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

Robert I. Horne[1], Andrea Possenti[1], Sean Chia[1,2], Faidon Brotzakis[1], Roxine Staats[1], Johnny Habchi[1] and 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 machine learning application, due to 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, demonstrating a need for novel and efficient approaches to drug discovery in this area. However, whether or not machine learning methods can fulfil this role remains to be demonstrated. To explore this possibility, 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 (PD) is the most common neurodegenerative movement disorder, estimated to affect 2–3% of the population ≥65 years of age[1–4]. The aggregation of the intrinsically disordered protein α-synuclein is thought to be responsible for some of the neurodegenerative processes underlying PD, in which the pathological accumulation of misfolded proteins 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 PD 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 therefore led to significant 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 increase the effectiveness of traditional drug discovery pipelines[17–19]. An approach based on the combination of docking simulations with in vitro screening was recently employed to identify novel compounds predicted to bind to the fibril structures of α-synuclein that could prevent autocatalytic proliferation of α-synuclein fibrils[20]. Building on that work, we employ here a machine learning approach to effectively explore the chemical space around initial hits from docking simulations and 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[21], 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 **SI Fig. 1**. Component 1 is a junction tree variational autoencoder[22], pre-trained on a set of

250,000 molecules[23]. 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[24,25]. Gaussian processes work effectively on small datasets and this approach enables the leveraging of the associated uncertainty measure when ranking molecules during acquirement priority[21]. Depending on how conservative the experimentalist wishes to be, a greater or lesser weighting can be applied to this uncertainty, allowing ranking based on both the predicted potency of the molecules and the uncertainty of that prediction. The last component is a chemical kinetics assay[10,26,27] which identifies the top compounds that can significantly inhibit the surface-catalysed secondary nucleation step in the aggregation of α-synuclein.

Chemical kinetics approaches have advanced to the point that specific mechanisms of α-synuclein aggregation can be targeted in a reproducible way[10,26,27]. 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[20,28,29]. Specific rate measures and other aggregation metrics were derived from these experiments allowing quantitative and reliable comparisons between molecules in terms of structure-activity relationship (SAR) and offering metrics to optimise structures of interest[10,30]. For this work, the half times of aggregation were used as the metric of efficacy to be used in machine learning as these gave a more robust measure of efficacy than extracted rate values which are more prone to influence by small fluctuations in the slope of the aggregation fluorescence trace[31] (**SI Fig. 2**). Any molecules that achieved a half time 1.5 times greater than that of the negative control under standard assay conditions (see Methods) were classed as hits[20].

A drawback, however, remains the resource intensive nature of the assays in terms of protein consumption and the time taken, as each aggregation assay takes place over a timescale of days. As such, it is of particular interest in this research area to screen molecules in the most effective way possible. Docking simulations are often able to provide initial hits at a lower experimental cost than high throughput screening (HTS)[32]. The docking simulation approach used here[20] was able to circumvent the conventional approach of high throughput screening of large compound libraries, which typically offer

hit rates in the range of 0.1-1%. In that work, 4 molecules of interest were identified from a selection of 68, 2 of which would have been considered hits as defined above, representing a hit rate of 3%.

This hit rate, though an improvement, remains modest, so other methods could be introduced at this point to improve the hit rate further. The reason for this low hit rate is that, in the case discussed here, the predicted binding scores do not correlate well with aggregation inhibition metrics as demonstrated previously[20] and so the docking approach is not suited to compound optimization where differences between compound efficacies are finely separated. To complete this task, conventional similarity searches were carried out alongside machine learning approaches to obtain new hits derived from the molecules initially identified via docking simulations.

The overall approach is shown schematically in **Fig. 1** and described here. Initial docking simulations were carried out using a combination of AutoDock Vina[33] and FRED[34] software, and subsequent experiments identified 4 compounds which showed activity, labelled 48, 52, 68 and 69 according to the cluster number label in the docking simulations (**Fig 1A**). Using the Tanimoto similarity metric[33], two similarity searches were carried out using these four structures as starting points. Firstly, a selection of very closely related molecules (Tanimoto similarity > 0.5) to the parent compounds was experimentally tested (**Fig. 1B**) followed by a larger selection of compounds with a looser cut-off of structural similarity (Tanimoto similarity > 0.45) to the parent compounds (**Fig. 1C**). The final step was applying machine learning (**Fig. 1D**) in an iterative exploration of the chemical space. The similarity searches removed the most obvious targets of the machine learning approach, but also increased the size of the dataset available for training, although the training set remained very small by typical machine learning standards (of the order of hundreds of molecules). Since, however, training set sizes such as this are typical in early-stage research, a further aim of this work is to demonstrate that machine learning can be used rather effectively even in such data sparse scenarios.

The primary aim of the machine learning approach presented here was to obtain greater hit rate than either HTS, the docking simulations or the similarity searches, as well as greater hit novelty than the similarity searches could supply. 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 the secondary nucleation of α-synuclein more effectively, which could be of use in therapeutic programs targeting synucleinopathies in the future.

Results

Methods of exploring the compound space

An initial close similarity search was carried out (Tanimoto similarity > 0.5), and the molecules were tested for aggregation inhibition yielding 8 new hits from 25 new molecules (**SI Fig. 3)**. Another 69 molecules were then brought in with Tanimoto similarity > 0.45 (**Fig. 1C**) to the hit structures in order to expand the training set beyond the initial docking compounds and their closely related derivatives. Although several 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 this loose set of 29 molecules related to compound 48, only 2 were hits, while of 24 molecules related to compound 52 only 5 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 this looser similarity search was less than half of that of the close search and involved testing three times as many compounds.

At this point it would have been challenging to further explore chemical space using conventional SAR techniques without significant attrition. The hit rate worsened as the similarity constraint to the hits was loosened. This trend would only be expected to continue as wider chemical space was explored, and as the similarity threshold dropped the number of molecules that could be screened increased from the 10s to the 1000s, which would have necessitated a larger scale screening.

As a result, more effective methods of hit location were needed to identify hit molecules with increased chemical variability. The machine learning approach outlined in the next sections was applied, yielding a hit rate that was a 78% improvement over the combined similarity search hit rates (**Figs. 2 and 3**), even after the easily identifiable targets in the

proximal chemical space had been removed from the test set. As a result of the removal of all the closely related chemical matter from the test set, the molecules predicted by this approach differed to a significant extent from the parental structures (Tanimoto similarity < 0.4), and so would represent new avenues for exploration. The following sections outline the machine learning approach used and the results obtained from it.

Narrowing the search space

One of the issues with applying machine learning to a data sparse scenario is that predictions are very likely to be overconfident. While this problem can be addressed to an extent by utilising Gaussian processes, another strategy is to restrict the search area or test set to a region of chemical space that is more likely to yield successful results. To this end, a structural similarity search of the 4 primary hits from the initial docking screen 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.

Application of the drug discovery approach

Different machine learning models were trialled against the docking scores calculated for the test set and the best performing model (**SI Figs. 4 and 5**) was trained on the whole aggregation data set and used to predict the top set of molecules; 64 of these molecules were then purchased for testing. 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 time were chosen for purchase. The result was a favouring of structures that bore a structural resemblance to compound 69. There was an inherent bias towards the structure of compound 69 in the dataset given the relative population sizes (**SI Fig. 6**), 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.

