File_S1: This file describes the contents of supplementary material associated with ASiDentify.

Table S1 Source for databases used in ASiD and corresponding analyses.

Data	Source	Link	Notes
HGNC Protein-coding genes	hgnc_complete_set_2024-04-01.tsv	https://www.genenames.org/download/archive/	
ASD susceptibility genes	Banerjee-Basu S., and A. Packer, 2010 SFARI Gene: an evolving database for the autism research community. Dis Model Mech 3: 133–135. doi:10.1242/dmm.005439.	https://gene.sfari.org/database/human-gene/	SFARI Gene: Q4 2023
	Trost B., B. Thiruvahindrapuram, A. J. S. Chan, W. Engchuan, E. J. Higginbotham, *et al.*, 2022 Genomic architecture of autism from comprehensive whole-genome sequence annotation. Cell 185: 4409–4427.e18. doi:10.1016/j.cell.2022.10.009.		Table S2
Cell type gene expression	Hodge R. D., T. E. Bakken, J. A. Miller, K. A. Smith, E. R. Barkan, *et al.*, 2019 Conserved cell types with divergent features in human versus mouse cortex. Nature 573: 61–68. doi:10.1038/s41586-019-1506-7.	https://portal.brain-map.org/atlases-and-data/rnaseq/human-multiple-cortical-areas-smart-seq	Gene Expression by Cluster, trimmed means'
Brain region gene expression	GTEx Consortium, 2017 Genetic effects on gene expression across human tissues. Nature 550: 204–213. doi:10.1038/nature24277.	https://gtexportal.org/home/downloads/adult-gtex/bulk_tissue_expression	V8 'GTEx_Analysis_2017-06-05_v8_RNASeQCv1.1.9_gene_median_tpm.gct.gz'
Cortical neuron differentiation gene expression	Burke E. E., J. G. Chenoweth, J. H. Shin, L. Collado-Torres, S.-K. Kim, *et al.*, 2020 Dissecting transcriptomic signatures of neuronal differentiation and maturation using iPSCs. Nat Commun 11: 462. doi:10.1038/s41467-019-14266-z.	https://stemcell.libd.org/scb/data_links.html	Gene Time-Course Data
Brain region protein expression	Carlyle B. C., R. R. Kitchen, J. E. Kanyo, E. Z. Voss, M. Pletikos, *et*		Table S1 & Table S5

	al., 2017 A multiregional proteomic survey of the postnatal human brain. Nat Neurosci 20: 1787–1795. doi:10.1038/s41593-017-0011-2.		
	Li M, Santpere G, Imamura Kawasawa Y, Evgrafov O V., Gulden FO, Pochareddy S, Sunkin SM, Li Zhen, Shin Y, Zhu Y, et al. 2018. Integrative functional genomic analysis of human brain development and neuropsychiatric risks. Science. 362(6420):eaat7615. doi:10.1126/science.aat7615.	https://www.brainspan.org/static/download.html	RNA-Seq Gencode v10 summarized to genes'
		https://download.alleninstitute.org/brainspan/RNASeq_Gencode_v10/	BrainSpan_RNAseq_Specimen_IDs.xlsx'
Mutational constraint	Karczewski K. J., L. C. Francioli, G. Tiao, B. B. Cummings, J. Alföldi, *et al.*, 2020 The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 581: 434–443. doi:10.1038/s41586-020-2308-7.	https://gnomad.broadinstitute.org/downloads	v2.1.1 'pLoF Metrics by Gene TSV'
Synapse localization	Koopmans F., P. van Nierop, M. Andres-Alonso, A. Byrnes, T. Cijsouw, *et al.*, 2019 SynGO: an evidence-based, expert-curated knowledge base for the synapse. Neuron 103: 217–234. doi:10.1016/j.neuron.2019.05.002.	https://www.syngoportal.org/	SynGO release "20231201"
Alzherimer's disease gene set	Bellenguez C., F. Küçükali, I. E. Jansen, L. Kleineidam, S. Moreno-Grau, *et al.*, 2022 New insights into the genetic etiology of Alzheimer's disease and related dementias. Nat Genet 54: 412–436. doi:10.1038/s41588-022-01024-z.		Table S20
Parkinson's disease gene set	Kim J. J., D. Vitale, D. V. Otani, M. M. Lian, K. Heilbron, *et al.*, 2024 Multi-ancestry genome-wide association meta-analysis of Parkinson's disease. Nat		Table S12

