PNAS
www.pnas.org

Supplementary Information for

Genome-scale reconstruction of metabolic networks of model animals represents a platform for translational research

Hao Wang[a,b,c], Jonathan L. Robinson[a,b], Pınar Kocabaş[a], Johan Gustafsson[a], Mihail Anton[b], Pierre-Etienne Cholley[b], Shan Huang[b], Johan Gobom[d], Thomas Svensson[b], Mathias Uhlén[e,f,g], Henrik Zetterberg[d,h,i,j], Jens B. Nielsen[a,e,k]*

[a]Department of Biology and Biological Engineering, Chalmers University of Technology, Kemivägen 10, Gothenburg, Sweden.
[b]Department of Biology and Biological Engineering, National Bioinformatics Infrastructure Sweden, Science for Life Laboratory, Chalmers University of Technology, Kemivägen 10, Gothenburg, Sweden.
[c]Wallenberg Center for Molecular and Translational Medicine, University of Gothenburg, Kemivägen 10, Gothenburg, Sweden.
[d]Department of Psychiatry and Neurochemistry, Institute of Neuroscience and Physiology, the Sahlgrenska Academy at the University of Gothenburg, Mölndal, Sweden.
[e]Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark.
[f]Department of Protein Science, Science for Life Laboratory, KTH–Royal Institute of Technology, Stockholm, Sweden.
[g]Wallenberg Center for Protein Research, KTH–Royal Institute of Technology, Stockholm, Sweden.
[h]Clinical Neurochemistry Laboratory, Sahlgrenska University Hospital, Mölndal, Sweden.
[i]Department of Neurodegenerative Disease, University College London Institute of Neurology, Queen Square, London, UK.
[j]UK Dementia Research Institute at UCL, London, UK.
[k]BioInnovation Institute, Ole Maaløes Vej 3, DK2200 Copenhagen, Denmark.

* Corresponding author: Jens B. Nielsen
Email**:** nielsenj@chalmers.se

This PDF file includes:

Supplementary Information Text
Figures S1 to S9
References for supplementary text

Other supplementary materials:

Datasets S1 to S4

Supplementary Information Text

Generation of animal GEMs

The GEM generation pipeline consists of two modules (Fig. S1): one for reconstruction of an orthology-based GEM using Human1 (1) as template, another for preparation of species-specific metabolic network by using the RAVEN 2.0 package (2).

In module 1, the ortholog pairs between human and animal, as well as associated features including *bestForward*, *bestReverse*, *methodCount* (the number of different methods used in determining the orthology), were first retrieved from the Alliance Genomes databases using the stringent criteria (3), and then processed by the function *extractAllianceGenomeOrthologs*, in which all one-to-one pairs were kept, while the one-to-multiple pairs were filtered with following criteria: 1) exclude orthologs that are neither the best forward nor the best reverse match to a human gene; 2) only keep orthologs that are both the best froward and reverse hits. If steps 1) and 2) exclude all hits for a query gene, then retrieve and keep the ortholog pair(s) with the highest *methodCount*. Subsequently, these processed ortholog pairs and Human-GEM (v1.5.0) were used as input of *getModelFromOrthology* function for obtaining an ortholog-GEM, in which the human genes and gene-reactions rules were replaced with ortholog genes of corresponding animal by function *replaceGrRules*.

In module 2, the RAVEN function *getModelFromKEGG* function was used to retrieve metabolic networks for a given model animal and human using the KEGG database (4). The metabolic network unique to an animal species was obtained by removing reactions shared in the human metabolic network, and then subjected to manual inspection of reaction compartment, reversibility, and annotations.

For each species, the GEM was obtained from integrating species-specific network into the ortholog-GEM by function *addMetabolicNetwork* and a follow-up gap-filling step, using *gapfill4essentialTasks* function, to ensure the resulting GEM in conducting essential metabolic tasks (Dataset S1) and biomass formation. Additionally, all the GEMs were also evaluated by "metabolicTasks_VerifyModel" (Dataset_S1), in which a list of 21 verification tasks were checked to ensure that there were no infeasible flux circles in the GEMs, e.g. generation of reducing power or re-phosphorylation of ATP for free or at physiologically infeasible yields.

Since a substantial amount of work and manual curation went into developing the recent GEM iCEL1314. We thus incorporated a total of 32 new Ascaroside biosynthesis and transport reactions, representing the major changes introduced into iCEL1314, as part of Worm1 species-specific network for more complete coverage of metabolism.

Metabolic Atlas 2.0

This is a major release that includes software architecture changes and upgrades since version 1.0 (1). All the underlying code and data are now publicly available on GitHub repositories, which facilitates the migration of Metabolic Atlas towards a fully automated pipeline that welcomes community contributions.

Graph database

Metabolic networks have previously been described as graphs (5), where metabolic insights may be investigated through the application of graph algorithms. The graph database engine Neo4j (https://neo4j.com/) has been used by various systems biology databases (6–8) and well-established biomedical resources (9, 10). In Metabolic Atlas 2.0, the backend was upgraded from the previously used Postgres relational database to the graph database Neo4j, which enables queries to compare GEMs, such as between multiple versions of the same GEM and between

different GEMs through the associated identifiers from external databases: e.g. MetaNetX (11), KEGG (4) and UniProt (12).

Processing input data

With the implementation of the graph database, all data files, such as integrated GEMs and manually-drawn 2D maps, have been centralized into a single public GitHub repository https://github.com/MetabolicAtlas/data-files. A pipeline that conducts automatic data processing and integration is available at https://github.com/MetabolicAtlas/neo4j-data-generation.

3D Map Viewer

The 3D Map Viewer was upgraded with improved performance, so that the visualization of metabolic networks is now accessible from computers and mobile devices, regardless of the network size. The source code is available at https://github.com/MetabolicAtlas/3d-network-viewer.

Memote test

The new and published GEMs were benchmarked with Memote (v0.12.0) (13), by which a 'snapshot' report was generated for each GEM using the same set of parameters. The obtained 'Total Score' were comparatively evaluated at species level.