Inspections of the chemical space via dimensionality reduction techniques such as PCA, t-SNE and UMAP were also carried out to investigate how the model was prioritising molecules. These methods are shown in **SI Fig. 7,** and the UMAP is shown in **Figs. 4-5**. **Fig. 4** shows the relative positioning of the training points and the parents within the chemical space. **Fig. 5** demonstrates how the stacked RFR-GPR model assigned low uncertainty to areas of the chemical space proximal to the observed data, and how the corresponding acquirement priority mirrored this when trained on the aggregation data. **Fig. 5** 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 aggregation data from the first 3 prediction cycles are shown in **Fig. 2** and appear promising. Of the 64 molecules from iteration 1, 16 were strong hits, representing a hit rate of 25%, the second iteration showed a further increase, with 24 strong hits representing a 38% hit rate and the third iteration showed a hit rate of 35%, though with stronger hits than any of the previous steps. These hit rates represent an order of magnitude improvement over HTS (~1%) and, remarkably, an overall 78% 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 loose similarity search (**Fig. 3**).

The changes in similarity of the hits to the parent structure and the structures of 2 of most potent hits from iteration 1 are shown in **Fig. 6**. Similarity dropped for all structures at successive stages of the investigation, reaching its lowest point at the initial iterations of the machine learning approach. Of the most potent hits, most retained the left-hand side scaffold of compound 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, Hit A replaced the right-hand side of compound 69 with a single substituted benzene ring, while Hit B 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 efficacy, which the model did effectively while also identifying alterations in the rest of the scaffold that enhanced the efficacy considerably beyond that of the parent.

Discussion and Conclusions

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 30-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 (14%) 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.

The primary limitations of this approach are the requirement to pick from a pre-existing library and focussing only on the assay metric of interest as a learning parameter. To address the former a generative model could be employed[36]. To address the latter is an aim of machine learning studies into drug discovery in general, as multi-parameter optimisation is challenging[37–39]. The topic of most interest in machine learning drug discovery approaches besides potency prediction is 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] (**SI Fig. 8**).

Overall, the work carried out provides confidence that the combination of two different in silico strategies for hit identification, docking simulations and machine learning, could be superior to HTS alone. This approach also gave a significant improvement over the hit rate obtained by combining the docking hits with a conventional similarity search, even after those hits with close similarity to the parent molecules had been removed from the test set. The hits identified by this combined approach were both a significant improvement over the parent molecule in terms of efficacy and represented a much more significant departure from the parent than any of the other potent hits identified beforehand. As a result, it is likely that using Bayesian and ensemble machine learning approaches could be of significant benefit to researchers working in this field and indeed early-stage research in general.

Materials and Methods

Preparation of compounds and chemicals

The centroids from the above 79 clusters are selected for experimental validation. Compounds were purchased from MolPort (Riga, Latvia) and in the case where centroids were not available for purchase, the molecule in the cluster with the closest chemical structure was used as the representative compound instead. A total of 68 compounds were purchased, and then prepared in DMSO to a stock of 5 mM. All chemicals used were purchased at the highest purity available. The same process was repeated for subsequent steps, with Mcule as the supplier for iterations 2 and 3.

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 °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.

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).

Machine learning

Junction tree neural network variational autoencoder

The autoencoder[22] was pretrained on a library of 250,000 compounds (as used by Kusner et al., (2017)[23], and was implemented as described in reference 22 using a pip installable version (source: https://github.com/LiamWilbraham/jtnnencoder).

Protein representation model

The 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 GPR, RFR and MLPs 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], MulticoreTSNE[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.

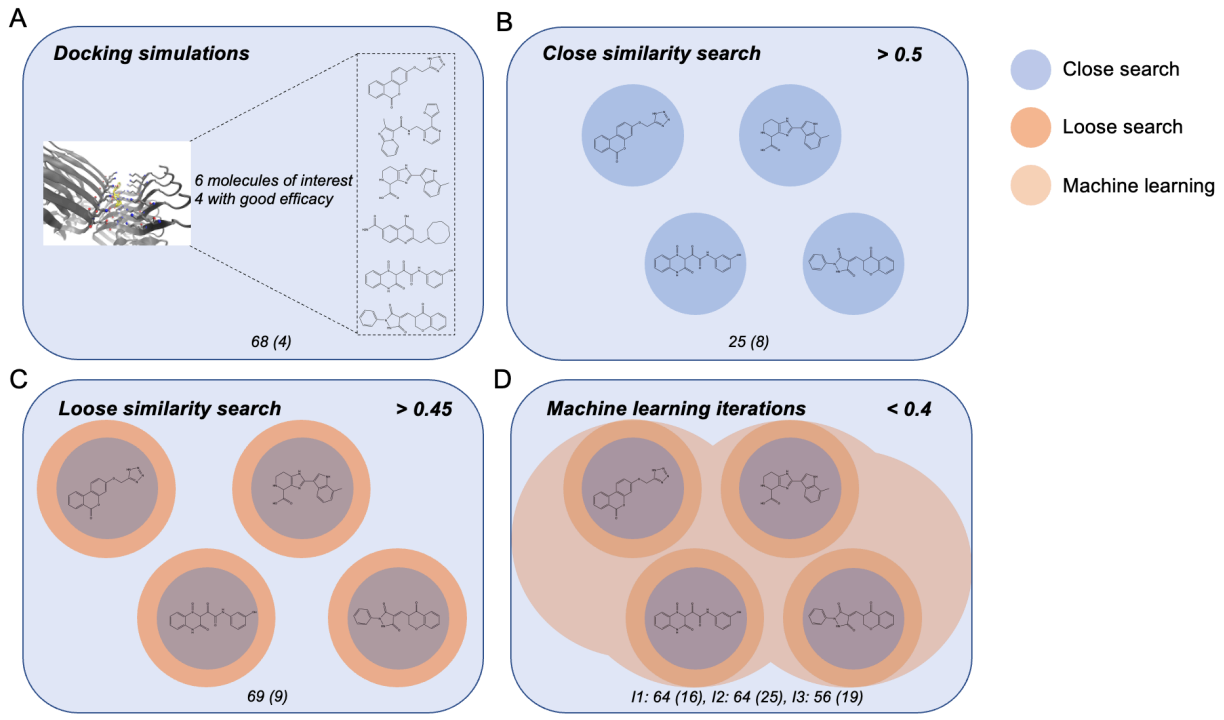

Figure 1: Stages of the compound space exploration. For (B)-(D) the number of molecules tested is shown at the bottom of each pane with the number of hits shown in brackets. **(A)** The initial docking simulations identified 4 molecules of interest from 68 tested (3 would be classified as hits with normalised half time > 1.5 at a molar ratio of 2.5 relative to α-synuclein monomer). **(B)** The initial close similarity search around the 4 parent compounds in chemical space selected molecules with Tanimoto similarity cut off > 0.5, as indicated in the top right of the pane. **(C)** The loose similarity search with Tanimoto similarity cut off > 0.45. **(D)** Machine learning was carried out on the resultant dataset of 162 molecules, picking hits from a set of molecules derived from the ZINC database with Tanimoto similarity > 0.3 to the parent structures. This yielded 16 new hits with an average Tanimoto similarity < 0.4 to the parent structures on the first iteration, 25 new hits on the second iteration and 19 new hits on the third iteration. These hits had higher potency on average than those identified in the previous similarity searches.

