Supplemental Methods

Normalisation of mRNA expression levels

Caenorhabditis elegans expression database (Golden et al.[1]). Expression data were reported for individual worms at different ages. Expression levels were listed with a variety of ID types, which were converted to reviewed UniProt IDs using UniProt[2] and DAVID[3]. We deleted entries for which a gene name pointed to multiple UniProt IDs and averaged all values for a given UniProt ID. Subsequently, we converted this average from $\log_2$ to $\log_{10}$, to give $\log_{10} C$, where C is concentration (See Equation $S3$). There were 22,490 values (various IDs) in the original data set. After conversion, there were 16,623 UniProt IDs.

BioGPS human expression database[4]. Expression data were reported for microarrays performed on a series of human tissues. For our analysis, most cancer and cell line expression levels were excluded. Expression levels were listed with Affymetrix and GNF probeset IDs. Affymetrix IDs were converted to reviewed UniProt ID using NetAffx[5] and DAVID. GNF probeset IDs were listed with corresponding IDs of one or more of the following kind: RefSeq[6], UniGene[7], Entrez Gene[8], ENSEMBL[9], and gene symbol[7]. Conversions to corresponding UniProt proteins were made from GNF probeset ID using DAVID, UniProt, and UniGene. Instances in which a given probeset ID mapped to multiple UniProt IDs were excluded from further analysis. Each expression value was $\log_{10}$-transformed. For each UniProt ID, transformed expression values across tissues and across probes were pooled and averaged, to give $\log_{10} C$, where C is concentration (See Equation $S3$). There were 44,775 values (probe set IDs) in the original data set. After conversion, there were 16,623 UniProt IDs (reviewed).

Normalisation of protein abundances

The human and *C. elegans* integrated data sets were downloaded from PaxDb[10] to obtain spectral count abundance data. We converted ENSEMBL protein IDs to UniProt IDs and the length of the corresponding proteins was obtained from UniProt. The Normalized Spectral Abundance Factor was calculated by dividing the spectral count by the protein length. We $\log_{10}$-transformed these values to give $\log_{10} C$, where C is concentration (See Equation $S3$). There were 12,803 values (ENSEMBL Protein) in the human data set and 11,719 values (various IDs) in the *C.*

elegans data set. This procedure resulted in 10,247 UniProt IDs (reviewed) in the human data set and 11,069 UniProt IDs in the *C. elegans* data set.

Aggregation propensity scores

For the proteome sets for *Homo sapiens* and *C. elegans* we calculated the Z_{agg} [11], Z_{agg}^{SC} [11], and TANGO[12] scores as previously described[11,12]; for the TANGO scores, we removed 0 values. In brief, the Z_{agg}^{SC} scores were calculated by combining the Z_{agg} and CamP[13] scores (http://www-vendruscolo.ch.cam.ac.uk/zyggregator.php). We first calculated the structurally-corrected aggregation propensity

$$Z^{pf} = \sum_{i=1}^{l} Z_{agg} \left(1 - {pf_i}/{pf_{max}}\right) \qquad (S1)$$

where the pf_i values are calculated using the CamP algorithm[13], and pf_{max} is the largest one among the pf_i values. The intrinsic aggregation propensity, Z_{agg}, for each residue *i* in a protein of length *l*, is estimated with the Zyggregator method[11]. The scores were normalized to an average of 0 and variance of 1 by rescaling them as

$$Z_{agg}^{SC} = \alpha + \beta Z^{pf} \qquad (S2)$$

The method was validated with *E. coli* solubility data[14,15].

Definition of the supersaturation score

The supersaturation score σ is calculated as the sum

$$\sigma = C + Z \qquad (S3)$$

where C is the $\log_{10}$ of the concentration and Z is the aggregation propensity score; the concentrations are derived from the mRNA expression or protein abundance levels (see Supplementary Tables). Our human σ_u database included 16,293 proteins; our σ_f database

included 6,155 proteins; and our σ_{u_T} database included 16,054 proteins (**Table S2**). These values were recentered such that the median σ score for each database was 0.

Protein data sets

We describe here the sources of the data and ID conversions made when assembling data sets for this study. In general, ID conversions were performed using UniProt or DAVID. For human protein data sets, only reviewed UniProt IDs were included, whereas for *C. elegans*, all UniProt IDs were considered. We followed this procedure because the coverage of review for the *C. elegans* proteome is substantially lower than that for the human proteome. In cases (such as RefSeq) in which an ID includes splicing variant information, this information was included in ID conversions initially. If this failed to result in an ID conversion, the splicing variant information was removed, because we used only canonical protein sequences for aggregation propensity prediction in this study. In many cases, conversions from multiple ID types were necessary. Unlike for the expression data, duplicate entries did not disqualify a protein from inclusion, but the database was cleansed of repeated entries of a given protein. In cases where a given non-UniProt ID mapped to multiple UniProt IDs, both were included in the data set. **Table S1** provides information about the number of proteins in the full set and the number in our supersaturation databases.

Amyloids. Amyloid proteins were identified by a search of the UniProt database[2] for reviewed human proteins with the keyword "KW-0034," corresponding to 27 unique, reviewed UniProt IDs.

Plaques. 26 proteins were reported[16] as enriched in amyloid plaques isolated by laser capture microdissection (LCM) and analyzed using mass spectrometry, relative to non-plaque control areas. We converted the names of these proteins from RefSeq ID to UniProt ID, corresponding to 26 unique, reviewed UniProt IDs.

Neurofibrillar tangles (NFTs). 72 proteins were reported[17] as present in neurofibrillar tangles isolated by laser capture microdissection (LCM) and analyzed using mass spectrometry. We

converted the names of these proteins from International Protein Index[18] (IPI) identifier to UniProt ID, corresponding to 88 unique, reviewed UniProt IDs.

Lewy bodies. 36 proteins were reported[19] as significantly enriched ($p < 0.01$) in Lewy bodies purified from patients with the Lewy body variant of Alzheimer's disease and analyzed by mass spectrometry, relative to controls from patients without Lewy body pathology. We converted the names of these proteins from RefSeq ID to UniProt ID, corresponding to 34 unique, reviewed UniProt IDs. In the lysate, 707 proteins were identified in total, corresponding to 557 unique, reviewed UniProt IDs.

Artificial ß peptides[20] (Artificial ß). We considered 133 proteins that were reported as enriched in studies that identified by using mass spectrometry interactors with designed ß sheet peptides that form amyloid-like aggregates, relative to a whole cell lysate[20]. We converted the names of these proteins from IPI identifier or Gene Name to UniProt ID using UniProt or DAVID, corresponding to 151 unique, reviewed UniProt IDs. 3,055 proteins were identified in the lysate in this study, corresponding to to 3,032 unique, reviewed UniProt IDs after conversion.

Worm ageing. In a recent study, 720 proteins were reportedas aggregating in aged *C. elegans* relative to younger worms, as determined from differential centrifugation and mass spectrometry[21]. We converted the names of these proteins from UniProt Accession (AC) to UniProt ID, corresponding to 719 unique UniProt IDs. In another study, 203 proteins were reported[22] (in UniProt IDs) as aggregating in aged *C. elegans* relative to younger worms, as determined from differential centrifugation and mass spectrometry. The combined set included a total of 761 unique UniProt IDs.

