules against the 2 different tau polymorphs present in these assays.

Returning to the diseased brain derived fibril project, further experiments were carried out to validate the mechanism of inhibition as had been done for αS in **Chapter 2**, albeit at an earlier phase of the iterative cycles. By modifying the input seed:monomer concentration ratio, experiments could be tailored to observe perturbations in secondary processes and heterologous nucleation (low seed experiments, **Figure 5.2B**) or elongation rates (high seed experiments, **Figure 5.2C**). Three compounds were chosen for detailed analysis, I1.21 (**Figure 5.2**), I1.51 (**Figure 5.3**) and I1.114 (**Figure 5.4**). All 3 exhibited significant dose-dependent inhibition of aggregation in the low seeded experiments, while exhibiting relatively little effect on elongation. This behaviour implies a degree of specificity for the inhibition of secondary aggregation processes. Given that secondary processes are considered to be the dominant mechanism in the production of oligomers associated with disease pathology, such behaviour would be desirable in potential therapeutics. All 3 molecules also exhibited significant improvements in potency over their parent compound as shown by the rate plots and corresponding KIC_{50} values[63] (**Figures 5.2D** and **5.3C, 5.4C**). Finally, binding affinities of the molecules to first generation brain-derived fibrils were obtained via fluorescence polarisation experiments to validate targeting of the higher order aggregates (**Figure 5.2E, Figure 5.3D**,

5.4D) leading to subsequent attenuation of fibril catalysed aggregation. A degree of specificity is implied by the fact that no effect is observed on the aggregation of Aß42 (**Figure 5.5**).

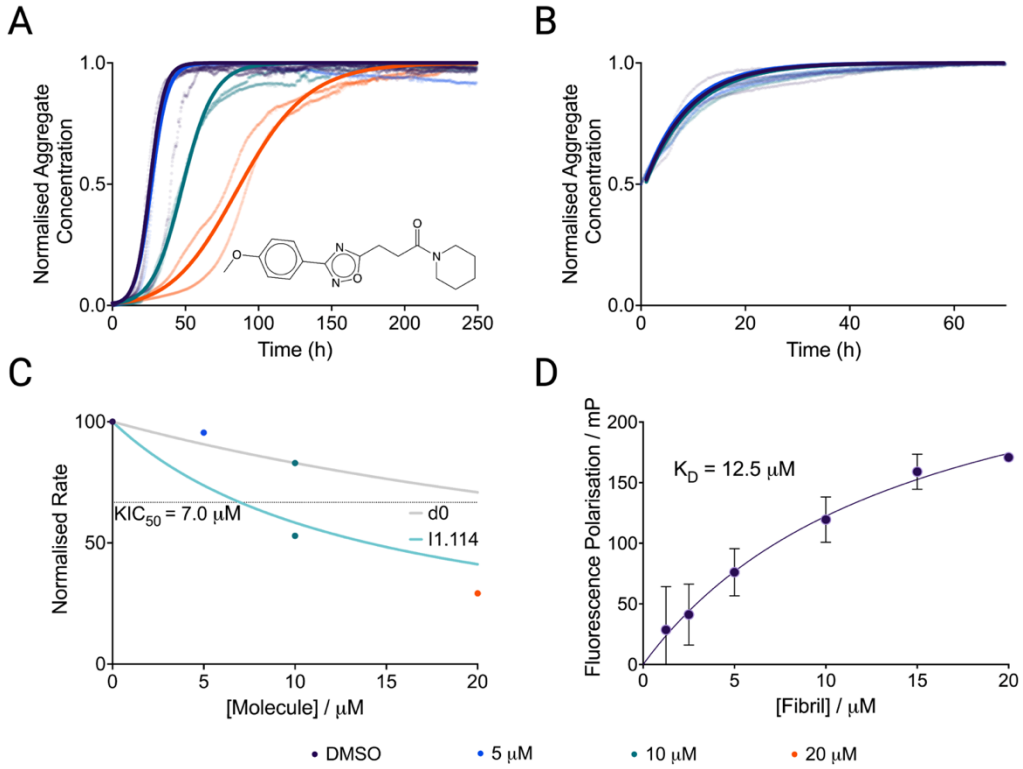


Figure 5.4. Kinetic analysis of a tau aggregation inhibitor (I.114) identified through iterative machine learning. (**A**) As above but for I.114. The molecular structure of I1.114 is shown (**B**) As above but for I.114. (**C**) Approximate rate of reaction (taken as $1/t_{1/2}$, normalised between 0 and 100) in the presence of 2 different molecules, the original docking hit d0 (grey), and I1.114 derived from it (light blue). The KIC_{50} of I1.114 (7.0 µM) is indicated by the intersection of the fit and the horizontal dotted line. (**D**) Change in fluorescence polarisation (in mP units) of 10 µM I1.114 with increasing concentrations of K12 fibrils (concentrations given in monomer equivalents). Error bars indicate the SD. The solid line is a fit to the points using a one-step binding curve, estimating a K_D of 12.5 µM ± 9.6 µM (SD) for I1.114. Kinetic aggregation experiments were run by Alessia Santambrogio.

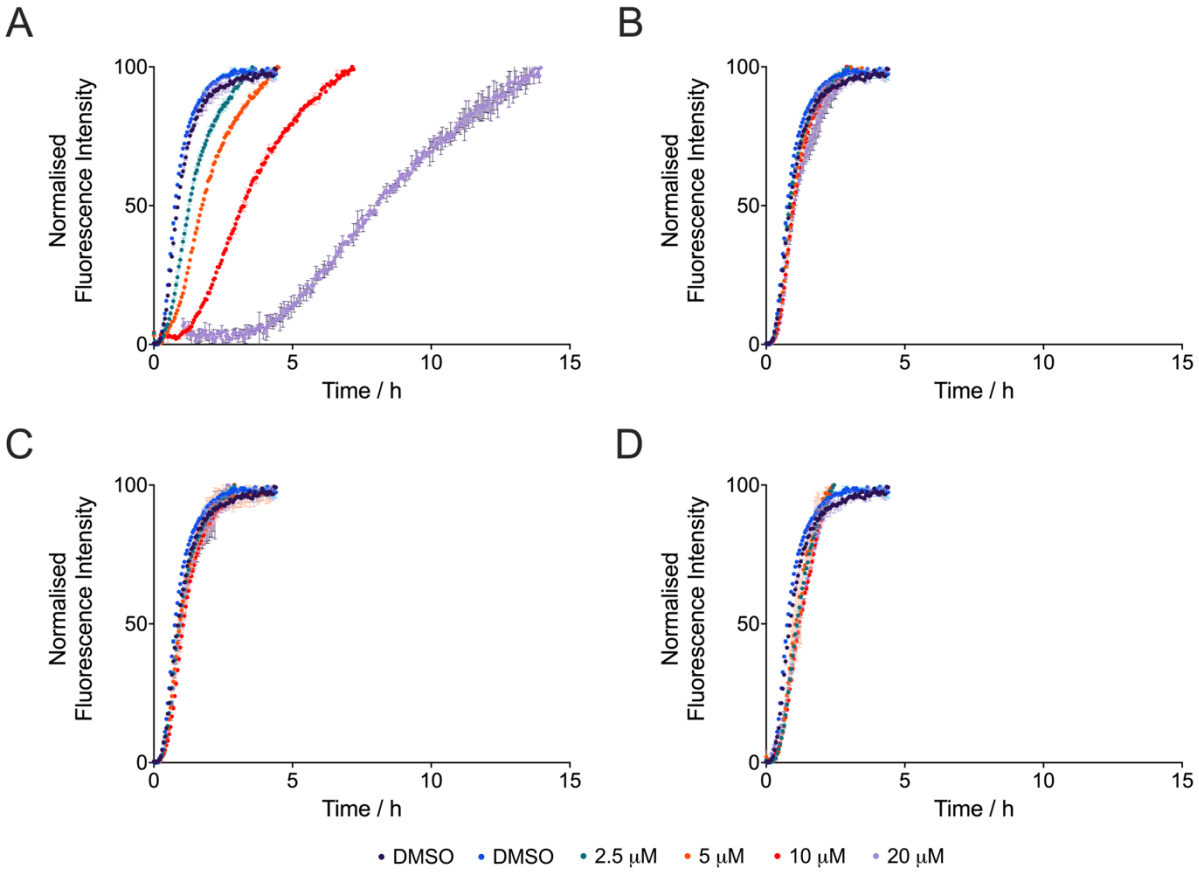


Figure 5.5. Kinetic traces for Aß42. Aß42 (40 nM seed, 2 μM monomer) was aggregated in the presence of 1% DMSO (purple and blue points), and molecules at 2.5 μM (teal), 5 μM (orange), 10 μM (red), and 20 μM (lilac). (**A**) Adapalene, a positive control with previously reported anti-aggregation potency[215]. (**B**) I1.21 (**C**) I1.51 (**D**) I1.114.

5.5. Discussion

Tau aggregation is a major target for disease-modifying AD therapies[216]. Since tau is known to self-assemble into a range of distinct amyloid fibril polymorphs underlying diverse neurodegenerative diseases[205-207], the study of aggregation *in vitro* should be concerned with the specific structure of the products of the process and the specific mechanism of aggregation. I have reported an approach to achieve this goal, using brain-derived fibrils as seeds in aggregation assays that can faithfully propagate their polymorphs, in the correct conditions. The results indicate that brain-seeded quiescent aggregation assays can provide insights into how tau strains might be targeted with small molecules. This was illustrated with an AD-derived polymorph. It may be possible to achieve strain-specific inhibition of aggregation processes with molecules obtained via this structure-based iterative machine learning method, which could be validated via testing against other tauopathies.

5.6. Materials and methods

5.6.1. Protein purification

K12 sequence:
MGSSHHHHHHSSGLVPRGSHMQTAPVPMPDLKNVKSKIGSTENLKHQPGGGKVQIV
YKPVDLSKVTSK**A**GSLGNIHHKPGGGQVEVKSEKLDFKDRVQSKIGSLDNITHVPGG
GNKKIETHKLTFRENAKAKTDHGAEIVYKSPVVS
0N3R tau sequence:
SSHHHHHHSSGLVPRGSHMAEPRQEFEVMEDHAGTYGLGDRKDQGGYTMHQDQE
GDTDAGLKAEEAGIGDTPSLEDEAAGHVTQARMVSKSKDGTGSDDKKAKGADGKT
KIATPRGAAPPGQKGQANATRIPAKTPPAPKTPPSSGEPPKSGDRSGYSSPGSPGTPGS
RSRTPSLPTPPTREPKKVAVVRTPPKSPSSAKSRLQTAPVPMPDLKNVKSKIGSTENL
KHQPGGGKVQIVYKPVLSKVTSK**A**GSLGNIHHKPGGGQVEVKSEKLDFKDRVQSKI
GSLDNITHVPGGGNKKIETHKLTFRENAKAKTDHGAEIVYKSPVVSGDTSPRHLSNV
SSTGSIDMVDSPQLATLADEVSASLAKQGL
K12 and 0N3R tau were purified as described previously[114]. Briefly, sequences for K12 and 0N3R tau with cysteine to serine mutations were cloned into PET-28a vectors and transformed into BL21(DE3) *E. coli*. Cells were grown and protein expression induced using an overnight autoinduction method described previously[217]. Crude lysate was prepared as described previously with the addition of a boiling step prior to application to carboxylmethyl fast flow (CMFF) capture. K12 or 0N3R was eluted from CMFF resin over a 20 column volumes (CV) linear gradient from 100 – 500 mM NaCl. Pooled CMFF eluate was added to Sepharose High Performance (SPHP) resin and eluted over 40 CV linear gradient from 100 – 600 mM NaCl. SPHP fractions were pooled, precipitated in acetone, and dissolved in 8M GuHCl prior to size-exclusion chromatography (SEC) separation on a 26 x 600 mm Superdex 75 column equilibrated in 20 mM sodium phosphate, pH 7.4. Proteins were lyophilised and frozen at -80 °C until use.

5.6.2. First-generation seed amplification

Generation of AD-derived and PiD-derived K12 seeds was conducted as described previously with several modifications. Heparin was avoided in this study despite being previously published due to failure to recapitulate the correct polymorph[218]. NaF was replaced with 250 mM Na_3Citrate. Brain homogenates utilised in this study include sporadic AD (sAD) 2 and PiD 5 listed in the supplement of reference [114]. First-generation reactions were seeded with $1x10^{-5}$ concentration of brain homogenates in the presence of 10 μM K12, 10 μM ThT, 250 mM Na_3citrate, 40 mM HEPES, pH 7.4. Reactions were subjected to rounds of 60 s shaking (500 rpm, orbital) and 60 s rest with periodic ThT readings every 15 min at 37 °C in a 384-well Nunc microplate (non-treated polymer base #242764) in a BMG FluoStar lite with aluminium sealing cover to prevent evaporation. Fibrils were harvested by scraping and pooling reaction contents once ThT fluorescence reached plateau > 20 h.

5.6.3. Second-generation 0N3R fibril amplification

0.3 μM of first-generation (brain homogenate seeded) K12 fibrils were added to monomeric recombinant 0N3R tau. The second-generation reaction buffer included 250 mM Na_3citrate, 40 mM HEPES, pH 7.4 with or without ThT depending on use for cryo-EM analysis. Reactions were again incubated in BMG FluoStar lite at 37 °C with rounds of shaking (500 rpm orbital, 60 s) and rounds of rest (60 s) with periodic ThT reads every 15 min. Fibrils were recovered for biophysical characterisation (GuHCl sensitivity, ATR-FTIR, protease resistant core sizing by MS) or structural analysis (cryo-EM) by gently aspirating the reaction solution to avoid fragmentation or clumping of fibrils; reactions were collected after reaching the ThT fluorescence plateau at > 80 h reaction time.

5.6.4. Computational docking, iterative ML methods and code availability

First, we selected a binding site on the tau fibrils. To achieve this goal, we analysed a structure of a tau fibril (PDB ID: 503l)[207] using Fpocket[213], which identifies potential binding pockets based on volume criteria. We identified a pocket on the fibril surface (encompassing residues Asp314-Leu315), which had high surface exposure, necessary for secondary nucleation and high hydrophobicity, as identified by CamSol[96] (**Figure A.2**), allowing the pocket to participate in aggregation. For the selection of screening compounds, we used the ZINC library, which contains a set of over 230 million purchasable compounds for screening[219]. To prioritise the

chemical space of small molecules considered in the docking calculations, central nervous system multiparameter optimisation (CNS MPO) criteria[220] were applied, effectively reducing the space to ~2 million compounds. In particular, CNS MPO has been shown to correlate with key in vitro attributes of drug discovery, and thus using this filter potentially enables the identification of compounds with better physicochemical and pharmacokinetic properties pertaining to brain penetration, where tau is localised. We further subjected these compounds to docking calculation against the binding site identified above using AutoDock Vina[221]. To increase the confidence of the calculations, the top-scoring 10000 small molecules were selected and docked against the same tau binding site, using FRED (OpenEye Scientific Software). The top-scoring, common 1000 compounds in both docking protocols were selected and clustered using Tanimoto clustering, leading to a list of 130. Molecules were then obtained and tested experimentally in aggregation and binding experiments, before application of the iterative machine learning procedure as outlined in the online repository https://github.com/rohorne07/Iterate.

5.6.5. Preparation of the compounds

The centroids from the above 130 clusters were selected for experimental validation. Compounds were purchased from MolPort (Riga, Latvia), and in the cases for which centroids were not available for purchase, the compounds in the clusters with the closest chemical structures were used as the representative compounds instead. In the end, a total of 102 compounds were purchased (centroids and alternative compounds in 28 clusters were all not available for purchase) and then prepared in DMSO to a stock of 5 mM. Stocks were diluted in DMSO to 100-fold above the final desired final concentration, before addition to aggregation reactions at 100-fold dilution (1% DMSO). All chemicals used were purchased at the highest purity available (>90% in purity).

5.6.6. Fluorescence polarisation

10 μM of each molecule was incubated with increasing concentrations of K12 fibrils in the same buffer as used for kinetic experiments, supplemented with 1% DMSO. After incubation, the samples were pipetted into a 96-well half-area, black/clear flat bottom polystyrene nonbinding surface (NBS) microplate (Corning 3881). The fluorescence polarisation of the

molecule was monitored using a plate reader (CLARIOstar, BMG Labtech, Aylesbury, UK) under quiescent conditions at room temperature, using a 360 nm excitation filter and a 520 nm emission filter.

5.6.7. Recombinant Aß42 expression

The recombinant Aß42 peptide (MDAEFRHDSGY EVHHQKLVFF AEDVGSNKGA IIGLMVGGVV IA), here called Aß42, was expressed in the *E. coli* BL21 Gold (DE3) strain (Stratagene, CA, U.S.A.) and purified as described previously. Briefly, the purification procedure involved sonication of *E. coli* cells, dissolution of inclusion bodies in 8 M urea, and ion exchange in batch mode on diethylaminoethyl cellulose resin followed by lyophylisation. The lyophilised fractions were further purified using Superdex 75 HR 26/60 column (GE Healthcare, Buckinghamshire, U.K.) and eluates were analysed using SDS-PAGE for the presence of the desired peptide product. The fractions containing the recombinant peptide were combined, frozen using liquid nitrogen, and lyophilised again.

5.6.8. Aß42 aggregation kinetics

Solutions of monomeric Aß42 were prepared by dissolving the lyophilised Aß42 peptide in 6 M guanidinium hydrocholoride (GuHCl). Monomeric forms were purified from potential oligomeric species and salt using a Superdex 75 10/300 GL column (GE Healthcare) at a flowrate of 0.5 mL/min, and were eluted in 20 mM sodium phosphate 200 µM EDTA, 0.02% NaN_3, pH 8. The centre of the peak was collected and the peptide concentration was determined from the absorbance of the integrated peak area using $\varepsilon_{280} = 1490$ l mol^{-1} cm^{-1}. The obtained monomer was diluted with buffer to the desired concentration and supplemented with 20 µM thioflavin T (ThT) from a 2 mM stock. Each sample was then pipetted into multiple wells of a 96- well half-area, low-binding, clear bottom and PEG coated plate (Corning 3881), 80 µL per well, in the absence and the presence of different molar-equivalents of small molecules in 1% DMSO or 1% DMSO alone as a negative control. Assays were initiated by placing the 96-well plate at 37 ºC under quiescent conditions in a plate reader (Fluostar Omega, Fluostar Optima or Fluostar Galaxy, BMGLabtech, Offenburg, Germany). The ThT fluorescence was measured through the bottom of the plate using a 440 nm excitation filter and a 480 nm emission filter.

5.7. Contributions

This chapter is substantially derived from a submission to *Nat. Chem. Bio*. Michael A. Metrick II, Alessia Santambrogio (A. S.), Michele Vendruscolo and I conceived the project and wrote the article. Tau variant synthesis and kinetics experiments were done by A. S. and initial docking work was carried out by Z. Faidon Brotzakis. I carried out the rest of the computational work, analysis, Aß42 kinetics and fibril binding experiments. Rebecca C. Gregory produced the Aß42.

6. Future directions

"You cannot plan the future. Only presumptuous fools plan. The wise man steers."
– Sir Terry Pratchett

6.1. Impact and developments

The future of therapeutics for neurodegeneration appears likely to consist of antibodies targeting amyloid in some manner in the near term and may be augmented with approaches such as those employed here in the mid-term. Looking any further beyond that is challenging however, and the methods used then will likely be determined by what progress can be made in understanding the steps in disease pathology. Given the rising star of CRISPR-Cas based systems, some limited genetic interventions to bolster the body's natural homeostasis systems may become available, given how effective such genetic interventions have already been shown to be in addressing other disease areas and preventing protein aggregation in simple model organisms[16,222]. Such treatments would face the same sorts of BBB penetrance and inflammatory roadblocks as current antibody treatments do, though their effect would presumably be significantly greater and inflammatory responses could be reduced by moving away from viral capsid delivery systems[223]. However this remains only a possibility on the horizon at present, and likely an extremely expensive one[224].

The work here was heavily geared towards optimising a single approach to tackling this disease class that could provide benefits to patients in the short term at lower expense than current methods. Hopefully this will contribute to the development of diagnostics and therapeutics that

can enable early detection and delay onset of disease, and in so doing give patients more quality years of life. As outlined in the following sections the work described has influenced various projects in academia and industry. There is of course significantly more work to be done before any of this becomes something that can provide tangible benefits to patients. Nonetheless, it hopefully represents a step on the long road to combating this formerly intractable set of conditions.

6.1.1. Iterative Learning

The iterative learning project has begun a more concerted effort within the Vendruscolo research group to applying machine learning to the problem of preventing protein misfolding. In many cases this work has outgrown what has been shown here. This includes the application of novel and improved ML architectures, 'deep docking' to accelerate physics based docking approaches with ML[197], selective docking to minimise off target binding, and development of learnable pharmacophore models. One particularly promising use case is based on AlphaFold, which produces contact maps that can be used to develop ligand binding models based on graph transformers[225,226].

A crucial missing piece of the work outlined in this thesis is the lack of comprehensive in vivo evidence of effectiveness, which is being addressed in collaboration with the Biomedical Research Foundation of the Academy of Athens who will be testing a number of the αS aggregation inhibitors in mouse models. Additionally, the Biomedical Sciences Research Centre in Vari will be testing a number of the tau aggregation inhibitors in Drosophila models. A very recent result came from a collaboration with the Aigbirhio group (Wolfson Brain Imaging Centre, Cambridge) involving the use of I3.08 and I4.05 as potential positron emission tomography (PET) tracers for αS fibrils in biological samples. An effective PET tracer binds to its target, in this case αS aggregates, allowing high resolution imaging in patients[227]. The Aigbirhio group began trialling the molecules' suitability for this purpose using fluorescence-based methods in mouse tissue expressing human A53T αS, a mutation known to accelerate PD. This shows promising colocalisation between the intrinsic fluorescence of I3.08 and I4.05, when bound to fibrils, and a GFP labelled anti-αS pSer129 antibody (**Figure 6.1**). The Aigbirhio group reported that this experiment needs optimisation as they believe the secondary anti-mouse antibody is staining structures which are not aggregate-like. This means it may be

binding off targets as well as the anti-αS primary antibody (bound to αS fibrils). The antibody staining will therefore require tuning, or else replacement of the secondary antibody with a more specific one. As the pSer129 primary antibody is somewhat specific for phosphorylated αS it may also be missing aggregates which are not phosphorylated. This experiment therefore gives a rough indication that the molecules are functioning as hoped in tissue, which will require further validation. The Aigbirhio group are hopeful about the potential of these molecules as tracers, but experiments measuring blood brain barrier penetrance will be the deciding factor on how far they progress.

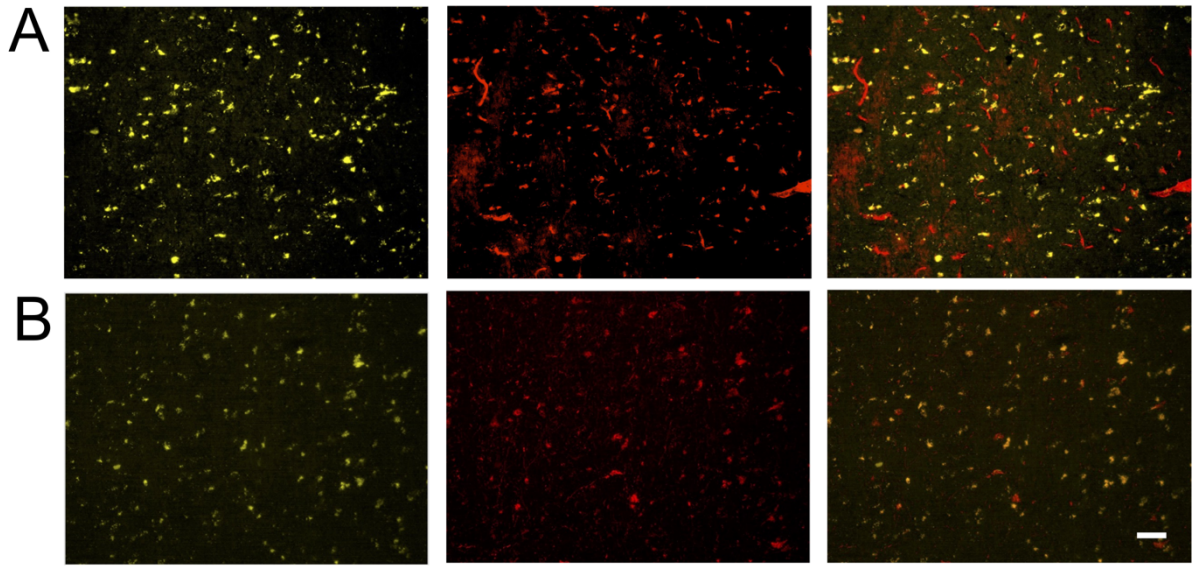


Figure 6.1. Preliminary data showing staining of mouse brain expressing human A53T αS. The secondary antibody is overstaining in this example and requires optimisation or replacement. Intrinsic fluorescence of **(A)** I4.05 and **(B)** I3.08 upon fibril binding (yellow, first column), and a GFP labelled pSer129 anti-αS antibody (red, second column) are shown. Colocalisation of each molecule and the GFP labelled antibody is shown in the third column. I4.05 has qualitatively greater coverage than I3.08, matching their relative binding efficacies. The scale bar is 45 μm. Carried out by Yanyan Zhao.