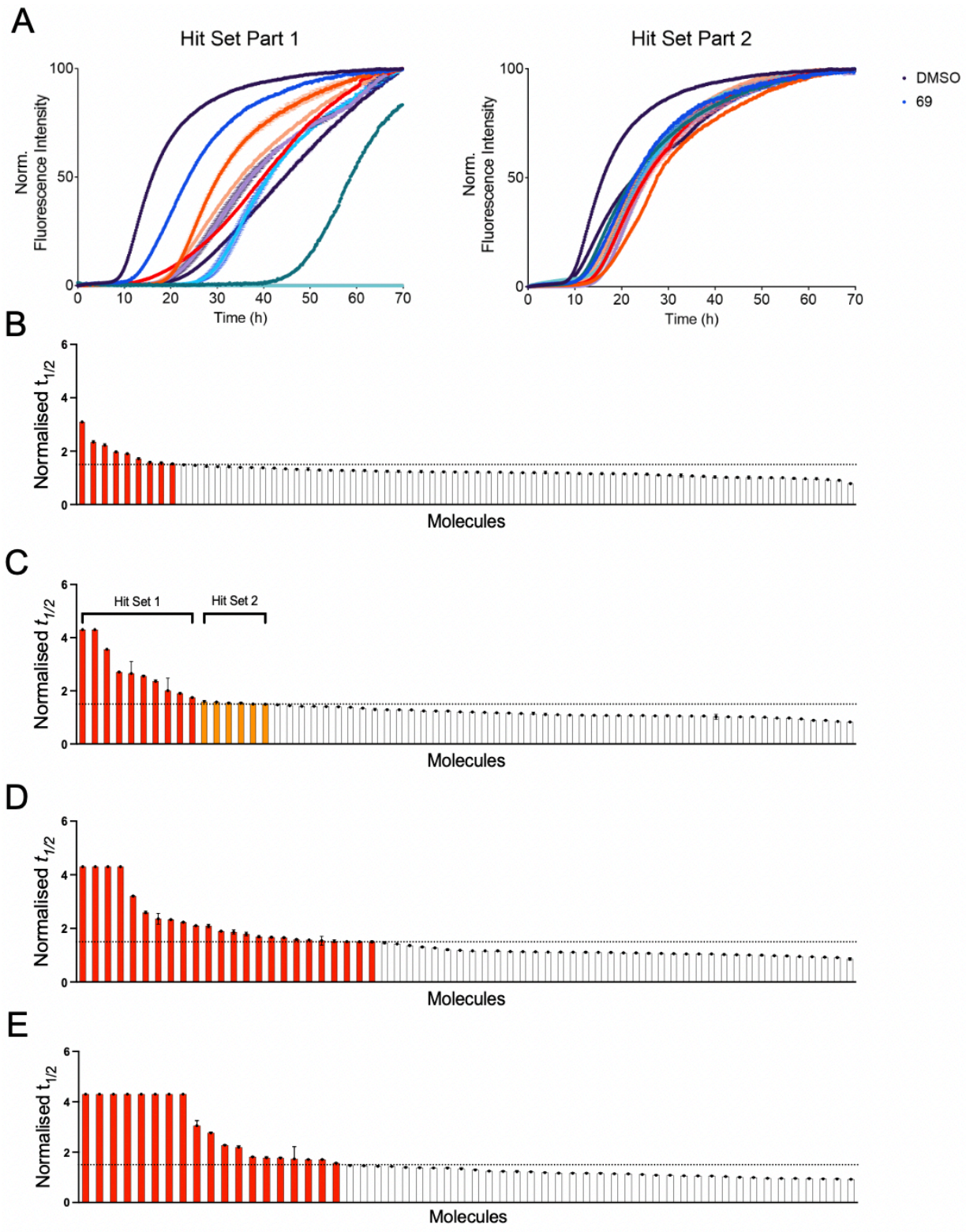

Figure 2: Application of the drug discovery approach. (A) Kinetic traces of a 10 μM solution of α-synuclein in the presence of 25 nM seeds at pH 4.8, 37 °C in the presence of molecules from iteration 1 (2.5 molar equivalents, different colours) versus 1% DMSO alone (dark purple) and the parent molecule 69 (blue), with endpoints normalised to the α-synuclein monomer concentration detected via the Pierce™ BCA Protein Assay at *t*

= 70 h. **(B)** Normalised half times of aggregation for the loose similarity search. The horizontal dotted line indicates the boundary for hit classification; $t_{1/2}$ = 1.5. **(C)** Normalised half times of aggregation for the molecule set predicted in iteration 1. Hit set 1 in (A) is highlighted in red, while hit set 2 in (A) is highlighted in orange. **(D)** Normalised half times of aggregation for the molecule set predicted in iteration 2. **(E)** Normalised half times of aggregation for the molecule set predicted in iteration 3. Note the most potent molecules exhibited complete inhibition of aggregation, so the $t_{1/2}$ has been presented as the length of the experiment.

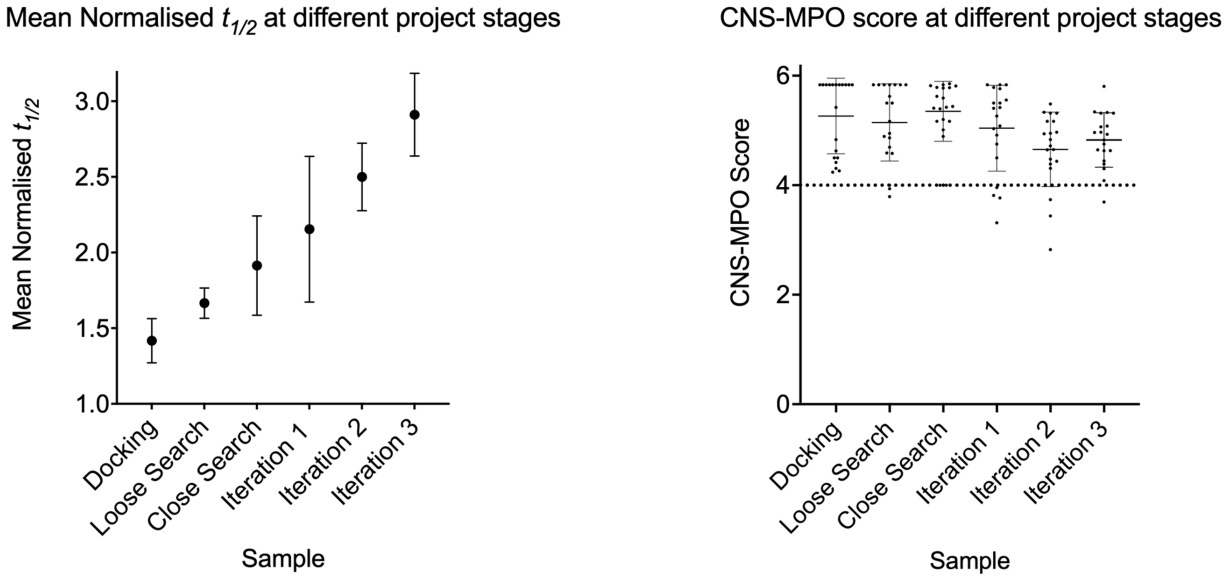

Figure 3: Average half time of aggregation and CNS-MPO scores for the top 20 molecules at each stage. The stages are the initial docking simulation (68 molecules tested), loose search (69 molecules tested), close search (25 molecules tested), iteration 1 (64 molecules tested), iteration 2 (64 molecules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

