nature machine intelligence

Article

https://doi.org/10.1038/s42256-023-00684-8

Hypergraph factorization for multi-tissue gene expression imputation

Received: 21 August 2022

Accepted: 2 June 2023

Published online: 17 July 2023

Check for updates

Ramon Viñas[1], Chaitanya K. Joshi[1], Dobrik Georgiev[1], Phillip Lin[2], Bianca Dumitrascu[3] ✉, Eric R. Gamazon[4] ✉ & Pietro Liò[1] ✉

Integrating gene expression across tissues and cell types is crucial for understanding the coordinated biological mechanisms that drive disease and characterize homoeostasis. However, traditional multi-tissue integration methods either cannot handle uncollected tissues or rely on genotype information, which is often unavailable and subject to privacy concerns. Here we present HYFA (hypergraph factorization), a parameter-efficient graph representation learning approach for joint imputation of multi-tissue and cell-type gene expression. HYFA is genotype agnostic, supports a variable number of collected tissues per individual, and imposes strong inductive biases to leverage the shared regulatory architecture of tissues and genes. In performance comparison on Genotype–Tissue Expression project data, HYFA achieves superior performance over existing methods, especially when multiple reference tissues are available. The HYFA-imputed dataset can be used to identify replicable regulatory genetic variations (expression quantitative trait loci), with substantial gains over the original incomplete dataset. HYFA can accelerate the effective and scalable integration of tissue and cell-type transcriptome biorepositories.

Sequencing technologies have enabled profiling of the transcriptome at tissue and single-cell resolutions, with great potential to unveil intra- and multi-tissue molecular phenomena such as cell signalling and disease mechanisms. Due to the invasiveness of the sampling process, gene expression is usually measured independently in easy-to-acquire tissues, leading to an incomplete picture of an individual's physiological state and necessitating effective multi-tissue integration methodologies.

A question of fundamental biological importance is to what extent the transcriptomes of difficult-to-acquire tissues and cell types can be inferred from those of accessible ones[1,2]. Due to their ease of collection, accessible tissues such as whole blood could have great utility for diagnosis and monitoring of pathophysiological conditions through metabolites, signalling molecules and other biomarkers, including possible transcriptome-level associations[3]. Moreover, all human somatic cells share the same genetic information, which may regulate expression in a context-dependent and temporal manner, partially explaining tissue- and cell-type-specific gene expression variation. Computational models that exploit these patterns could therefore be used to impute the transcriptomes of uncollected cell types and tissues, with potential to elucidate the biological mechanisms regulating a diverse range of developmental and physiological processes.

Multi-tissue imputation is a central problem in transcriptomics with broad implications for fundamental biological research and translational science. The methodological problem can powerfully influence downstream applications, including performing differential expression analysis, identifying regulatory mechanisms, determining co-expression networks and enabling drug target discovery. In practice, in experimental follow-up or clinical application, the task includes the special case of determining a good proxy or easily assayed system for causal tissues and cell types. Multi-tissue integration methods can also be applied to harmonize large collections of RNA-seq datasets

[1]Department of Computer Science and Technology, University of Cambridge, Cambridge, UK. [2]Division of Genetic Medicine, Vanderbilt University Medical Center, Nashville, TN, USA. [3]Department of Statistics and Irving Institute for Cancer Dynamics, Columbia University, New York City, NY, USA. [4]Vanderbilt Genetics Institute and Data Science Institute, MRC Epidemiology Unit, University of Cambridge, Cambridge, UK.
✉e-mail: bianca.dumitrascu@columbia.edu; eric.gamazon@vumc.org; pl219@cam.ac.uk

from diverse institutions, consortia and studies[4]—each potentially affected by technical artifacts—and to characterize gene expression co-regulation across tissues. Reconstruction of unmeasured gene expression across a broad collection of tissues and cell types from available reference transcriptome panels may expand our understanding of the molecular origins of complex traits and of their context specificity.