From these experiments I will discover how well the in vitro assays can predict efficacy in vivo, a notorious difficulty for most drug and diagnostic programs but especially for those targeting protein misfolding[228]. While the optimisation methods work in principle, a predictive assay is key, and that is something that ML is not yet able to help us identify given the paucity of publicly available data on this topic. Relating in vitro assay results to efficacy in animal models, and then relating animal model results to human patients, is a uniquely challenging task in the field of neurodegeneration. Indeed, relating effectiveness in animal models to human patients has proved problematic for trials such as those involved in aducanumab's development due to animal models displaying very different responses to gene mutations and

overexpressions than human patients appear to[229]. Utilising these relatively poor models means the difficulties of designing experiments with good predictive power for outcomes in human patients remain considerable, but these models are nonetheless essential. As our understanding develops, improved models of disease will be invaluable in improving treatments.

6.1.2. Exploration and exploitation via generative modelling

The generative models employed here were state of the art at the time of writing, but represent two computational approaches out of many. Alternatives include diffusional modelling, currently the most novel technique for molecule generation. This approach can create novel ligands within their binding pockets, simultaneously generating the pocket if desired[230]. This approach has yielded some promising results against αS already within the Vendruscolo group, something that will hopefully be expanded to other protein misfolding systems.

Generative modelling in combination with effective QSAR models are likely key to vastly reducing the resource cost of high throughput screens by carrying out the majority of the work in silico. A pitfall of the approaches used here is that the QSAR models were not trained on enough data to give a very accurate prediction of efficacy, especially in areas of the chemical outside the main chemotypes in the training data. While this level of accuracy was still effective for in silico library screening, this made effective molecule generation challenging, given that any poor structural predictions by the QSAR models would be amplified by rounds of reinforcement learning in the case of the GraphINVENT model. Nonetheless, we were able to generate molecules in the same region of the chemical space as the more potent leads, and the QSAR models identified one effective molecule from among the set of available training compounds for the generative model. As such some useful information was learned. This approach would be more powerful with the larger training sets, showing the need for greater data sharing in academia and industry, or at least better curation of data within organisations for use with these methods.

6.1.3. Developing oligomer detection methods

Nanopore detection methods provide ways of testing the effectiveness of therapies against the key pathogenic agent in protein aggregation, the smaller oligomers. If this approach could be

upscaled in the manner it has been for nanopore DNA sequencing, this could provide a high throughout oligomer detection method, something sorely lacking from the field currently. It should be noted that high throughout nanopore methods currently rely on protein nanopores embedded in membranes which cannot be tuned in terms of size the way solid state nanopores can be[231]. Solid state nanopores, though tuneable, have not yet been upscaled although this may become feasible in the future[232]. Similarly, there is potential to engineer protein pores to create larger pore diameters and tune diameter by design. Solid state nanopores are currently at the forefront of state of the art research into protein sequencing, which has some bearing on the task of oligomer detection[233,234]. Each nanopore type therefore has its strengths. Weaknesses may be addressed by future advances, so it remains to be seen which of the two types come to dominate over the other in the area of protein characterisation. The requirement for DNA tags currently restricts this approach to in vitro experiments and model organisms where artificial amino acids can be introduced. This would already be a significant boost to early-stage research, which currently rely on proxy measurements followed by laborious validation assays to measure actual oligomer populations in the presence of a small number of lead candidate inhibitors.

6.1.4. A common mechanism?

Expanding this approach into other disease areas carries the assumption that similar mechanisms are occurring in other diseases featuring misfolded aggregating proteins. This assumption is contentious, but reasonably well established for many conditions such as tau in AD, the other main case study besides αS described here. Efforts are being made to target similar scenarios elsewhere, using the same strategy of blocking production of toxic aggregates and removing inert aggregates as burdens on cellular function. IAPP aggregates were tentatively addressed here, as a likely pathogenic agent in beta islet cell death. Similarly, other proteins susceptible to aggregation such as FUS, TDP-43, medin, and others could be addressed in this way. Time will tell whether the approach of blocking oligomer formation and removal of aggregates will be useful only in specific scenarios, or in a more general sense.

It seems most likely at present that misfolding diseases are caused by failures in the protein homeostasis systems as the body ages. Stresses from sources such as reactive oxygen species and inflammation accumulate, impairing these systems, which partially explains the sporadic

and varied phenotypes of disease as each case is therefore unique and largely determined by environmental factors, albeit with influence from genetics. Addressing the disease at its 'source' may involve repairing a system consisting of many modules, each of which could be functioning at a different level for each patient. This may require tailored interventions, a treatment paradigm termed 'personalised medicine'[235]. However, identification of particular nodes of vulnerability and strengthening them may be key to providing a permanent solution in a more pan-patient manner. Such interventions would however require better knowledge of the systems in question which are extremely complex, and could potentially require genetic treatments which are currently out of clinical reach but may prove pivotal in the coming decades.

6.2. A more general outlook

While the field of machine learning certainly moves the fastest, experimental developments in the area of protein misfolding are also appearing at a high rate. The most interesting recent development is the discovery that tau protein fibrils may convert between different structural polymorphs as the aggregation proceeds[212]. If this occurs in patient brains this adds an additional layer of complexity to the problem of targeting fibrils. Rather than a static target there is an ensemble of sequentially formed structures. Whether other misfolding proteins also follow this behaviour remains to be seen and is the subject of current work. If this behaviour is indeed common, it may be a better approach to target structures earlier in this polymorph maturation cascade or develop inhibitors that can promiscuously target any of the on-pathway polymorphs, rather than the dominant final species, as I have done in this work for αS and tau.

More generally this demonstrates the need to update our assumptions about the mechanisms of protein misfolding as soon as this information becomes available and to alter our approaches accordingly. This makes neurodegeneration a challenging moving target, where assays may be rendered obsolete as new information becomes available, and shows the importance of avoiding sunken cost fallacies that waste resources and time. This issue arises from the effort to prevent a disease without full understanding of it, entailing inevitable, but necessary attrition. For example, it was previously considered of minimal importance to replicate the exact structures formed within patient brains. As long as the protein aggregated, and fibrils were formed the assay was considered a viable model of the aggregation. This was in part due to the

inability to ascertain the fibril structure with good accuracy, something that has been addressed with the development of cryo-EM techniques for this task. With high trial attrition over the past few decades, increasing emphasis has been placed on directly recapitulating the disease process and the associated structures that are formed. Whether it is possible to achieve this, in vitro, in a way that can be converted to a high throughput assay is not known. The assay reported for tau was the only condition found in that study capable of eventually recapitulating the dominant tau fibril polymorph observed in the latter stages of Alzheimer's disease. It uses an extremely high protein concentration of tau (6 mg/ml), rendering it effectively useless for screening unless tau production could be vastly upscaled. As is always the case, some degree of compromise between accurate modelling and practicality must be made, but not in such a way that the correlation between the assay and disease is reduced to an insignificant level. It may be possible to recreate high local concentration conditions via LLPS, although this introduces further complications due to the need for crowding agents which add questions about physiological relevance. Alternatively, a less resource costly assay could be developed that recreates an important polymorph intermediate on the maturation pathway, which exhibits benefits in animal models when targeted. Such compromises are especially necessary in neurodegeneration research, where the cost of replicating disease conditions faithfully often renders drug screening impractical. This also means alleviating the experimental resources required with computational techniques is especially important in this field.

Providing therapies at the correct point in time is also a key issue, and what has not been touched upon as much in this thesis is diagnostics. To counter the diseases effectively we need to detect them prior to symptoms appearing, at which point irreversible neuronal death has already occurred. Current diagnostics vary greatly in their methods and have differing predictive abilities given the sporadic nature of most neurodegenerative disease cases. These include the RT-QuIC assay already mentioned, which has had some success diagnosing Alzheimer's disease and Parkinson's disease via detection of aggregates in the CSF. This is not an especially easy test to distribute widely as CSF extraction generally requires a day in the hospital. Other methods include machine learning medical smart watch data, which in one trial led to accurate predictions of Parkinson's disease 7 years in advance of symptoms appearing. Furthermore, blood measurements of biomarkers such as TNFα and neurofilament light chain, which are released into the extracellular space by dying neurons, can provide evidence of neuronal damage, though not of its cause[236].

The progress described may seem rather small compared to that in other disease areas, but set against a backdrop of decades without a single positive clinical trial it is a significant improvement. As such it is an exciting time for neurodegenerative research, after decades of bleak attrition it seems headway is finally being made in supplying both diagnostics and treatments. Some may argue that the goal posts have simply been moved, and that the FDA has become increasingly lenient to companies attempting to deliver treatments against neurodegeneration[237]. This does however miss the point that until relatively recently there was little consensus on how exactly to even quantify an effective drug in this area. Most diseases develop over comparatively short spans of time and treatment outcomes are clearer cut. Neurodegeneration is an entirely different paradigm that requires altered expectations of outcomes. While it is of course unacceptable to deliver drugs without significant efficacy, it is important to acknowledge that we will likely only observe small perturbations in a clinical trial spanning a couple of years when trying to treat a disease spanning decades, especially given the extreme heterogeneity of pathologies observed. Relatively small perturbations delivered early on could have a significant effect over time, and encouraging pharmaceutical companies to keep working on these diseases is essential if any progress is to be made. The growing understanding of regulatory bodies that this disease area is uniquely challenging and requires a different approach to other diseases is both welcome and essential to further progress in this area. This project has aimed to demonstrate the development and application of tools to accelerate the speed with which we can make this progress, through both computational and experimental techniques.

6.3. Materials and methods

6.3.1. Staining of A53T expressing mouse brain tissue

Brain tissue slides were washed in cold acetone and then twice again in buffer (10mM PBS, 0.1% Triton X100). The slides were then blocked with 5% w/v (BSA) in buffer (10mM PBS, 0.1% Triton X100) for 1 h at RT. The excess BSA solution was removed and the primary antibody (pSer129, 1 mL, 1:1000 fold diluted) was added to each tissue slice. The slides were incubated overnight at 4°C. Slides were washed twice in buffer (10mM PBS, 0.1% Triton X100). Secondary antibody (Alexafluor555/647 donkey anti-rabbit, 1:1000 fold diluted) was added as well as each molecule at 20 uM in 1% DMSO. The slides were then incubated at RT

in a foil-covered tray for 1 h. Slides were washed twice in buffer (10mM PBS, 0.1% Triton X100). Counterstaining with DAPI (Sigma-Aldrich, MBD0015; 1:1000 fold diluted) was carried out for 15 minutes. 1-2 drops (depending on how much of the slide is covered by brain slices) of FluorSave (or a mountant with DAPI) were then added before addition of a coverslip and curing overnight at 4°C. Samples were imaged with a Leica CTR 6000.

Appendix

A. Targeting Parkinson's disease with iterative learning

i. Docking and Machine Learning Implementation

A full description of the initial docking approaches can be found in the previous work[64], using AutoDock Vina[90] and FRED[91] docking software, but is also explained in overview here. As described in the main text, the binding site encompassing residues His50−Lys58 and Thr72−Val77 on PDB 6CU7[98] was selected due to its propensity to form a pocket according to Fpocket[92] software and its simultaneous mid to low solubility according to CamSol[96] (**Figure 2.2**). Additionally, a key histidine residue in this site was predicted to protonate below the pH value where αS more readily aggregates (pH 5.8). A binding box was selected that had size 12 Å by 12 Å by 9 Å centred at 10.00 Å, 9.89 Å, 11.52 Å on the 6CU7 PDB, encompassing the site of interest. The target protein was left rigid, while the ligand was flexible, able to translate and rotate (including rotation of internal bonds). We prepared (added hydrogens) the target protein using Autodock tools. To increase the accuracy of the docking energy estimate, the exhaustiveness was increased compared to the default value of 8, to 20. 5 poses were output, and the best pose binding energy was selected as the binding energy label for that ligand. The choice of rigid target was made in order to decrease the computational cost of the high throughput screen of the 2 million compounds in phase 1.

Inspired by the increasing usage of consensus scoring, i.e combining multiple docking energy estimates by different docking programs, we performed docking of the 100,000 best binding molecules from AutoDock Vina, using FRED in phase 2. For each of the top 100,000 best

AutoDock Vina ligands, we combine the ligand with the target into a single .pdb file, and from that supply the information of the ligand to Openye's Spruce module to prepare an .oedu file that contains the grid position of the binding site. Then, the compound is bound to the target site and a single best pose and binding energy is output, that constitutes the FRED binding energy label for this compound. The top 10,000 are then clustered to obtain 79 representative centroids for testing. The pipeline is modular, and it is possible to incorporate any type of docking software the user might choose. In this study we have used AutoDock Vina, which is a publicly available software that is efficient at scale, and FRED. AutoDock Vina is relatively computationally efficient at scale, and we chose to use FRED since the top scoring pose prediction of FRED has been shown[238] to be able predict within 2Å of the native pose in 70% of examples tested. However, alternative open-source or free for academic use docking software such as rDock, LeDock and others can be used instead of FRED, with relatively little difference in performance as shown previously[239]. The performance of AutoDock Vina is comparable with other open-source software.

The code for testing the ML models on aggregation or docking data are available at https://github.com/rohorne07/Iterate. We initially tested the machine learning strategy on docking data (best R^2 ~0.6-0.7) before moving to experimental aggregation data (best R^2 ~0.2-0.3) to get an impression of the feasibility of the project, given the larger data sets available for the docking scores (**Figure A.4** and **Figure A.5**). The docking scores were calculated for the 'evaluation set', the in silico library that was used for iterative experimental screening in the main text. Both AutoDock Vina and FRED simulations were carried out on the evaluation set, giving binding scores for each molecule against the αS 6CU7 fibril structure pocket. The compound encoder was implemented as in Hie et al.[81] to obtain representations of all the molecules. The next sections briefly summarise the functioning and output of the prediction module.

Prediction module. The prediction module consisted of a shallow model designed to be appropriate for small data sets and easily applicable on standard hardware available for most laboratory workers over a short timescale. As a first line test Gaussian process regression (GPR) was employed alone, following Hie et al.[81] with training and testing carried out with cross validation on 4000 molecules from within the evaluation set. The metric used to evaluate performance in this case was the R^2 score or coefficient of determination. This score measures the goodness of fit between a set of predictions and the ground truth values. This score ranges

from 1, in a perfect fit, to arbitrarily negative values as a fit becomes worse, and is 0 when the predictions are equivalent to the expectation of the ground truth values of the training set[240]. This was compared with a naïve Bayes, which failed to score above 0 for any training set size on both docking and aggregation data.

The GPR kernel was initially the same as that utilised by Hie et al.[81] , i.e. a combination of a constant kernel and a radial basis function (RBF). Using these initial settings, R^2 scores of ~0.2 were obtained for the docking data. Hyperparameter optimisation yielded only marginal improvements in this performance. A selection of other kernels was tested, and all models were optimised via hyperparameter tuning before implementation, but most did not offer an improvement in performance. The Matérn kernel, a generalisation of the RBF with an extra parameter controlling the smoothness of the function, did however show a marginal improvement. These flexible functions are the most likely to be able to fit shallow energy minima problems such as those encountered here. The R^2 scores were still low, especially for smaller training sets as would be available from experiment, but represented a viable starting point.

At this point a 2-layer model was applied. This reflected the strategy used by Hie et al. [81] in fitting a Gaussian process regressor (GPR) to the residuals of another model, in that case a multi-layer perceptron (MLP). An MLP did not show a dramatic improvement over the GPR alone both in that work or when tested with the docking scores here, however a random forest regressor (RFR) with stacked GPR did show a further improvement both in terms of the R^2 (~0.6-0.7) and the quality of the molecule sets predicted during the simulation, as can be seen in **Figure A.5**.

This set up gave improved results in both R^2 and hit rate, while retaining an easy to implement and efficient model. The average Pearson's coefficient of correlation ranged between 0.25 and 0.3 for both the coupled (GPR+RFR) and uncoupled models (RFR alone), which while modest matched the values obtained by Hie et al. during their testing. RFR was more demanding computationally, but given the small size of the experimental training sets in this scenario this was not a hindrance.

A simulation was created to mimic how the experimental cycle of testing might work using the docking scores as a surrogate for aggregation data. In the simulation, a random subset of 100

molecules was selected and the model trained on these molecules and their binding scores. The resultant model was then used to predict binding scores for the remaining molecules and rank them using a combination of the predicted value and the associated uncertainty value. The top 100 were then selected and their binding scores added to the training set as would occur in the experimental scenario, and this process was repeated 10 times. The ideal scenario would be that molecule sets with improved mean binding energy relative to the mean of the test set would be selected, and that selections would improve as the training set expanded, and this is what is shown in **Figure A.4** (though improvement is not drastic as further data is added, possibly due to the relative ease with which strong dockers are selected).

Different uncertainty penalties were tested during this process. We found that a low uncertainty penalty produced better results by removing the most overconfident predictions without placing too many limitations on the model. At the early stages most predictions with low uncertainty were those with predicted binding scores close to the mean of the training set. An excessive uncertainty penalty during these stages would cause the model to only predict molecules that it was confident in, which were also likely to be mild.

The same process was utilised using different parts of the molecular feature set (the latent vector consists of a tree vector representation of clusters within a molecule, plus a graph representation of the molecule), and it was found that GPR performance metrics were better when using the molecular graph alone compared with using the entire representation. In general, it is to be expected that fitting fewer features to a predicted value is easier for a regressor to achieve and so higher scores are obtained. However, a better average R^2 score across the data set does not necessarily lead to a better result in terms of the actual molecules picked, and we found using the full representation led to more hits being identified (**Figure A.4**).

A snapshot of the results of this testing is shown in **Figures A.4** and **A.5**. **Figure A.4** demonstrates 2 points: the performance was slightly improved using the Matérn kernel in place of the RBF kernel both in terms of overall hit selection and performance improvement with increasing training set size, and the full-length molecular representation gave a significant boost in terms of number of hits selected vs the truncated representation, despite lower R^2 scores. These results also provided some evidence that Gaussian process learning might work reasonably effectively even in this data sparse scenario albeit at a modest level. It was expected

that fitting experimental data would prove more challenging, however, and so a boost in performance was sought for that would not compromise the simplicity of the model, through use of the coupled RFR-GPR model. Correlation values of 0.6-0.7 were obtained using this set up on docking energies and a large portion of the data set (4000 molecules), and this fell to between 0.2 and 0.3 for the aggregation data (**Figure A.1**), which while low was encouraging given the much smaller data set and noisier data.

A

Model	Parameter	
	Parameter label	*Parameter value*
LR	fit_intercept	True
LR	copy_X	True
LR	n_jobs	1
LR	positive	False
MLP	hidden_layer_sizes	100
MLP	activation	relu
MLP	solver	adam
MLP	alpha	0.0001
MLP	batch_size	auto
MLP	learning_rate	constant
MLP	learning_rate_init	0.001
MLP	max_iter	200
MLP	tol	0.0001
MLP	warm_start	False
MLP	early_stopping	False
MLP	beta_1	0.9
MLP	beta_2	0.999
MLP	epsilon	1×10^{-8}
MLP	n_iter_no_change	10

LR = linear regressor, MLP = multi layer perceptron

B

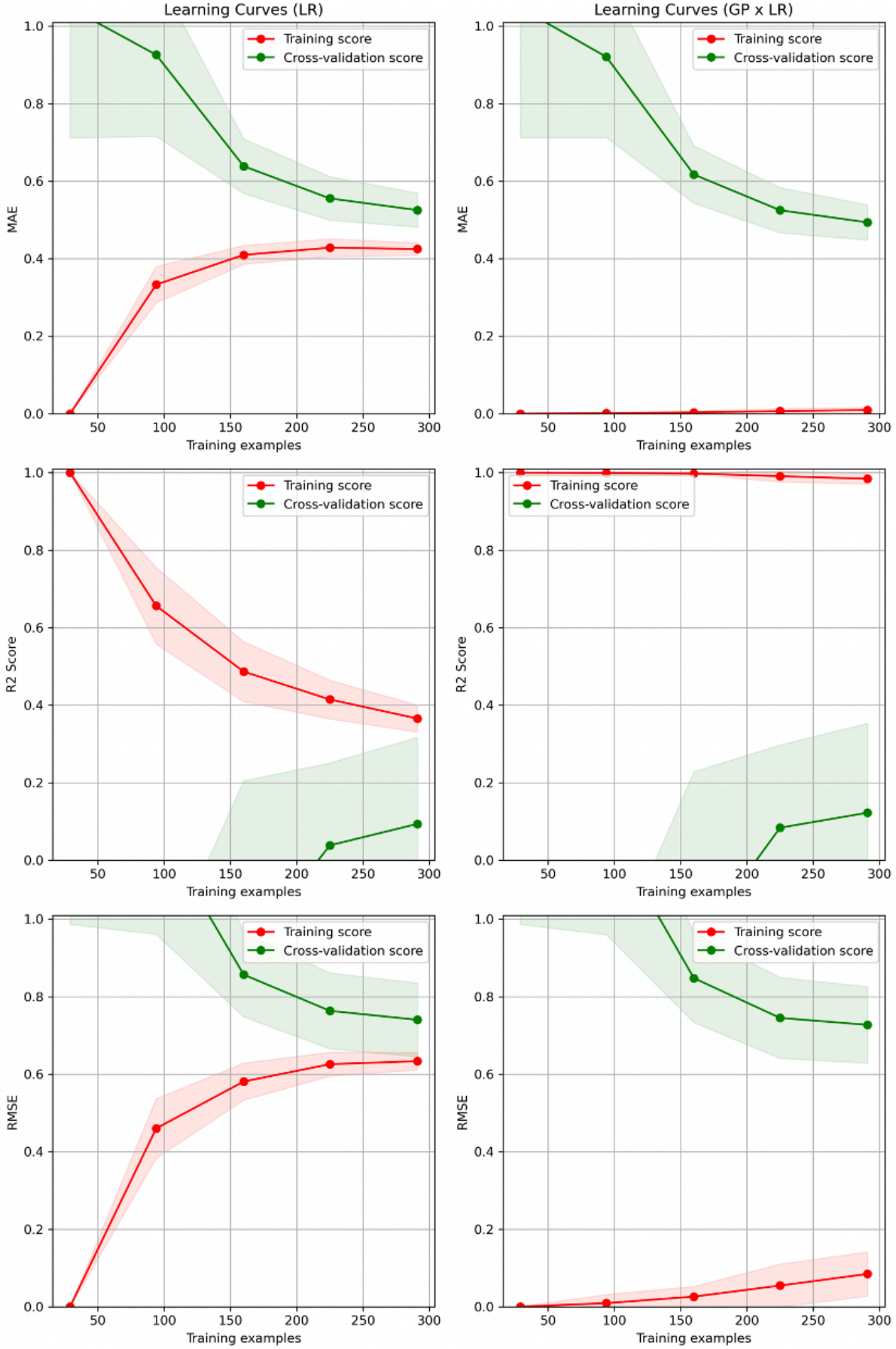
Model	Parameters		
	Kernel (if applicable)	*Parameter label*	*Parameter value*
GP	Constant	length_scale	1.0
GP	Constant	length_scale_bounds	fixed
GP	RBF	length_scale	1.0
GP	RBF	length_scale_bounds	fixed
GP	Matern	length_scale	1.0
GP	Matern	length_scale_bounds	fixed
GP	Matern	nu	1.5
RF	n/a	n_estimators	950
RF	n/a	criterion	squared_error
RF	n/a	max_depth	50
RF	n/a	min_samples_split	2
RF	n/a	min_samples_leaf	2
RF	n/a	min_weight_fraction_leaf	0.0
RF	n/a	max_features	log2
RF	n/a	max_leaf_nodes	None
RF	n/a	min_impurity_decrease	0.0
RF	n/a	bootstrap	False
RF	n/a	oob_score	False
RF	n/a	warm_start	False
RF	n/a	max_samples	None

