Energetic portrait of the amyloid beta nucleation transition state

Anna Arutyunyan*[1], Mireia Seuma*[2,3,7], Andre J. Faure[3,4,6], Benedetta Bolognesi†[,2], Ben Lehner†[,1,3,4,5]

1 Wellcome Sanger Institute, Cambridge, UK
2 Institute for Bioengineering of Catalonia (IBEC), The Barcelona Institute of Science and Technology (BIST) , Baldiri Reixac 10-12, 08028, Barcelona, Spain
3 Centre for Genomic Regulation (CRG), The Barcelona Institute for Science and Technology (BIST), Barcelona, Spain
4 Universitat Pompeu Fabra (UPF), Barcelona, Spain
5 Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain
6 Current address: ALLOX, C/ Dr. Aiguader, 88, PRBB Building, 08003 Barcelona, Spain
7 Current address: Medical Research Council Laboratory of Molecular Biology, Cambridge, UK

* Contributed equally
† e-mail: bbolognesi@ibecbarcelona.eu, bl11@sanger.ac.uk

Abstract

Amyloid protein aggregates are pathological hallmarks of more than fifty human diseases including the most common neurodegenerative disorders. The atomic structures of amyloid fibrils have now been determined, but the process by which soluble proteins nucleate to form amyloids remains poorly characterised and difficult to study, even though this is the key step to understand to prevent the formation and spread of aggregates. Here we use massively parallel combinatorial mutagenesis, a kinetic selection assay, and machine learning to reveal the transition state of the nucleation reaction of amyloid beta, the protein that aggregates in Alzheimer's disease. By quantifying the nucleation of >140,000 proteins we infer the changes in activation energy for all 798 amino acid substitutions in amyloid beta and the energetic couplings between >600 pairs of mutations. This unprecedented dataset provides the first comprehensive view of the energy landscape and the first large-scale measurement of energetic couplings for a protein transition state. The energy landscape reveals that the amyloid beta nucleation transition state contains a short structured C-terminal hydrophobic core with a subset of interactions similar to mature fibrils. This study demonstrates the feasibility of using mutation-selection-sequencing experiments to study transition states and identifies the key molecular species that initiates amyloid beta aggregation and, potentially, Alzheimer's disease.

Introduction

Amyloid fibrils are supramolecular fibrous protein assemblies in which β-strands are stacked along the long axis of each fibril in an ordered 'cross-β' structure[1]. Specific amyloid assemblies define most human neurodegenerative diseases, including Alzheimer's disease, Parkinson's disease and frontotemporal dementia. Fibrils of amyloid beta are, for example, a pathological hallmark of Alzheimer's disease and variants in amyloid beta cause familial Alzheimer's disease [2,3]. In total, at least 50 human disorders are associated with the formation of amyloid fibrils of more than 30 different proteins[4]. Beyond human pathology, amyloids are present in all kingdoms of life and possess functions in a wide variety of organisms, including humans[5].

Thanks primarily to developments in cryogenic electron microscopy (cryo-EM), the atomic structures of many amyloid fibrils have now been determined, including fibrils extracted from human brains. These structures have revealed that proteins can adopt different filament structures (fibril polymorphs) with, at least in some cases, different amyloid folds associated with different clinical conditions[6]. Time-resolved cryo-EM has been used to characterise the *in vitro* assembly of amyloids, revealing a diversity of folds that appear and disappear as fibrillation proceeds [7,8]. Mature amyloid fibrils are extremely stable and normally irreversible states that likely represent the thermodynamically favoured state of many proteins at high concentration[9]. The conversion of soluble proteins to amyloid fibrils occurs through nucleation-and-growth processes, with nucleation the rate-limiting step[10]. That most proteins never form amyloids under physiological conditions is due to kinetic control, with a very high free energy barrier to nucleation meaning fibrils never form on human timescales [11,12]. Nucleation reactions are therefore the key processes to understand and prevent in order to stop amyloid formation in human diseases.

The rate of a nucleation reaction depends on the difference in energy (the activation energy, E_a) between the highest energy state along the reaction coordinate (the transition state) and the initial soluble state (Fig. 1a). Due to their high energy and transient nature, transition states are notoriously difficult to characterise. One disruptive approach for probing transition states, pioneered by Fersht and colleagues, is to use mutations [13–15]. Mutations that stabilise a transition state will lower E_a and accelerate a reaction whereas mutations that destabilise the transition state will slow it. Kinetic measurements therefore allow the importance of individual residues and structural contacts in a transition state to be probed. Moreover, comparing changes in kinetics to changes in stability can provide information on the degree of native structure around a mutated residue in a transition state [15,16].

Small numbers of mutations have been used to study the transition states of enzymes and protein folding pathways [17–19], providing many important structural insights [13,14]. More recently ten mutations were used to probe the transition state of an amyloid elongation reaction, revealing that monomers that successfully bind to the fibril ends have fibril-like contacts [20]. Another study combined molecular dynamics with kinetic experiments to suggest a hairpin trimer structure for the transition state of Tau amyloids [21]. However, in all cases, due to the difficulty of making kinetic measurements, only a very small number of carefully chosen mutations have been characterised.

Here we use massively parallel mutagenesis and a kinetic selection assay to infer the change in nucleation reaction activation energy (E_a) for all possible mutations in an aggregating protein, amyloid beta. Quantifying the nucleation rates of >100,000 double mutants and >40,000 higher order combinatorial mutants allows us to infer the change in activation energy for all amino acid substitutions and the energetic couplings between >600 pairs of substitutions. This unprecedented in scale energy landscape reveals that the nucleation transition state is a short C-terminal hydrophobic region and identifies the key structural contacts that initiate amyloid beta nucleation and, potentially, Alzheimer's disease.

Results

Quantifying amyloid beta nucleation kinetics and activation energies at scale

The rate of an amyloid nucleation reaction depends exponentially on the activation energy (E_a) of the reaction, as described by the Arrhenius equation ($k = A * exp(-\frac{E_a}{RT})$, where k is the nucleation rate constant, A is a pre-exponential factor, T is the absolute temperature, and R is the universal gas constant, Fig. 1b,c) [22–24]. To quantify changes in k for all mutations in amyloid beta (Aβ42) we employed a kinetic selection assay in which the rate of Aβ42 nucleation controls the growth rate of yeast cells (Fig. 1a, Extended Data Fig. 1a). Comparison with *in vitro* nucleation rate constants from different studies shows that growth rates are proportional to the logarithm of nucleation rate constants (Extended Data Fig. 1b,c, Supplementary Table 1) [25–29]. However, the dynamic range of the assay is bounded by the maximum and minimum quantifiable growth rates, limiting measurement precision.

To expand the measurement range of the selection assay we employed double mutant libraries. Testing mutations in fast nucleating variants of Aβ42 provides an expanded dynamic range for quantifying decreased nucleation, whereas testing mutations in slow nucleating Aβ42 variants expands the dynamic range for quantifying increased nucleation. Using a large number of double mutants reduces the influence of specific interactions between mutations [30,31].

In total, we quantified the growth rates of 101,888 double mutants of Aβ42 using three separate libraries and triplicate selection experiments (Fig. 2a-c, Extended Data Fig. 2b Supplementary Table 2 and Methods). The effect of each amino acid (AA) substitution was quantified in a median of 215 Aβ42 backgrounds, providing highly reproducible measurements of changes in nucleation kinetics in Aβ42 proteins with many different rates of nucleation.

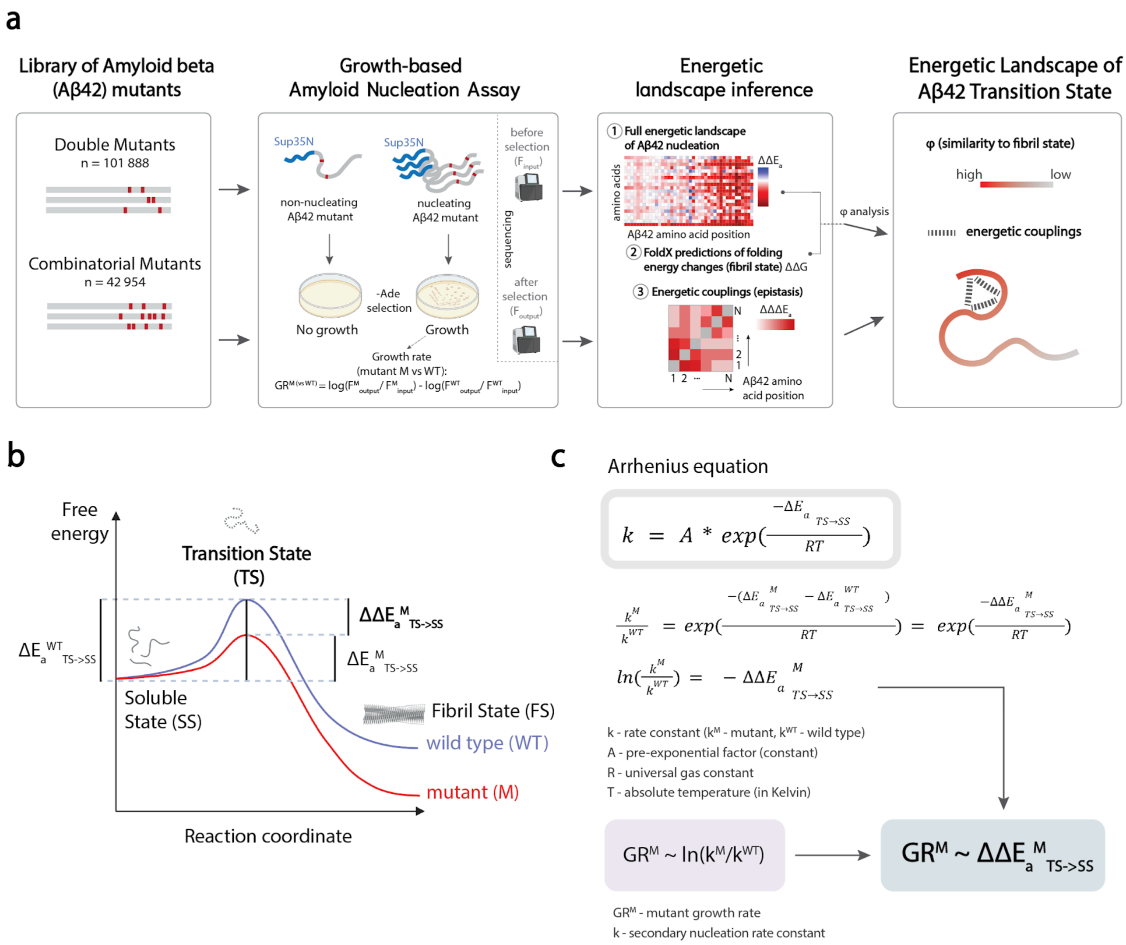

a
Library of Amyloid beta (Aβ42) mutants
Double Mutants
n = 101 888
Combinatorial Mutants
n = 42 954
Growth-based Amyloid Nucleation Assay
Sup35N
non-nucleating Aβ42 mutant
Sup35N
nucleating Aβ42 mutant
before selection (F_{input})
sequencing
-Ade selection
No growth
Growth
after selection (F_{output})
Growth rate (mutant M vs WT):
$GR^{M(vs WT)} = \log(F^M_{output}/F^M_{input}) - \log(F^{WT}_{output}/F^{WT}_{input})$
Energetic landscape inference
1 Full energetic landscape of Aβ42 nucleation
amino acids
Aβ42 amino acid position
$\Delta\Delta E_a$
2 FoldX predictions of folding energy changes (fibril state) ΔΔG
3 Energetic couplings (epistasis)
$\Delta\Delta\Delta E_a$
N
2
1
1 2 ··· N
Aβ42 amino acid position
φ analysis
Energetic Landscape of Aβ42 Transition State
φ (similarity to fibril state)
high
low
energetic couplings
b
Free energy
Transition State (TS)
$\Delta E_a{}^{WT}{}_{TS\text{->}SS}$
$\Delta\Delta E_a{}^{M}{}_{TS\text{->}SS}$
$\Delta E_a{}^{M}{}_{TS\text{->}SS}$
Soluble State (SS)
Fibril State (FS)
wild type (WT)
mutant (M)
Reaction coordinate
c
Arrhenius equation
$k = A * exp(\frac{-\Delta E_{a\,TS\to SS}}{RT})$
$\frac{k^M}{k^{WT}} = exp(\frac{-(\Delta E^M_{a\,TS\to SS} - \Delta E^{WT}_{a\,TS\to SS})}{RT}) = exp(\frac{-\Delta\Delta E^M_{a\,TS\to SS}}{RT})$
$ln(\frac{k^M}{k^{WT}}) = -\Delta\Delta E^M_{a\,TS\to SS}$
k - rate constant (k^M - mutant, k^{WT} - wild type)
A - pre-exponential factor (constant)
R - universal gas constant
T - absolute temperature (in Kelvin)
$GR^M \sim ln(k^M/k^{WT})$
$GR^M \sim \Delta\Delta E_a{}^{M}{}_{TS\text{->}SS}$
GR^M - mutant growth rate
k - secondary nucleation rate constant

Figure 1. Quantifying amyloid nucleation activation energies at scale. **a** Schematic overview of the study: mutant libraries are selected through an assay where cell growth depends on amyloid nucleation and is quantified by sequencing. The resulting growth rates are used to fit a mechanistic model to infer activation energy terms and energetic couplings which can be used to map the energetic landscape of the Aβ42 transition state. **b**, Free energy landscape of the amyloid nucleation reaction. **c**, Arrhenius equation and derivation of linear relationship between measured growth rates and change in activation energy; k - rate constant (k_M for mutant, k_{WT} for wild type), A - pre-exponential constant, ΔE_a- activation energy, R - universal gas constant, T - absolute temperature (Kelvin), TS - transition state, FS - fibril state, SS - soluble state, $\Delta\Delta E_a{}^M$- change in activation energy for mutant M (vs WT), GR^M - mutant growth rate, κ - secondary nucleation rate constant.

The transition state activation energy landscape of Aβ42

We used MoCHI, a flexible toolkit for fitting models to deep mutational scanning data [30,31] to infer the change in activation energy $\Delta\Delta E_a$ for all possible AA substitutions. The model assumes additivity of free energy changes across the entire double mutant dataset, with a sigmoidal function accounting for the upper and lower bounds of the growth rate measurements (Fig. 2d, Extended Data Fig. 2c). The resulting free energies were scaled linearly using *in vitro* measurements (Fig. 2e, Extended Data Fig. 2a and Methods) and provide an excellent prediction of the double mutant nucleation rates (Spearman's ρ = 0.83, evaluated by ten-fold cross validation, total of 60,510 variants) (Fig. 2b,d, Supplementary

Table 3). They also agree very well with *in vitro* measurements from three independent studies using two different experimental techniques (Extended Data Fig. 2d)[25,28,29].

To our knowledge, this is the first complete map of activation energies quantified for mutations in any protein (Fig. 2g). The map reveals many interesting features of the mutational landscape of Aβ42. In total, 720 AA substitutions (86%) affect the nucleation activation energy ($\Delta\Delta E_a \neq 0$ kcal/mol, Z-test, FDR < 0.05, Benjamini-Hochberg correction) spanning across all positions of Aβ42. 605 substitutions (72%) increase the activation energy ($\Delta\Delta E_a > 0$ kcal/mol, FDR < 0.05) and 115 (14%) decrease $\Delta\Delta E_a$ ($\Delta\Delta E_a < 0$ kcal/mol, FDR < 0.05) (Fig. 2g). 482 substitutions (57%) cause a larger than 1 kcal/mol change in ΔE_a, with 443 substitutions increasing and 39 substitutions decreasing ΔE_a by > 1 kcal/mol. The energy changes are very well correlated with the effects of individual substitutions ($\rho = 0.94$, $p < 2e\text{-}308$, Fig. 2e, Extended Data Fig. 3a), further highlighting the additive nature of the assay, but with an expanded dynamic range for mutations that slow nucleation (Extended Data Fig. 3a-d, Supplementary Table 4, particularly mutations of G-37, G-38, L-34 and M-35).

