Epitope scaffolding using α-synuclein cyclic peptides to generate oligomer-selective antibodies for Parkinson's disease

Shawn C.C. Hsueh,[†] Adekunle Aina,[†] Neil R. Cashman,[‡] Xubiao Peng,[*,†,§] and Steven S. Plotkin[*,†,¶]

†*Department of Physics and Astronomy, The University of British Columbia, Vancouver, BC, V6T 1Z1, Canada*

‡*Djavad Mowafaghian Centre for Brain Health, The University of British Columbia, Vancouver, BC, V6T 2B5, Canada*

¶*Genome Science and Technology Program, The University of British Columbia, Vancouver, BC, V6T 1Z1, Canada*

§*Current address: Center for Quantum Technology Research, School of Physics, Beijing Institute of Technology, Beijing, China*

E-mail: xubiaopeng@bit.edu.cn; steve@phas.ubc.ca

Abstract

Effectively scaffolding epitopes on immunogens, in order to raise conformationally selective antibodies through active immunization, is a central problem in treating protein misfolding diseases, particularly neurodegenerative diseases such as Alzheimer's disease or Parkinson's disease. We seek to selectively target conformations enriched in toxic, oligomeric propagating species while sparing the healthy forms of the protein that are often more abundant. To this end, we scaffolded epitopes in cyclic peptides by

varying the number of flanking glycines, to best mimic a misfolding-specific conformation of an epitope of α-synuclein enriched in the oligomer ensemble, as characterized by a region most readily disordered and solvent-exposed in a stressed, partially denatured protofibril. We screen and rank the cyclic peptide scaffolds of α-synuclein *in silico* based on their ensemble overlap properties with the fibril, oligomer-model, and isolated monomer ensembles. We introduce a method for screening against structured off-pathway targets in the human proteome, by selecting scaffolds with minimal conformational similarity between their epitope and the same solvent-exposed primary sequence in structured human proteins. Different cyclic peptide scaffolds with variable numbers of glycines can have markedly different conformational ensembles. Ensemble comparison and overlap was quantified by the Jensen-Shannon Divergence, and a new measure introduced here—the embedding depth, which determines the extent to which a given ensemble is subsumed by another ensemble, and which may be a more useful measure in sculpting the conformational-selectivity of an antibody.

Keywords

cyclic peptides; epitope scaffolding; molecular dynamics; ensemble similarity; protein misfolding; virtual screening.

1 Introduction

A key step in the development of a therapeutic antibody or active vaccine is the immunization strategy (*1*), namely, the choice of protein epitope and how it will be presented to an animal or human immune system. Both primary sequence and conformation of the epitope determine the particular protein morphologies to which the resulting antibodies will be selective.

Nowhere has conformational-selectivity been more important to immunotherapies than in protein-misfolding diseases (*2*). For this class of diseases, an effective antibody must be able

to spare healthy protein and discriminate misfolded protein species that lead to molecular and cellular pathology (*1*). Since the primary sequences of healthy and aberrant protein are generally the same, barring splice variants and perhaps post-translational modifications, the efficacy of an antibody is then due to its selective preference for binding to a misfolded conformation over healthy in-vivo "native" conformations.

For many proteins involved in misfolding disease however, the native conformational ensemble is intrinsically disordered (*3*). Examples include amyloid-β (Aβ), tau protein, α-synuclein, and the low-complexity domains in FUS and TDP43. Raising an antibody that avoids the majority of diverse conformations of an intrinsically-disordered protein's ensemble is a challenge. For example, a peptide consisting of a contiguous fragment of native primary sequence tethered to an immunogen such as keyhole limpet hemocyanin (KLH) will likely exhibit overlap in its presented ensemble with the ensemble of isolated native monomer.

Many neurodegenerative diseases, including Alzheimer's disease (AD), Parkinson's disease (PD), chronic traumatic encephalopathy (CTE) and amyotropic lateral sclerosis (ALS), spread throughout the brain via a prion-like mechanism involving soluble oligomers (*1*, *4–10*) Soluble oligomers contain roughly 4-40 chains of protein, which exist in a misfolded conformational ensemble that is conformationally labile and very difficult to experimentally characterize. We computationally model some key aspects of the oligomer ensemble using molecular dynamics, in order to predict disease-specific epitopes.

Cyclic peptides or macrocycles are polymers that have been conjugated to form a ring-like topology (Fig. 1). They have been increasingly used as therapeutics, often as small molecule drugs to bind targets (*11*, *12*). They have also been used as immunogens to raise oligomer-selective antibodies in Alzheimer's disease using the method that we describe here (*13*, *14*).

In this paper, we do not address the non-trivial problem of misfolding-specific epitope prediction, which we have treated elsewhere (*15*) (see Methods Section 2.1). We focus instead on the problem of how to properly present a predicted epitope to the immune system of an animal (or human, in the case of active immunization), so that the resulting antibodies

generated by the animal are selective to disease-specific forms of the protein.

We start from computationally generated ensembles of the α-synuclein fibril, isolated monomer, and stressed, partially disrupted proto-fibril as defined below, which is used as a model for the toxic oligomer − a species we wish to target with conformationally selective antibodies. We briefly review the epitope prediction method described in (*15*), which is applied here to an α-synuclein protofibril. Given an epitope and an oligomer model, we then construct various cyclic peptide immunogen scaffolds of that epitope by varying the number of glycines flanking the epitope on the N- and C- termini. We determine the optimal scaffolding of the cyclic peptide construct by maximizing the overlap with the stressed fibril ensemble and minimizing the ensemble overlap with either the monomer or the unstressed "native" fibril. We also minimize the tendency of antibodies to a scaffolded epitope to elicit off-pathway targeting, by minimizing the conformational similarity to unrelated structured proteins in the human proteome that contain the epitope's primary sequence. Using the above features as screening criteria, we used a ranking method we have developed previously (*16*) involving multi-criteria decision making analysis to systematically rank the scaffolds from best to worst. Such a ranking allows for the experimentalist to construct a reduced number of cyclic peptide scaffolds that are most likely to have the desired features of eliciting antibodies with conformationally selective binding to toxic oligomer species.

The pipeline of the *in silico* screening method is given in Fig. 2. Each step is described in the Methods. In the Results, we visually depict the projection of simulated ensembles in reduced conformational space, and we formalize the ensemble overlap. We then develop the above-described method to find off-pathway targets for a given scaffold, and rank 48 candidate cyclic peptide scaffolds. We finally discuss the non-trivial question of weight assignments for ranking candidates, alternate scaffolding methods, the benefits of *in silico* screening to facilitate more efficient *in vitro* and *in vivo* screening, and the sensitivity or robustness of epitope prediction depending on the structural model of the fibril used in the epitope prediction algorithm.

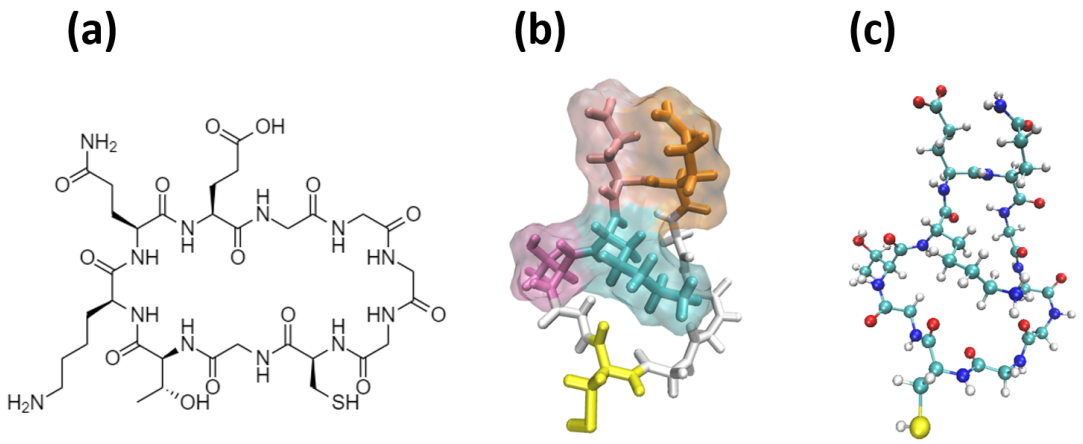


Figure 1: Cyclic peptide renderings for cyclo(CGTKEQGGGG), a scaffold of TKEQ. (**a**) 2D representation of the cyclic peptide. (**b**) 3-dimensional rendering of the cyclic peptide in licorice, also showing the surface of the TKEQ epitope (**c**) Ball and Stick (CPK) rendering with color assigned by the atom identity.

2 Methods

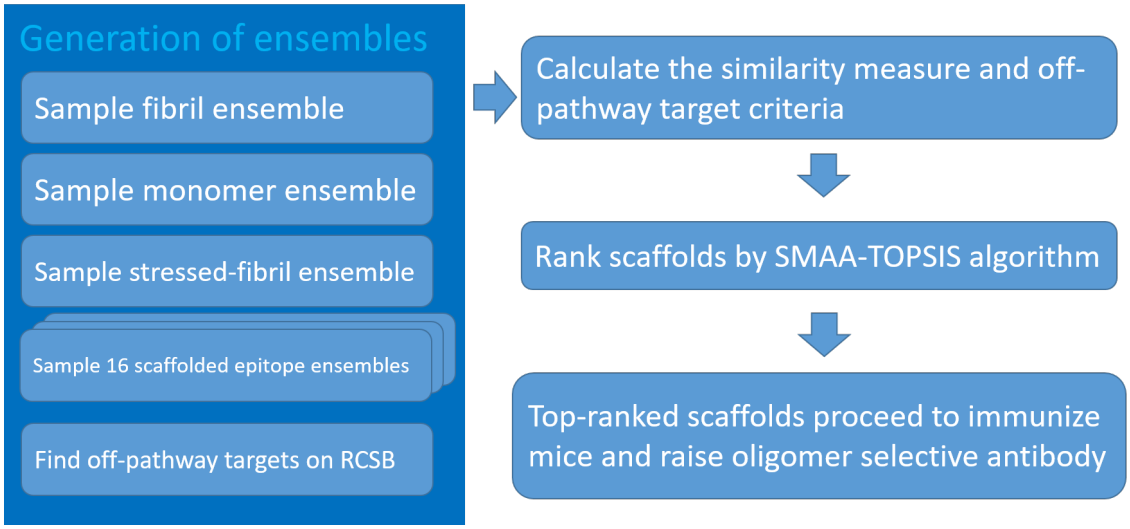


Figure 2: The workflow of *in silico* screening.

2.1 α-synuclein epitope prediction

In our approach, we operate from the hypothesis that a partially disordered protofibril ensemble is enriched in oligomer-selective conformational epitopes. That is, similar regions

are exposed in both toxic oligomers and stressed protofibrils. If so, the stressed fibril may be used to predict oligomer-selective epitopes. This hypothesis is supported by previous *in vitro* and *in vivo* evidence (*13–15*).

The EKTKEQ epitope in α-synuclein was predicted by the Collective Coordinate (CC) algorithm, which has been used previously to predict misfolding-specific epitopes in superoxide dismutase 1 (SOD1) (*15*). The CC algorithm predicts epitopes by identifying local unfolding events when a global denaturing stress is applied, where regions of a protein or multi-protein aggregate deviate structurally from their native structural conformation. Three metrics are used to measure local disorder: increased solvent accessible surface area (ΔSASA), loss of native contacts (ΔQ), and increased root mean squared fluctuations (ΔRMSF).

The procedure for predicting an epitope involves implementing a global unfolding potential, which biases the system to have 0.65 of its total native contacts. Ten independent biased ensembles of an α-synuclein fibril structure (PDB ID: 2N0A) using the CC algorithm, as well as a single equilibrium ensemble of the fibril using molecular dynamics (MD) simulation (*15*), are generated to calculate ΔSASA, ΔQ and ΔRMSF. The equilibrium ensemble serves as a reference for calculating the above three difference values. Multiple independent biasing simulations were performed in order to ensure consistency in the regions of the fibril structure that are observed to have relatively higher values of ΔSASA, ΔQ and ΔRMSF, thereby avoiding predictions based on rare fluctuations that might be present in a single biased simulation.

Fig. 3 bottom panel shows the sequence motifs larger than 3 amino acids that are predicted as epitopes by each of the three metrics, for each of the 5 chains in the protofibril structure (PDB 2N0A). Several epitopes were predicted by each metric; however, the consesus epitope EKTKEQ is predicted by all the three metrics, and was taken as the final predicted epitope for α-synuclein (Fig. 3 top panel).

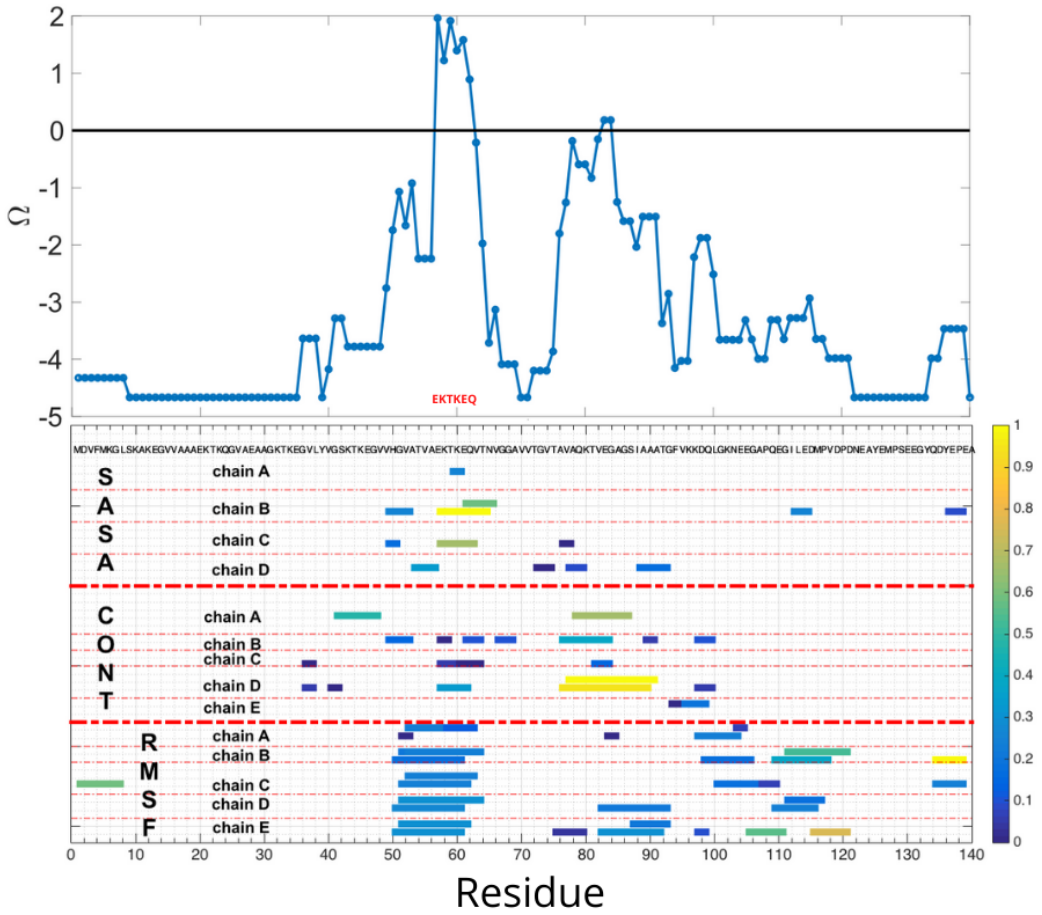


Figure 3: Collective coordinate epitope prediction for α-synuclein, using three criteria of increased SASA, loss of native contacts, and increased fluctuations (RMSF). Several epitopes were predicted by each criterion; however, only a single consensus epitope EKTKEQ was predicted. Chain E is not shown for ΔSASA because no epitope is predicted.