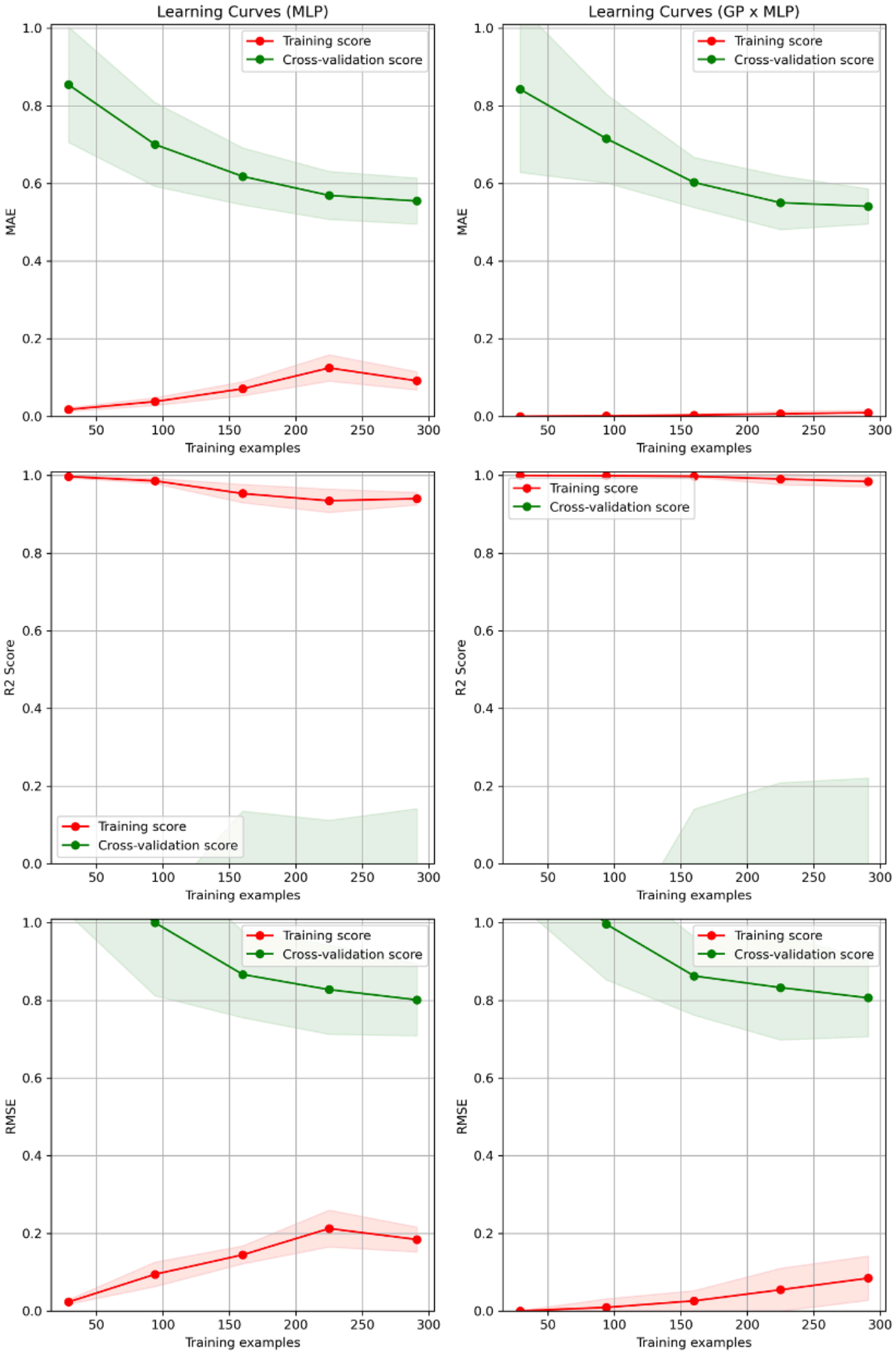
GP = Gaussian Process, RF = random forest, RBF = Radial Basis Function

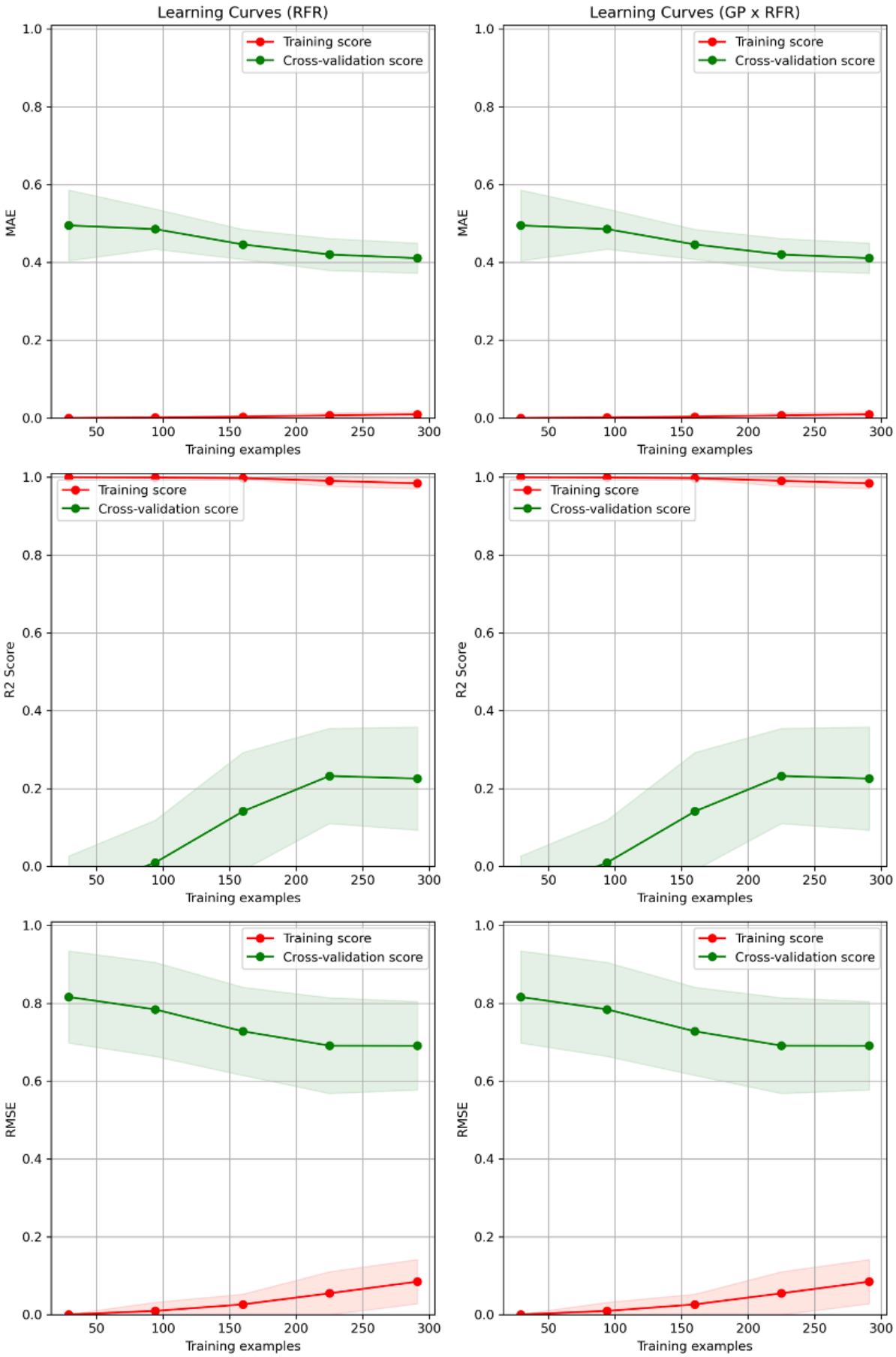
Table A.1. Parameters used in QSAR model optimisation. (**A**) Models such as LR and MLP were trialled with their default parameters either alone or in conjunction with a GP, but showed poor performance so were not further investigated. (**B**) GP and RF models were the best performing and so were subjected to hyperparameter optimisation via grid search cross validation using the R^2 score as the optimisation metric. The best performing parameters are shown. The performance of these models is shown in **Figure A.1**.

Brain Sample	Source	Gender (M/F)	Age at Death (years)	Disease duration (years)	Postmortem interval (h)	Primary diagnosis	Additional diagnosis
DLB1	Ghetti	M	81	N.A.	20	Diffuse Lewy body disease	Senile changes, Cerebrovascular disease
DLB2	BSFRC	M	70	6	N.A.	Lewy body Dementia	Senile changes, Cerebrovascular disease
DLB3	BSFRC	M	75	4	N.A.	Lewy body Dementia	Senile changes, Cerebrovascular disease
MSA1	BSFRC	F	62	N.A.	N.A.	Multiple system atrophy	N.A.
MSA2	Ghetti	F	71	N.A.	N.A.	Multiple system atrophy	Senile changes, Cerebrovascular disease
MSA3	Ghetti	M	52	N.A.	3	Multiple system atrophy	Senile changes, Cerebrovascular disease
CBD	Ghetti	F	51	10	9.6	Corticobasal Degeneration	N.A.

Table A.2. Clinical and neuropathological characteristics of synucleinopathy and non-synucleinopathy brain tissue samples used in the study.


Learning Curves (LR)
Learning Curves (GP x LR)
Training score
Cross-validation score
MAE
R2 Score
RMSE
Training examples


Learning Curves (MLP)
Learning Curves (GP x MLP)
Training score
Cross-validation score
MAE
R2 Score
RMSE
Training examples


Learning Curves (RFR)
Learning Curves (GP x RFR)
Training score
Cross-validation score
MAE
R2 Score
RMSE
Training examples
0.0
0.2
0.4
0.6
0.8
1.0
50
100
150
200
250
300

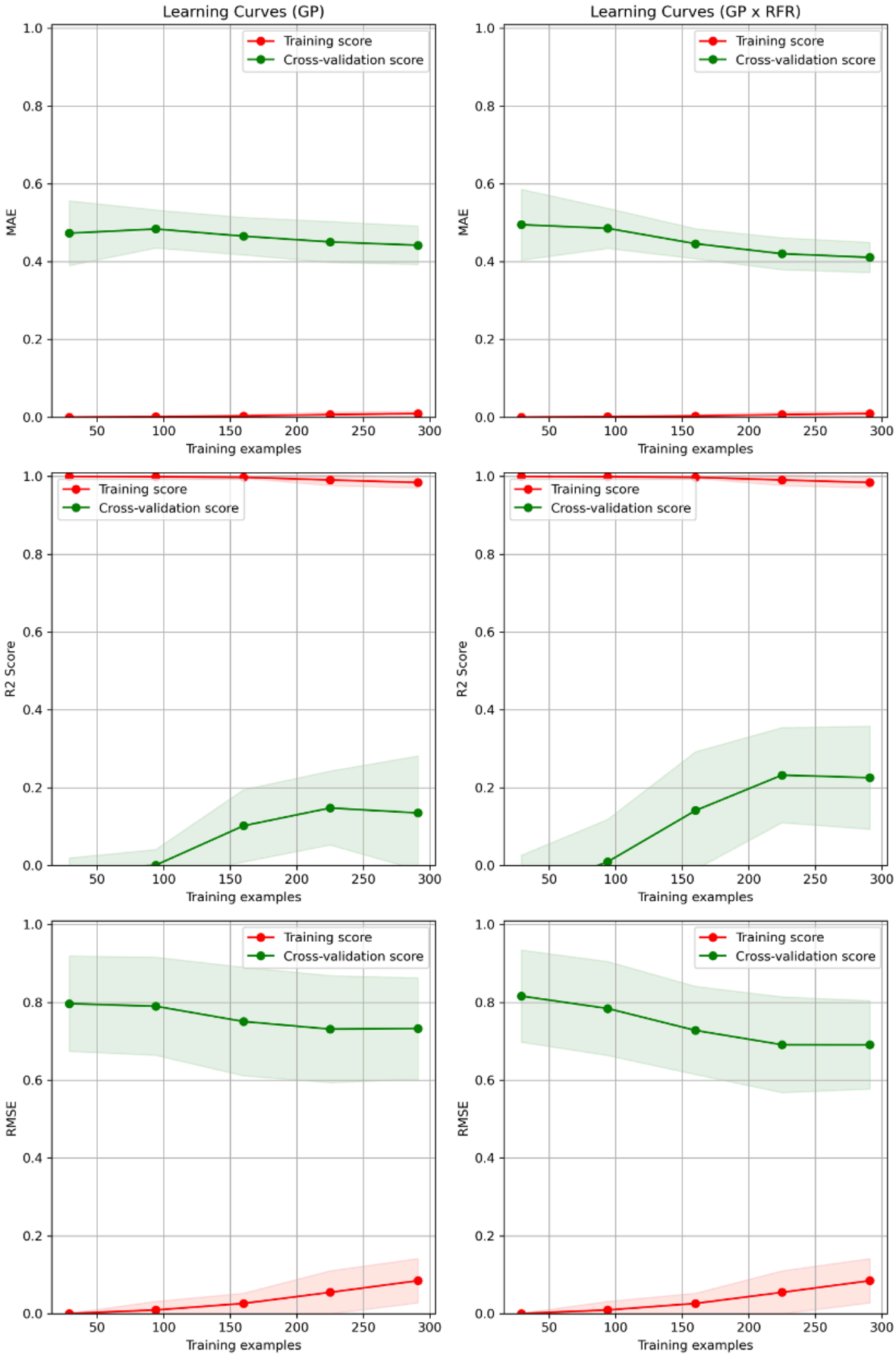


Figure A.1. MAE, RMSE and R^2 for different models trained on the latent features of the variational

autoencoder and the aggregation data. The y-axis reports the respective scoring metric, and the x-axis the number of molecules included in the training set, of a total sample of 360 molecules. In each case, the performance of the model in isolation is shown in the left column, while the performance of the model when used in tandem with the GPR fitted to the residuals of the first model is shown in the right column. The labels are as follows: LR = linear regressor, GP = Gaussian process, MLP = multilayer perceptron, RFR = random forest regressor. Model parameters were chosen using a grid search of possible parameters while cross validating on 5 stratified K folds of the aggregation data, and selecting the parameters that gave the best performance in terms of R^2 score. The parameters for the models shown here are displayed in **Table A.1**.

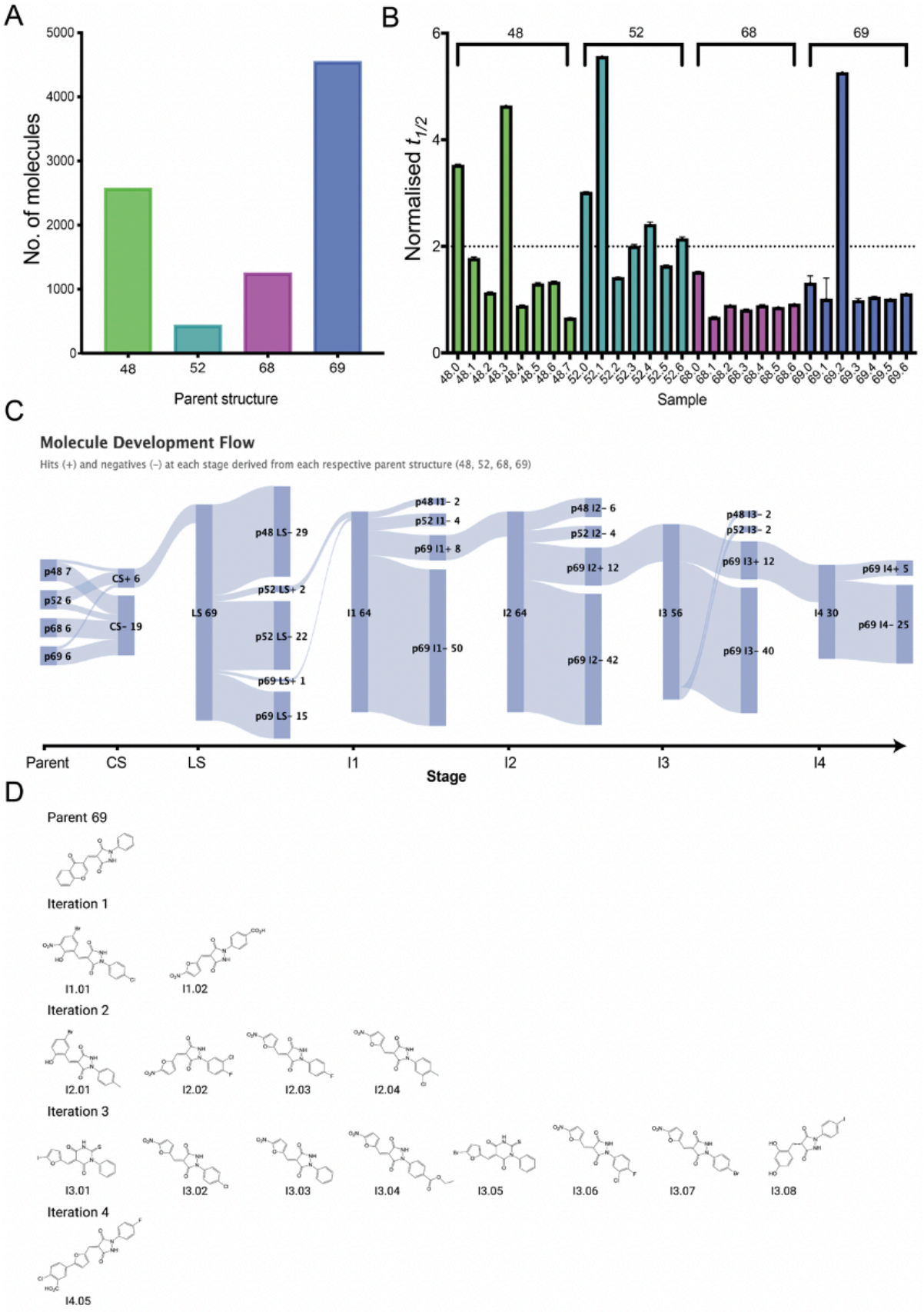


Figure A.2. Summary of the molecules described in this work. (A) Number of molecules derived from 1 of the 4 docking hits (48, 52, 68, 69) within the evaluation set (see **Figure 1**). There were more structures derived from molecules 69 and 48 compared with molecules 68 and 52. **(B)** Normalised half time of aggregation ($t_{1/2}$) for

the 25 molecules in the close similarity docking set (25 μM), i.e. those closely related (Tanimoto similarity > 0.5) to the 4 molecules in the docking set (labelled as 48.0, 52.0, 68.0 and 69.0 on the x-axis). Leads were defined as molecules that more than double $t_{1/2}$, as indicated by the horizontal line that marks 2 times the half time (y-axis) in the absence of the molecules. Some derivatives of molecules 48, 52 and 69 showed good potency, in particular 48.3, 52.1 and 69.2, but these effects were outstripped by future leads such as I4.05 which yielded the same effects at 50 fold lower concentration. **(C)** Flow chart of leads (+) and negatives (-) in the project starting from the close search (CS), moving to the loose search (LS) then iterations 1, 2, 3 and 4 (I1, I2, I3, I4). Each branch is labelled with the molecule source (e.g. parent 48 = p48) whether it was a lead or a negative, and the number of molecules in the branch. Attrition reached its highest point at the loose search before gradually improving with each subsequent iteration. Iteration 4 is included but not directly comparable as a model was trained on the lower dose inhibition for this step. **(D)** Structures of the most potent leads at each stage, which flatlined aggregation at 25 μM, all of which were derived from p69. The structures gradually converged as the core pyrazolidine-3,5-dione structure and RHS aromatic ring were largely retained (with some exceptions for ring expanded derivatives in iteration 3) with addition of electron withdrawing groups to the benzene ring. The LHS was altered more significantly, replacing the parent bicyclic system with substituted furans, which were further elaborated in iteration 4 with an additional benzoic acid group.

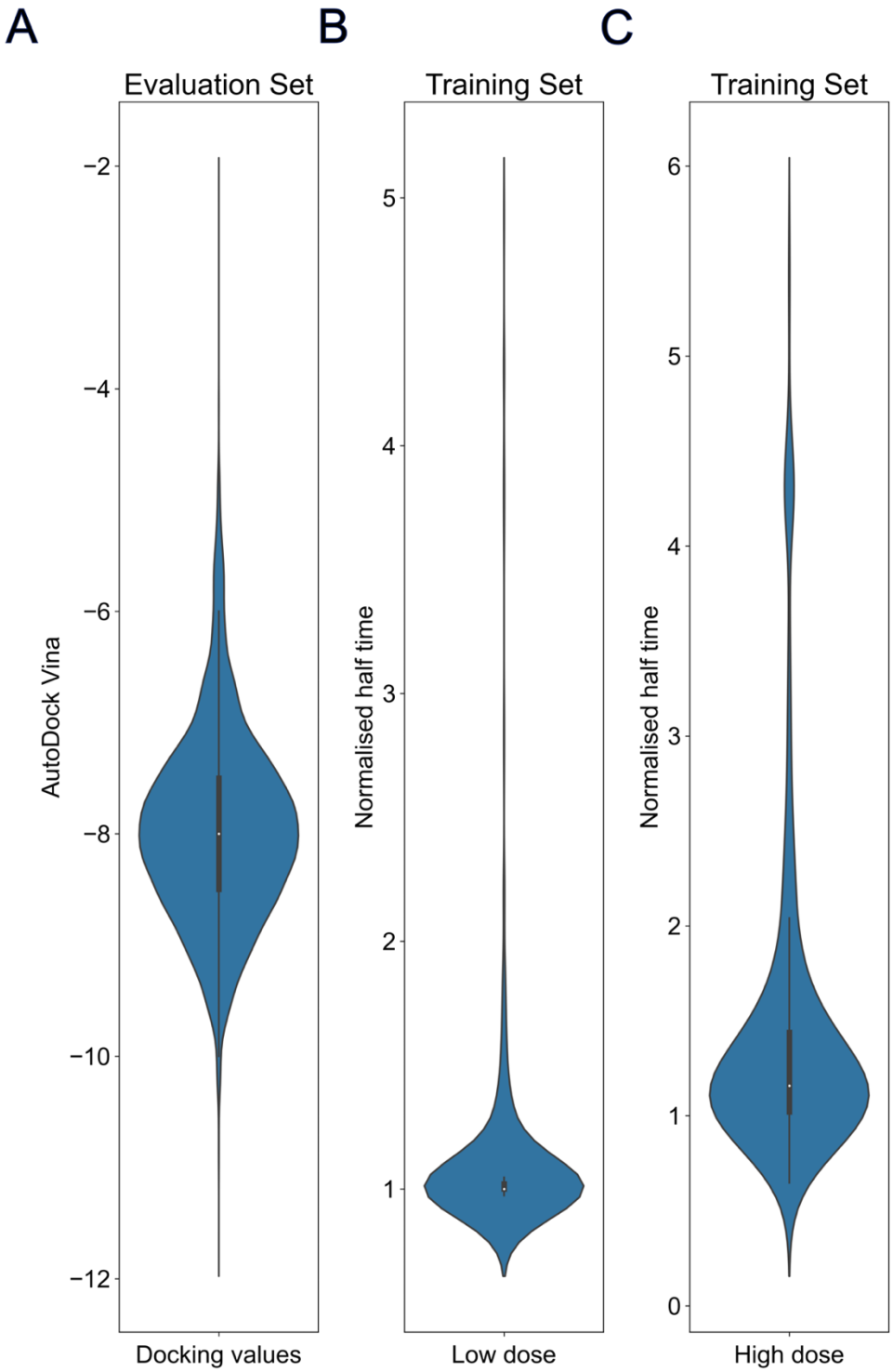


Figure A.3. Distributions of the data sets used. (A) AutoDock Vina binding energies (kcal mol^{-1}) for the evaluation set (~9000 molecules). The values are narrowly distributed between -6 and -10 kcal mol^{-1} as the data set consists of 4 key structures predicted to have good binding. Normalised half times of aggregation at **(B)** 3.12 µM and **(C)** 25 µM for the whole training set (~400 molecules), including docking molecules and initial similarity searches and after all iterations had been added. The high dose was used for training in iterations 1-3 and the low dose for iteration 4.

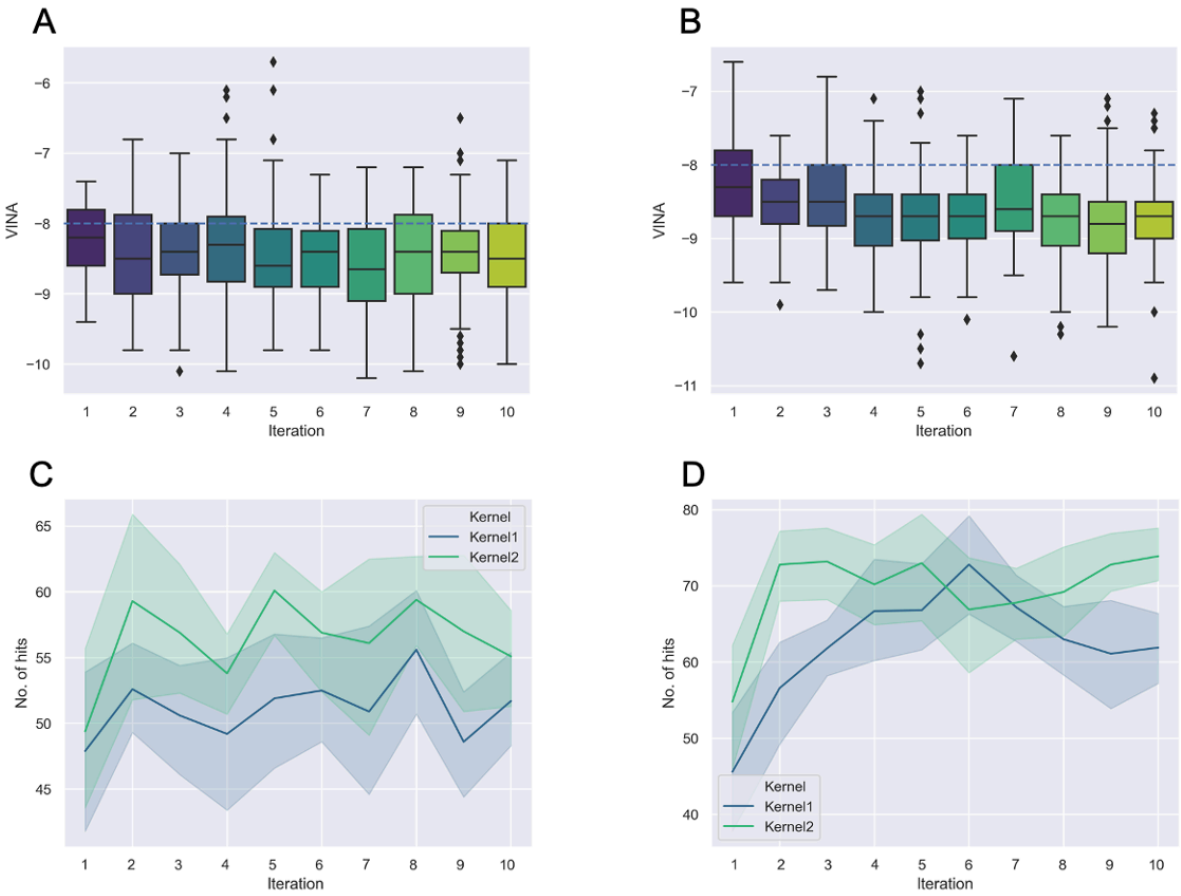


Figure A.4. A simulation of the experimental scenario using docking energies as a proxy for aggregation experiments. (A) Starting from a single random sample, the GP with RBF kernel was tested. AutoDock Vina binding energies in kcal mol^{-1} are plotted against iteration number. Each boxplot visualises the distribution of binding scores for the top 100 molecules predicted by the algorithm at each iteration. The dotted line indicates the mean binding energy of the test set. **(B)** Same process as in panel A, but employing the GP with a Matérn Kernel. **(C)** Aggregated average number of hits out of the top 100 predicted molecules from 10 different random starts of the process shown in panels A and B for the RBF kernel (Kernel 1, in blue) and the Matérn kernel (Kernel 2, in green). A hit was taken as a molecule falling in the lower quartile of the test set distribution (<-9 kcal/mol). Results were obtained using the half-length representation of the molecules. **(D)** Same process as described in panel C, but employing the full-length molecule representation.