Aβ42 nucleation transition state is structured at the C-terminus

The complete activation energy landscape reveals a striking asymmetry in Aβ42: mutations in the C-terminus have much larger effects on the activation energy than mutations in the N-terminus (Fig. 2f,g, Extended Data Fig. 2e). Not all residues are structurally resolved in mature Aβ42 fibrils[32–38], but this distinction between structured and unstructured positions does not explain the differences in activation energies as mutations in many residues that are structured in mature fibrils have only small effects on the activation energy (Fig. 2f,g, Extended Data Fig. 3d).

The primary sequence of Aβ42 contains two hydrophobic regions: residues 17–21 (previously referred to as aggregation prone region 1, APR1) and residues 29–42 (APR2)[27,39–41]. Both APR1 and APR2 form hydrophobic cores in mature fibril structures (Fig. 2f,h, Extended Data Fig. 2e) and both have been proposed to be important for fibril stability[41]. In our data, however, mutations in APR2 have much stronger effects on E_a than mutations in APR1 (Extended Data Fig. 2f), indicating that APR2 is much more important for the rate-limiting step of the Aβ42 nucleation reaction.

The comprehensive activation energy measurements therefore strongly suggest that the C-terminus of Aβ42—but not more N-terminal residues that are also structured in mature fibrils —constitutes the structured region of the nucleation transition state.

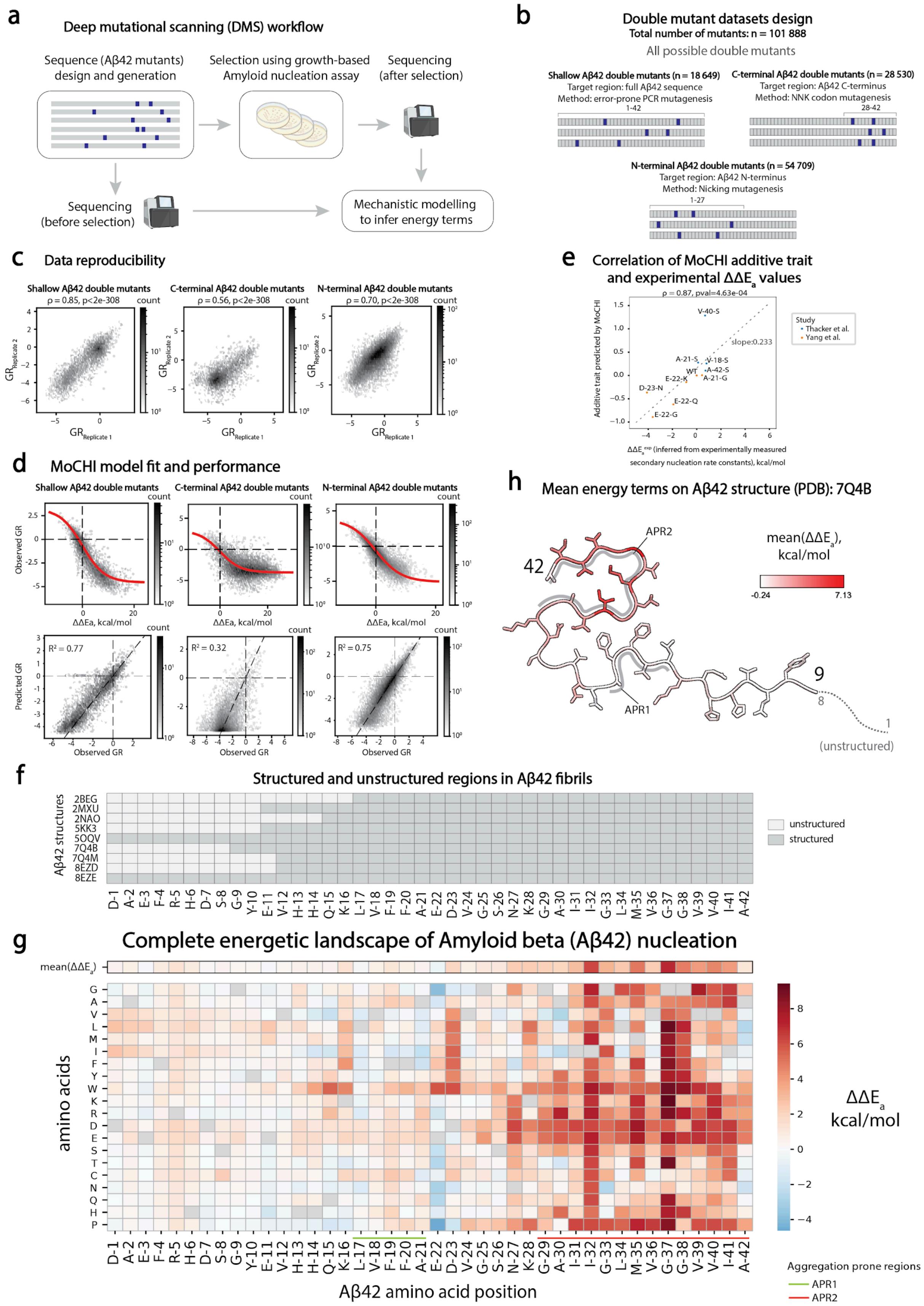


Figure 2. **Complete activation energy landscape for Aβ42 nucleation. a**, Schematic overview of the Deep Mutational Scanning (DMS) approach. **b**, Aβ42 double mutant libraries' design. **c**, Growth rate (GR) replicate correlations for shallow, C-terminal and N-terminal

Aβ42 double mutant libraries (from left to right); Spearman's ρ (correlation) coefficients and associated p-values are reported. **d**, MoCHI model fit (top panels, red trend line represents sigmoid function fit, dashed black lines indicate 0 in both axes) and correlation between observed and predicted growth rates (bottom panels) for shallow, C-terminal and N-terminal Aβ42 double mutants datasets. **e**, Scatterplot of additive trait values predicted by MoCHI[31] model trained on double mutant datasets (Y axis) and experimentally-derived $\Delta\Delta E_a$ values (X axis, derived from secondary nucleation rate constants) used to calibrate MoCHI-inferred terms to kcal/mol units (see Methods for details). Dashed grey line represents linear regression fit for the data. **f**, Overview of structured and unstructured regions in Aβ42 fibrils. **g**, Complete energetic landscape of Aβ42 nucleation: heatmap of inferred activation energy terms ($\Delta\Delta E_a$) for all possible substitutions in Aβ42. Mean $\Delta\Delta E_a$ values for each position are displayed in the top row of the heatmap (outlined in black). Aggregation prone regions 1 (APR1) and 2 (APR2) are highlighted in light green and orange, respectively. **h**, Cross section of 7Q4B PDB structure of Aβ42 fibrils with residues coloured by mean $\Delta\Delta E_a$ per position. Aggregation prone regions 1 (APR1) and 2 (APR2) are highlighted in grey.

Comparison to mature fibril stabilities

We next compared the importance of residues for nucleation to their importance for the stability of mature fibrils. For each mature fibril polymorph of Aβ42 we calculated the effect of every mutation on the thermodynamic stability of the fibril as the change in Gibbs free energy, ΔΔG, (see Methods, Fig. 3a, Extended Data Fig.4) [20,41]. The stability energy matrices show that both APR1 and APR2 are important for structural stability across polymorphs, with mutations in both hydrophobic cores reducing thermodynamic stability (Fig. 3b, Extended Data Fig. 5i).

To more formally compare the changes in activation energy, $\Delta\Delta E_a$, to changes in fibril stability, ΔΔG, we calculated the ratio of $\Delta\Delta E_a$ to ΔΔG, an approach inspired by phi-value analysis [42–46] . A $\Delta\Delta E_a/\Delta\Delta G$ ratio of one indicates that a mutation affects the activation energy as much as the stability of the mature fibril, whereas a ratio near zero means the mutation has very small effects on the activation energy. When mutations only mildly perturb a reaction path, $\Delta\Delta E_a/\Delta\Delta G$ ratios can be interpreted as quantifying the degree of structural conservation between the transition state and the mature fibril (Fig. 3a) [42–46].

Calculating $\Delta\Delta E_a/\Delta\Delta G$ ratios for all moderate effect mutations (see Methods) and all mature fibril structures of Aβ42 shows that the changes in activation energy are similar to changes in stability for mutations in APR2 but much smaller than changes in stability for mutations in APR1 (Fig. 3b-d, Extended Data Fig. 6c, Supplementary Table 6, Supplementary Table 7).

The mature Aβ42 fibril structure with $\Delta\Delta E_a/\Delta\Delta G$ ratios closest to 1 (smallest root mean square distance to 1 across all ratios, Supplementary Table 8) in the APR2 region is 7Q4B, which is a cryo-EM structure of Aβ42 fibrils from familial Alzheimer's disease brains[37], suggesting the structured region of the Aβ42 transition state is most similar to these mature fibril polymorphs. However, most Aβ42 mature fibril polymorphs are structured very similarly in the C-terminal APR2 region (Extended Data Fig. 5a-h), and the $\Delta\Delta E_a/\Delta\Delta G$ ratios are consistently high in this region across the polymorphs (Fig. 3c, Extended Data Fig. 6a,b).

In summary, the comprehensive activation energy measurements and comparisons to mature fibril stabilities suggest that APR1 is largely unstructured in the Aβ42 nucleation transition state whereas APR2 is structured and in a manner that is similar to the structure of this region in mature fibrils (Fig. 3c,d).

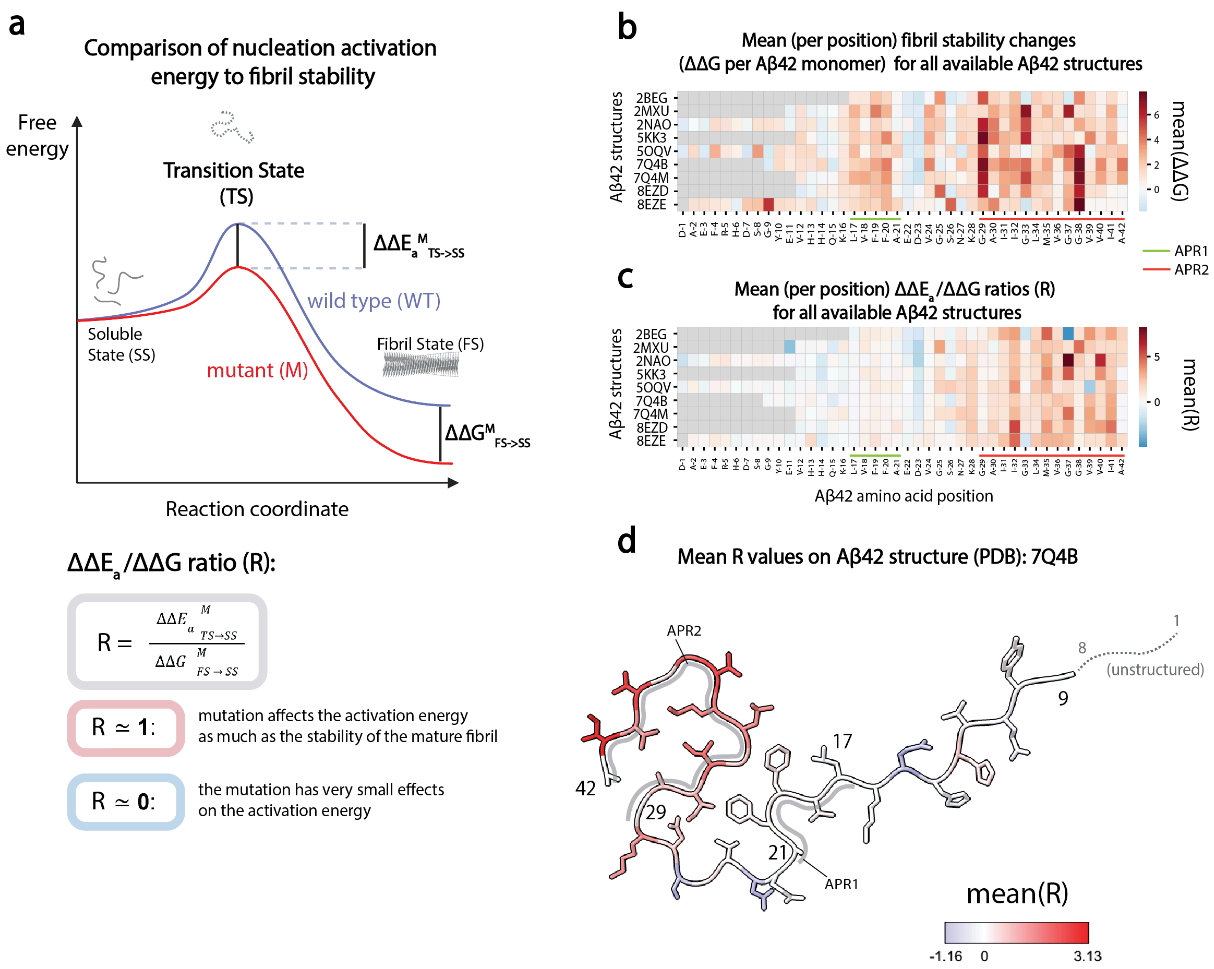


Figure 3. **Comparing activation energies to mature fibril stabilities**. **a**, Free energy landscape of the amyloid nucleation reaction (top). Changes in activation energy ($\Delta\Delta E_a{}^{M}{}_{TS\rightarrow SS}$) and in fibril stability ($\Delta\Delta G^{M}{}_{FS\rightarrow SS}$) upon mutation (M) are highlighted. Principles of $\Delta\Delta E_a/\Delta\Delta G$ ratio analysis (bottom). **b**, Mean (per position) fibril stability changes (ΔΔG per Aβ42 monomer) as predicted by FoldX for all available Aβ42 structures for those substitutions that are later considered in ratio analysis (ΔΔG values are selected using the following criteria: 0.6 kcal/mol < |ddG| < 10 kcal/mol, see Methods). Aggregation prone regions 1 (APR1) and 2 (APR2) are highlighted in light green and orange, respectively. **c**, Mean (per position) $\Delta\Delta E_a/\Delta\Delta G$ ratios (R) for all available Aβ42 structures. Aggregation prone regions 1 (APR1) and 2 (APR2) are highlighted in light green and orange, respectively. **d**, Cross section of 7Q4B PDB structure of Aβ42 fibrils with residues coloured by mean R-values per position.

Combinatorial double mutant cycles

To further probe the structure of the Aβ42 transition state, we performed double mutant cycles [13,14]. In double mutants, the additivity of energy changes can be used to report on structural proximity: while mutations in non-contacting positions normally result in additive changes in free energy, mutations in structurally contacting residues are often energetically coupled, causing non-additive changes in energy when combined. Quantifying energy changes in double mutants and comparing these to energy changes in single mutants therefore provides a powerful approach to probe the structures of proteins, including short-lived high energy transition states[13,14,47,48].

To quantify energetic couplings between mutations in Aβ42 we designed combinatorial mutagenesis libraries in which the effects of single and double mutants were quantified in many different variants of Aβ42 (Fig. 4a,b, Extended Data Fig. 7a). In total we quantified the nucleation of 42,954 variants, with each specific pair of mutations tested in a median of 1,024 different genotypes of Aβ42. As for our comprehensive single mutant energy measurements, quantifying double mutants in Aβ42 variants with fast and slow nucleation rates allows more precise inference of changes in activation energy and across an expanded dynamic range. To reduce the chances of inducing large-structural changes we deliberately employed mutations that conserve side chain hydrophobicity (see Methods).

We used MoCHI to infer both the changes in free energy ($\Delta\Delta E_a$) and the energetic couplings ($\Delta\Delta\Delta E_a$, Fig. 4c, Extended Data Fig. 7b,c). In total, we quantified 640 energetic couplings between 40 pairs of residues (Extended Data Fig. 8a-j, Supplementary Table 9). Consistent with expectations of energetic additivity for most combinations of mutations, the vast majority of energetic couplings are very small (median $\Delta\Delta\Delta E_a$ = 0.3 kcal/mol), with 535/640 (84%) having an absolute value < 1 kcal/mol (Extended Data Fig. 8k). Less than 4% of couplings have an absolute value > 2 kcal/mol, with 15 positive and 9 negative couplings.