2.2 Scaffolding of epitopes

We scaffold epitopes by constructing cyclic peptides of the epitopes. First, the predicted epitope (EKTKEQ), or a shorter sequence that is subsumed by the epitope such as EKTK, is flanked on both sides with consecutive glycines. This is implemented computationally by mutating the native flanking residues to glycine using SCWRL4 (*17*). In addition, one more residue is mutated to cysteine on the N-terminal side of the sequence. Although, in principle, other amino acids or a combination thereof can be utilized, our choice of using glycines helps to focus immunogenic effects to just the epitope, since glycine is relatively chemically

inert compared to other amino acids (*18*), and has minimal immunogenicity (*19*). The cysteine is present to conjugate the peptides to immunogens such as Bovine serum albumin (BSA) or Keyhole limpet hemocyanin (KLH) via a disulfide bond, increasing the likelihood of generating antibodies to the scaffolded epitope. The topology of the cyclic peptide is obtained by head-to-tail linkage of the termini using a locally written Python script. We refer to a cyclic peptide epitope with n N-terminal glycines and m C-terminal glycines as cyclo($C - G_n -$ (Epitope) $- G_m$) or (n, m)Epitope. The cyclized peptide is then energy minimized in GROMACS using steepest descent algorithm, before running equilibrium MD simulations.

2.3 Generation of ensembles

Generally, we collect four different equilibrium molecular dynamics (MD) ensembles for each epitope. These correspond to the epitope in the context of an isolated monomer, the fibril, the stressed fibril oligomer model, and the cyclic peptide scaffold. We perform MD simulations with the open-source GROMACS (*20*) package and the community developed PLUMED library (*21*). The force field used here for all ensemble sampling is CHARMM36m (*22*), which is a modified version of CHARMM36 with improved modeling for disordered proteins. Each epitope ensemble is obtained specifically as follows.

2.3.1 Fibril ensemble

The conformational sampling starts from an existing experimentally resolved α-synuclein structure determined by solid-state NMR (*23*) (PDB ID: 2N0A). A protofibril composed of 5 chains (chains A-E of 2N0A) is solvated in 150mM Na and Cl aqueous solution, such that the system is neutral. After 50ps constant volume (NVT) and 150ps constant pressure (NPT) equilibrium simulations under positional restraints on heavy atoms, we perform an equilibration MD simulation up to 100ns until the convergence is seen from RMSD during the simulation. A 20ns equilibration MD continued from the previous 100ns simulation is

then performed, for collecting the equilibrium ensemble of the fibril, from which we sample snapshots at 20ps intervals. Only configurations of the middle chains (chain B-D) are collected in the ensemble, in order to reduce edge effects in the simulation. In total, there were 3003 configurations in the fibril ensemble obtained by this procedure. The above procedures of solvation and NVT-equilibration and NPT-equilibration are also implemented for the other ensembles described below, before the MD production runs.

2.3.2 Stressed fibril ensemble

The stressed fibril ensemble is a partially disordered fibril ensemble used to predict conformational epitopes similar to what might be presented by the oligomer. To generate this ensemble, a time-dependent global bias potential $V(Q,t)$ is implemented to partially unfold the α-synuclein fibril, where

$$V(Q,t) = \frac{1}{2}k(Q - Q_c(t))^2 \tag{1}$$

Q in equation (1) is a collective coordinate defined as the normalized count of the native contacts (*15*): Native contacts are pairs of heavy (non-hydrogen) atoms within 4.8Å of each other that are in different amino acids labeled by primary sequence residue index α, β, that satisfy $|\alpha - \beta| \geq 3$, and persist over 5 percent of the time in the first 100ns of equilibrium fibril simulation. Formally, the collective coordinate Q for any structure characterized by a set of heavy atom distances, r_{ij}, is defined as follows:

$$Q = \frac{\sum_{ij}^{N} Q_{ij}(r_{ij})}{\sum_{ij}^{N} \langle Q_{ij}(r_{ij}) \rangle_{native}} \tag{2}$$

where

$$Q_{ij}(r_{ij}) = \frac{1 - (\frac{r_{ij}}{r_0})^n}{1 - (\frac{r_{ij}}{r_0})^m} \tag{3}$$

where we take $r_0 = 4.8$Å, $n = 6$ and $m = 12$. A contact function $Q_{ij}(r_{ij})$ is defined for each heavy-atom pair i, j in the list of native contacts, which rapidly goes to one when r_{ij} is slightly lower than r_0 and rapidly goes to zero when r_{ij} is slightly larger than r_0. The

quantity in the numerator of Q in equation (2) is the sum of Q_{ij} in an arbitrary structure, where $\sum_{ij}^{N}$ is summation over the native contact list, and the quantity in the denominator is the Boltzmann average of the Q_{ij} in the fibril equilibrium ensemble. Thus, Q is typically a number between zero and unity.

$Q_c(t)$ in equation (1) stands for the target value of the collective coordinate, Q. It is a time-dependent quantity that starts from a value corresponding to the counts of all native contacts in the fibril equilibrium ensemble ($Q_c = 1$), which is then taken to linearly decrease with time, from $Q_c = 1$ to $Q_c = 0.65$ over 50 ns. Afterward, the bias is held fixed at $Q_c = 0.65$ for 150ns. During the last 50ns of this second period when Q_c is held fixed at 0.65, snapshots from the stressed fibril ensemble are collected at an interval of 15ps. The above process is repeated 10 times, including the ramp down and subsequent biased equilibration and sampling, to average the stochastic unfolding process of the protofibril and obtain a reliable biased ensemble. The capping chains at the ends of the fibril are discarded to avoid edge effects, yielding a total ensemble consisting of $50ns/15ps * 3\,\text{chains} * 10\,\text{repeats} = 100000$ sampled structures. A subset with 3500 configurations is randomly sampled from the total structural ensemble for efficiency of the calculation. In order to obtain an ensemble of the epitope that is partially disordered and exposed to solvent, we add on additional constraint on SASA of the epitope in the ensemble, as follows. Among the 3500 configurations, we further discard those structures that have a lower epitope SASA than the equilibrium epitope SASA, generating a stressed fibril ensemble with exposed epitope consisting of 3407 configurations. We note that at this point the epitope has already been predicted, and the above calculations are for comparison of the stressed/exposed protofibril ensemble with cyclic peptide scaffold motifs.

2.3.3 Monomer ensemble

Since α-synuclein is an intrinsically disordered protein (IDP) (*24*), its equilibrium ensemble is relatively difficult to sample sufficiently by normal MD simulation (*25*). We have previously

developed a method for generating equilibrium ensembles for IDPs (*26*), and have applied this method to generate equilibrium ensembles for several IDPs such as α-synuclein, Aβ peptide, and prothymosin α. We have also used the method to model the unfolded ensembles of natively folded proteins, including disulfide bonded unfolded ensembles such as that for SOD1 (*26*).

Here, we modify the method in reference (*26*) (which included a coarse-graining step) to generate an equilibrium monomer ensemble for α-synuclein. Here we do not employ a coarse-graining step. The steps involved in our ensemble generation include: **1.)** Generation of a conformationally diverse ensemble via the pivot algorithm and a "crankshaft" move wherein two randomly selected end points are fixed (*26*, *27*), and **2.)** Equilibration of each pivot– and crankshaft–randomized structure for a short simulation time. The descriptions for these steps follows.

The initial α-synuclein monomer configuration is obtained by extracting a single chain from the fibril structure (PDB ID: 2N0A). This structure is then altered by employing a generalization of the pivot algorithm (*28–31*), which is an efficient algorithm for generating ensembles for a self-avoiding random walk, as well as a "crankshaft" move (*26*, *27*) that randomizes ϕ and ψ angles between two randomly-selected backbone atoms along the peptide chain, such that the randomly-selected backbone atoms remain fixed. Pivot moves and crankshaft moves are attempted with equal probability. If a move results in a steric clash, it is rejected, and the next randomly selected move is then attempted. We take a simple approximation wherein ϕ and ψ are sampled from a (2π periodic) von Mises probability distribution with mean 0 and variance 1, i.e. $\mathrm{e}^{\cos\phi}/(2\pi I_0(1))$, where I_0 is the modified Bessel function of order 0. Pivot/crankshaft moves are repeatedly attempted until there are N successful pivot/crankshaft moves, where N is the length of the chain. This process completely randomizes the conformation such that information about the original structure is lost.

After a randomized conformation is obtained, the structure is then solvated in a box

determined by the principal axes wherein the peptide is at closest 1.2 nm from the box edges. The structure is then energy minimized, equilibrated in explicit modified TIP3P water and 150mM salt in an NVT ensemble for 100ps, and then equilibrated in an NPT ensemble for 300ps. This process generated 1587 different structures as initial configurations. We then performed a 3ns equilibrium simulation starting from each of the above initial configurations, collecting a snapshot every 1ns to be added to the monomer ensemble. As a result, we obtained a monomer ensemble with 4607 configurations (i.e. 93 simulations did not reach 3ns and 16 simulations did not reach 2ns within the simulation wall-clock time, due to variable simulation box sizes). IDP ensemble generation is an active area of research and several other methods of computational ensemble generation have been developed (*32*).

2.3.4 Scaffolded epitope ensembles

The construction of the initial structure used for scaffold simulations is described in Methods Section 2.2. After minimization, solvation, NVT-equilibrium and NPT-equilibrium simulation, a 300ns MD simulation is carried out, from which an initial ensemble is collected with constant sampling interval of 40ps. Since the cysteine side chain has to be solvent-accessible to form a disulfide bond with the carrier protein (KLH or BSA), some configurations were discarded if the SG atom in the cysteine was buried or if the sidechain of the cysteine pointed inside the cyclic peptide: The SG atom is defined as buried if its SASA is lower than 70% of that in an isolated Gly-Cys-Gly tripeptide (80Å^2). The cysteine side chain is defined as inward pointing if its dihedral angle ψ (angle between CB atom and the backbone plane) is within [-90, 90] degrees. In practice, the scaffolded cyclic peptide ensembles thus processed have a different number of configurations for each cyclic construct, ranging from 2352 to 5468 configurations. The fraction of the ensemble that satisfied the above criteria varied from about 39% to 97%.

2.4 Projecting ensemble distributions to lower dimension

We first let N be the sum of the number of structures in all four ensembles: Monomer, fibril, stressed fibril, and cyclic peptide scaffolds (Here N ranges from 13369 to 16485). We then construct a pairwise root mean squared deviation (RMSD) matrix that consists of the RMSD between any two epitope (EKTK, KTKE, or TKEQ) structures in any of the four ensembles. Each row of the $N \times N$ matrix contains the distances (RMSD values) from one structure. Two rows then constitute a 2-dimensional space giving the distances to two structures determined by which rows have been chosen. N rows, corresponding to the whole matrix, thus constitute an N-dimensional space where the location of each of N points, corresponding to each structure, is determined by the set of distances (RMSD) to all other structures (including itself). The set of N structures is thus represented by a distribution of points in an N-dimensional space.

Multidimensional Scaling (MDS) (*33*) or Stochastic Proximity Embedding (SPE) (*34*) was performed on the RMSD matrix to reduce the dimension from $N \times N$ to $N \times D$ where D is the desired lower dimension. The value of D depends on the application as described below. As well, whether MDS or SPE is used depends on the application, and will be described in detail in Sections 2.5, 2.6 and 2.7 below.

Applying MDS or SPE results in an ensemble of points in a reduced number of D dimensions. We then fit the effective distribution of the N points in D dimensions using kernel density estimation (KDE). (*34*) Since KDE performance worsens exponentially with higher dimensional data sets as a result of the so-called "curse of dimensionality" (*35*), a dimensional reduction step is essential. We call the structural distribution (technically the structural distance distribution) of a given ensemble, say, the monomer structural distribution, to be the KDE distribution of the points that correspond to the monomer structures in the RMSD matrix.

2.5 Calculating ensemble overlap in 1D

We visualize the similarity between two ensembles intuitively by representing them in 1D (this is not used as a ranking criterion). The overlap is the percentage of a KDE distribution that overlaps with another KDE distribution. The formula we use here for the overlap is: Overlap $= 1 - \frac{1}{2}\int_{-\infty}^{\infty} |P - Q|\, dx$, where x is the first MDS coordinate, and $P(x)$ and $Q(x)$ are two KDE distributions in this 1D coordinate. For example, for two gaussians of unit variance and mean separation a, the overlap is $1 - \mathrm{erf}\left(a/(2\sqrt{2})\right)$.

2.6 Ensemble similarity measures

We compare conformational ensembles of the epitope in the four different contexts, including monomer, fibril, stressed fibril and cyclic scaffold, using two measures: Embedding Depth and Jensen-Shannon Divergence. These are defined as follows.

2.6.1 Embedding depth

The embedding depth is most easily understood when taken as a measure between a single structure and a given ensemble. In this context, the embedding depth quantifies how deeply that structure is embedded within the ensemble, when both structure and ensemble are projected onto some coordinate metric such as MDS coordinates. The structure-to-ensemble embedding depth $\mathcal{D}_{\delta(\mathbf{x}-\mathbf{x_o})|P}$, for a structure characterized by a point at $\mathbf{x} = \mathbf{x_o}$ embedded in a distribution $P(\mathbf{x})$, is defined as the fraction of the ensemble that has a lower KDE probability than the particular structure (Fig. 4):

$$\mathcal{D}_{\delta(\mathbf{x}-\mathbf{x_o})|P} = \int_{-\infty}^{\infty} d\mathbf{x} P(\mathbf{x}) \Theta\left(P(\mathbf{x_o}) - P(\mathbf{x})\right). \tag{4}$$

In equation (4), $\Theta(P)$ is the Heaviside step function, which returns 1 if $P > 0$, otherwise 0.