Several methods have traditionally been employed to impute uncollected gene expression. Leveraging a surrogate tissue has been widely used in studies of biomarker discovery, diagnostics and expression quantitative trait loci (eQTLs), and in the development of model systems[5–9]. Nonetheless, gene expression is known to be tissue and cell-type specific, limiting the utility of a proxy tissue. Other related studies impute tissue-specific gene expression from genetic information[10]. Wang et al.[11] propose a mixed-effects model to infer uncollected data in multiple tissues from eQTLs. Sul et al.[12] introduce a model termed Meta-Tissue, which aggregates information from multiple tissues to increase the statistical power of eQTL detection. However, these approaches do not model the complex nonlinear relationships between measured and unmeasured gene expression traits among tissues and cell types, and individual-level genetic information (for example, at eQTLs) is subject to privacy concerns and often unavailable.

Computationally, multi-tissue transcriptome imputation is challenging because the data dimensionality scales rapidly with the number of genes and tissues, often leading to overparameterized models. TEEBoT[1] addresses this issue by employing principal component analysis—a non-parametric dimensionality reduction method—to project the data into a low-dimensional manifold, followed by linear regression to predict target gene expression from the principal components. However, this technique does not account for nonlinear effects and can only handle a single reference tissue, that is, whole blood. Approaches such as standard multilayer perceptrons (MLPs) can exploit nonlinear patterns, but are massively overparameterized and computationally infeasible.

To address these challenges, we present HYFA (hypergraph factorization), a parameter-efficient graph representation learning approach for joint multi-tissue and cell-type gene expression imputation. HYFA represents multi-tissue gene expression in a hypergraph of individuals, metagenes and tissues, and learns factorized representations via a specialized message-passing neural network operating on the hypergraph. In contrast to existing methods, HYFA supports a variable number of reference tissues, increasing the statistical power over single-tissue approaches, and incorporates inductive biases to exploit the shared regulatory architecture of tissues and genes. In performance comparison, HYFA attains improved performance over TEEBoT and standard imputation methods across a broad range of tissues from the Genotype-Tissue Expression (GTEx) project (v8) (ref. 2). Through transfer learning on a paired single-nucleus RNA-seq dataset (GTEx-v9) (ref. 13), we further demonstrate the ability of HYFA to resolve cell-type signatures—average gene expression across cells for a given cell type, tissue and individual—from bulk gene expression. Thus, HYFA may provide a unifying transcriptomic methodology for multi-tissue imputation and cell-type deconvolution. In post-imputation analysis, application of eQTL mapping on the fully imputed GTEx data yields a substantial increase in number of detected replicable eQTLs. HYFA is publicly available at https://github.com/rvinas/HYFA.

Results

HYFA (hypergraph factorization)

We developed HYFA, a framework for inferring the transcriptomes of unmeasured tissues and cell types from bulk expression collected in a variable number of reference tissues (Fig. 1 and Methods). HYFA receives as input gene expression measurements collected from a set of reference tissues, as well as demographic information, and outputs gene expression values in a tissue of interest (for example uncollected). The first step of the workflow is to project the input gene expression into low-dimensional metagene representations[14,15] for every collected tissue. Each metagene summarizes abstract properties of groups of genes, for example sets of genes that tend to be expressed together[16], that are relevant for the imputation task. In a second step, HYFA employs a custom message-passing neural network[17] that operates on a 3-uniform hypergraph, yielding factorized individual, tissue and metagene representations. Finally, HYFA infers latent metagene values for the target tissue—a hyperedge-level prediction task—and maps these representations back to the original gene expression space. Through higher-order hyperedges (for example, a 4-uniform hypergraph), HYFA can also incorporate cell-type information and infer finer-grained cell-type-specific gene expression (Methods). Altogether, HYFA offers features to reuse knowledge across tissues and genes, capture nonlinear cross-tissue patterns of gene expression, learn rich representations of biological entities and account for variable numbers of reference tissues.