Energetic couplings in the Aβ42 transition state

We used the average of the absolute couplings $\Delta\Delta\Delta E_a$ between different mutations in the same residues as an overall metric of the energetic non-independence of two sites (interaction score)[49]. The median of the interaction scores distribution is 0.53 kcal/mol, with only 4/40 pairs of residues with an average coupling > 1 kcal/mol (Fig. 4d, Extended Data Fig. 8l). All four of these strongly coupled sites are located in APR2: M35-V36, I32-M35, I32-V40, M35-V40.

To compare the energetic couplings to the structures of mature fibrils, we plotted the mean absolute $\Delta\Delta\Delta E_a$ values against the distance between amino acid residues in 3D space (minimal side chain heavy atom distance, scHA_min) separately for each of nine Aβ42 fibril polymorphs (Fig. 4e, Extended Data Fig. 9a, Supplementary Table 10). The couplings were significantly correlated with the inverse of the distances in four polymorphs (8EZE, 5OQV, 2BEG, 7Q4M, $p < 0.02$) (Fig. 4f). In Fig. 4e we show this plot for fibrils extracted from human brains affected by familial Alzheimer's diseases (PDB 7Q4B)[37], which has a striking L-shaped distribution of coupling strength versus 3D distances, with mutations in a subset of structural contacts in the mature fibrils strongly energetically coupled in the transition state. Visualising these transition state energetic couplings on the mature fibril structure further highlights that a subset of structural contacts in the C-terminal APR2 region of mature fibrils are strongly energetically coupled in the transition state (Fig. 4g, Extended Data Fig. 9b).

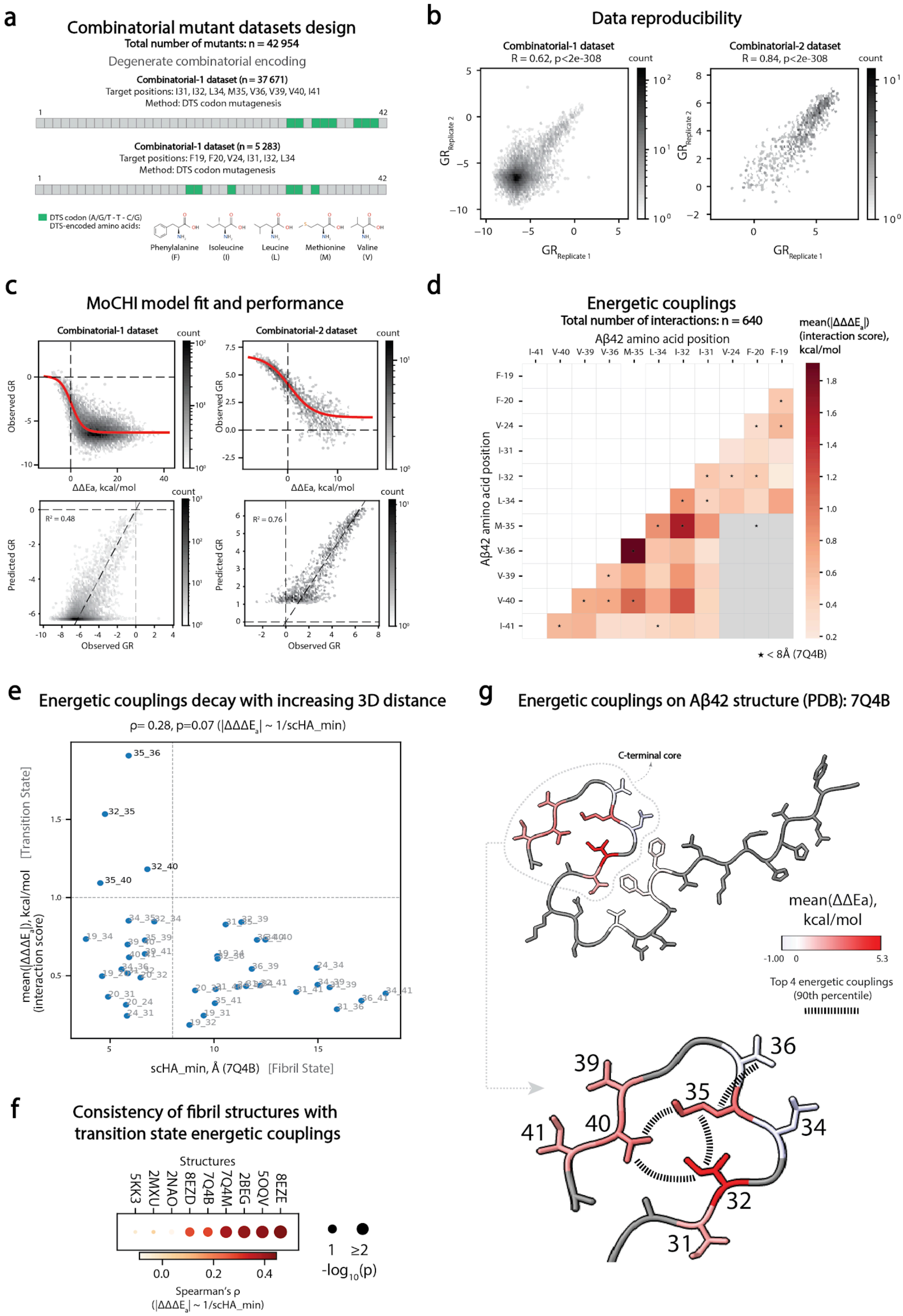

a
Combinatorial mutant datasets design
Total number of mutants: n = 42 954
Degenerate combinatorial encoding
Combinatorial-1 dataset (n = 37 671)
Target positions: I31, I32, L34, M35, V36, V39, V40, I41
Method: DTS codon mutagenesis
1
42
Combinatorial-1 dataset (n = 5 283)
Target positions: F19, F20, V24, I31, I32, L34
Method: DTS codon mutagenesis
1
42
DTS codon (A/G/T - T - C/G)
DTS-encoded amino acids:
Phenylalanine (F)
Isoleucine (I)
Leucine (L)
Methionine (M)
Valine (V)
b
Data reproducibility
Combinatorial-1 dataset
R = 0.62, p<2e-308
Combinatorial-2 dataset
R = 0.84, p<2e-308
count
$GR_{Replicate\ 2}$
$GR_{Replicate\ 1}$
c
MoCHI model fit and performance
Combinatorial-1 dataset
Combinatorial-2 dataset
Observed GR
ΔΔEa, kcal/mol
Predicted GR
Observed GR
R² = 0.48
R² = 0.76
count
d
Energetic couplings
Total number of interactions: n = 640
Aβ42 amino acid position
I-41 V-40 V-39 V-36 M-35 L-34 I-32 I-31 V-24 F-20 F-19
F-19 F-20 V-24 I-31 I-32 L-34 M-35 V-36 V-39 V-40 I-41
Aβ42 amino acid position
mean(|ΔΔΔE$_a$|) (interaction score), kcal/mol
1.8 1.6 1.4 1.2 1.0 0.8 0.6 0.4 0.2
★ < 8Å (7Q4B)
e
Energetic couplings decay with increasing 3D distance
ρ= 0.28, p=0.07 (|ΔΔΔE$_a$| ~ 1/scHA_min)
35_36
32_35
32_40
35_40
mean(|ΔΔΔE$_a$|), kcal/mol (interaction score) [Transition State]
scHA_min, Å (7Q4B) [Fibril State]
f
Consistency of fibril structures with transition state energetic couplings
Structures
5KK3 2MXU 2NAO 8EZD 7Q4B 7Q4M 2BEG 5OQV 8EZE
0.0 0.2 0.4
Spearman's ρ (|ΔΔΔE$_a$| ~ 1/scHA_min)
1 ≥2
-log$_{10}$(p)
g
Energetic couplings on Aβ42 structure (PDB): 7Q4B
C-terminal core
mean(ΔΔEa), kcal/mol
-1.00 0 5.3
Top 4 energetic couplings (90th percentile)
39 36 35 41 40 34 32 31

Figure 4. **Energetic couplings in the Aβ42 transition state**. **a**, Combinatorial Aβ42 mutant libraries' design; DTS codons encode Phenylalanine (F), Methionine (M), L (Leucine), I (Isoleucine), V (Valine). **b**, Growth rate (GR) replicate correlations for Combinatorial-1 (left) and Combinatorial-2 (right) Aβ42 combinatorial mutant libraries. Pearson's correlation coefficients (R) and associated p-values are indicated. **c**, MoCHI model fit (top panels, red trend line represents sigmoid function fit, dashed black lines indicate 0 in both axes) and and correlation between observed and predicted growth rates (bottom panels) for Combinatorial-1 and Combinatorial-2 Aβ42 combinatorial mutant libraries. **d**, Interaction scores (mean absolute value of energetic couplings for a pair of positions, mean(|$\Delta\Delta\Delta E_a$|)) for mutagenised Aβ42 position pairs. Star symbols (⋆) indicate residue pairs closer than 8 Å in 7Q4B Aβ42 structure. **e**, Scatterplot of interaction scores for pairs of positions and inter-residue distance for the corresponding pairs of amino acids in 3D space (scHA_min, minimum heavy atom side chain distance) as measured in the Aβ42 fibril structure 7Q4B; dashed light grey vertical line marks 8 Å, dashed light grey horizontal line marks interaction score (mean(|$\Delta\Delta\Delta E_a$|) of 1 kcal/mol. **f**, Dotplot representing consistency of transition state interactions with fibril structures (2BEG, 2MXU, 2NAO, 5KK3, 5OQV, 7Q4B, 7Q4M, 8EZD, 8EZE) according to energetic couplings and amino acid inter-residue distances; size of the dot reflects statistical significance ($-\log_{10}$(p)) of the correlation between interaction score (mean(|$\Delta\Delta\Delta E_a$|)) and inverse of the inter-residue distance (1/scHA_min); colour of the dot represents Spearman's ρ (correlation) coefficient. **g**, Cross section of 7Q4B PDB structure of Aβ42 fibrils with residues coloured by mean activation energy terms ($\Delta\Delta E_a$) inferred from the MoCHI model trained on combinatorial mutants datasets (residues with no inferred $\Delta\Delta E_a$ are shown in grey). Positions mutagenised in combinatorial datasets (Combinatorial-1 and Combinatorial-2) are labelled on the PDB structure. Top 4 interacting position pairs (in 90th percentile of interaction score distribution) are connected with dashed black lines.

Taken together, therefore, both the complete activation energy mutational landscape and the large-scale measurement of energetic couplings strongly suggest that the nucleation transition state of Aβ42 is structured in the short APR2 C-terminal hydrophobic region with strong energetic interactions for a subset of the structural contacts observed in mature fibrils.

Discussion

We have demonstrated here the feasibility of using mutation-selection-sequencing experiments to study protein transition states and presented an ‘energetic portrait’ of the transition state that initiates Aβ42 aggregation and, potentially, Alzheimer’s disease.

In total we measured the nucleation rates of >140,000 protein sequences to quantify the change in nucleation activation energy for all 798 amino acid substitutions in Aβ42 and, in addition, the energetic couplings between 640 pairs of mutations. This is, to our knowledge, the first comprehensive measurement of changes in activation energy for any protein and also the first large-scale measurement of energetic couplings for a kinetic process. Our work builds on pioneering studies by Fersht and others[16,19,20], expanding activation energy measurements and energetic couplings to the scale of all mutations in a sequence.

A key aspect of our approach is the use of double mutants and combinatorial mutagenesis. Quantifying mutational effects and interactions in variants with faster and slower kinetics expands the dynamic range of an assay and so the precision of energy measurements. We inferred energies and couplings by fitting models to the data using fast and flexible neural networks, but alternative strategies can also be used [31]. The approach assumes that changes in free energy are largely additive when mutations are combined, and the good predictive performance of our models even when many different mutations are combined suggests that this is indeed the case for Aβ42 nucleation.

A second key aspect of our approach is the use of a kinetic selection assay where enrichments report on the rate of a reaction. This is not true for most mutation-selection-sequencing experiments, where enrichments depend on thermodynamic stabilities [30,50,51]. An important challenge for the future is the development of kinetic selection assays that report on reactions other than amyloid nucleation. Approaches using microfluidics [52] and droplet-based selections [53,54] may allow this.

Amyloid plaques of Aβ42 are a pathological hallmark of all forms of Alzheimer’s disease and mutations in Aβ42 cause familial forms of the disease[3]. The only drugs demonstrated to slow the progression of Alzheimer’s disease in clinical trials are antibodies targeting Aβ42 [55,56]. We believe that our results identify the key molecular species that initiates the formation of Aβ42 amyloid fibrils and so, potentially, also Alzheimer’s disease. The ‘energetic structure’ of this transition state consists of a partially-structured C-terminal hydrophobic core in the APR2 region. Put bluntly, we believe that preventing the formation of this C-terminal structure should stop the nucleation and spread of Aβ42 fibrils, and so, potentially, also the development of Alzheimer’s disease.

Methods

Library designs and construction

N-terminal Aβ42 double mutant library

To obtain the N-terminal Aβ42 double mutant library (532 possible single aa mutants and 151,200 possible double aa mutants) we used a nicking mutagenesis approach described in Mighell et al. [57]. The Aβ42 region with 25 bp upstream and 21 bp downstream was amplified from plasmid PCUP1-Sup35N-Aβ42 [58] kindly provided by the Chernoff lab using oligos AB_TS_003-004 (Supplementary Table 11) to introduce AvrII and HindIII restriction sites. The target region for mutagenesis was then digested and ligated (T4 DNA Ligase, Thermo Scientific) into the nicking plasmid pGJJ057, previously digested with the same restriction sites.

There are 4 main steps in the nicking mutagenesis protocol: i) obtention of ssDNA template, ii) synthesis of a mutant strand by annealing and extension of mutagenic oligos (oligos AB_TS_005-032, Supplementary Table 11), iii) degradation of the wild type (WT) template strand and iv) synthesis of the 2nd mutant strand (oligo AB_TS_033, Supplementary Table 11). We used 28 mutagenic oligos synthesised by IDT, each of them containing one NNK (N=A/T/C/G; K=T/G) degenerate codon at every targeted position for mutagenesis, flanked by 7 upstream and 7 downstream wild-type codons.

In order to obtain a double mutant library, we ran the protocol twice. In the first round, we obtained a library with 66% of single amino acid mutant and 33% of WT sequences (estimated by sanger sequencing of individual clones). In the second round - in which we used the single mutant library as template - we obtained a library with 66% of double and 33% of single amino acid mutant sequences.

As described in Mighell et al [57], the N-terminal Aβ42 double mutant library was finally transformed into 10-beta Electrocompetent *E. coli* (NEB). Cells were recovered in SOC and plated on LB with ampicillin to assess transformation efficiency. A total of 2.88 million transformants were estimated, representing each variant of the library more than 18x. 50 ml of overnight culture were harvested to purify the library in the nicking plasmid pGJJ057 (QIArep Miniprep Kit, Qiagen).

In order to clone the library back inside the PCUP1-Sup35N plasmid for selection, the library in the nicking plasmid was digested with EcorI and XbaI restriction enzymes (Thermo Scientific) for 4h at 37C and purified from a 2% agarose gel (QIAquick Gel Extraction Kit, Qiagen). At this stage, the purified product was ligated into the previously digested PCUP1-Sup35N and transformed into 10-beta Electrocompetent *E. coli* (NEB). For this library, a total of 3.1 million transformants were estimated.

C-terminal Aβ42 double mutant library

To obtain a library of double mutants in the Aβ28-42 region (285 possible single aa mutants and 37,905 possible double aa mutants), we used a NNK codon mutagenesis approach. We ordered an oligo pool (IDT) containing 105 oligos of Aβ28-42 (15 AA positions), each with 2 positions containing an NNK (N=A/T/C/G; K=T/G) degenerate codon (oligo pool AB_TS_034, Supplementary Table 11). Each NNK degenerate codon encoded 32 possible codons, resulting in a library of 107,520 unique nt sequences. Each oligo also contained 5' (CTTTGCAGAAGATGTGGGTTCAAAC) and 3' (TAATCTAGAGCGGCCGCCACC) constant regions for cloning purposes.