	Genet 56: 27–36. doi:10.1038/s41588-023-01584-8.		
Previous ASD gene prediction models	Krishnan A., R. Zhang, V. Yao, C. L. Theesfeld, A. K. Wong, *et al.*, 2016 Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. Nat Neurosci 19: 1454–1462. doi:10.1038/nn.4353		Table S3
	Duda M., H. Zhang, H.-D. Li, D. P. Wall, M. Burmeister, *et al.*, 2018 Brain-specific functional relationship networks inform autism spectrum disorder gene prediction. Transl Psychiatry 8: 56. doi:10.1038/s41398-018-0098-6.		Table S1
	Lin Y., S. Afshar, A. M. Rajadhyaksha, J. B. Potash, and S. Han, 2020 A machine learning approach to predicting autism risk genes: Validation of known genes and discovery of new candidates. Front Genet 11: 500064. doi:10.3389/fgene.2020.500064.		Table S3
	Brueggeman L., T. Koomar, and J. J. Michaelson, 2020 Forecasting risk gene discovery in autism with machine learning and genome-scale data. Sci Rep 10: 4569. doi:10.1038/s41598-020-61288-5.	https://github.com/LeoBman/forecASD	Table S1/forecASD_table.csv
Curated RBP gene set	Gebauer F., T. Schwarzl, J. Valcárcel, and M. W. Hentze, 2021 RNA-binding proteins in human genetic disease. Nat Rev Genet 22: 185–198. doi:10.1038/s41576-020-00302-y.	https://apps.embl.de/rbpbase/	v0.2.1 alpha
Curated CR gene set	Marakulina D., I. E. Vorontsov, I. V Kulakovskiy, A. Lennartsson,	https://epifactors.autosome.org/	2022 release

	F. Drabløs, *et al.*, 2023 EpiFactors 2022: expansion and enhancement of a curated database of human epigenetic factors and complexes. Nucleic Acids Res 51: D564–D570. doi:10.1093/nar/gkac989.		

- Data: Type of data.
- Source: Source for the data, such as websites or papers.
- Link: Link to the data.
- Notes: Additional information, including version or supplementary table from a paper.

Table S2 Canonical RBDs used in RBPbase. Pfam_ID list from 'DOMAIN.BC' file at https://git.embl.de/grp-hentze/grp_rbpbase/rbpbasebackend. Pfam_Name and Pfam_Description were added.

- Pfam_ID: Pfam ID for each canonical RNA-binding domain. From 'DOMAIN.BC' file at https://git.embl.de/grp-hentze/grp_rbpbase/rbpbasebackend.
- Pfam_Name: Name of the Pfam domain. Generated using the *PFAM.db* R package.
- Pfam_Description: Description of the Pfam domain. Generated using the *PFAM.db* R package.

Table S3 Pairwise comparisons of ASiD genes and predictions from Krishnan, Duda, Lin and Brueggeman. Deciles were re-calculated using genes present in both ASiD and each of the other models. Each ASiD decile was compared to each corresponding decile in Krishnan et al. 2016, Duda et al. 2018, Lin et al. 2020 and Brueggeman et al. 2020. We performed one-sided Fisher's exact tests with Bonferroni correction, with the background gene set being the complete intersecting gene lists.