Complexes. A database of complexes was downloaded from the Molecular Interaction (MINT) database[23]. This list was a compilation of experimental data and computational predictions using the spoke expansion. Only those pairwise combinations in which both components had a human taxonomic ID were used. This procedure resulted in 1,729 unique identifiers. RefSeq IDs were converted to UniProt IDs. The 3 ENSEMBL IDs did not map to UniProt IDs. This resulted in 1,637 unique, reviewed UniProt IDs. The list of complex proteins consisted only of those human

proteins from the MINT database also found in our σ_f database, on the reasoning that the supersaturation score of complex components was only relevant in terms of the prediction for the folded components. This method does not take into account the full structure of a complex, but only predictions of the hydrogen protection factors for the individual protein components of the complex.

Nuclear complexes. Nuclear complex proteins were identified using the Database of Nuclear Protein Complexes (PINdb)[24]. Protein names, of which there were 604 unique ones, were converted to UniProt IDs using DAVID. This resulted in 630 unique, reviewed UniProt IDs. As above, the list of nuclear complex proteins consisted only of those human proteins from the PINdb database also found in our σ_f database.

Nuclear proteome. For the purposes of this study, the nuclear proteome data set included those proteins from our σ_f set (since the nuclear proteome was only used as a control for the nuclear complexes data) that were included in the Gene Ontology[25] (GO) term "nucleus." This resulted in 1,942 UniProt IDs.

Calculation of the fold changes

The fold change is calculated as the linear difference between the logarithmic medians of two sets. The linear fold difference d between the medians of the supersaturation scores of the control set C and experiment set E being tested

$$d = 10^{median(E) - median(C)} \qquad (S4)$$

Statistical tests for distributions

To assess the difference in distributions of σ scores between various data sets, we used the non-parametric Wilcoxon/Mann-Whitney U test[26]. Non-parametric tests such as this one provide robust statistical comparisons when samples are not derived from a normal distribution, as is the case for our data. These tests were performed using R and the SciPy package for Python.

Identification of enriched KEGG[27] pathways

We used the DAVID bioinformatics software[3] to identify enriched KEGG pathways for proteins at or above the 95th percentile of our supersaturation databases. To determine enrichment, we used DAVID to perform a conservative variant of the Fisher Exact Test to generate the EASE score[28] for enrichment of this set of top supersaturation proteins relative to the proteins in the database as a whole, for all KEGG pathways. DAVID also computes Bonferroni-corrected p-values, which represent the most conservative correction for multiple hypothesis testing. Further information on our multiple hypothesis testing regime can be found below.

To identify enriched pathways for low expression proteins, we first considered the half of our database with the lowest expression. We ranked these based on σ_u score and selected proteins at or above the 95th percentile. We then performed the above statistical tests to compare these 406 proteins to the 16,263 proteins in the full σ_u database.

Gaussian noise generation

Fifty noise levels were defined by varying, between $\log_{10}(1.1)$ and $\log_{10}(500)$, the standard deviation of Gaussian distributions centered at 0. At each noise level l, we performed 100 trials. For each trial t, a random number $n_{l,t,p}$ was drawn from that noise level's distribution for each of the p proteins in the database. The noise-introduced supersaturation score $\sigma_{p,l,t}$ was defined as

$$\sigma_{p,l,t} = \sigma_p + n_{l,t,p} \qquad (S5)$$

For trial t of noise level l, the set $S_{l,t}$ of noise values is

$$S_{l,t} = \{n_{l,t,1}, n_{l,t,2}, \dots, n_{l,t,p}\} \qquad (S6)$$

The set $m_{l,t}$ of linear magnitudes of noise for trial t of noise level l is

$$m_{l,t} = \{10^{n_{l,t,1}}, 10^{n_{l,t,2}}, \dots, 10^{n_{l,t,p}}\} \qquad (S7)$$

For noise level l, the set M_l of median noise values for its constituent trials is

$$M_l = \{median(m_{l,1}), median(m_{l,2}), \ldots, median(m_{l,100})\} \quad (S8)$$

In each Gaussian noise plot, the values plotted on the x-axis were the median of M_l with error bars whose upper and lower bounds were the 75th percentile and 25th percentile values, respectively, as calculated using default settings in the statistical programming environment R.

Gaussian noise significance testing

For each trial at each noise level, the sets of σ scores for the proteome/lysate and for the experimental category being tested were determined. A one-tailed Wilcoxon/Mann-Whitney U test was performed for each of these trials. At each noise level, the median of the p-values for the 100 trials was plotted with error bars whose upper and lower bounds were the 75th percentile and 25th percentile values, respectively.

Gaussian noise fold change testing

For each trial at each noise level, the sets of σ scores for the proteome/lysate and for the experimental category being tested were determined. The linear difference $d_{l,t}$ between the medians of the supersaturation scores of the control set $C_{l,t}$ and experiment set $E_{l,t}$ being tested at noise level l and trial t is

$$d_{l,t} = 10^{median(E_{l,t}) - median(C_{l,t})} \quad (S9)$$

At noise level l, we plotted the median of set $\{d_{l,1}, d_{l,2}, \ldots, d_{l,100}\}$ with error bars whose upper and lower bounds were the 75th percentile and 25th percentile values of this set, respectively.

Gaussian noise KEGG pathway enrichment testing

For each trial at each noise level, we determined the Bonferroni-corrected p-value derived from the hypergeometric mean, as described above. For each noise level, we plotted the median p-value of the 100 trials with error bars whose upper and lower bound were the 75th percentile and 25th percentile values of this set, respectively.

Aggregation propensity re-weighting

To test the robustness of our results to increased weighting of the aggregation propensity scores, we recalculated supersaturation scores such that the re-weighted supersaturation score σ_{p_r} of protein p is

$$\sigma_{p_r} = C_p + wZ_p \qquad (S10)$$

where C_p is the logarithm of the concentration of protein p (see Supplementary Tables), Z_p is the aggregation propensity of protein p (calculated using the Zyggregator method[11]) and w is a re-weighting value between 1 and 5. This was performed on un-centered scores in order to avoid having to apportion the centering values between concentration and aggregation propensity. Because σ is a logarithmic value, w constitutes an exponential re-weighting on the linear score. We subsequently recomputed significance tests and fold changes for each re-weighted score.

Multiple hypothesis testing

The significance tests performed in this analysis are subject to the problem of multiple hypothesis testing, which can lead to a higher than expected number of false positives when considering unmodified p-values. Methods to correct for multiple hypothesis testing require the definition of a family of hypotheses to be considered together for adjustment. It has been suggested that such a family should consist of tests performed on a single experiment[29,30]. A number of methods exist to modify p-values or significance thresholds in order to control the family-wise error rate. Our study includes several different scenarios requiring control for multiple hypothesis testing, which necessitated the selection of appropriate families and correction methods.

DAVID pathway analysis. The DAVID software has built-in adjustments for multiple hypothesis testing performed on its database of pathways and gene categories. The two methods available through DAVID are the Bonferroni correction and the calculation of Benjamini-Hochberg False Discovery Rates. The former is the most conservative method for multiple hypothesis testing available. DAVID defines a family as a given category of gene lists (for the purposes of the

results described here, the set of all KEGG pathways). Given that there are other categories of gene lists available from DAVID, including Gene Ontology, biochemical pathway databases, and structural databases, the DAVID definition may underestimate the number of hypotheses tested. We therefore employed the highly conservative Bonferroni correction so as to reduce false positives.

Median σ tests. For each database of σ scores, we tested a series of data sets for elevated scores. We defined a family as the data sets for which median comparisons were performed on a single database. For σ_f, these were the six aggregating protein data sets, the two protein complex data sets, and the KEGG pathway overlap set (N=9). For σ_u, these were the six aggregating protein data sets and the KEGG pathway overlap set (N=7). For σ_{u_T}, these were the six aggregating protein data sets and the KEGG pathway overlap set (N=7). To correct for multiple hypotheses, we employed the Holm-Bonferroni method[31] as implemented in R. This method has greater statistical power than the Bonferroni method, yet still provides strong control of the family-wise error rate.