500 ng of the oligo pool were amplified by PCR (Q5 high-fidelity DNA polymerase, NEB) for 10 cycles with primers annealing to the constant regions (oligos AB_TS_037,039, Supplementary Table 11). The PCR product was incubated with ExoSAP-IT (Thermo Fisher Scientific) at 37C for 1h and purified by column purification (MinElute PCR Purification Kit, Qiagen). This library did not contain the EcoRI restriction site between Sup35N and Aβ42, which was previously used for cloning of the shallow and N-terminal Aβ42 double mutant libraries.

In parallel, empty plasmids PCUP1-Sup35N-Aβ(1-27) (i.e. missing a specific Aβ region for mutagenesis) was constructed by PCR linearisation PCUP1-Sup35N-Aβ42 using oligos AB_TS_040-041 (Supplementary Table 11). These was then linearised by PCR for 35 cycles (oligos AB_TS_044-045, Supplementary Table 11), treated with DpnI (FastDigest, Thermo Scientific) overnight and purified from a 1% agarose gel (QIAquick Gel Extraction Kit, Qiagen).

The purified library was then ligated into 200 ng of PCUP1-Sup35N-Aβ(1-27) in a 10:1 (library:plasmid) ratio by Gibson assembly with 3h of incubation. The resulting product was then dialysed for 3h by using a membrane filter (MF-Millipore 0.025 µm membrane, Merck) and concentrated 10X using a speed vacuum concentrator. Finally, the library was transformed into 10-beta Electrocompetent *E. coli* (NEB). Cells were recovered in SOC and plated on LB with ampicillin to assess transformation efficiency. 50 ml of overnight culture were harvested to purify the plasmid library (QIArep Miniprep Kit, Qiagen). A total of 1.8 million transformants were estimated, representing each variant of the library more than 47 times.

Shallow Aβ42 double mutant library

We built a library containing double mutants across the entire Aβ42 sequence (798 possible single aa mutants and 310,821 possible double aa mutants) in a previous study[26]. Briefly, the library was obtained by error-prone PCR and cloned by digestion and ligation into the PCUP1-Sup35N plasmid (oligos AB_TS_001-002). Additional details on library construction are provided in Seuma et al.[26]. For this library, a total of 4.1 million transformants were estimated, representing each variant in the library >10x.

Combinatorial Aβ42 libraries

We designed 2 combinatorial libraries (Combinatorial-1 and Combinatorial-2) by mutagenizing specific positions of Aβ1-42 (I31, I32, L34, M35, V36, V39, V40 and I41 in Combinatorial-1; F19, F20, V24, I31, I32 and L34 in Combinatorial-2) to a specific subset of hydrophobic amino acids. We used the DTS degenerate codon (D=A/T/G; T=C/G) to encode methionine, valine, leucine, isoleucine and phenylalanine. For each library, we purchased a 4 nmole ultramer from IDT (oligos AB_TS_035-036, Supplementary Table 11) covering the target region for mutagenesis (Aβ29-42 for Combinatorial-1 and Aβ12-42 for Combinatorial-2) containing DTS codons at the above mentioned positions. The oligos also contain 5' and 3' constant regions for cloning purposes. The number of possible unique amino acid sequences is 390,625 for Combinatorial-1 and 15,625 for Combinatorial-2.

These libraries were amplified and cloned following the same steps as for the C-terminal Aβ42 double mutant library (oligos AB_TS_037-045). The purified Combinatorial-1 library was ligated into PCUP1-Sup35N-Aβ(1-27) while Combinatorial-2 into PCUP1-Sup35N-Aβ(1-11). Combinatorial-1 and Combinatorial-2 also did not contain the EcoRI restriction site between Sup35N and Aβ42, which was previously used for cloning of the shallow and N-terminal Aβ42 double mutant libraries.

A total of 0.38 million transformants were estimated for Combinatorial-2 representing each variant of the library more than 24 times. The Combinatorial-1 library was bottlenecked to 0.94 million transformants due to high complexity.

Yeast transformation

Saccharomyces cerevisiae GT409 [psi-pin-] (MATa ade1-14 his3 leu2-3,112 lys2 trp1 ura3-52) strain (kindly provided by the Chernoff lab) was used in all experiments in this study. Lithium acetate transformation was used to transform each plasmid library in yeast cells, in three biological replicates. An individual colony for each transformation tube was grown overnight in 20 ml YPDA at 30 C and 200 rpm. Once at saturation, cells were diluted to OD600= 0.25 in 175 ml YPDA and grown until exponential phase, for ~5h. Cells were then harvested, washed with milliQ and resuspended in 8.5ml sorbitol mixture (100 mM LiOAc, 10 mM Tris pH 8, 1 mM EDTA, 1M sorbitol). 5ug of plasmid library, 175ul of ssDNA (10mg/ml, UltraPure, Thermo Scientific) and 35 ml of PEG mixture (100 mM LiOAc, 10 mM Tris pH 8, 1 mM EDTA pH 8, 40% PEG3350) were added to the cells, and incubated for 30 min at RT. Heat-shock was performed for 15 min at 42C in a water bath. Cells were then harvested, washed and resuspended in 50ml recovery medium (YPDA, 0.5M sorbitol) for 1h at 30C 200 rpm. Finally, cells were harvested, washed and resuspended in 350 ml plasmid selection medium (-URA 10% glucose) and grown for 48h. Once at saturation, cells were diluted in 200 ml plasmid selection medium (-URA 10% glucose) to OD600= 0.05 and grown until they reached the exponential phase, for ~15h. Lastly, cells were harvested and stored at -80C in 25% glycerol. A minimum of 3.7, 3, 2.2, 0.99 and 0.43 million transformants were estimated for each of the three biological replicates of shallow Aβ42 double mutant, N-terminal Aβ42 double mutant, C-terminal Aβ42 double mutant, Combinatorial-1 and Combinatorial-2 combinatorial libraries, respectively.

Selection experiments

Each transformation replicate was later used for selection. Tubes were thawed from -80C, washed and resuspended in 100-1000ml plasmid selection medium (-URA 10% glucose) - ensuring 100 cells each variant in the library - at OD=0.05 and grown until exponential at 30C 200 rpm. Once at exponential, cells were diluted in 100-1000 ml protein induction medium (-URA 10% glucose, 100µM Cu2SO4) at OD=0.05 and grown for 24 h at 30C 200 rpm. After 24h, 25-50ml of cells were harvested for input pellets and a minimum of 18.5, 75, 12, 230 and 1.5 million cells / replica (for shallow Aβ42 double mutant, N-terminal Aβ42 double mutant, C-terminal Aβ42 double mutant, Combinatorial-1 and Combinatorial-2 libraries, respectively) were plated on -URA-ADE selection medium plates (145cm2, Nunc, Thermo Scientific) and incubated for 6 days (for C-terminal Aβ42 double mutant, Combinatorial-1 and Combinatorial-2 combinatorial libraries) or 7 days (shallow and

N-terminal Aβ42 double mutant libraries) at 30C. The number of cells plated ensures a minimum coverage of 10 times each variant in the library. Finally, colonies were scraped off the plates and harvested for output pellets. Inputs and output pellets were stored at -20C and later treated for DNA extraction.

DNA extraction and sequencing library preparation

Input and output pellets (3 biological replicates each, 6 tubes in total) were resuspended in 2 ml extraction buffer (2% Triton-X, 1% SDS, 100 mM NaCl, 10 mM Tris pH 8, 1 mM EDTA pH 8) and subjected to two cycles of freezing-thawing in an ethanol-dry ice bath and water bath at 62C for 10 min each. 1.5 ml of phenol:chloroform:isoamyl 25:24:1 was then added to the pellets and vortexed for 10 min. The aqueous phase was recovered by means of 30min centrifugation at 3000 rpm. DNA was precipitated with 1:10 V NaOAc 3M and 2.2 V of 99% cold ethanol and incubated at -20C for 2h. After centrifugation, pellets were dried overnight at RT. The next day, pellets were resuspended in TE 1X buffer (10mM Tris pH8, 1mM EDTA pH8) and treated with 10 ul RNase A (Thermo Scientific) for 1h at 37C. DNA was purified using a silica beads extraction kit (QIAEX II Gel Extraction Kit, Qiagen). Plasmid library concentration was quantified by quantitative PCR using primers annealing to the origin of the replication site of the plasmid (oligos AB_TS_046-047, Supplementary Table 11).

Sequencing library preparation

Each sequencing library was prepared in a two-step PCR (Q5 high-fidelity DNA polymerase, NEB). First, the mutagenised Aβ region was amplified for 15 cycles with frame-shifted primers (oligos AB_TS_048-080, Supplementary Table 11), using 300M (for Shallow Aβ42 double mutant library), 100M (for N-terminal Aβ42 double mutant, C-terminal Aβ42 double mutant and Combinatorial-1 combinatorial libraries) and 50M (for Combinatorial-2 combinatorial library) molecules as template for each sample (3 inputs, 3 outputs). PCR products were treated with ExoSAP-IT (Thermo Fisher Scientific) at 37C for 1h, purified by column purification (MinElute PCR Purification Kit, Qiagen) and eluted in 5ul TE 1X buffer (10mM Tris, 1mM EDTA). 4ul of purified product were used as template for the second PCR. In this case, samples were amplified for 10 cycles with primers containing Illumina sequencing indexes (oligos AB_TS_081-111, Supplementary Table 11). The three input samples were pooled together equimolarly, and the same for the three output samples. The final pools were purified from a 2% agarose gel using a silica beads extraction kit (QIAEX II Gel Extraction Kit, Qiagen).

Libraries were sequenced on an Illumina HiSeq2500 sequencer using 125 paired-end reads (shallow Aβ42, N-terminal and C-terminal Aβ42 double mutant libraries) or 150 paired-end sequencing on an Illumina NextSeq500 (Combinatorial-1 and Combinatorial-2 combinatorial libraries) at the CRG Genomics core facility.

Data processing, growth rates and error estimates

FastQ files from paired end sequencing were processed using DiMSum[59], an R pipeline for the analysis of deep mutational scanning data. A total of 251.6M (for Shallow Aβ42 double mutant library), 352.9M (for N-terminal Aβ42 double mutant library), 106.3M (for C-terminal Aβ42 double mutant library), 109.8M (for Combinatorial-1) and 26.8M (for Combinatorial-2, 15.9M for input samples and 10.9M for output samples) paired-end reads were obtained from sequencing.

5' and 3' constant regions were trimmed, allowing an error rate of 20% mismatches compared to the reference sequence. Reads were aligned and sequences with a Phred base quality score below 30 and non-designed sequences were discarded. Variants with fewer than 10 input reads in any of the replicates were discarded. Estimates from DiMSum were used to choose the filtering thresholds.

Relative growth rates and their associated error estimates (available in Supplementary Table 2) were also calculated using DiMSum. Growth rate for a specific variant i ($GR_{i\,(vs\,WT)}$) in the library in each biological replicate is defined in the following way:

$GR_{i\,(vs\,WT)} = ES_i - ES_{WT}$, where ES_i is the enrichment score for the variant i, and ES_{WT} is the enrichment score of the WT Aβ42, both defined as follows: $ES_i = log(F_i^{OUTPUT} - F_i^{INPUT})$.

Growth rates for each variant were merged using error-weighted mean of each variant across replicates and centered using the error-weighted mean frequency of WT Aβ42 synonymous substitutions. Due to large possible sequence space (16-391K variants at the AA level) in combinatorial mutant datasets, WT Aβ42 was not present in there, so the normalisation was done using a random variant present in those datasets. These data were later centered after modelling with MoCHI using common variants between double and combinatorial mutant datasets (see next section).

Inferring changes in activation energy with MoCHI

To infer activation energy terms from growth rates of cells carrying nucleating Aβ42 variants, we fitted an energy model to our datasets using MoCHI [30,31]. Briefly, MoCHI takes as input amino acid sequences, measured growth rates and error estimates of each variant in the library. MoCHI then predicts growth rates for the given variants based on a particular fit (sigmoid in this case), while correcting for global non-linearities (non-specific epistasis) that in this case are due to the upper and lower limits of the growth assay. The effects of individual mutations and mutation combinations (genetic interactions) are modelled additively at the energetic level. Using the coefficients derived by the model, one then obtains the change in activation energy associated with each mutation for the phenotype of interest (in our case, amyloid nucleation).

For double mutant datasets (shallow Aβ42 double mutant library, N-terminal Aβ42 double mutant library and C-terminal Aβ42 double mutant library) we used MoCHI with default parameters for a two-state model with one phenotype (nucleation) for the three double mutants datasets, using L1 and L2 regularization with a lambda of 10^{-5} and allowing only first order (additive) energy terms. We evaluated the model using the held-out "fold" from the 10 times that the model was run on the dataset.

In order to translate (or calibrate) additive traits inferred with MoCHI[31] into units of energy (kcal/mol), we used experimentally measured nucleation rate constants from previously published studies[28,60] . We derived changes in activation energy for Aβ42 variants from the available kinetic data, using the following relationship (derived from the Arrhenius equation, see Fig. 1c):

$\Delta\Delta E^{experimental}_{a\,(i\ vs\ WT)} = RT * ln(\frac{k_{WT}}{k_i})$, for Aβ42 variant i, where R is the universal gas constant and T is temperature in Kelvin (303 K, as used in MoCHI model training). For the Yang et al. dataset[28] we used primary and secondary nucleation rate constants (k_n and k_2) reported in the study to directly calculate changes in activation energy for reported Aβ42 variants (D23N, E22G, E22Q, E22K and A21G) relative to WT. For the Thacker[25] dataset, we derived multiplicative terms k_+k_n and k_+k_2 (primary and secondary nucleation rate constants multiplied by the rate of elongation) from the reported λ and κ values (the rate at which new fibril mass is formed via primary and secondary nucleation, respectively) and the exact model description authors provided in the supplementary material of the publication, for the reported Aβ42 variants (V18S_A21S, V40S_A42S, A-21-S, A-42-S, V-18-S and V-40-S) relative to WT, and used these multiplicative terms in the same equation (above) to calculate the activation energy change. We fitted a linear regression model to be able to predict our MoCHI-inferred additive trait values from the experimentally-derived $\Delta\Delta E_a$ values for Aβ42 variants common across the two datasets (Fig. 2e, Extended Data Fig. 2a), and used the resulting slope (0.233, fitting with secondary nucleation rate-derived $\Delta\Delta E_a$ values) to calibrate the MoCHI terms to kcal/mol units.

For combinatorial mutant datasets (Combinatorial-1 and Combinatorial-2), as WT Aβ42 variant was not present in any of the two datasets, we introduced it artificially and added to the data prior to training the MoCHI model, declaring its growth rate as 0 and its error estimate as an arbitrary big number, 100 in this case. We used MoCHI with default parameters for a two-state model with one phenotype (nucleation) for the two combinatorial mutants datasets, using L1 and L2 regularization with a lambda of 10^{-5}, and allowing first order (additive) and second order (non-additive) energy terms to account for energetic couplings. We evaluated the model using the held-out “fold” from the 10 times that the model was run on the dataset. Since we used an artificially introduced WT Aβ42 variant in MoCHI model training, we then re-centered the resulting predicted activation energy terms. To do this, we chose all common variants between double and combinatorial mutant datasets and fit a linear regression model that inferred a slope and intercept for the activation energy terms in double vs combinatorial mutant datasets. We used the derived slope (0.416) and intercept (-0.155 kcal/mol) values to recenter the energy terms of the combinatorial mutants.

Visualisation of numeric values on Aβ42 structures

For purposes of visualisation of numeric values (energy terms, $\Delta\Delta E_a/\Delta\Delta G$ ratios and energetic couplings) on Aβ42 structures in figures, we used the following chains for each of the PDB files: B for 2BEG; D for 2MXU; C for 2NAO; D for 5KK3; F for 5OQV; E for 7Q4B; G for 7Q4M; E for 8EZD; and E for 8EZE. Molecular graphics and analyses performed with UCSF ChimeraX, developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco, with support from National Institutes of Health R01-GM129325 and the Office of Cyber Infrastructure and Computational Biology, National Institute of Allergy and Infectious Diseases[61].

