A Machine Learning Approach to Identify Small Molecule Inhibitors of Secondary Nucleation in α-Synuclein Aggregation

Robert I. Horne[1], Andrea Possenti[1], Sean Chia[1,2], Z. Faidon Brotzakis[1], Roxine Staats[1], Magdalena Nowinska[1], Michele Vendruscolo[1+]

[1]Centre for Misfolding Diseases, Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, UK

*[2]Bioprocessing Technology Institute, Agency of Science, Technology and Research (A*STAR), Singapore*

[+]mv245@cam.ac.uk

Keywords

Parkinson's disease; α-synuclein; protein aggregation; machine learning; computational docking; structure-based drug discovery; kinetic-based drug discovery

Abstract

Drug development is an increasingly active area of application of machine learning methods, due to the need to overcome the high attrition rates of conventional drug discovery pipelines. This issue is especially pressing for neurodegenerative diseases where very few disease-modifying drugs have been approved. To address this problem, we describe a machine learning approach to identify specific inhibitors of the proliferation of α-synuclein aggregates through secondary nucleation, a process that has been implicated in Parkinson's disease and related synucleinopathies. We use a combination of docking simulations followed by machine learning to first identify initial hit compounds and then explore the chemical space around these compounds. Our results demonstrate that this approach leads to the identification of novel chemical matter with an improved hit rate and potency over conventional similarity search approaches.

Introduction

Parkinson's disease is the most common neurodegenerative movement disorder, estimated to affect 2–3% of the population over 65 years of age[1–4]. The aggregation of the intrinsically disordered protein α-synuclein is thought to be responsible for the initial neurodegenerative processes underlying this disease, in which the pathological accumulation of misfolded protein molecules results in neuronal toxicity beginning in the substantia nigra region of the brain[1–3,5]. This hypothesis is supported by genetic evidence and by observations of the accumulation of α-synuclein in inclusions known as Lewy bodies in the neurons of Parkinson's disease patients[5–7]. The aggregates of α-synuclein, including insoluble fibrils and misfolded soluble oligomers, have been shown to exhibit various mechanisms of cellular toxicity including cell membrane disruption[8,9]. The pathological relevance of the aberrant aggregation of α-synuclein has led to major efforts being invested into studying its mechanisms and identifying molecules that can inhibit those aggregation mechanisms associated with neurotoxicity[10–13]. This is a particularly pressing need given the lack of therapies and diagnostics currently available to PD patients[14–16].

Computational methods in this area could in principle reduce the time and cost of traditional drug discovery pipelines[17–19]. In this area, machine learning is rapidly emerging as a powerful drug discovery strategy[20]. To explore the potential of this strategy in drug discovery programs for Parkinson's disease and other protein misfolding diseases, we describe here a machine learning approach to effectively explore the chemical space to identify compounds that inhibit the aggregation of α-synuclein. Our starting point is an approach based on the combination of docking simulations with in vitro screening that was recently employed to identify a set of compounds that bind to the fibril structures of α-synuclein, and prevent the autocatalytic proliferation of α-synuclein fibrils as a result[21]. Here, we used this initial set of compounds as input for a machine learning approach to identify chemical matter that is both efficacious and represents a significant departure from the parent structures, providing novel compounds that conventional similarity searches would have failed to efficiently identify.

The machine learning approach used here is derived from one recently developed to identify kinase inhibitors[22], and consists of 3 main components: (1) the algorithms required to represent the structures of the molecules, (2) a model for training and prediction using these representations and the assay metrics, and (3) an experimental method to test the model and feed results back into it. This set up is illustrated in **Fig. S1**. Component 1 is a junction tree variational autoencoder[23], pre-trained on a set of 250,000 molecules[24]. Component 2 consists of a shallow learning algorithm utilising an initial random forest regressor (RFR) with a Gaussian process regressor (GPR) fitted to the residuals[25,26]. Since Gaussian processes work effectively on small datasets, this approach enables the leveraging of the associated uncertainty measure when ranking molecules during acquirement prioritisation[22]. Depending on how conservative the experimentalist wishes to be, a greater or lesser weighting can be applied to this uncertainty, allowing a ranking based on both the predicted potency of the molecules and the uncertainty of that prediction. The last component is a chemical kinetics assay[10,27,28], which identifies the top compounds that significantly inhibit the surface-catalysed secondary nucleation step in the aggregation of α-synuclein.

The primary aim of the machine learning approach presented here is to open a route to obtain a greater hit rate and novelty than high-throughput screening, docking simulations and similarity searches. Overall, this work represents the rational development of an approach that combines chemical kinetics and machine learning for identification and optimisation of compounds that can inhibit potently and selectively the secondary nucleation of α-synuclein, which could be of use in therapeutic programs targeting synucleinopathies in the future.

Results

A machine learning method of exploring the compound space

The overall approach is illustrated schematically in **Fig. 1**. A docking screening was previously carried out[21] using a combination of AutoDock Vina[29] and FRED[30] software, and subsequent experiments identified 4 active compounds (labelled 48, 52, 68 and 69, referred to as the 'docking set') (**Fig. 1A**). Here, using the Tanimoto similarity metric[29],

two similarity searches were then carried out using these four structures as starting points. A selection of closely related molecules (Tanimoto similarity > 0.5) to the parent compounds (referred to as the 'close similarity docking set', **Fig. 1B**) was experimentally tested in an aggregation assay (**Fig. 1C**). 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', **Fig. 1B**). The compounds resulting from these experiments were then used as input for a machine learning method for an iterative exploration of the chemical space (**Fig. 1B**). The similarity searches removed the most obvious targets of the machine learning approach, but also increased the size of the dataset available for training. The training set, however, remained small by typical machine learning standards, as it consisted in a few hundred molecules. Since training set sizes such as the one used here are typical in early-stage research, a further aim of this work was to demonstrate that machine learning can be used rather effectively even in such data sparse scenarios.

Selected molecules in the close similarity docking set were tested in an aggregation assay, which yielded 5 new hits from 25 new molecules (**Fig. S2**). Additional selected molecules in the loose similarity docking set were also tested in this assay in order to expand the training set beyond the initial docking compounds and their closely related derivatives. Although new hits featured amongst this set, the hit rate was low, 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 compound 48 in the loose similarity docking set, none were hits, while from the 24 molecules related to compound 52, only 2 were hits. The functional range of molecules 48 and 52 appeared narrowly limited around the chemical space of the parent structures. Overall, the hit rate from the loose similarity docking set was less than a quarter of that of the close similarity docking set and involved testing three times as many compounds.

At this point it would have been challenging to further explore the chemical space using conventional structure-activity relationship (SAR) techniques without significant attrition, since as discussed above the hit rate worsened as the similarity constraint to the hits was loosened.

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 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 one 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 dataset of ~9000 compounds which comprised the 'test set'.

Different machine learning models were then trialled against the docking scores calculated for the test set, and the best performing model was trained on the whole aggregation data set and used to predict the top set of molecules (see Machine Learning Implementation in Supplementary Information, and **Figs. S3** and **S4**). For this work, the half time of aggregation 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[31] (**Fig. S5**). Molecules that achieved a half time 2-fold greater than that of the negative control under standard assay conditions (see Methods) were classed as hits[21]. The algorithm was run 10 times from different random starting states and those molecules that appeared in the top 100 ranked molecules more than 50% of the times (64 molecules) were chosen for purchase (first iteration). In this first iteration, there was an inherent bias towards the structure of compound 69 in the dataset given the relative population sizes (**Fig. S6**), but with the caveat that many of these structures were only loosely related to the parent (Tanimoto similarity < 0.4). Many of the hit molecules came from this group, suggesting significant chemical departures from the parent structure.

After the first iteration, the compounds were pooled together to extend the training set and further two iterations were carried out. Example kinetic traces for one molecule are shown in **Fig. 2A**, the molecules being labelled according to iteration number and hit

identifier within that iteration, I1.01 being the first hit (01) within iteration 1 (I1). The aggregation data from the first 3 prediction iterations are also shown in **Fig. 2B**. Of the 64 molecules from iteration 1, 8 were strong hits, representing a hit rate of 12.5%, the second iteration showed a further increase, with 11 strong hits representing a 17% hit rate and the third iteration, with 12 hits, exhibited a hit rate of 21%. These hit rates represent an order of magnitude improvement over HTS (~1%) and, remarkably, an overall 40% improvement over the combined similarity search hit rates, which removed the most likely hit candidates. The potency of these hits was also significantly higher on average than those identified by the similarity searches (**Fig. 3A**), without compromising the CNS-MPO[41] scores (**Fig. 3B**).

Inspections of the chemical space explored by the machine learning approach via dimensionality reduction techniques such as PCA, t-SNE and UMAP (see Methods) were also carried out to investigate how the model was prioritising molecules (**Fig. S7**). The relative positioning of the training points and the parents within the chemical space is shown in **Fig. 4**. 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 (**Fig. 5**). 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 hit acquirement while still exploring the chemical space.

The changes in similarity of the hits to the parent structures are shown in **Fig. 6**. The similarity dropped for all structures at successive stages of the investigation, reaching its lowest point at the initial iterations of the machine learning approach. The most potent hits mostly retained the left-hand side scaffold of molecule 69 with the addition of polar groups to the benzene ring, but with significant alterations to the right-hand side of the scaffold. For example, from iteration 1, I1.01 replaced the right-hand side of molecule 69 with a single substituted benzene ring, while I1.02 replaced the right-hand side with a substituted furan ring. These changes were reflected in the Tanimoto similarity values, which were at the lower end of what was permitted in the test 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.

Validation experiments

A series of validation experiments were carried out on the most potent hits from the machine learning iterations. We first tested the binding to fibrils using the change in fluorescence polarisation under a titration of fibrils (**Fig. S8**). One example, molecule I3.08, which had the appropriate fluorescence properties for the assay, was investigated and exhibited a K_D of ~ 700 nM.

The dose dependent potency in a light seeded aggregation assay was also investigated (**Fig. S9**), with all hit molecules exhibiting substoichiometric potency. As transient oligomeric species are considered the most damaging of the aggregate species *in vivo*[8,9] oligomer flux simulations are also shown in **Fig. S9**, with all hits demonstrating dose dependent delay and reduction of the oligomer peak. The calculations and input to derive these are shown in the Methods. The last metric extracted from these light seeded experiments was an approximate overall rate at each molecule concentration, obtained by taking $1/t_{1/2}$, as shown in **Fig. S10**. These approximate rates were fitted to a Hill slope which was then used to obtain the 50% kinetic inhibitory concentration (KIC_{50}), also shown in **Fig. S10**. The more potent hits had KIC_{50}s in the region of 1 μM, translating to a significant potency at a ratio of 1:10 molecule to α-synuclein monomer.

Finally, transmission electron microscopy imaging of the fibrils produced at the end of the light seeded aggregation reaction was carried out (**Fig. S11**), to verify fibrils were produced, and mass spectrometry was carried out on α-synuclein samples after incubation with hit molecules. No change was observed in the mass of the α-synuclein peak for the molecule incubations compared with a 1% DMSO control, excluding covalent modification as a mechanism.

Discussion

Chemical kinetics approaches have advanced to the point that specific mechanisms of α-synuclein aggregation can be targeted in a reproducible way[10,27,28]. The mechanism targeted in this work is the surface-catalysed secondary nucleation step, which is responsible for the autocatalytic proliferation of α-synuclein fibrils. In a recent proof-of-principle report, initial hit molecules identified via docking simulations were shown to bind competitively with α-synuclein monomers along specific sites on the surface of α-synuclein fibrils[21,32,33]. 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[10,34].

The aim of this work was to develop a machine learning approach to drug discovery for protein aggregation diseases that could improve both the hit rate of the in vitro assays employed and provide novel chemical matter more efficiently than conventional approaches. As of the first iterations, the hit 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%)[35]. 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 are ~4000 molecules in the test set that have Tanimoto similarity values in the range of these hits, and all of these would potentially have had to be screened to locate these hits using similarity searches alone, as demonstrated by the looser similarity search approach which exhibited a comparatively poor hit rate (4%) despite 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 hit rate.