- Paper: The paper used for each pairwise comparison. 'Krishnan' refers to Krishnan et al. 2016. 'Lin' refers to Lin et al. 2020. 'Duda' refers to Duda et al. 2018. 'Brueggeman' refers to Brueggeman et al. 2020.
- Decile: The deciles being compared. Deciles were calculated according to prediction scores from each paper.
- Genes_in_ASiD_decile: The number of genes in the decile from ASiD.
- Genes_in_Paper_decile: The number of genes in the decile from the compared paper. Note: This number will always be the same as 'Genes_in_ASiD_decile' because deciles were calculated using only genes present in both lists of the comparison.
- Intersection: The number of genes present in both deciles.
- Pvalue: The *P*-value calculated from a one-sided Fisher's exact test. The background gene set being the complete intersecting gene lists.
- p_adjusted: The adjusted *P*-value after Bonferroni correction.

Table S4 GO term enrichment analysis of the top decile of ASiD genes with the background genes as all 19,290 protein-coding genes. Benjamini-Hochberg corrections were used for multiple comparisons and GO terms with *P*-values <0.05 were deemed significant. The results were filtered to display terms with a term size <2,000.

- query: List of genes to be queried. The top decile of ASiD gene predictions from all protein-coding genes was used.
- significant: Indicator for statistically significant results. Only significant results, where the threshold used was $P < 0.05$ are shown.
- p_value: Hypergeometric *P*-value after correction for multiple testing.
- term_size: The number of genes that are annotated to the term.
- query_size: The number of genes that were included in the query.
- intersection_size: The number of genes in the query that are annotated to the corresponding term.
- precision: The proportion of genes in the input list that are annotated to the function. Defined as intersection_size/query_size.
- recall: The proportion of functionally annotated genes that the query recovers. Defined as intersection_size/term_size.
- term_id: GO term identifier.
- source: Datasources used for query. Includes: GO:BP, GO:MF, and GO:CC.
- term_name: GO term name.
- effective_domain_size: The total number of genes "in the universe " which is used as one of the four parameters for the hypergeometric probability function of statistical significance.
- source_order: The numeric order for the term within its datasource. Important for drawing reproducible manhattan plots across different platforms.
- parents: List of native IDs that are hierarchically above the term.

Note: the descriptions are adapted from: https://biit.cs.ut.ee/gprofiler/page/apis

PC_genes_output.xlsx ASiD output for all protein-coding genes.

Sheet 1 – 'Predictions': Gene features and ASiD prediction scores for all protein-coding genes.

- Ensembl ID: Ensembl gene identifier.
- HGNC ID: HGNC gene identifier.
- HGNC status: Status of the gene symbol report, which can be either "Approved" or "Entry Withdrawn".
- Gene name: Gene name.
- Gene description: Gene description.
- Excitatory Expression: Mean gene expression ($\log_2$[CPM + 1]) in excitatory neurons. The data is from Allen Brain Map (see Table S1 for details).
- Inhibitory Expression: Mean gene expression ($\log_2$[CPM + 1]) in inhibitory neurons. The data is from Allen Brain Map (see Table S1 for details).