Gaussian noise threshold significance testing. These tests were intended to simulate random errors that might occur under replication of the experiments and predictions necessary to produce supersaturation scores. First, the Holm-Bonferroni method is used to correct for the multiple hypotheses described for median σ tests. This results in 100 p-values and median differences. Our interest is in whether the overall distribution of p-values is significantly skewed below p=0.05. For this purpose, the multiple hypothesis corrections previously described are insufficient. They control the rate of false positives for individual tests, but say nothing about the overall significance of the replicated trials. Another approach is to combine probabilities using Fisher's method[32] or one of its variants[33], but these methods estimate the probability that at least one trial can reject the null hypothesis[34]. We opted instead to evaluate the null hypothesis that the distribution of p-values at each noise level was less than 0.05 and the distribution of fold changes at each noise level was greater than 1.

For the Gaussian noise significance and enrichment testing described above, we determined whether the set of p-values associated with each noise level was significantly lower than 0.05. To

do so, we performed a Wilcoxon/Mann-Whitney U test between the set of p-values of interest and a set of size 100, all of whose elements had a value of 0.05. The latter would be the set obtained if each of the 100 trials at a given noise level resulted in a p-value at the threshold of 0.05 (that is, minimally insignificant). If this significance test resulted in $p < 0.05$, we plotted a green point to denote a statistically significant result for that noise level.

For the Gaussian noise fold change testing described above, we determined whether the set of fold changes associated with each noise level was significantly greater than 1. To do so, we performed a Wilcoxon/Mann-Whitney U test between the set of fold changes of interest and a set of size 100, all of whose element had a value of 1. The latter would be the set obtained if each of the 100 trials at a given noise level resulted in a fold change at the threshold of 1 (that is, no change). If this significance test resulted in $p < 0.05$, we plotted a green point to denote a statistically significant result for that noise level.

The maximum robustness level is the highest significant level for which all lower levels are also significant. We report robustness as a multiple of the underlying signal. For example, 20% noise would be reported as 1.2X.

Because the significance and enrichment values take into account multiple hypothesis testing, but by definition the median differences do not, the former are more stringent criteria to evaluate significance.

Supplemental References

1 Golden, T. R., Hubbard, A., Dando, C., Herren, M. A. & Melov, S. Age-related behaviors have distinct transcriptional profiles in *Caenorhabditis elegans*. *Aging Cell* **7**, 850-865 (2008).

2 UniProt Consortium. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucl. Acids Res.* **40**, D71-D75 (2012).

3 Huang da, W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Prot.* **4**, 44-57 (2009).

4 Su, A. I. *et al.* A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. USA* **101**, 6062-6067 (2004).

5 Liu, G. Y. *et al.* NetAffx: Affymetrix probesets and annotations. *Nucl. Acids Res.* **31**, 82-86 (2003).

6 Pruitt, K. D., Tatusova, T., Brown, G. R. & Maglott, D. R. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucl. Acids Res.* **40** (2012).

7 Wheeler, D. L. *et al.* Database resources of the National Center for Biotechnology. *Nucl. Acids Res.* **31**, 28-33 (2003).

8 Maglott, D., Ostell, J., Pruitt, K. D. & Tatusova, T. Entrez Gene: Gene-centered information at NCBI. *Nucl. Acids Res.* **33**, D54-D58 (2005).

9 Flicek, P. *et al.* Ensembl 2012. *Nucl. Acids Res.* **40**, D84-D90 (2012).

10 Schrimpf, S. P. *et al.* The initiative on Model Organism Proteomes (iMOP). *Proteomics* **12**, 346-350 (2012).

11 Tartaglia, G. G. *et al.* Prediction of aggregation-prone regions in structured proteins. *J. Mol. Biol.* **380**, 425-436 (2008).

12 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. Biotech.* **22**, 1302-1306 (2004).

13 Tartaglia, G. G., Cavalli, A. & Vendruscolo, M. Prediction of local structural stabilities of proteins from their amino acid sequences. *Structure* **15**, 139-143 (2007).

14 Agostini, F., Vendruscolo, M. & Tartaglia, G. G. Sequence-based prediction of protein solubility. *J. Mol. Biol.* **421**, 237-241 (2012).

15 Niwa, T. *et al.* Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of Escherichia coli proteins. *Proc. Natl. Acad. Sci. USA* **106**, 4201-4206 (2009).

16 Lujian, L. *et al.* Proteomic Characterization of Postmortem Amyloid Plaques Isolated by Laser Capture Microdissection. *J. Biol. Chem.* **279**, 37061-37068 (2004).

17 Wang, Q. *et al.* Proteomic analysis of neurofibrillary tangles in Alzheimer disease identifies GAPDH as a detergent-insoluble paired helical filament tau binding protein. *FASEB J.* **19**, 869-871 (2005).

18 Kersey, P. J. *et al.* The International Protein Index: An integrated database for proteomics experiments. *Proteomics* **4**, 1985-1988 (2004).

19 Xia, Q. *et al.* Proteomic identification of novel proteins associated with Lewy bodies. *Front. Biosci.* **13**, 3850-3856 (2008).

20 Olzscha, H. *et al.* Amyloid-like aggregates sequester numerous metastable proteins with essential cellular functions. *Cell* **144**, 67-78 (2011).
21 David, D. C. *et al.* Widespread protein aggregation as an inherent part of aging in C. elegans. *PLoS Biol.* **8**, e1000450 (2010).
22 Reis-Rodrigues, P. *et al.* Proteomic analysis of age-dependent changes in protein solubility identifies genes that modulate lifespan. *Aging Cell* **11**, 120-127 (2012).
23 Licata, L. *et al.* MINT, the molecular interaction database: 2012 update. *Nucl. Acids Res.* **40**, D857-D861 (2012).
24 Luc, P. V. & Tempst, P. PINdb: a database of nuclear protein complexes from human and yeast. *Bioinformatics* **20**, 1413-1415 (2004).
25 Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25-29 (2000).
26 Mann, H. B. & Whitney, D. R. On a test of whether one of two random variables is stochastically larger than the other *Ann. Math. Stat.* **18**, 50-60 (1947).
27 Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M. & Hirakawa, M. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucl. Acids Res.* **38**, D355-D360 (2010).
28 Hosack, D. A., Dennis, G., Sherman, B. T., Lane, H. C. & Lempicki, R. A. Identifying biological themes within lists of genes with EASE. *Genome Biol.* **4** (2003).
29 Bender, R. & Lange, S. Adjusting for multiple testing - when and how? *J. Clin. Epidem.* **54**, 343-349 (2001).
30 Hochberg, Y. & Tamhane, A. C. *Multiple comparison procedures*. (Wiley, 1987).
31 Holm, S. A Simple Sequentially Rejective Multiple Test Procedure. *Scand. J. Stat.* **6**, 65-70 (1979).
32 Birnbaum, A. Combining Independent Tests of Significance. *J. Am. Stat. Ass.* **49**, 559-574 (1954).
33 Whitlock, M. C. Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. *J. Evol. Biol* **18**, 1368-1373 (2005).
34 Cooper, H. & Hedges, L. V. *The handbook of research synthesis*. (Russell Sage Foundation, 1994).