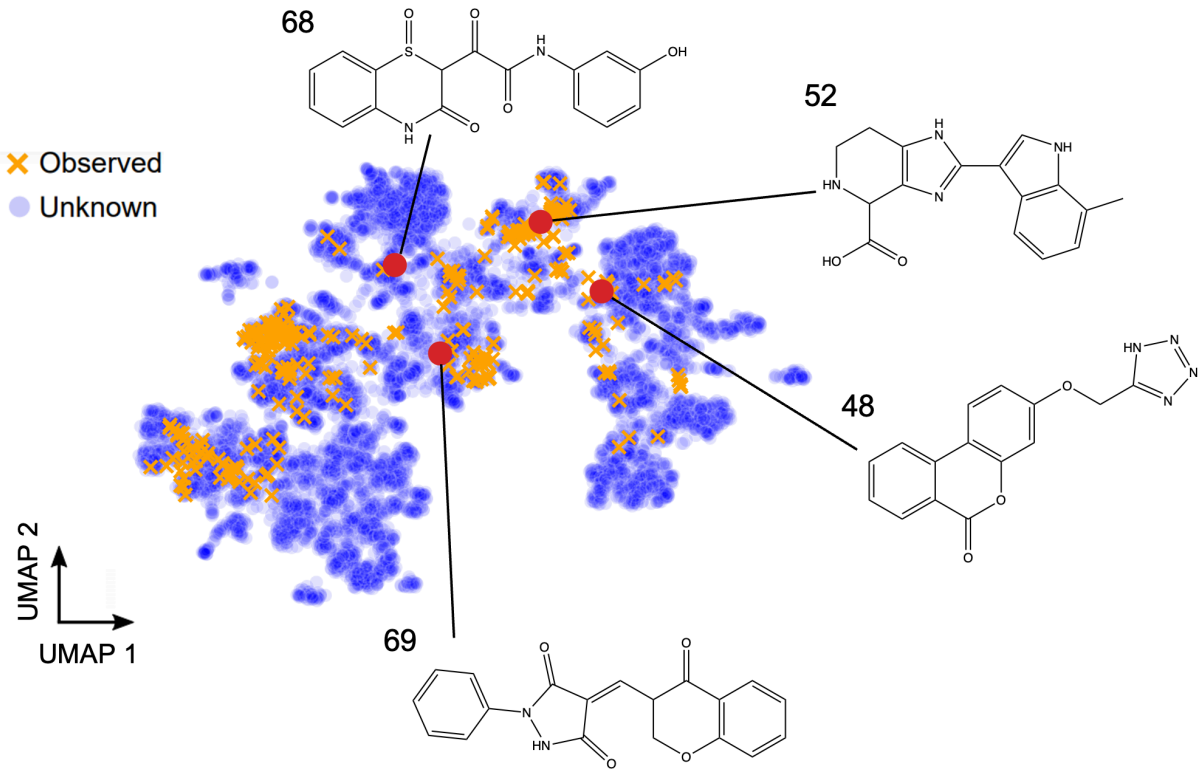

Figure 4: Visualisations of the compound feature space. UMAP visualisations of the compound space indicating which areas of the chemical space have been explored (orange crosses) and the relative positioning of the parent structures in this space.

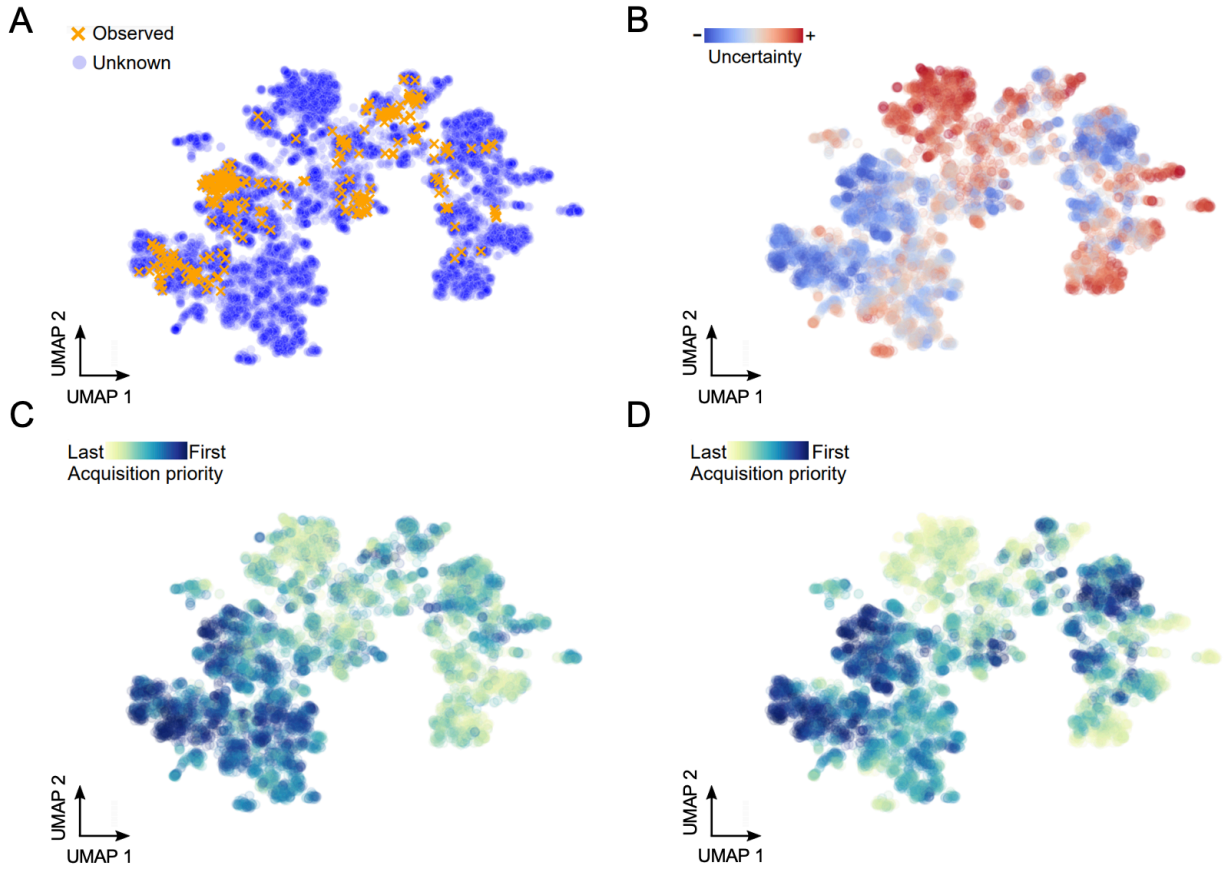

Figure 5: UMAP visualisations of the compound feature space using uncertainty. (A) Indicates which areas of the compound space have been explored (orange crosses) and which have not (blue circles). (B) GPR assigned lower uncertainty (blue) to regions of the chemical space near to the observed data and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