- Astrocyte Expression: Mean gene expression (log_2[CPM + 1]) in astrocytes. The data is from Allen Brain Map (see Table S1 for details).
- Microglia Expression: Mean gene expression (log_2[CPM + 1]) of microglia. The data is from Allen Brain Map (see Table S1 for details).
- Oligodendrocyte Expression: Mean gene expression (log_2[CPM + 1]) in oligodendrocytes. The data is from Allen Brain Map (see Table S1 for details).
- Amygdala expression: Gene expression (log_2[TPM + 1]) in the amygdala. The data is from GTEx (see Table S1 for details).
- Basal ganglia expression: Median gene expression (log_2[TPM + 1]) in the caudate (basal ganglia), nucleus accumbens (basal ganglia), putamen and substantia nigra. The data is from GTEx (see Table S1 for details).
- Cerebellum expression: Median gene expression (log_2[TPM + 1]) in the cerebellar hemisphere and cerebellum. The data is from GTEx (see Table S1 for details).
- Cortex expression: Median gene expression (log_2[TPM + 1]) in the anterior cingulate cortex (BA24), cortex, and frontal Cortex (BA9). The data is from GTEx (see Table S1 for details).
- Hippocampus expression: Gene expression (log_2[TPM + 1]) in the hippocampus. The data is from GTEx (see Table S1 for details).
- Hypothalamus expression: Gene expression (log_2[TPM + 1]) in the hypothalamus. The data is from GTEx (see Table S1 for details).
- Non-brain expression: Median gene expression (log_2[TPM + 1]) in all non-brain tissues. The data is from GTEx (see Table S1 for details).
- Accelerated dorsal expression (D2): Gene expression (log_2[RPKM + 1]) at day 2 of the *in vitro* neuron differentiation time course, calculated by taking the mean of biological replicates and median across cell lines. The data is from Burke et al. 2020 (see Table S1 for details).
- NPC expression (D15): Gene expression (log_2[RPKM + 1]) at day 15 of the *in vitro* neuron differentiation time course, calculated by taking the mean of biological replicates and median across cell lines. The data is from Burke et al. 2020 (see Table S1 for details).
- Neural rosette expression (D21): Gene expression (log_2[RPKM + 1]) at day 21 of the *in vitro* neuron differentiation time course, calculated by taking the mean of biological replicates and median across cell lines. The data is from Burke et al. 2020 (see Table S1 for details).
- Neuron expression (D77): Gene expression (log_2[RPKM + 1]) at day 77 of the *in vitro* neuron differentiation time course, calculated by taking the mean of biological replicates and median across cell lines. The data is from Burke et al. 2020 (see Table S1 for details).
- CBC protein expression: Protein expression (log_{10}) in the cerebellar cortex. The data is from Carlyle et al. 2017 (see Table S1 for details).
- MD protein expression: Protein expression (log_{10}) in the mediodorsal thalamic nucleus. The data is from Carlyle et al. 2017 (see Table S1 for details).
- STR protein expression: Protein expression (log_{10}) in the striatum. The data is from Carlyle et al. 2017 (see Table S1 for details).
- AMY protein expression: Protein expression (log_{10}) in the amygdala. The data is from Carlyle et al. 2017 (see Table S1 for details).
- HIP protein expression: Protein expression (log_{10}) in the hippocampus. The data is from Carlyle et al. 2017 (see Table S1 for details).

- V1C protein expression: Protein expression (log_{10}) in the primary visual cortex. The data is from Carlyle et al. 2017 (see Table S1 for details).
- DFC protein expression: Protein expression (log_{10}) in the dorsolateral prefrontal cortex. The data is from Carlyle et al. 2017 (see Table S1 for details).
- LOEUF: Loss-of-function observed/expected upper bound fraction. Low LOEUF scores indicate selection against predicted loss-of-function variation in a given gene. The data is from gnomAD (see Table S1 for details).
- Synapse localization: A binary feature for whether a gene's protein product localizes to synapses. The data is from SynGO (see Table S1 for details).
- Autism susceptibility: The dependent variable in ASiDentify. There are 1,163 known ASD susceptibility genes, and the remaining 18,127 genes were labelled as non-ASD genes.
- predictions: ASiDentify prediction scores for each gene. Prediction scores are derived from the fold where the gene is held-out in training.

Sheet 2 – 'Odds Ratios': Beta coefficients, and corresponding odds ratios, for each feature with an odds ratio that differs from 1.0 in at least one fold.

- coef: Beta coefficients from ASiDentify.
- variable: The feature for the corresponding beta coefficient.
- OR: Odds ratios calculated from the beta coefficients.
- group: The group of genes the model was run on.

Sheet 3 – 'CIs': Mean odds ratio and 95% confidence interval for features from Sheet 2.

- variable: Feature name.
- Mean: Mean odds ratio of the feature.
- CI_low: Lower limit of the 95% confidence interval.
- CI_high: Upper limit of the 95% confidence interval.
- group: The group of genes the model was run on.