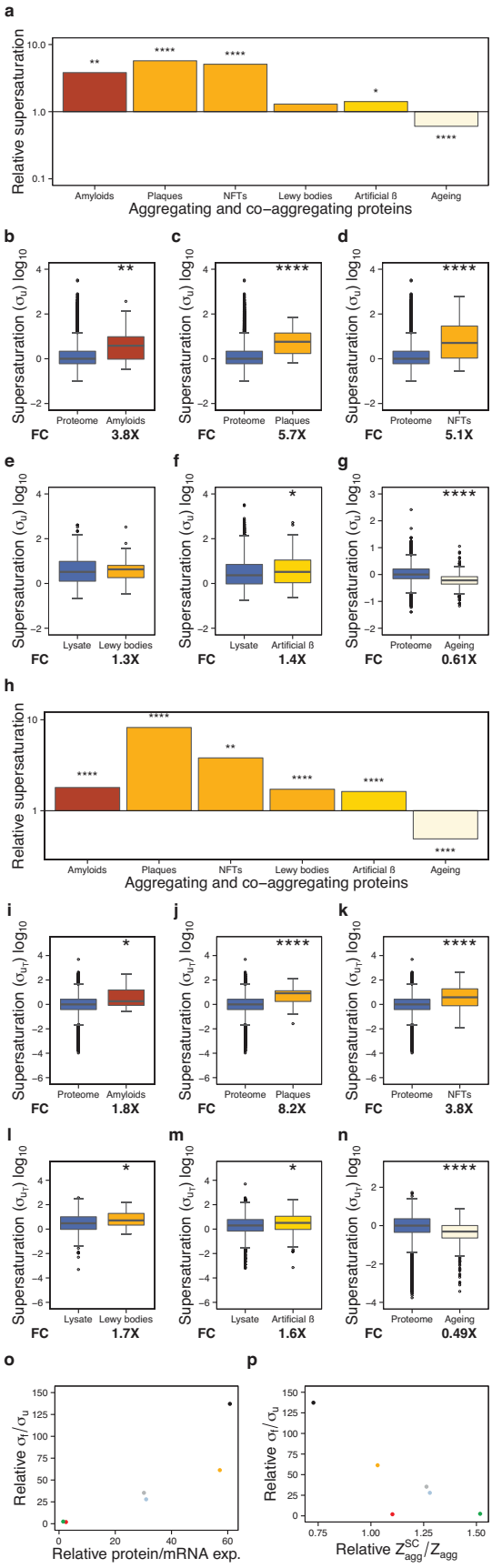
a
Relative supersaturation
Amyloids
Plaques
NFTs
Lewy bodies
Artificial ß
Ageing
Aggregating and co−aggregating proteins
b
c
d
Supersaturation (σ_u) $\log_{10}$
Proteome
Amyloids
Plaques
NFTs
FC
3.8X
5.7X
5.1X
e
f
g
Lysate
Lewy bodies
Artificial ß
1.3X
1.4X
0.61X
h
i
j
k
σ_{u_T}
1.8X
8.2X
3.8X
l
m
n
1.7X
1.6X
0.49X
o
p
Relative σ_f/σ_u
Relative protein/mRNA exp.
Relative Z^{SC}_{agg}/Z_{agg}

Figure S1. Prediction of aggregating proteins with σ_u and σ_{u_T}, related to Figure 2. (a) Summary of results for various classes of proteins given in terms of the increase in the σ_u score over the average value for the whole proteome or an experimental lysate ("fold change"). σ_u scores were compared for the **(b)** proteome and amyloid proteins ($p=5.4\cdot10^{-3}$), **(c)** proteome and proteins that co-precipitate with Aß plaques ($p=4.2\cdot10^{-8}$), **(d)** proteome and proteins that co-precipitate with neurofibrillary tangles ($p=1.6\cdot10^{-12}$), **(e)** lysate to proteins that co-precipitate with Lewy bodies ($p=0.23$), **(f)** lysate and proteins that co-precipitate with artificial ß-peptide aggregates ($p=1.4\cdot10^{-2}$), and **(g)** proteome and proteins found to aggregate in *C. elegans* during ageing ($p=2.0\cdot10^{-85}$). **(h)** Summary of results for various classes of proteins given in terms of the increase in the σ_{u_T} score over the average value for the whole proteome or an experimental lysate ("fold change"). σ_{u_T} scores were compared for: **(i)** proteome and amyloid proteins ($p=2.0\cdot10^{-2}$), **(c)** proteome and proteins that co-precipitate with Aß plaques ($p=4.0\cdot10^{-6}$), **(k)** proteome and proteins that co-precipitate with neurofibrillary tangles ($p=1.2\cdot10^{-8}$), **(l)** lysate to proteins that co-precipitate with Lewy bodies ($p=2.0\cdot10^{-2}$), **(m)** lysate and proteins that co-precipitate with artificial ß-peptide aggregates ($p=1.6\cdot10^{-2}$), and **(n)** proteome and proteins found to aggregate in *C. elegans* during ageing ($p=5.7\cdot10^{-37}$). For the six widespread protein aggregation data sets described in Figs. 2 and S1, the ratio of the median σ_f and σ_u relative to control are plotted against the ratio of the median **(o)** protein abundance and mRNA level relative to control or **(p)** Z_{agg}^{SC} and Z_{agg} relative to control, for amyloid proteins (black), proteins aggregating in ageing in *C. elegans* (orange), proteins that co-precipitate with Lewy bodies (red), proteins that co-precipitate with Aß plaques (grey), proteins that co-precipitate with neurofibrillary tangles (blue), and proteins that co-precipitate with artificial ß-peptide aggregates (green). Boxes and significance tests are as in **Fig. 2**.

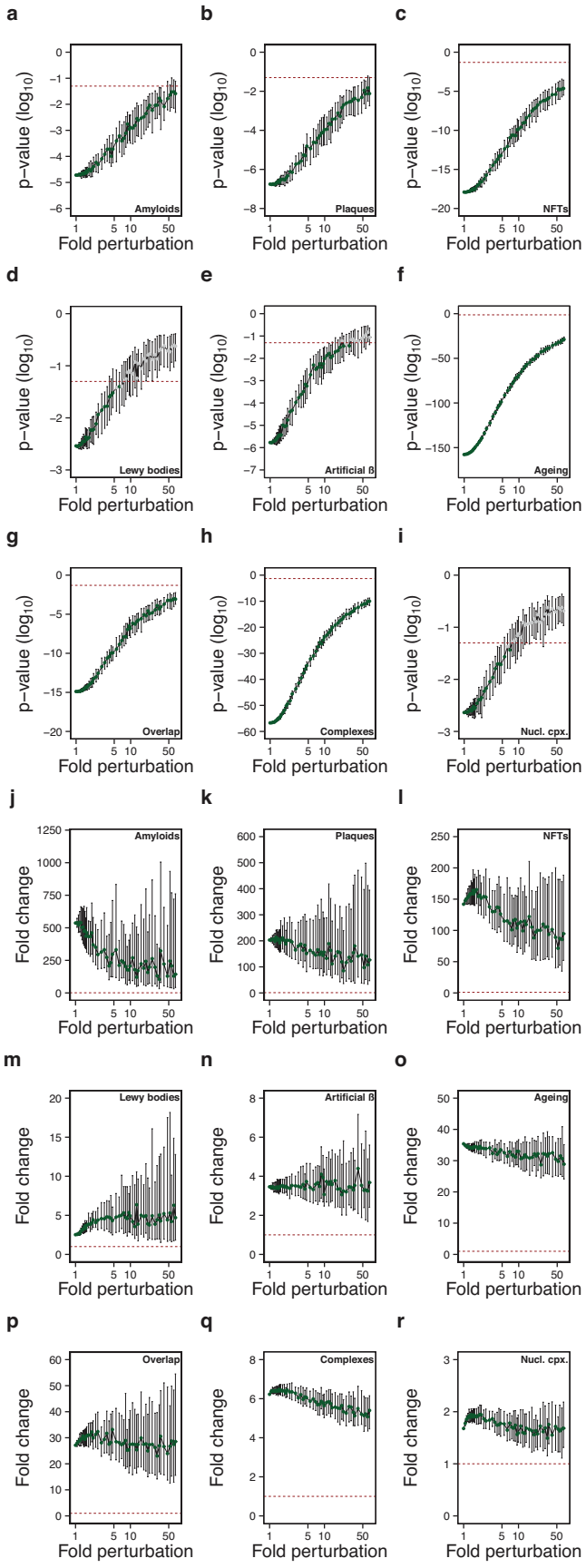
a
b
c
d
e
f
g
h
i
j
k
l
m
n
o
p
q
r
p−value (log10)
Fold change
Fold perturbation
Amyloids
Plaques
NFTs
Lewy bodies
Artificial ß
Ageing
Overlap
Complexes
Nucl. cpx.