Gene essentiality analysis

Genome-scale essentiality data for mouse, fruit fly, and worm were retrieved from the Online Gene Essentiality (OGEE) database (14), from which a full list of genes classified as either "essential" or "non-essential" were extracted (Dataset S2). Genes classified as "conditional" – essential in some but not all of the tested conditions – were grouped with the essential genes in our analysis. A computational gene deletion analysis was then performed to generate a predicted classification (essential or non-essential) for each GEM by function *evalGeneEssentialityPred*, in which each gene was individually "deleted" by inactivating all reactions encoded by the gene (excluding reactions that could be catalyzed by a compensatory isozyme) and then flux balance analysis was performed by setting the biomass production as the objective. For GEMs with pre-defined default media conditions (FlySilico, iCEL1273, and iCEL1314), only metabolites present in the default media conditions were allowed to be consumed when maximizing biomass (Dataset S2). For the other GEMs, metabolites present in Ham's medium were allowed to be consumed (Dataset S2). For all GEMs, the maximum allowed consumption rate of each media component was set to 1000 mmol per gram dry cell weight per hour (mmol/gDW/h). Genes whose deletion disabled the production of biomass (to a value less than 1 mmol/gDW/h) were classified as "essential", whereas those with little or no effect on biomass flux were classified as "non-essential". All GEM simulations were carried out with the Gurobi solver (Gurobi Optimization, LLC).

The predicted essentiality classifications were then compared to those from the OGEE database, where true positives (TP) and true negatives (TN) were defined as genes that were correctly predicted as essential and non-essential, respectively, and false positives (FP) and false negatives (FN) were incorrect predictions. For each GEM, these values were used to calculate different performance metrics (Dataset S2), including Sensitivity, Specificity, Accuracy, F1 score, Matthew's correlation coefficient (MCC) and a hypergeometric test (Fisher's exact test) that was performed to evaluate the significance (*p*-value) of true positives among the predicted essential genes, with following equations:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

$$\text{F1} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}$$

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}$$

Omics data retrieval

The GEO database (15) was screened for RNA-seq datasets sequenced from rodent models of Alzheimer's disease (AD). Initially, we collected 60 datasets that were derived from various mouse models; however, there were no datasets available from AD rat models. These datasets were further refined according to the following criteria: i) contains raw gene counts; ii) has paired disease and wild type samples for each experimental condition; iii) availability of extensive metadata (gender, age, tissue/cell type source); and iv) contains at least three biological replicates for each condition. Finally, a total of 11 datasets with 404 samples under 54 different experimental conditions were selected from 6 representative AD mouse models (Dataset S3). These models were manually inspected according to the Alzforum database (16), from which the model names, mutated risk genes, Aβ plaque formation and disease progression information were retrieved (Dataset S3).

Proteomics datasets produced from brain tissues of APP overexpression mouse models 5xFAD (17), hAPP and hAPP/PS1 (18), ADLP^{APT} and $\text{ADLP}^{\text{APP/PS1}}$ (19) were used in validating the differentially expressed lysosomal enzymes. Peptidomics data measured from nondegenerative patients and healthy controls were retrieved from a study targeting for the quantification of abundant peptides in cerebrospinal fluid samples (20).

Generation and comparison of tissue-specific GEMs

The raw gene counts from various datasets were joined into a single matrix and the genes that were not present in all datasets were discarded. The counts were normalized using the function *estimateSizeFactors* in DESeq2 (v1.26.0) (21) with default parameters. The counts for each gene were then divided by the average transcript lengths which were retrieved from BioMart (22) (v. 2.41.9, genome version 100) using the R-package GenomicFeatures (23) (v. 1.37.6). The gene counts of all samples were then linearly scaled to the average across all samples, to resemble TPM values. The tissue- and cell type-specific GEMs were finally generated using function *getINITModel2* with the essential metabolic tasks (Dataset S1) and modified gene counts as input and by setting expression threshold as 1. GEM structures in t-distributed stochastic neighbor embedding (*t*SNE) and reaction content, as well as functional differences in subsystem coverage and metabolic task performance were compared using the RAVEN function *compareMultipleModels*.

Differential expression analysis

Differential expression analysis for paired transgenic mice versus wild type controls were performed using DESeq2 (v1.26.0) with default parameters (21). Raw gene (integer) counts were used as input for each analysis, where gene counts of multiple variants of the same gene were merged by summing them up. The variance-mean dependence in count data was estimated using the Wald

test based on a model using the negative binomial distribution, and *p*-values were corrected for multiple testing using the Benjamini-Hochberg method.

Gene set analysis

Multiple gene-set enrichment analysis were performed using GSAM package (https://github.com/JonathanRob/GeneSetAnalysisMatlab), which is a MATLAB implementation of the approach developed by Väremo et al (24). Gene sets were defined as either the set of genes constituting pathways in Mouse1 (reporter subsystems), or as the set of genes that encode for any reactions involving a certain metabolite (reporter metabolites) (5), or retrieved from the Molecular Signatures Database (MSigDB) (25), specifically the Hallmark (26), KEGG (4), and Reactome (9) gene set collections. The gene identifiers of MSigBD gene sets were replaced with corresponding mouse ortholog genes according to the information retrieved from the Alliance Genomes database (3).

The significance of directional gene set enrichment (p_{enrich}) was estimated as described previously (24). First, the significance estimates (*p*-values) from the differential expression analysis were converted to directional p-values (p_{dir}) for each gene *i*:

$$p_{dir,i} = \frac{(p_i - 1) \cdot sign(FC_i) + 1}{2}$$

where *sign(FC)* corresponds to the sign of the log fold-change of each gene. This transformation yields p_{dir} values that are near zero for genes exhibiting a very significant increased expression, near one for genes that significantly decreased expression, and approximately 0.5 for genes with negligible change in expression.

Gene sets were scored based on the p_{dir} values of their associated genes. For the reporter metabolite and reporter subsystem analyses, gene sets were scored using the reporter method (5), which involves a conversion of the gene p_{dir} values into Z-scores. A Wilcoxon rank-sum test was used to score gene sets from MSigDB. After calculating the gene set scores, the enrichment significance of each gene set (p_{enrich}) was estimated by comparing the gene set scores to those of 50,000 randomly shuffled gene sets of equal size.