Sheet 4 – 'AUROC_AUPRC': AUROC and AUPRC values for each fold.

- AUROC: The Area Under the Receiver Operating Characteristic for each fold.
- AUPRC: The Area Under the Precision-Recall Curve for each fold.

RBP_output.xlsx Output from modified ASiD models for RBPs.

Sheets 1, 5, 9, and 13 – 'Comprehensive RBP Predictions', 'HC RBP Predictions', 'C RBP Predictions' and 'NC RBP Predictions': Gene features and prediction scores for comprehensive RBPs, high-confidence RBPs, canonical RBPs and non-canonical RBPs, respectively.

- Gene name: Gene name.
- Ensembl ID: Ensembl gene identifier.
- Gene description: Gene description.
- Canonical RBD: Binary data for if the RBP contains a canonical RNA-binding domain (see Table S2).

- Pfam ID: Pfam domain identifier.
- Protein Domains: Pfam domain name.
- Pfam Description: Pfam domain description.
- Excitatory Expression: Mean gene expression (log_2[CPM + 1]) in excitatory neurons. The data is from Allen Brain Map (see Table S1 for details).
- Inhibitory Expression: Mean gene expression (log_2[CPM + 1]) in inhibitory neurons. The data is from Allen Brain Map (see Table S1 for details).
- Astrocyte Expression: Mean gene expression (log_2[CPM + 1]) in astrocytes. The data is from Allen Brain Map (see Table S1 for details).
- Microglia Expression: Mean gene expression (log_2[CPM + 1]) of microglia. The data is from Allen Brain Map (see Table S1 for details).
- Oligodendrocyte Expression: Mean gene expression (log_2[CPM + 1]) in oligodendrocytes. The data is from Allen Brain Map (see Table S1 for details).
- Amygdala expression: Gene expression (log_2[TPM + 1]) in the amygdala. The data is from GTEx (see Table S1 for details).
- Basal ganglia expression: Median gene expression (log_2[TPM + 1]) in the caudate (basal ganglia), nucleus accumbens (basal ganglia), putamen and substantia nigra. The data is from GTEx (see Table S1 for details).
- Cerebellum expression: Median gene expression (log_2[TPM + 1]) in the cerebellar hemisphere and cerebellum. The data is from GTEx (see Table S1 for details).
- Cortex expression: Median gene expression (log_2[TPM + 1]) in the anterior cingulate cortex (BA24), cortex, and frontal Cortex (BA9). The data is from GTEx (see Table S1 for details).
- Hippocampus expression: Gene expression (log_2[TPM + 1]) in the hippocampus. The data is from GTEx (see Table S1 for details).
- Hypothalamus expression: Gene expression (log_2[TPM + 1]) in the hypothalamus. The data is from GTEx (see Table S1 for details).
- Non-brain expression: Median gene expression (log_2[TPM + 1]) in all non-brain tissues. The data is from GTEx (see Table S1 for details).
- Accelerated dorsal expression (D2): Gene expression (log_2[RPKM + 1]) at day 2 of the *in vitro* neuron differentiation time course, calculated by taking the mean of biological replicates and median across cell lines. The data is from Burke et al. 2020 (see Table S1 for details).
- NPC expression (D15): Gene expression (log_2[RPKM + 1]) at day 15 of the *in vitro* neuron differentiation time course, calculated by taking the mean of biological replicates and median across cell lines. The data is from Burke et al. 2020 (see Table S1 for details).
- Neural rosette expression (D21): Gene expression (log_2[RPKM + 1]) at day 21 of the *in vitro* neuron differentiation time course, calculated by taking the mean of biological replicates and median across cell lines. The data is from Burke et al. 2020 (see Table S1 for details).
- Neuron expression (D77): Gene expression (log_2[RPKM + 1]) at day 77 of the *in vitro* neuron differentiation time course, calculated by taking the mean of biological replicates and median across cell lines. The data is from Burke et al. 2020 (see Table S1 for details).
- CBC protein expression: Protein expression (log_{10}) in the cerebellar cortex. The data is from Carlyle et al. 2017 (see Table S1 for details).