Figure S2. Noise tests for σ_f, related to Figures 2, 5, and 6. Test of the robustness against noise of the significance of difference in distribution between the proteome/lysate and various data sets. Gaussian noise was introduced 100 independent times into the proteome scores at 50 different noise levels (1X = no noise) and plotted against **(a-i)** p-value or **(j-r)** fold change: **(a)** amyloid proteins v. proteome (robust up to 66X), **(b)** Aß plaque co-aggregators v. lysate (robust up to 66X), **(c)** neurofibrillary tangle co-aggregators v. proteome (robust up to 66X), **(d)** Lewy body co-aggregators v. lysate (robust up to 4.7X), **(e)** artificial ß-peptide aggregate co-aggregators v. proteome (robust up to 19X), **(f)** proteins aggregating over ageing in *C. elegans* v. proteome (robust up to 66X), **(g)** the overlap proteins of the Alzheimer's, Parkinson's, and Huntington's disease KEGG pathways v. proteome (robust up to 66X), **(h)** proteins that form complexes v. proteome (robust up to 66X), **(i)** proteins that form nuclear complexes v. nuclear proteome (robust up to 5.5X), **(j)** amyloid proteins v. proteome (robust up to 66X), **(k)** Aß plaque co-aggregators v. lysate (robust up to 66X), **(l)** neurofibrillary tangle co-aggregators v. proteome (robust up to 66X), **(m)** Lewy body co-aggregators v. lysate (robust up to 66X), **(n)** artificial ß-peptide aggregate co-aggregators v. proteome (robust up to 66X), **(o)** proteins aggregating over ageing in *C. elegans* v. proteome (robust up to 66X), **(p)** the overlap proteins of the Alzheimer's, Parkinson's, and Huntington's disease KEGG pathways v. proteome (robust up to 66X), **(q)** proteins that form complexes v. proteome (robust up to 66X), and **(r)** proteins that form nuclear complexes v. nuclear proteome (robust up to 66X). Green points are noise levels **(a-i)** significantly below p=0.05 significance (red dashed line) or **(j-r)** significantly above fold change of 1 (red dashed line) by Wilcoxon/Mann Whitney U. Plotted points are **(a-i)** p-value median error bars represent the interquartile range of Wilcoxon/Mann Whitney U p-values; **(j-r)** fold change: median, error bars represent the interquartile range of fold change; and fold perturbation: median, error bars represent the interquartile range of fold perturbation.

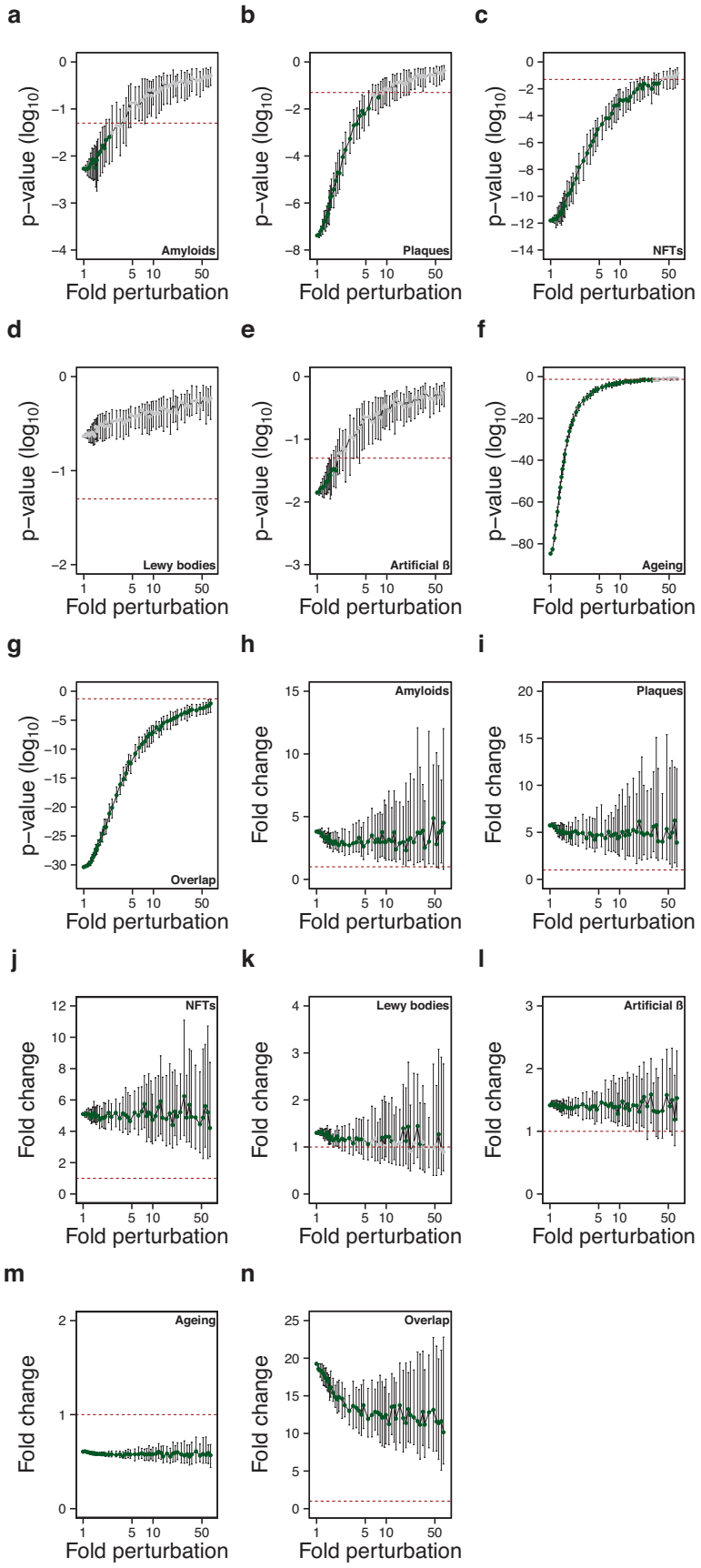
a
b
c
d
e
f
g
h
i
j
k
l
m
n
p−value (log10)
Fold change
Fold perturbation
Amyloids
Plaques
NFTs
Lewy bodies
Artificial ß
Ageing
Overlap
1
5
10
50