Network analysis

The integrative analysis of reporter metabolite gene sets with the metabolic network of Mouse1 was carried out with the Kiwi package (27), in which gene sets were considered related if the mutual shortest path length was no more than 2 in metabolic network. The parameters *p*-value cutoff, maximum number, and maximum degree of gene sets were adjusted between 0.00001-0.002, 50-100, and 50-100, respectively, to avoid having networks with over-crowded nodes. A directionality score that indicates the differential direction of each gene set was calculated and represented as node color scaling from blue (down-) to red (up-regulation). The interactions between metabolite gene sets and genes were extracted from the Mouse1. The input gene set and gene level statistics files were generated using the PIANO R-package (24). The network visualization diagrams were adjusted by using Cytoscape (28).

Statistical analysis

All analysis was performed using R (ver 4.0.2) or Matlab (R2019b). Unless otherwise stated, boxplots show the relative abundance of lysosomal peptides. The non-parametric Wilcoxon 2-sample rank sum test was used for pairwise comparison between patient and healthy control groups. Statistical significance in this study was defined as $p < 0.05$.

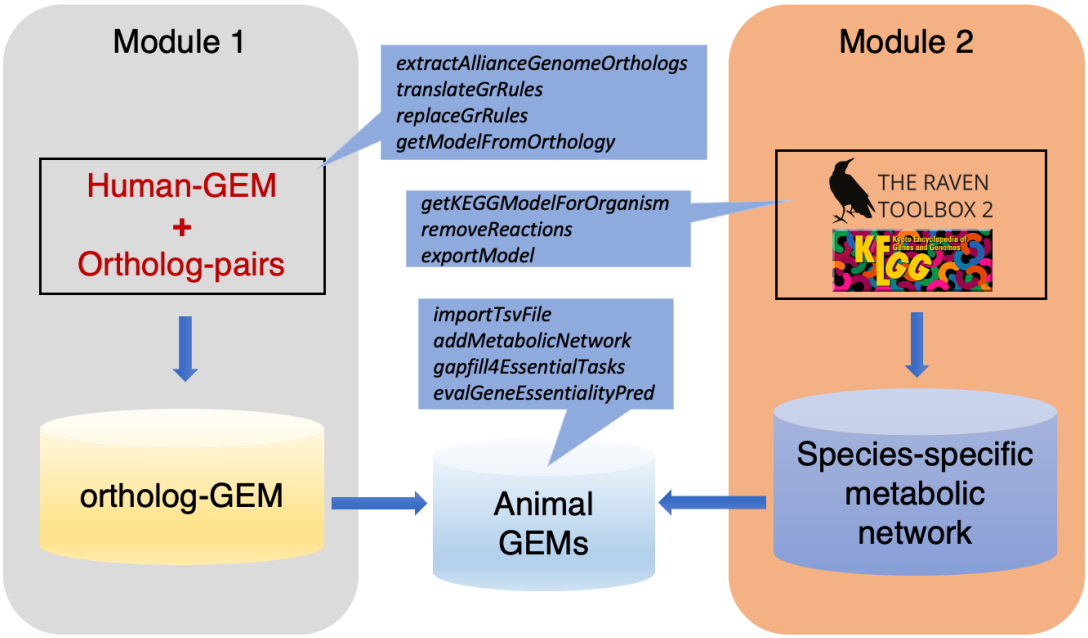

Figure S1. The pipeline for the generation, validation and maintenance of animal GEMs. This pipeline consists of two modules: one for developing the ortholog-GEM based on template Human-GEM (ver 1.5.0) and provided ortholog pairs; another for extracting species-specific pathways by using the RAVEN package and KEGG database. The functions undertaking corresponding steps in the pipeline are shown in italic and detailed in SI text.