Figure A.5. Performance of the RFR method coupled to the Matérn kernel compared to the Matérn kernel alone. (A) R^2 score with increasing training set size (up to 4000) for both models, using the full-length representation. On the left is the GP with Matérn kernel alone, and on the right is the GP with Matérn kernel + RFR. Cross validation with 10 random shuffle splits and 20% of the data randomly selected as a validation set. **(B)** Aggregated average hit data from 10 different random starts of the experimental simulation for the iterative approach, starting from 100 randomly selected molecules and successively adding the actual docking data of the predicted top 100 hits to the training set with each iteration. GP with Matérn kernel alone (Kernel 2 = Matérn) vs GP with Matérn kernel + RFR. **(C)** Average Pearson's correlation coefficient (pcorr) between the predicted binding score values and the real scores at each iteration.

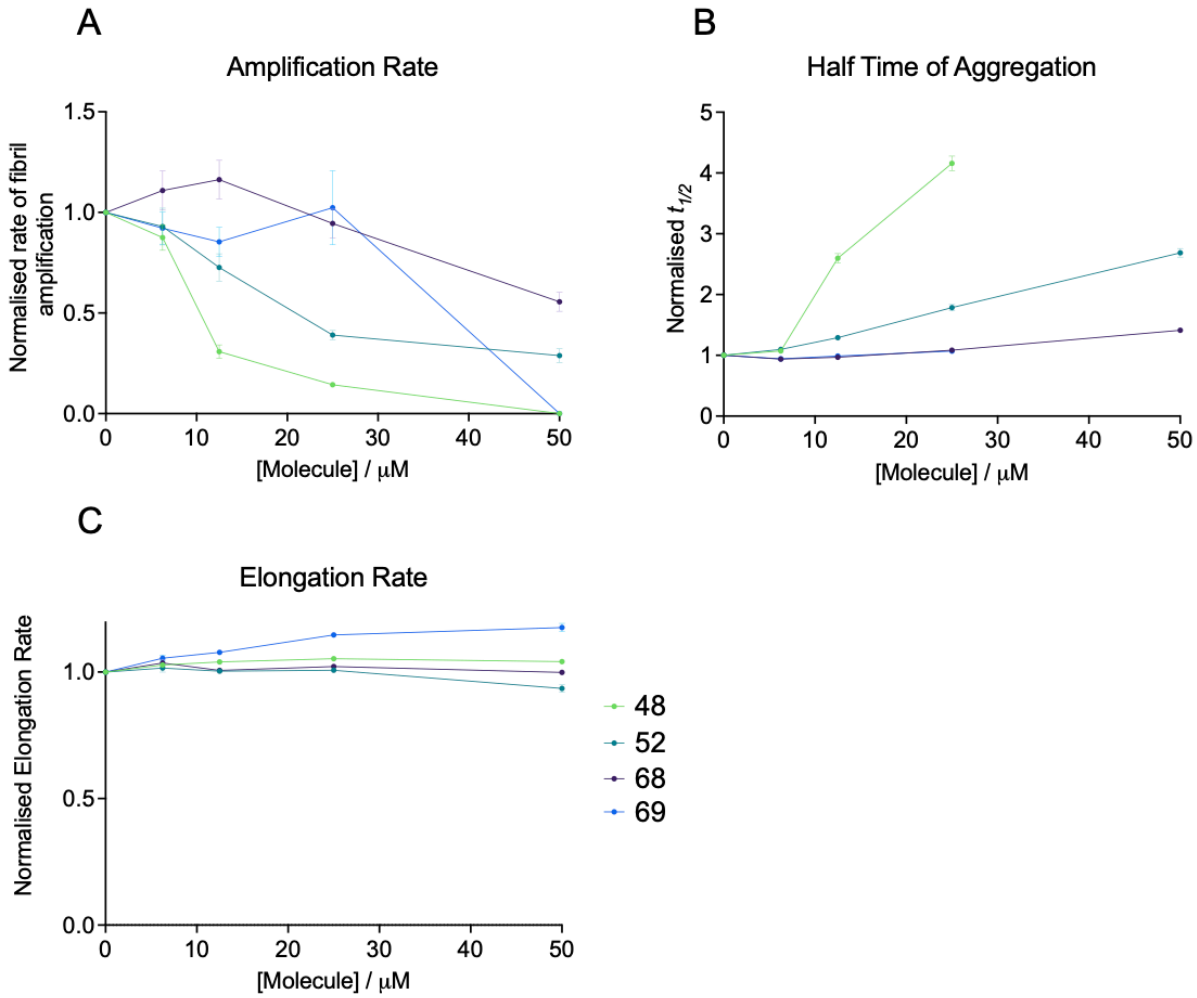
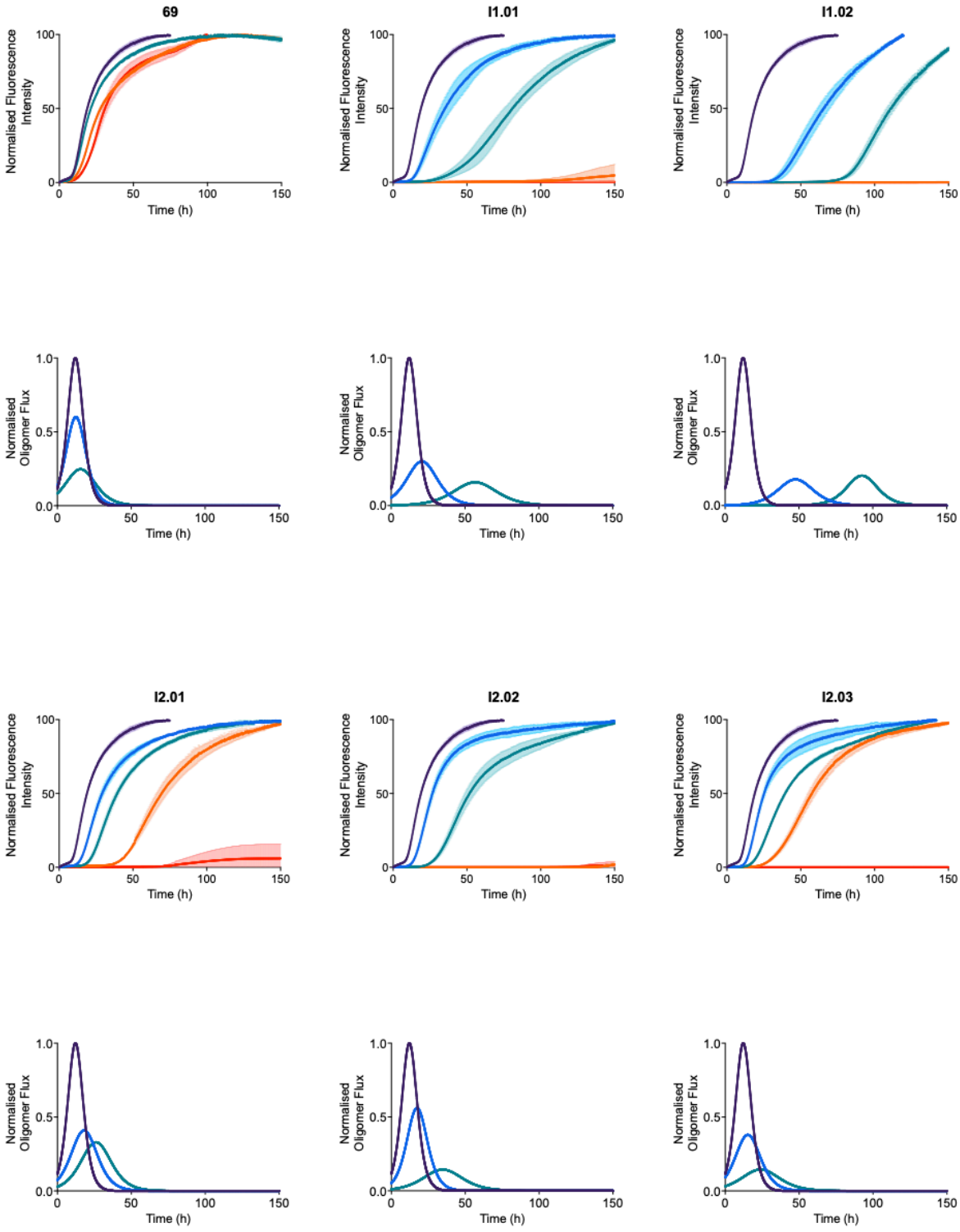
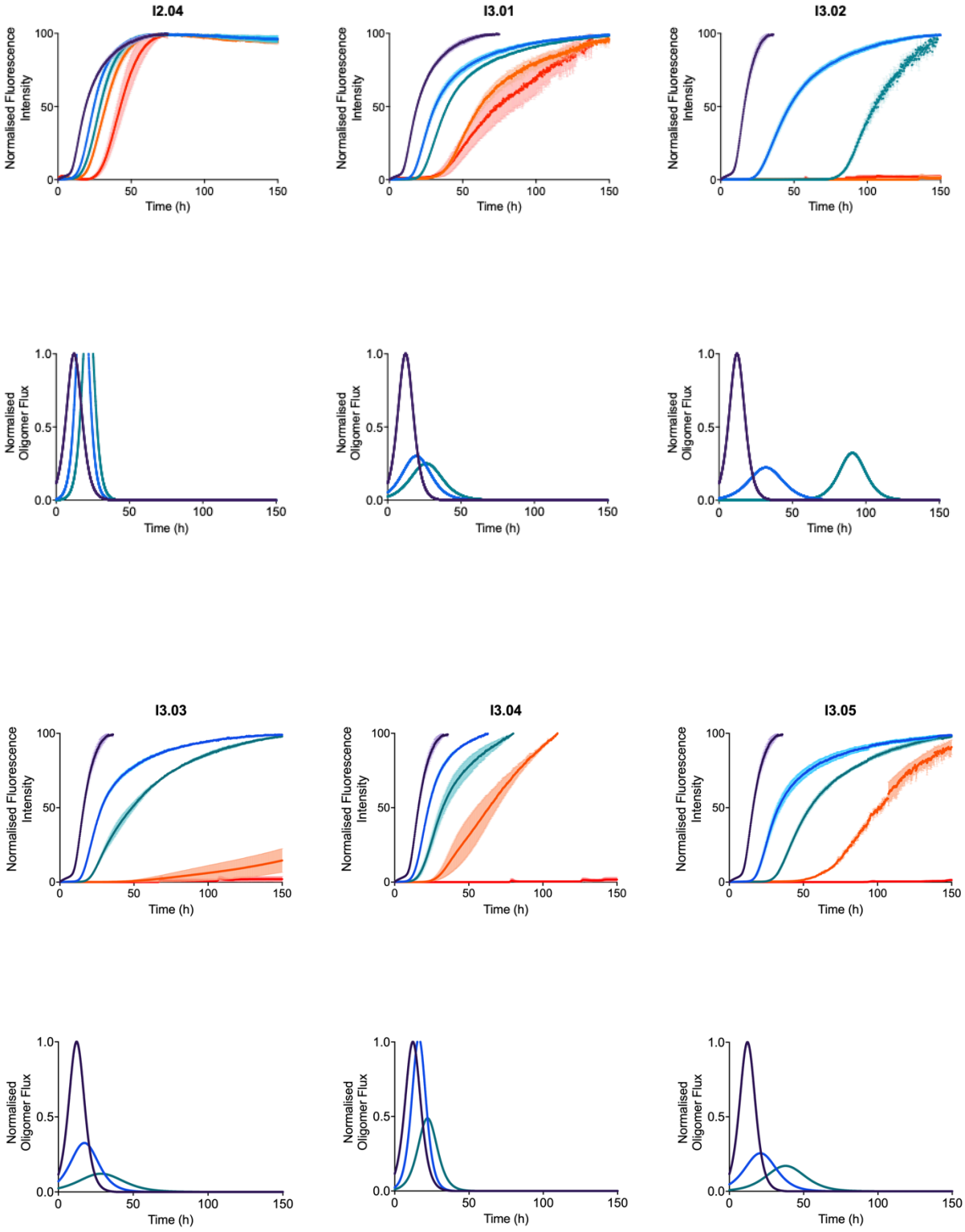


Figure A.6. Amplification rate and half time of aggregation of αS in the presence of the 4 molecules in the docking set. (A) Relative rate of fibril amplification of αS in the presence of the 4 docking molecules (labelled as 48, 52, 68 and 69) in the docking set; the kinetic traces are normalised to the DMSO control. **(B)** Half times of aggregation derived from the same experiment. (**C**) Relative rate of fibril elongation normalised to the DMSO control. The amplification rate (A) and half time of aggregation (B) were tested in the machine learning method as parameters to describe the potency of a molecule. The amplification rate tends to be more affected by perturbations to the early slope of the exponential phase can have large effects on the derived rate value. The half time, although a simpler measure, is more robust and so was chosen for the machine learning approach. Data obtained from reference[64].


69
I1.01
I1.02
I2.01
I2.02
I2.03
Normalised Fluorescence Intensity
Normalised Oligomer Flux
Time (h)
0
50
100
150
0.0
0.5
1.0


I2.04
I3.01
I3.02
Normalised Fluorescence Intensity
Normalised Oligomer Flux
Time (h)
I3.03
I3.04
I3.05

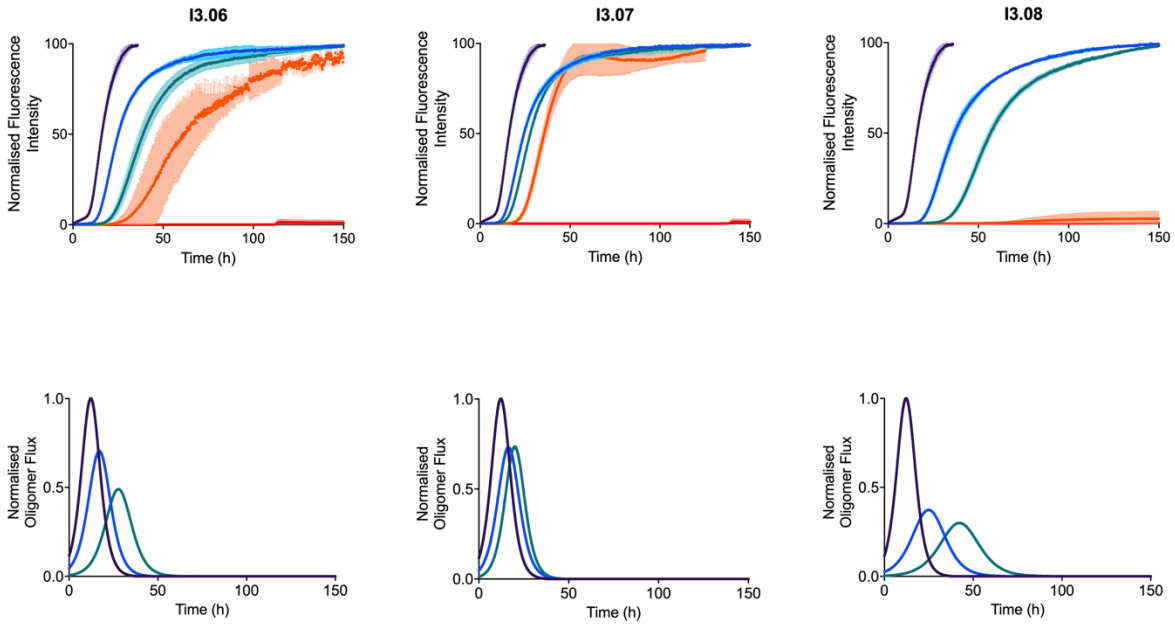
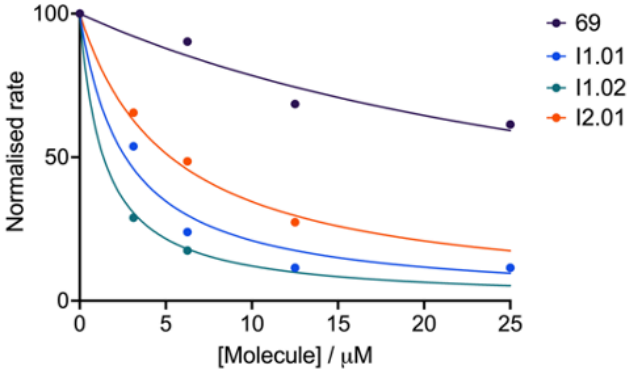
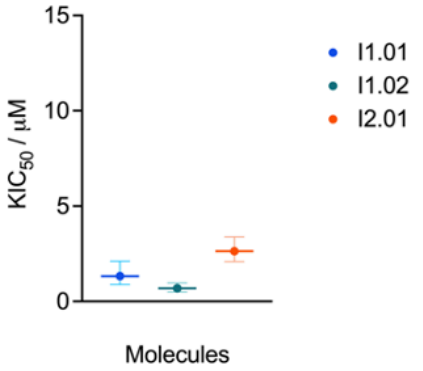
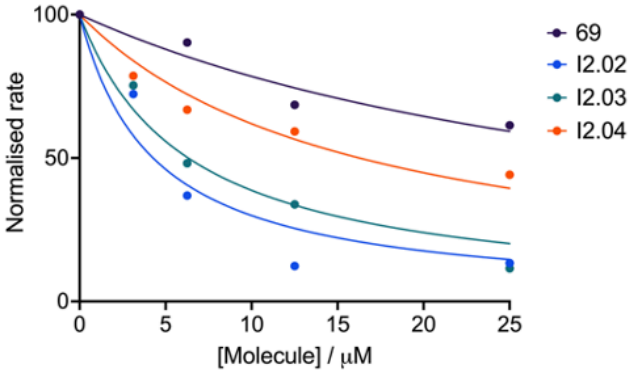
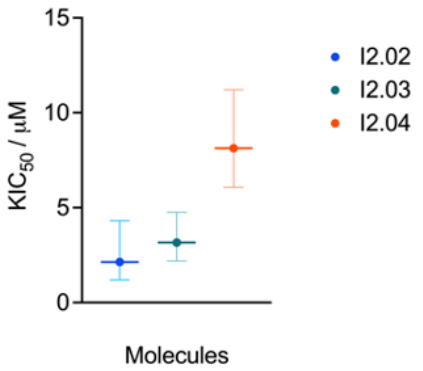


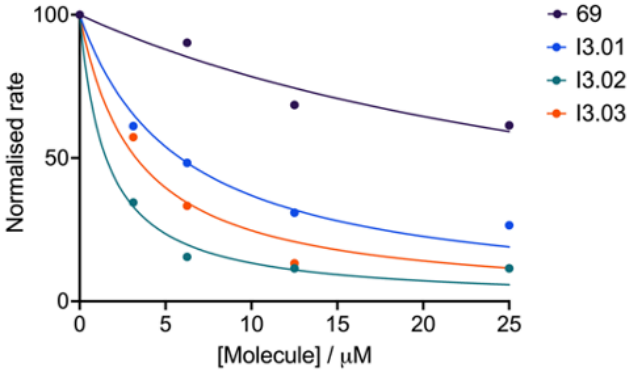
Figure A.7. Aggregation curves (top) and oligomer flux simulations (bottom) for the most potent compounds from all of the iterations. The kinetic traces show a 10 µM solution of αS in the presence of 25 nM seeds at pH 4.8, 37 ºC in the presence of molecules at 3.12 µM (blue), 6.25 µM (teal), 12.5 µM (orange) and 25 µM (red) versus 1% DMSO alone (dark purple), with endpoints normalised to the αS monomer concentration detected via the Pierce™ BCA Protein Assay at the end of the experiment. Oligomer simulations were carried out only for the lower 2 concentrations, as full aggregation curves were only consistently obtained for all molecules in the secondary nucleation assay at these concentrations.

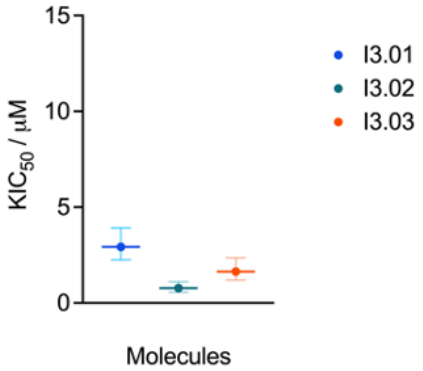

Normalised rate
100
50
0
0
5
10
15
20
25
[Molecule] / µM
69
I1.01
I1.02
I2.01


KIC50 / µM
15
10
5
0
Molecules
I1.01
I1.02
I2.01


Normalised rate
100
50
0
0
5
10
15
20
25
[Molecule] / µM
69
I2.02
I2.03
I2.04


KIC50 / µM
15
10
5
0
Molecules
I2.02
I2.03
I2.04


Normalised rate
100
50
0
0
5
10
15
20
25
[Molecule] / µM
69
I3.01
I3.02
I3.03


KIC50 / µM
15
10
5
0
Molecules
I3.01
I3.02
I3.03

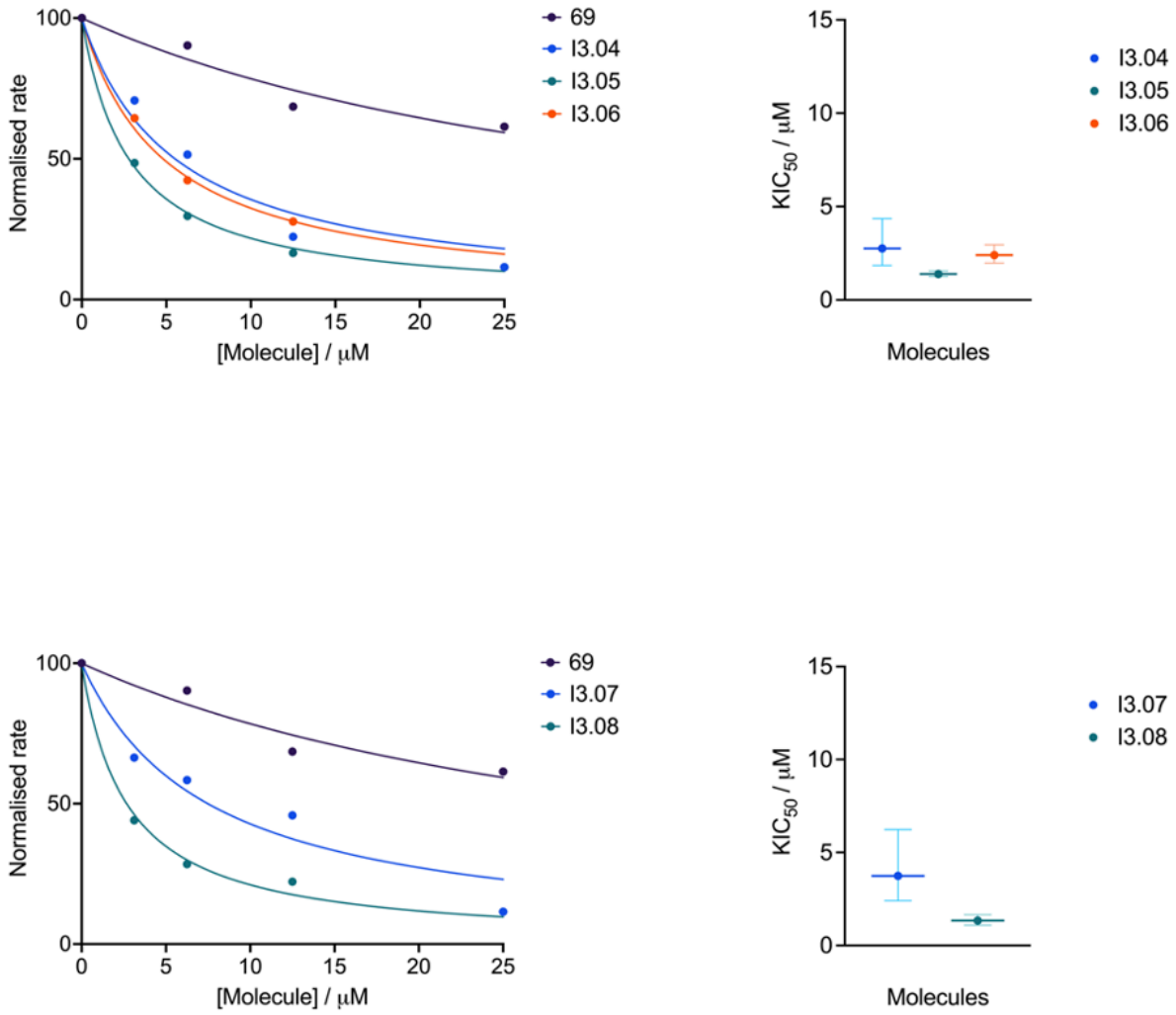


Figure A.8. Concentration dependence of the reaction rate and corresponding 50% kinetic inhibitory concentration (KIC_{50}) values for the most potent compounds. The approximate normalised rate of reaction (taken as $1/t_{1/2}$) is shown on the left for each molecule at each concentration for which a half time could be obtained. For molecules that completely inhibited the aggregation process on the timescale of the experiment, the $t_{1/2}$ in the presence of the highest concentration of molecule (25 μM) was taken to be the length of the experiment. The approximate rates are fitted using an [Inhibitor] vs. normalised response Hill slope. The KIC_{50} values are shown on the right with the 95% confidence interval.