A limitation of this approach is the requirement to pick from a pre-existing library. To address this limitation generative models could be employed[36]. A second limitation is the focus on one assay metric of interest as a learning parameter. Addressing this limitation involves multi-parameter optimisation, which is a challenging area in rapid development[37–39]. Another topic of great interest in machine learning drug discovery approaches besides potency prediction is the prediction of pharmacokinetics and toxicity[40]. It could be possible to achieve this multi-parameter optimisation utilising

multiple models in parallel and then employing a joint ranking metric, or architectures such as generative adversarial networks may also be capable of achieving this in a single model, although this has primarily been demonstrated with predicted chemical properties such as clogP and QED rather than experimental results[37–39]. The molecules in this work were derived from a set that passed CNS-MPO[41] criteria in the initial docking simulation, and so the CNS-MPO score of the whole aggregation inhibitor set is relatively favourable with most molecules exceeding the common cut off value of 4[41] (**Figs**. **5B** and **S12**).

Conclusions

The results that we have presented illustrate the potential of a drug discovery approach that involves the combination of chemical kinetics and machine learning. The hits identified by this combined approach offered a significant improvement in potency over the parent molecules and represented a major structural departure from them. We anticipate that using machine learning approaches of the type described here could be of significant benefit to researchers working in this field and indeed early-stage research in general.

Materials and Methods

Compounds and chemicals

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

Recombinant α-synuclein expression

Recombinant α-synuclein was purified as described previously[42–44]. The plasmid pT7-7 encoding for human α-synuclein was transformed into BL21-competent cells. Following transformation, competent cells were grown in LB in the presence of ampicillin (100 µg/mL). Cells were induced with IPTG and grown overnight at 37 ^{o}C and harvested by

centrifugation in a Beckman Avanti J25 centrifuge with a JA-20 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 multiple freeze–thaw cycles and sonication. The cell suspension was boiled for 20 min and centrifuged at 13,500 rpm with a JA-20 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 °C. After centrifugation at 13,500 rpm, the supernatant was taken with an addition of 0.36 g/mL ammonium sulfate. The solution was stirred for 30 min at 4 °C and centrifuged again at 13,500 rpm. The pellet was resuspended in 25 mM Tris, pH 7.7, and ion-exchange chromatography was performed using a HQ/M-column of buffer A (25 mM Tris, pH 7.7) and buffer B (25 mM Tris, pH 7.7, 600 mM NaCl). The fractions containing α-synuclein (≈ 300 μM) were dialysed overnight against the appropriate buffer. The protein concentration was determined spectrophotometrically using $\varepsilon 280 = 5600\ M^{-1}\ cm^{-1}$.

Seed fibril preparation

α-Synuclein fibril seeds were produced as described previously[43,44]. Samples of α-synuclein (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.

Measurement of aggregation kinetics

α-Synuclein 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 α-synuclein 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 °C 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.

Transmission electron microscopy

10 μM α-synuclein 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.

Determination of the elongation rate constant

In the presence of high concentrations of seeds (≈ μM), the aggregation of α-synuclein is dominated by the elongation of the added seeds[43,44]. Under these conditions where other microscopic processes are negligible, the aggregation kinetics for α-synuclein can be described by[43,44]

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

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 α-synuclein 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).

Determination of the 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[10,45]

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

where m_{tot} denotes the total concentration of α-synuclein 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[45]

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

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 α-synuclein in the presence of the compounds as compared to in its absence (1% DMSO).

Determination of the oligomer flux over time

The theoretical prediction of the reactive flux towards oligomers over time was calculated as[10,45]

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

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.

Fluorescence polarization

10 μM of Hit C was incubated with increasing concentrations of either pre-formed α-synuclein fibrils (in 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 Hit C 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 450 nm emission filter.

Mass spectrometry

10 μM of preformed α-synuclein 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).

Machine learning

Junction tree neural network variational autoencoder. The autoencoder[23] was pretrained on a library of 250,000 compounds[24], and was implemented as described in ref. 23 using a pip installable version (https://github.com/LiamWilbraham/jtnnencoder).

Protein representation model. A protein representation model was tested although not used in this work. A soft symmetric alignment model (λ= 0.1)[46] was pre-trained on

the Pfam[47] database, and the implementation used here involved averaging over output positions.

Prediction module. All coding was carried out in Python 3. Scikit-learn[48] implementations of the Gaussian process regressor (GPR), random forest regressor (RFR) and multiple linear regression MLP methods were tested in various combinations, and the results are outlined in the results section. For data handling, calculations and graph visualisation the following software and packages were used: pandas[49], seaborn[50], matplotlib[51], numpy[52], scipy[53], umap-learn[54], Multicore-TSNE[55] 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.

Acknowledgements

The authors would like to give especial thanks to Dr Becky Gregory for providing α-synuclein, Dr Laila Sakhnini for carrying out the mass spectrometry, Dr Heather Greer for assisting with the transmission electron microscopy (TEM) and EPSRC Underpinning Multi-User Equipment Call (EP/P030467/1) for funding the TEM. We would also like to thank ARCHER and MARCOPOLO high performance computing resources for the computer time. Z. Faidon Brotzakis would like to acknowledge the Federation of European Biochemical Societies (FEBS) for financial support (LTF).

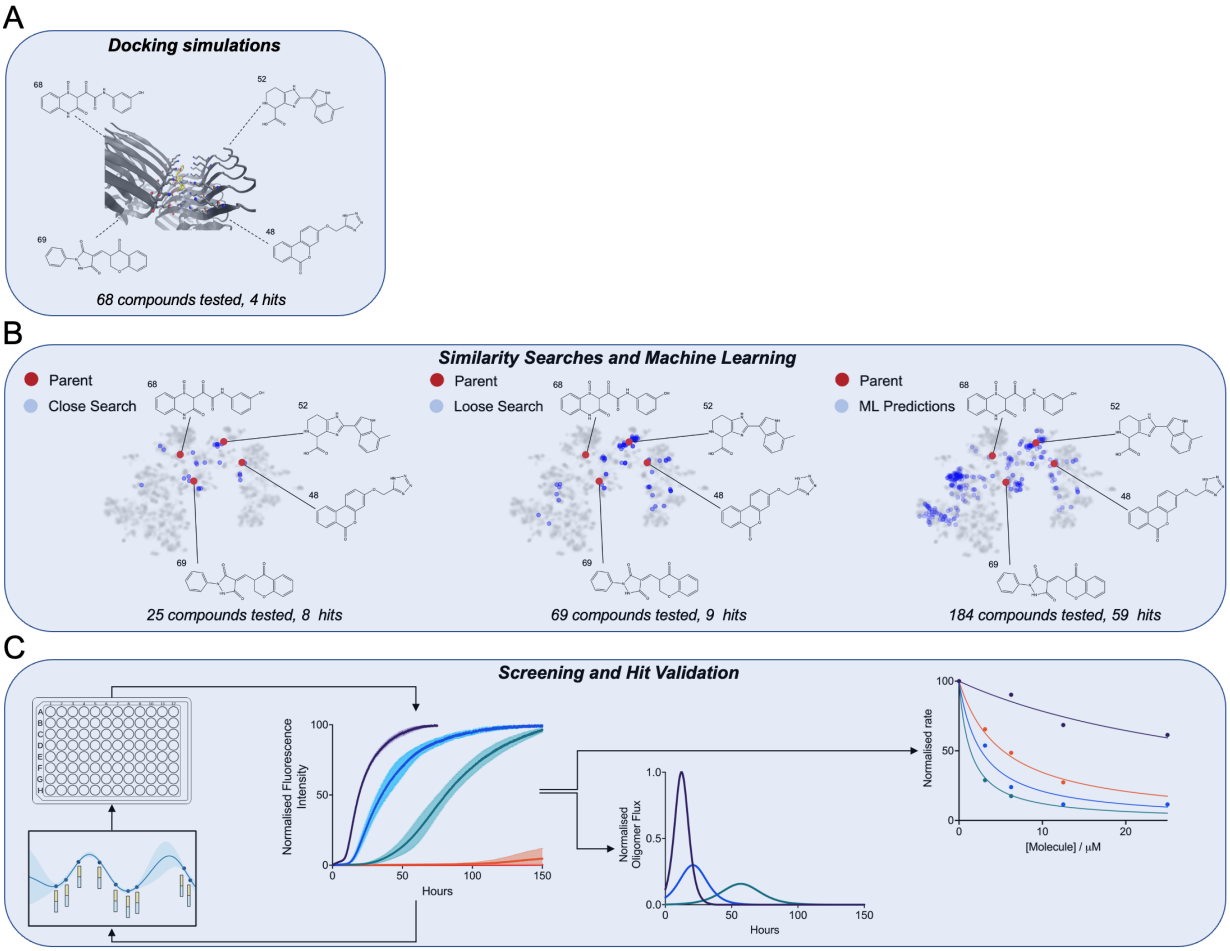


Figure 1. Illustration of the three stages of exploration of the chemical space. (A) We initially identified 4 active molecules (the 'docking set') from the 68 molecules tested experimentally, which were selected from the top hits from a docking screening. These 4 molecules increase the half time of α-synuclein aggregation. **(B)** We then performed a close similarity search around the 4 parent compounds in chemical space selected molecules with Tanimoto similarity cut off > 0.5 (the 'close similarity docking set') and a loose similarity search with Tanimoto similarity cut off > 0.4 (the 'loose similarity docking set'). A machine learning method was then applied using the observed data to predict hits from a compound library derived from the ZINC database with Tanimoto similarity > 0.3 to the parent structures (the 'test set'). **(C)** Successive iterations of prediction and experimental testing yielded higher hit rates, and molecules with higher potency on average than those identified in the previous similarity searches. Validation experiments were also carried out on the hits identified.