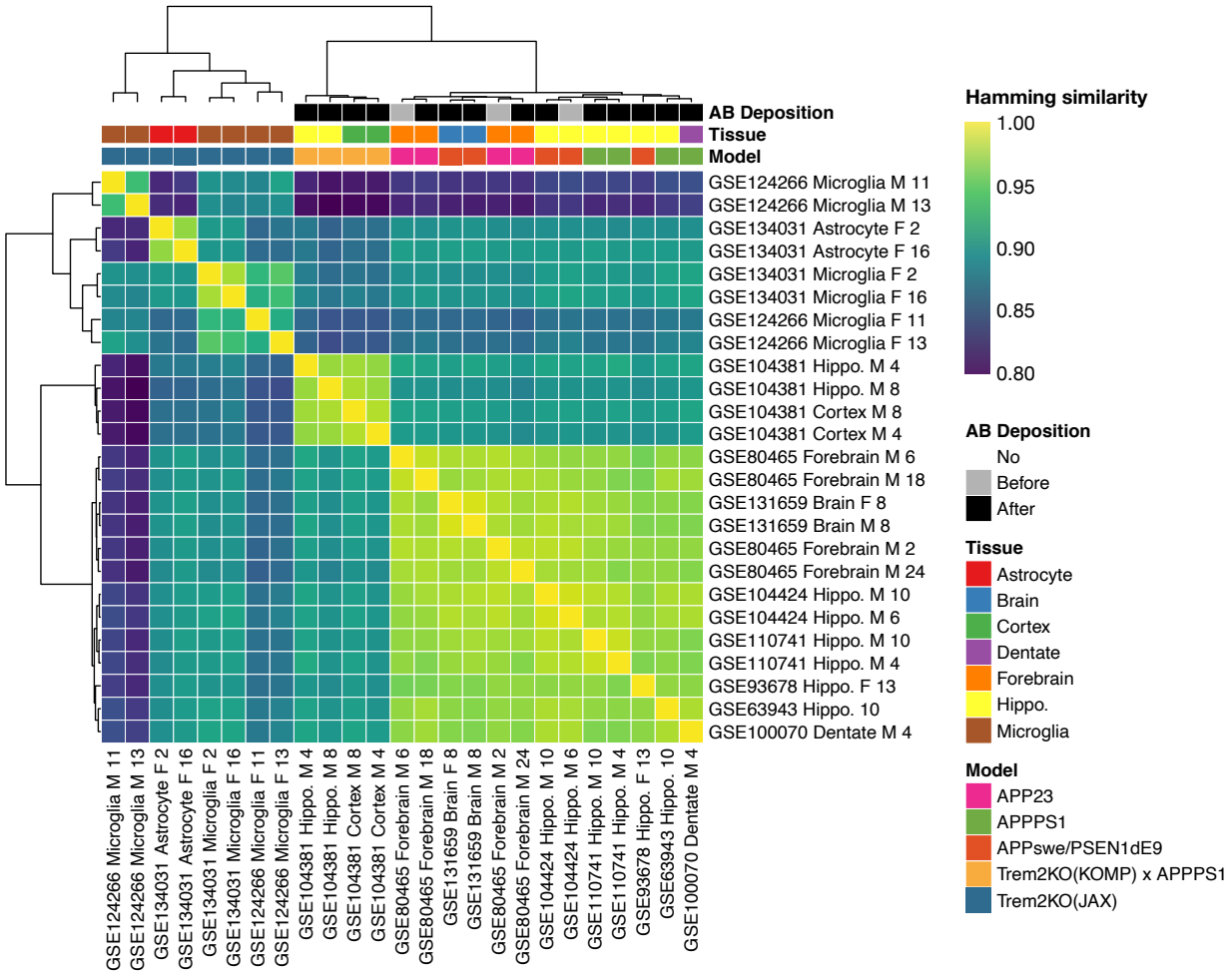

Figure S2. Heatmap comparing the reaction content between tissue- and cell type-specific GEMs reconstructed from RNA-seq data of different AD mouse models.

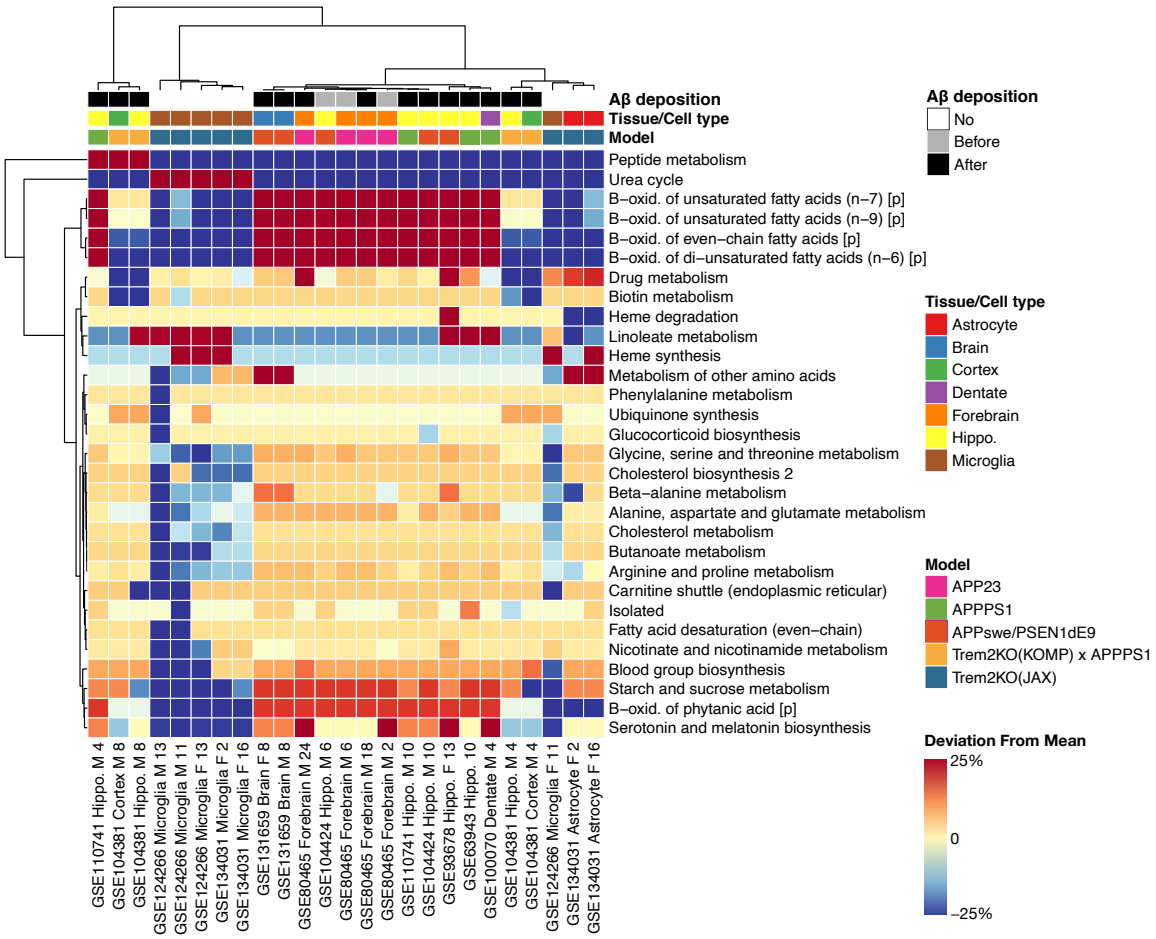

Figure S3. Heatmap shows the variations in subsystem coverage among the tissue/cell type-specific GEMs.