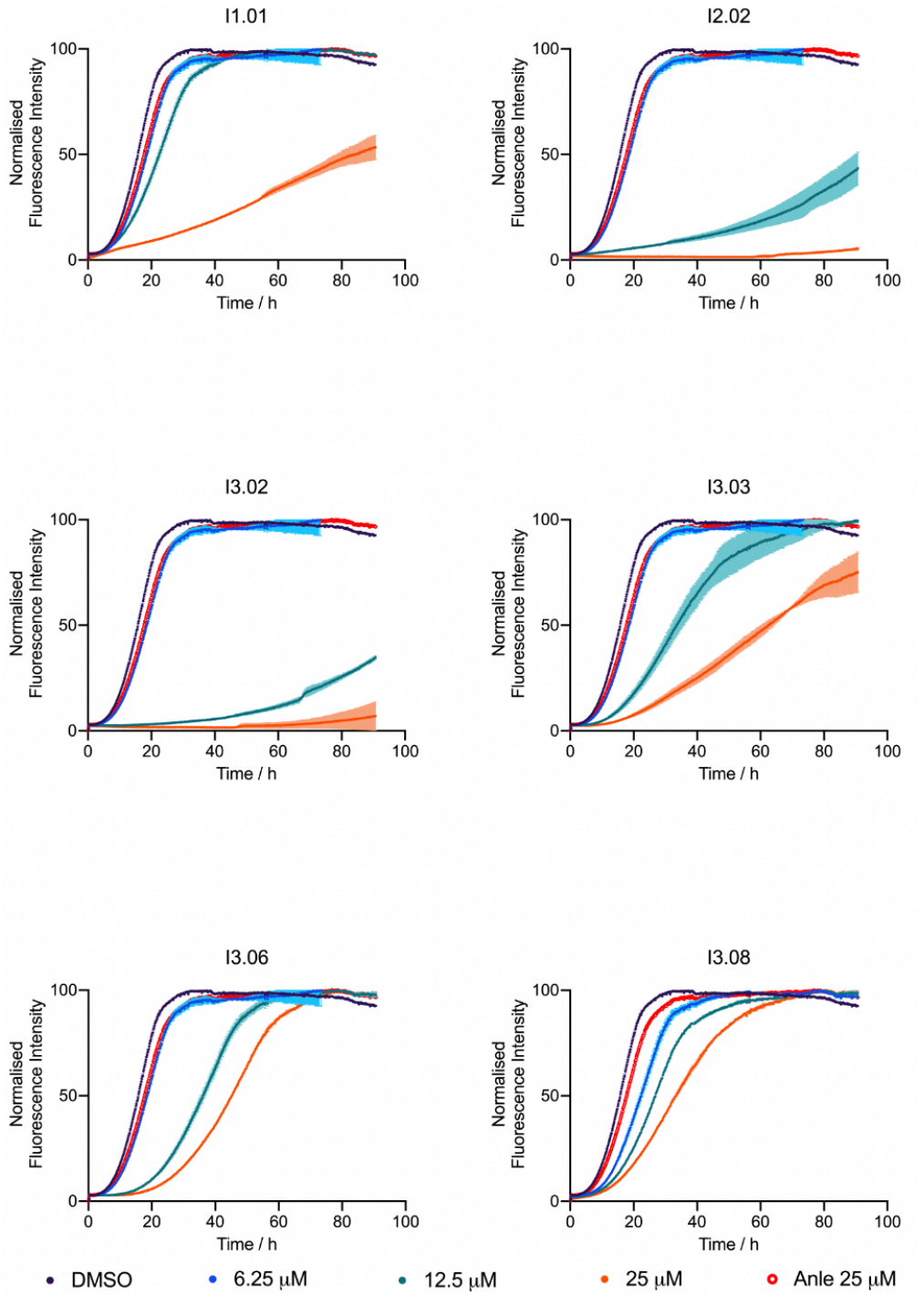


Figure A.9. Lipid induced aggregation curves in the presence of the early leads from the project. The kinetic traces show a 20 µM solution of αS in the presence of 100 µM DMPS vesicles (monomer equivalent) at pH 6.5, 30 ºC in the presence of molecules at 6.25 µM (blue), 12.5 µM (teal), 25 µM (orange) and Anle-138b at 25 µM (red circles) versus 1% DMSO alone (dark purple), with endpoints normalised to the αS monomer concentration detected via the Pierce™ BCA Protein Assay at the end of the experiment.

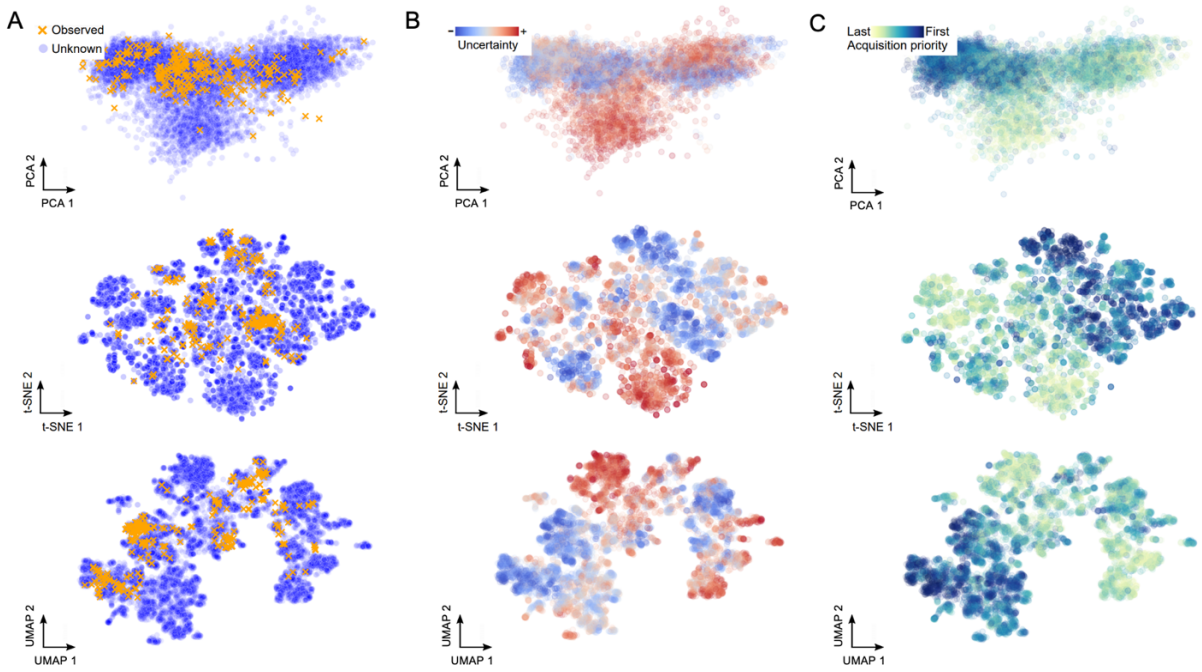


Figure A.10. PCA, t-SNE and UMAP visualisations of the compound feature space using uncertainty. (A) From top to bottom: PCA, t-SNE and UMAP visualisations of the compound space indicating which areas of the chemical space have been explored (orange crosses) and which have not (blue circles). **(B)** GPR assigned lower uncertainty (blue) to regions of the chemical space near to the observed data and high uncertainty (red) to areas which were further away. **(C)** The lower uncertainty compounds were prioritised (dark blue) during acquirement ranking.

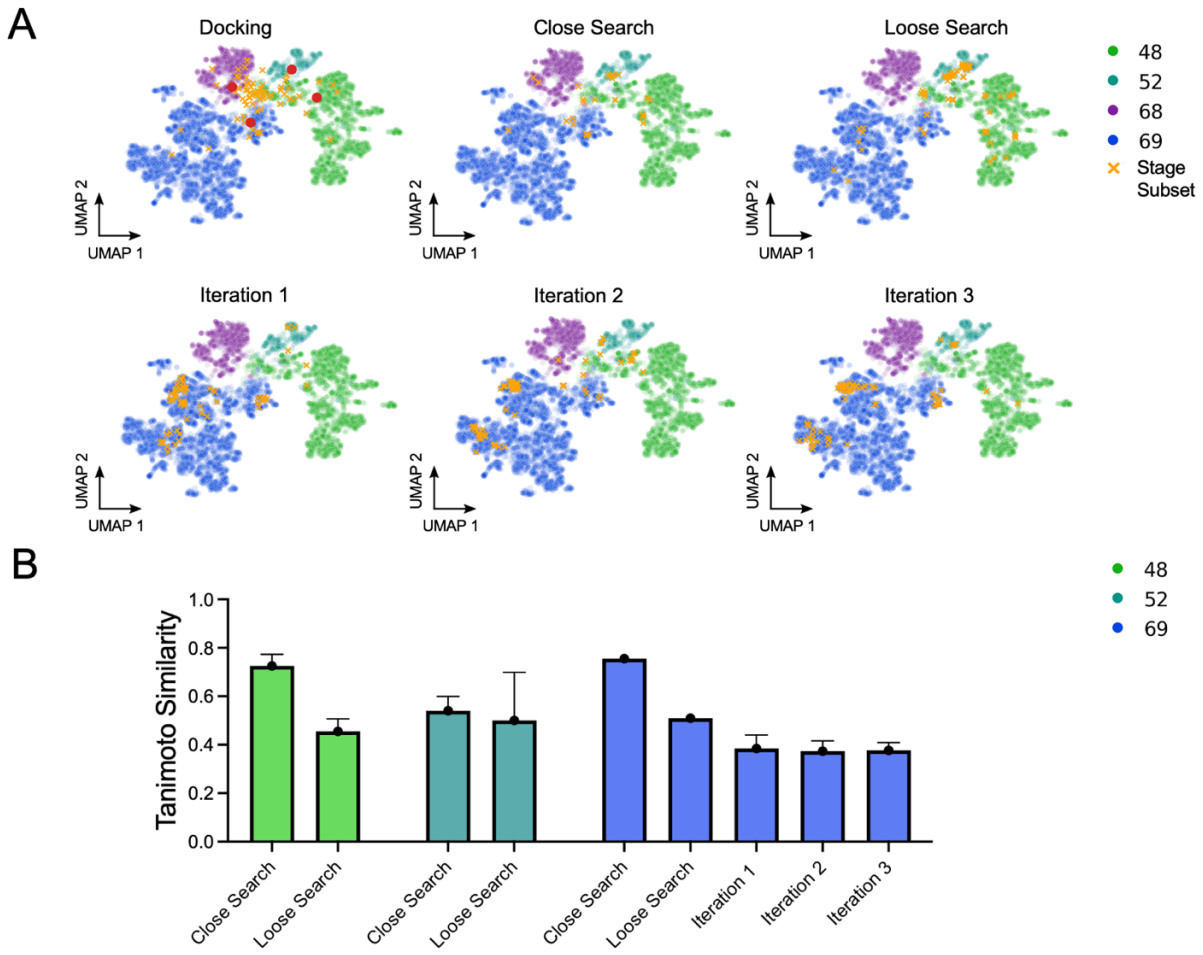


Figure A.11. Analysis of the structural changes in the compound optimisation. (A) UMAP visualisation of the compound space indicating how the positioning of each new molecule subset (orange crosses) changed at each stage of the project as well as how the chemical landscape was split between the parent molecules (different colours) The locations of the parent molecules are also indicated in the 'Docking' pane (red circles). **(B)** Average Tanimoto similarity of the lead molecules to their respective parents at each stage of the project. At iterations 1, 2 and 3 all of the leads were derived from molecule 69, albeit with lower similarity than any of the previous stages. Molecule 68 failed to produce any leads.

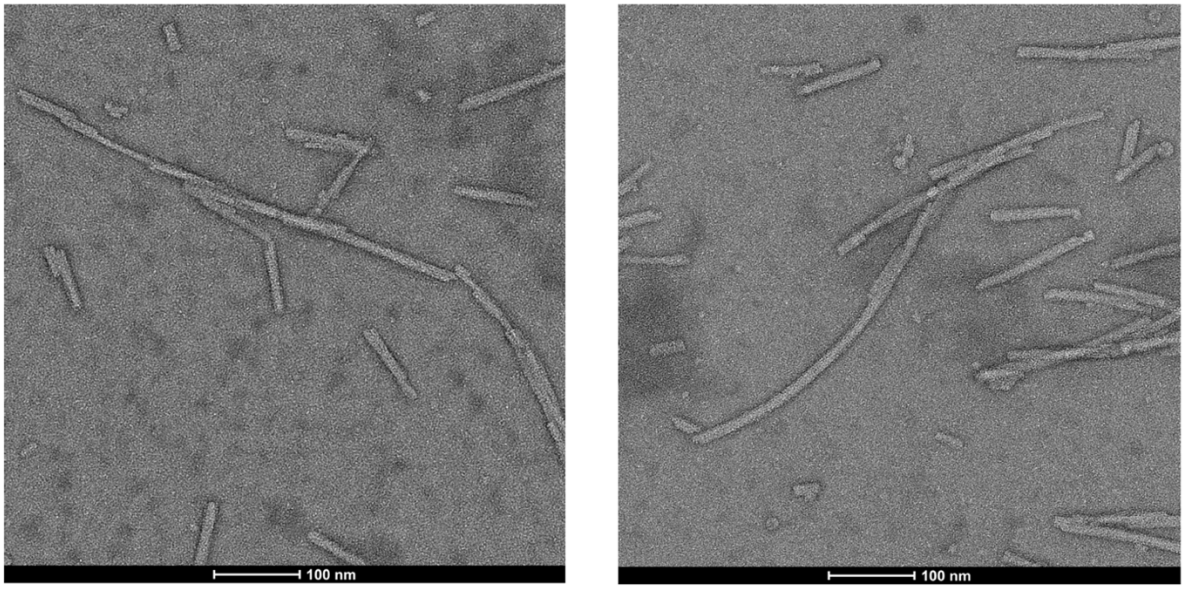


Figure A.12. Transmission electron microscopy images of the fibrils at the end of the secondary nucleation assay. Two representative images are shown, the scale bar is 100 nm.

B. Exploration and exploitation approaches based on generative learning

Number of estimators	1800
Minimum samples split	5
Minimum samples at a leaf node	1
Maximum depth	70
Bootstrap	False

Table B.1. RF parameters used during genetic algorithm selection

Maximum number of features in subset	5
No. of individuals is starting population	100
Probability of crossover	0.5
Probability of mutation	0.2
Number of generations	50

Table B.2. Genetic algorithm parameters. Computational work was carried out by Mhd Hussein Murtada under my supervision.

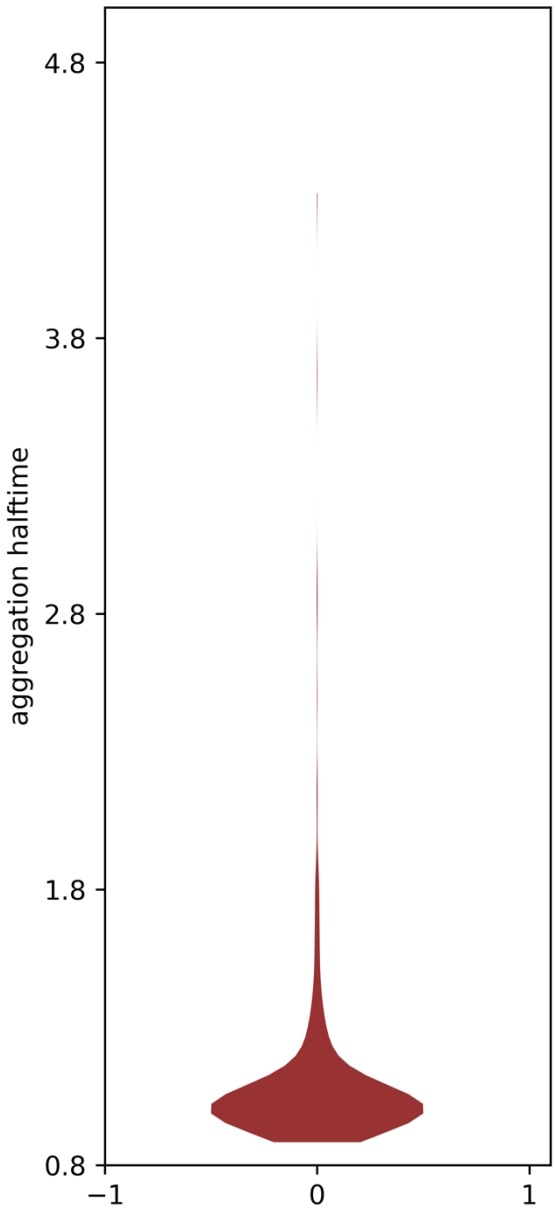


Figure B.1. Distribution of normalised aggregation half times in the αS aggregation inhibitor data set. The data set of known aggregation inhibitors was unbalanced towards having many more inactive than active compounds.

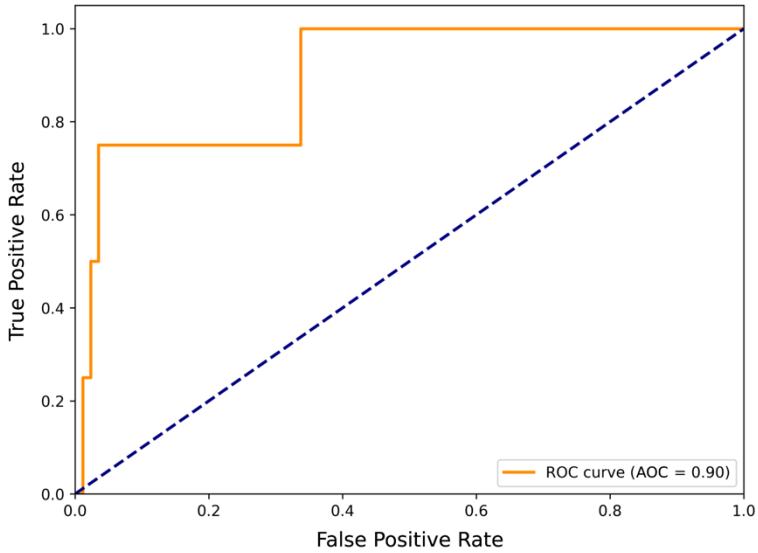


Figure B.2. ROC AUC curve for the SMILES embedding model initially appeared promising with an AUC of 0.9. Computational work was carried out by Mhd Hussein Murtada under my supervision.

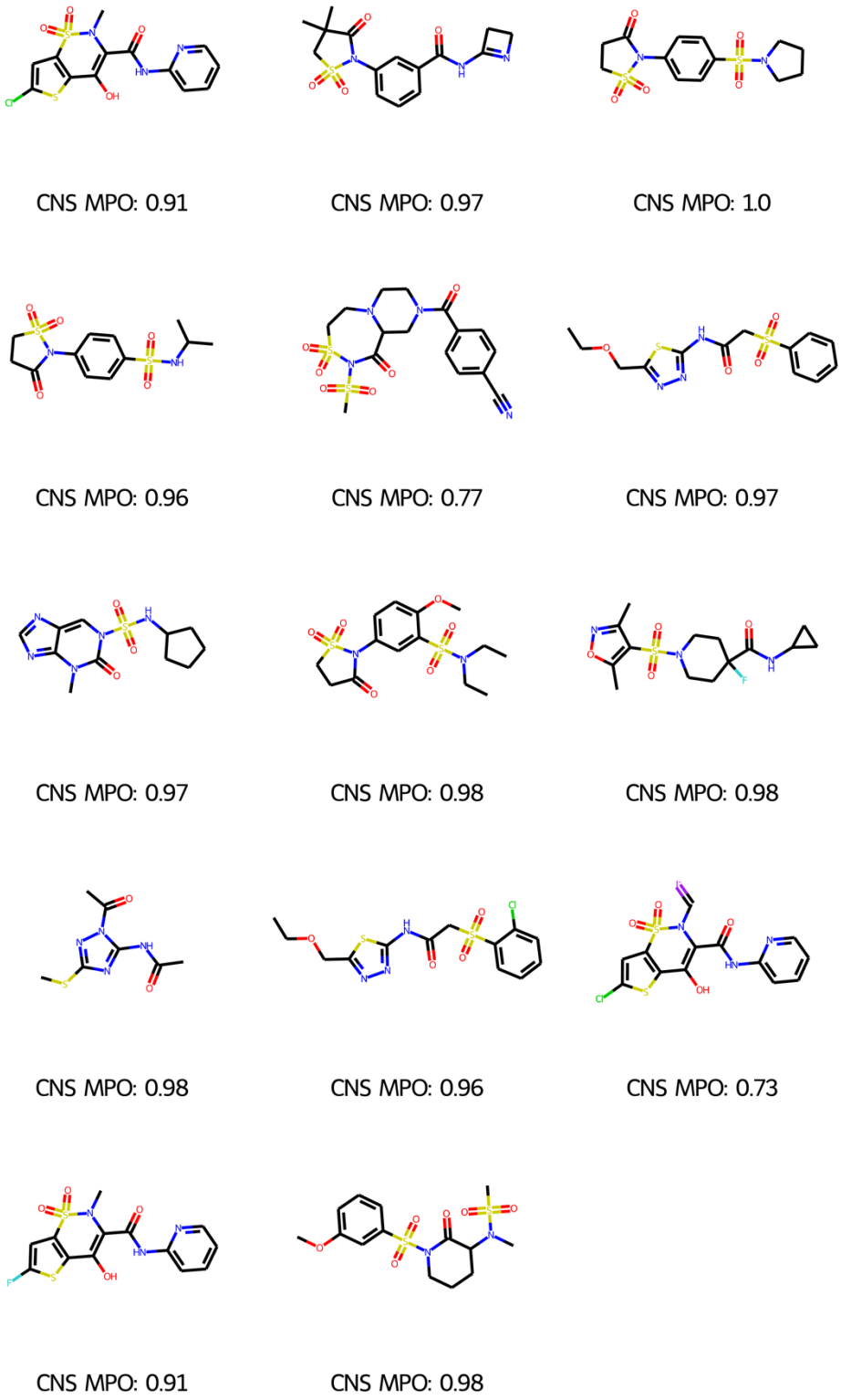


Figure B.3. Structures generated by the final pipeline and their respective calculated CNS MPO scores. Computational work was carried out by Mhd Hussein Murtada under my supervision.

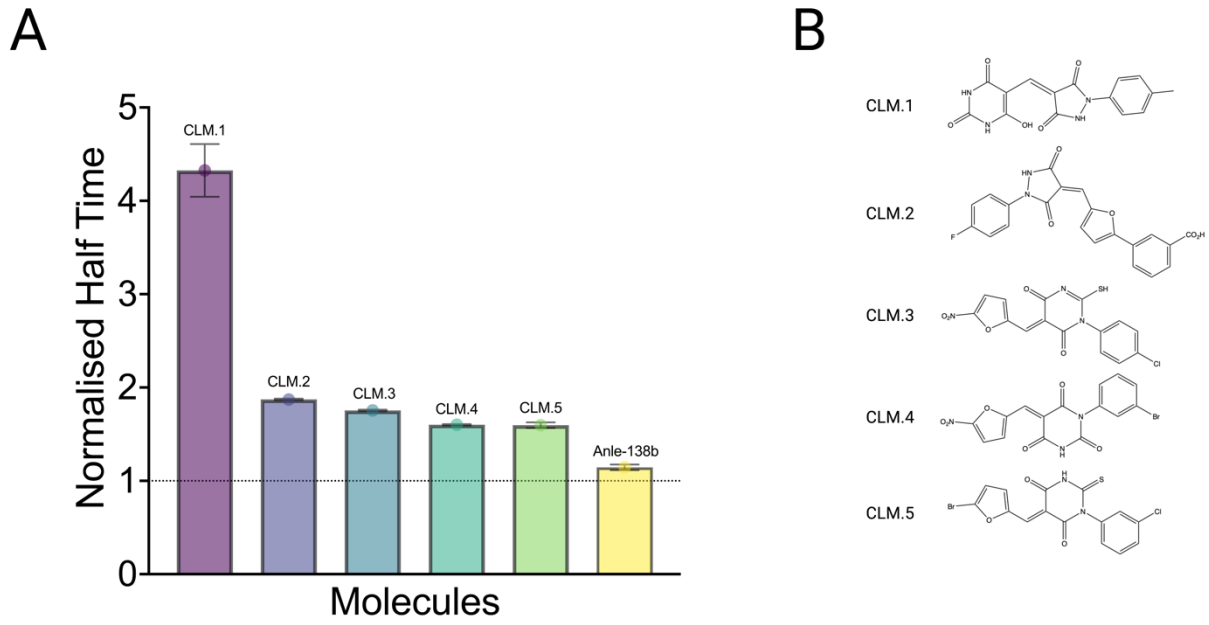


Figure B.4. Aggregation data from the new leads generated via the exploitation pipeline. (A) Normalised half times for a 10 µM solution of αS with 25 nM seeds at pH 4.8, 37°C in the presence of CLM generated molecules at 3.12 µM. The horizontal dotted line indicates the normalised half time of a 1% DMSO negative control. Anle-138b at 25 µM is also shown for comparison. (B) Structures of the CLM generated molecules.