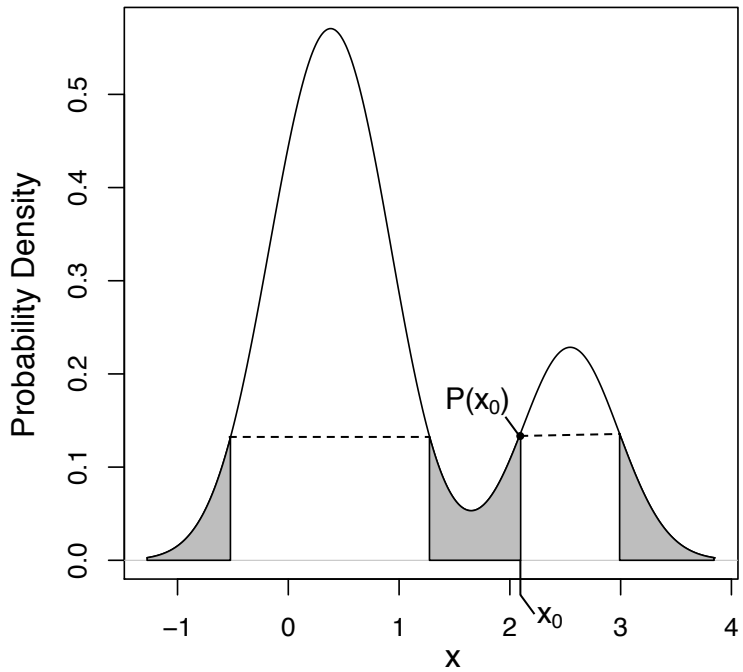


Figure 4: Illustration of the embedding depth of the point x_o in a multimodal distribution $P(x)$. The embedding depth of point x_o is given by the integral over all parts of the distribution with probability less than $P(x_o)$. Note there are 4 points with the same $P(x_i) = P(x_o)$ and thus the same embedding depth.

The mode of a distribution has an embedding depth of one because all of the ensemble has a lower KDE probability than the mode. On the other hand, outliers of a distribution have embedding depths close to zero.

The embedding depth of one distribution $Q(\mathbf{x})$ within another $P(\mathbf{x})$ can be found by integrating equation (4) over the distribution $Q(\mathbf{x})$:

$$\mathcal{D}_{Q|P} = \int_{-\infty}^{\infty} d\mathbf{x}\, Q(\mathbf{x}) \int_{-\infty}^{\infty} d\mathbf{x}'\, P(\mathbf{x}')\Theta\left(P(\mathbf{x}) - P(\mathbf{x}')\right). \tag{5}$$

The embedding depth between two distributions is in general non-reciprocal, in that $\mathcal{D}_{Q|P} \neq \mathcal{D}_{P|Q}$. It is straightforward to show that the embedding depth of a distribution within itself $\mathcal{D}_{P|P} = 1/2$. This follows because the special case of $\mathcal{D}_{P|P}$ must be symmetric

(reciprocal), so that switching $\mathbf{x}$ and $\mathbf{x}'$ in $\mathcal{D}_{P|P}$ must yield the same equation. Therefore,

$$\begin{aligned}
\mathcal{D}_{P|P} &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} d\mathbf{x}\, d\mathbf{x}'\; P(\mathbf{x})P(\mathbf{x}')\Theta\left(P(\mathbf{x}) - P(\mathbf{x}')\right) \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} d\mathbf{x}\, d\mathbf{x}'\; P(\mathbf{x})P(\mathbf{x}')\frac{1}{2}\left[\Theta\left(P(\mathbf{x}) - P(\mathbf{x}')\right) + \Theta\left(P(\mathbf{x}') - P(\mathbf{x})\right)\right] \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} d\mathbf{x}\, d\mathbf{x}'\; P(\mathbf{x})P(\mathbf{x}')\frac{1}{2}\left(1\right) = \frac{1}{2}\,. \qquad (6)
\end{aligned}$$

When used to compare two conformational ensembles, $\mathcal{D}_{A|B}$ or $\mathcal{D}_{\text{A-in-B}}$ represents the extent to which the ensemble A is subsumed by ensemble B. A similar concept in measuring the embedding of one ensemble in the other has been introduced before in the context of statistics (*36*), but we are not aware of such a measure being previously applied to conformational ensemble comparison. The embedding depth measures used for ranking ($\mathcal{D}_{\text{cyclic-in-stress}}$, $\mathcal{D}_{\text{cyclic-in-monomer}}$, and $\mathcal{D}_{\text{cyclic-in-fibril}}$) are calculated in an MDS-reduced 3D conformational space, where 95% of the matrix information can be preserved (Fig. S1b).

2.6.2 Jensen-Shannon Divergence (JSD)

The similarity between two different ensemble distributions is measured using the Jensen-Shannon Divergence (JSD) (*37–39*), which is implemented here using the ENCORE software (*34*). JSD is a symmetrized and smoothed version of the Kullback–Leibler divergence (*40*, *41*), D_{KL}, which is a difference measure between two distributions. JSD is defined by

$$JSD(P,Q) = \frac{1}{2}D_{KL}(P,M) + \frac{1}{2}D_{KL}(Q,M) \qquad (7)$$

where

$$D_{KL}(P,Q) = \int_{-\infty}^{\infty} P\left(x\right)\log\left(\frac{P\left(x\right)}{Q\left(x\right)}\right)dx \qquad (8)$$

where $P(x)$ and $Q(x)$ are two conformational ensembles in SPE-reduced dimensional space, and M is defined by the average of the two distributions P and Q: $M(x) = (P(x)+Q(x))/2$.

The value of the JSD measure depends on the dimensionality of the SPE-reduced space,

with lower dimension tending to have smaller JSD (i.e. more overlap). The JSD values used for ranking cyclic peptide scaffolds ($\text{JSD}_{\text{cyclic-fibril}}$, $\text{JSD}_{\text{cyclic-monomer}}$, and $\text{JSD}_{\text{cyclic-stress}}$) are produced by weight averaging the JSD from 3D to 11D by the inverse of the SPE residuals in each dimension (Fig. S1 for these residuals). By this measure, the higher dimensions contribute more to the weighted average JSD. The convergence test in Section S.2 shows that the information loss is about 3% in 3D and the information loss is less than 1% in 11D for scaffold (1,4)TKEQ for example (Fig. S1). JSD as calculated in equation (7) ranges from 0 to log(2), but in this paper, we normalize JSD to lie between 0 to 1.

2.7 Evaluating off-pathway target criteria

We scrape through the RCSB database of resolved protein structures, to find potential off-pathway targets of antibodies that could be generated by each cyclic peptide scaffold. An off-pathway target of concern meets three conditions: **1.)** same primary sequence as the epitope in the PDB structure, **2.)** structural similarity to scaffolded epitope, and **3.)** sufficiently high solvent exposure so that it is likely to be antibody-accessible.

The first criterion is assessed by finding "hits" when scraping through the RCSB database (e.g. any protein that contains one or more TKEQ motifs). The second criterion is assessed by the PDB entry's structural embedding depth in the scaffolded cyclic peptide ensemble. The embedding depth is calculated in the MDS-reduced 5D space of the RMSD matrix. The third criterion is assessed by the finding the fraction of the scaffolded epitope ensemble with less solvent-accessible surface area (SASA) than the PDB entry (i.e. $f(\text{SASA}_{\text{off-target-exceed-cyclic}})$).

The severeness of a potential off-pathway target for a given scaffold is determined by summing, over all PDB structure "hits", the product of the structural embedding depth ($\mathcal{D}_{\text{off-target-in-cyclic}}$) and the fraction of cyclic peptide ensemble SASA exceeded ($f(\text{SASA}_{\text{off-target-exceed-cyclic}})$). The criteria used for ranking (before a linear rescaling) is the negative of this severeness,

which is represented by equation (9) below:

$$\mathrm{OP}_{\text{unscaled}} = -\sum_{\text{PDB "hits"}} \mathcal{D}_{\text{off-target-in-cyclic}} \times f(\mathrm{SASA}_{\text{off-target-exceed-cyclic}}) \tag{9}$$

By this convention, a higher (less negative) criteria value corresponds to a more favorable scaffold. Since $\mathcal{D}_{\text{off-target-in-cyclic}}$ and $f(\mathrm{SASA}_{\text{off-target-exceed-cyclic}})$ are very small for most RCSB entries (i.e. most entries are not of significant concern), we employ a cut-off criterion wherein $\mathcal{D}$ and f have to both exceed 5% for the entry to be included in a criteria calculation. Finally, the criteria are linearly rescaled to be within the range [0,1].

Sometimes, the epitope appears multiple times in a single RCSB PDB entry because of multiple deposited models, multiple chains or repetitive epitope occurrence on a single chain. If such multiple occurrence happens, the $\mathcal{D}_{\text{off-target-in-cyclic}}$ and $f(\mathrm{SASA}_{\text{off-target-exceed-cyclic}})$ in equation (9) are taken as the average of the values over all epitope occurrences in the corresponding PDB entry.

The search result in the RCSB database will also contain the fibrillar and structured monomer α-synuclein entries. The fibril entries are excluded from the off-pathway target calculation, because we have already compared cyclic and fibril ensembles using $\mathrm{JSD}_{\text{cyclic-fibril}}$. We have also compared cyclic and unstructured monomer ensembles using $\mathrm{JSD}_{\text{cyclic-monomer}}$, however monomer structures in the RCSB database are kept because they are assumed to be at least partially structured due to effects such as peptide-membrane interactions.

2.8 SMAA-TOPSIS ranking algorithm

Scaffolds are ranked using an adaptation of the so-called SMAA-TOPSIS algorithms, which stand for a combination of Stochastic Multi-criteria Acceptability Analysis and Technique for Order Performance by Similarity to Ideal Solution. (*42*) The ranking algorithm we employ here uses a cumulative measure of the retention probability that a candidate is retained in the top n leads (*16*). The method generates a ranking of candidates when multiple criteria

are used for screening the candidates, and when those screening criteria have a unknown predictive importance *a priori*. The screening criteria as well as the importance weights for those criteria are both subject to error, and an exact weight assignment for the importance of the screening criteria is not needed. Instead, weight distributions are assigned, where the mean weight $\overline{w}$ represents the relative importance of each criterion and the standard deviation of the weight distribution represents the uncertainty in the importance for that criterion. The weight distribution is taken to be a uniform distribution:

$$f(w) = \begin{cases} \frac{1}{2\Delta} \left(if\ \overline{w} - \Delta < w < \overline{w} + \Delta\right) \\ 0 \left(\text{elsewhere}\right) \end{cases} \tag{10}$$

where $\overline{w}$ stands for the mean (or center) and Δ stands for the distribution width. All the criteria have $\Delta = 0.5$ in this paper (or equivalently standard deviation σ of 0.288). The mean weights for each criterion are given in Table 1:

Table 1: Mean weights of ranking criteria.

Criterion	**Mean weight $\overline{\mathbf{w}}$**
JSD_{cf}	1
JSD_{cm}	1
1 - JSD_{cs}	0.5
1- $\mathcal{D}_{c\|f}$	1
1- $\mathcal{D}_{c\|m}$	1
$\mathcal{D}_{c\|s}$	0.5
Off-target	1.5

c=cyclic, f=fibril, s=stressed fibril, m=monomer. See also Table S1.

The ranking algorithm rescales all screening criterion to unity so weights can be properly applied. The Ideal Best and Ideal Worst entries are given similarity screening criteria of either all 1's or all 0's respectively; These numbers are given for each scaffold in Table S1. All 48 scaffolds show strong dissimilarity to fibril, whereas other screening criteria show more variance (Table S1). The compression of values for JSD_{cf} around unity means that all scaffolds perform similarly. For this criterion, the ideal worst entry is still taken to still

be 0 (although no scaffold performed that poorly), rather than the lowest value followed by a rescaling to 0. This procedure thus prevents insignificant discrepancies between scaffolds from being amplified. Similarly, the values of $\mathcal{D}_{c|f}$ are all either essentially zero or very small, so we take the ideal worst entry to be 1 and the ideal best to be 0, to avoid ranking on statistically-insignificant differences. A server where the SMAA-TOPSIS algorithm may be run to rank the users' own data can be found at `http://bjork.phas.ubc.ca`. The server can select and rank candidate leads from multiple competing screening tests that would otherwise make the selection of leads a nontrivial maximum-likelihood ranking problem. (*16*)

3 Results

Using our previously developed epitope prediction algorithm (*15*), amino acid sequence EKTKEQ (residues 57-62 in α-synuclein) was predicted as a misfolding-specific epitope (*43*) (Fig. 3). To dissect the key amino acids in this epitope, we have analyzed three separate contiguous sequences subsumed by the epitope, namely EKTK, KTKE and TKEQ.

3.1 Comparing ensembles in reduced conformational space

In order to design epitope scaffolds that are capable of eliciting antibodies that are conformationally selective to soluble oligomers, the conformational ensemble of an epitope in a cyclic peptide scaffold is compared with three other ensembles of the epitope, namely, the monomer, fibril, and the stressed (partially disordered) fibril. The stressed fibril is taken as a proxy for the conformational ensemble of the epitope in a soluble oligomer, as successfully demonstrated previously (*13*, *14*). A desired cyclic peptide scaffolded epitope construct that is oligomer-selective would have high ensemble similarity to the stressed fibril, and as well, low ensemble similarity to the equilibrium fibril and isolated monomer.

To determine the similarity between the four epitope ensembles (i.e. scaffold, stressed fibril, fibril, and monomer), we first generated these ensembles by performing molecular

dynamics (MD) simulations (see Methods Section 2.3). The conformational similarity of the epitope between ensembles is quantified in a reduced conformational space of the epitopes (see Methods Section 2.4). For example, Fig. 5 shows the structural distributions in 1D for the four ensembles of epitope TKEQ, in two scaffolds. The two scaffolds chosen for this illustration are cyclo(CGTKEQGGGG) or (1,4)TKEQ, which stands for 1 glycine N-terminal to the TKEQ epitope and 4 glycines C-terminal to the epitope, and cyclo(CGGTKEQGGG), or (2,3)TKEQ.