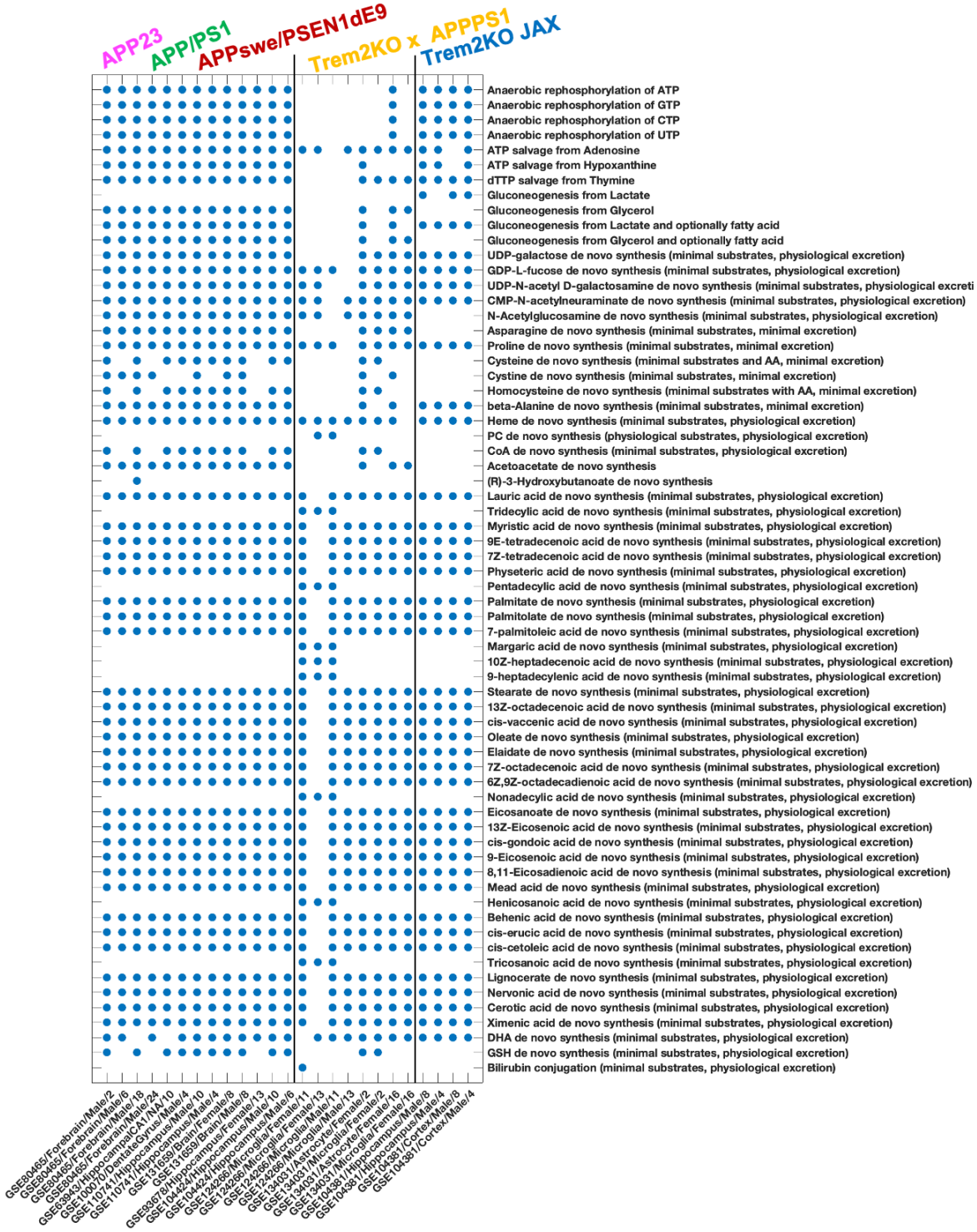

Figure S4. Comparison of metabolic task performance among the tissue- and cell type- specific GEMs. Metabolic tasks are listed along the vertical axis and the samples from which the GEMs were reconstructed are on the horizontal axis. A blue circle represents that the model was able to perform the metabolic task, whereas a white space indicates that the model failed to perform the task.

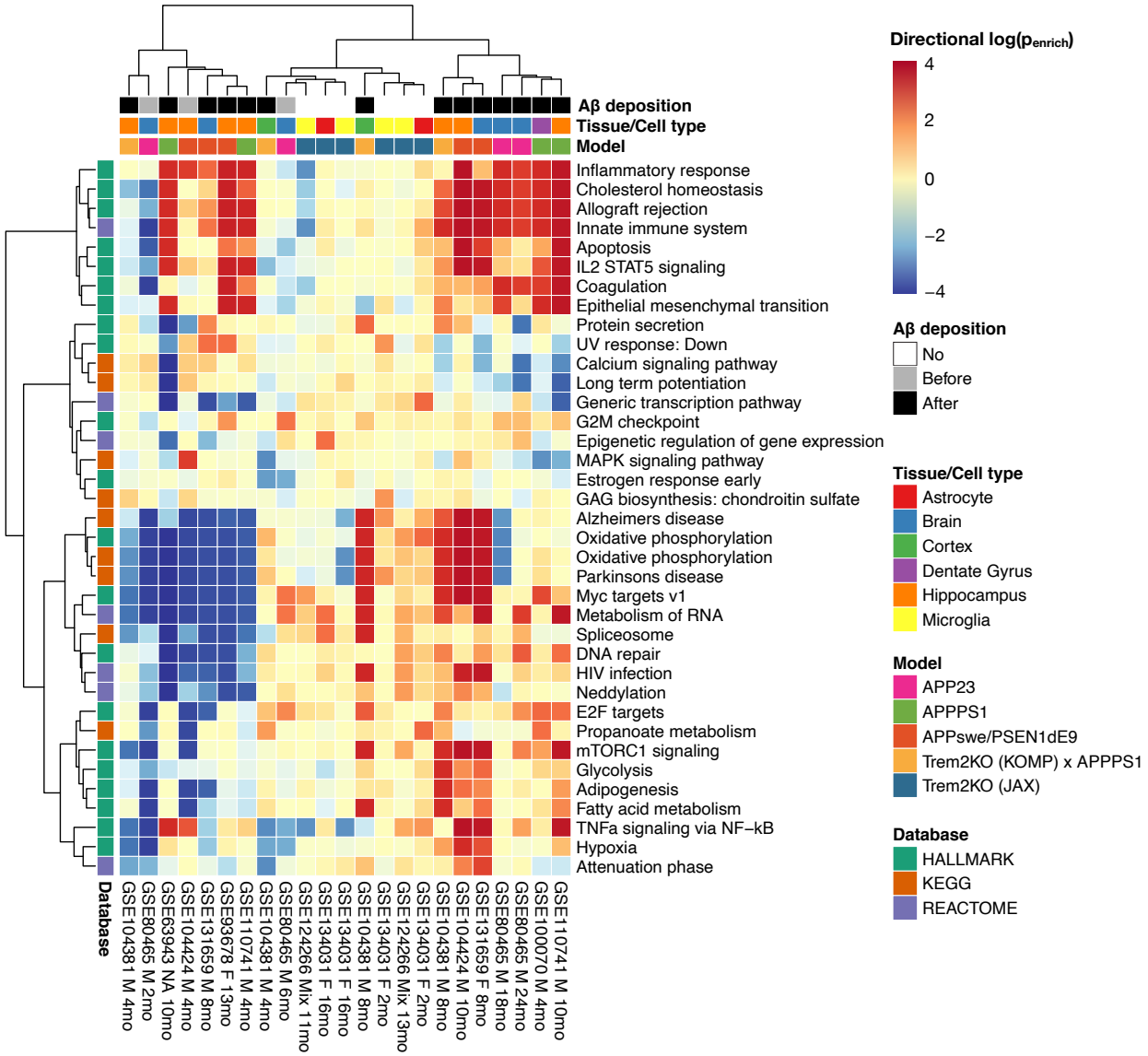

Figure S5. Gene-set analysis of RNA-seq data from different mouse AD models revealed distinct metabolic changes under various experimental conditions. Gene sets from three major biochemical databases (KEGG, Hallmark, and Reactome) were included in the enrichment analysis. The log-transformed P_{enrich} value quantifies the significance of substantially up- (in positive values) or down-regulated (in negative values) gene sets between diseased and normal conditions.