Characterization of cross-tissue relationships

Characterizing cross-tissue relationships at the transcriptome level can help elucidate coordinated gene regulation and expression, a fundamental phenomenon with direct implications for health homoeostasis, disease mechanisms and comorbidities[18–20]. We trained HYFA on bulk gene expression from the GTEx project (GTEx-v8; Methods)[2] and assessed the cross-tissue gene expression predictability—measured using the Pearson correlation between the observed and the predicted gene expression across individuals—and quality of tissue embeddings (Fig. 2). Application of Uniform Manifold Approximation and Projection (UMAP)[21] on the learnt tissue representations revealed strong clustering of biologically related tissues (Fig. 2a), including the gastrointestinal system (for example, oesophageal, stomach, colonic and intestinal tissues), the female reproductive tissues (that is, uterus, vagina and ovary) and the central nervous system (that is, the 13 brain tissues). For every pair of reference and target tissues in GTEx, we then computed the Pearson correlation coefficient ρ between the predicted and actual gene expression, averaged the scores across individuals and used a cutoff of $\rho > 0.5$ to depict the top pairwise associations (Fig. 2b and Extended Data Fig. 1). We observed connections between most GTEx tissues and whole blood, which suggests that blood-derived gene expression is highly informative on (patho)physiological processes in other tissues[22]. Notably, brain tissues and the pituitary gland were strongly associated with several tissues ($\rho > 0.5$), including gastrointestinal tissues (that is oesophagus, stomach and colon), the adrenal gland and skeletal muscle, which may account for known disease comorbidities.

Imputation of gene expression from whole-blood transcriptome

Knowledge about tissue-specific patterns of gene expression can increase our understanding of disease biology, facilitate the development of diagnostic tools and improve patient subtyping[1,23], but most tissues are inaccessible or difficult to acquire. To address this challenge, we studied to what extent HYFA can recover tissue-specific gene expression from whole-blood transcriptomic measurements (Fig. 3). For each test individual with measured whole-blood gene expression, we predicted tissue-specific gene expression in the remaining collected tissues of the individual. We evaluated performance using the Pearson correlation between the inferred gene expression and the ground-truth samples. We observed strong prediction performance for oesophageal tissues (muscularis, $\rho = 0.49$; gastro, $\rho = 0.46$; mucosa, $\rho = 0.36$), heart tissues (left ventricle, $\rho = 0.48$; atrial, $\rho = 0.46$) and lung ($\rho = 0.47$), while Epstein Barr virus-transformed lymphocytes ($\rho = 0.06$), an accessible and renewable resource for functional genomics, was a notable outlier. We noted that the per-gene prediction scores followed smooth, unimodal distributions (Extended Data Fig. 2). The blood-imputed gene expression also predicted disease-relevant genes in the hard-to-access central nervous system (Extended Data Fig. 3). These include *APP*, *PSEN1*