Fibril stability analyses

Inspired by phi-value analysis[20,41], we calculated activation energy to fibril stability energy ratios for Aβ42 by dividing our inferred activation energy terms $\Delta\Delta E_a$ (Fig. 2f) by the change in free energy ΔΔG of fibril state structures of Aβ42 (separately for 2BEG, 2MXU, 2NAO, 5KK3, 5OQV, 7Q4B, 7Q4M, 8EZD, 8EZE, see Supplementary Table 5 for structure details). Following the example of a recent study on PI3K-SH3 amyloids, where FoldX predictions of ΔΔG were shown to correlate well with *in vitro* measurements of fibrils stability for 15 PI3K-SH3 variants [20], we ran FoldX on a stacked single filament tetramer fibril structures of Aβ42 peptides. For the 2NAO structure we ran it on a stacked single filament trimer conformation, as the PDB structure did not contain more than three stacked chains in any filament. We used the following Aβ42 chains: B, C, D, E for 2BEG; A, B, C for 2NAO; A, B, C, D for 2MXU; A, B, C, D for 5KK3; A, C, F, H for 5OQV; B, D, F, R for 7Q4B; A, C, E, G for 7Q4M; E, F, G, H for 8EZD; and E, F, G, H for 8EZE. Resulting ΔΔG values were then divided by 4 (or 3 in case of 2NAO) to obtain per-monomer ΔΔG values. In our downstream analysis, we further only considered those $\Delta\Delta E_a/\Delta\Delta G$ ratios, for which the predicted ΔΔG values were above 0.6 kcal/mol (literature-motivated threshold [19]) and below 10 kcal/mol, to only consider mutations that do not significantly perturb FS structure (Fig. 3b,c, Extended Data Fig. 4, Extended Data Fig. 5a-h). The main assumption underlying $\Delta\Delta E_a/\Delta\Delta G$ ratio analysis is that the introduced mutations do not perturb these structures drastically.

Calculation of inter-residue distances in 3D space

Distances between AA side chains in 3D space were calculated using the DMS2structure toolkit (available at https://github.com/lehner-lab/DMS2structure/ and published in a previous study [47]). Briefly, for two given amino acids in a PDB structure, DMS2structure calculates distances between all pairs of heavy atoms (any atoms other than hydrogen) in their side chains across the two amino acids. The minimum of these distances is then reported as the minimal side chain heavy atom distance (scHA_min), further used in downstream analysis. In absence of a side chain for Glycine, the central carbon atom (C-α) is used for calculations. We specifically used *contact_matrix_from_pairdistances.R* script to calculate inter-residue distances for monomer conformations of Aβ42 structures, and *pairdistances_from_PDB_crystal.R* script to calculate inter-residue distances for dimer (2 monomers in different filaments facing each other) conformations of Aβ42 structures.

Data availability

Raw sequencing data and the processed data table (Supplementary Table 2) are deposited in NCBI's Gene Expression Omnibus (GEO) with accession number GSE247583 (N-terminal Aβ42 double mutant, C-terminal Aβ42 double mutant and combinatorial libraries) and GSE151147 (shallow double mutants library [26]).

Code availability

All the code used for analyses presented in this work are available at: https://github.com/lehner-lab/amyloids_energy_modelling.

Author contributions

A.A. and M.S. analysed the data. M.S. performed all experiments. M.S., B.B. and B.L. designed the experiments. A.J.F. developed the modelling approach and performed preliminary data analyses with M.S.. B.B. and B.L. conceived the project and supervised the research. A.A., B.B. and B.L. wrote the manuscript with input from M.S..

Competing interests

The authors declare no competing interests.

Acknowledgements

This work was funded by the La Caixa Research Foundation project ‘DeepAmyloids’ (LCF/PR/HR21/52410004). Work in the lab of BB was also funded by the Spanish Ministry of Science, Innovation and Universities (PID2021-127761OB-I00, RYC2020-028861-I funded by MCIN/AEI/ 10.13039/501100011033, “ERDF A way of making Europe” and “ESF Investing in your future”) and by the European Union (ERC Consolidator, Glam-MAP, 101125484). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. Work in the lab of B.L. was funded by a European Research Council (ERC) Advanced (883742) grant, the Spanish Ministry of Science and Innovation (LCF/PR/HR21/52410004, EMBL Partnership, Severo Ochoa Centre of Excellence), the Bettencourt Schueller Foundation, the AXA Research Fund, Agència de Gestió d’Ajuts Universitaris i de Recerca (AGAUR, 2017 SGR 1322), the CERCA Program/Generalitat de Catalunya and Wellcome (Grant reference: 220540/Z/20/A, 'Wellcome Sanger Institute Quinquennial Review 2021-2026'). M.S was funded by a fellowship from Agencia de Gestio d’Ajuts Universitaris i de Recerca (2019FI_B 01311). A.J.F. was funded by a Ramón y Cajal fellowship (RYC2021-033375-I) financed by the Spanish Ministry of Science and Innovation (MCIN/AEI/10.13039/501100011033) and the European Union (NextGenerationEU/PRTR).

Description of Supplementary Tables

Supplementary Table 1

Description: Primary and secondary nucleation rate constants measured independently in previous studies[25,28,29] for Aβ42 mutants (with each of the studies' data in separate sheets), and their corresponding growth rates measured in this study in the double mutants datasets (GR_shallow_double_mutants, GR_C_terminal_Ab42_double_mutants and GR_N_terminal_Ab42_double_mutants) and change in additive trait as inferred by MoCHI[30,31] (ddEa_joint_model) trained on all double mutants datasets. For the Yang et al. dataset[28] primary and secondary nucleation rate constants (k_n and k_2) reported in the study are in 'kn_yang' and 'k2_yang' columns, respectively. For the Thacker[25] and Illes-Toth[29] datasets, multiplicative terms k_+k_n and k_+k_2 (primary and secondary nucleation rate constants multiplied by the rate of elongation) were derived from the reported λ and κ values (the rate at which new fibril mass is formed via primary and secondary nucleation, respectively, in columns 'lambda' and 'kappa') using the exact model description authors provided in the supplementary material of the publications, and are in 'k+kn_thacker' or 'k+kn_illestoth' and 'k+k2_thacker' or 'k+k2_illestoth' columns, respectively.

Supplementary Table 2

Description: Growth rates (relative to WT) and their associated error estimates from DiMSum[59] for all the Aβ42 mutants (double and combinatorial) generated and analysed in this study.

Supplementary Table 3

Description: Inferred changes in activation energy (column 'ddEa_scaled', in kcal/mol units) for all the substitutions in Aβ42 (column 'id') with associated outputs from MoCHI[30,31] model (fit on all the Aβ42 double mutant datasets jointly). Columns 'zscore_unaffected', 'p.adjust_mode', 'category_affected' and 'category_incr_decr_nucleation' present the statistics from the Z-test asking whether inferred $\Delta\Delta E_a$ values are different from $\Delta\Delta E_a = 0$ kcal/mol (unaffected nucleation) and in which direction (> 0 or < 0). For the first entry (WT) ΔE_a is reported.

Supplementary Table 4

Description: (ddEa_and_GR_values sheet) inferred changes in activation energy (column 'ddEa'), corresponding mean growth rate for all Aβ42 double mutants (for each mutant, the mean growth rate is calculated across all three or less Aβ42 double mutants datasets where this mutant is present, with column 'dataset' indicating which dataset the mutant is present in). (ddEa_and_GR_Zscores sheet) Z-scores of changes in activation energy (X column) and Z-scores of growth rates necessary to reproduce heatmaps in ED Fig.3b-d (columns 'ddEa_norm', 'neg_GR_mean_norm' and 'ddEa_norm_minus_neg_GR_mean_norm', respectively, for panels b,c and d).

Supplementary Table 5

Description: Aβ42 structures used in analyses of the study, indicating their PDB ID, technique employed for structural determination and peptide origin. H-D - Hydrogen Deuterium, NMR - Nuclear Magnetic Resonance, SS - solid state, EM - electron microscopy, Cryo EM - cryogenic electron microscopy, MAS NMR - magic-angle spinning NMR.

Supplementary Table 6
Description: Processed FoldX output table containing predicted changes in free energy ($\Delta\Delta G$) of Aβ42 fibril structures (each sheet in the table corresponds to a given Aβ42 PDB structure). Column 'total energy' contains total $\Delta\Delta G$ predicted for a tetramer (or in case of 2NAO structure - trimer, see Methods for details) of Aβ42 fibrils, whereas column 'ddG_per_monomer' contains $\Delta\Delta G$ values predicted by FoldX for a single monomer fibril of Aβ42. All energies are in kcal/mol units.

Supplementary Table 7
Description: Ratios of activation energy change $\Delta\Delta E_a$ to fibril stability change $\Delta\Delta G$ (column 'ddEa_to_ddG_ratio') for all the relevant substitutions in Aβ42 (those with $\Delta\Delta G$ satisfying the following condition: 0.6 kcal/mol < $|\Delta\Delta G|$ < 10 kcal/mol, see Methods for more details); corresponding changes in activation energy (column 'ddEa') and free energy of fibril state (column 'ddG') that were used to calculate the phi values, for all available Aβ42 structures (separately in each sheet of the table). All energies are in kcal/mol units.

Supplementary Table 8
Description: Root mean square distance to 1 across all $\Delta\Delta E_a$ /$\Delta\Delta G$ ratios (R) in APR2 region of Aβ42 (AA 29-42) ('rms_indiv_ddEa_to_ddG_ratios' column) for each PDB structure of Aβ42 used in the analysis.

Supplementary Table 9
Description: Inferred changes in activation energy and energetic couplings (column 'ddEa_scaled', in kcal/mol units) with associated outputs from MoCHI[30,31] model (fit on all the Aβ42 combinatorial mutant datasets jointly) for all the mutations and their pairwise combinations introduced in the combinatorial Aβ42 mutant datasets.

Supplementary Table 10
Description: Interaction scores (calculated as mean of absolute energetic couplings, 'mean_abs_dddEa' column) and the corresponding 3D distance between residues ('scHA_min' column) for every pair of positions ('Ab42_position_pair' column) mutagenised in Aβ42 combinatorial mutant datasets (these data are used to produce scatterplots in Fig. 4e and ED Fig. 9a).

Supplementary Table 11
Description: Complete list of oligonucleotides employed for cloning, mutagenesis and sequencing library preparation in this study.

References

1. Riek, R. & Eisenberg, D. S. The activities of amyloids from a structural perspective. *Nature* **539**, 227–235 (2016).
2. Campion, D. *et al.* Early-onset autosomal dominant Alzheimer disease: prevalence, genetic heterogeneity, and mutation spectrum. *Am. J. Hum. Genet.* **65**, 664–670 (1999).
3. O'Brien, R. J. & Wong, P. C. Amyloid precursor protein processing and Alzheimer's disease. *Annu. Rev. Neurosci.* **34**, 185–204 (2011).
4. Dobson, C. M. The Amyloid Phenomenon and Its Links with Human Disease. *Cold Spring Harb. Perspect. Biol.* **9**, (2017).
5. Fowler, D. M., Koulov, A. V., Balch, W. E. & Kelly, J. W. Functional amyloid--from bacteria to humans. *Trends Biochem. Sci.* **32**, (2007).
6. Scheres, S. H. W., Ryskeldi-Falcon, B. & Goedert, M. Molecular pathology of neurodegenerative diseases by cryo-EM of amyloids. *Nature* **621**, 701–710 (2023).
7. Lövestam, S. *et al.* Disease-specific tau filaments assemble via polymorphic intermediates. *Nature* **625**, 119–125 (2023).
8. Wilkinson, M. *et al.* Structural evolution of fibril polymorphs during amyloid assembly. *Cell* **186**, 5798–5811.e26 (2023).
9. Dobson, C. M. Protein misfolding, evolution and disease. *Trends Biochem. Sci.* **24**, 329–332 (1999).
10. Buell, A. K. The nucleation of protein aggregates - from crystals to amyloid fibrils. *Int. Rev. Cell Mol. Biol.* **329**, 187–226 (2017).
11. Baldwin, A. J. *et al.* Metastability of Native Proteins and the Phenomenon of Amyloid Formation. (2011) doi:10.1021/ja2017703.
12. Knowles, T. P. J., Vendruscolo, M. & Dobson, C. M. The amyloid state and its association with protein misfolding diseases. *Nat. Rev. Mol. Cell Biol.* **15**, 384–396 (2014).
13. Carter, P. J., Winter, G., Wilkinson, A. J. & Fersht, A. R. The use of double mutants to detect structural changes in the active site of the tyrosyl-tRNA synthetase (Bacillus stearothermophilus). *Cell* **38**, 835–840 (1984).
14. Horovitz, A. Double-mutant cycles: a powerful tool for analyzing protein structure and function. *Fold. Des.* **1**, (1996).
15. Fersht, A. R., Leatherbarrow, R. J. & Wells, T. N. C. Quantitative analysis of structure–activity relationships in engineered proteins by linear free-energy relationships. *Nature* **322**, 284–286 (1986).
16. Fersht, A. R. From covalent transition states in chemistry to noncovalent in biology: from β- to Φ-value analysis of protein folding. *Q. Rev. Biophys.* **57**, e4 (2024).
17. Winter, G., Fersht, A. R., Wilkinson, A. J., Zoller, M. & Smith, M. Redesigning enzyme structure by site-directed mutagenesis: tyrosyl tRNA synthetase and ATP binding. *Nature* **299**, 756–758 (1982).
18. Serrano, L., Kellis, J. T., Jr, Cann, P., Matouschek, A. & Fersht, A. R. The folding of an enzyme. II. Substructure of barnase and the contribution of different interactions to protein stability. *J. Mol. Biol.* **224**, 783–804 (1992).
19. Fersht, A. R. & Sato, S. Phi-value analysis and the nature of protein-folding transition states. *Proc. Natl. Acad. Sci. U. S. A.* **101**, (2004).