Figure S3. Noise tests for σ_u, related to Figures 2 and 5. Test of the robustness against noise of the significance of difference in distribution between the proteome/lysate and various data sets. Gaussian noise was introduced 100 independent times into the proteome scores at 50 different noise levels (1X = no noise) and plotted against **(a-g)** p-value or **(h-o)** fold change: **(a)** amyloid proteins v. proteome (robust up to 2.3X), **(b)** Aß plaque co-aggregators v. lysate (robust up to 5.5X), **(c)** neurofibrillary tangle co-aggregators v. proteome (robust up to 36X), **(d)** Lewy body co-aggregators v. lysate (not significant), **(e)** artificial ß-peptide aggregate co-aggregators v. proteome (robust up to 1.9X noise), **(f)** proteins aggregating over ageing in *C. elegans* v. proteome (robust up to 28X noise), and **(g)** the overlap proteins of the Alzheimer's, Parkinson's, and Huntington's disease KEGG pathways v. proteome (robust up to 66X noise), **(h)** amyloid proteins v. proteome (robust up to 66X noise), **(i)** Aß plaque co-aggregators v. lysate (robust up to 66X noise), **(j)** neurofibrillary tangle co-aggregators v. proteome (robust up to 66X noise), **(k)** Lewy body co-aggregators v. lysate (robust up to 4.4X noise), **(l)** artificial ß-peptide aggregate co-aggregators v. proteome (robust up to 66X noise), **(m)** proteins aggregating over ageing in *C. elegans* v. proteome (robust up to 66X noise), and **(n)** the overlap proteins of the Alzheimer's, Parkinson's, and Huntington's disease KEGG pathways v. proteome (robust up to 66X noise). Green points are noise levels **(a-g)** significantly below p=0.05 significance (red dashed line) or **(h-n)** significantly above fold change of 1 (red dashed line) by Wilcoxon/Mann-Whitney U. Plotted points are **(a-g)** p-value median error bars represent the interquartile range of Wilcoxon/Mann Whitney U p-values; **(h-n)** fold change: median, error bars represent the interquartile range of fold change; and fold perturbation: median, error bars represent the interquartile range of fold perturbation.

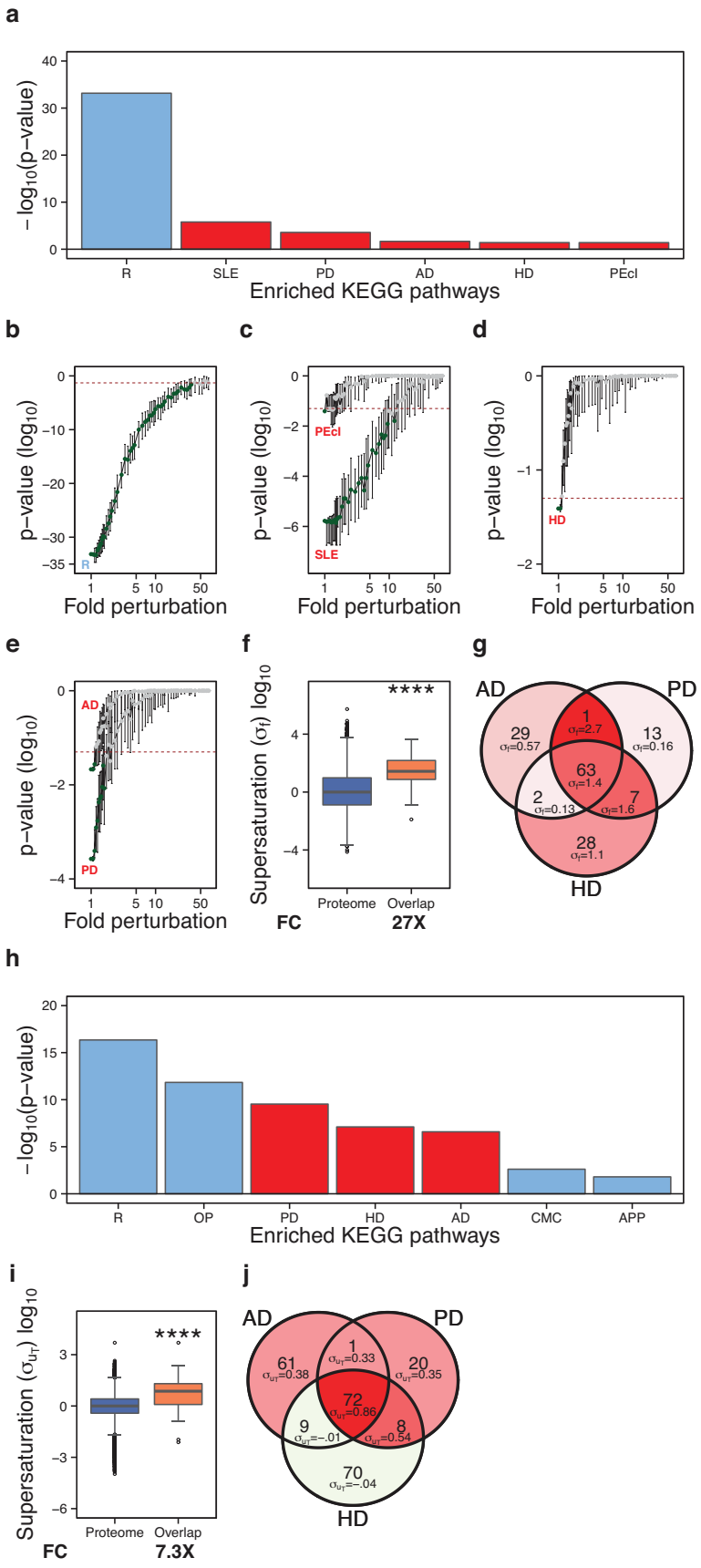
a
− log10(p−value)
Enriched KEGG pathways
R
SLE
PD
AD
HD
PEcl
b
c
d
p−value (log10)
Fold perturbation
e
f
g
Supersaturation (σf) log10
Proteome
Overlap
FC
27X
****
AD
PD
HD
h
OP
CMC
APP
i
j
Supersaturation (σUT) log10
7.3X

Figure S4. Neurodegenerative disease pathways are enriched in supersaturated proteins (σ_f, σ_{u_T}), related to Figures 3 and 5. (a) List of the KEGG pathways[24] identified here as significantly enriched (Bonferroni-corrected p-values) in proteins at or above the 95th percentile of supersaturation (σ_f): (R) Ribosome, (SLE) Systemic Lupus Erythematosus, (PD) Parkinson's disease, (AD) Alzheimer's disease, (HD) Huntington's disease, (PEcI) Pathogenic *Escherichia coli* infection; physiological and pathological pathways are shown in blue and red, respectively. **(b,c)** Test of the robustness against noise of the significance of enrichment of the KEGG pathways according to their supersaturation score. Gaussian noise was introduced 100 independent times into the proteome scores at 50 different noise levels and plotted (1X = no noise) for significant: **(b)** physiological pathway, which was robust up to 36X (R) and **(c-e)** pathological pathways, which were robust up to **(c)** 9.4X SLE, 1X (PEcI), 1X; **(d)** 1.1X (HD), **(e)** 1.6X (PD), and 1.1X (AD). Error bars indicate interquartile ranges, green points indicate error levels below the p=.05 significance (red dashed line) by the Wilcoxon/Mann-Whitney U test. **(f)** Comparison of the σ_f scores for the proteome and the set of 63 proteins common among the Alzheimer's, Parkinson's, and Huntington's KEGG pathways[24]; this set of proteins is denoted as 'overlap,' see panel (g) ($p=1.3\bullet10^{-15}$). **(g)** Comparison of the σ_f scores for the proteins in the Alzheimer's, Parkinson's, and Huntington's pathways. **(h)** List of the KEGG pathways[24] identified here as significantly enriched (Bonferroni-corrected p-values) at or above the 95th percentile of supersaturation (σ_{u_T}): (R) Ribosome, (OP) Oxidative phosphorylation, (PD) Parkinson's disease, (AD) Alzheimer's disease, (HD) Huntington's disease, (CMC) Cardiac muscle contraction, (APP) Antigen processing and presentation; physiological and pathological pathways are shown in blue and red, respectively. **(i)** Comparison of the σ_{u_T} scores for the proteome and the set of 72 proteins common among the Alzheimer's, Parkinson's, and Huntington's KEGG pathways[24]; this set of proteins is denoted as 'overlap,' see panel (c) ($p=9.7\bullet10^{-14}$). **(j)** Comparison of the σ_{u_T} scores for the proteins in the Alzheimer's, Parkinson's, and Huntington's pathways. Colours are assigned based on the division of the σ scores into deciles from low (green) to high (red). Boxplots and significance tests are as in **Fig. 2**.