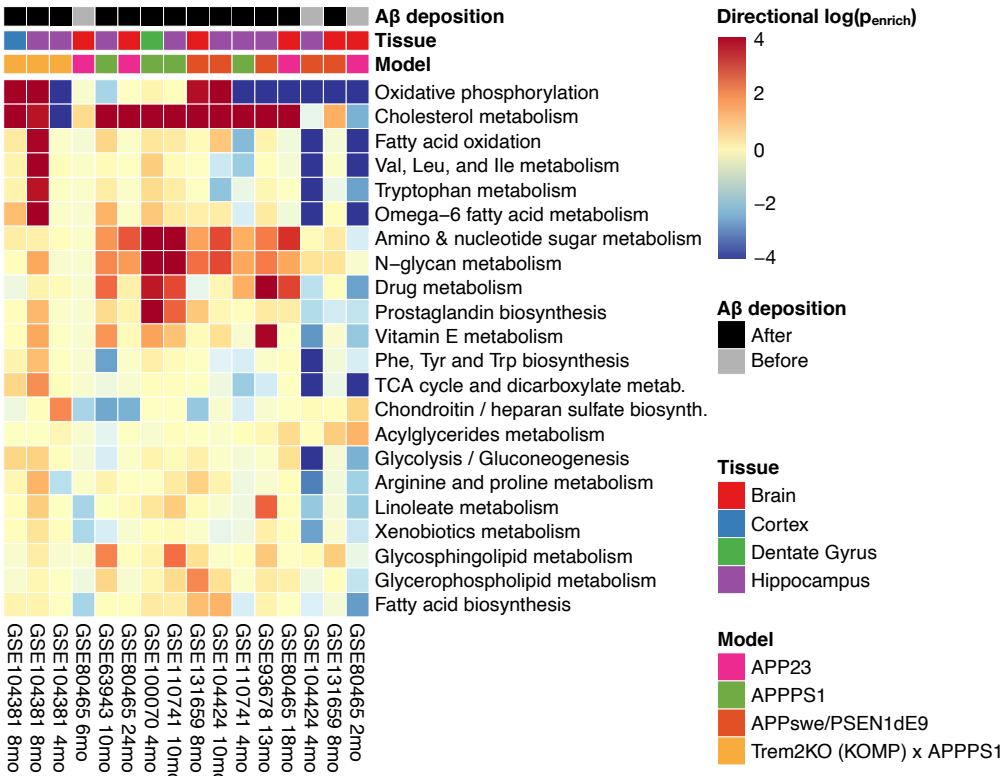

Figure S6. Enrichment analysis of subsystems gene sets extracted from Mouse1. Substantial metabolic changes are observed in the subsystems of oxidative phosphorylation and cholesterol metabolism along with Aβ deposition among the APP overexpression models. The log-transformed P_{enrich} value quantifies the significance of substantially up- (in positive values) or down-regulated (in negative values) gene sets between diseased and normal conditions.