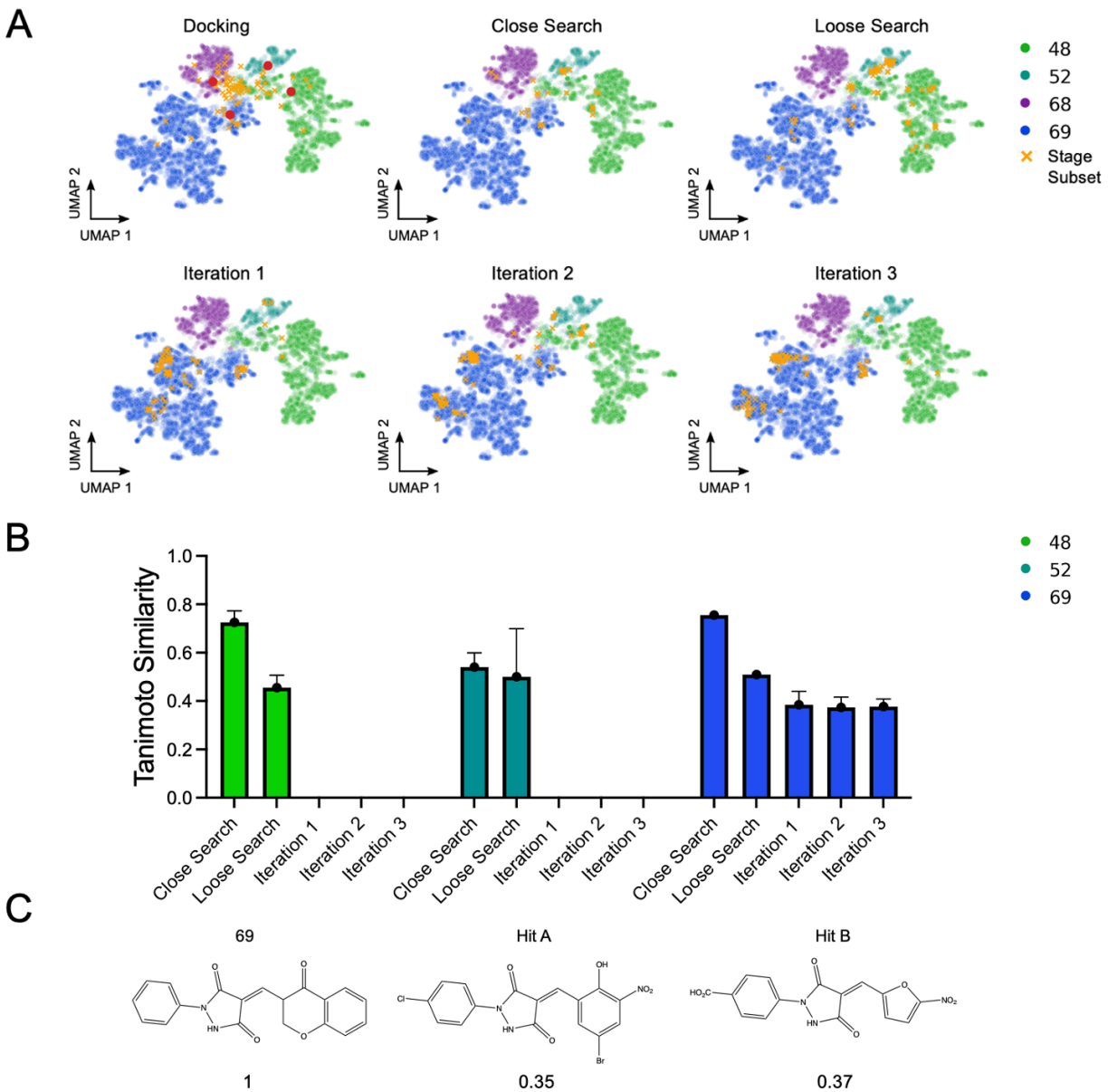

Figure 6: Structural changes during the optimisation. (A) UMAP visualisations of the compound space indicating how the positioning of each new molecule subset (orange crosses) changed at each stage of the project as well as how the chemical landscape was split between the parent molecules (different 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 48 (green), 52 (teal) and 69 (blue) at each stage of the project. At iteration 1, 2 and 3 all of the hits were derived from structure 69, albeit with lower similarity than any of the previous stages. 68 has not been included as it failed to produce any hits outside of the parent molecule. **(C)** Structures of the parent molecule (molecule 69) and of the 2 strongest hits from iteration 1, which completely prevent aggregation until t = 70 h. Tanimoto similarities to molecule 69 are given beneath each structure.

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 a-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* **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. 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).
21. Hie, B., Bryson, B. D. & Berger, B. Leveraging Uncertainty in Machine Learning Accelerates Biological Discovery and Design. *Cell Systems* **11**, 461-476 (2020).
22. Jin, W., Barzilay, R. & Jaakkola, T. Chapter 11: Junction Tree Variational Autoencoder for Molecular Graph Generation. *arXiv: 1802.04364v4* (2019).
23. Kusner, M. J., Paige, B. & Hemández-Lobato, J. M. Grammar variational autoencoder. *34th International Conference on Machine Learning, ICML 2017* **4**, 3072–3084 (2017).
24. Rasmussen, C. E. & Williams, C. K. I. *Gaussian Processes for Machine Learning*. **7**, (2006).
25. Breiman, L. Random forests. *Machine Learning* **45**, 5–32 (2001).
26. Knowles, T. P. J. *et al.* An Analytical Solution to the Kinetics of Breakable Filament Assembly. *Science* **326**, 1533–1537 (2009).

27. 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).
28. 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).
29. 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).
30. Chia, S. *et al.* SAR by kinetics for drug discovery in protein misfolding diseases. *Proc. Natl. Acad. Sci. USA* **115**, 10245–10250 (2018).
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. 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**, 146–157 (2011).
33. 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).
34. Le Guilloux, V., Schmidtke, P. & Tuffery, P. Fpocket: An open source platform for ligand pocket detection. *BMC Bioinformatics* **10**, 1–11 (2009).
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,* (2018).

SUPPLEMENTARY INFORMATION

Machine Learning Implementation

Before embarking on experimental work, we tested the machine learning strategy on simulated data. The docking scores already calculated for the training set were obtained 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 have been included below as the FRED results followed very similar trends. The compound encoder was implemented as in Hie et al. [21] to obtain representations of all the compounds, 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 that was 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, matching the approach used in Hie et al.[21] 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.[21] , 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. It was 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 **SI Figs. 4** and **5**.

SI Fig. 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 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 looked 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.[21] in fitting a Gaussian process regressor (GPR) to the residuals of another model, in that case an 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 **SI Fig. 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 correlation coefficient oscillated around 0.25 - 0.3 for both the coupled and uncoupled models which while modest did match 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.

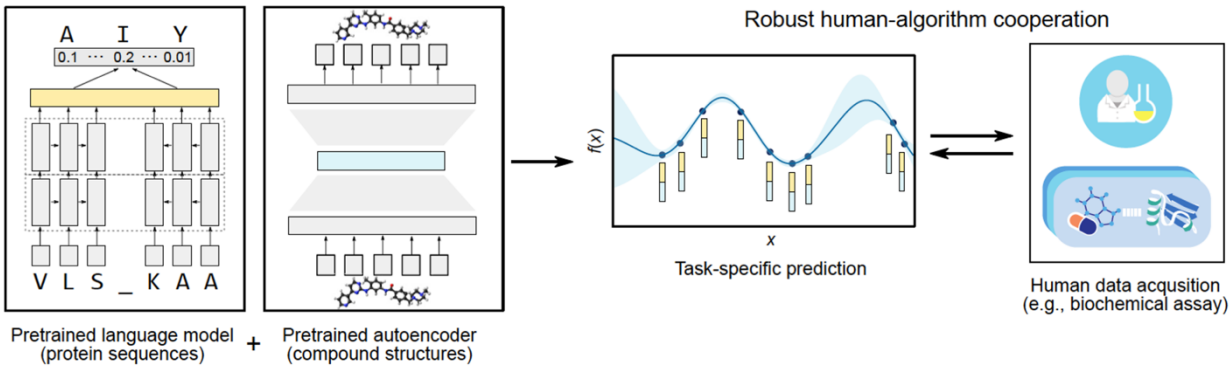

Figure S1: A schematic representation of the workflow. Starting with the pretrained models that convert the molecule and protein into a feature input that can be utilised by the prediction model and ending with an iterative loop between the prediction model and the experimentalist. Adapted from reference[21].