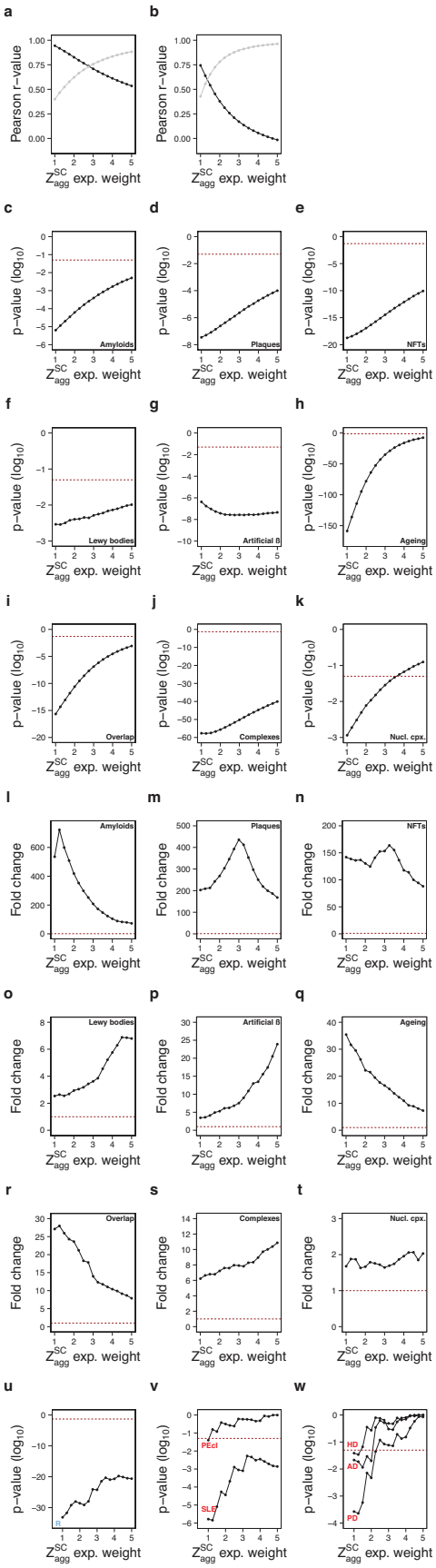
a
b
Pearson r−value
1.00
0.75
0.50
0.25
0.00
1 2 3 4 5
Z^{SC}_{agg} exp. weight
c
d
e
p−value (log_{10})
Amyloids
Plaques
NFTs
f
g
h
Lewy bodies
Artificial β
Ageing
i
j
k
Overlap
Complexes
Nucl. cpx.
l
m
n
Fold change
Amyloids
Plaques
NFTs
o
p
q
Lewy bodies
Artificial β
Ageing
r
s
t
Overlap
Complexes
Nucl. cpx.
u
v
w
R
PEcl
SLE
HD
AD
PD

Figure S5. Widespread aggregation predictions from the folded state with increasing weight of Z_{agg}^{SC}, related to Figures 2, 5 and 6. (a) Human and **(b)** *C. elegans* Z_{agg}^{SC} scores were multiplied by values ranging from 1 to 5 when calculating σ_f effectively increasing the weight of aggregation propensity in the supersaturation score. The Pearson correlation coefficients for supersaturation scores with abundance (black) and aggregation propensity (grey) were plotted. The Wilcoxon/Mann-Whitney U p-value was plotted for: **(c)** amyloid proteins v. proteome (robust up to 5X), **(d)** Aß plaque co-aggregators v. lysate (robust up to 5X), **(e)** neurofibrillary tangle co-aggregators v. proteome (robust up to 5X), **(f)** Lewy body co-aggregators v. lysate (robust up to 5X), **(g)** artificial ß-peptide aggregate co-aggregators v. proteome (robust up to 5X), **(h)** proteins aggregating over ageing in *C. elegans* v. proteome (robust up to 5X), **(i)** the overlap proteins of the Alzheimer's, Parkinson's, and Huntington's disease KEGG pathways v. proteome (robust up to 5X), **(j)** proteins that form complexes v. proteome (robust up to 5X), and **(k)** proteins that form nuclear complexes v. nuclear proteome (robust up to 3.5X). The fold change was plotted for: **(l)** amyloid proteins v. proteome (robust up to 5X), **(m)** Aß plaque co-aggregators v. lysate (robust up to 5X), **(n)** neurofibrillary tangle co-aggregators v. proteome (robust up to 5X), **(o)** Lewy body co-aggregators v. lysate (robust up to 5X), **(p)** artificial ß-peptide aggregate co-aggregators v. proteome (robust up to 5X), **(q)** proteins aggregating over ageing in *C. elegans* v. Proteome (robust up to 5X), **(r)** the overlap proteins of the Alzheimer's, Parkinson's, and Huntington's disease KEGG pathways v. Proteome (robust up to 5X), **(s)** proteins that form complexes v. proteome (robust up to 5X), and **(t)** proteins that form nuclear complexes v. nuclear proteome (robust up to 5X). Proteins at or above the 95th percentile in our databases were tested for KEGG pathway enrichment in comparison to the databases as a whole, using the DAVID bioinformatics software. The Bonferroni-corrected p-value was plotted for **(u)** the significant σ_f physiological pathway (R) Ribosome (robust up to 5X), and **(v-w)** significant σ_f pathological pathways (SLE) Systemic lupus erythematosus (robust up to 5X), (PEcI) Pathogenic *Escherichia coli* infection (robust up to 1X), (PD) Parkinson's disease (robust up to 2.25X), (AD) Alzheimer's disease (robust up to 2X), and (HD) Huntington's disease (robust up to 1.25X).