Although information may be lost when projecting high dimensional ensembles to lower dimension, the distributions in Fig. 5 do illustrate how the ensembles overlap. The fibril ensemble consists of a predominant sharp spike because of its rigid structure. On the other hand, the monomer ensemble is broadly distributed because it is natively unstructured and conformationally diverse. The stressed fibril ensemble distributes around the fibril ensemble because it is generated by forcing the partial unfolding of fibril by biased MD. Each cyclic peptide scaffold possesses a different distribution. As well, the stressed fibril, monomer, and fibril ensembles are slightly different in each case of Fig. 5, because all structural distributions are distance distributions based on RMSD, which are different for each scaffold ensemble. From the degree of overlap or similarity between a scaffolded cyclic peptide ensemble and the other ensembles, we can assess whether the scaffold has the potential to raise oligomer-selective antibodies while sparing fibril and monomer.

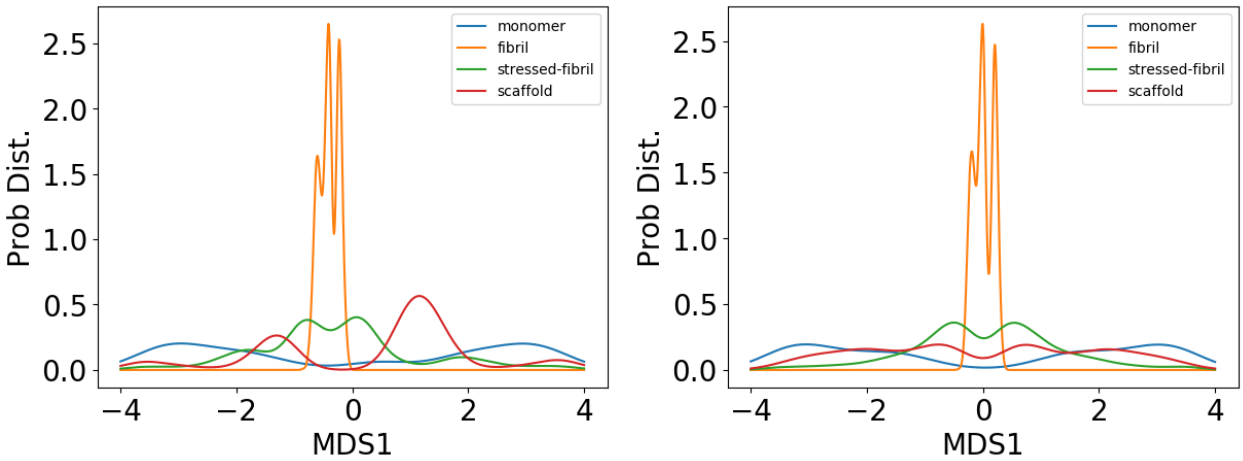
For example, both (1,4)TKEQ and (2,3)TKEQ scaffold ensembles have very low overlap with the fibril ensemble (3% and 8% respectively) (see Methods Section 2.5). Also, they both have high overlap with the stressed fibril, where (2,3)TKEQ is somewhat higher than (1,4)TKEQ (68% and 47%). The above (low fibril overlap, high stressed fibril overlap) are desired properties of a conformationally selective immunogen. On the other hand, the scaffolds also have high overlap with isolated monomer ensemble, with (2,3)TKEQ having much higher overlap than that in (1,4)TKEQ (74% and 47%). High overlap with the monomer ensemble is not a favorable property because of the possibility of targeting healthy protein.

Naïvely from the 1D overlap measure, (1,4)TKEQ may be better than (2,3)TKEQ, since it has a higher ratio of stressed fibril ensemble overlap to monomer ensemble overlap (1.0 for (1,4)TKEQ *vs.* 0.92 for (2,3)TKEQ). However, with a more rigorous similarity measure analysis allowing for higher projected dimensions, and the ranking algorithm in Section 2.8, (2,3)TKEQ actually ranks higher than (1,4)TKEQ.

The similarity or overlap between ensembles is rigorously quantified by two measures, Jensen-Shannon Divergence (JSD) and embedding depth (see Methods Section 2.6). JSD is a measure that represents the dissimilarity between two ensembles, which has been used previously to compare protein ensembles (*34*, *39*). Embedding depth represents the extent to which a given ensemble is subsumed by another, i.e. to what extent conformations in one ensemble are contained within another ensemble. It is a non-reciprocal measure that is introduced in Methods Section 2.6.1 to compare protein conformational ensembles (see also Ref. (*36*)). For example, the embedding depth of the fibril ensemble within the stressed fibril ensemble in Fig. 5a is 0.338, because fibril conformations are contained within the stressed fibril ensemble. Note that embedding depth between two identical ensembles is 0.5, so 0.338 has represented a large degree of embedding. On the other hand, the JSD between the two ensembles is 0.984, which represents two almost completely dissimilar ensembles. The efficacy of a scaffold to target a conformational species may thus likely be better represented by the embedding depth than the JSD.

3.2 Scaffold screening criteria and scaffold ranking

In Section 3.1, we showed that the various similarity measures that we had defined between a scaffolded cyclic peptide epitope and other ensembles can be optimized by varying the number of flanking N-terminal and C-terminal glycines (See Methods Section 2.2). The ensemble overlap of a given epitope with other ensembles can be tuned by changing the scaffolding residues. We thus computationally constructed $4 \times 4 = 16$ scaffolds for each epitope, corresponding to the epitope being flanked by anywhere between 1 and 4 glycines



(a) (1,4)TKEQ scaffold

(b) (2,3)TKEQ scaffold

Figure 5: The equilibrium ensemble distributions for the TKEQ epitope projected by the multidimensional scaling (MDS) method (*33*) onto the first MDS dimension. For a given epitope, different cyclic peptide scaffolds possess different distributions, which will result in different overlap with the other three ensembles. By comparing the degree of ensemble overlap, the conformational selectivity of a scaffold can be assessed. The scaffolds shown are (1,4)TKEQ in panel (a), and (2,3)TKEQ in panel (b).

on each terminus or "side", or $16 \times 3 = 48$ total for epitopes EKTK, KTKE, and TKEQ.

Fig. 6 shows the results for three α-synuclein epitopes; EKTK, KTKE and TKEQ, illustrating changes that can occur in ensemble similarity across all 48 scaffolds, as measured by Jensen-Shannon Divergence JSD (see Methods Section 2.6.2), embedding depth $\mathcal{D}_{A|B}$ (see Methods Section 2.6.1) and off-pathway target criteria OP (see Methods Section 2.7), by varying the number of flanking N-terminal and C-terminal glycines. Detailed values are in Table S1. The changes in the values corresponding to the fibril ensemble ($\text{JSD}_{\text{cyclic-fibril}}$, and $\mathcal{D}_{\text{cyclic-in-fibril}}$) are all very small on the scale of the plots, so they appear to remain unchanged across all scaffolds. In practice, this means that these two criteria do not contribute significantly to the ranking between scaffolds. For epitopes EKTK and KTKE, there is a generally decreasing trend of JSD with increasing scaffold length (total residue number in the cyclic peptide), and a generally increasing trend for $\mathcal{D}$ with increasing scaffold length. These trends are less significant for epitope TKEQ. We address this phenomenon further in Section 3.4.

The difficulty in manually assessing good scaffolds from multiple similarity measures led

us to apply a systematic ranking method for the selection of the best performing scaffolds. The performance of each scaffold is assessed by 7 ranking criteria, including $\text{JSD}_{\text{cyclic-fibril}}$, $\text{JSD}_{\text{cyclic-monomer}}$, 1-$\text{JSD}_{\text{cyclic-stress}}$, 1-$\mathcal{D}_{\text{cyclic-in-fibril}}$, 1-$\mathcal{D}_{\text{cyclic-in-monomer}}$, $\mathcal{D}_{\text{cyclic-in-stress}}$ and OP (the off-pathway target criterion). We formulate the ranking such that large values in all criteria are desired, so we thus subtract some JSD values from 1 is to convert dissimilarity to similarity, and we subtract some and $\mathcal{D}$ values from 1 to convert similarity to dissimilarity.

Scaffolds are ranked using the SMAA-TOPSIS algorithm (*16*) using these 7 criteria (see Methods Section 2.8). In the ranking algorithm, each criterion is assigned a weight for its relative importance; The weights used here are given in Table 1. Since $\text{JSD}_{\text{cyclic-monomer}}$ and 1-$\mathcal{D}_{\text{cyclic-in-monomer}}$ both represent the effect of monomer dissimilarity, the importance of monomer in the ranking amounts effectively to the sum of the two weights. The same applies to 1-$\text{JSD}_{\text{cyclic-stress}}$ and $\mathcal{D}_{\text{cyclic-in-stress}}$, as well as $\text{JSD}_{\text{cyclic-fibril}}$ and 1- $\mathcal{D}_{\text{cyclic-in-fibril}}$. Thus, the importance respectively for fibril, monomer, stressed fibril, and off-pathway target criteria are 2, 2, 1, 1.5. The ranking for all 48 scaffolds is shown in Table S1. Such a ranking can guide which cyclic peptide scaffolds are most promising to use in an active immunization for antibody generation to oligomer targets of α-synuclein.

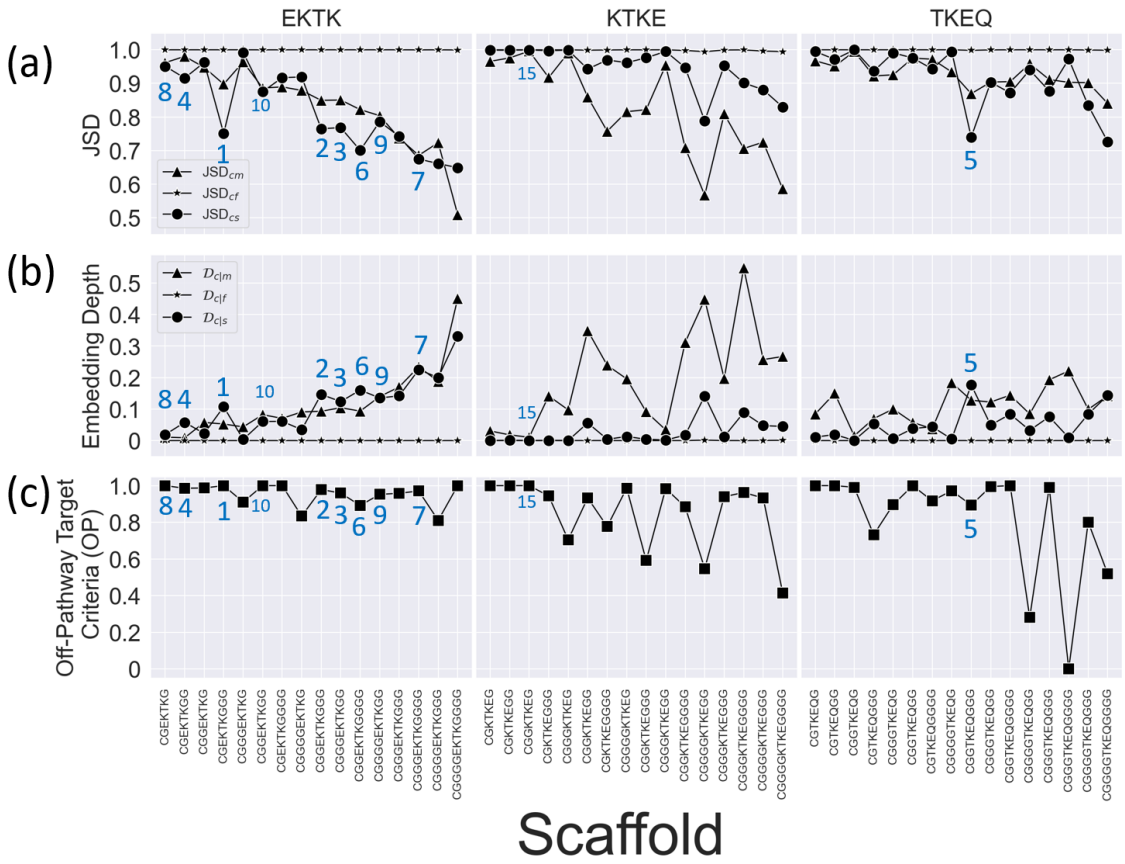


Figure 6: Measures for the rankings of all 16 epitope scaffolds, for three overlapping 4 residue sub-epitopes of EKTKEQ in α-synuclein (top of each column). (a) Scaffolded cyclic peptide ensemble dissimilarity to monomer (triangle), fibril (star), and stressed fibril (circle) ensembles, as measured by Jensen-Shannon Divergence (JSD), showing the changes in ensemble overlap with varying numbers of flanking glycines. (b) Scaffolded cyclic peptide ensemble embedding depth within the monomer (triangle), fibril (star), and stressed fibril (circle) ensembles, showing the changes in ensemble embedding with varying number of flanking glycines. (c) Normalized off-pathway targeting values (OP) for scaffolds with varying number of flanking glycines. Higher values indicate there is less predicted off-pathway targeting by a given scaffold. The ranks of the top 10 scaffolds are indicated in the figure panels, along with the rank for the highest ranking scaffold (15) for epitope KTKE.

3.3 Embedding depth as a similarity measure compared with JSD

The various ensemble similarity measures are compared with each other in Fig. 7, which shows matrices of the Pearson's correlation coefficient values, across all the scaffolds in Table S1. JSD and embedding depth $\mathcal{D}$ are compared for all scaffolds, along with three other

scaffold properties: Root Mean Squared Fluctuation of the epitope (RMSF), scaffold total residue length, and the ranking of each scaffold. JSD is calculated between ensembles by weight averaging values from 3 to 11 dimensions as described in Section 2.6.2; Embedding depth $\mathcal{D}$ is calculated in 3-dimensions as described in Section 2.6.1. The three JSDs ($\text{JSD}_{\text{cyclic-stress}}$, $\text{JSD}_{\text{cyclic-fibril}}$, $\text{JSD}_{\text{cyclic-monomer}}$) have mutually positive correlations, as shown in the top-left 3×3 matrices in Fig. 7a, b and c. This indicates that these JSD values may have implicit dependencies on one another and may contain some degree of redundant information. For epitope TKEQ, the JSD correlations between cyclic peptide ensembles and fibril, stressed fibril, and monomer ensembles may result at least in part from their flexibility, i.e. more flexibility of a cyclic peptide allows more conformational diversity, allowing in turn for more similar conformations to the other ensembles, reducing the JSD. Conversely, a rigid scaffold would have a narrow structural distribution, and thus minimal overlap with any of the other ensembles, resulting in larger JSD. The JSD values of various scaffolds thus have strong negative correlation with the RMSF for this epitope (8th column/row in Fig. 7c). This anticorrelation of RMSF with JSD is recapitulated partially by the (weaker) correlation of RMSF with $\mathcal{D}$. This effect may indicate a systematic deviation of conformational similarity for scaffolds of this epitope, and may suggest better targeting can be achieved for EKTK and KTKE, where the anti-correlation of JSD with RMSF is less significant.