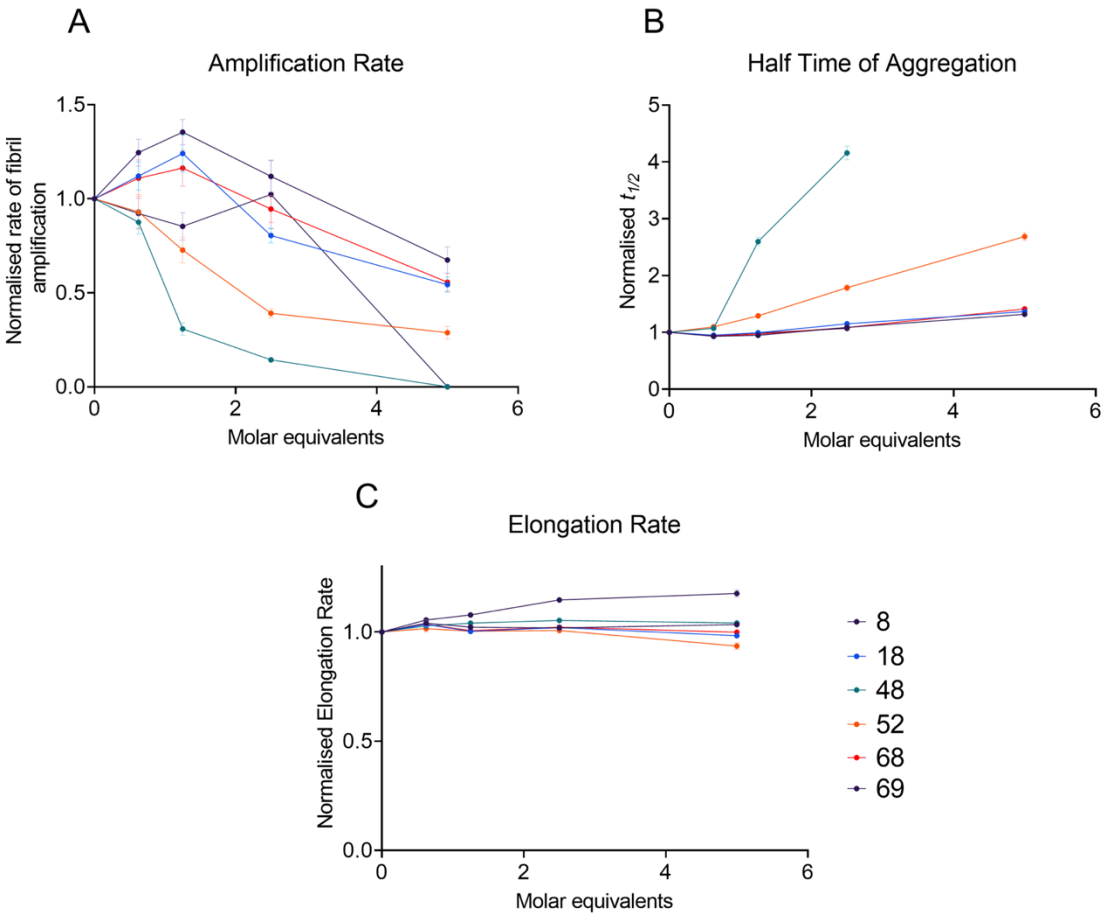

Figure S2: Extracted kinetic rates and half times of aggregation. (A) Relative rate of fibril amplification of α-synuclein in the presence of molecules 8-69 (different colours) normalised to the DMSO control (purple). **(B)** The half times of aggregation derived from the same experiment (**C**) The relative rate of fibril elongation of α-synuclein in the presence of molecules 8-69 normalised to the DMSO control. The amplification rate (A) and half time of aggregation (B) derived from the secondary nucleation experiment were 2 candidates for the measure of molecule efficacy to be used in machine learning. As can be seen from (A) the amplification rate is more erratic in its behavior as any perturbation to the early slope of the exponential phase can have large effects on the derived rate value. The half time, though a simpler measure, is more robust and so was chosen for machine learning. Data obtained from reference 20.

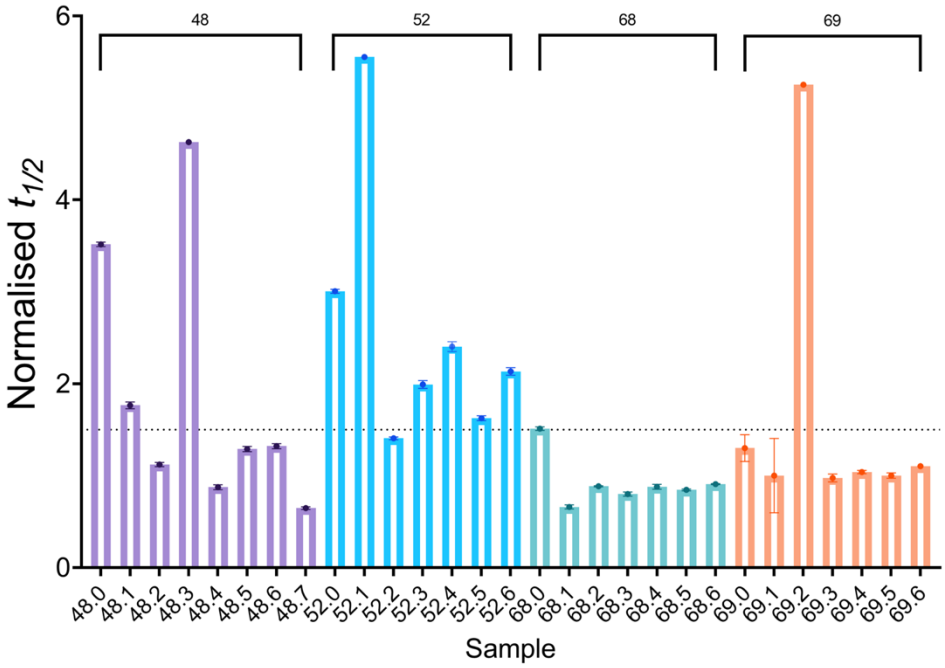

Figure S3: Half times of aggregation in the fibril amplification assay. We show the results for molecules (2.5 molar equivalents) very closely related to the 4 hits of interest (Tanimoto similarity > 0.5). Hits were defined as molecules that increase the half time by 50% or more as indicated by the horizontal line. Molecules 48 and 52 (labelled as 48.0 and 52.0 here) showed good inhibition among their respective derivatives while molecule 69 (69.0) exhibited one strong hit.