- MD protein expression: Protein expression (log_{10}) in the mediodorsal thalamic nucleus. The data is from Carlyle et al. 2017 (see Table S1 for details).
- STR protein expression: Protein expression (log_{10}) in the striatum. The data is from Carlyle et al. 2017 (see Table S1 for details).
- AMY protein expression: Protein expression (log_{10}) in the amygdala. The data is from Carlyle et al. 2017 (see Table S1 for details).
- HIP protein expression: Protein expression (log_{10}) in the hippocampus. The data is from Carlyle et al. 2017 (see Table S1 for details).
- V1C protein expression: Protein expression (log_{10}) in the primary visual cortex. The data is from Carlyle et al. 2017 (see Table S1 for details).
- DFC protein expression: Protein expression (log_{10}) in the dorsolateral prefrontal cortex. The data is from Carlyle et al. 2017 (see Table S1 for details).
- LOEUF: Loss-of-function observed/expected upper bound fraction. Low LOEUF scores indicate selection against predicted loss-of-function variation in a given gene. The data is from gnomAD (see Table S1 for details).
- Synapse localization: A binary feature for whether a gene's protein product localizes to synapses. The data is from SynGO (see Table S1 for details).
- Autism susceptibility: The dependent variable in ASiDentify.
- predictions: ASiDentify prediction scores for each gene. Prediction scores are derived from the fold where the gene is held-out in training.

Sheets 2, 6, 10 and 14 – 'Comprehensive RBP Odds Ratios', 'HC RBP Odds Ratios', 'C RBP Odds Ratios', and 'NC RBP Odds Ratios': Beta coefficients, and corresponding odds ratios, for each feature with an odds ratio that differs from 1.0 in at least one fold for comprehensive RBPs, high-confidence RBPs, canonical RBPs and non-canonical RBPs, respectively.

- coef: Beta coefficients from ASiDentify.
- variable: The feature for the corresponding beta coefficient.
- OR: Odds ratios calculated from the beta coefficients.
- group: The group of genes the model was run on.

Sheets 3, 7, 11, and 15 – 'Comprehensive RBP CIs', 'HC RBP CIs', 'C RBP CIs' and 'NC RBP CIs': Mean odds ratio and 95% confidence interval for features from Sheets 2, 6, 10 and 14, for comprehensive RBPs, high-confidence RBPs, canonical RBPs and non-canonical RBPs, respectively.

- variable: Feature name.
- Mean: Mean odds ratio of the feature.
- CI_low: Lower limit of the 95% confidence interval.
- CI_high: Upper limit of the 95% confidence interval.
- group: The group of genes the model was run on.

Sheets 4, 8, 12 and 16 – 'Comprehensive RBP AUROC_AUPRC', 'HC RBP AUROC_AUPRC', 'C RBP AUROC_AUPRC', and 'NC RBP AUROC_AUPRC': AUROC and AUPRC values for each fold for comprehensive RBPs, high-confidence RBPs, canonical RBPs and non-canonical RBPs, respectively.

- AUROC: The Area Under the Receiver Operating Characteristic for each fold.

- AUPRC: The Area Under the Precision-Recall Curve for each fold.

RBP_RIC_Annotation: details of the evidence supporting an RBP's classification for the 'comprehensive' list of 3,300 RBPs identified using RBPbase.

See file for detailed description.

Random_output.xlsx Ten runs of ASiD model with 82 randomly selected *bona fide* ASD risk genes and 737 randomly selected non-ASD risk genes, from all protein-coding genes.

Sheets 1-10: Mean and 95% confidence intervals for odds ratios of features from all ten random sets of genes. Odds ratios shown for each feature with an odds ratio that differs from 1.0 in at least one fold.