https://doi.org/10.1038/s42256-023-00684-8

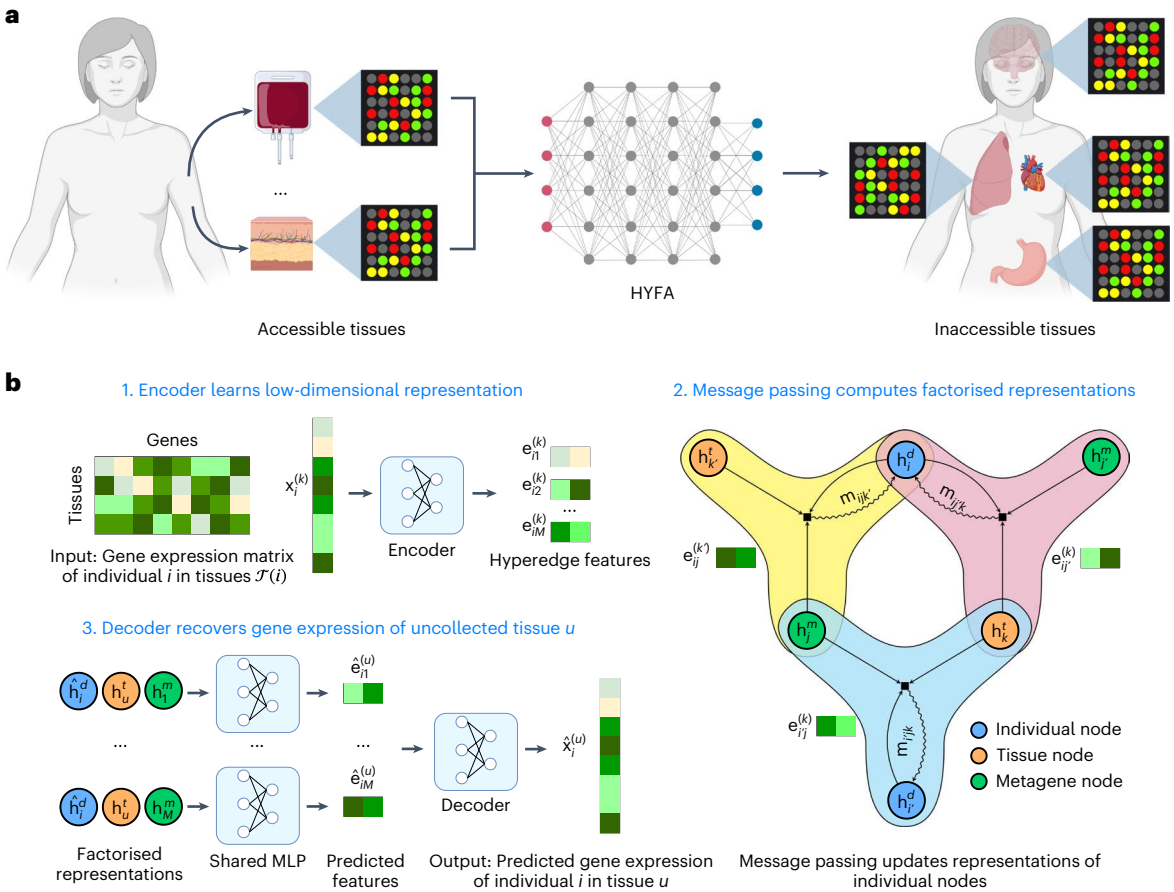


Fig. 1 | Overview of HYFA. a, HYFA processes gene expression from a number of collected tissues (for example, accessible tissues) and infers the transcriptomes of uncollected tissues. **b**, Workflow of HYFA. The model receives as input a variable number of gene expression samples $\mathbf{x}_i^{(k)}$ corresponding to the collected tissues $k \in \mathcal{T}(i)$ of a given individual i. The samples $\mathbf{x}_i^{(k)}$ are fed through an encoder that computes low-dimensional representations $\mathbf{e}_{ij}^{(k)}$ for each metagene $j \in 1, \ldots, M$. A metagene is a latent, low-dimensional representation that captures certain gene expression patterns of the high-dimensional input sample. These representations are then used as hyperedge features in a message-passing neural network that operates on a hypergraph. In the hypergraph representation, each hyperedge labelled with $\mathbf{e}_{ij}^{(k)}$ connects an individual i with metagene j and tissue k if tissue k was collected for individual i, that is $k \in \mathcal{T}(i)$. Through message passing, HYFA learns factorized representations of individual, tissue and metagene nodes. To infer the gene expression of an uncollected tissue u of individual i, the corresponding factorized representations are fed through an MLP that predicts low-dimensional features $\mathbf{e}_{ij}^{(u)}$ for each metagene $j \in 1, \ldots, M$. HYFA finally processes these latent representations through a decoder that recovers the uncollected gene expression sample $\hat{\mathbf{x}}_i^{(u)}$.

and *PSEN2*, that is, the causal genes for autosomal dominant forms of early-onset Alzheimer's disease[24], and Alzheimer's disease genetic risk factors such as *APOE*[25]. We compared our method with TEEBoT[1] (without expression single-nucleotide polymorphism information), which first projects the high-dimensional blood expression data into a low-dimensional space through principal component analysis (30 components; 75–80% explained variance) and then performs linear regression to predict the gene expression of the target tissue. Overall, TEEBoT and HYFA attained comparable scores when a single tissue (that is whole blood) was used as reference and both methods outperformed standard imputation approaches (mean imputation, blood surrogate and k-nearest neighbours; Fig. 3c).