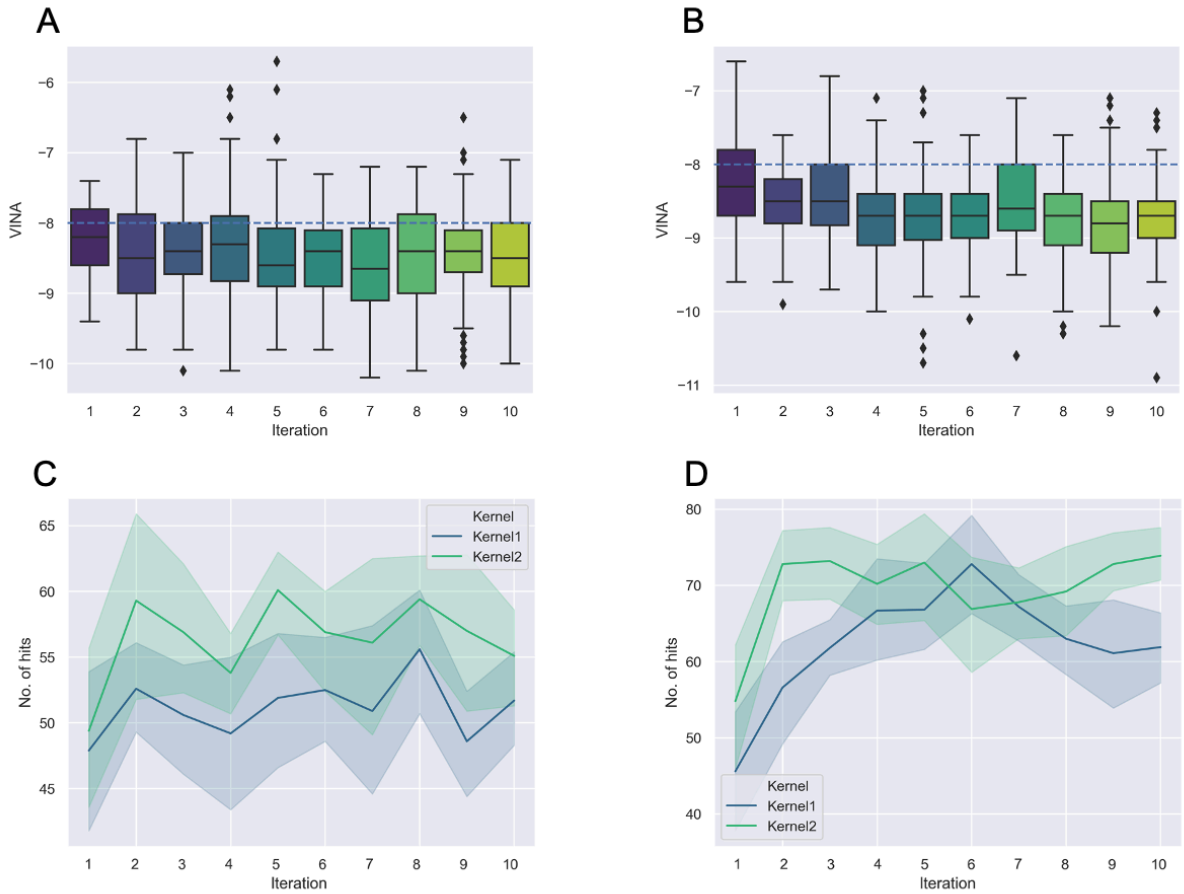

Figure S4: **A simulation of the experimental scenario using docking energies as a proxy for aggregation data. (A)** Starting from a single random sample the RBF kernel was tested as described in the text. AutoDock Vina binding energies in kcal mol^{-1} are plotted against iteration number. Each boxplot 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)** The 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^{-1}). Results were obtained using the half-length representation of the molecules. **(D)** The same process as described in panel C, but employing the full-length molecule representation.

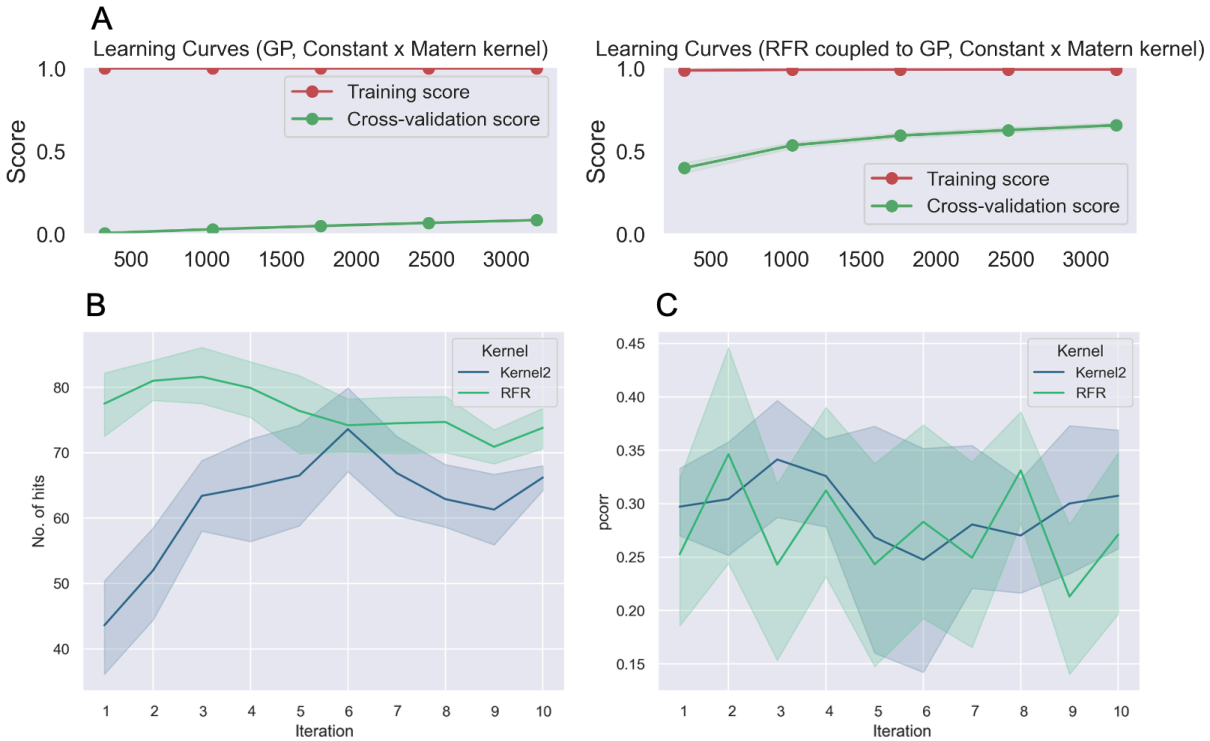

Figure S5: **Performance of RFR coupled to the Matérn kernel compared to the Matérn kernel alone. (A)** R^2 score with increasing training set size for both models, using the full-length representation. On the left is the Matérn kernel alone, and on the right is the RFR model. Cross validation with 10 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.