Embedding depth can help elucidate the implicit effects of scaffold rigidity. A rigid scaffold that is deeply embedded within the stressed fibril would be a desirable candidate. Compared across scaffolds for a given epitope as in Fig. 7, an anticorrelation between $\mathcal{D}_{c|s}$ and RMSF would be ideal for oligomer targeting. This is not observed for any of our epitopes, but the correlations are not significant for EKTK and KTKE, ($p = 0.33$ and $p = 0.25$) while for TKEQ the significance of the correlation of 0.76 is $p = 0.0007$. It is also worth noting that we are integrating a significant amount of information when investigating correlations across all scaffolds for a given epitope, in order to compare one epitope with another. The variance scaffold to scaffold for a given epitope is sufficiently large to allow for highly ranking

scaffolds for any of the epitopes. That said, the median rankings for EKTK, KTKE, and TKEQ are 10.5, 37.5, and 25 respectively, and 9 of the top 10 scaffolds are for EKTK.

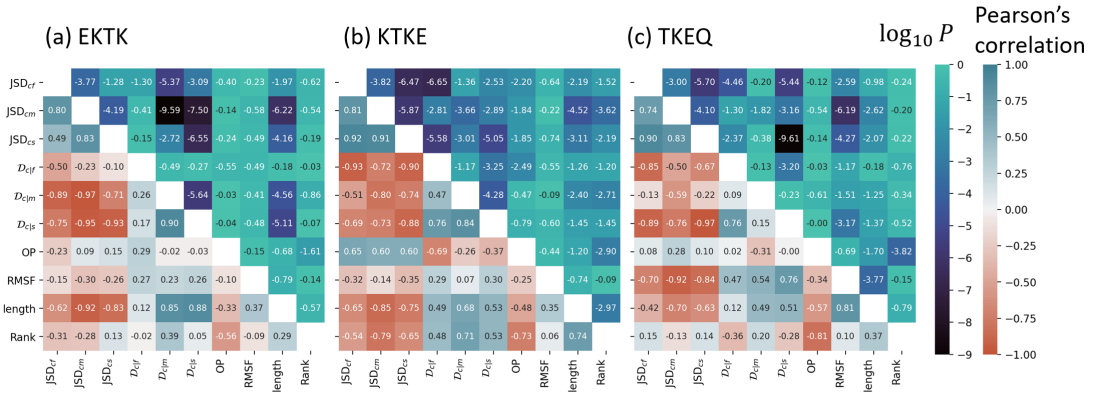


Figure 7: Correlation matrices of the ensemble comparison metrics JSD and $\mathcal{D}$, and three other scaffold properties: Dynamic flexibility (RMSF) of an epitope, total residue length of the cyclic peptide scaffold, and the ranking, for α-synuclein epitopes (a) EKTK, (b) KTKE, and (c) TKEQ. The number in each square is the Pearson correlation coefficient.

3.4 Scaffold properties and their impact on performance

The top 5-ranked scaffolds—(1,3)EKTK, (2,3)EKTK, (3,2)EKTK, (1,2)EKTK, and (2,3)TKEQ—all have substantially lower $\text{JSD}_{\text{cyclic-stress}}$ than $\text{JSD}_{\text{cyclic-monomer}}$, higher $\mathcal{D}_{\text{cyclic-in-stress}}$ than $\mathcal{D}_{\text{cyclic-in-monomer}}$, and high off-pathway (OP) target values (less off-pathway target severeness) (Fig. 6 and Table S1). $\text{JSD}_{\text{cyclic-fibril}}$ is approximately 1 for all scaffolds, and likewise $\mathcal{D}_{\text{cyclic-in-fibril}} \approx 0$, so these screening criteria cannot discriminate scaffolds.

We found that scaffolds with more residues tend to have smaller JSDs, and larger embedding depths $\mathcal{D}$ (Fig. 6 and Fig. 7 "length" row/column). The trend is particularly apparent for EKTK scaffolds. It is likely that this trend is due to scaffold rigidity, since larger cyclic peptides have less structural constraint and more flexibility, allowing them to have higher overlap with the other disordered ensembles. On the other hand, while RMSF correlated positively with peptide length for all epitopes, only TKEQ achieved statistical significance (Fig. 8a), suggesting that the primary sequence of the epitope is as important as scaffold size in determining its dynamics and flexibility, and thus its potential similarity to other en-

sembles. As another indicator that primary sequence determines epitope properties, scaffold size itself does not correlate significantly with a scaffold's ranking performance for TKEQ and EKTK (Fig. 8b), primarily because it is the *relative* overlap with the stressed fibril and monomer ensembles that determines the ranking. For KTKE, the correlation of scaffold size with ranking is such that smaller scaffolds tend to perform significantly better (Fig. 8b), mostly because the strongest trend with increasing scaffold length is increased overlap with the monomer ensemble (Fig. 6), which is an undesirable trait. The rankings themselves for KTKE are generally lower than those for the other two epitopes.

Plotting the scaffold rank *vs.* the *difference* in the $\mathrm{JSD}_{cm} - \mathrm{JSD}_{cs}$, as well as the difference $\mathcal{D}_{c|s} - \mathcal{D}_{c|m}$ (Fig. 8c,d) gives an indication of the importance of the relative overlap between the stressed-fibril *vs.* the monomer ensembles. Both abscissa in Fig. 8c,d are plotted as desirable quantities—the larger the value, the better the performance (higher rank or lower numerical rank value). For epitopes EKTK and KTKE, there is a significant correlation between a scaffold's rank and these relative overlap measures. For epitope TKEQ, the correlation is not significant however, essentially because the variation in the off-pathway (OP) target criterion strongly affects the ranking for this epitope. This can be seen by retaining only those TKEQ scaffolds with OP above a threshold such as 0.85 for example, and measuring the correlation for this reduced, filtered dataset. This reduced dataset consists of 11 out of 16 scaffolds and yields $r = -0.68$, $p = 0.019$ in Fig. 8c and $r = -0.88$, $p = 0.0002$ in Fig. 8d.

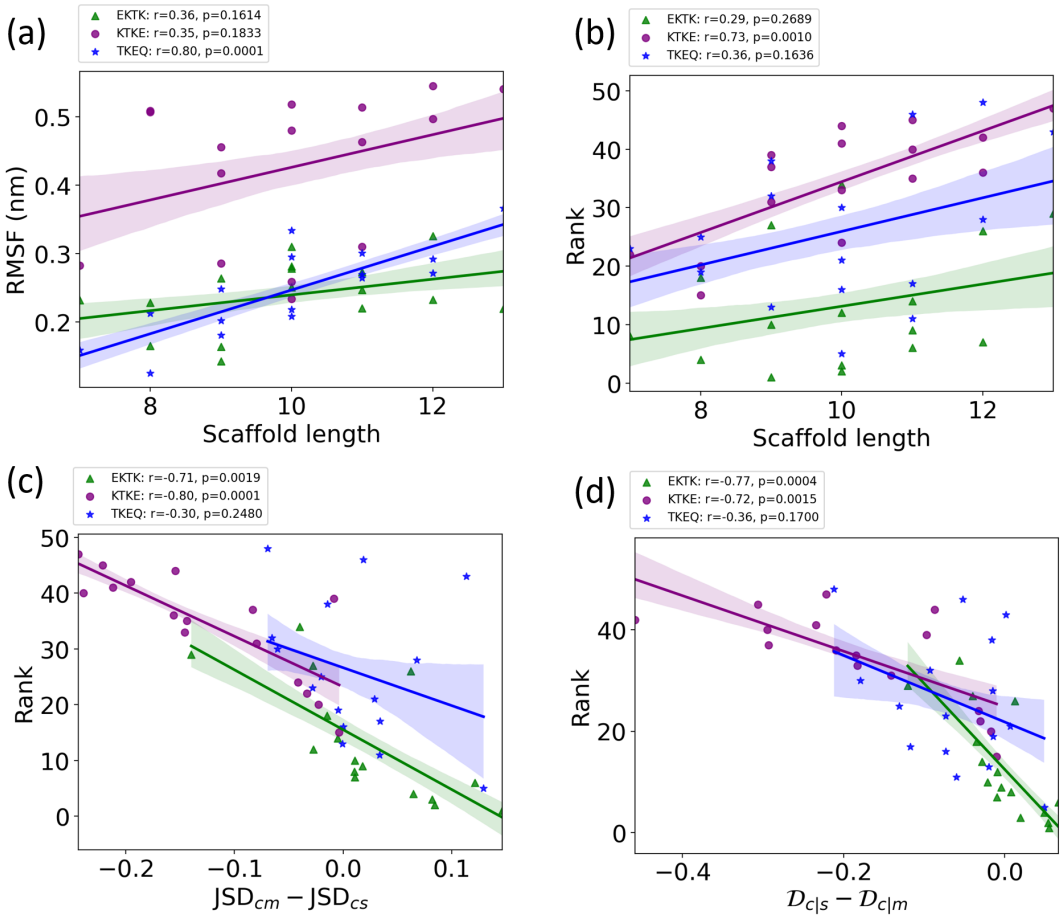


Figure 8: Epitope-dependent correlation between (a) RMSF and scaffold length (number of residues in the scaffold), (b) ranking and scaffold length, (c) rank and the quantity $\mathrm{JSD}_{cm} - \mathrm{JSD}_{cs}$, and (d) rank and $\mathcal{D}_{c|s} - \mathcal{D}_{c|m}$, for α-synuclein epitope scaffolds. The Pearson correlation coefficient, r and the corresponding p-values are given for EKTK (green triangles), KTKE (purple circles) and TKEQ (blue stars) scaffolds. The shaded areas around the fitted lines are the 68% confidence intervals corresponding to the standard errors.

Some epitopes tend to have better performance than others. EKTK and TKEQ scaffolds are generally ranked higher than KTKE scaffolds. As mentioned above, the median rankings for EKTK, TKEQ and KTKE are 10.5, 25, and 37.5 respectively. For the top 20 scaffolds, there are 12 EKTK scaffolds, 6 TKEQ scaffolds, and only 2 KTKE scaffolds, and 9 of the top 10 scaffolds are for EKTK, with only one TKEQ scaffold ranked 5th (Fig. 6). The poor performance of KTKE scaffolds appears to be due to their systematically lower structural similarity to the stressed fibril than to the unstructured monomer ensemble, as shown in

Fig. 6. The higher monomer similarity may itself be due to the higher conformational flexibility (RMSF) of KTKE scaffolds, which is roughly twice as high on average as the RMSF of epitopes EKTK or TKEQ (Fig. 8a). I.e. flexibility in the scaffold construction could favor greater similarity to the monomer ensemble, because the monomer ensemble is inherently more conformationally diverse than that of the stressed fibril.

Since a cyclic peptide has an inherent curvature which may more strongly resemble the conformations of an epitope in a turn or bend, the averaged virtual bond angles representing the local curvatures of epitopes in the stressed fibril ensemble are calculated and compared. The virtual bond angle is defined as $\pi - \angle C^{\alpha}_{i-1} C^{\alpha}_{i} C^{\alpha}_{i+1}$ in the kink model (*44–46*), where C^{α}_{i} is the C_{α} atom of the ith amino acid. This angle has been shown to represent curvature in the continuum limit (*47*). However, the curvature of KTKE (1.13 ± 0.14) is not significantly lower than that of EKTK (1.18 ± 0.18) or TKEQ (1.10 ± 0.16). As a result, the curvature itself does not explain the lower performance of KTKE.

3.5 Finding potential off-pathway targets

It is desirable to avoid off-pathway targets to minimize unwanted side-effects. To find the prevalence of unwanted targets in the proteome for a given epitope, we scrape through the RCSB database of resolved protein structures, to find proteins that might be potential off-pathway targets of antibodies that could be generated by each cyclic peptide scaffold (see Methods Section 2.7).

To illustrate the procedure, we demonstrate the off-pathway target analysis of (1,4)TKEQ scaffold here. Fig. 9a shows the scaffold ensemble for (1,4)TKEQ, projected on the first MDS coordinate, along with all entries of the human proteome having known PDB strutures and containing TKEQ motifs. The degree a PDB entry is embedded in the cyclic peptide ensemble is quantified as the embedding depth $\mathcal{D}_{\text{off-target-in-cyclic}}$, and is recorded as a percentage. Note that Fig. 9a is only for visualization; The formal calculation of embedding depth, $\mathcal{D}_{\text{off-target-in-cyclic}}$, is performed in 5 dimensions (5D). Nevertheless, we can see from this figure

that the epitope in the context of the various PDB structures is conformationally distinct from the epitope in the context of the cyclic peptide.

Fig. 9b indicates the distribution of the SASA for the (1,4)TKEQ ensemble, along with the SASA of TKEQ for all PDB entries in the human proteome with TKEQ motifs. The SASA of the scaffolded cyclic peptide ensemble ranges from 400Å^2 to 600Å^2, and serves as a good reference from which to compare. The fraction of the cyclic peptide ensemble that has lower SASA than a given PDB entry is defined as $f(\text{SASA}_{\text{off-target-exceed-cyclic}})$.

We apply cutoff thresholds of 5% for both $\mathcal{D}_{\text{off-target-in-cyclic}}$ and $f(\text{SASA}_{\text{off-target-exceed-cyclic}})$ to all PDB entries containing the epitope. In the case of (1,4)TKEQ, two PDB entries (1XQ8 (*48*) and 2KKW (*49*)) show *both* noticeable structural similarity ($\mathcal{D}_{\text{off-target-in-cyclic}} > 5\%$) and solvent exposure ($f(\text{SASA}_{\text{off-target-exceed-cyclic}}) > 5\%$) relative to the (1,4)TKEQ ensemble (Fig. 9c and (1,4) entry in Fig. S2c). Other off-pathway targets identified for all scaffolds can be found in Fig. S2. Many off-pathway targets, including 1XQ8 and 2KKW above, are themselves deposited structures of monomeric α-synuclein. Fibril PDB entries of α-synuclein are excluded from the off-pathway target criterion calculation to avoid double counting the contribution of fibril to the ranking. Of the many α-synuclein fibril structures in the PDB, the only one identified as an off-pathway target by the above cutoff criteria was 2N0A, which happens to be the PDB entry we used to generate the fibril ensemble and perform epitope prediction. On the other hand, the structured monomer entries still contribute to the off-pathway target criterion because they are distinct from the simulated monomer ensemble. These monomer PDB entries are partly structured (containing α-helices) by binding to micelles (e.g. PDB 2KKW and 1XQ8), and are thought to be involved in various aspects of α-synuclein physiology (*48–53*). Thus, membrane-bound PDB monomer structures are treated separately from the unstructured isolated monomer ensemble.