- variable: Feature name.
- Mean: Mean odds ratio of the feature.
- CI_low: Lower limit of the 95% confidence interval.
- CI_high: Upper limit of the 95% confidence interval.
- group: The group of genes the model was run on.

CR_output.xlsx Output from modified ASiD model for chromatin regulators.

Sheet 1 – 'Predictions': Gene features and ASiD prediction scores for chromatin regulators.

- Ensembl ID: Ensembl gene identifier. The column data and description is from EpiFactors (see Table S1 for details).
- HGNC ID: HGNC gene identifier. The column data and description is from EpiFactors (see Table S1 for details).
- Gene name: Gene name. The column data and description is from EpiFactors (see Table S1 for details).
- Gene description: Gene description. The column data and description is from EpiFactors (see Table S1 for details).
- Gene ID: Entrez gene identifier. The column data and description is from EpiFactors (see Table S1 for details).
- UniProt AC: UniProt accession number. The column data and description is from EpiFactors (see Table S1 for details).
- UniProt ID: UniProt identifier. The column data and description is from EpiFactors (see Table S1 for details).
- Domain: Pfam domains. The column data and description is from EpiFactors (see Table S1 for details).
- GeneTag: HGNC gene family tag. The column data and description is from EpiFactors (see Table S1 for details).
- GeneDesc: HGNC gene family description. The column data and description is from EpiFactors (see Table S1 for details).

- Function: General function. The column data and description is from EpiFactors (see Table S1 for details).
- Modification: Which modification the function is targeted towards. The column data and description is from EpiFactors (see Table S1 for details).
- PMID Function: PMID for information on class and function. The column data and description is from EpiFactors (see Table S1 for details).
- Complex name: Protein complex name. The column data and description is from EpiFactors (see Table S1 for details).
- Target: Target molecular (e.g. histone, DNA, RNA). The column data and description is from EpiFactors (see Table S1 for details).
- Specific target: Target entity (e.g. histone with residue). The column data and description is from EpiFactors (see Table S1 for details).
- Product: Product (e.g. type of modification). The column data and description is from EpiFactors (see Table S1 for details).
- PMID target: PMID for reference on target. The column data and description is from EpiFactors (see Table S1 for details).
- Comment: Notes or comments. The column data and description is from EpiFactors (see Table S1 for details).
- Excitatory Expression: Mean gene expression ($\log_2$[CPM + 1]) in excitatory neurons. The data is from Allen Brain Map (see Table S1 for details).
- Inhibitory Expression: Mean gene expression ($\log_2$[CPM + 1]) in inhibitory neurons. The data is from Allen Brain Map (see Table S1 for details).
- Astrocyte Expression: Mean gene expression ($\log_2$[CPM + 1]) in astrocytes. The data is from Allen Brain Map (see Table S1 for details).
- Microglia Expression: Mean gene expression ($\log_2$[CPM + 1]) of microglia. The data is from Allen Brain Map (see Table S1 for details).
- Oligodendrocyte Expression: Mean gene expression ($\log_2$[CPM + 1]) in oligodendrocytes. The data is from Allen Brain Map (see Table S1 for details).
- Amygdala expression: Gene expression ($\log_2$[TPM + 1]) in the amygdala. The data is from GTEx (see Table S1 for details).
- Basal ganglia expression: Median gene expression ($\log_2$[TPM + 1]) in the caudate (basal ganglia), nucleus accumbens (basal ganglia), putamen and substantia nigra. The data is from GTEx (see Table S1 for details).
- Cerebellum expression: Median gene expression ($\log_2$[TPM + 1]) in the cerebellar hemisphere and cerebellum. The data is from GTEx (see Table S1 for details).
- Cortex expression: Median gene expression ($\log_2$[TPM + 1]) in the anterior cingulate cortex (BA24), cortex, and frontal Cortex (BA9). The data is from GTEx (see Table S1 for details).
- Hippocampus expression: Gene expression ($\log_2$[TPM + 1]) in the hippocampus. The data is from GTEx (see Table S1 for details).
- Hypothalamus expression: Gene expression ($\log_2$[TPM + 1]) in the hypothalamus. The data is from GTEx (see Table S1 for details).
- Non-brain expression: Median gene expression ($\log_2$[TPM + 1]) in all non-brain tissues. The data is from GTEx (see Table S1 for details).