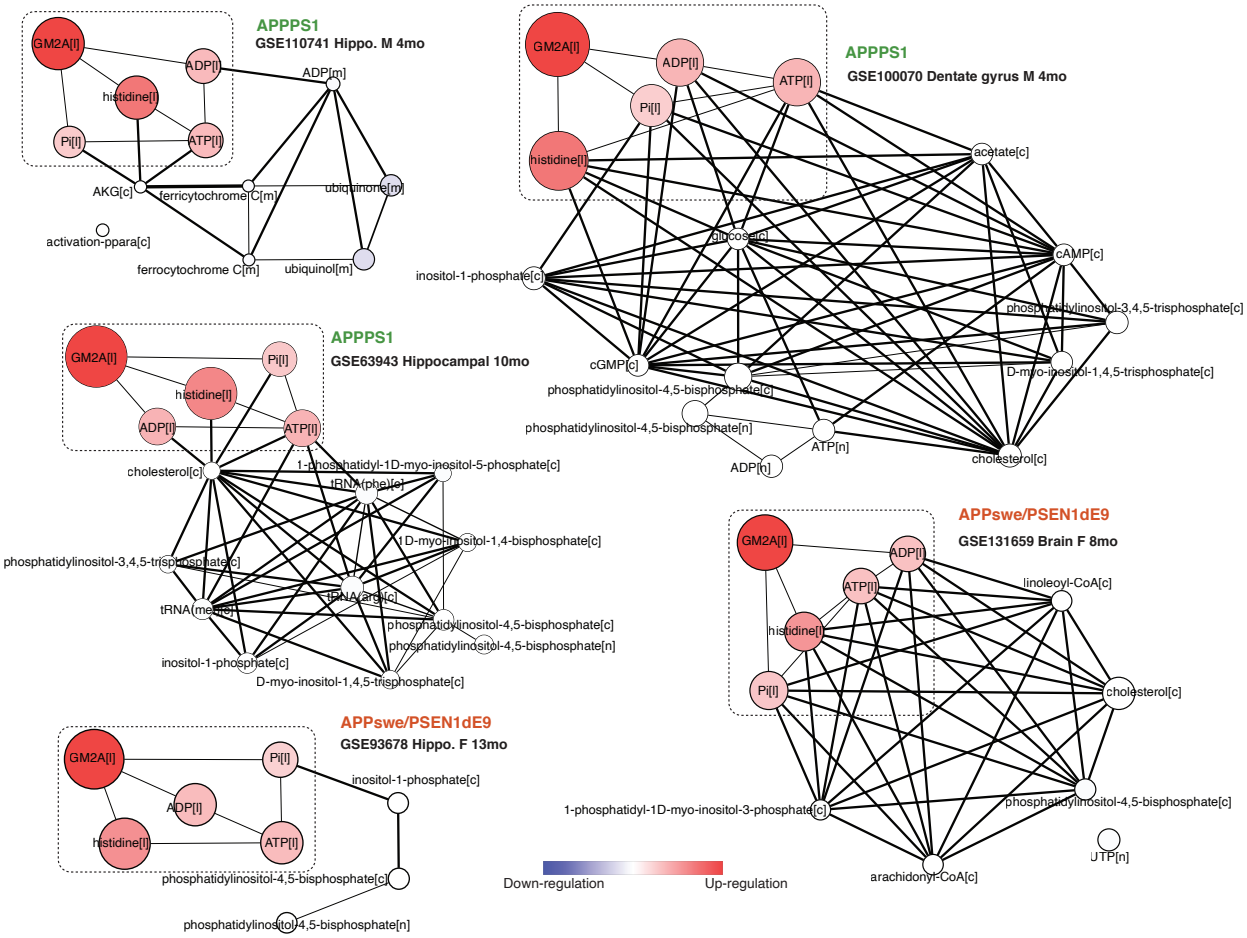

Figure S7. Integrative analysis of reporter metabolite gene sets in APP overexpression mouse models using RNA-seq data sampled after Aβ deposition.

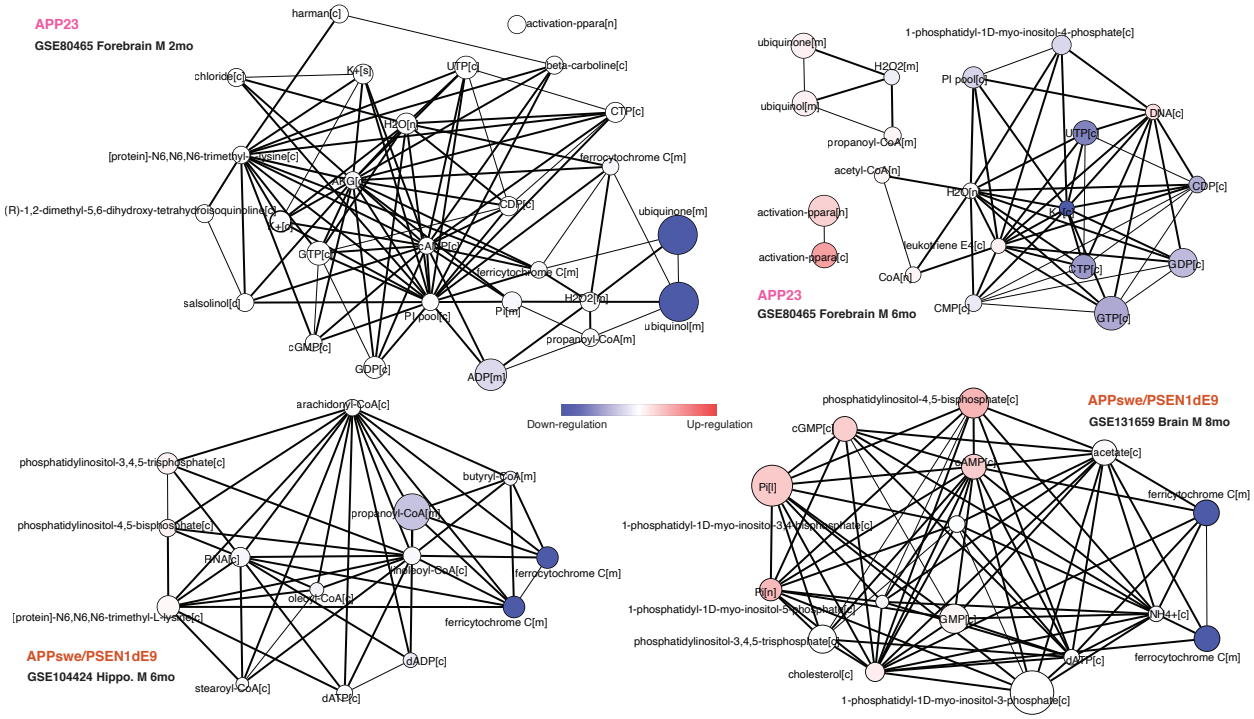

Figure S8. Integrative analysis of reporter metabolite gene sets in APP overexpression mouse models using RNA-seq data sampled prior to Aβ deposition.