C. Developing nanopore oligomer detection methods

i. PAGE gel

Polyacrylamide gels (10% v/v, with 0.5X, pH 8, Tris-Borate-EDTA and 11 mM $MgCl_2$) were hand-cast on a PAGE loading gel setup. Details on the PAGE gels recipes can be found in **Table C.1**. Once all the gel mixture additions (**Table C.1**) were mixed together, 1% (w/v) APS and 0.07% (w/v) TEMED were added and the mixture was immediately vortexed and poured in between two PAGE glass slides using a Pasteur pipette. The gel comb was promptly inserted and the gel was left to polymerise for at least 45 minutes and a maximum of 1 hour. Gels were run for 120 minutes at 100 V in a running buffer containing 0.5x TBE and 11 mM $MgCl_2$. DNA was stained using GelRed® (Biotium) for 15 minutes under constant shaking. Imaging was performed using a Gel-DocIt imaging system by UPV using Visionworks software under UV excitation light and exposure times varying from 5 to 10 seconds. To check for the binding of the copper-free click chemistry reaction between DBCO-DNA and azide-labelled αS and secondarily to optimise the reaction conditions, DBCO-DNA was incubated in a 1:1 ratio (monomer : DBCO) for 1h, 3h, and overnight **(Figure 4.6)**.

Gel mixture addition	**Amount [mL]**	**Final concentration**
Acrylamide/bis-acrylamide, 30 % solution	5	10 %
10x Tris-Borate-EDTA (TBE)	0.75	0.5x
0.5 M $MgCl_2$	0.33	11 mM $MgCl_2$
MilliQ water	8.92	
Total volume	15	
Polymerisation initiators		
10 % (w/v) ammonium persulfate (APS)	0.15	0.1 %
N,N,N',N'-Tetramethylethylendiamine (TEMED)	0.01	0.07 %

Table C.1: Recipe for a 10% PAGE gel.

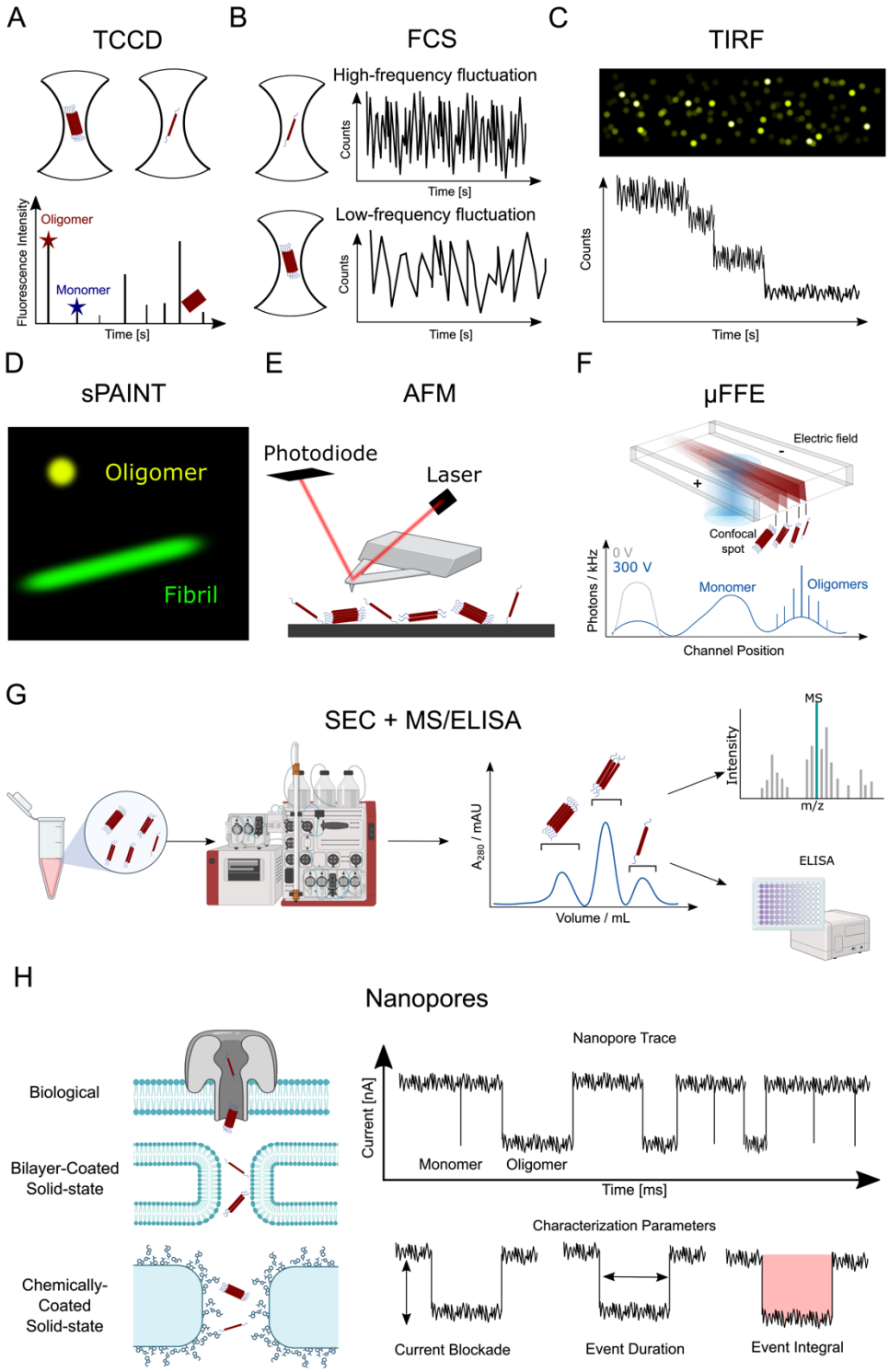


Figure C.1. Methods to characterise oligomers include **(A)** confocal two-colour coincidence detection (TCCD)[168], **(B)** fluorescence correlation spectroscopy (FCS) measurements[169], **(C)** single molecule total internal reflection fluorescence (TIRF) imaging[170], **(D)** single-molecule spectrally-resolved points accumulation for imaging in nanoscale topography (sPAINT)[171], **(E)** atomic force microscopy (AFM)[172], and **(F)** micro free flow electrophoresis (µFFE)[116]. **(G)** Size exclusion chromatography followed by mass spectrometry or enzyme linked immunosorbent assays (SEC/MS or ELISA) provide a bulk method to probing oligomer levels[36]. **(H)** Additionally, biological nanopores[241] have shown to be useful for characterizing protein sizes, as well as lipid bilayer-coated[179] and chemically-coated tween-20 nanopores[181].

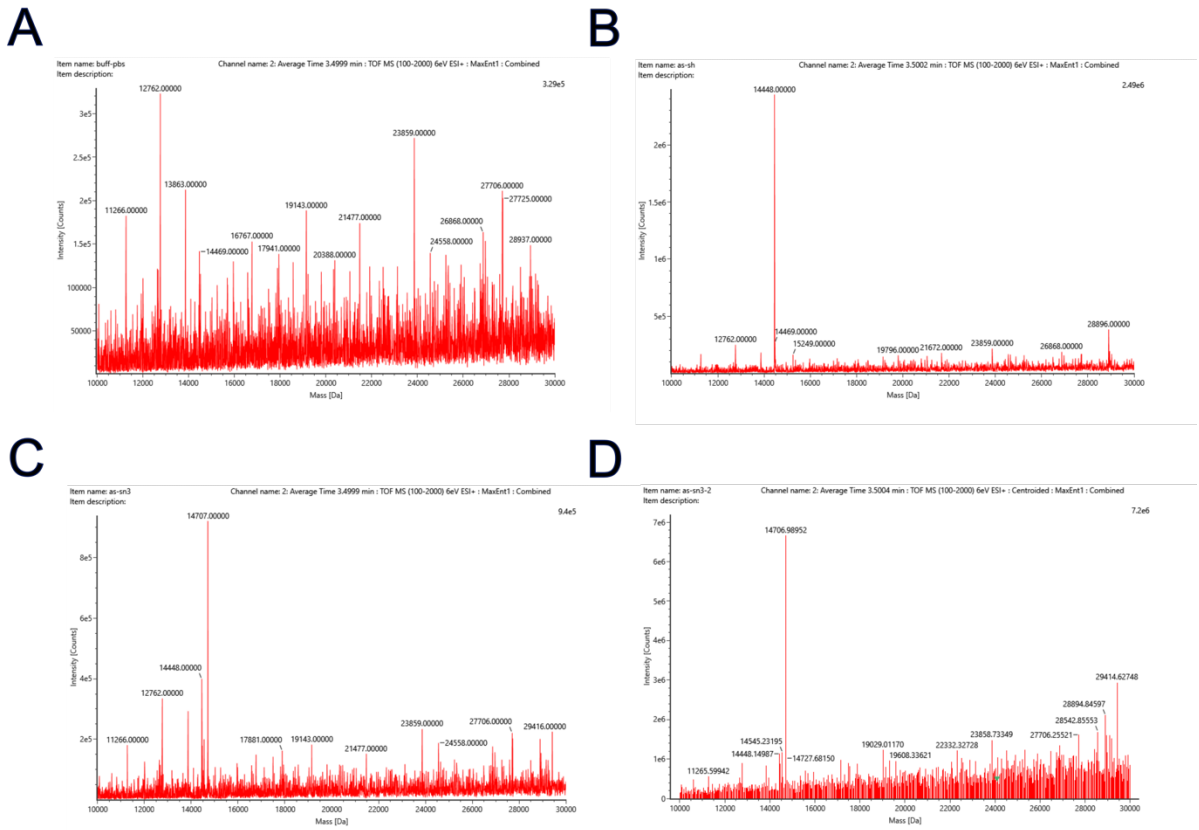


Figure C.2. LC-MS data for the reaction progress of the azide linking step to N122C-αS. (A) PBS buffer control. (**B**) N122C (1 µM, PBS buffer) after reduction with TCEP to remove dimers, showing a single peak at 14448 Da. (**C**) Reduced N122C (1 µM, PBS) after a 2 h incubation with iodoacetamide-PEG$_3$-azide. The labelled peak, at 14707 Da, is prominent but residual unlabelled N122C remains. (**D**) After 3 h almost all of the monomer has been labelled.

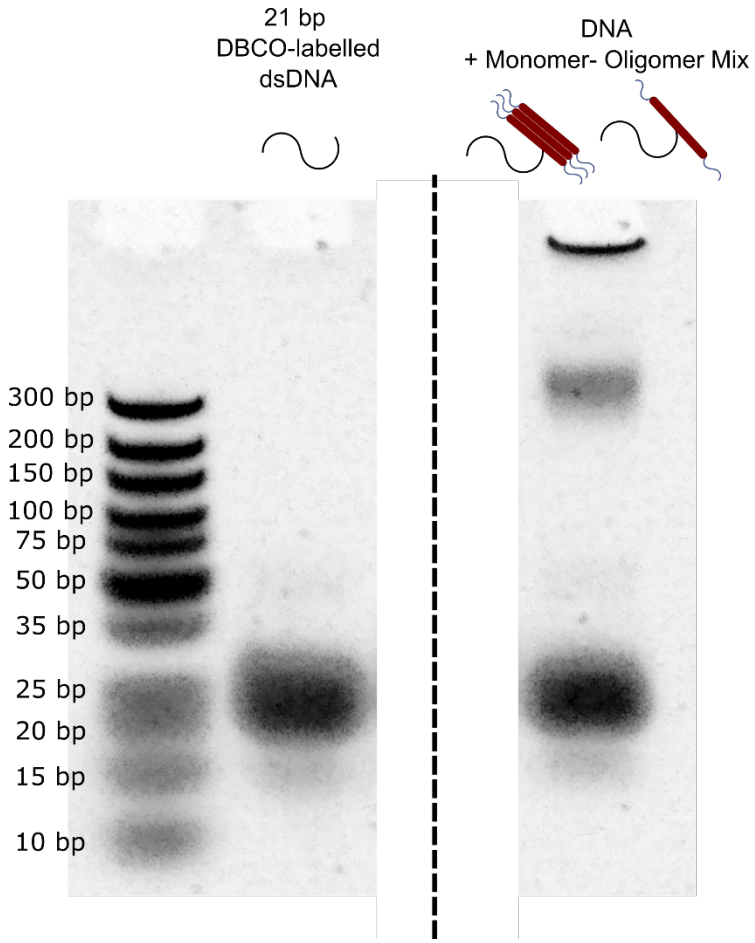


Figure C.3. Monomer-bound DNA observed using PAGE. Column 1 shows 21 bp dsDNA and Column 2 shows a mixture of 21 bp dsDNA mixed with a partially converted oligomer and monomer sample. The monomer is 14 kDa (10 kDa ~ 270 bp), which matches the strongly stained band when added to the 21 bp DNA. The DNA retained in the well may be due to aggregates formed in the sample that cannot enter the gel. Gel was run by Sara Rocchetti.

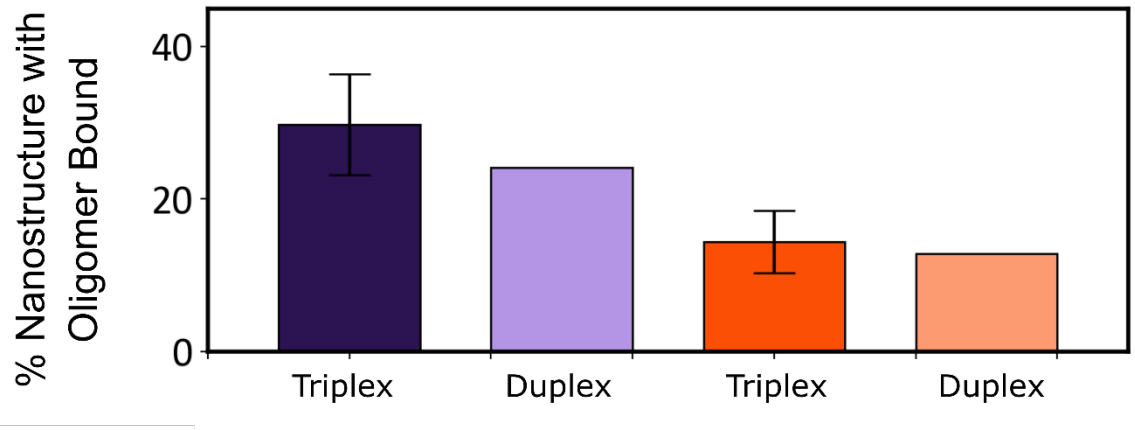


Figure C.4. Comparison of Duplex and Triplex measurements. The samples that were measured in duplex show similar % of nanostructure with protein bound highlighting that monomer interchange is unlikely. The fraction of events with an oligomer bound to the DNA barcode; triplexed DMSO (purple) (N= 114, SD=6.62), duplexed DMSO light purple (N=54), triplexed I3.08 (orange) (N=90, SD=4.07), duplexed I3.08 (light orange) (N=39). Nanopore experiments were carried out by Sarah E. Sandler.

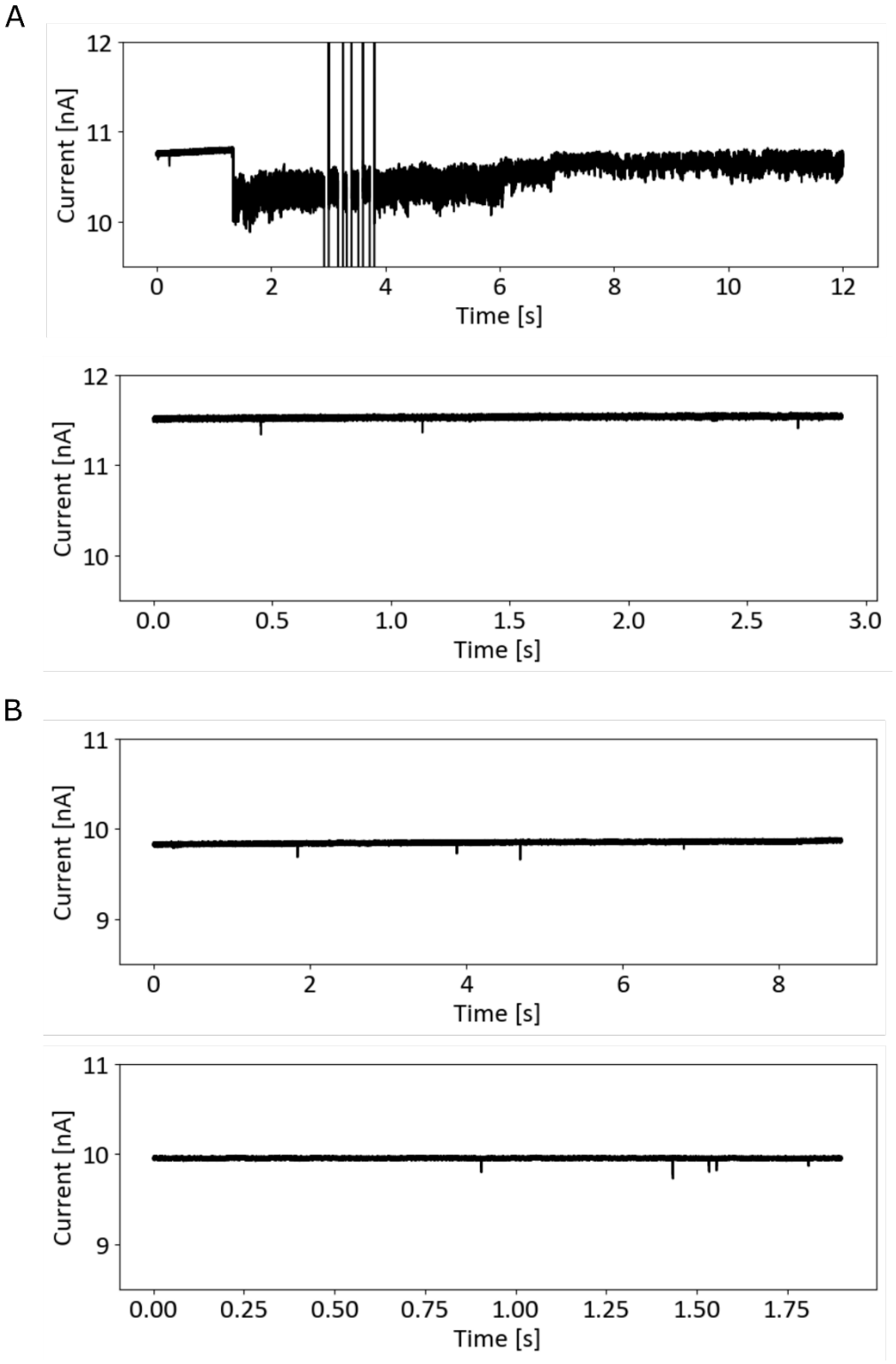


Figure C.5. Nanopore traces with and without presence of Anle-138b. **(A)** Raw current of the nanopore trace with Anle-138b shows that upon mixture and measurement in the nanopore after 1 sec, a lot of noise is created. After 3 min (trace below) the noise level resumes back to normal. The vertical lines represent kick outs to remove protein from clogging the pore. **(B)** Raw current trace without Anle-138b shows similar noise and baseline both upon mixture and measurement in the pore and after 3 min of measurement. Nanopore experiments were carried out by Sarah E. Sandler.

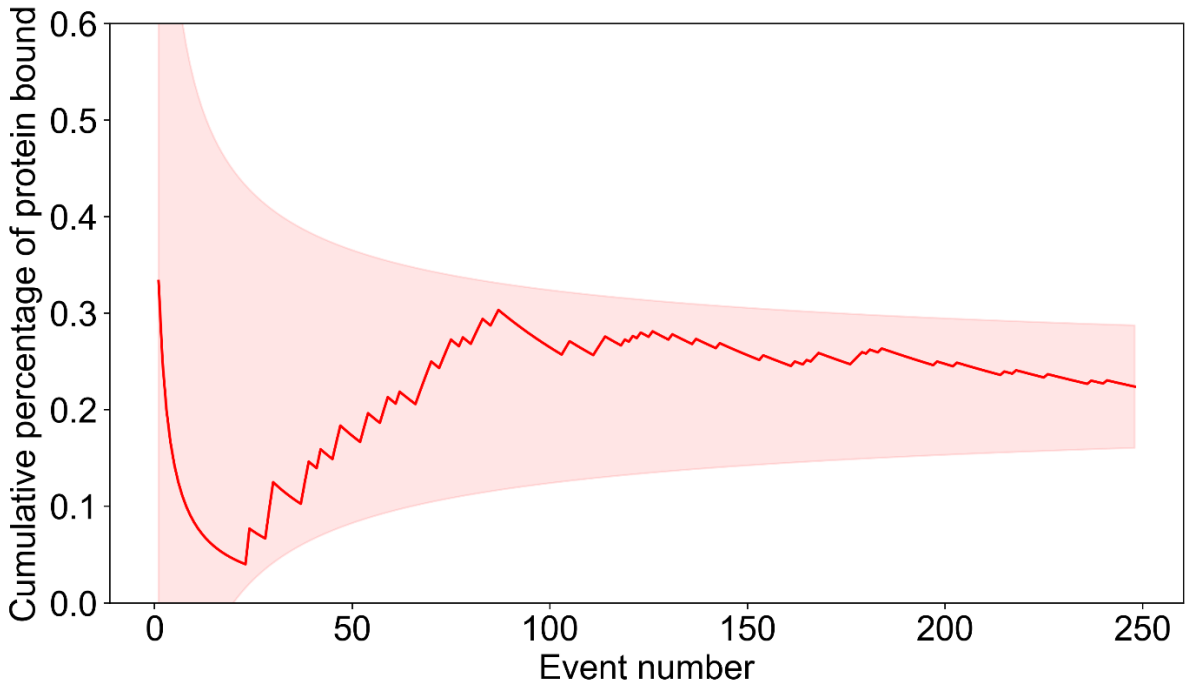


Figure C.6. Cumulative percentage of events with clear barcode and protein bound as measured in 4 M LiCl. The number of events with the protein bound remains the same for the stabilised oligomer over the course of the 1 h measurement time. N=250 because the original trace was filtered to remove folded or knotted events with unreadable barcodes. The error fluctuation represents 1σ deviation. This becomes smaller as the number of events increases. The increase in number of events corresponds to measurement time. Nanopore experiments were carried out by Sarah E. Sandler.

The following replacements are made to create the "1" bits in the barcode portion of the nanostructure as previously shown and used in **Figure 4.2**[182]. All design work was done by Sarah E. Sandler.