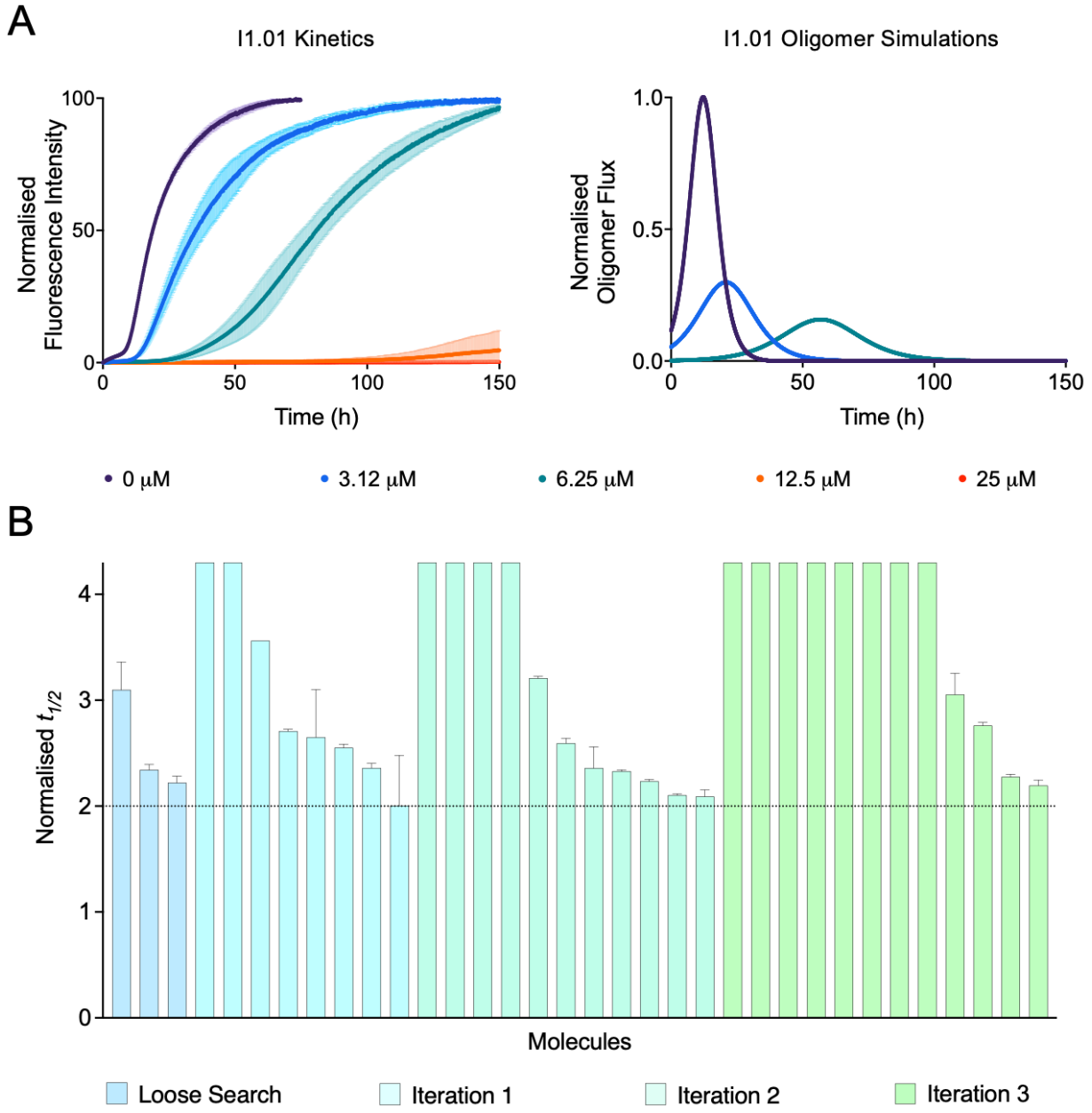


Figure 2. Results of the three iterations of the machine learning drug discovery approach. (A) Kinetic traces of a 10 μM solution of α-synuclein in the presence of 25 nM seeds at pH 4.8, 37 °C in the presence of molecules or 1% DMSO. During the initial screening, all molecules were screened at 2.5 molar equivalents (25 μM), and hits were then taken for further validation at 4 different concentrations: 25 μM (red), 12.5 μM (orange), 6.25 μM (teal) and 3.12 μM (blue). Molecule I1.01 is shown as an exemple. The endpoints are normalised to the α-synuclein monomer concentration at the end of the experiment, which was detected via the Pierce™ BCA Protein Assay at t = 70 h.

Oligomer flux simulations derived from the kinetics are also shown. **(B)** Normalised half times of aggregation for the hits from the different stages: loose search, iteration 1, iteration 2 and iteration 3. The horizontal dotted line indicates the boundary for hit classification, which was normalized $t_{1/2}$ = 2. For the loose search, 69 molecules were tested, while for iterations 1, 2 and 3, the number of molecules tested were 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.

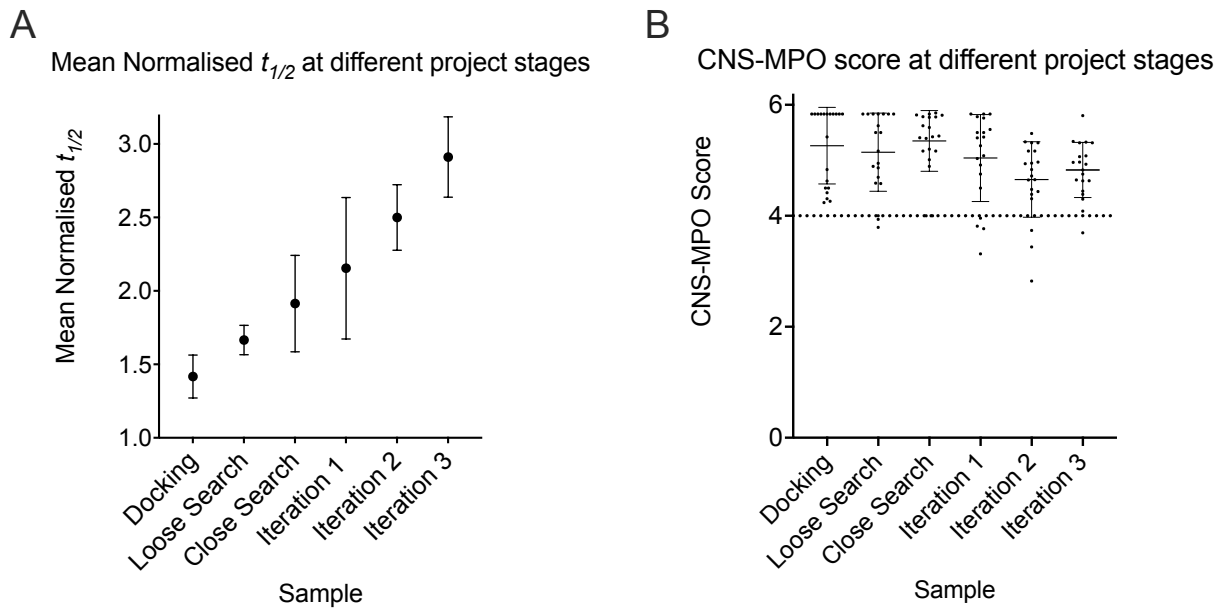


Figure 3. Average half time of aggregation and CNS-MPO scores for the top 20 molecules at each stage. **(A)** The stages are the initial docking simulation (68 molecules tested), loose search (69 molecules tested), close search (25 molecules tested), iteration 1 (64 molecules tested), iteration 2 (64 molecules tested) and iteration 3 (56 molecules tested). **(B)** A common cut off for CNS-MPO score is 4, as indicated by the horizontal dotted line.

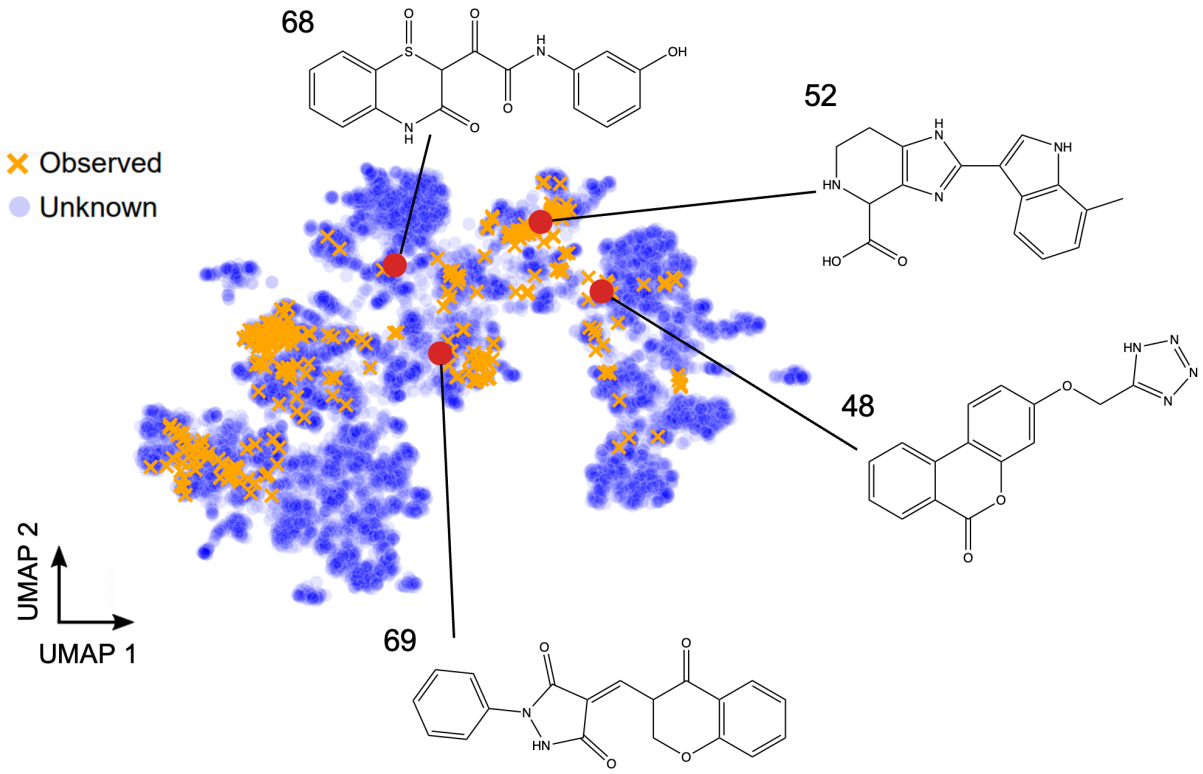


Figure 4. UMAP visualisation of the compound feature space. The visualisation indicates the areas of the chemical space that have been explored (orange crosses) starting from the 4 initial molecules inthe docking set, and the relative positioning of the parent structures in this space.

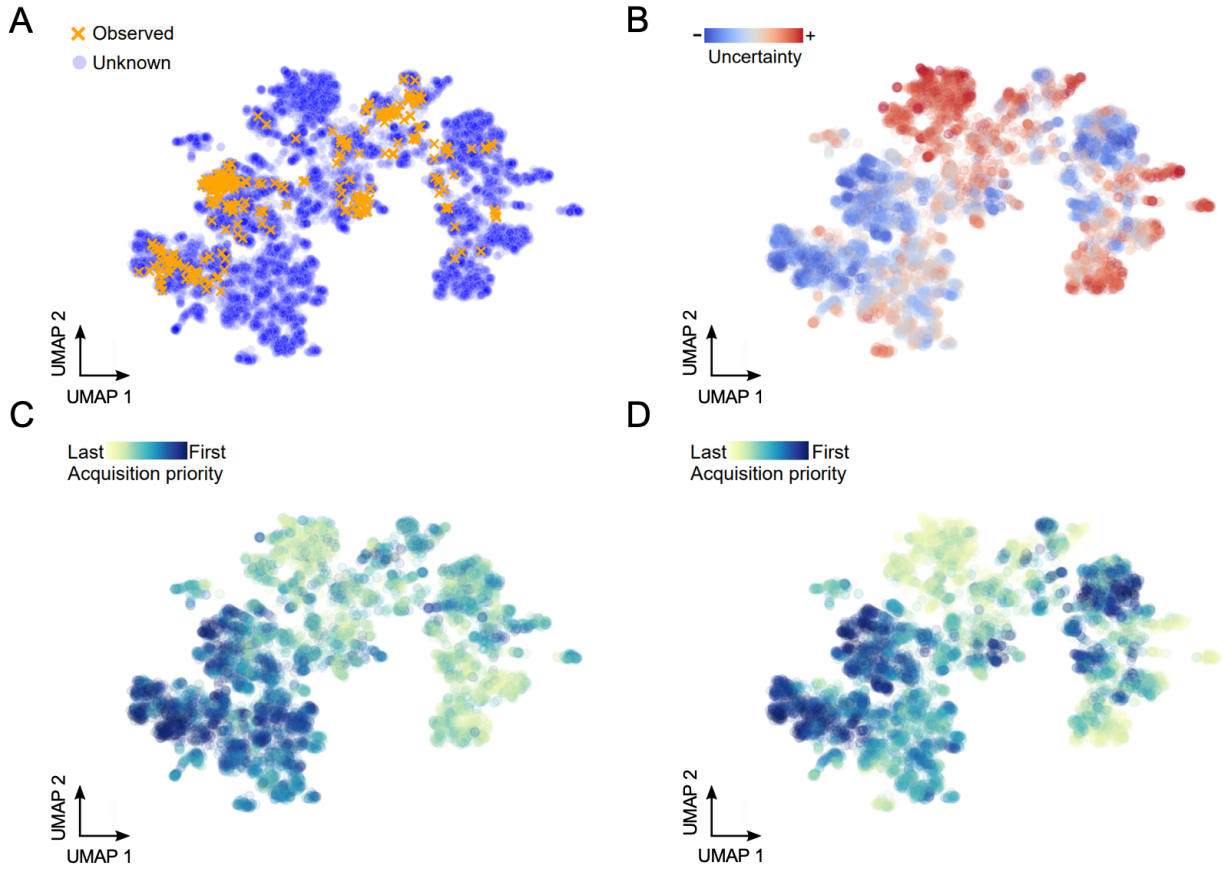


Figure 5. UMAP visualisation of the compound feature space using uncertainty. (A) Areas of the compound space that have been explored (orange crosses) and that have not been explored (blue circles). **(B)** GPR assigned lower uncertainty (blue) to regions of the chemical space near to the observed data and high uncertainty (red) to areas which were further away. **(C)** 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.