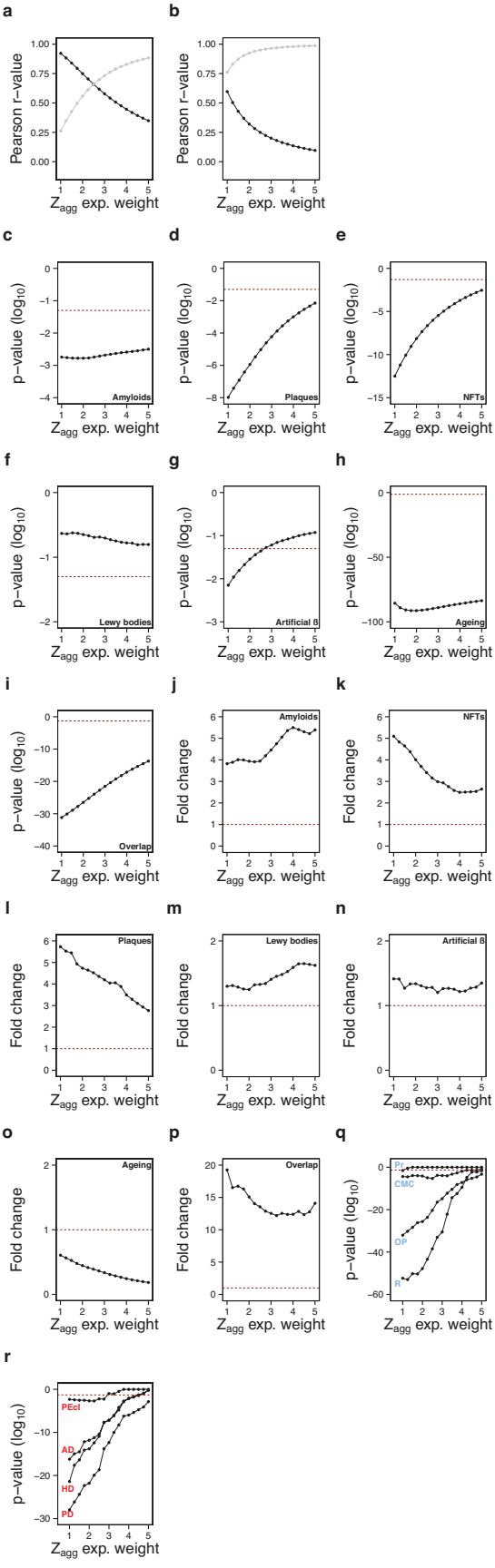
a
b
Pearson r−value
1.00
0.75
0.50
0.25
0.00
1 2 3 4 5
Zagg exp. weight
c
d
e
p−value (log10)
Amyloids
Plaques
NFTs
f
g
h
Lewy bodies
Artificial ß
Ageing
i
j
k
Overlap
Fold change
Amyloids
NFTs
l
m
n
Plaques
Lewy bodies
Artificial ß
o
p
q
Ageing
Overlap
PI
CMC
OP
R
r
PEcI
AD
HD
PD

Figure S6. Widespread aggregation predictions from the unfolded state with increasing weight of Z_{agg}, related to Figures 2, 3 and 5. (a) Human and **(b)** *C. elegans* Z_{agg} scores were multiplied by values ranging from 1 to 5 when calculating σ_u effectively increasing the weight of aggregation propensity in the supersaturation score. The Pearson correlation coefficients for supersaturation scores with abundance (black) and aggregation propensity (grey) were plotted. The Wilcoxon/Mann-Whitney U p-value was plotted for: **(c)** amyloid proteins v. proteome (robust up to 5X), **(d)** Aß plaque co-aggregators v. lysate (robust up to 5X), **(e)** neurofibrillary tangle co-aggregators v. proteome (robust up to 5X), **(f)** Lewy body co-aggregators v. lysate (not significant), **(g)** artificial ß peptide aggregate co-aggregators v. proteome (robust up to 2.5X), **(h)** proteins aggregating over ageing in *C. elegans* v. proteome (robust up to 5X), and **(i)** the overlap proteins of the Alzheimer's, Parkinson's, and Huntington's disease KEGG pathways v. proteome (robust up to 5X). The fold change was plotted for: **(j)** amyloid proteins v. proteome (robust up to 5X), **(k)** Aß plaque co-aggregators v. lysate (robust up to 5X), **(l)** neurofibrillary tangle co-aggregators v. proteome (robust up to 5X), **(m)** Lewy body co-aggregators v. lysate (robust up to 5X), **(n)** artificial ß peptide aggregate co-aggregators v. proteome (robust up to 5X), **(o)** proteins aggregating over ageing in *C. elegans* v. proteome (robust up to 5X), and **(p)** the overlap proteins of the Alzheimer's, Parkinson's, and Huntington's disease KEGG pathways v. proteome (robust up to 5X). Proteins at or above the 95th percentile in our databases were tested for KEGG pathway enrichment in comparison to the databases as a whole, using the DAVID bioinformatics software. The Bonferroni-corrected p-value was plotted for **(q)** significant σ_u physiological pathways (R) Ribosome (robust up to 5X), (OP) Oxidative phosphorylation (robust up to 5X), (CMC) Cardiac muscle contraction (robust up to 4.75X), and (Pr) Proteasome (robust up to 1X), and **(r)** significant σ_u pathological pathways (PD) Parkinson's disease (robust up to 5X), (HD) Huntington's disease (robust up to 4.5X), (AD) Alzheimer's disease (robust up to 4.25X), and (PEcI) Pathogenic *Escherichia coli* infection (robust up to 2.75X)

(See **Table S1.xlsx**)

Table S1. Supersaturation database, related to Figure 1. This database includes a total of 34,544 *C. elegans* (top) and human proteins (bottom). Human amyloid proteins are marked with an X in the Amyloid column (bottom) (the 7 amyloid proteins not included in our database are listed at the bottom of this table). Database values between the *C. elegans* and human sets are not directly comparable as they were separately normalized. Different types of supersaturation scores are not directly comparable as they were also separately normalized. All values are in log_{10}. Supersaturation scores have been recentered to a median of 0. Values are rounded to two decimal places.

Dataset	References	Species	Original #	# UniProt IDs	# $\boldsymbol{\sigma_u}$	# $\boldsymbol{\sigma_f}$	# $\boldsymbol{\sigma_{u_T}}$
mRNA expression	(Su et al., 2004)	Human	44,775	16,293	16,263	–	16,054
mRNA expression	(Golden et al., 2008)	Worm	22,490	16,623	16,623	–	16,432
Protein abundance	(Schrimpf et al., 2012)	Human	12,803	10,247	–	6,155	–
Protein abundance	(Schrimpf et al., 2012)	Worm	11,719	11,069	–	10,149	–
Aggregation propensity ($\boldsymbol{z_{agg}}$)	(Tartaglia et al., 2009)	Human	16,268	16,268	16,263	–	–
Aggregation propensity ($\boldsymbol{Z_{agg}^{SC}}$)	(Tartaglia et al., 2009)	Human	10,552	10,552	–	6,155	–
Aggregation propensity (TANGO)	(Fernandez-Escamilla et al., 2004)	Human	16,054	16,054	–	–	16,054
Aggregation propensity ($\boldsymbol{z_{agg}}$)	(Tartaglia et al., 2009)	Worm	17,197	17,197	16,623	–	–
Aggregation propensity ($\boldsymbol{Z_{agg}^{SC}}$)	(Tartaglia et al., 2009)	Worm	10,149	10,149	–	10,149	–
Aggregation propensity (TANGO)	(Fernandez-Escamilla et al., 2004)	Worm	16,434	16,434	–	–	16,432

Table S2. Summary of data sets, related to Table 1. Ten data sets were used to build the supersaturation algorithms for human and *C. elegans* proteins. Dataset: description of the data used. References: reference for the data included. Species: species to which the data refers. Original #: the number of data points listed in the original set. # UniProt IDs: after conversion to Uniprot ID, the number of data points listed; for human proteins, only reviewed Uniprot IDs are counted. # σ_u, # σ_f, # σ_{u_T}: of those proteins in the previous column, the number included in the given database.

	Human			*C. elegans*		
	σ_f	σ_u	σ_{u_T}	σ_f	σ_u	σ_{u_T}
100 Most Abundant Proteins	3.35	2.11	1.77	2.60	0.63	0.42
100 Least Abundant Proteins	-2.32	-0.49	-0.39	-2.20	-0.84	-0.89

Table S3. Supersaturation scores of the most and least abundant proteins, related to Figure 1. Listed here are the median supersaturation scores of the 100 most and least abundant proteins (or mRNA transcripts). σ_f, σ_u, and σ_{u_T}: median supersaturation scores, after centering the median to 0. Values are rounded to two decimal places.