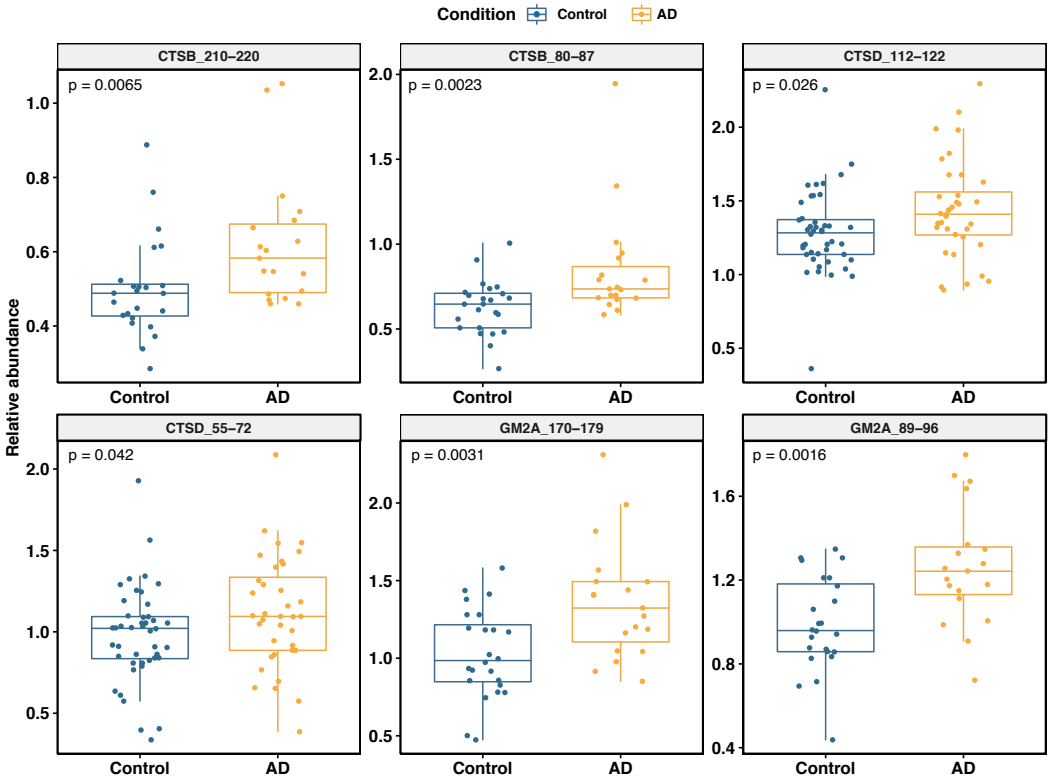

Fig. S9. Enrichment of lysosomal peptides in the cerebrospinal fluid samples of AD patients. Six peptides of lysosomal digestive enzymes exhibited significantly elevated concentrations in cerebrospinal fluid samples of AD patients versus healthy controls. The CTSB peptides 80-87 and 210-220, GM2A peptides 89-96 and 170-179 were measured from the pilot study of ref 20, while CTSD peptides 55-72 and 112-122 were quantified from the clinical study II of ref 20.

References

1. J. L. Robinson, *et al.*, An atlas of human metabolism. *Sci. Signal.* **13**, 1–12 (2020).

2. H. Wang, *et al.*, RAVEN 2.0: A versatile toolbox for metabolic network reconstruction and a case study on Streptomyces coelicolor. *PLoS Comput. Biol.* **14**, 1–17 (2018).

3. J. Agapite, *et al.*, Alliance of Genome Resources Portal: Unified model organism research platform. *Nucleic Acids Res.* **48**, D650–D658 (2020).

4. M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, K. Morishima, KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **45**, D353–D361 (2017).

5. K. R. Patil, J. Nielsen, Uncovering transcriptional regulation of metabolism by using metabolic network topology. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 2685–2689 (2005).

6. R. Henkel, O. Wolkenhauer, D. Waltemath, Combining computational models, semantic annotations and simulation experiments in a graph database. *Database* **2015**, 1–16 (2015).

7. I. Balaur, *et al.*, Recon2Neo4j: Applying graph database technologies for managing comprehensive genome-scale networks. *Bioinformatics* **33**, 1096–1098 (2017).

8. N. Swainston, *et al.*, biochem4j: Integrated and extensible biochemical knowledge through graph databases. *PLoS One* **12**, 1–14 (2017).

9. B. Jassal, *et al.*, The reactome pathway knowledgebase. *Nucleic Acids Res.* **48**, D498–D503 (2020).

10. M. Varadi, *et al.*, PDBE-KB: A community-driven resource for structural and functional annotations. *Nucleic Acids Res.* **48**, D344–D353 (2020).

11. S. Moretti, *et al.*, MetaNetX/MNXref - Reconciliation of metabolites and biochemical reactions to bring together genome-scale metabolic networks. *Nucleic Acids Res.* **44**, D523–D526 (2016).

12. The UniProt Consortium, UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).

13. C. Lieven, *et al.*, MEMOTE for standardized genome-scale metabolic model testing. *Nat. Biotechnol.* **38**, 272–276 (2020).

14. W. H. Chen, G. Lu, X. Chen, X. M. Zhao, P. Bork, OGEE v2: An update of the online gene essentiality database with special focus on differentially essential genes in human cancer cell lines. *Nucleic Acids Res.* **45**, D940–D944 (2017).

15. T. Barrett, *et al.*, NCBI GEO: Archive for functional genomics data sets - Update. *Nucleic Acids Res.* **41**, 991–995 (2013).

16. ALZFORUM, AD research models (2020).

17. D. K. Kim, *et al.*, Deep proteome profiling of the hippocampus in the 5XFAD mouse model reveals biological process alterations and a novel biomarker of Alzheimer's disease. *Exp. Mol. Med.* **51** (2019).

18. J. N. Savas, *et al.*, Amyloid Accumulation Drives Proteome-wide Alterations in Mouse

Models of Alzheimer's Disease-like Pathology. *Cell Rep.* **21**, 2614–2627 (2017).

19. D. K. Kim, *et al.*, Molecular and functional signatures in a novel Alzheimer's disease mouse model assessed by quantitative proteomics. *Mol. Neurodegener.* **13**, 1–19 (2018).

20. S. Sjödin, *et al.*, Endo-lysosomal proteins and ubiquitin CSF concentrations in Alzheimer's and Parkinson's disease. *Alzheimer's Res. Ther.* **11**, 1–16 (2019).

21. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 1–21 (2014).

22. S. Durinck, P. T. Spellman, E. Birney, W. Huber, Mapping identifiers for the integration of genomic datasets with the R/ Bioconductor package biomaRt. *Nat. Protoc.* **4**, 1184–1191 (2009).

23. M. Lawrence, *et al.*, Software for Computing and Annotating Genomic Ranges. *PLoS Comput. Biol.* **9**, 1–10 (2013).

24. L. Väremo, J. Nielsen, I. Nookaew, Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods. *Nucleic Acids Res.* **41**, 4378–4391 (2013).

25. A. Subramanian, *et al.*, Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–15550 (2005).

26. A. Liberzon, *et al.*, The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Syst.* **1**, 417–425 (2015).

27. L. Väremo, F. Gatto, J. Nielsen, Kiwi: A tool for integration and visualization of network topology and gene-set analysis. *BMC Bioinformatics* **15**, 4–9 (2014).

28. M. S. Cline, *et al.*, Integration of biological networks and gene expression data using cytoscape. *Nat. Protoc.* **2**, 2366–2382 (2007).

Dataset S1. Comparative evaluation of model animal GEMs.

Dataset S2. Gene essentiality analysis data and results.

Dataset S3. AD mouse models and RNA-seq data samples investigated in this study.

Dataset S4. Differentially expressed lysosomal enzymes detected from proteomics datasets of APP overexpression mouse models.