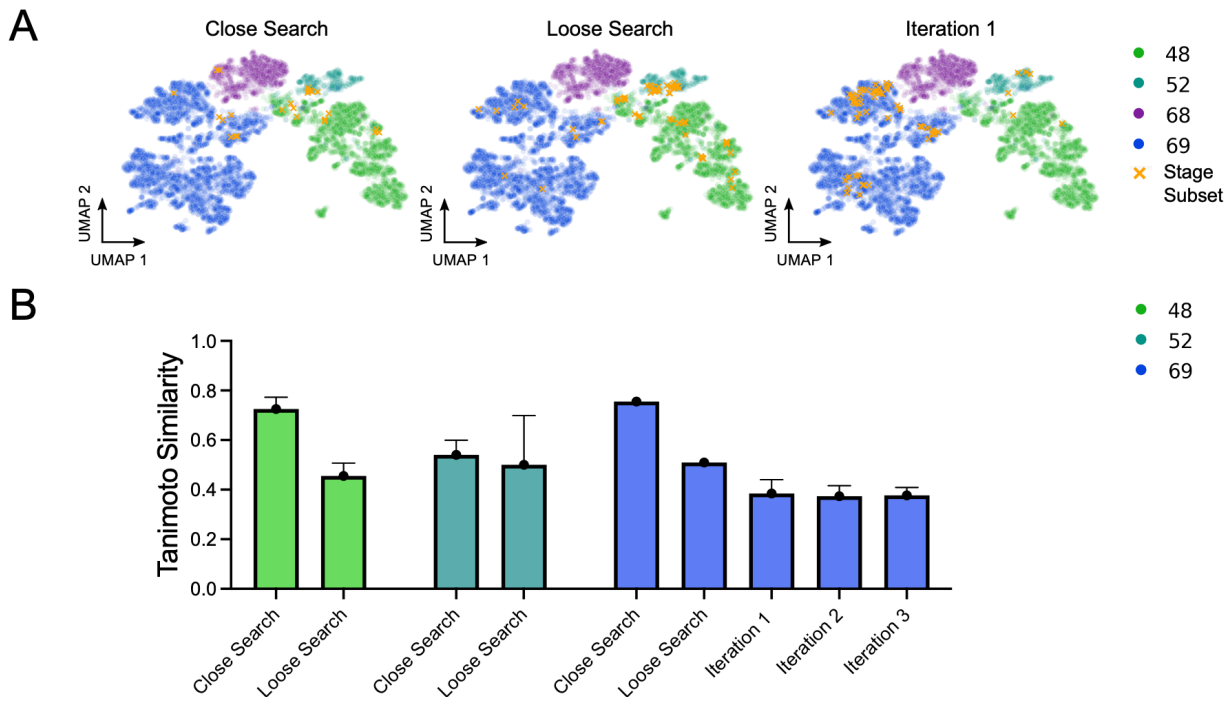


Figure 6. Representation of the structural changes during the machine learning compound optimisation. (A) UMAP visualisation of the compound space indicating how the positioning of each new molecule subset (orange crosses) changed at each stage of the project as well as how the chemical landscape was split between the parent molecules (different colours) The locations of the parent molecules are also indicated in the 'Docking' pane (red circles). **(B)** Tanimoto similarity of the hit molecules derived from molecules 48 (green), 52 (teal) and 69 (blue) at each stage of the project. At iterations 1, 2 and 3 all of the hits were derived from molecule 69, albeit with lower similarity than any of the previous stages. Molecule 68 was not included as it failed to produce any hits outside of the parent molecule.

References

1. Poewe, W. *et al.* Parkinson disease. *Nature Reviews Disease Primers* **3**, 1–21 (2017).
2. Balestrino, R. & Schapira, A. H. V. Parkinson disease. *European Journal of Neurology* **27**, 27–42 (2020).
3. Aarsland, D. *et al.* Parkinson disease-associated cognitive impairment. *Nature Reviews Disease Primers* **7**, (2021).
4. Ray Dorsey, E. *et al.* Global, regional, and national burden of Parkinson's disease, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *The Lancet Neurology* **17**, 939–953 (2018).
5. Spillantini, M. G., Crowther, R. A., Jakes, R., Hasegawa, M. & Goedert, M. Synuclein in filamentous inclusions of Lewy bodies from Parkinson's disease and dementia with Lewy bodies. *Proc. Natl. Acad. Sci. USA* **95**, 6469-6473 (1998).
6. Savica, R., Bradley, B. F. & Mielke, M. M. When do α-Synucleinopathies start? An epidemiological timeline a review. *JAMA Neurology* **75**, 503–509 (2018).
7. Chiti, F. & Dobson, C. M. Protein misfolding, functional amyloid, and human disease. *Annual Review of Biochemistry* **75**, 333–366 (2006).
8. Fusco, G. *et al.* Structural basis of membrane disruption and cellular toxicity by α-synuclein oligomers. *Science (New York, N.Y.)* **358**, 1440–1443 (2017).
9. Lashuel, H. A., Overk, C. R., Oueslati, A. & Masliah, E. The many faces of α-synuclein: from structure and toxicity to therapeutic target. *Nature reviews Neuroscience* **14**, 38–48 (2013).
10. 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**, 1–9 (2020).
11. 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. *Scientific Reports* **8**, (2018).
12. Wagner, J. *et al.* Anle138b: A novel oligomer modulator for disease-modifying therapy of neurodegenerative diseases such as prion and Parkinson's disease. *Acta Neuropathologica* **125**, 795–813 (2013).

13. Pujols, J., Peña-Díaz, S., Pallarès, I. & Ventura, S. Chemical Chaperones as Novel Drugs for Parkinson's Disease. *Trends in Molecular Medicine* **26**, 408–421 (2020).
14. Oertel, W. & Schulz, J. B. Current and experimental treatments of Parkinson disease: A guide for neuroscientists. *Journal of Neurochemistry* **139**, 325–337 (2016).
15. McFarthing, K., Rafaloff, G., Baptista, M. A. S., Wyse, R. K. & Stott, S. R. W. Parkinson's Disease Drug Therapies in the Clinical Trial Pipeline: 2021 Update. *Journal of Parkinson's Disease* **11**, 891–903 (2021).
16. Tolosa, S. *et al.* Challenges in the diagnosis of Parkinson's disease. *Lancet Neurol* **20**, (2021).
17. Vamathevan, J. *et al.* Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery* **18**, 463–477 (2019).
18. Panteleev, J., Gao, H. & Jia, L. Recent applications of machine learning in medicinal chemistry. *Bioorganic and Medicinal Chemistry Letters* **28**, 2807–2815 (2018).
19. Meng, X.-Y., Zhang, H.-X., Mezei, M. & Cui, M. Molecular Docking: A powerful approach for structure-based drug discovery. *Current Computer-Aided Drug Design* **7**, (2011).
20. Myszczynska, M. A. *et al.* Applications of machine learning to diagnosis and treatment of neurodegenerative diseases. *Nature Reviews Neurology* **16**, 440–456 (2020).
21. Chia, S. *et al.* Structure-based discovery of small molecule inhibitors of the autocatalytic proliferation of alpha-synuclein aggregates. *Proc. Natl. Acad. Sci. USA, submitted.* (2021).
22. Hie, B., Bryson, B. D. & Berger, B. Leveraging Uncertainty in Machine Learning Accelerates Biological Discovery and Design. *Cell Systems* **11**, 461-476 (2020).
23. Jin, W., Barzilay, R. & Jaakkola, T. Chapter 11: Junction Tree Variational Autoencoder for Molecular Graph Generation. *arXiv: 1802.04364v4* (2019).
24. Kusner, M. J., Paige, B. & Hemández-Lobato, J. M. Grammar variational autoencoder. *34th International Conference on Machine Learning, ICML 2017* **4**, 3072–3084 (2017).
25. Rasmussen, C. E. & Williams, C. K. I. *Gaussian Processes for Machine Learning*. **7**, (2006).

26. Breiman, L. Random forests. *Machine Learning* **45**, 5–32 (2001).
27. Knowles, T. P. J. *et al.* An Analytical Solution to the Kinetics of Breakable Filament Assembly. *Science* **326**, 1533–1537 (2009).
28. Knowles, T. P. J., Vendruscolo, M. & Dobson, C. M. The amyloid state and its association with protein misfolding diseases. *Nature Reviews Molecular Cell Biology* **15**, 384–396 (2014).
29. 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).
30. Le Guilloux, V., Schmidtke, P. & Tuffery, P. Fpocket: An open source platform for ligand pocket detection. *BMC Bioinformatics* **10**, 1–11 (2009).
31. Cohen, S. I. A., Vendruscolo, M., Dobson, C. M. & Knowles, T. P. J. From macroscopic measurements to microscopic mechanisms of protein aggregation. *Journal of Molecular Biology* **421**, 160–171 (2012).
32. Perni, M. *et al.* Multistep Inhibition of α-Synuclein Aggregation and Toxicity in Vitro and in Vivo by Trodusquemine. *ACS Chemical Biology* **13**, 2308–2319 (2018).
33. Brown, J. W. P. *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).
34. Chia, S. *et al.* SAR by kinetics for drug discovery in protein misfolding diseases. *Proc. Natl. Acad. Sci. USA* **115**, 10245–10250 (2018).
35. Zhu, T. *et al.* Hit Identification and Optimization in Virtual Screening: Practical Recommendations Based on a Critical Literature Analysis. *Journal of Medicinal Chemistry* **56**, 6560–6572 (2013).
36. 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).
37. Maziarka, Ł. *et al.* Mol-CycleGAN: A generative model for molecular optimization. *Journal of Cheminformatics* **12**, (2020).
38. Zhou, Z., Kearnes, S., Li, L., Zare, R. N. & Riley, P. Optimization of Molecules via Deep Reinforcement Learning. *Scientific Reports* **9**, (2019).
39. You, J., Liu, B., Ying, R., Pande, V. & Leskovec, J. Graph Convolutional Policy Network for Goal-Directed Molecular Graph Generation. *arXiv 1806.02473* (2019)

40. Allen, C. H. G. *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).
41. 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).
42. Galvagnion, C. *et al.* Lipid vesicles trigger α-synuclein aggregation by stimulating primary nucleation. *Nature Chemical Biology* **11**, 229–234 (2015).
43. Flagmeier, P. *et al.* Mutations associated with familial Parkinson's disease alter the initiation and amplification steps of α-synuclein aggregation. *Proc. Natl. Acad. Sci. USA* **113**, 10328–10333 (2016).
44. Buell, A. K. *et al.* Solution conditions determine the relative importance of nucleation and growth processes in α-synuclein aggregation. *Proc. Natl. Acad. Sci. USA* **111**, 7671–7676 (2014).
45. Michaels, T. C. T., Cohen, S. I. A., Vendruscolo, M., Dobson, C. M. & Knowles, T. P. J. Hamiltonian Dynamics of Protein Filament Formation. *Physical Review Letters* **116**, 3–8 (2016).
46. Bepler, T. & Berger, B. Learning Protein Sequence Embeddings Using Information From Structure. *arXiv* 1902.08661 (2019).
47. Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Research* **49**, 412–419 (2021).
48. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).
49. McKinney, W. Data Structures for Statistical Computing in Python. *Proceedings of the 9th Python in Science Conference* **1**, 56–61 (2010).
50. Waskom, M. Seaborn: Statistical Data Visualization. *Journal of Open Source Software* **6**, 3021 (2021).
51. Hunter, J. D. Matplotlib: a 2D Graphics Environment. *Computing in Science and Engineering* **9**, 90–95 (2007).
52. Harris, C. R. *et al.* Array programming with NumPy. *Nature* **585**, 357–362 (2020).
53. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods* **17**, 261–272 (2020).
54. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv: 1802.04364v4* (2018).