- Accelerated dorsal expression (D2): Gene expression (log_2[RPKM + 1]) at day 2 of the *in vitro* neuron differentiation time course, calculated by taking the mean of biological replicates and median across cell lines. The data is from Burke et al. 2020 (see Table S1 for details).
- NPC expression (D15): Gene expression (log_2[RPKM + 1]) at day 15 of the *in vitro* neuron differentiation time course, calculated by taking the mean of biological replicates and median across cell lines. The data is from Burke et al. 2020 (see Table S1 for details).
- Neural rosette expression (D21): Gene expression (log_2[RPKM + 1]) at day 21 of the *in vitro* neuron differentiation time course, calculated by taking the mean of biological replicates and median across cell lines. The data is from Burke et al. 2020 (see Table S1 for details).
- Neuron expression (D77): Gene expression (log_2[RPKM + 1]) at day 77 of the *in vitro* neuron differentiation time course, calculated by taking the mean of biological replicates and median across cell lines. The data is from Burke et al. 2020 (see Table S1 for details).
- CBC protein expression: Protein expression (log_{10}) in the cerebellar cortex. The data is from Carlyle et al. 2017 (see Table S1 for details).
- MD protein expression: Protein expression (log_{10}) in the mediodorsal thalamic nucleus. The data is from Carlyle et al. 2017 (see Table S1 for details).
- STR protein expression: Protein expression (log_{10}) in the striatum. The data is from Carlyle et al. 2017 (see Table S1 for details).
- AMY protein expression: Protein expression (log_{10}) in the amygdala. The data is from Carlyle et al. 2017 (see Table S1 for details).
- HIP protein expression: Protein expression (log_{10}) in the hippocampus. The data is from Carlyle et al. 2017 (see Table S1 for details).
- V1C protein expression: Protein expression (log_{10}) in the primary visual cortex. The data is from Carlyle et al. 2017 (see Table S1 for details).
- DFC protein expression: Protein expression (log_{10}) in the dorsolateral prefrontal cortex. The data is from Carlyle et al. 2017 (see Table S1 for details).
- LOEUF: Loss-of-function observed/expected upper bound fraction. Low LOEUF scores indicate selection against predicted loss-of-function variation in a given gene. The data is from gnomAD (see Table S1 for details).
- Synapse localization: A binary feature for whether a gene's protein product localizes to synapses. The data is from SynGO (see Table S1 for details).
- Autism susceptibility: The dependent variable in ASiDentify.
- predictions: ASiDentify prediction scores for each gene. Prediction scores are derived from the fold where the gene is held-out in training.

Sheet 2 – 'Odds Ratios': Beta coefficients, and corresponding odds ratios, for each feature with an odds ratio that differs from 1.0 in at least one fold.

- coef: Beta coefficients from ASiDentify.
- variable: The feature for the corresponding beta coefficient.
- OR: Odds ratios calculated from the beta coefficients.
- group: The group of genes the model was run on.

Sheet 3 – 'CIs': Mean odds ratio and 95% confidence interval for features from Sheet 2.

- variable: Feature name.
- Mean: Mean odds ratio of the feature.
- CI_low: Lower limit of the 95% confidence interval.
- CI_high: Upper limit of the 95% confidence interval.
- group: The group of genes the model was run on.

Sheet 4 – ‘AUROC_AUPRC’: AUROC and AUPRC values for each fold.

- AUROC: The Area Under the Receiver Operating Characteristic for each fold.
- AUPRC: The Area Under the Precision-Recall Curve for each fold.