First bit

Replace oligos 26,27,28,29,30,31 and 32

CTGAAAGCGTAAGAATACGTGGCACAGACAATATTTTTGAATGGCT
ACATCACTTGTCCTCTTTTGAGGAACAAGTTTTCTTGTCCTGAGTAGA
AGAACTCAAATCCTCTTTTGAGGAACAAGTTTTCTTGTCTATCGGCCT
TGCTGGTAATTCCTCTTTTGAGGAACAAGTTTTCTTGTATCCAGAACA
ATATTACCGCTCCTCTTTTGAGGAACAAGTTTTCTTGTCAGCCATTGC
AACAGGAAAATCCTCTTTTGAGGAACAAGTTTTCTTGTACGCTCATGG
AAATACCTACTCCTCTTTTGAGGAACAAGTTTTCTTGTATTTTGACGC
TCAATCGTCTTCCTCTTTTGAGGAACAAGTTTTCTTGTGAAATGGATT
ATTTACATTGTCCTCTTTTGAGGAACAAGTTTTCTTGTGCAGATTCAC
CAGTCACACGTCCTCTTTTGAGGAACAAGTTTTCTTGTACCAGTAATA
AAAGGGACATTCCTCTTTTGAGGAACAAGTTTTCTTGTTCTGGCCAAC
AGAGATAGAATCCTCTTTTGAGGAACAAGTTTTCTTGTCCCTTCTGAC

Second bit

Replace oligos 40,41,42,43,44,45 and 46

AATATAATCCTGATTGTTTGGATTATACTTCTGAATAATGGAAGGG
CACTAACAACTCCTCTTTTGAGGAACAAGTTTTCTTGTTAATAGATTA
GAGCCGTCAATCCTCTTTTGAGGAACAAGTTTTCTTGTTAGATAATAC
ATTTGAGGATTCCTCTTTTGAGGAACAAGTTTTCTTGTTTAGAAGTAT
TAGACTTTACTCCTCTTTTGAGGAACAAGTTTTCTTGTAAACAATTCG
ACAACTCGTATCCTCTTTTGAGGAACAAGTTTTCTTGTTTAAATCCTT
TGCCCGAACGTCCTCTTTTGAGGAACAAGTTTTCTTGTTTATTAATTT
TAAAAGTTTGTCCTCTTTTGAGGAACAAGTTTTCTTGTAGTAACATTA
TCATTTTGCGTCCTCTTTTGAGGAACAAGTTTTCTTGTGAACAAAGAA
ACCACCAGAATCCTCTTTTGAGGAACAAGTTTTCTTGTGGAGCGGAAT
TATCATCATATCCTCTTTTGAGGAACAAGTTTTCTTGTTTCCTGATTA
TCAGATGATGTCCTCTTTTGAGGAACAAGTTTTCTTGTGCAATTCATC

Third bit

Replace oligos 54,55,56,57,58,59 and 60

TGTAAATGCTGATGCAAATCCAATCGCAAGACAAAGAACGCGAGAA
TACCTTTTTTTCCTCTTTTGAGGAACAAGTTTTCTTGTAATGGAAACA
GTACATAAATTCCTCTTTTGAGGAACAAGTTTTCTTGTCAATATATGT
GAGTGAATAATCCTCTTTTGAGGAACAAGTTTTCTTGTCCTTGCTTCT
GTAAATCGTCTCCTCTTTTGAGGAACAAGTTTTCTTGTGCTATTAATT
AATTTTCCCTTCCTCTTTTGAGGAACAAGTTTTCTTGTTAGAATCCTT
GAAAACATAGTCCTCTTTTGAGGAACAAGTTTTCTTGTCGATAGCTTA
GATTAAGACGTCCTCTTTTGAGGAACAAGTTTTCTTGTCTGAGAAGAG
TCAATAGTGATCCTCTTTTGAGGAACAAGTTTTCTTGTATTTATCAAA
ATCATAGGTCTCCTCTTTTGAGGAACAAGTTTTCTTGTTGAGAGACTA
CCTTTTTAACTCCTCTTTTGAGGAACAAGTTTTCTTGTCTCCGGCTTA
GGTTGGGTTATCCTCTTTTGAGGAACAAGTTTTCTTGTTATAACTATA

Fourth bit

Replace oligos 68,69,70,71,72,73 and 74

TCATCGAGAACAAGCAAGCCGTTTTTATTTTCATCGTAGGAATCAT
AGAATATAAATCCTCTTTTGAGGAACAAGTTTTCTTGTGTACCGACAA
AAGGTAAAGTTCCTCTTTTGAGGAACAAGTTTTCTTGTAATTCTGTCC
AGACGACGACTCCTCTTTTGAGGAACAAGTTTTCTTGTAATAAACAAC
ATGTTCAGCTTCCTCTTTTGAGGAACAAGTTTTCTTGTAATGCAGAAC
GCGCCTGTTTTCCTCTTTTGAGGAACAAGTTTTCTTGTATCAACAATA
GATAAGTCCTTCCTCTTTTGAGGAACAAGTTTTCTTGTGAACAAGAAA
AATAATATCCTCCTCTTTTGAGGAACAAGTTTTCTTGTCATCCTAATT
TACGAGCATGTCCTCTTTTGAGGAACAAGTTTTCTTGTTAGAAACCAA
TCAATAATCGTCCTCTTTTGAGGAACAAGTTTTCTTGTGCTGTCTTTC
CTTATCATTCTCCTCTTTTGAGGAACAAGTTTTCTTGTCAAGAACGGG
TATTAAACCATCCTCTTTTGAGGAACAAGTTTTCTTGTAGTACCGCAC

Fifth bit

Replace oligos 82,83,84,85,86,87 and 88

AGATAGCCGAACAAAGTTACCAGAAGGAAACCGAGGAAACGCAATA
AAAAATGAAATCCTCTTTTGAGGAACAAGTTTTCTTGTATAGCAGCCT
TTACAGAGAGTCCTCTTTTGAGGAACAAGTTTTCTTGTAATAACATAA
AAACAGGGAATCCTCTTTTGAGGAACAAGTTTTCTTGTGCGCATTAGA
CGGGAGAATTTCCTCTTTTGAGGAACAAGTTTTCTTGTAACTGAACAC
CCTGAACAAATCCTCTTTTGAGGAACAAGTTTTCTTGTGTCAGAGGGT
AATTGAGCGCTCCTCTTTTGAGGAACAAGTTTTCTTGTTAATATCAGA
GAGATAACCCTCCTCTTTTGAGGAACAAGTTTTCTTGTACAAGAATTG
AGTTAAGCCCTCCTCTTTTGAGGAACAAGTTTTCTTGTAATAATAAGA
GCAAGAAACATCCTCTTTTGAGGAACAAGTTTTCTTGTATGAAATAGC
AATAGCTATCTCCTCTTTTGAGGAACAAGTTTTCTTGTTTACCGAAGC
CCTTTTTAAGTCCTCTTTTGAGGAACAAGTTTTCTTGTAAAAGTAAGC

Table C.2. DNA Dumbbell Bits

142	GATGGTTTAATTTCAACTTTAATCATTGTGAATTACCT
143	actgactgactgactgactgaTTTATGCGATTTTAAGAACTGGCTCATTATACCAGT CAGG
DBCO	tcagtcagtcagtcagtcagt*DBCO*

Table C.3. DNA overhang sequences

D. Generalising to other misfolded proteins

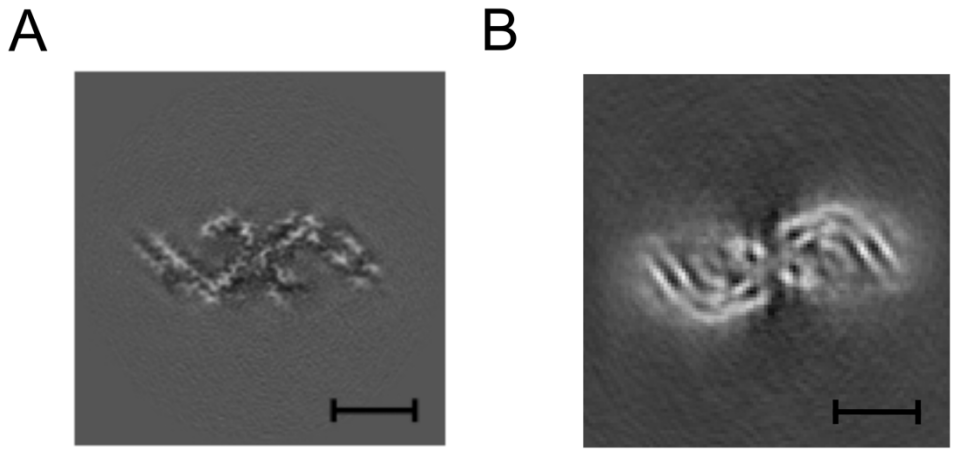


Figure D.1. Reported cryo-EM reconstruction of AD derived tau fibrils[207] vs second-generation in vitro 0N3R tau fibrils generated using AD fibrils. (A) Z-axis cross-sections of AD derived fibrils after complete 3D refinement. **(B)** Preliminary z-axis cross-sections of second-generation in vitro 0N3R tau fibrils after first stage 3D refinement. Scale bar is 50 Å. Cryo-EM was carried out by Alessia Santambrogio and Thomas Löhr.

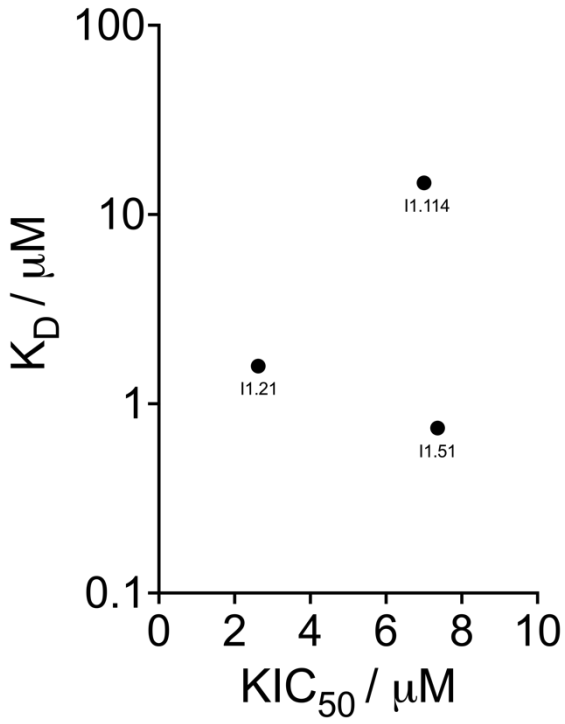


Figure D.2. KIC_{50} values vs K_D values for the 3 ML derived molecules that were fully characterised. Kinetics were run by Alessia Santambrogio.

Bibliography

1. O'Baugh, J., Wilkes, L.M., Luke, S. & George, A. 'Being positive': perceptions of patients with cancer and their nurses. *Journal of Advanced Nursing* **44**, 262-270 (2003).
2. Gove, D., Downs, M., Vernooij-Dassen, M. & Small, N. Stigma and GPs' perceptions of dementia. *Aging & mental health* **20**, 391-400 (2016).
3. De Magalhães, J.P. How ageing processes influence cancer. *Nature Reviews Cancer* **13**, 357-365 (2013).
4. Hou, Y. et al. Ageing as a risk factor for neurodegenerative disease. *Nature Reviews Neurology* **15**, 565-581 (2019).
5. Sundquist, M., Brudin, L. & Tejler, G. Improved survival in metastatic breast cancer 1985–2016. *The Breast* **31**, 46-50 (2017).
6. Baguley, B.C. Multiple drug resistance mechanisms in cancer. *Molecular biotechnology* **46**, 308-316 (2010).
7. Small, D.H. & Cappai, R. Alois Alzheimer and Alzheimer's disease: a centennial perspective. *Journal of neurochemistry* **99**, 708-710 (2006).
8. Sevigny, J. et al. The antibody aducanumab reduces Aβ plaques in Alzheimer's disease. *Nature* **537**, 50-56 (2016).
9. van Dyck, C.H. et al. Lecanemab in early Alzheimer's disease. *New England Journal of Medicine* (2022).
10. Chiti, F. & Dobson, C.M. Protein misfolding, functional amyloid, and human disease. *Annu Rev Biochem* **75**, 333-66 (2006).
11. Ow, S.Y. & Dunstan, D.E. A brief overview of amyloids and Alzheimer's disease. *Protein Science* **23**, 1315-1331 (2014).

12. Haass, C. et al. The Swedish mutation causes early-onset Alzheimer's disease by β-secretase cleavage within the secretory pathway. *Nature medicine* **1**, 1291-1296 (1995).
13. Jellinger, K.A. Basic mechanisms of neurodegeneration: a critical update. *Journal of cellular and molecular medicine* **14**, 457-487 (2010).
14. Devi, G. & Scheltens, P. Heterogeneity of Alzheimer's disease: consequence for drug trials? *Alzheimer's research & therapy* **10**, 1-3 (2018).
15. Kundra, R., Ciryam, P., Morimoto, R.I., Dobson, C.M. & Vendruscolo, M. Protein homeostasis of a metastable subproteome associated with Alzheimer's disease. *Proceedings of the National Academy of Sciences* **114**, E5703-E5711 (2017).
16. Morley, J.F., Brignull, H.R., Weyers, J.J. & Morimoto, R.I. The threshold for polyglutamine-expansion protein aggregation and cellular toxicity is dynamic and influenced by aging in Caenorhabditis elegans. *Proceedings of the National Academy of Sciences* **99**, 10417-10422 (2002).
17. Labbadia, J. & Morimoto, R.I. The biology of proteostasis in aging and disease. *Annual review of biochemistry* **84**, 435-464 (2015).
18. Groveman, B.R. et al. Rapid and ultra-sensitive quantitation of disease-associated α-synuclein seeds in brain and cerebrospinal fluid by αSyn RT-QuIC. *Acta Neuropathol Commun* **6**, 7 (2018).
19. Schalkamp, A.-K., Peall, K.J., Harrison, N.A. & Sandor, C. Wearable movement-tracking data identify Parkinson's disease years before clinical diagnosis. *Nature Medicine* **29**, 2048-2056 (2023).
20. Gaetani, L. et al. Neurofilament light chain as a biomarker in neurological disorders. *Journal of Neurology, Neurosurgery & Psychiatry* **90**, 870-881 (2019).
21. Choi, M.L. et al. Pathological structural conversion of α-synuclein at the mitochondria induces neuronal toxicity. *Nature neuroscience*, 1-15 (2022).
22. Balusu, S. et al. MEG3 activates necroptosis in human neuron xenografts modeling Alzheimer's disease. *Science* **381**, 1176-1182 (2023).
23. Yang, Y. et al. Structures of α-synuclein filaments from human brains with Lewy pathology. *Nature* **610**, 791-795 (2022).
24. Knopman, D.S., Jones, D.T. & Greicius, M.D. Failure to demonstrate efficacy of aducanumab: An analysis of the EMERGE and ENGAGE trials as reported by Biogen, December 2019. *Alzheimer's & Dementia* **17**, 696-701 (2021).

25. Brockmann, R., Nixon, J., Love, B.L. & Yunusa, I. Impacts of FDA approval and Medicare restriction on antiamyloid therapies for Alzheimer's disease: patient outcomes, healthcare costs, and drug development. *The Lancet Regional Health–Americas* **20**(2023).
26. Sims, J.R. et al. Donanemab in early symptomatic Alzheimer disease: the TRAILBLAZER-ALZ 2 randomized clinical trial. *JAMA* (2023).
27. Global status report on the public health response to dementia. *World Health Organisation* (2021).
28. Cummings, J., Reiber, C. & Kumar, P. The price of progress: Funding and financing Alzheimer's disease drug development. *Alzheimer's & Dementia: Translational Research & Clinical Interventions* **4**, 330-343 (2018).
29. Ke, P.C. et al. Half a century of amyloids: past, present and future. *Chemical Society Reviews* **49**, 5473-5509 (2020).
30. National audit of dementia care in general hospitals 2018–2019. Round four audit report. (Royal College of Psychiatrists London, 2019).
31. Freer, R. et al. A protein homeostasis signature in healthy brains recapitulates tissue vulnerability to Alzheimer's disease. *Science advances* **2**, e1600947 (2016).
32. Ciryam, P., Tartaglia, G.G., Morimoto, R.I., Dobson, C.M. & Vendruscolo, M. Widespread aggregation and neurodegenerative diseases are associated with supersaturated proteins. *Cell reports* **5**, 781-790 (2013).
33. Thacker, D. et al. The role of fibril structure and surface hydrophobicity in secondary nucleation of amyloid fibrils. *Proceedings of the National Academy of Sciences* **117**, 25272-25283 (2020).
34. Cerf, E. et al. Antiparallel β-sheet: a signature structure of the oligomeric amyloid β-peptide. *Biochemical Journal* **421**, 415-423 (2009).
35. Celej, M.S. et al. Toxic prefibrillar α-synuclein amyloid oligomers adopt a distinctive antiparallel β-sheet structure. *Biochemical journal* **443**, 719-726 (2012).
36. Michaels, T.C.T. et al. Dynamics of oligomer populations formed during the aggregation of Alzheimer's Abeta42 peptide. *Nat Chem* **12**, 445-451 (2020).
37. Vecchi, G. et al. Proteome-wide observation of the phenomenon of life on the edge of solubility. *Proceedings of the National Academy of Sciences* **117**, 1015-1020 (2020).
38. Balch, W.E., Morimoto, R.I., Dillin, A. & Kelly, J.W. Adapting proteostasis for disease intervention. *science* **319**, 916-919 (2008).

39. Hipp, M.S., Kasturi, P. & Hartl, F.U. The proteostasis network and its decline in ageing. *Nature reviews Molecular cell biology* **20**, 421-435 (2019).
40. Arosio, P. et al. Kinetic analysis reveals the diversity of microscopic mechanisms through which molecular chaperones suppress amyloid formation. *Nature communications* **7**, 10948 (2016).
41. Haass, C. & Selkoe, D.J. Soluble protein oligomers in neurodegeneration: lessons from the Alzheimer's amyloid β-peptide. *Nature reviews Molecular cell biology* **8**, 101-112 (2007).
42. Benilova, I., Karran, E. & De Strooper, B. The toxic Aβ oligomer and Alzheimer's disease: an emperor in need of clothes. *Nature neuroscience* **15**, 349-357 (2012).
43. Knowles, T.P., Vendruscolo, M. & Dobson, C.M. The amyloid state and its association with protein misfolding diseases. *Nat Rev Mol Cell Biol* **15**, 384-96 (2014).
44. Goedert, M. & Spillantini, M.G. A century of Alzheimer's disease. *science* **314**, 777-781 (2006).
45. Spillantini, M.G. & Goedert, M. Tau pathology and neurodegeneration. *The Lancet Neurology* **12**, 609-622 (2013).
46. Fusco, G. et al. Structural basis of membrane disruption and cellular toxicity by alpha-synuclein oligomers. *Science* **358**, 1440-1443 (2017).
47. Lashuel, H.A., Overk, C.R., Oueslati, A. & Masliah, E. The many faces of alpha-synuclein: from structure and toxicity to therapeutic target. *Nat Rev Neurosci* **14**, 38-48 (2013).
48. Campioni, S. et al. A causative link between the structure of aberrant protein oligomers and their toxicity. *Nature chemical biology* **6**, 140-147 (2010).
49. Vendruscolo, M. Thermodynamic and kinetic approaches for drug discovery to target protein misfolding and aggregation. *Expert Opinion on Drug Discovery*, 1-11 (2023).
50. Goedert, M., Spillantini, M.G., Del Tredici, K. & Braak, H. 100 years of Lewy pathology. *Nature Reviews Neurology* **9**, 13-24 (2013).
51. Spillantini, M.G., Crowther, R.A., Jakes, R., Hasegawa, M. & Goedert, M. alpha-Synuclein in filamentous inclusions of Lewy bodies from Parkinson's disease and dementia with lewy bodies. *Proc Natl Acad Sci U S A* **95**, 6469-73 (1998).
52. Savica, R., Boeve, B.F. & Mielke, M.M. When Do alpha-Synucleinopathies Start? An Epidemiological Timeline: A Review. *JAMA Neurol* **75**, 503-509 (2018).

53. Rambaran, R.N. & Serpell, L.C. Amyloid fibrils: abnormal protein assembly. *Prion* **2**, 112-117 (2008).
54. Ray, S. et al. α-Synuclein aggregation nucleates through liquid–liquid phase separation. *Nature chemistry* **12**, 705-716 (2020).
55. Jacobs, T.M. & Kuhlman, B. Using anchoring motifs for the computational design of protein–protein interactions. *Biochemical Society Transactions* **41**, 1141-1145 (2013).
56. Michaels, T.C., Dear, A.J., Cohen, S.I., Vendruscolo, M. & Knowles, T.P. Kinetic profiling of therapeutic strategies for inhibiting the formation of amyloid oligomers. *The Journal of Chemical Physics* **156**, 164904 (2022).
57. Staats, R. et al. Screening of small molecules using the inhibition of oligomer formation in α-synuclein aggregation as a selection parameter. *Communications Chemistry* **3**, 191 (2020).
58. Price, D.L. et al. The small molecule alpha-synuclein misfolding inhibitor, NPT200-11, produces multiple benefits in an animal model of Parkinson's disease. *Sci Rep* **8**, 16165 (2018).
59. Pujols, J., Pena-Diaz, S., Pallares, I. & Ventura, S. Chemical Chaperones as Novel Drugs for Parkinson's Disease. *Trends Mol Med* **26**, 408-421 (2020).
60. Wagner, J. et al. Anle138b: a novel oligomer modulator for disease-modifying therapy of neurodegenerative diseases such as prion and Parkinson's disease. *Acta Neuropathol* **125**, 795-813 (2013).
61. Emin, D. et al. Small soluble α-synuclein aggregates are the toxic species in Parkinson's disease. *Nature Communications* **13**, 1-15 (2022).
62. Gaspar, R. et al. Secondary nucleation of monomers on fibril surface dominates α-synuclein aggregation and provides autocatalytic amyloid amplification. *Quarterly reviews of biophysics* **50**, E6 (2017).
63. Chia, S. et al. SAR by kinetics for drug discovery in protein misfolding diseases. *Proc Natl Acad Sci U S A* **115**, 10245-10250 (2018).
64. Chia, S. et al. Structure-Based Discovery of Small-Molecule Inhibitors of the Autocatalytic Proliferation of alpha-Synuclein Aggregates. *Mol Pharm* **20**, 183–193 (2022).
65. Horne, R.I. et al. Discovery of Potent Inhibitors of α-Synuclein Aggregation Using Structure-Based Iterative Learning. *bioRxiv*, 2021.11. 10.468009 (2021).
66. van Dyck, C.H. et al. Lecanemab in Early Alzheimer's Disease. *N Engl J Med* (2022).

67. McFarthing, K. et al. Parkinson's Disease Drug Therapies in the Clinical Trial Pipeline: 2022 Update. *J Parkinsons Dis* **12**, 1073-1082 (2022).
68. Oertel, W. & Schulz, J.B. Current and experimental treatments of Parkinson disease: A guide for neuroscientists. *J Neurochem* **139 Suppl 1**, 325-337 (2016).
69. Tolosa, E., Garrido, A., Scholz, S.W. & Poewe, W. Challenges in the diagnosis of Parkinson's disease. *Lancet Neurol* **20**, 385-397 (2021).
70. Aarsland, D. et al. Parkinson disease-associated cognitive impairment. *Nat Rev Dis Primers* **7**, 47 (2021).
71. Balestrino, R. & Schapira, A.H.V. Parkinson disease. *Eur J Neurol* **27**, 27-42 (2020).
72. Collaborators, G.B.D.P.s.D. Global, regional, and national burden of Parkinson's disease, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet Neurol* **17**, 939-953 (2018).
73. Poewe, W. Parkinson disease Primer - a true team effort. *Nat Rev Dis Primers* **6**, 31 (2020).
74. Flagmeier, P. et al. Mutations associated with familial Parkinson's disease alter the initiation and amplification steps of alpha-synuclein aggregation. *Proc Natl Acad Sci U S A* **113**, 10328-33 (2016).
75. Man, W.K. et al. The docking of synaptic vesicles on the presynaptic membrane induced by alpha-synuclein is modulated by lipid composition. *Nature Communications* **12**(2021).
76. Panteleev, J., Gao, H. & Jia, L. Recent applications of machine learning in medicinal chemistry. *Bioorg Med Chem Lett* **28**, 2807-2815 (2018).
77. Vamathevan, J. et al. Applications of machine learning in drug discovery and development. *Nat Rev Drug Discov* **18**, 463-477 (2019).
78. Meng, X.Y., Zhang, H.X., Mezei, M. & Cui, M. Molecular docking: a powerful approach for structure-based drug discovery. *Curr Comput Aided Drug Des* **7**, 146-57 (2011).
79. Myszczynska, M.A. et al. Applications of machine learning to diagnosis and treatment of neurodegenerative diseases. *Nat Rev Neurol* **16**, 440-456 (2020).
80. Brown, J.W. et al. β-Synuclein suppresses both the initiation and amplification steps of α-synuclein aggregation via competitive binding to surfaces. *Scientific reports* **6**, 1-10 (2016).

81. Hie, B., Bryson, B.D. & Berger, B. Leveraging Uncertainty in Machine Learning Accelerates Biological Discovery and Design. *Cell Syst* **11**, 461-477 e9 (2020).
82. Knowles, T.P. et al. An analytical solution to the kinetics of breakable filament assembly. *Science* **326**, 1533-7 (2009).
83. Jin, W., Barzilay, R. & Jaakkola, T. Junction tree variational autoencoder for molecular graph generation. in *International conference on machine learning* 2323-2332 (PMLR, 2018).
84. Kusner, M.J., Paige, B. & Hernández-Lobato, J.M. Grammar variational autoencoder. in *International conference on machine learning* 1945-1954 (PMLR, 2017).
85. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of chemical information and computer sciences* **28**, 31-36 (1988).
86. Bento, A.P. et al. An open source chemical structure curation pipeline using RDKit. *Journal of Cheminformatics* **12**, 1-16 (2020).
87. Breiman, L. Random forests. *Machine learning* **45**, 5-32 (2001).
88. Rasmussen, C.E. & Williams, C. Gaussian processes for machine learning, vol. 1. (MIT press Cambridge MA, 2006).
89. Svozil, D., Kvasnicka, V. & Pospichal, J. Introduction to multi-layer feed-forward neural networks. *Chemometrics and intelligent laboratory systems* **39**, 43-62 (1997).
90. Trott, O. & Olson, A.J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem* **31**, 455-61 (2010).
91. McGann, M. FRED pose prediction and virtual screening accuracy. *Journal of chemical information and modeling* **51**, 578-596 (2011).
92. Le Guilloux, V., Schmidtke, P. & Tuffery, P. Fpocket: an open source platform for ligand pocket detection. *BMC Bioinformatics* **10**, 168 (2009).
93. Kelley, B.P., Brown, S.P., Warren, G.L. & Muchmore, S.W. POSIT: Flexible Shape-Guided Docking For Pose Prediction. *J Chem Inf Model* **55**, 1771-80 (2015).
94. Wager, T.T., Hou, X., Verhoest, P.R. & Villalobos, A. Central Nervous System Multiparameter Optimization Desirability: Application in Drug Discovery. *ACS Chem Neurosci* **7**, 767-75 (2016).
95. Wager, T.T., Hou, X., Verhoest, P.R. & Villalobos, A. Moving beyond rules: the development of a central nervous system multiparameter optimization (CNS MPO)

approach to enable alignment of druglike properties. *ACS chemical neuroscience* **1**, 435-449 (2010).