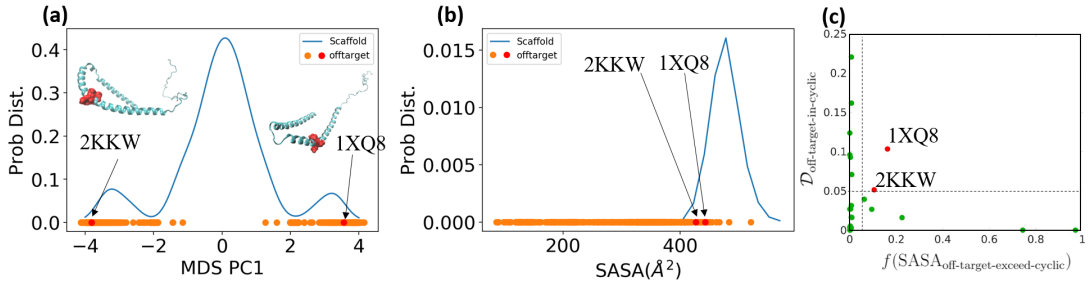


Figure 9: Off-pathway target analysis for (1,4)TKEQ. (a) Structural ensemble distribution of cyclic peptide (1,4)TKEQ in 1D along the first MDS component of the ensemble, along with the projected embedding of potential off-pathway targets. Most of the off-pathway targets are located at the periphery of the scaffold distribution. The actual calculation is performed in 5D. Structures of PDB entries 2KKW and 1XQ8 are rendered in ribbon schematics, and the epitope is rendered in red Van der Waals surface. (b) The SASA distribution of (1,4)TKEQ, along with the SASA for the off-pathway target structures. (c) Only 1XQ8 and 2KKW show both noticeable structural similarity ($\mathcal{D}_{\text{off-target-in-cyclic}}(5D) > 5\%$) and SASA exposure ($f(\text{SASA}_{\text{off-target-exceed-cyclic}}) > 5\%$)

4 Discussion and Conclusion

Our problem in this work was to find a cyclic peptide scaffold(s) of an epitope that best resembled the epitope in the context of an oligomer, as modelled by a stressed, partially disordered protofibril. These cyclic peptides could then be used as immunogens to raise oligomer-selective antibodies. This involved extensive conformational analysis and a ranking procedure with weighted criteria based on this analysis.

Assigning appropriate weights for each criterion is a non-trivial task, which depends on biological intuition, and analyzing the consequences arising from screening criteria with a given set of weights. For this specific ranking, we put twice the weight on dissimilarity from the fibril and the monomer than similarity to the stressed fibril. We chose this weighting scheme because the stressed fibril ensemble generation process contains more prior assumptions, in which we hypothesize that a partially disordered fibril ensemble is enriched in oligomer-selective conformational epitopes.

The fibril ensemble is obtained by generating the equilibrium ensemble of the solid-

state NMR resolved structure (PDB entry 2N0A). The monomer ensemble is obtained by generating disordered conformations of an isolated monomer of α-synuclein. The confidence in the accuracy of the fibril ensemble and monomer ensemble are greater than that of the stressed fibril oligomer model ensemble. The off-pathway target criteria have a slightly higher weight than the similarity of stressed fibril, because minimal off-pathway target effects are desired and are key feature of monoclonal antibody therapy.

The *in silico* screening method introduced in this paper can generally be applied to various different scaffolding methods. There are many other scaffolding methods that have been previously developed (*54–57*) besides cyclizing the epitope with a variable number of glycine residues. Apart from these methods, several possible extensions of our cyclic peptide scaffolding method may be directly implemented. For example, proline can be used for scaffolding instead of glycine, since it is relatively inert chemically compared to other amino acids, and it is also able to constrain the cyclic peptide ensemble by adding conformational rigidity (*58*), which may be exploited to bias the scaffold ensemble more towards the stressed fibril ensemble.

As another extension of the method developed here, multiple cyclic peptides can be stacked along an axis normal to the macrocycle. Such a structure can be stabilized either by non-covalent sidechain interactions, or by cysteine disulfide bonds to make multi-ring designs. Such an assembled structure has the potential to better mimic a stressed fibril ensemble of the epitope via these inter-macrocycle interactions. Similar designs have previously been proposed for the functional reconstruction of structurally complex epitopes (*59*).

Different carrier proteins that conjugate to the scaffolded epitope such as nanoparticles or lipid micelles have also been successfully implemented previously (*60–63*). Similarly, a linear peptide epitope of CGTKEQGGGG conjugated to a much larger carrier can be simulated as a peptide with cysteine SG atom spatially fixed at $x = 0$, and a boundary constraint such that it is forbidden to access the half-space region $x < 0$. Such a peptide in fact exhibits desirable ensemble overlap properties ($\mathcal{D}_{\text{ensemble-in-monomer}}$=0.142,

$\mathcal{D}_{\text{ensemble-in-stress}}$=0.207, $\mathcal{D}_{\text{ensemble-in-fibril}}$=0.000). Thus, conjugating on a non-interacting surface might be sufficient to scaffold the epitope to mimic the model oligomer. The above possible extension strategies are interesting future tests for computational and experimental studies.

The *in silico* immunogen screening method developed here provides an additional route to optimize a peptide immunogen that can be faster and cheaper than experiments. Experimental screening methods can then be performed on a subset of top-ranked scaffolds to save available resources. Some examples of screening experiments are surface plasmon resonance (SPR) assays to measure binding affinity to antibodies and conformational selectivity (*16*), Thioflavin T (ThT) assays (*14*, *64*, *65*), and Förster resonance energy transfer (FRET) assays (*66*) to measure aggregation tendency, since scaffolds that can trigger aggregation of normal α-synuclein may have more contribution to the pathology because of greater seeding propensity.

The aggregation of α-synuclein is critical to the pathogenesis of Parkinson's disease and other synucleinopathies (*23*, *67–74*). The SNCA gene encodes four isoforms of α-synuclein: SNCA140, SNCA126, SNCA112 and SNCA98 (*75*). The four isoforms have been shown to have varying degrees of aggregation, with the SNCA140 isoform being dominant and having a significantly higher levels of fibril formation (*75*). The fibril structure (PDB ID: 2N0A) used in this paper is of the SNCA140 isoform with 140 amino acid residues (*23*). Subsequent cryo-electron microscopy (cryo-EM) structures of the SNCA140 α-synuclein isoform shows that it exhibits two predominant fibril polymorphs; one rod and the other twister with a common protofilament structural kernel (*76*). Thus, the two polymorphs differ not in the protofilament unit but in the location of the protofilament packing interface. The different polymorphic structures formed by α-synuclein due to different packing of similar kernel has also been observed in other aggregation prone proteins like Aβ (*1*, *77*) and tau (*1*, *78*).

There are several alternative α-synuclein fibril structures with different polymorphs deposited in RCSB protein data bank (PDB) (`www.rcsb.org`) that we could have used for this

study (e.g. PDB entries 6CU7, 6CU8, 6H6B and 6FLT), which raises the question of whether the same epitopes would have been predicted with these alternative structures. There is a precedent for some epitope commonality among polymorphic Aβ fibril structures (*79*). The single protofilament solid-state NMR structure of the α-synuclein fibril (PDB ID: 2N0A) used here for epitope prediction is similar to both the dominant rod (PDB ID: 6CU7) and twister (PDB ID: 6CU8) polymorphs for 38 matched residues of the α-synuclein fibril structures, with an RMSD of 3.5Å and 3.8Å, respectively (*76*). Another cryo-EM structure of a truncated α-synuclein fibril, which includes the first 121 residues, has also been shown to have structural similarity to the rod polymorph with an RMSD of 2.1Å (*76*, *77*). These similarities between the aggregating units of different polymorphs of α-synuclein suggests that there may be some commonalities in collective coordinate-predicted epitopes. That said, the fact that these other fibril structures did not appear as strong hits in our off-pathway analysis indicates that if there was overlap in the exposed epitopes, the conformational ensemble would not be well-modelled by any of the cyclic peptides we investigated. Systematic analysis explicitly analyzing the predictions for these structures would have to be done to validate this hypothesis, and is a topic for future work.

5 Author Information

5.1 Author Contributions

Conceptualization, S.S.P.; Methodology, S.H., X.P., A.A. and S.S.P.; Software, X.P., S.H., and A.A.; Formal analysis, S.H., S.S.P., X.P., and A.A.; Resources, S.S.P and N.R.C.; Writing—original draft preparation, S.H., A.A., and S.S.P.; Writing—review and editing, S.S.P.; Visualization, S.H., X.P., A.A., and S.S.P.; Supervision, S.S.P.; Project administration, S.S.P.; Funding acquisition, S.S.P.

5.2 Funding Sources

This research was funded by the Canadian Institute of Health Research Transitional Operating Grant 2682, and by Alberta Innovates Research Team Program Grant PTM13007, and Compute Canada Resources for Research Groups RRG 3071.

5.3 Conflicts of Interest

S.S.P. was Chief Physics Officer of ProMIS Neurosciences until October 2020. N.R.C. is Chief Scientific Officer of ProMIS Neurosciences. S.S.P., N.R.C., and X.P. are the inventors on patent application PCT/CA2019/051434 (applicant being University of British Columbia). The patent application describes immunogens, antibodies and methods of their making as well as their use. Patent applications owned by the University of British Columbia are licensed to ProMIS Neurosciences. The work presented was financially supported in part by ProMIS Neurosciences. S.H., X.P., A.A., N.R.C. and S.S.P. have received consultation compensation from ProMIS Neurosciences.

6 Abbreviations

Aβ	Amyloid-β
KLH	Keyhole limpet hemocyanin
AD	Alzheimer 's disease
PD	Parkinson's disease
CTE	Chronic traumatic encephalopathy
ALS	Amyotropic lateral sclerosis
MD	Molecular dynamics
JSD	Jensen Shannon Divergence
SASA	Solvent accessible surface area
PDB	Protein data bank
RMSF	Root mean squared fluctuations
ThT	Thioflavin T
FRET	Förster resonance energy transfer
RMSD	Root mean squared deviation
CC	Collective coordinate
CG	Coarse grained
BSA	Bovine serum albumin
KDE	Kernel density estimation
MDS	Multidimensional scaling
SPE	Stochastic proximity embedding
D_{KL}	Kullback–Leibler divergence
SOD1	Superoxide dismutase 1
Cryo-EM	Cryo-electron microscopy
NMR	Nuclear magnetic resonance

Supporting Information Available

Supporting Information: Ranking criteria and rankings for α-synuclein epitope scaffolds; Convergence of JSD and embedding depth; Off-pathway targets for each epitope

Shawn C.C. Hsueh, Adekunle Aina, Neil R. Cashman, Xubiao Peng, Steven S. Plotkin

S.1 The ranking criteria and rankings for α-synuclein epitope scaffolds

Table S1: Ranking criteria and rankings for α-synuclein epitope scaffolds. The abbreviations JSD_{cf}, JSD_{cm}, JSD_{cs}, $\mathcal{D}_{c|m}$, and $\mathcal{D}_{c|s}$ correspond to $\mathrm{JSD}_{\text{cyclic-fibril}}$, $\mathrm{JSD}_{\text{cyclic-monomer}}$, $\mathrm{JSD}_{\text{cyclic-stress}}$, $\mathcal{D}_{\text{cyclic-in-monomer}}$, and $\mathcal{D}_{\text{cyclic-in-stress}}$, respectively. OP corresponds to off-pathway targeting.