20. Larsen, J. A. *et al.* The mechanism of amyloid fibril growth from Φ-value analysis. *ChemRxiv* (2024) doi:10.26434/chemrxiv-2024-fxvmk.
21. Sari, L., Bali, S., Joachimiak, L. A. & Lin, M. M. Hairpin trimer transition state of amyloid fibril. *Nat. Commun.* **15**, 2756 (2024).
22. Arrhenius, S. Über die Dissociationswärme und den Einfluss der Temperatur auf den Dissociationsgrad der Elektrolyte. *Zeitschrift für Physikalische Chemie* **4U**, 96–116 (1889).
23. Arrhenius, S. Über die Reaktionsgeschwindigkeit bei der Inversion von Rohrzucker durch Säuren. *Zeitschrift für Physikalische Chemie* **4U**, 226–248 (1889).
24. Laidler, K. J. The development of the Arrhenius equation. (1984) doi:10.1021/ed061p494.
25. Thacker, D. *et al.* The role of fibril structure and surface hydrophobicity in secondary nucleation of amyloid fibrils. *Proc. Natl. Acad. Sci. U. S. A.* **117**, (2020).
26. Seuma, M., Faure, A. J., Badia, M., Lehner, B. & Bolognesi, B. The genetic landscape for amyloid beta fibril nucleation accurately discriminates familial Alzheimer's disease mutations. *Elife* **10**, (2021).
27. Seuma, M., Lehner, B. & Bolognesi, B. An atlas of amyloid aggregation: the impact of substitutions, insertions, deletions and truncations on amyloid beta fibril nucleation. *Nat. Commun.* **13**, 1–13 (2022).
28. Yang, X. *et al.* On the role of sidechain size and charge in the aggregation of A42 with familial mutations. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E5849–E5858 (2018).
29. Illes-Toth, E., Meisl, G., Rempel, D. L., Knowles, T. P. J. & Gross, M. L. Pulsed Hydrogen-Deuterium Exchange Reveals Altered Structures and Mechanisms in the Aggregation of Familial Alzheimer's Disease Mutants. *ACS Chem. Neurosci.* **12**, (2021).
30. Faure, A. J. *et al.* Mapping the energetic and allosteric landscapes of protein binding domains. *Nature* **604**, 175–183 (2022).
31. Faure, A. J. & Lehner, B. MoCHI: neural networks to fit interpretable models and quantify energies, energetic couplings, epistasis and allostery from deep mutational scanning data. *bioRxiv* 2024.01.21.575681 (2024) doi:10.1101/2024.01.21.575681.
32. Lührs, T. *et al.* 3D structure of Alzheimer's amyloid-β(1–42) fibrils. *Proceedings of the National Academy of Sciences* **102**, 17342–17347 (2005).
33. Xiao, Y. *et al.* Aβ(1–42) fibril structure illuminates self-recognition and replication of amyloid in Alzheimer's disease. *Nat. Struct. Mol. Biol.* **22**, 499–505 (2015).
34. Wälti, M. A. *et al.* Atomic-resolution structure of a disease-relevant Aβ(1–42) amyloid fibril. *Proceedings of the National Academy of Sciences* **113**, E4976–E4984 (2016).
35. Colvin, M. T. *et al.* Atomic Resolution Structure of Monomorphic Aβ42 Amyloid Fibrils. (2016) doi:10.1021/jacs.6b05129.
36. Gremer, L. *et al.* Fibril structure of amyloid-β(1–42) by cryo–electron microscopy. *Science* (2017) doi:10.1126/science.aao2825.
37. Yang, Y. *et al.* Cryo-EM structures of amyloid-β 42 filaments from human brains. *Science* **375**, 167–172 (2022).
38. Lee, M., Yau, W.-M., Louis, J. M. & Tycko, R. Structures of brain-derived 42-residue amyloid-β fibril polymorphs with unusual molecular conformations and intermolecular interactions. *Proc. Natl. Acad. Sci. U. S. A.* **120**, e2218831120 (2023).
39. Fernandez-Escamilla, A.-M., Rousseau, F., Schymkowitz, J. & Serrano, L. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat. Biotechnol.* **22**, 1302–1306 (2004).
40. Goldschmidt, L., Teng, P. K., Riek, R. & Eisenberg, D. Identifying the amylome, proteins

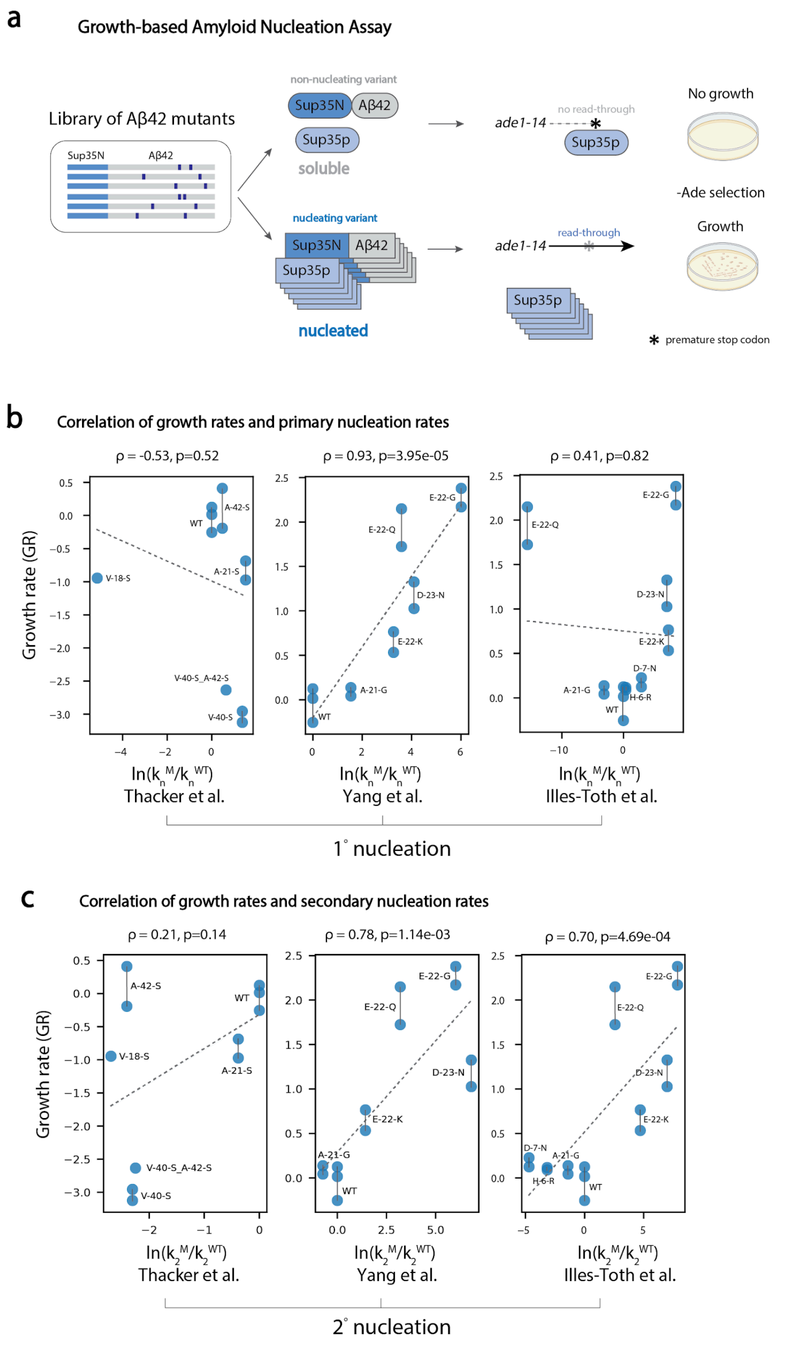
capable of forming amyloid-like fibrils. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 3487–3492 (2010).

41. van der Kant, R., Louros, N., Schymkowitz, J. & Rousseau, F. Thermodynamic analysis of amyloid fibril structures reveals a common framework for stability in amyloid polymorphs. *Structure* (2022) doi:10.1016/j.str.2022.05.002.
42. Fersht, A. R., Leatherbarrow, R. J. & Wells, T. N. C. Quantitative analysis of structure–activity relationships in engineered proteins by linear free-energy relationships. *Nature* **322**, 284–286 (1986).
43. Fersht, A. R., Leatherbarrow, R. J. & Wells, T. N. C. Structure-activity relationships in engineered proteins: analysis of use of binding energy by linear free energy relationships. (2002) doi:10.1021/bi00393a013.
44. Matouschek, A., Kellis, J. T., Serrano, L. & Fersht, A. R. Mapping the transition state and pathway of protein folding by protein engineering. *Nature* **340**, 122–126 (1989).
45. Fersht, A. R., Matouschek, A. & Serrano, L. The folding of an enzyme. I. Theory of protein engineering analysis of stability and pathway of protein folding. *J. Mol. Biol.* **224**, (1992).
46. Fersht, A. R. Optimization of rates of protein folding: the nucleation-condensation mechanism and its implications. *Proc. Natl. Acad. Sci. U. S. A.* **92**, 10869 (1995).
47. Schmiedel, J. M. & Lehner, B. Determining protein structures using deep mutagenesis. *Nat. Genet.* **51**, 1177–1186 (2019).
48. Rollins, N. J. *et al.* Inferring protein 3D structure from deep mutation scans. *Nat. Genet.* **51**, 1170–1176 (2019).
49. Salinas, V. H. & Ranganathan, R. Coevolution-based inference of amino acid interactions underlying protein function. (2018) doi:10.7554/eLife.34300.
50. Tsuboyama, K. *et al.* Mega-scale experimental analysis of protein folding stability in biology and design. *Nature* **620**, 434–444 (2023).
51. Beltran, A., Jiang, X. ’er, Shen, Y. & Lehner, B. Site saturation mutagenesis of 500 human protein domains reveals the contribution of protein destabilization to genetic disease. *bioRxiv* 2024.04.26.591310 (2024) doi:10.1101/2024.04.26.591310.
52. Markin, C. J. *et al.* Revealing enzyme functional architecture via high-throughput microfluidic enzyme kinetics. *Science* **373**, (2021).
53. Neun, S., van Vliet, L., Hollfelder, F. & Gielen, F. High-Throughput Steady-State Enzyme Kinetics Measured in a Parallel Droplet Generation and Absorbance Detection Platform. *Anal. Chem.* **94**, 16701–16710 (2022).
54. Ma, F. *et al.* Efficient molecular evolution to generate enantioselective enzymes using a dual-channel microfluidic droplet screening platform. *Nat. Commun.* **9**, 1–8 (2018).
55. Brenman, J. E. Lecanemab in Early Alzheimer’s Disease. *The New England journal of medicine* vol. 388 1631 (2023).
56. Schneider, L. A resurrection of aducanumab for Alzheimer’s disease. *Lancet Neurol.* **19**, 111–112 (2020).
57. Mighell, T. L., Toledano, I. & Lehner, B. SUNi mutagenesis: Scalable and uniform nicking for efficient generation of variant libraries. *PLoS One* **18**, e0288158 (2023).
58. Chandramowlishwaran, P. *et al.* Mammalian amyloidogenic proteins promote prion nucleation in yeast. *J. Biol. Chem.* **293**, 3436–3450 (2018).
59. Faure, A. J., Schmiedel, J. M., Baeza-Centurion, P. & Lehner, B. DiMSum: an error model and pipeline for analyzing deep mutational scanning data and diagnosing common experimental pathologies. *Genome Biol.* **21**, 207 (2020).
60. Thacker, D. *et al.* The role of fibril structure and surface hydrophobicity in secondary

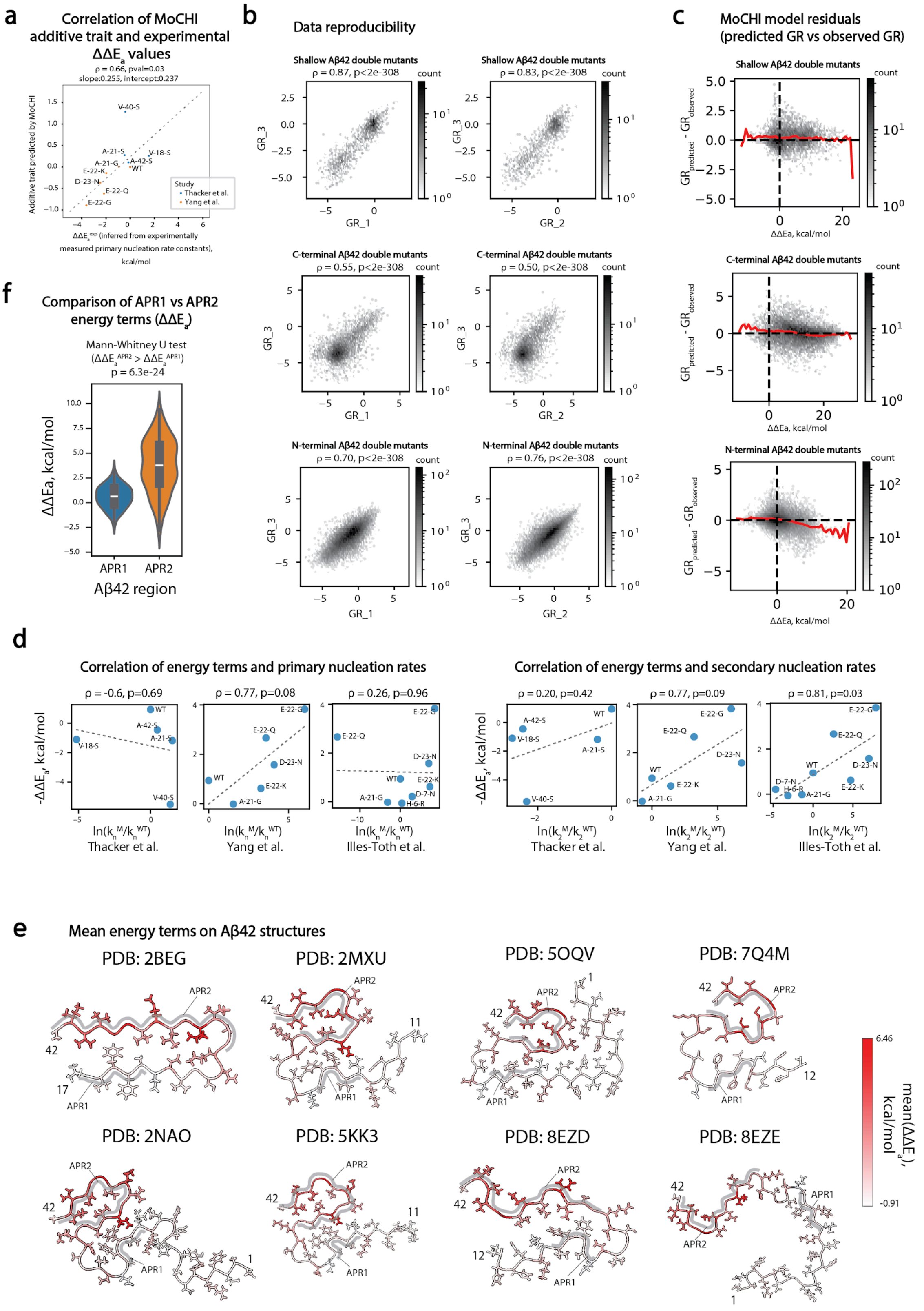
nucleation of amyloid fibrils. *Proc. Natl. Acad. Sci. U. S. A.* **117**, (2020).

61. Meng, E. C. *et al.* UCSF ChimeraX: Tools for structure building and analysis. *Protein Sci.* **32**, e4792 (2023).
62. Thacker, D. *et al.* The role of fibril structure and surface hydrophobicity in secondary nucleation of amyloid fibrils. *Proc. Natl. Acad. Sci. U. S. A.* **117**, (2020).

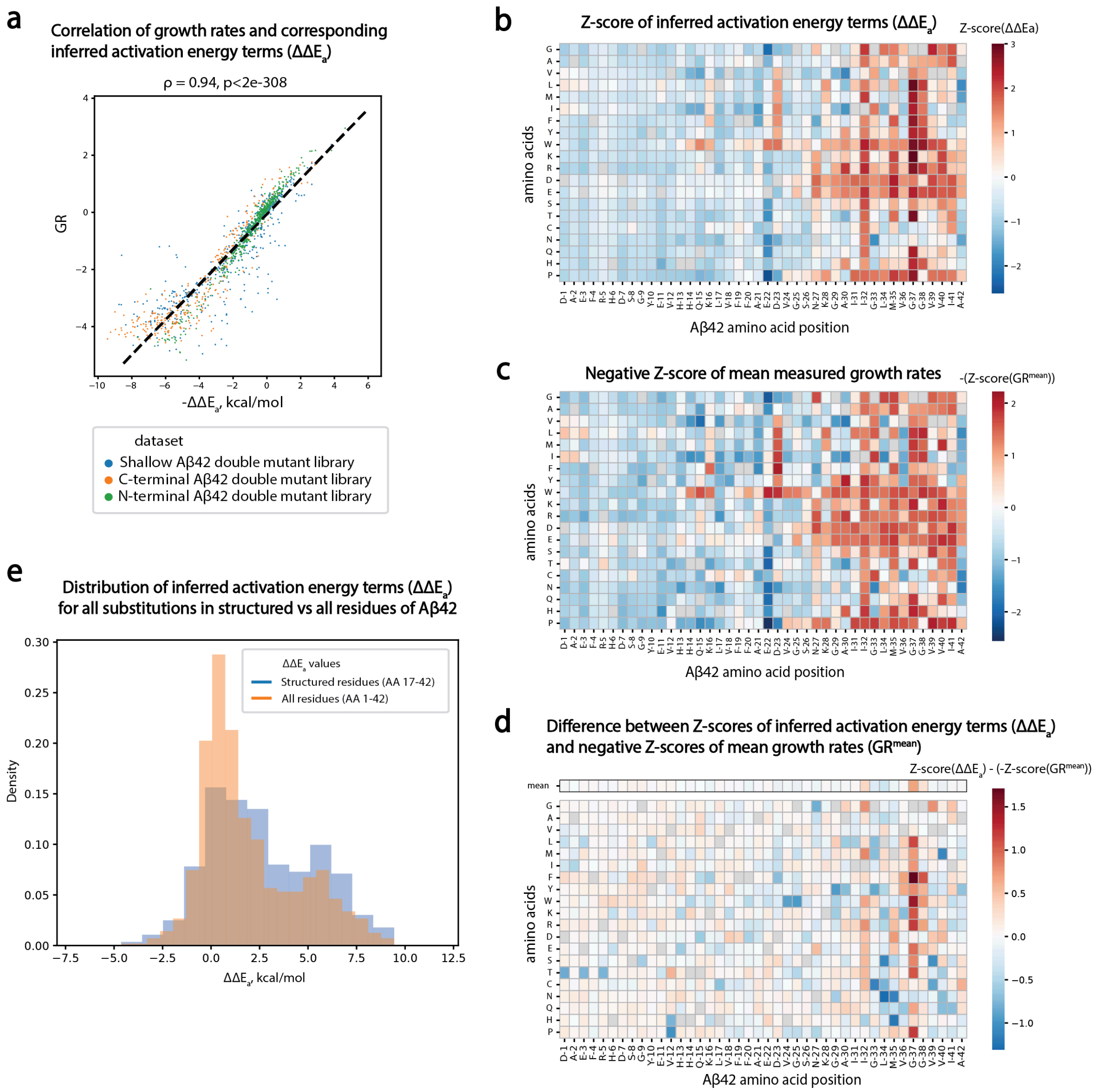
Extended Data Figures



Extended Data Figure 1. High throughput quantification of amyloid nucleation kinetics. **a**, Schematic overview of the amyloid nucleation assay: Aβ42, fused to the nucleation domain of Sup35 (Sup35N), seeds aggregation of the yeast prion Sup35p, causing read-through of a premature stop codon in the *ade1* reporter gene, thus allowing growth in medium lacking adenine. **b-c**, Correlation of growth rates (GR) with logarithm of primary (b) and secondary (c) nucleation rate constants from *in vitro* measurements[28,6229]. Spearman's ρ (correlation) coefficients and associated p-values are reported. Data points for repeated measurements (mutations for which GR was measured with the amyloid nucleation assay in more than one dataset) are connected with a solid grey vertical line. Dashed grey lines represent linear regression fit for the data.