96. Sormanni, P., Aprile, F.A. & Vendruscolo, M. The CamSol method of rational design of protein mutants with enhanced solubility. *Journal of molecular biology* **427**, 478-490 (2015).
97. Buell, A.K. et al. Solution conditions determine the relative importance of nucleation and growth processes in alpha-synuclein aggregation. *Proc Natl Acad Sci U S A* **111**, 7671-6 (2014).
98. Li, B. et al. Cryo-EM of full-length α-synuclein reveals fibril polymorphs with a common structural kernel. *Nature communications* **9**, 1-10 (2018).
99. Butina, D. Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way To Cluster Small and Large Data Sets. *Journal of Chemical Information and Computer Sciences* **39**, 747-750 (1999).
100. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *Journal of chemical information and modeling* **50**, 742-754 (2010).
101. Horne, R.I. et al. Exploration and Exploitation Approaches Based on Generative Machine Learning to Identify Potent Small Molecule Inhibitors of α-Synuclein Secondary Nucleation. *Journal of Chemical Theory and Computation* **19**, 4701–4710 (2023).
102. McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).
103. Stumpfe, D., Hu, H. & Bajorath, J.r. Evolving concept of activity cliffs. *ACS omega* **4**, 14360-14368 (2019).
104. Hill, A.V. The possible effects of the aggregation of the molecules of hemoglobin on its dissociation curves. *j. physiol.* **40**, iv-vii (1910).
105. Kurnik, M. et al. Potent α-synuclein aggregation inhibitors, identified by high-throughput screening, mainly target the monomeric state. *Cell chemical biology* **25**, 1389-1402. e9 (2018).
106. Horne, R.I. et al. Secondary Processes Dominate the Quiescent, Spontaneous Aggregation of α-Synuclein at Physiological pH with Sodium Salts. *ACS Chemical Neuroscience* **14**, 3125-3131 (2023).
107. Van Der Maaten, L. Accelerating t-SNE using tree-based algorithms. *The Journal of Machine Learning Research* **15**, 3221-3245 (2014).

108. Lundberg, S.M. & Lee, S.-I. A unified approach to interpreting model predictions. *Advances in neural information processing systems* **30**(2017).

109. Cooper, A., Doyle, O. & Bourke, A. Supervised Clustering for Subgroup Discovery: An Application to COVID-19 Symptomatology. in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* 408-422 (Springer, 2021).

110. Furukawa, K. et al. Isoelectric point-amyloid formation of α-synuclein extends the generality of the solubility and supersaturation-limited mechanism. *Current Research in Structural Biology* **2**, 35-44 (2020).

111. Atarashi, R. et al. Ultrasensitive human prion detection in cerebrospinal fluid by real-time quaking-induced conversion. *Nature medicine* **17**, 175-178 (2011).

112. Atarashi, R. et al. Ultrasensitive human prion detection in cerebrospinal fluid by real-time quaking-induced conversion. *Nat Med* **17**, 175-8 (2011).

113. Wilham, J.M. et al. Rapid end-point quantitation of prion seeding activity with sensitivity comparable to bioassays. *PLoS Pathog* **6**, e1001217 (2010).

114. Metrick, M.A., 2nd et al. A single ultrasensitive assay for detection and discrimination of tau aggregates of Alzheimer and Pick diseases. *Acta Neuropathol Commun* **8**, 22 (2020).

115. Grazia Spillantini, M. et al. A novel tau mutation (N296N) in familial dementia with swollen achromatic neurons and corticobasal inclusion bodies. *Annals of neurology* **48**, 939-943 (2000).

116. Arter, W.E. et al. Rapid Structural, Kinetic, and Immunochemical Analysis of Alpha-Synuclein Oligomers in Solution. *Nano Lett* **20**, 8163-8169 (2020).

117. Zhu, T. et al. Hit identification and optimization in virtual screening: Practical recommendations based on a critical literature analysis: Miniperspective. *Journal of medicinal chemistry* **56**, 6560-6572 (2013).

118. Blaschke, T. et al. REINVENT 2.0: an AI tool for de novo drug design. *Journal of chemical information and modeling* **60**, 5918-5922 (2020).

119. Maziarka, Ł. et al. Mol-CycleGAN: a generative model for molecular optimization. *Journal of Cheminformatics* **12**, 1-18 (2020).

120. You, J., Liu, B., Ying, Z., Pande, V. & Leskovec, J. Graph convolutional policy network for goal-directed molecular graph generation. *Advances in neural information processing systems* **31**(2018).

121. Zhou, Z., Kearnes, S., Li, L., Zare, R.N. & Riley, P. Optimization of Molecules via Deep Reinforcement Learning. *Sci Rep* **9**, 10752 (2019).
122. Chandra, R., Horne, R.I. & Vendruscolo, M. Bayesian Optimization in the Latent Space of a Variational Autoencoder for the Generation of Selective FLT3 Inhibitors. *Journal of Chemical Theory and Computation* **20**, 469-476 (2023).
123. Allen, C.H. et al. Improving the prediction of organism-level toxicity through integration of chemical, protein target and cytotoxicity qHTS data. *Toxicology research* **5**, 883-894 (2016).
124. Horne, R.I. et al. Using Generative Modeling to Endow with Potency Initially Inert Compounds with Good Bioavailability and Low Toxicity. *Journal of Chemical Information and Modeling* (2024).
125. Perni, M. et al. Multistep Inhibition of alpha-Synuclein Aggregation and Toxicity in Vitro and in Vivo by Trodusquemine. *ACS Chem Biol* **13**, 2308-2319 (2018).
126. Galvagnion, C. et al. Lipid vesicles trigger α-synuclein aggregation by stimulating primary nucleation. *Nature chemical biology* **11**, 229-234 (2015).
127. Michaels, T.C., Cohen, S.I., Vendruscolo, M., Dobson, C.M. & Knowles, T.P. Hamiltonian Dynamics of Protein Filament Formation. *Phys Rev Lett* **116**, 038101 (2016).
128. Pedregosa, F. et al. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* **12**, 2825-2830 (2011).
129. McKinney, W. Data structures for statistical computing in python. in *Proceedings of the 9th Python in Science Conference* Vol. 445 51-56 (Austin, TX, 2010).
130. Waskom, M.L. Seaborn: statistical data visualization. *Journal of Open Source Software* **6**, 3021 (2021).
131. Hunter, J.D. Matplotlib: A 2D graphics environment. *Computing in science & engineering* **9**, 90-95 (2007).
132. Harris, C.R. et al. Array programming with NumPy. *Nature* **585**, 357-362 (2020).
133. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature methods* **17**, 261-272 (2020).
134. Kramer, O. *Machine learning for evolution strategies*, (Springer, 2016).
135. Mazutis, L. et al. Single-cell analysis and sorting using droplet-based microfluidics. *Nature protocols* **8**, 870-891 (2013).

136. McDonald, J.C. et al. Fabrication of microfluidic systems in poly (dimethylsiloxane). *ELECTROPHORESIS: An International Journal* **21**, 27-40 (2000).
137. Challa, P.K., Kartanas, T., Charmet, J. & Knowles, T.P. Microfluidic devices fabricated using fast wafer-scale LED-lithography patterning. *Biomicrofluidics* **11**, 014113 (2017).
138. Tan, S.H., Nguyen, N.-T., Chua, Y.C. & Kang, T.G. Oxygen plasma treatment for reducing hydrophobicity of a sealed polydimethylsiloxane microchannel. *Biomicrofluidics* **4**, 032204 (2010).
139. Saar, K.L. et al. On-chip label-free protein analysis with downstream electrodes for direct removal of electrolysis products. *Lab on a Chip* **18**, 162-170 (2018).
140. Mahul-Mellier, A.-L. et al. The process of Lewy body formation, rather than simply α-synuclein fibrillization, is one of the major drivers of neurodegeneration. *Proceedings of the National Academy of Sciences* **117**, 4971-4982 (2020).
141. Mercado, R. et al. Graph networks for molecular design. *Machine Learning: Science and Technology* **2**, 025023 (2021).
142. Moret, M., Friedrich, L., Grisoni, F., Merk, D. & Schneider, G. Generative molecular design in low data regimes. *Nature Machine Intelligence* **2**, 171-180 (2020).
143. Neves, B.J. et al. QSAR-based virtual screening: advances and applications in drug discovery. *Frontiers in pharmacology* **9**, 1275 (2018).
144. Kwon, S., Bae, H., Jo, J. & Yoon, S. Comprehensive ensemble in QSAR prediction for drug discovery. *BMC bioinformatics* **20**, 1-12 (2019).
145. Schonlau, M. & Zou, R.Y. The random forest algorithm for statistical learning. *The Stata Journal* **20**, 3-29 (2020).
146. ChemDiv CNS BBB Library. (2022).
147. ChemDiv CNS MPO Library. (2022).
148. Meng, F., Xi, Y., Huang, J. & Ayers, P.W. A curated diverse molecular database of blood-brain barrier permeability with chemical descriptors. *Sci Data* **8**, 289 (2021).
149. Kumar, R. et al. DeePred-BBB: A Blood Brain Barrier Permeability Prediction Model With Improved Accuracy. *Frontiers in Neuroscience* **16**(2022).
150. Brown, N., Fiscato, M., Segler, M.H. & Vaucher, A.C. GuacaMol: benchmarking models for de novo molecular design. *Journal of chemical information and modeling* **59**, 1096-1108 (2019).

151. Mansouri, K. et al. Open-source QSAR models for pKa prediction using multiple machine learning approaches. *Journal of cheminformatics* **11**, 1-20 (2019).
152. Yap, C.W. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *Journal of computational chemistry* **32**, 1466-1474 (2011).
153. Yasui, Y. & Fujisawa, K. Fast and scalable NUMA-based thread parallel breadth-first search. in *2015 International Conference on High Performance Computing & Simulation (HPCS)* 377-385 (IEEE, 2015).
154. Joyce, J.M. Kullback-leibler divergence. in *International encyclopedia of statistical science* 720-722 (Springer, 2011).
155. Kingma, D.P. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
156. Jaeger, S., Fulle, S. & Turk, S. Mol2vec: unsupervised machine learning approach with chemical intuition. *Journal of chemical information and modeling* **58**, 27-35 (2018).
157. Fukushima, K. Cognitron: A self-organizing multilayered neural network. *Biological cybernetics* **20**, 121-136 (1975).
158. Van Houdt, G., Mosquera, C. & Nápoles, G. A review on the long short-term memory model. *Artificial Intelligence Review* **53**, 5929-5955 (2020).
159. Holland, J.H. Genetic algorithms. *Scientific american* **267**, 66-73 (1992).
160. Alam, T., Qamar, S., Dixit, A. & Benaida, M. Genetic algorithm: Reviews, implementations, and applications. *arXiv preprint arXiv:2007.12673* (2020).
161. Szenkovits, A. et al. Feature selection with a genetic algorithm for classification of brain imaging data. *Advances in feature selection for data and pattern recognition*, 185-202 (2018).
162. Gow, S., Niranjan, M., Kanza, S. & Frey, J. A Review of Reinforcement Learning in Chemistry. *Digital Discovery* (2022).
163. McFarthing, K. et al. Parkinson's disease drug therapies in the clinical trial pipeline: 2022 update. *Journal of Parkinson's Disease*, 1-10 (2022).
164. Cohen, S.I. et al. Proliferation of amyloid- β 42 aggregates occurs through a secondary nucleation mechanism. *Proceedings of the National Academy of Sciences* **110**, 9758-9763 (2013).
165. Linse, S. et al. Kinetic fingerprints differentiate the mechanisms of action of anti-A β antibodies. *Nature Structural & Molecular Biology* **27**, 1125-1133 (2020).

166. Aprile, F.A. et al. Rational design of a conformation-specific antibody for the quantification of A β oligomers. *Proceedings of the National Academy of Sciences* **117**, 13509-13518 (2020).
167. Kulenkampff, K., Wolf Perez, A.-M., Sormanni, P., Habchi, J. & Vendruscolo, M. Quantifying misfolded protein oligomers as drug targets and biomarkers in Alzheimer and Parkinson diseases. *Nature Reviews Chemistry* **5**, 277-294 (2021).
168. Orte, A. et al. Direct characterization of amyloidogenic oligomers by single-molecule fluorescence. *Proceedings of the National Academy of Sciences* **105**, 14424-14429 (2008).
169. Sahoo, B., Drombosky, K.W. & Wetzel, R. Fluorescence Correlation Spectroscopy: A Tool to Study Protein Oligomerization and Aggregation In Vitro and In Vivo. in *Protein Amyloid Aggregation: Methods and Protocols* (ed. Eliezer, D.) 67-87 (Springer New York, New York, NY, 2016).
170. Dresser, L. et al. Amyloid-β oligomerization monitored by single-molecule stepwise photobleaching. *Methods* **193**, 80-95 (2021).
171. Lee, J.E. et al. Mapping Surface Hydrophobicity of α-Synuclein Oligomers at the Nanoscale. *Nano Lett* **18**, 7494-7501 (2018).
172. Ruggeri, F.S. et al. Identification and nanomechanical characterization of the fundamental single-strand protofilaments of amyloid α-synuclein fibrils. *Proceedings of the National Academy of Sciences* **115**, 7230-7235 (2018).
173. Chen, K., Bell, N.A.W., Kong, J., Tian, Y. & Keyser, U.F. Direction- and Salt-Dependent Ionic Current Signatures for DNA Sensing with Asymmetric Nanopores. *Biophysical Journal* **112**, 674-682 (2017).
174. Liu, H., Zhou, Q., Wang, W., Fang, F. & Zhang, J. Solid-State Nanopore Array: Manufacturing and Applications. *Small* **19**, 2205680 (2023).
175. Afshar Bakshloo, M. et al. Nanopore-Based Protein Identification. *Journal of the American Chemical Society* **144**, 2716-2725 (2022).
176. Afshar Bakshloo, M. et al. Discrimination between Alpha-Synuclein Protein Variants with a Single Nanometer-Scale Pore. *ACS Chemical Neuroscience* **14**, 2517-2526 (2023).
177. Kowalczyk, S.W., Hall, A.R. & Dekker, C. Detection of Local Protein Structures along DNA Using Solid-State Nanopores. *Nano Letters* **10**, 324-328 (2010).

178. Zeng, X. et al. Nanopore technology for the application of protein detection. *Nanomaterials* **11**, 1942 (2021).
179. Yusko, E.C. et al. Real-time shape approximation and fingerprinting of single proteins using a nanopore. *Nature Nanotechnology* **12**, 360-367 (2017).
180. Larkin, J., Henley, R.Y., Muthukumar, M., Rosenstein, Jacob K. & Wanunu, M. High-Bandwidth Protein Analysis Using Solid-State Nanopores. *Biophysical Journal* **106**, 696-704 (2014).
181. Hu, R. et al. Intrinsic and membrane-facilitated α-synuclein oligomerization revealed by label-free detection through solid-state nanopores. *Scientific reports* **6**, 20776 (2016).
182. Bell, N.A. & Keyser, U.F. Digitally encoded DNA nanostructures for multiplexed, single-molecule protein sensing with nanopores. *Nature nanotechnology* **11**, 645-651 (2016).
183. Sandler, S.E. et al. Sensing the DNA-mismatch tolerance of catalytically inactive Cas9 via barcoded DNA nanostructures in solid-state nanopores. *Nat Biomed Eng* (2023).
184. Chen, K. et al. Digital Data Storage Using DNA Nanostructures and Solid-State Nanopores. *Nano Letters* **19**, 1210-1215 (2019).
185. Chen, S.W. et al. Structural characterization of toxic oligomers that are kinetically trapped during α-synuclein fibril formation. *Proceedings of the National Academy of Sciences* **112**, E1994-E2003 (2015).
186. Chen, S.W. & Cremades, N. Preparation of α-synuclein amyloid assemblies for toxicity experiments. *Amyloid Proteins: Methods and Protocols*, 45-60 (2018).
187. Xu, C.K. et al. α-Synuclein oligomers form by secondary nucleation. *bioRxiv* (2023).
188. Krainer, G. et al. Direct digital sensing of protein biomarkers in solution. *Nature Communications* **14**, 653 (2023).
189. Dada, S.T. et al. Spontaneous nucleation and fast aggregate-dependent proliferation of α-synuclein aggregates within liquid condensates at neutral pH. *Proceedings of the National Academy of Sciences* **120**, e2208792120 (2023).
190. Patterson, J.T. et al. Chemically generated IgG2 bispecific antibodies through disulfide bridging. *Bioorganic & Medicinal Chemistry Letters* **27**, 3647-3652 (2017).
191. Gong, H. et al. Simple method to prepare oligonucleotide-conjugated antibodies and its application in multiplex protein detection in single cells. *Bioconjugate chemistry* **27**, 217-225 (2016).

192. Dear, A.J. et al. Kinetic diversity of amyloid oligomers. *Proceedings of the National Academy of Sciences* **117**, 12087-12094 (2020).

193. Neumann, H., Wang, K., Davis, L., Garcia-Alai, M. & Chin, J.W. Encoding multiple unnatural amino acids via evolution of a quadruplet-decoding ribosome. *Nature* **464**, 441-444 (2010).

194. Hampel, H. et al. The amyloid-β pathway in Alzheimer's disease. *Molecular psychiatry* **26**, 5481-5503 (2021).

195. Xiao, Y. et al. Aβ (1–42) fibril structure illuminates self-recognition and replication of amyloid in Alzheimer's disease. *Nature structural & molecular biology* **22**, 499-505 (2015).

196. Jiang, D., Rauda, I., Han, S., Chen, S. & Zhou, F. Aggregation pathways of the amyloid β (1–42) peptide depend on its colloidal stability and ordered β-sheet stacking. *Langmuir* **28**, 12711-12721 (2012).

197. Gentile, F. et al. Deep docking: a deep learning platform for augmentation of structure based drug discovery. *ACS central science* **6**, 939-949 (2020).

198. Janson, J. et al. Spontaneous diabetes mellitus in transgenic mice expressing human islet amyloid polypeptide. *Proceedings of the National Academy of Sciences* **93**, 7283-7288 (1996).

199. Permert, J. et al. Islet amyloid polypeptide in patients with pancreatic cancer and diabetes. *New England Journal of Medicine* **330**, 313-318 (1994).

200. Rodriguez Camargo, D.C. et al. Surface-catalyzed secondary nucleation dominates the generation of toxic IAPP aggregates. *Frontiers in Molecular Biosciences* **8**, 1037 (2021).

201. Röder, C. et al. Cryo-EM structure of islet amyloid polypeptide fibrils reveals similarities with amyloid-β fibrils. *Nature structural & molecular biology* **27**, 660-667 (2020).

202. Jack Jr, C.R. et al. NIA-AA research framework: toward a biological definition of Alzheimer's disease. *Alzheimer's & Dementia* **14**, 535-562 (2018).

203. Braak, H. & Braak, E. Staging of Alzheimer's disease-related neurofibrillary changes. *Neurobiology of aging* **16**, 271-278 (1995).

204. Brettschneider, J., Tredici, K.D., Lee, V.M.-Y. & Trojanowski, J.Q. Spreading of pathology in neurodegenerative diseases: a focus on human studies. *Nature Reviews Neuroscience* **16**, 109-120 (2015).

205. Shi, Y. et al. Structure-based classification of tauopathies. *Nature* **598**, 359-363 (2021).
206. Lövestam, S. et al. Assembly of recombinant tau into filaments identical to those of Alzheimer's disease and chronic traumatic encephalopathy. *Elife* **11**, e76494 (2022).
207. Fitzpatrick, A.W. et al. Cryo-EM structures of tau filaments from Alzheimer's disease. *Nature* **547**, 185-190 (2017).
208. Lövestam, S. et al. Seeded assembly in vitro does not replicate the structures of α-synuclein filaments from multiple system atrophy. *FEBS open bio* **11**, 999-1013 (2021).
209. Schweighauser, M. et al. Structures of α-synuclein filaments from multiple system atrophy. *Nature* **585**, 464-469 (2020).
210. Meisl, G. et al. Molecular mechanisms of protein aggregation from global fitting of kinetic models. *Nature protocols* **11**, 252-272 (2016).
211. Rodriguez Camargo, D.C. et al. Proliferation of tau 304–380 fragment aggregates through autocatalytic secondary nucleation. *ACS Chemical Neuroscience* **12**, 4406-4415 (2021).
212. Lövestam, S. et al. Disease-specific tau filaments assemble via polymorphic intermediates. *bioRxiv*, 2023.07. 24.550295 (2023).
213. Le Guilloux, V., Schmidtke, P. & Tuffery, P. Fpocket: an open source platform for ligand pocket detection. *BMC bioinformatics* **10**, 1-11 (2009).
214. Nizynski, B., Dzwolak, W. & Nieznanski, K. Amyloidogenesis of Tau protein. *Protein Science* **26**, 2126-2150 (2017).
215. Habchi, J. et al. Systematic development of small molecules to inhibit specific microscopic steps of Aβ42 aggregation in Alzheimer's disease. *Proceedings of the National Academy of Sciences* **114**, E200-E208 (2017).
216. Cummings, J. et al. Alzheimer's disease drug development pipeline: 2022. *Alzheimer's & Dementia: Translational Research & Clinical Interventions* **8**, e12295 (2022).
217. Studier, F.W. Protein production by auto-induction in high-density shaking cultures. *Protein expression and purification* **41**, 207-234 (2005).
218. Zhang, W. et al. Heparin-induced tau filaments are polymorphic and differ from those in Alzheimer's and Pick's diseases. *Elife* **8**, e43584 (2019).
219. Irwin, J.J. & Shoichet, B.K. ZINC− a free database of commercially available compounds for virtual screening. *Journal of chemical information and modeling* **45**, 177-182 (2005).

220. Wager, T.T., Hou, X., Verhoest, P.R. & Villalobos, A. Central nervous system multiparameter optimization desirability: application in drug discovery. *ACS chemical neuroscience* **7**, 767-775 (2016).

221. Trott, O. & Olson, A.J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry* **31**, 455-461 (2010).

222. Frangoul, H. et al. CRISPR-Cas9 gene editing for sickle cell disease and β-thalassemia. *New England Journal of Medicine* **384**, 252-260 (2021).

223. Ramamoorth, M. & Narvekar, A. Non viral vectors in gene therapy-an overview. *Journal of clinical and diagnostic research: JCDR* **9**, GE01 (2015).

224. Carr, D.R. & Bradshaw, S.E. Gene therapies: the challenge of super-high-cost treatments and how to pay for them. *Regenerative medicine* **11**, 381-393 (2016).

225. Brotzakis, Z.F., Zhang, S. & Vendruscolo, M. AlphaFold Prediction of Structural Ensembles of Disordered Proteins. *bioRxiv*, 2023.01. 19.524720 (2023).

226. Vaswani, A. et al. Attention is all you need. *Advances in neural information processing systems* **30**(2017).

227. Xiang, J. et al. Development of an α-synuclein positron emission tomography tracer for imaging synucleinopathies. *Cell* **186**, 3350-3367. e19 (2023).

228. McGonigle, P. Animal models of CNS disorders. *Biochemical pharmacology* **87**, 140-149 (2014).

229. Koprich, J.B., Kalia, L.V. & Brotchie, J.M. Animal models of α-synucleinopathy for Parkinson disease drug development. *Nature Reviews Neuroscience* **18**, 515-529 (2017).

230. Schneuing, A. et al. Structure-based drug design with equivariant diffusion models. *arXiv preprint arXiv:2210.13695* (2022).

231. Howorka, S. Building membrane nanopores. *Nature nanotechnology* **12**, 619-630 (2017).

232. Goto, Y., Akahori, R., Yanagi, I. & Takeda, K.-i. Solid-state nanopores towards single-molecule DNA sequencing. *Journal of human genetics* **65**, 69-77 (2020).

233. Restrepo-Pérez, L., Joo, C. & Dekker, C. Paving the way to single-molecule protein sequencing. *Nature nanotechnology* **13**, 786-796 (2018).

234. Alfaro, J.A. et al. The emerging landscape of single-molecule protein sequencing technologies. *Nature methods* **18**, 604-617 (2021).

235. Vicente, A.M., Ballensiefen, W. & Jönsson, J.-I. How personalised medicine will transform healthcare by 2030: the ICPerMed vision. *Journal of Translational Medicine* **18**, 1-4 (2020).

236. Bermejo, P. et al. Differences of peripheral inflammatory markers between mild cognitive impairment and Alzheimer's disease. *Immunology letters* **117**, 198-202 (2008).

237. Grill, J.D. & Karlawish, J. Implications of FDA approval of a first disease-modifying therapy for a neurodegenerative disease on the design of subsequent clinical trials. *Neurology* **97**, 496-500 (2021).

238. McGann, M. FRED and HYBRID docking performance on standardized datasets. *Journal of computer-aided molecular design* **26**, 897-906 (2012).

239. Wang, Z. et al. Comprehensive evaluation of ten docking programs on a diverse set of protein–ligand complexes: the prediction accuracy of sampling power and scoring power. *Physical Chemistry Chemical Physics* **18**, 12964-12975 (2016).

240. Robinson, C. & Dilkina, B. A machine learning approach to modeling human migration. in *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies* 1-8 (2018).

241. Straathof, S. et al. Protein Sizing with 15 nm Conical Biological Nanopore YaxAB. *ACS Nano* **17**, 13685-13699 (2023).