Rank	**scaffold**	**scaffold symbol**	$\mathbf{JSD_{cf}}$	$\mathbf{JSD_{cm}}$	$\mathbf{JSD_{cs}}$	$\mathcal{D}_{\mathbf{c\|f}}(\times 10^{-5})$	$\mathcal{D}_{\mathbf{c\|m}}$	$\mathcal{D}_{\mathbf{c\|s}}$	**OP**
0	I_{best}	I_{best}	1.00000	1.000	0.000	0	0.000	1.000	1.000
1	CGEKTKGGG	(1,3)EKTK	0.99979	0.897	0.751	0	0.052	0.107	1.000
2	CGGEKTKGGG	(2,3)EKTK	0.99978	0.849	0.765	0	0.093	0.147	0.979
3	CGGGEKTKGG	(3,2)EKTK	0.99979	0.850	0.768	0	0.104	0.124	0.959
4	CGEKTKGG	(1,2)EKTK	0.99979	0.980	0.916	0	0.009	0.058	0.986
5	CGGTKEQGGG	(2,3)TKEQ	0.99833	0.869	0.740	12	0.128	0.177	0.893
6	CGGEKTKGGGG	(2,4)EKTK	0.99978	0.822	0.700	0	0.094	0.160	0.892
7	CGGGEKTKGGGG	(3,4)EKTK	0.99955	0.686	0.675	0	0.235	0.225	0.971
8	CGEKTKG	(1,1)EKTK	0.99979	0.961	0.951	0	0.011	0.019	1.000
9	CGGGGEKTKGG	(4,2)EKTK	0.99977	0.804	0.786	0	0.140	0.135	0.952
10	CGGEKTKGG	(2,2)EKTK	0.99967	0.886	0.875	0	0.082	0.061	1.000
11	CGGTKEQGGGG	(2,4)TKEQ	0.99974	0.905	0.871	0	0.144	0.084	1.000
12	CGEKTKGGGG	(1,4)EKTK	0.99974	0.889	0.917	0	0.070	0.061	1.000
13	CGGTKEQGG	(2,2)TKEQ	0.99979	0.975	0.976	0	0.058	0.038	1.000
14	CGGGEKTKGGG	(3,3)EKTK	0.99979	0.737	0.742	0	0.170	0.142	0.958
15	CGGKTKEG	(2,1)KTKE	0.99979	0.996	0.999	0	0.010	0.000	1.000
16	CGGGTKEQGG	(3,2)TKEQ	0.99951	0.904	0.903	0	0.122	0.049	0.995
17	CGGGTKEQGGG	(3,3)TKEQ	0.99974	0.911	0.877	0	0.193	0.076	0.991
18	CGGEKTKG	(2,1)EKTK	0.99979	0.948	0.962	0	0.058	0.022	0.987
19	CGGTKEQG	(2,1)TKEQ	0.99979	0.995	1.000	0	0.015	0.001	0.991
20	CGKTKEGG	(1,2)KTKE	0.99979	0.976	0.998	0	0.018	0.001	1.000
21	CGTKEQGGGG	(1,4)TKEQ	0.99979	0.972	0.943	0	0.037	0.044	0.917
22	CGKTKEG	(1,1)KTKE	0.99978	0.966	0.999	0	0.030	0.000	1.000
23	CGTKEQG	(1,1)TKEQ	0.99979	0.967	0.995	0	0.084	0.011	1.000
24	CGGGKTKEGG	(3,2)KTKE	0.99979	0.954	0.995	0	0.034	0.002	0.984
25	CGTKEQGG	(1,2)TKEQ	0.99979	0.951	0.971	0	0.150	0.019	1.000
26	CGGGGEKTKGGG	(4,3)EKTK	0.99973	0.723	0.661	0	0.187	0.200	0.811
27	CGGGEKTKG	(3,1)EKTK	0.99979	0.964	0.991	0	0.043	0.004	0.910
28	CGGGGTKEQGGG	(4,3)TKEQ	0.99925	0.902	0.834	1	0.099	0.084	0.801
29	CGGGGEKTKGGGG	(4,4)EKTK	0.99931	0.509	0.649	0	0.452	0.331	1.000
30	CGGGGTKEQG	(4,1)TKEQ	0.99979	0.934	0.994	0	0.184	0.005	0.973
31	CGKTKEGGG	(1,3)KTKE	0.99956	0.917	0.997	7	0.142	0.000	0.944
32	CGGGTKEQG	(3,1)TKEQ	0.99979	0.925	0.991	0	0.100	0.007	0.896
33	CGGGGKTKEG	(4,1)KTKE	0.99969	0.816	0.962	0	0.196	0.013	0.987
34	CGGGGEKTKG	(4,1)EKTK	0.99979	0.879	0.919	0	0.091	0.035	0.834
35	CGGGKTKEGGG	(3,3)KTKE	0.99922	0.809	0.953	5	0.197	0.012	0.940
36	CGGGGKTKEGGG	(4,3)KTKE	0.99599	0.725	0.881	35	0.257	0.047	0.932
37	CGGKTKEGG	(2,2)KTKE	0.99873	0.859	0.942	13	0.349	0.056	0.933
38	CGTKEQGGG	(1,3)TKEQ	0.99978	0.922	0.936	0	0.070	0.054	0.732
39	CGGGKTKEG	(3,1)KTKE	0.99974	0.991	0.999	0	0.097	0.000	0.705
40	CGGKTKEGGGG	(2,4)KTKE	0.99791	0.708	0.947	6	0.312	0.017	0.884
41	CGKTKEGGGG	(1,4)KTKE	0.99929	0.757	0.969	0	0.239	0.004	0.777
42	CGGGKTKEGGGG	(3,4)KTKE	0.99958	0.706	0.901	0	0.548	0.089	0.963
43	CGGGGTKEQGGGG	(4,4)TKEQ	0.99861	0.840	0.726	2	0.142	0.144	0.519
44	CGGKTKEGGG	(2,3)KTKE	0.99966	0.822	0.976	0	0.091	0.004	0.591
45	CGGGGKTKEGG	(4,2)KTKE	0.99420	0.567	0.788	138	0.448	0.142	0.547
46	CGGGGTKEQGG	(4,2)TKEQ	0.99979	0.958	0.940	0	0.084	0.032	0.281
47	CGGGGKTKEGGGG	(4,4)KTKE	0.99445	0.586	0.829	90	0.267	0.045	0.415
48	CGGGTKEQGGGG	(3,4)TKEQ	0.99979	0.903	0.972	0	0.221	0.009	0.000
49	I_{worst}	I_{worst}	0.00000	0.000	1.000	10^5	1.000	0.000	0.000

S.2 The convergence of JSD and embedding depth

Here, we show the convergence of the dimension reduction performance on the pairwise RMSD matrix that describes fibril, stressed fibril, monomer and (1,4)TKEQ ensemble. Stochastic Proximity Embedding (SPE) (*34*) is performed to calculate JSD in reduced dimension. The different degrees of convergence in 3-11 dimensions (See Supplemental Fig. S1a) motivated us to use a weighted average JSD across several dimensions, by weighting the JSD in each dimension by the inverse of the residual in that dimension. In practice this weights higher dimensions more than lower dimensions.

Multidimensional scaling (MDS) is performed to calculate embedding depth in reduced dimension. Since the residual quickly converged in dimensions of 3 or larger, (See Supplemental Fig. S1b), we simply perform the embedding depth measure in 3D.

The residual of the MDS and SPE are defined respectively as follows:

$$\text{Residual}_{MDS} = \frac{\sum_{i<j} (d_{ij} - r_{ij})^2}{\sum_{i<j} r_{ij}^2} \tag{S1}$$

$$\text{Residual}_{SPE} = \sum_{i<j} \frac{(d_{ij} - r_{ij})^2}{r_{ij}} / \sum_{i<j} r_{ij} \tag{S2}$$

In equations (S1) and (S2), index i and j correspond to two different rows in a pairwise RMSD matrix including the fibril ensemble, the stressed fibril ensemble, the monomer ensemble, and a given cyclic peptide scaffold ensemble. Before any dimensional reduction, the RMSD matrix is an $N \times N$ matrix, where N is the sum of the number of configurations of the above four ensembles, and the indices i and j run from 1 to N. d_{ij} and r_{ij} are the Euclidean distance between i^{th} and j^{th} row of the RMSD matrix in reduced dimension and in the original dimension, respectively.

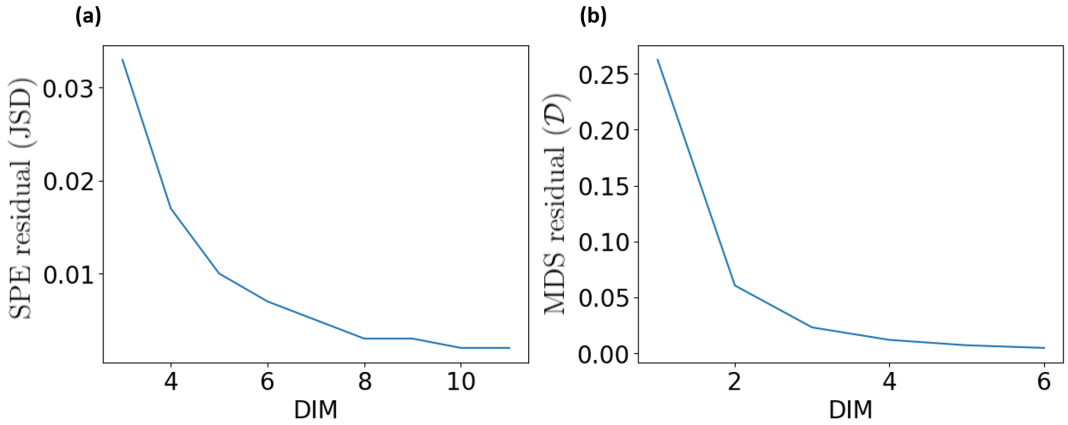


Figure S1: Convergence of dimensional reduction for (a) stochastic proximity embedding (SPE) for JSDs and (b) multidimensional scaling (MDS) for embedding depth. Both quantities are plotted vs. the dimension of the reduction. The scaffold used here for illustration is (1,4)TKEQ.

S.3 The off-pathway target for each epitope

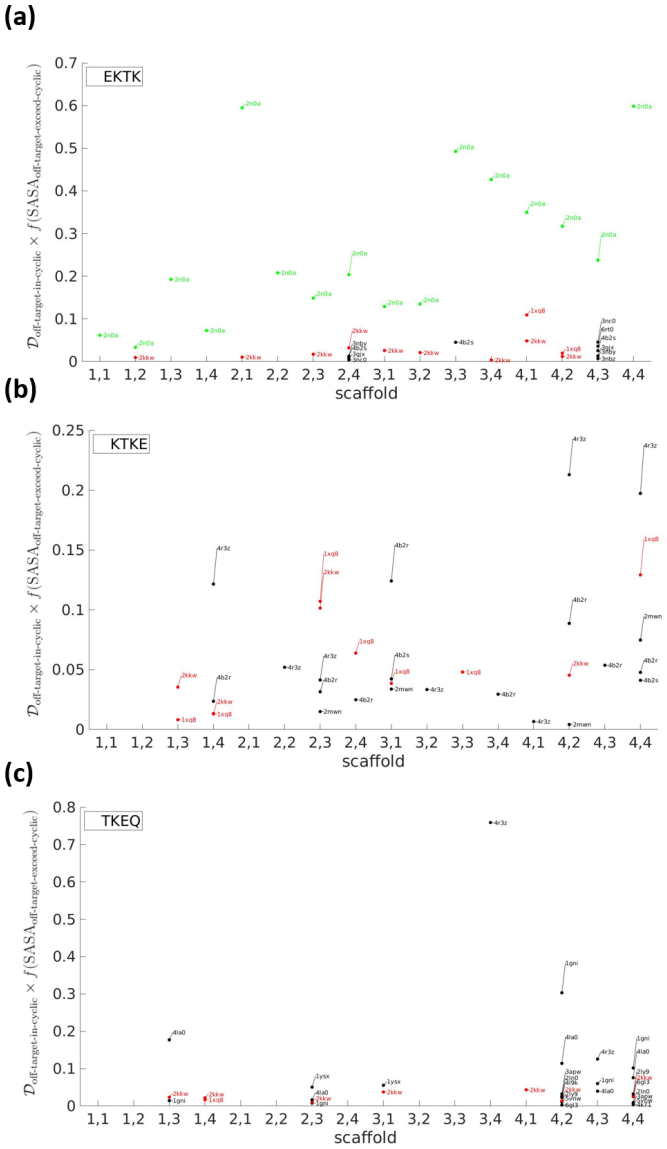


Figure S2: Off-pathway targets found in the RCSB, for each 4 amino acid epitope sequence and corresponding scaffold. E.g. 1,1 on the x-axis of the top panel is (1,1)EKTK. An off-pathway target is identified if both its $\mathcal{D}_{\text{off-target-in-cyclic}}$ and $f(\text{SASA}_{\text{off-target-exceed-cyclic}})$ are greater than 0.05, see e.g. Fig. 9c. For such PDB entries, the value of each point on the y-axis is given by the product of $\mathcal{D}_{\text{off-target-in-cyclic}} \times f(\text{SASA}_{\text{off-target-exceed-cyclic}})$ for that off-pathway target, plotted here as a decimal value. The summation of values for points in the vertical column corresponding to each given scaffold is taken here to be the severeness of the corresponding scaffold; This value is used to rank each scaffold. Results are presented for (a) EKTK (b) KTKE and (c) TKEQ epitopes. Entries colored green stand for fibrillar α-synuclein (PDB 2N0A), while entries colored red are membrane-bound α-synuclein monomers in the PDB, which typically contain α-helical structure. The fibril α-synuclein entries are specifically excluded from the off-pathway calculation because they have been treated already using JSD and embedding depth $\mathcal{D}$, while the structured monomers are retained in the off-pathway calculation.

References

1. Plotkin, S. S., and Cashman, N. R. (2020) Passive immunotherapies targeting A-beta and tau in Alzheimer's disease. *Neurobiology of Disease* 105010.

2. Chiti, F., and Dobson, C. M. (2017) Protein misfolding, amyloid formation, and human disease: a summary of progress over the last decade. *Annual review of biochemistry 86*, 27–68.

3. Uversky, V. N., Oldfield, C. J., and Dunker, A. K. (2005) Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling. *Journal of Molecular Recognition 18*, 343–384.

4. McAlary, L., Plotkin, S. S., and Cashman, N. R. (2019) Emerging developments in targeting proteotoxicity in neurodegenerative diseases. *CNS drugs* 1–22.

5. Cashman, N. R., and Caughey, B. (2004) Prion diseases—close to effective therapy? *Nature Reviews Drug Discovery 3*, 874–884.

6. Rasmussen, J., Jucker, M., and Walker, L. C. (2017) Aβ seeds and prions: how close the fit? *Prion 11*, 215–225.

7. Watts, J. C., and Prusiner, S. B. (2018) β-amyloid prions and the pathobiology of Alzheimer's disease. *Cold Spring Harbor perspectives in medicine 8*, a023507.

8. Clavaguera, F., Hench, J., Goedert, M., and Tolnay, M. (2015) Invited review: Prion-like transmission and spreading of tau pathology. *Neuropathology and applied neurobiology 41*, 47–58.

9. Dujardin, S., and Hyman, B. T. (2019) Tau prion-like propagation: state of the art and current challenges. *Tau Biology* 305–325.

10. Goedert, M., Jakes, R., and Spillantini, M. G. (2017) The synucleinopathies: twenty years on. *Journal of Parkinson's disease 7*, S51–S69.

11. Zorzi, A., Deyle, K., and Heinis, C. (2017) Cyclic peptide therapeutics: past, present and future. *Current opinion in chemical biology 38*, 24–29.

12. Mulligan, V. K. (2020) The emerging role of computational design in peptide macrocycle drug discovery. *Expert Opinion on Drug Discovery* 1–19.

13. Silverman, J. M. et al. (2018) A Rational Structured Epitope Defines a Distinct Subclass of Toxic Amyloid-beta Oligomers. *ACS Chemical Neuroscience 9*, 1591–1606.

14. Gibbs, E., Silverman, J. M., Zhao, B., Peng, X., Wang, J., Wellington, C. L., Mackenzie, I. R., Plotkin, S. S., Kaplan, J. M., and Cashman, N. R. (2019) A rationally designed humanized antibody selective for amyloid beta oligomers in Alzheimer's disease. *Scientific reports 9*, 1–14.

15. Peng, X., Cashman, N. R., and Plotkin, S. S. (2018) Prediction of Misfolding-Specific Epitopes in SOD1 Using Collective Coordinates. *The Journal of Physical Chemistry B 122*, 11662–11676.

16. Peng, X., Gibbs, E., Silverman, J. M., Cashman, N. R., and Plotkin, S. S. (2021) A method for systematically ranking therapeutic drug candidates using multiple uncertain screening criteria. *Statistical Methods in Medical Research* https://doi.org/10.1177/09622802211002861.

17. Krivov, G. G., Shapovalov, M. V., and Dunbrack Jr, R. L. (2009) Improved prediction of protein side-chain conformations with SCWRL4. *Proteins: Structure, Function, and Bioinformatics 77*, 778–795.