55. Maaten Van Der, L. Accelerating t-SNE using tree-based algorithms. *Journal of Machine Learning Research* **15**, 3221–3245 (2014).
56. Robinson, C. & Dilkina, B. A machine learning approach to modeling human migration. *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies, COMPASS 2018* (2018) doi:10.1145/3209811.3209868.

SUPPLEMENTARY INFORMATION

Machine Learning Implementation

Before embarking on the experimental work, we tested the machine learning strategy on simulated data. The docking scores already calculated for the training set were calculated also for the test set. Both AutoDock Vina and FRED simulations were carried out on the test set, giving binding scores for each molecule against α-synuclein fibrils. Only the results from the implementation using AutoDock Vina are included below as the FRED results followed similar trends. The compound encoder was implemented as in Hie et al. [22] to obtain representations of all the molecules, although in our work we found the full representation to be the most effective for hit picking compared to the truncated version used in that work. 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 datasets 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.[22] A binding score training set was then used to fit the GPR, which was in turn used to predict the binding scores of the test 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[56]. The training set size was subsequently expanded, and the same process carried out. This was compared with a naïve Bayes, which failed to score above 0 for any training set size.

The GPR kernel was initially the same as that utilised by Hie et al.[22] , 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 small training set sizes. 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 the smaller training sets, but represented a viable starting point.

The same process was utilised using different proportions of the molecular feature set, and it was found that GPR performance metrics were better when using truncated feature sets 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.

To investigate this issue, 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.

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.

A snapshot of the results of this testing is shown in **Figs. S3** and **S4**. **Fig. S3** 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 the 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.

At this point a two-layer model was applied. This reflected the strategy used by Hie et al.[22] in fitting a Gaussian process regressor (GPR) to the residuals of another model, in that case an multiple linear regression (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 and the quality of the molecule sets predicted during the simulation, as can be seen in **Fig. S4**.

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 and uncoupled models, 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 training sets in this scenario this was not a hindrance. RFR also allowed calculation of feature importance, so this was investigated to observe whether any features of the molecular representation were either unnecessary or essential, however none were consistently favoured so the feature set was left unaltered.

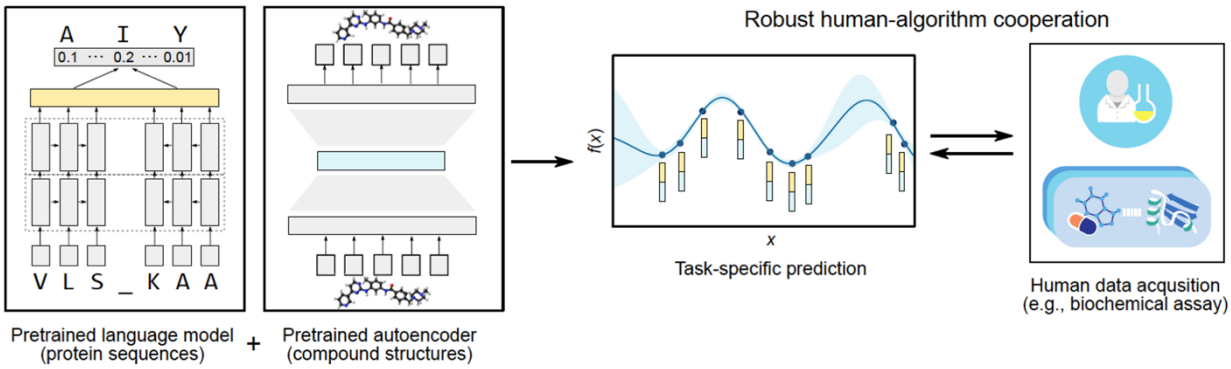


Figure S1. Schematic representation of the machine learning workflow. Starting with the pretrained models that convert the small molecule and protein into a feature input that can be utilised by the prediction model and ending with an iterative loop between the prediction model and the experimentalist. Adapted from reference[22].

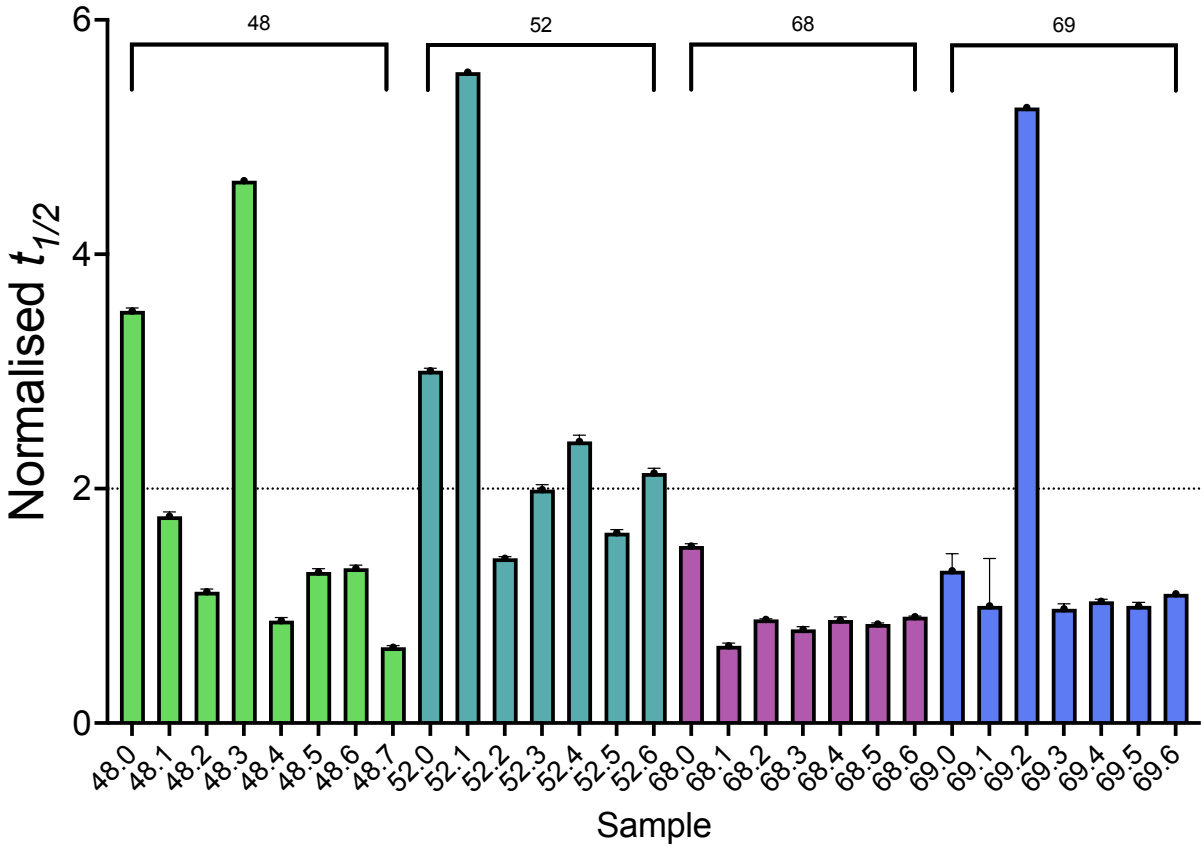


Figure S2. Half time of α-synuclein aggregation in the fibril amplification assay of the close similarity docking set. We show the results for the 25 molecules in the close similarity docking set (2.5 molar equivalents), i.e. those closely related (Tanimoto similarity > 0.5) to the 4 molecules in the docking set (labelled as 48.0, 52.0, 68.0 and 69.0 on the x-axis). Hits were defined as molecules that more than double the half time of α-synuclein aggregation, as indicated by the horizontal line that marks 2 times the half time (normalised $t_{1/2}$, y-axis) in the absence of small molecules. Some derivatives of molecules 48, 52 and 69 showed good potency, in particular 48.3, 52.1 and 69.2.

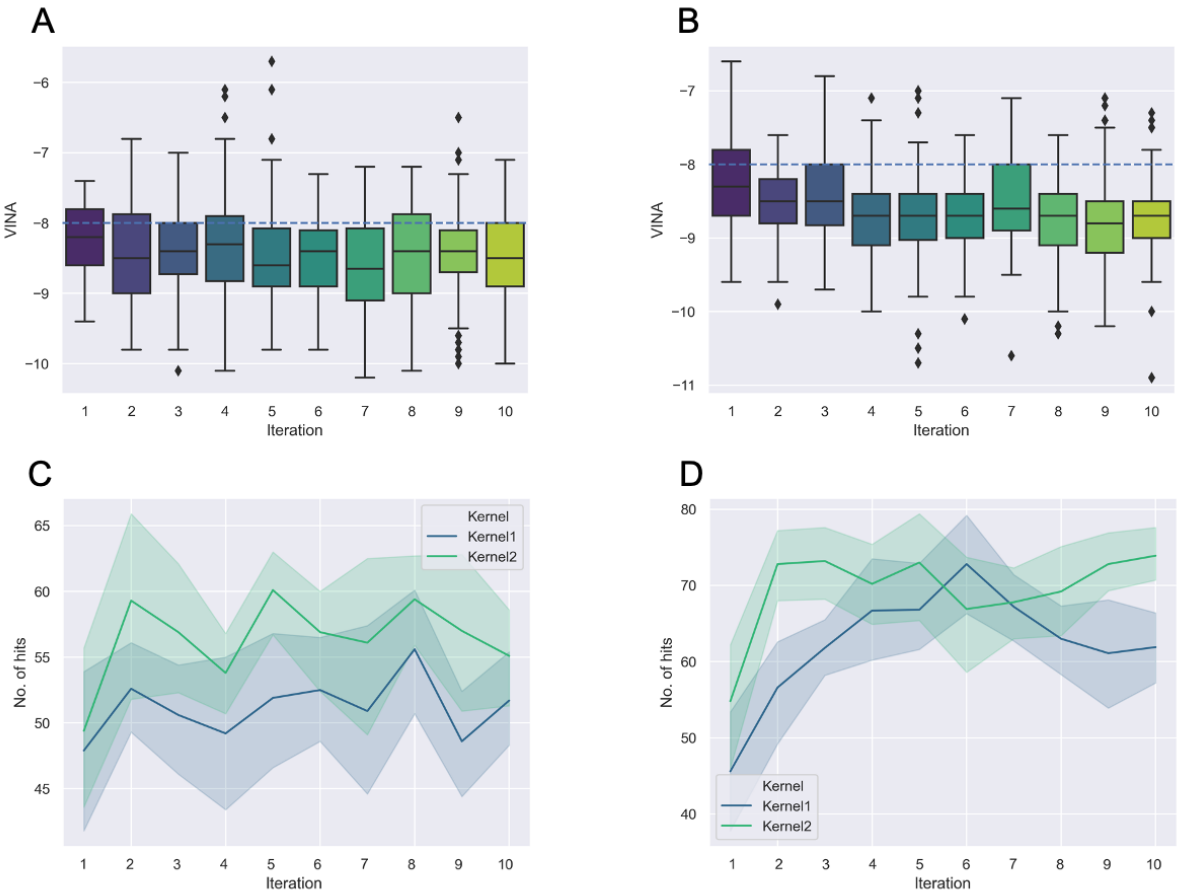


Figure S3. **A simulation of the experimental scenario using docking energies as a proxy for aggregation experiments. (A)** Starting from a single random sample, the RBF kernel was tested as described in the text. AutoDock Vina binding energies in kcal/mol are plotted against iteration number. Each boxplot 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 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.