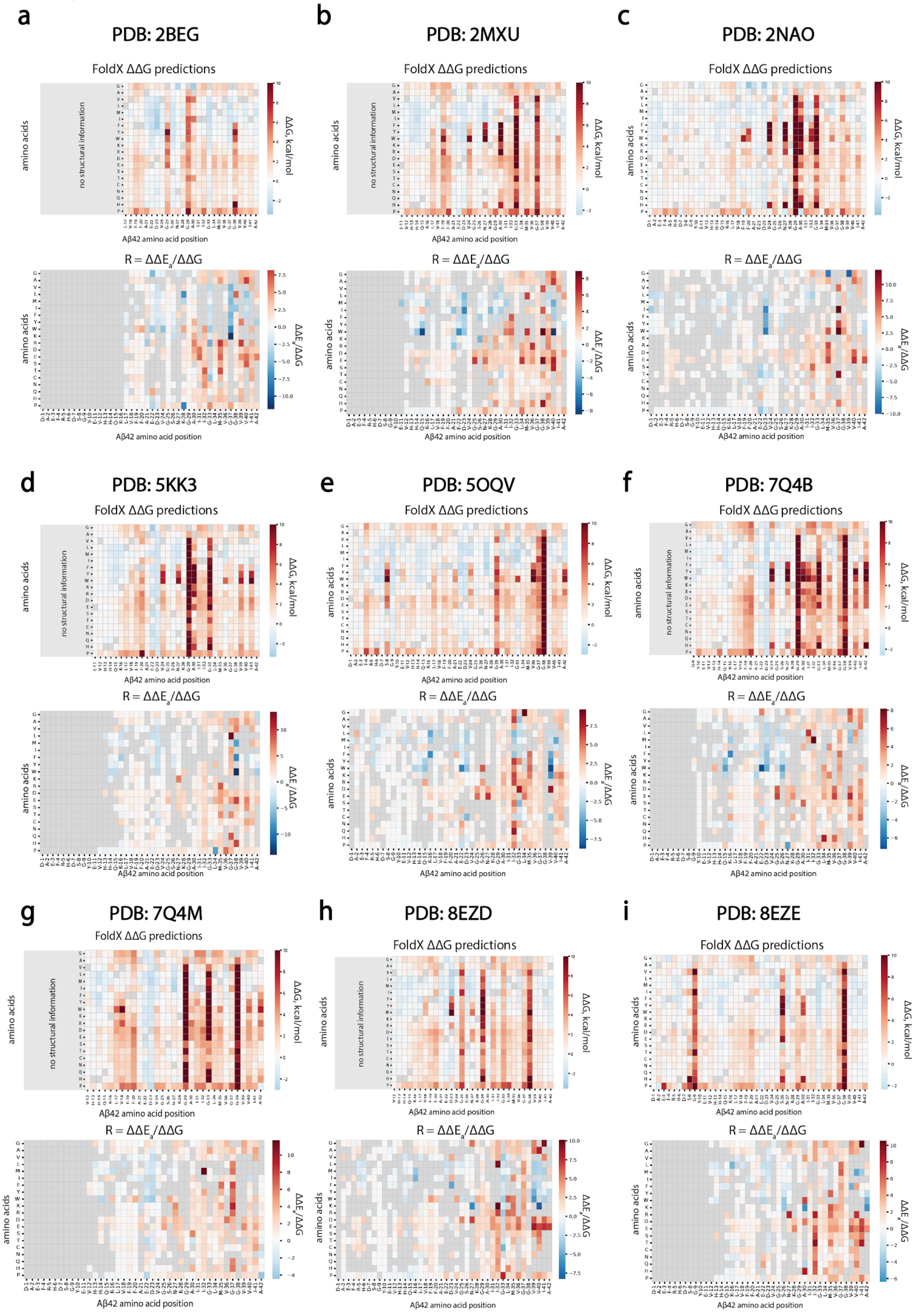


Extended Data Figure 2. **Overview of double mutant libraries and inferred activation energy changes**. **a**, Scatterplot of additive trait values predicted by MoCHI[31] model trained on double mutant datasets (Y axis) and experimentally-derived $\Delta\Delta E_a$ values (X axis, using primary nucleation rate constants). Dashed grey line represents linear regression fit for the data. **b**, Replicate growth rates correlations for shallow, C-terminal and N-terminal Aβ42 double mutant libraries (from top to bottom). Spearman's ρ (correlation) coefficients and associated p-values are reported. **c**, MoCHI residuals (predicted vs observed growth rates, red line is following mean residuals in 50 equally spaced bins across x axis, dashed black lines indicate 0 in both axes) for shallow, C-terminal and N-terminal Aβ42 double mutant libraries (from top to bottom). **d**, Correlation of the inferred activation energy terms ($-\Delta\Delta E_a$) with the logarithm of Aβ42 primary (three panels on the left) and secondary (three panels on the right) nucleation rate constants measured in independent studies[25,28,29]; Spearman's ρ (correlation) coefficients and associated p-values are reported. Dashed grey lines represent linear regression fit for the data. **e**, Cross sections of 2BEG, 2MXU, 5OQV, 7Q4B, 2NAO, 5KK3, 8EZD and 8EZE PDB structures of Aβ42 fibrils coloured by mean $\Delta\Delta E_a$ per position. Aggregation prone regions 1 (APR1) and 2 (APR2) are highlighted in grey. **f**, Violinplot comparing inferred activation energy terms ($\Delta\Delta E_a$) between APR1 region (AA 17-21) and APR2 region (AA 29-42) of Aβ42 (p=6.3e-24, one-sided Mann-Whitney U test ($\Delta\Delta E_a^{APR2} > \Delta\Delta E_a^{APR1}$)).

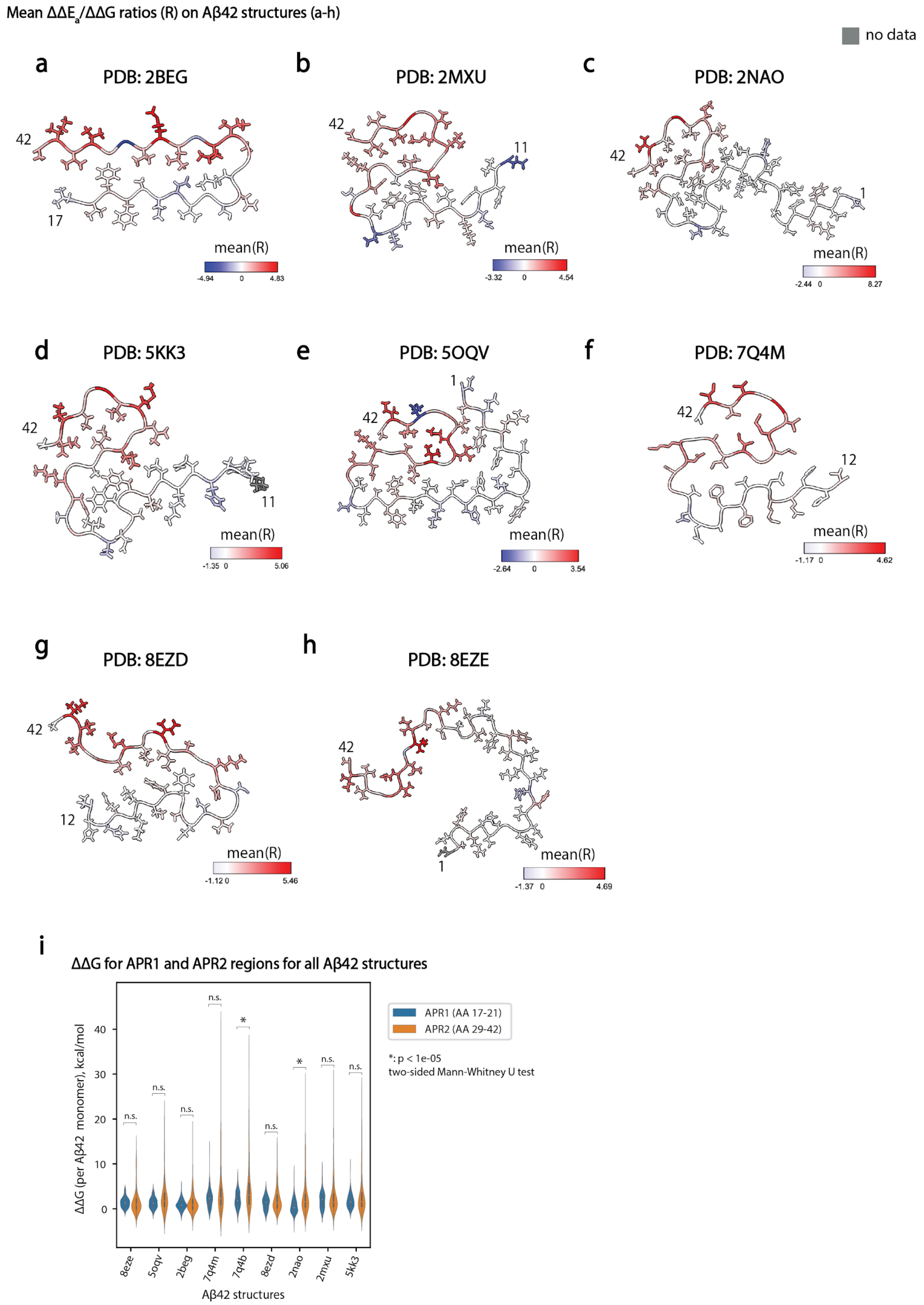


Extended Data Figure 3. **Comparison of activation energies and cellular growth rates** **a**, Correlation of growth rates of shallow, C-terminal and N-terminal Aβ42 double mutants and corresponding inferred activation energy terms ($\Delta\Delta E_a$). Spearman's ρ (correlation) coefficients and associated p-values are reported. Dashed black line represents linear regression fit to the data. **b**, Z-scores of inferred activation energy terms ($\Delta\Delta E_a$) (scaled to zero mean and unit variance). **c**, Negative Z-scores of mean (across all double mutant datasets) growth rates (GR^{mean}) (scaled to zero mean and unit variance). **d**, Difference between Z-scores of inferred activation energy terms ($\Delta\Delta E_a$) and negative Z-scores of mean growth rates (GR^{mean}). Means of difference values for each position are displayed in the top row of the heatmap (outlined in black). **e**, Distribution of inferred activation energy terms ($\Delta\Delta E_a$) for all substitutions across either all positions in Aβ42 (AA 1-42, in orange) or only those Aβ42 positions that are structurally resolved in all available fibril structures (AA 17-42, in blue).

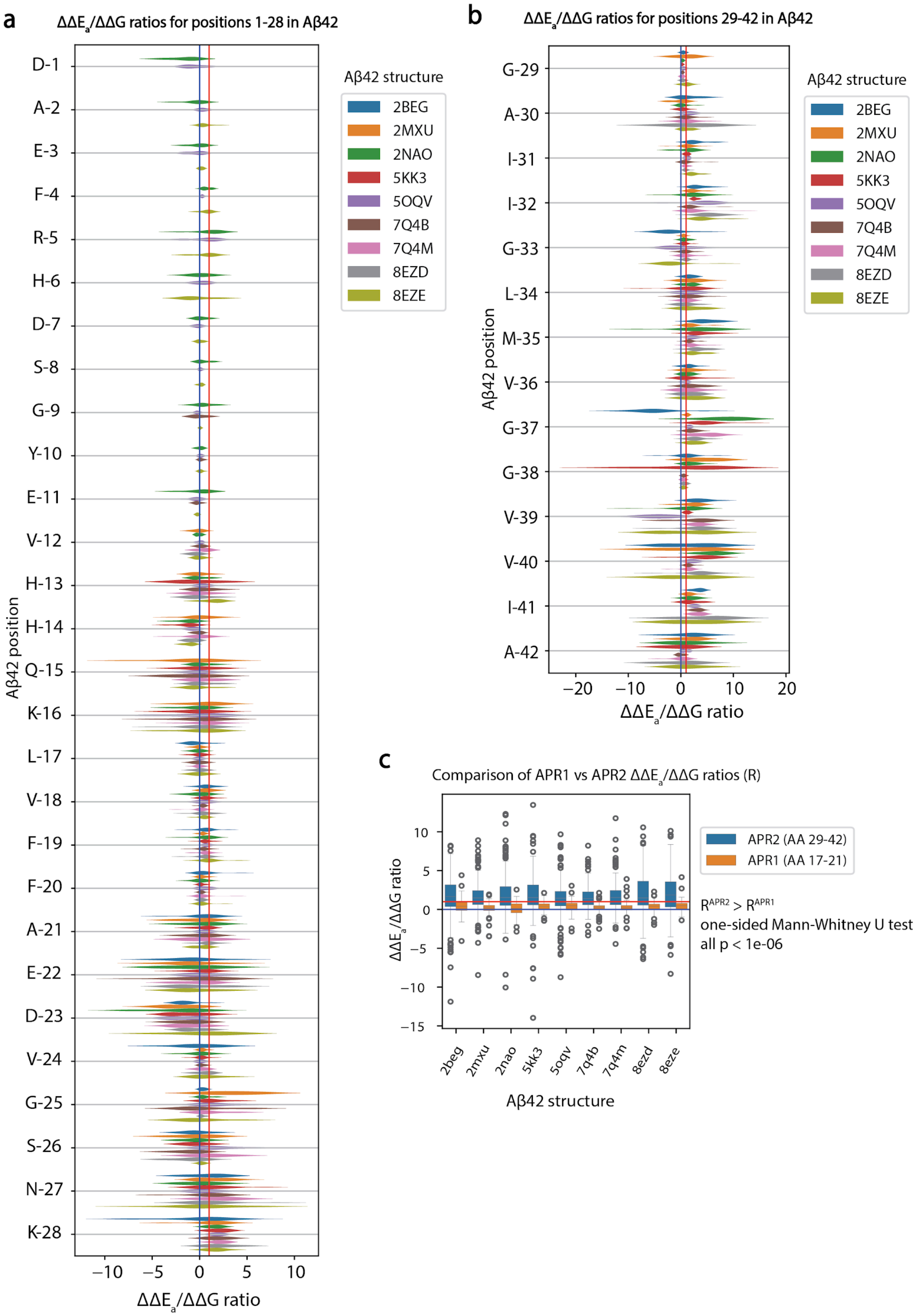
Fibril stability changes (ΔΔG per Aβ42 monomer) predicted with FoldX (top panels) and $\Delta\Delta E_a/\Delta\Delta G$ ratios (bottom panels) for all Aβ42 structures