Multiple reference tissues improve performance

We hypothesized that using multiple tissues as reference would improve downstream imputation performance. To evaluate this, we selected individuals with measured gene expression both at the target tissue and four reference accessible tissues (whole blood, skin sun exposed, skin not sun exposed and adipose subcutaneous) and employed HYFA to impute target expression values (Fig. 3 and Extended Data Fig. 4). We discarded under-represented target tissues with fewer than 25 test individuals. Relative to using whole blood in isolation, using all accessible tissues as reference resulted in improved performance for 32 out of 38 target tissues (Extended Data Fig. 4). This particularly boosted imputation performance for oesophageal tissues (muscularis, $\Delta\rho = 0.068$; gastro, $\Delta\rho = 0.061$; mucosa, $\Delta\rho = 0.048$), colonic tissues (transverse, $\Delta\rho = 0.065$; sigmoid, $\Delta\rho = 0.056$) and artery tibial ($\Delta\rho = 0.079$). In contrast, performance for the pituitary gland ($\Delta\rho = -0.011$), lung ($\Delta\rho = -0.003$) and stomach ($\Delta\rho = -0.002$) remained stable or dropped slightly. Moreover, the performance gap between HYFA and TEEBoT (trained on the set of complete multi-tissue samples) widened relative to the single-tissue scenario (Fig. 3 and Extended Data Fig. 5)—HYFA obtained better performance in all target tissues, with statistically significant improvements in 26 out of 38 tissues (two-sided Mann–Whitney–Wilcoxon $P < 0.05$). We attribute the improved scores to HYFA's ability to process a variable number of reference tissues, reuse knowledge across tissues and capture nonlinear patterns.

Inference of cell-type signatures

We next investigated the potential of HYFA to predict cell-type-specific signatures—average gene expression across cells from a given cell type—in a given tissue of interest. We first selected GTEx donors with collected bulk (v8) and single-nucleus RNA-seq profiles (v9, Methods). Next, we trained HYFA to infer cell-type signatures from the multi-tissue bulk