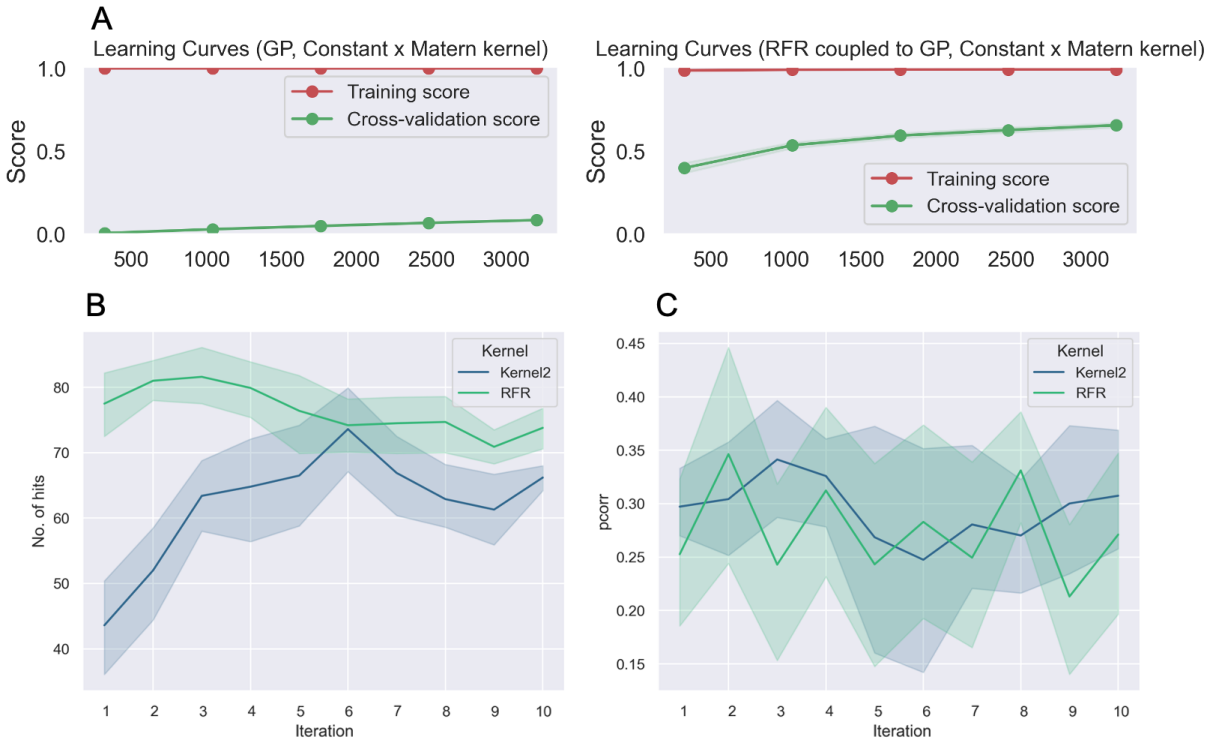


Figure S4. **Performance of the RFR method coupled to the Matérn kernel compared to the Matérn kernel alone. (A)** R^2 score with increasing training set size for both models, using the full-length representation. On the left is the Matérn kernel alone, and on the right is the RFR model. Cross validation 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. **(C)** Average Pearson's correlation coefficient (pcorr) between the predicted binding score values and the real scores at each iteration.

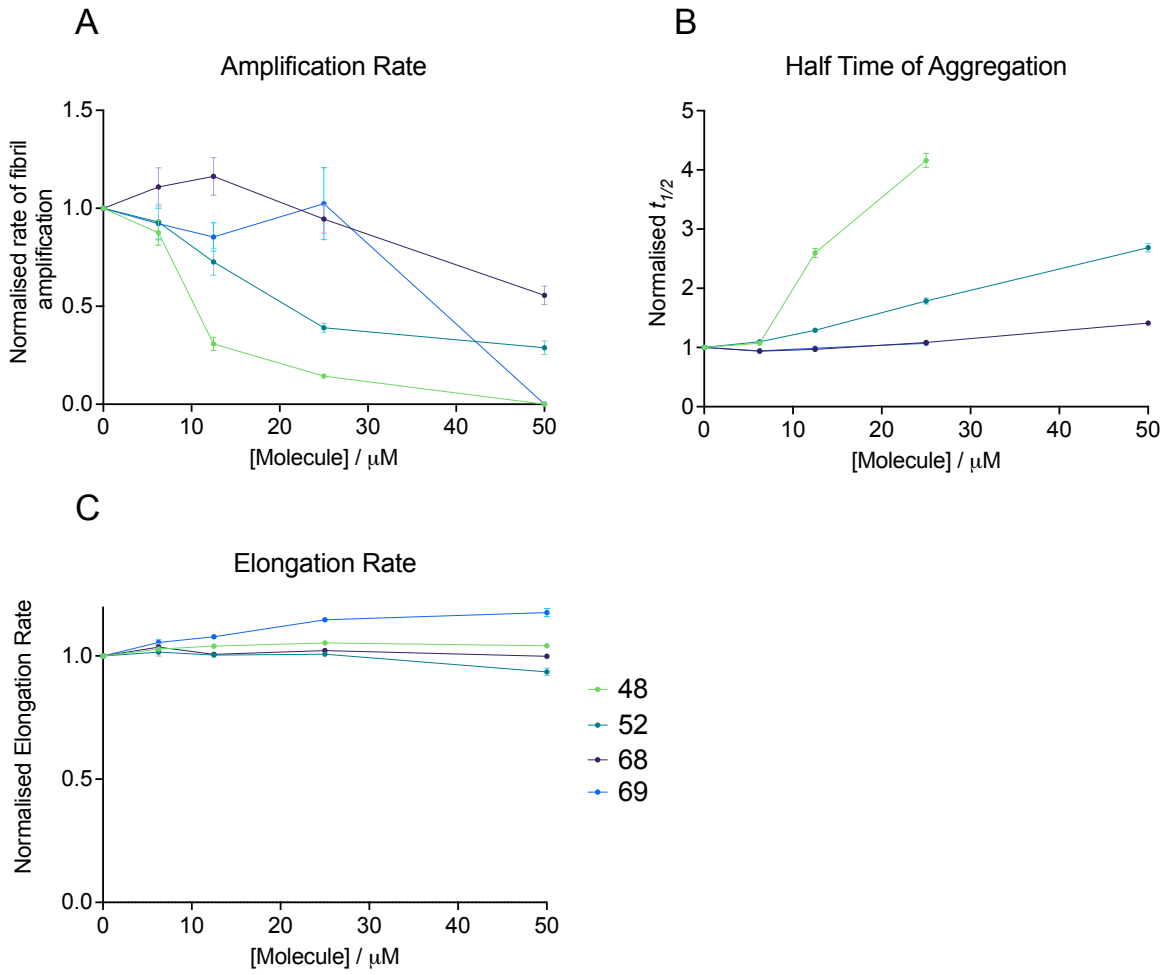


Figure S5. Amplification rate and half time of aggregation of α-synuclein in the presence of the 4 molecules in the docking set. (A) Relative rate of fibril amplification of α-synuclein in the presence of 4 molecules (labeled as 48, 52, 68 and 69) in the docking set; the kinetic traces are normalised to the DMSO control (purple). **(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 21.

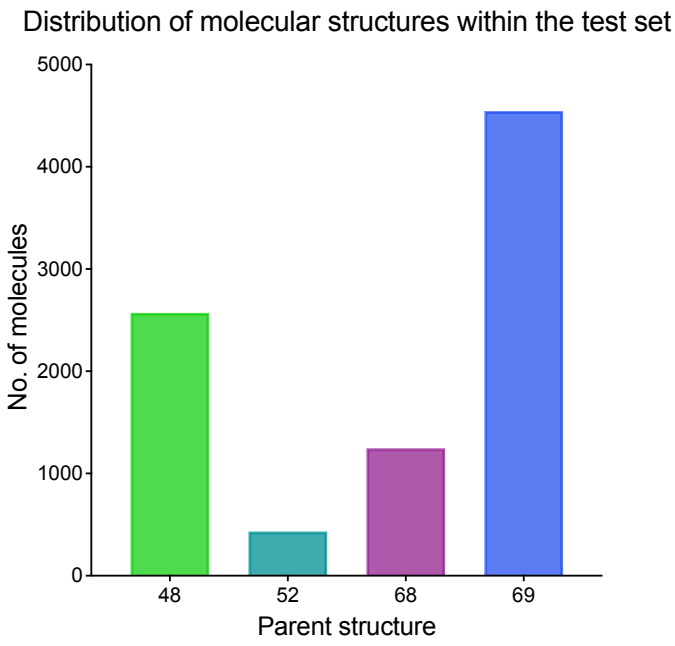


Figure S6. Distribution of molecular structures within the test set. There were more structures derived from molecules 69 and 48 compared with molecules 68 and 52.

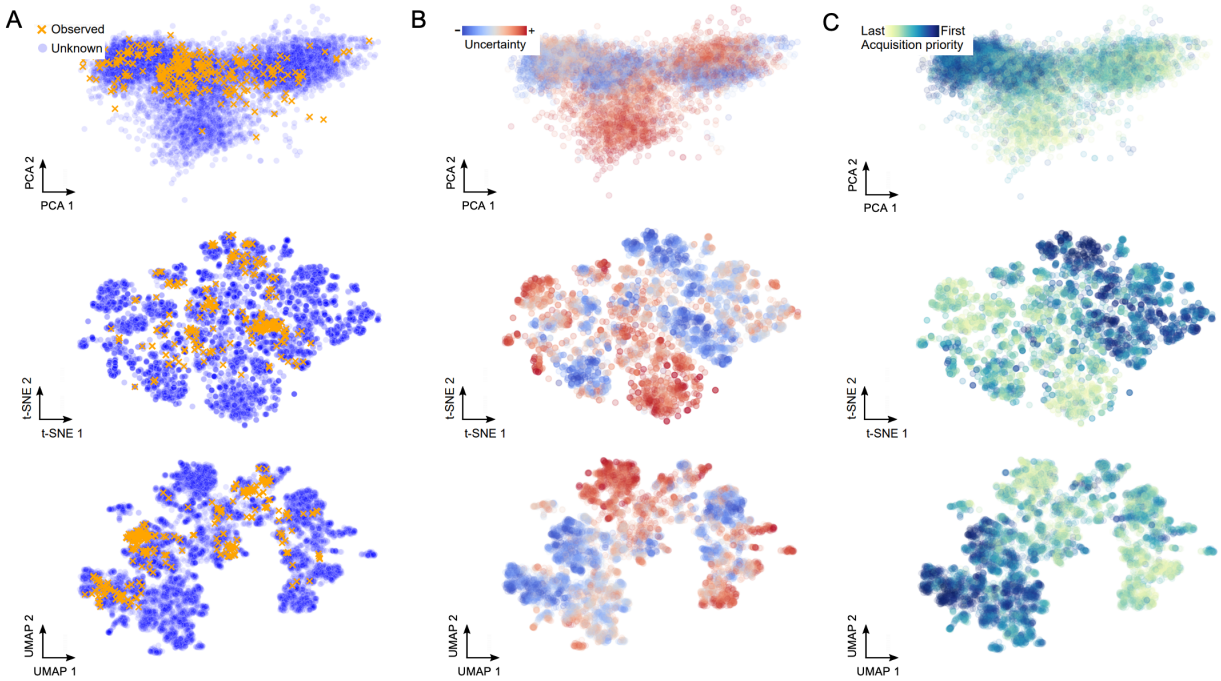


Figure S7. PCA, t-SNE and UMAP visualisations of the compound feature space using uncertainty. (A) From top to bottom: PCA, t-SNE and UMAP visualisations of the compound space indicating which areas of the chemical space have been explored (orange crosses) and which have not (blue circles). **(B)** GPR 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 (red) during acquirement ranking.