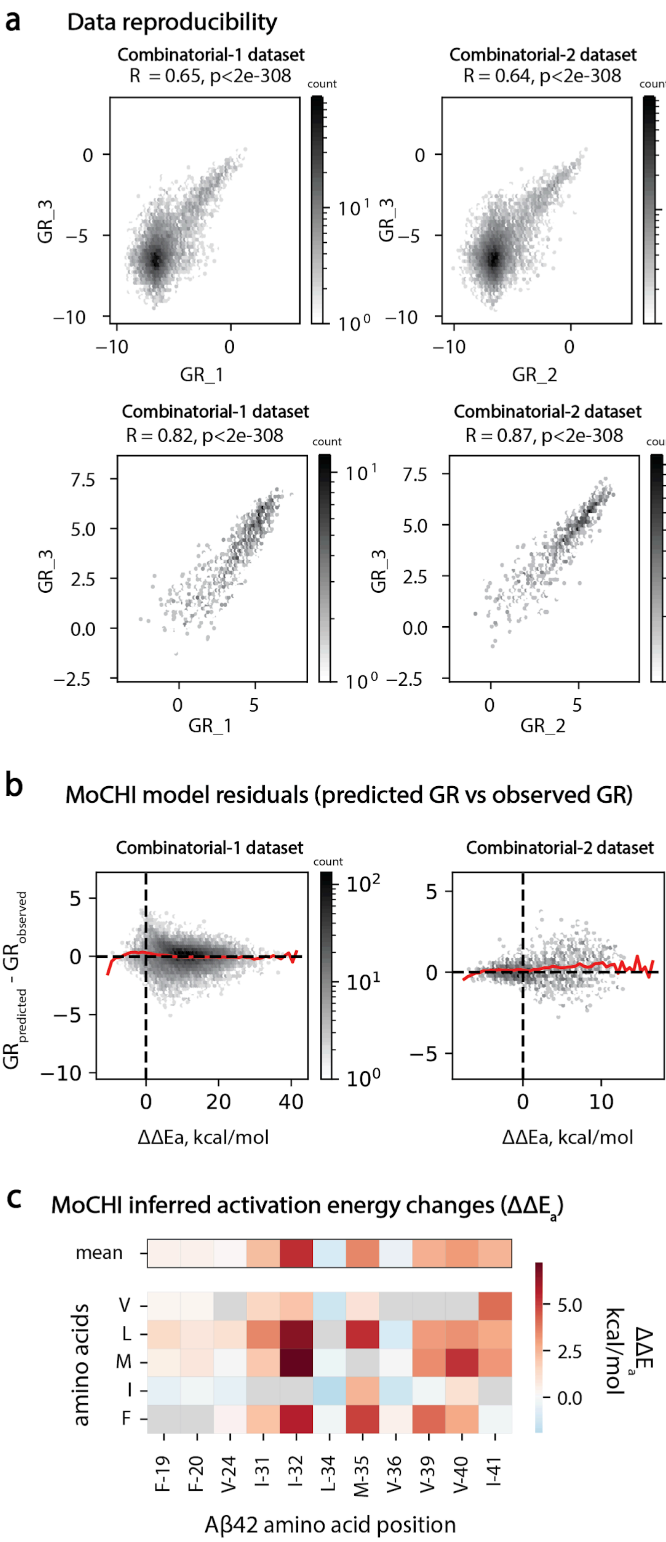
Extended Data Figure 4. **Fibril stability and $\Delta\Delta E_a/\Delta\Delta G$ ratios (R) for all Aβ42 structures**. **a-i**, (top panels) Fibril stability changes ($\Delta\Delta G$ per Aβ42 monomer) predicted with FoldX upon all possible substitutions and (bottom panels) R-values for structures 2BEG (a), 2MXU (b), 2NAO (c), 5KK3 (d), 5OQV (e), 7Q4B (f), 7Q4M; (g) 8EZD (h), 8EZE (i). FoldX was run using a stack of four Aβ42 monomers for all structures, apart from 2NAO where a stacked trimer was used (see Methods).



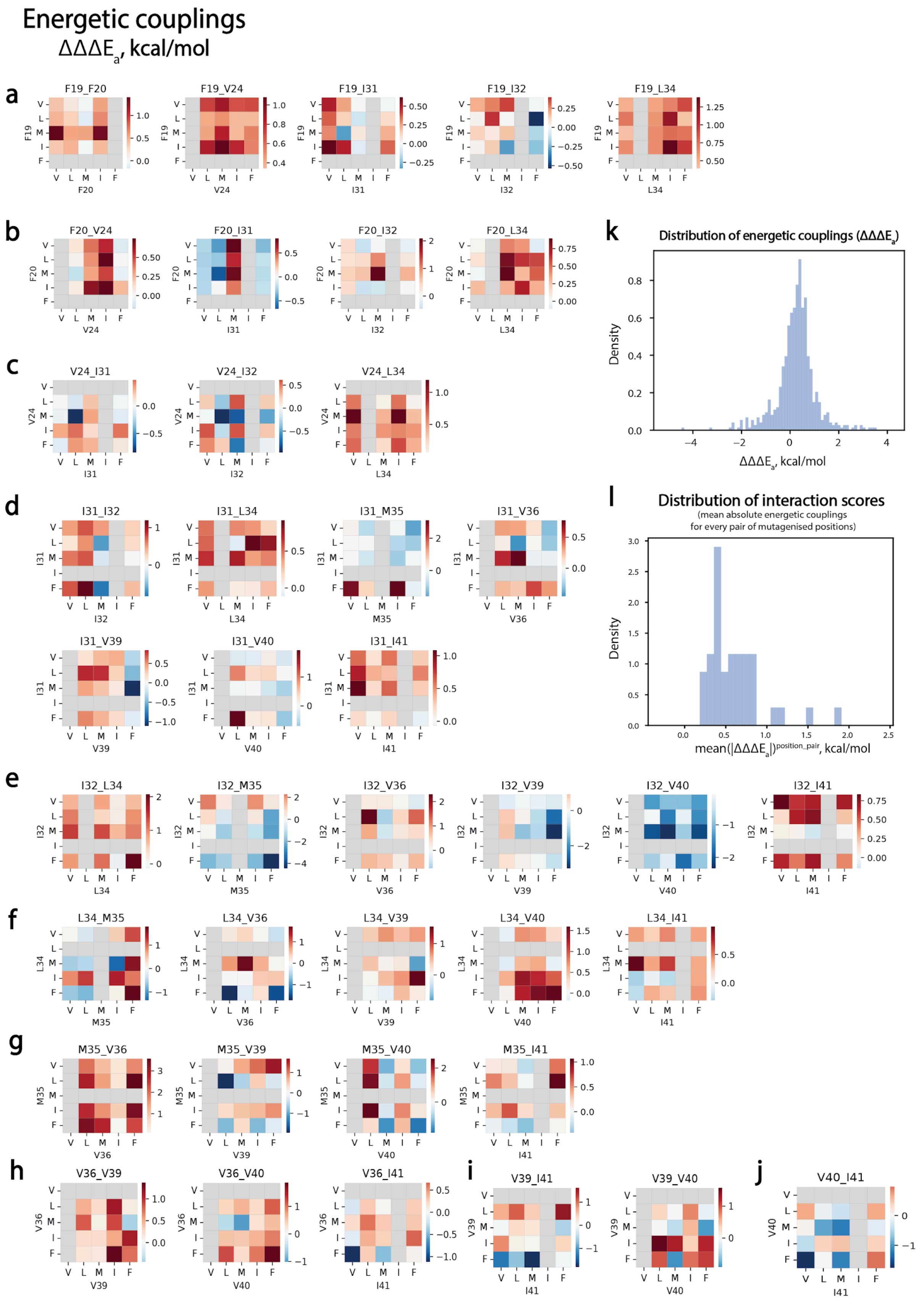
Extended Data Figure 5. **Mean ΔΔE$_a$/ΔΔG ratios (R) on all Aβ42 structures**. **a-h**, Cross sections of PDB structures of Aβ42 fibrils coloured by mean R-values at that position: 2BEG (a), 2MXU (b), 2NAO (c), 5KK3 (d), 5OQV (e), 7Q4B (f), 8EZD (g), 8EZE (h). Residues with no R-values are coloured in grey. **i**, Violin plot comparing fibril stability changes (ΔΔG per Aβ42 monomer predicted with FoldX) for all possible substitutions in APR1 vs APR2 regions of Aβ42, for all available Aβ42 fibril structures (one-sided Mann-Whitney U test statistics are reported).



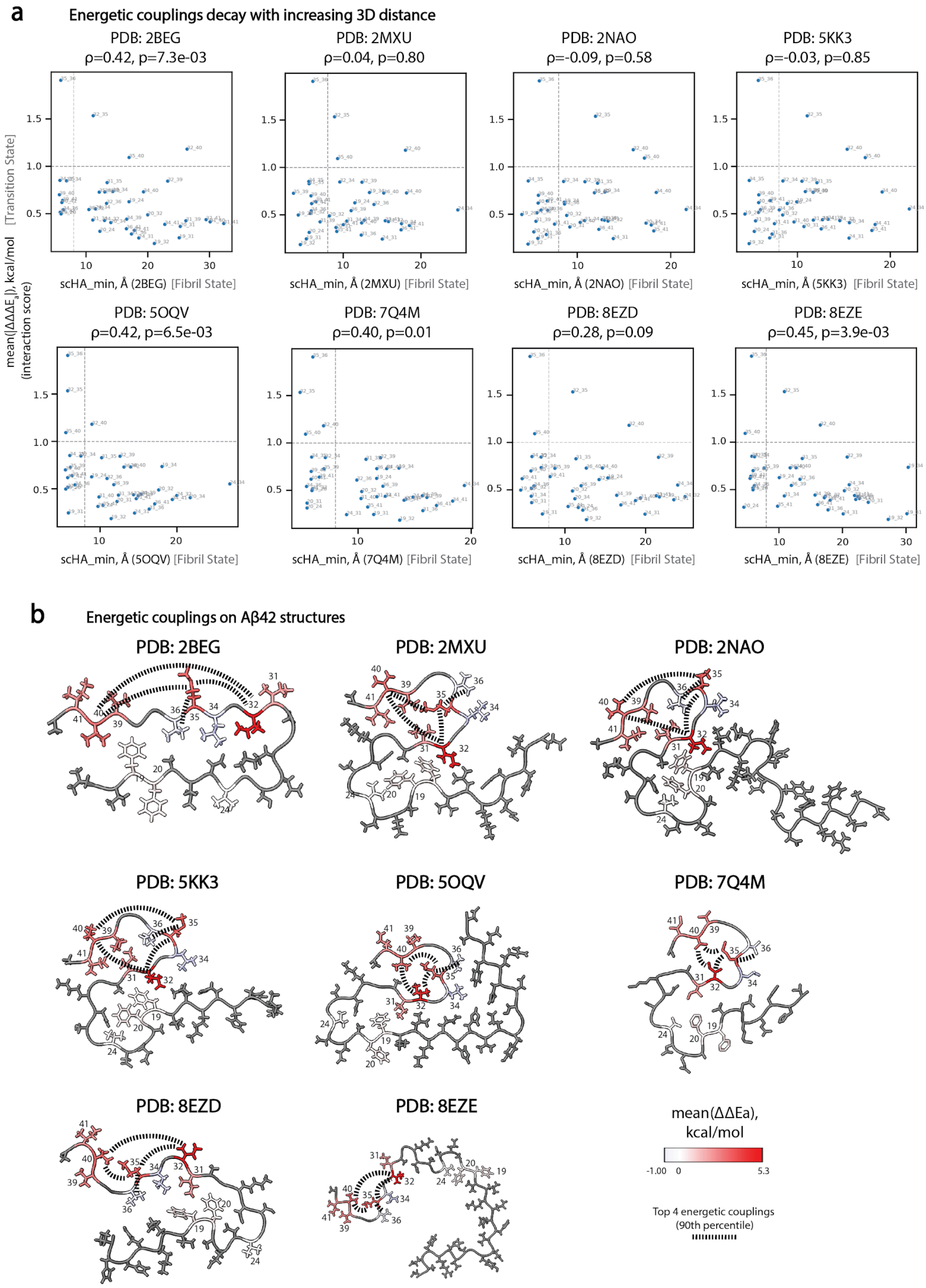
Extended Data Figure 6. Overview of all R-values and comparison of APR1 and APR2 R-values. **a-b**, Violinplot of R-values for all the Aβ42 structures (2BEG, 2MXU, 2NAO, 5KK3, 5OQV, 7Q4B, 7Q4M, 8EZD, 8EZE) for each position at the (a) N-terminus (AA 1-27) and (b) C-terminus (AA 28-42). Red vertical lines mark R = 1, and blue vertical lines mark R = 0. **c**, Boxplots comparing APR1 region (AA 17-21) and APR2 region (AA 29-42) R-values for all the Aβ42 structures (2BEG, 2MXU, 2NAO, 5KK3, 5OQV, 7Q4B, 7Q4M, 8EZD, 8EZE). One-sided Mann-Whitney U test statistics are reported ($p<1e-06$ for all comparisons).



Extended Data Figure 7. **Overview of combinatorial mutant libraries and energetic couplings**. **a**, Replicate growth rates correlations for Combinatorial-1 (top panels) and Combinatorial-2 (bottom panels) combinatorial mutant libraries. Pearson's correlation coefficients (R) and associated p-values are indicated. **b**, MoCHI residuals (predicted vs observed growth rates, red line is following mean residuals in 50 equally spaced bins across x axis, dashed black lines indicate 0 in both axes) for Combinatorial-1 (left) and Combinatorial-2 (right) combinatorial mutant libraries. **c**, Heatmap displaying the inferred activation energy changes ($\Delta\Delta E_a$) for all Aβ42 substitutions present in combinatorial mutagenesis datasets (Combinatorial-1, Combinatorial-2). Mean $\Delta\Delta E_a$ values for each position are displayed in the top row of the heatmap (outlined in black).



Extended Data Figure 8. Energetic coupling landscape of Aβ42. a-j, Heatmaps displaying the energetic couplings between all individual mutations introduced in the combinatorial mutagenesis datasets (Combinatorial-1, Combinatorial-2) for each pair of mutated positions (F-19 (a), F-20 (b), V-24 (c), I-31(d), I-32 (e), L-34 (f), M-35 (g), V-36 (h), V-39 (i), V-40 (j)). **k**, Distribution of 640 energetic couplings ($\Delta\Delta\Delta E_a$) between mutations introduced in combinational Aβ42 mutant datasets. **l**, Distribution of 40 interaction scores (mean $|\Delta\Delta\Delta E_a|$) for pairs of positions mutagenised in combinatorial Aβ42 mutant datasets.



Extended Data Figure 9. Energetic couplings compared to Aβ42 fibril polymorphs. a, Scatterplots of interaction scores for pairs of positions and the inter-residue distance for corresponding pairs of amino acids in 3D space (scHA_min, minimum heavy atom side chain distance) of Aβ42 structures 2BEG, 2MXU, 2NAO, 5KK3, 5OQV, 7Q4M, 8EZD, 8EZE; dashed light grey vertical line marks 8 Å, dashed light grey horizontal line marks interaction score (mean($|\Delta\Delta\Delta E_a|$) of 1 kcal/mol. **b**, Cross sections of PDB structures (2BEG, 2MXU, 2NAO, 5KK3, 5OQV, 7Q4M, 8EZD, 8EZE) of Aβ42 fibrils coloured by mean inferred activation energy terms (ΔΔEa) from the MoCHI model trained on combinatorial mutants datasets (residues with no inferred ΔΔEa values are in grey). Positions mutagenised in combinatorial datasets (Combinatorial-1 and Combinatorial-2) are labelled on the PDB structure. Top 4 interacting position pairs (in 90th percentile by their interaction scores) are connected with dashed black lines.