18. Masucci, M. G. Glycine-containing sequences conferring invisibility to the immune system. 1998; US Patent 5,833,991.

19. Khrustalev, V. V., and Barkovsky, E. V. (2011) Percent of highly immunogenic amino

acid residues forming B-cell epitopes is higher in homologous proteins encoded by GC-rich genes. *Journal of theoretical biology 282*, 71–79.

20. Van Der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A. E., and Berendsen, H. J. (2005) GROMACS: fast, flexible, and free. *Journal of computational chemistry 26*, 1701–1718.

21. Bonomi, M., Branduardi, D., Bussi, G., Camilloni, C., Provasi, D., Raiteri, P., Donadio, D., Marinelli, F., Pietrucci, F., Broglia, R. A., and Parrinello, M. (2009) PLUMED: A portable plugin for free-energy calculations with molecular dynamics. *Computer Physics Communications 180*, 1961–1972.

22. Huang, J., Rauscher, S., Nawrocki, G., Ran, T., Feig, M., Groot, B. L. d., Grubmüller, H., and Jr, A. D. M. (2017) CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nature Methods 14*, 71–73.

23. Tuttle, M. D. et al. (2016) Solid-state NMR structure of a pathogenic fibril of full-length human α-synuclein. *Nature structural & molecular biology 23*, 409–415.

24. Uversky, V. N. (2003) A protein-chameleon: conformational plasticity of α-synuclein, a disordered protein involved in neurodegenerative disorders. *Journal of Biomolecular Structure and Dynamics 21*, 211–234.

25. Jain, K., Ghribi, O., and Delhommelle, J. (2021) Folding Free-Energy Landscape of α-Synuclein (35–97) Via Replica Exchange Molecular Dynamics. *Journal of Chemical Information and Modeling 61*, 432–443.

26. Das, A., Sin, B. K., Mohazab, A. R., and Plotkin, S. S. (2013) Unfolded protein ensembles, folding trajectories, and refolding rate prediction. *J. Chem. Phys. 139*, 121925.

27. Favrin, G., Irbäck, A., and Sjunnesson, F. (2001) Monte Carlo update for chain

molecules: biased Gaussian steps in torsional space. *The Journal of Chemical Physics 114*, 8154–8158.

28. Madras, N., and Sokal, A. D. (1988) The pivot algorithm: A highly efficient Monte Carlo method for the self-avoiding walk. *Journal of Statistical Physics 50*, 109–186.

29. Hadizadeh, S., Linhananta, A., and Plotkin, S. S. (2011) Improved measures for the shape of a disordered polymer to test a mean-field theory of collapse. *Macromolecules 44*, 6182–6197.

30. Wang, X., Vitalis, A., Wyczalkowski, M. A., and Pappu, R. V. (2006) Characterizing the conformational ensemble of monomeric polyglutamine. *Proteins: Structure, Function, and Bioinformatics 63*, 297–311.

31. Mohazab, A. R., and Plotkin, S. S. (2013) Polymer uncrossing and knotting in protein folding, and their role in minimal folding pathways. *PloS one 8*, e53642.

32. Bhattacharya, S., and Lin, X. (2019) Recent advances in computational protocols addressing intrinsically disordered proteins. *Biomolecules 9*, 146.

33. Cox, M. A., and Cox, T. F. (2008) Handbook of data visualization - Multidimensional scaling. *Springer* 315–347.

34. Tiberti, M., Papaleo, E., Bengtsen, T., Boomsma, W., and Lindorff-Larsen, K. (2015) ENCORE: Software for Quantitative Ensemble Comparison. *PLOS Computational Biology 11*, 1–16.

35. Bellman, R. (1966) Dynamic programming. *Science 153*, 34–37.

36. Liu, R. Y., and Singh, K. (1993) A quality index based on data depth and multivariate rank tests. *Journal of the American Statistical Association 88*, 252–260.

37. Sibson, R. (1969) Information radius. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 14*, 149–160.

38. Lin, J. (1991) Divergence measures based on the Shannon entropy. *IEEE Transactions on Information theory 37*, 145–151.

39. Lindorff-Larsen, K., and Ferkinghoff-Borg, J. (2009) Similarity measures for protein ensembles. *PloS one 4*, e4203.

40. Kullback, S., and Leibler, R. A. (1951) On information and sufficiency. *The annals of mathematical statistics 22*, 79–86.

41. Cover, T., and Thomas, J. (1991) Elements ofInfonnaiion The ory. *New York: Wiley*

42. Hwang, C.-L., and Masud, A. S. M. (2012) Multiple objective decision making—methods and applications: a state-of-the-art survey. *Springer Science & Business Media 164*.

43. Cashman, N. R., Plotkin, S. S., Peng, X., and Kaplan, J. Conformation-specific epitopes in alpha-synuclein, antibodies thereto and methods related thereof. 2020; Patent WO/2020/073121.

44. Chernodub, M., Hu, S., and Niemi, A. J. (2010) Topological solitons and folded proteins. *Physical Review E 82*, 011916.

45. Molkenthin, N., Hu, S., and Niemi, A. J. (2011) Discrete nonlinear Schrödinger equation and polygonal solitons with applications to collapsed proteins. *Physical Review Letters 106*, 078102.

46. Krokhotin, A., Niemi, A. J., and Peng, X. (2012) Soliton concepts and protein structure. *Physical Review E 85*, 031906.

47. Hu, S., Lundgren, M., and Niemi, A. J. (2011) Discrete Frenet frame, inflection point solitons, and curve visualization with applications to folded proteins. *Physical Review E 83*, 061908.

48. Ulmer, T. S., Bax, A., Cole, N. B., and Nussbaum, R. L. (2005) Structure and dynamics of micelle-bound human α-synuclein. *Journal of Biological Chemistry 280*, 9595–9603.

49. Rao, J. N., Jao, C. C., Hegde, B. G., Langen, R., and Ulmer, T. S. (2010) A combinatorial NMR and EPR approach for evaluating the structural ensemble of partially folded proteins. *Journal of the American Chemical Society 132*, 8657–8668.

50. Jensen, P. H., Nielsen, M. S., Jakes, R., Dotti, C. G., and Goedert, M. (1998) Binding of α-synuclein to brain vesicles is abolished by familial Parkinson's disease mutation. *Journal of Biological Chemistry 273*, 26292–26294.

51. Kahle, P. J., Haass, C., Kretzschmar, H. A., and Neumann, M. (2002) Structure/function of α-synuclein in health and disease: rational development of animal models for Parkinson's and related diseases. *Journal of neurochemistry 82*, 449–457.

52. Vekrellis, K., Rideout, H. J., and Stefanis, L. (2004) Neurobiology of α-synuclein. *Molecular neurobiology 30*, 1–21.

53. Amos, S.-B. T. A., Schwarz, T. C., Shi, J., Cossins, B. P., Baker, T. S., Taylor, R. J., Konrat, R., and Sansom, M. S. P. (2021) Membrane Interactions of α-Synuclein Revealed by Multiscale Molecular Dynamics Simulations, Markov State Models, and NMR. *The Journal of Physical Chemistry B 125*, 2929–2941.

54. Bhardwaj, G. et al. (2016) Accurate de novo design of hyperstable constrained peptides. *Nature 538*, 329–335.

55. Luther, A., Moehle, K., Chevalier, E., Dale, G., and Obrecht, D. (2017) Protein epitope mimetic macrocycles as biopharmaceuticals. *Current Opinion in Chemical Biology 38*, 45–51.

56. Wang, C. K., and Craik, D. J. (2018) Designing macrocyclic disulfide-rich peptides for biotechnological applications. *Nature Chemical Biology 14*, 417–427.

57. Nimrod, G., Fischman, S., Austin, M., Herman, A., Keyes, F., Leiderman, O., Hargreaves, D., Strajbl, M., Breed, J., Klompus, S., Kevin, M., Jennifer, S., Andrew, B.,

Tristan, J. V., and Yanay, O. (2018) Computational design of epitope-specific functional antibodies. *Cell reports 25*, 2121–2131.

58. Rawi, R., Shen, C.-H., Kwong, P. D., and Chuang, G.-Y. (2018) CRISPro: An Automated Pipeline for Protein Conformation Stabilization by Proline. *Journal of chemical information and modeling 58*, 2189–2192.

59. Timmerman, P., Puijk, W., Boshuizen, R., Dijken, P. v., Slootstra, J., Beurskens, F., Parren, P., Huber, A., Bachmann, M., and Meloen, R. (2009) Functional reconstruction of structurally complex epitopes using CLIPS™ Technology. *The Open Vaccine Journal 2*, 56–67.

60. Adolfsson, O., Pihlgren, M., Toni, N., Varisco, Y., Buccarello, A. L., Antoniello, K., Lohmann, S., Piorkowska, K., Gafner, V., Atwal, J. K., et al. (2012) An effector-reduced anti-β-amyloid (Aβ) antibody with unique aβ binding properties promotes neuroprotection and glial engulfment of Aβ. *Journal of Neuroscience 32*, 9677–9689.

61. Muhs, A., Hickman, D. T., Pihlgren, M., Chuard, N., Giriens, V., Meerschman, C., Van der Auwera, I., van Leuven, F., Sugawara, M., Weingertner, M.-C., et al. (2007) Liposomal vaccines with conformation-specific amyloid peptide antigens define immune response and efficacy in APP transgenic mice. *Proceedings of the National Academy of Sciences 104*, 9810–9815.

62. Theunis, C., Crespo-Biel, N., Gafner, V., Pihlgren, M., López-Deber, M. P., Reis, P., Hickman, D. T., Adolfsson, O., Chuard, N., Ndao, D. M., et al. (2013) Efficacy and safety of a liposome-based vaccine against protein Tau, assessed in tau. P301L mice that model tauopathy. *PLoS One 8*, e72301.

63. Nicolau, C., Greferath, R., Balaban, T. S., Lazarte, J. E., and Hopkins, R. J. (2002) A liposome-based therapeutic vaccine against β-amyloid plaques on the pancreas of

transgenic NORBA mice. *Proceedings of the National Academy of Sciences 99*, 2332–2337.

64. LeVine III, H. (1999) Quantification of β-sheet amyloid fibril structures with thioflavin T. *Methods in enzymology 309*, 274–284.

65. Xue, C., Lin, T. Y., Chang, D., and Guo, Z. (2017) Thioflavin T as an amyloid dye: fibril quantification, optimal concentration and effect on aggregation. *Royal Society open science 4*, 160696.

66. Forster, T. (1946) Energiewanderung und fluoreszenz. *Naturwissenschaften 33*, 166–175.

67. Peelaerts, W., Bousset, L., Van der Perren, A., Moskalyuk, A., Pulizzi, R., Giugliano, M., Van den Haute, C., Melki, R., and Baekelandt, V. (2015) α-Synuclein strains cause distinct synucleinopathies after local and systemic administration. *Nature 522*, 340–344.

68. Chen, S. W. et al. (2015) Structural characterization of toxic oligomers that are kinetically trapped during α-synuclein fibril formation. *Proceedings of the National Academy of Sciences 112*, E1994–E2003.

69. Ingelsson, M. (2016) Alpha-synuclein oligomers—neurotoxic molecules in Parkinson's disease and other Lewy body disorders. *Frontiers in neuroscience 10*, 408.

70. Ono, K. (2017) The Oligomer Hypothesis in α-Synucleinopathy. *Neurochemical Research 42*, 3362–3371.

71. Peng, C., Gathagan, R. J., and Lee, V. M.-Y. (2018) Distinct α-Synuclein strains and implications for heterogeneity among α-Synucleinopathies. *Neurobiology of Disease 109*, 209–218.

72. Xu, L., Bhattacharya, S., and Thompson, D. (2018) The fold preference and thermodynamic stability of α-synuclein fibrils is encoded in the non-amyloid-β component region. *Physical Chemistry Chemical Physics 20*, 4502–4512.

73. Meade, R. M., Fairlie, D. P., and Mason, J. M. (2019) Alpha-synuclein structure and Parkinson's disease – lessons and emerging principles. *Molecular Neurodegeneration 14*, 29.

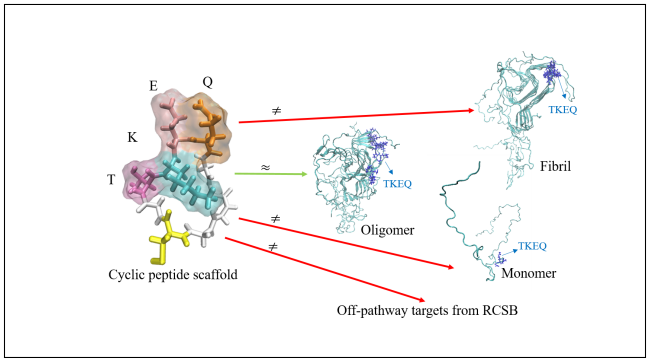
74. Schweighauser, M., Shi, Y., Tarutani, A., Kametani, F., Murzin, A. G., Ghetti, B., Matsubara, T., Tomita, T., Ando, T., Hasegawa, K., et al. (2020) Structures of α-synuclein filaments from multiple system atrophy. *Nature 585*, 464–469.

75. Bungeroth, M., Appenzeller, S., Regulin, A., Völker, W., Lorenzen, I., Grötzinger, J., Pendziwiat, M., and Kuhlenbäumer, G. (2014) Differential aggregation properties of alpha-synuclein isoforms. *Neurobiology of Aging 35*, 1913–1919.

76. Li, B., Ge, P., Murray, K. A., Sheth, P., Zhang, M., Nair, G., Sawaya, M. R., Shin, W. S., Boyer, D. R., Ye, S., Eisenberg, D. S., Zhou, Z. H., and Jiang, L. (2018) Cryo-EM of full-length α-synuclein reveals fibril polymorphs with a common structural kernel. *Nature Communications 9*, 3609.

77. Gremer, L., Schölzel, D., Schenk, C., Reinartz, E., Labahn, J., Ravelli, R. B. G., Tusche, M., Lopez-Iglesias, C., Hoyer, W., Heise, H., Willbold, D., and Schröder, G. F. (2017) Fibril structure of amyloid-β(1–42) by cryo–electron microscopy. *Science 358*, 116–119.

78. Fitzpatrick, A. W. P., Falcon, B., He, S., Murzin, A. G., Murshudov, G., Garringer, H. J., Crowther, R. A., Ghetti, B., Goedert, M., and Scheres, S. H. W. (2017) Cryo-EM structures of tau filaments from Alzheimer's disease. *Nature 547*, 185–190.

79. Cashman, N. R., Plotkin, S. S., and Guest, W. C. Methods and systems for predicting misfolded protein epitopes. 2010; US Patent App. 12/574,637.

Graphical TOC Entry