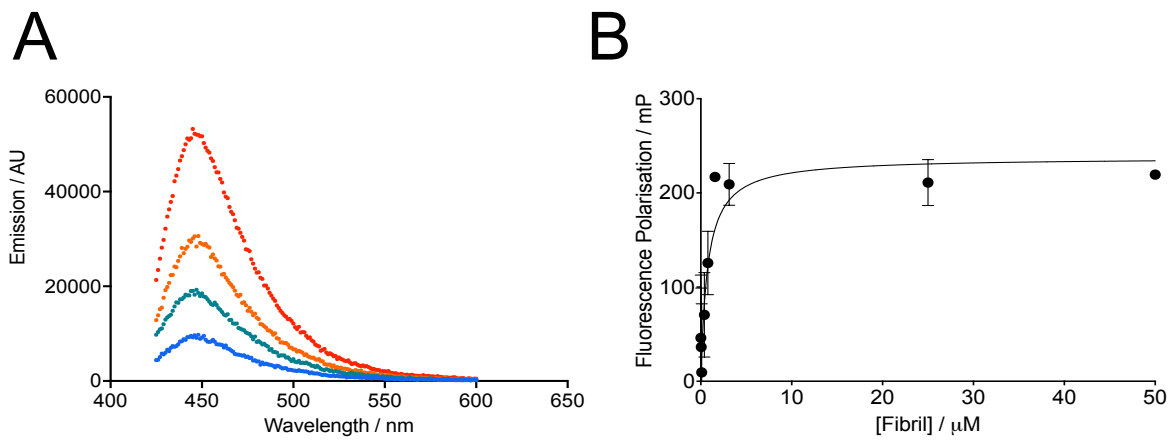
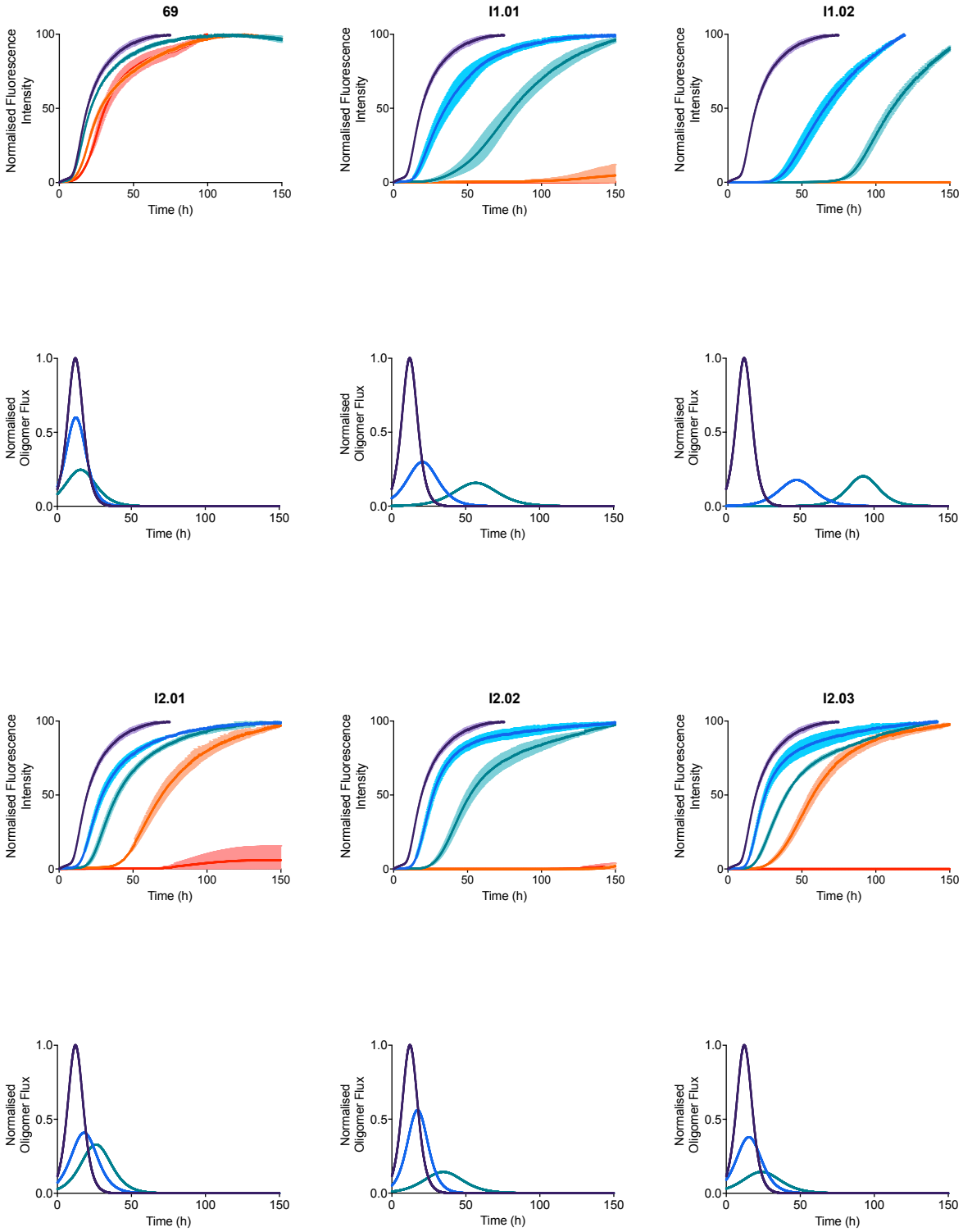
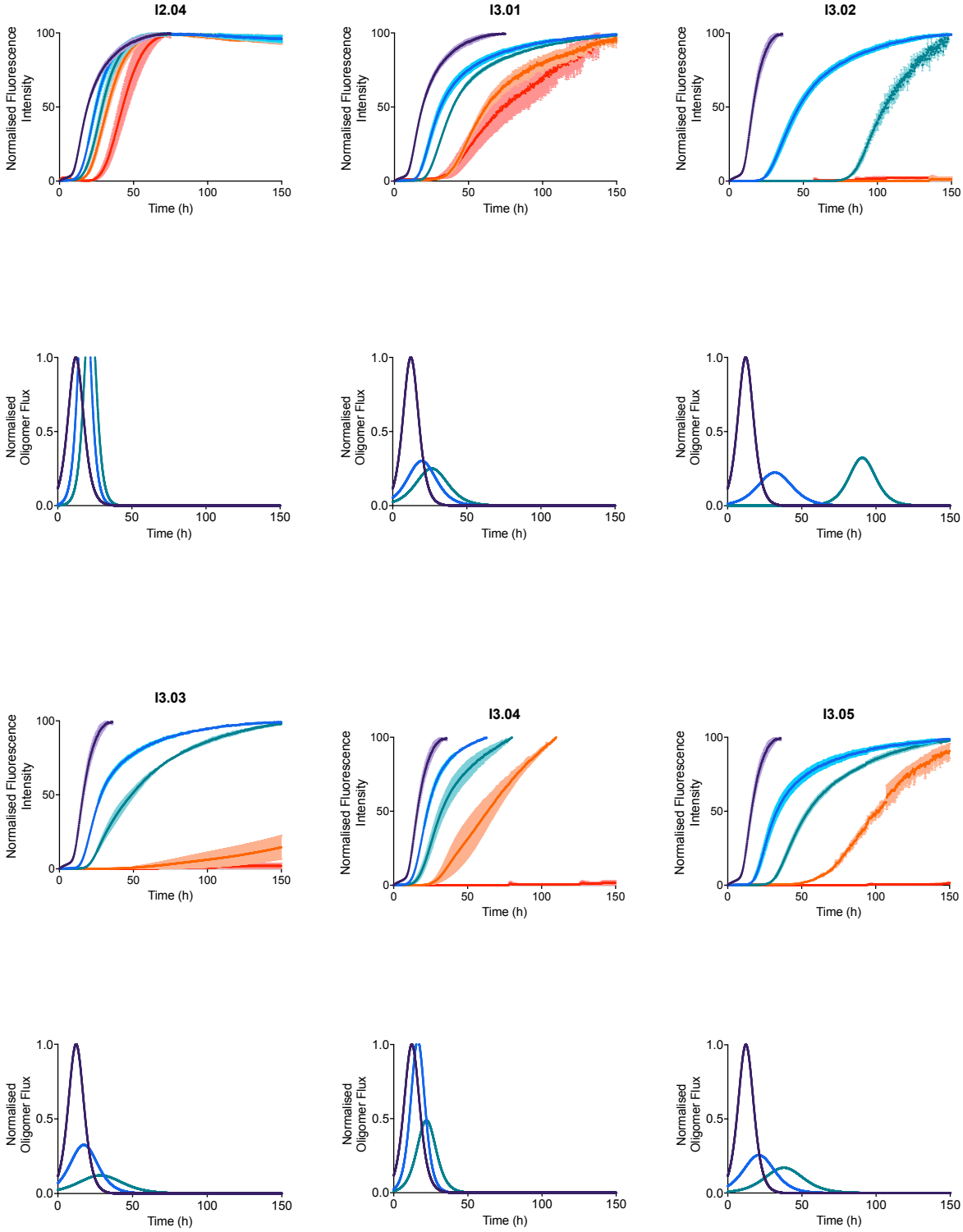


Figure S8. Fluorescence emission spectra and binding curve of molecule I3.08. (A) Emission spectra of molecule I3.08 at 4 different concentrations: 25 μM (red), 12.5 μM (orange), 6.25 μM (teal) and 3.12 μM (blue) μM using an excitation wavelength of 360 nm. **(B)** Change in fluorescence 38olarization (in mP units) of 10 μM I3.08 with increasing concentrations of α-synuclein fibrils (concentrations given in monomer equivalents). The solid line is a fit to the points using a one-step binding curve, estimating a K_D of about 700 ± 300 nM for molecule I3.08 when binding α-synuclein fibrils.