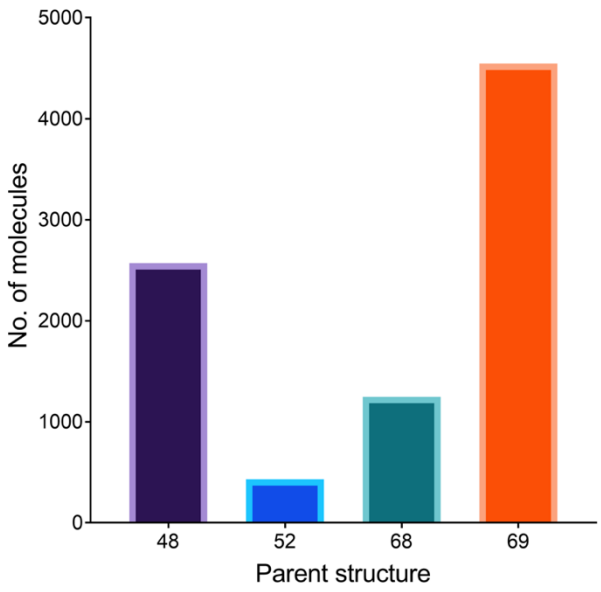

Figure S6: Distribution of molecular structures within the test set. There were significantly more structures derived from molecules 69 and 48 available compared with molecule 68 and especially molecule 52.

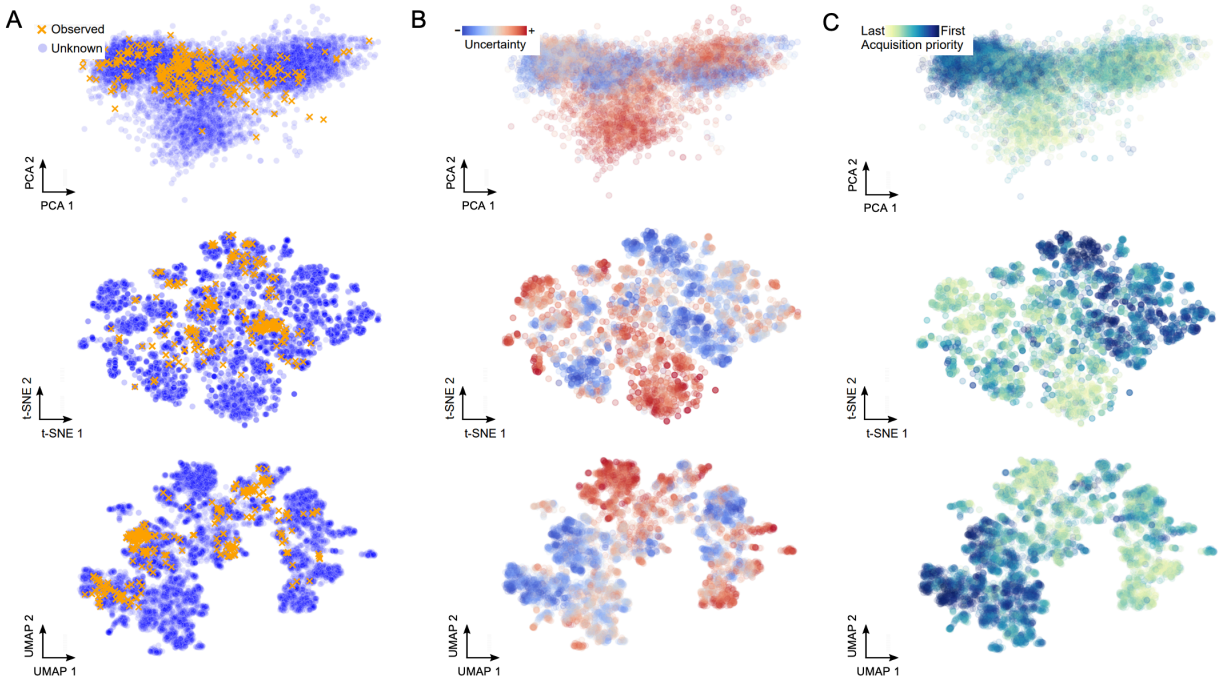

Figure S7: Visualisations of the compound feature space using uncertainty. (A) From top to bottom; PCA, t-SNE and UMAP visualisations of the compound space indicating which areas of the chemical space have been explored (orange crosses) and which have not (blue circles). **(B)** GPR assigned lower uncertainty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

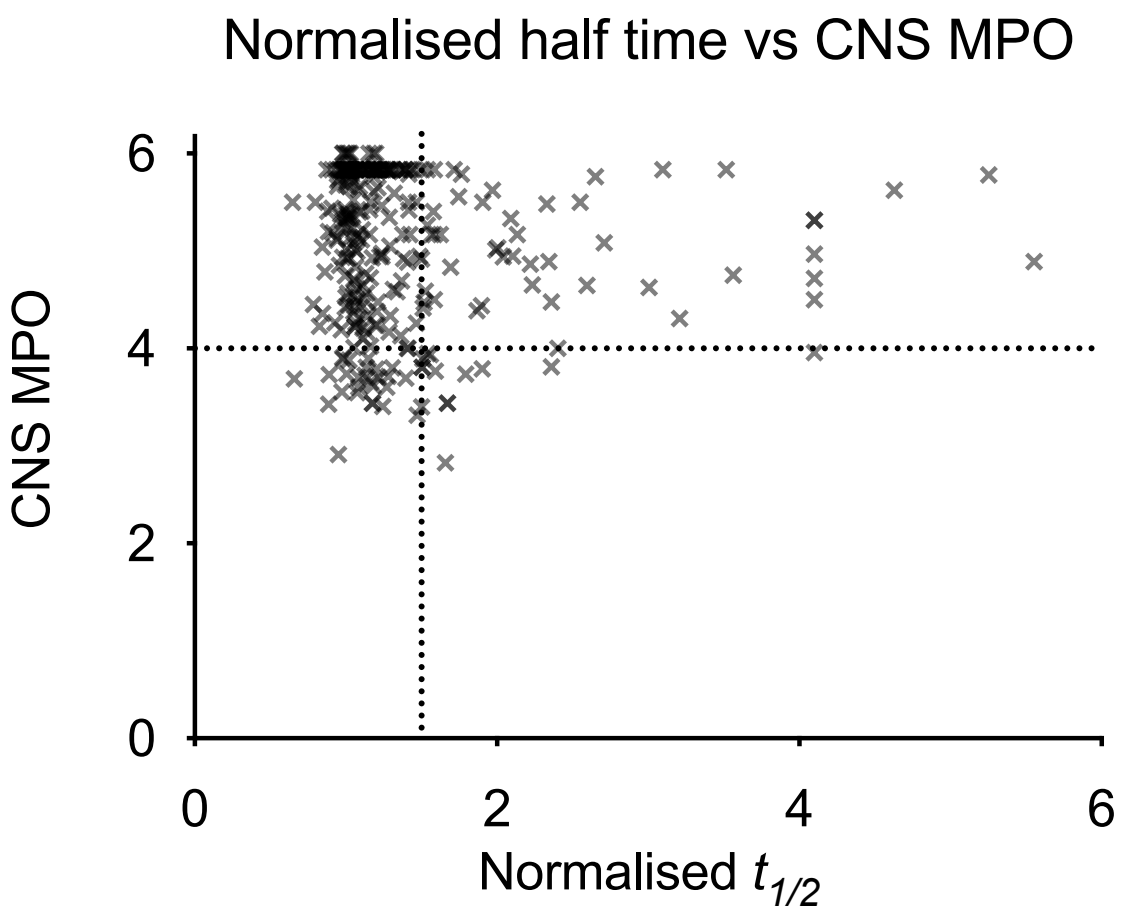

Figure S8: CNS-MPO scores vs normalised half time. The vertical dotted line indicates the half time cut off (1.5) for a molecule to be defined as a hit, while the horizontal dotted line indicates the common cut off of 4 for CNS MPO scores. The majority of the hits are retained as a result.