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
I3.03
I3.04
I3.05
Normalised Fluorescence Intensity
Normalised Oligomer Flux
Time (h)
0
50
100
150
0.0
0.5
1.0

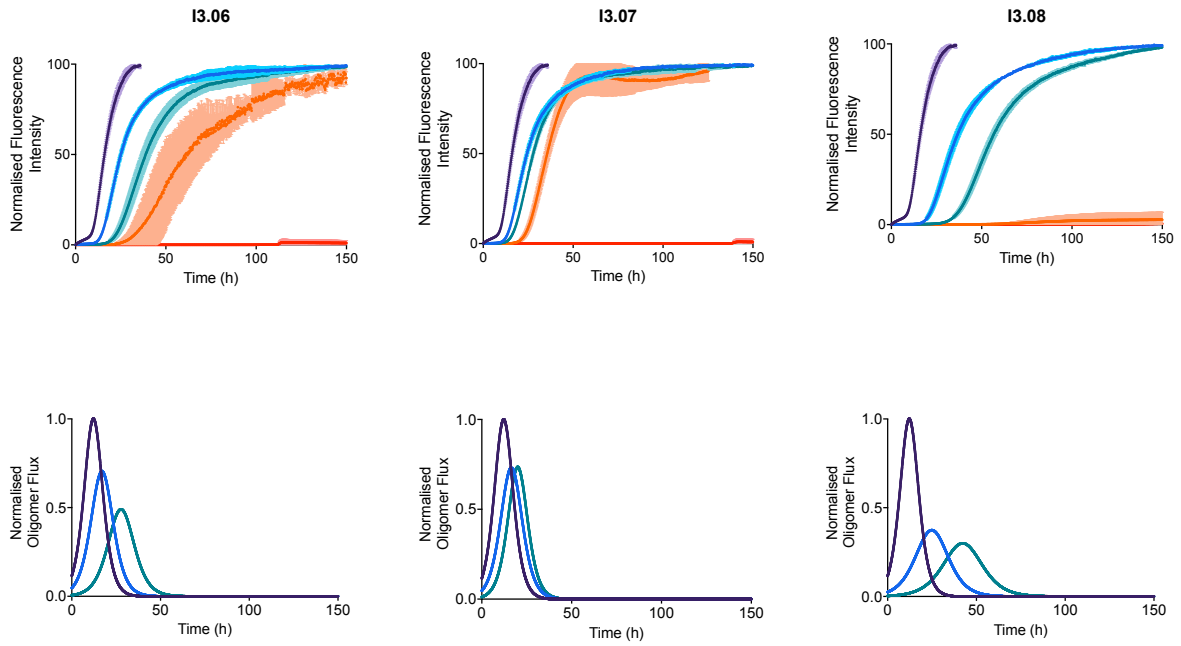
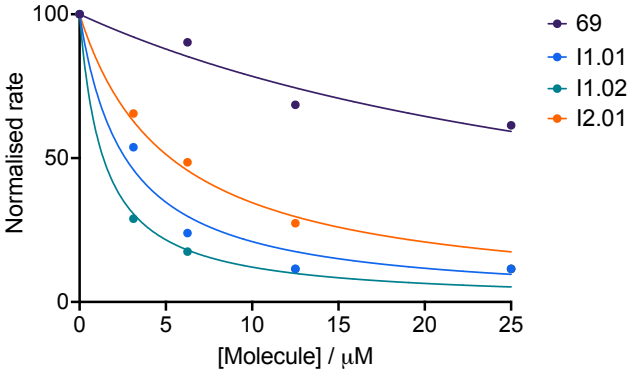
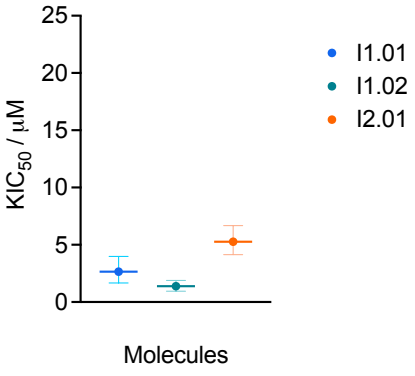
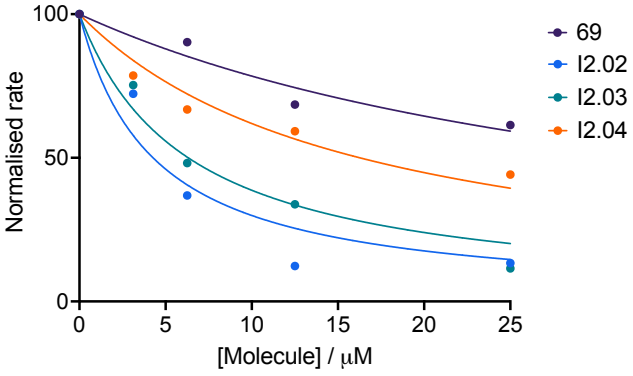
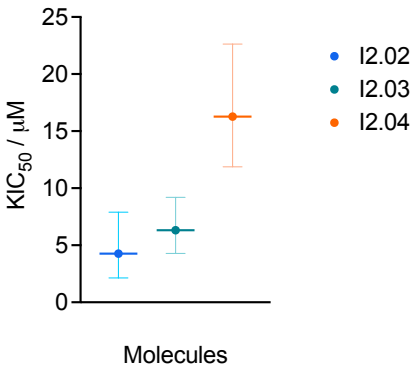


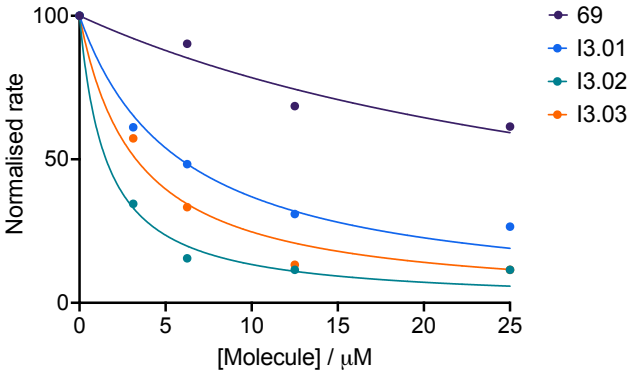
Figure S9. Aggregation curves (top) and oligomer flux simulations (bottom) for the most potent compounds from all of the iterations. The kinetic traces show a 10 μM solution of α-synuclein in the presence of 25 nM seeds at pH 4.8, 37 ºC in the presence of molecules at 2.5 (red), 1.25 (orange), 0.625 (teal) and 0.312 molar equivalents versus 1% DMSO alone (dark purple), with endpoints normalised to the α-synuclein 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.

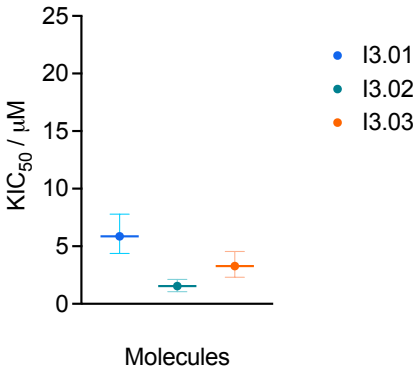

Normalised rate
100
50
0
0
5
10
15
20
25
[Molecule] / μM
69
I1.01
I1.02
I2.01


KIC50 / μM
25
20
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
25
20
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
25
20
15
10
5
0
Molecules
I3.01
I3.02
I3.03

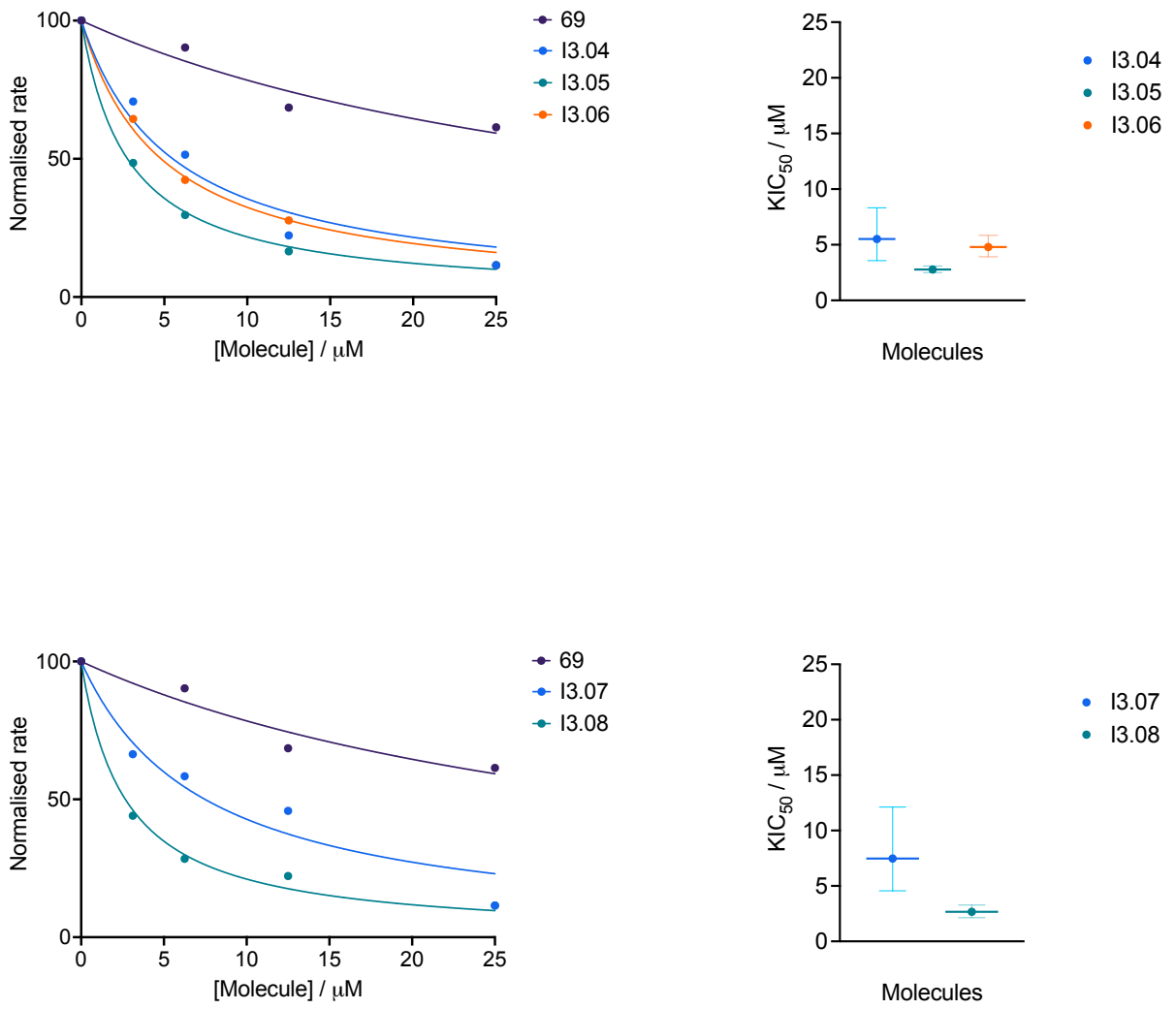


Figure S10. Concentration dependence of the reaction rate and corresponding 50% kinetic inhibitory concentration (KIC_{50}) values for the most potent compounds. The approximate normalised rate of reaction (taken as 1 / $t_{1/2}$) is shown on the left for each molecule at each concentration for which a half time could be obtained. For molecules that inhibited completely the aggregation process on the timescale of the experiment, the highest concentration (2.5 molar equivalents) was taken to be the length of the experiment. The approximate rates are fitted using an [Inhibitor] vs. normalized response Hill slope. The KIC_{50} values are shown on the right with the 95% confidence interval.

Figure S11. 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.

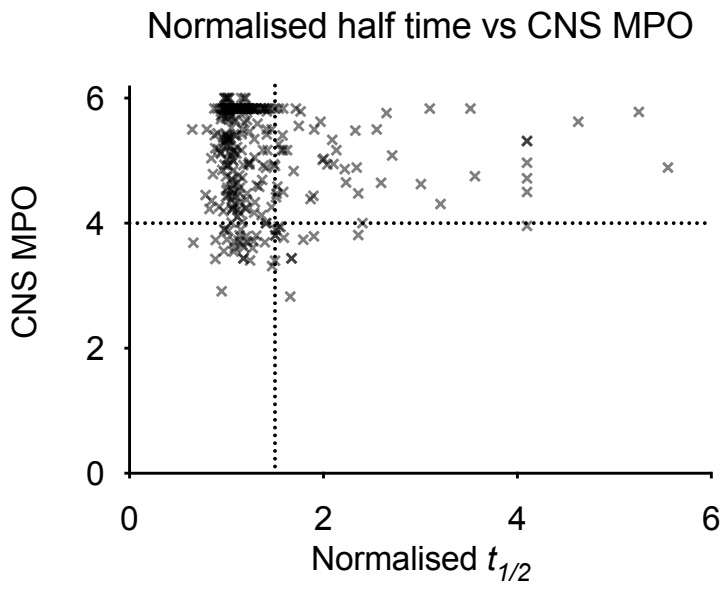


Figure S12. CNS-MPO scores vs normalised half time. The vertical dotted line indicates the half time cut off (1.5) for a molecule to be defined as a hit, while the horizontal dotted line indicates the common cut off of 4 for CNS MPO scores. The majority of the hits are retained as a result.