https://doi.org/10.1038/s42256-023-00684-8

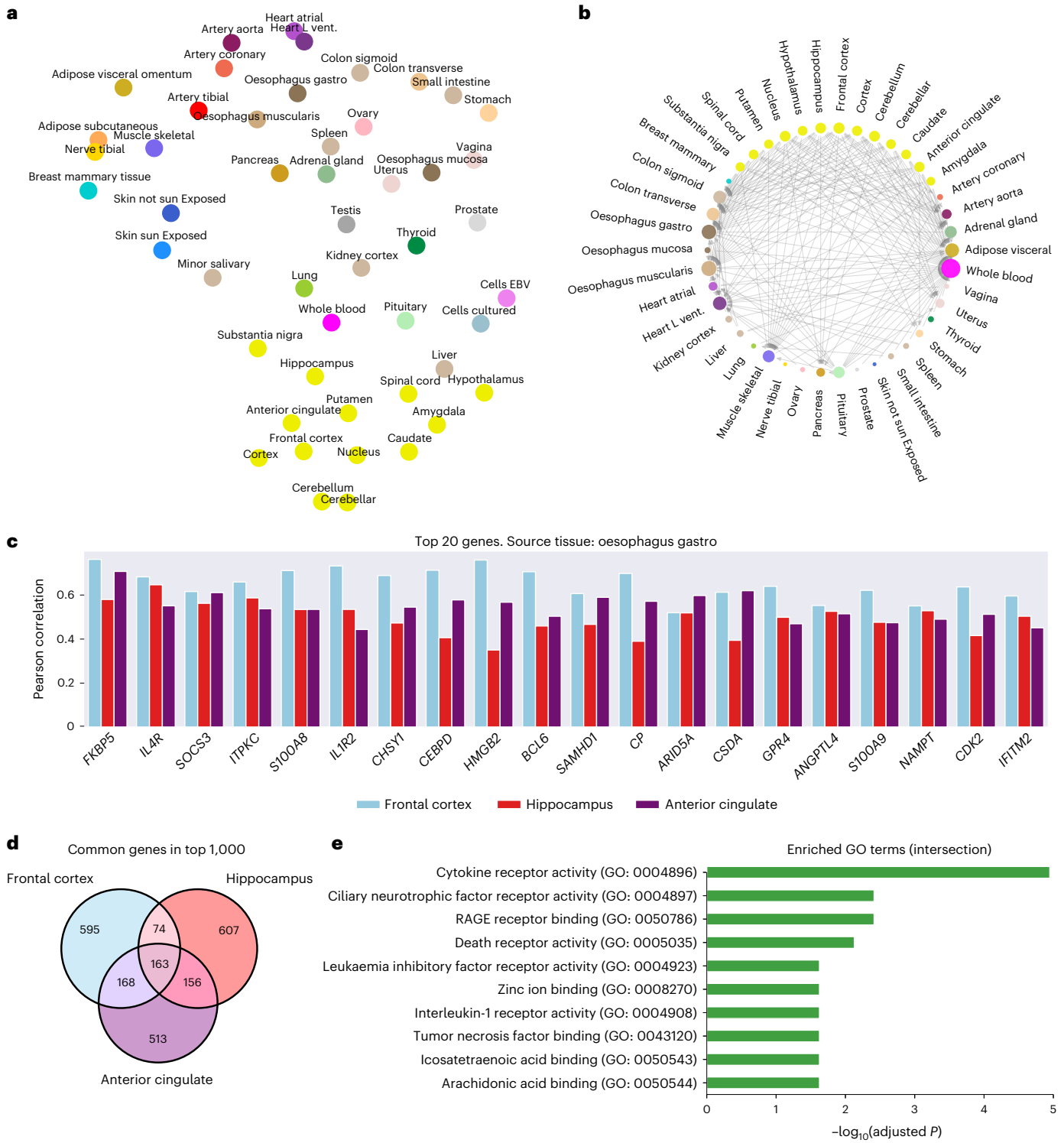


Fig. 2 | Analysis of cross-tissue relationships. a–**e**, Colours are assigned to conform to the GTEx Consortium conventions. **a**, UMAP representation of the tissue embeddings learnt by HYFA. Note that human body systems cluster in the embedding space (for example, digestive system—stomach, small intestine, colon, oesophagus—and central nervous system). EBV, Epstein–Barr virus. **b**, Network of tissues depicting the predictability of target tissues with HYFA using the average per-sample ρ. The dimension of each node is proportional to its degree. Edges from reference to target tissues indicate an average $\rho > 0.5$. Interestingly, central nervous system tissues strongly correlate with several non-brain tissues such as gastrointestinal tissues and skeletal muscles. **c**, Top predicted genes in multiple brain regions with the oesophago-gastric junction as the reference tissue, ranked by average Pearson correlation. **d**, Common genes in the top 1,000 predicted genes for each brain tissue. **e**, Enriched Gene Ontology (GO) terms for the top shared genes at the intersection. The top predicted genes were enriched in signalling pathways (FDR < 0.05), consistent with studies reporting that gut microbes communicate to the central nervous system through endocrine and immune mechanisms. These results depict the cross-tissue associations and highlight the potential connection between the elements of the oesophago-gastric junction and the ciliary neurotrophic factor, which has been linked to the survival of neurons[33] and the control of body weight[35].

https://doi.org/10.1038/s42256-023-00684-8

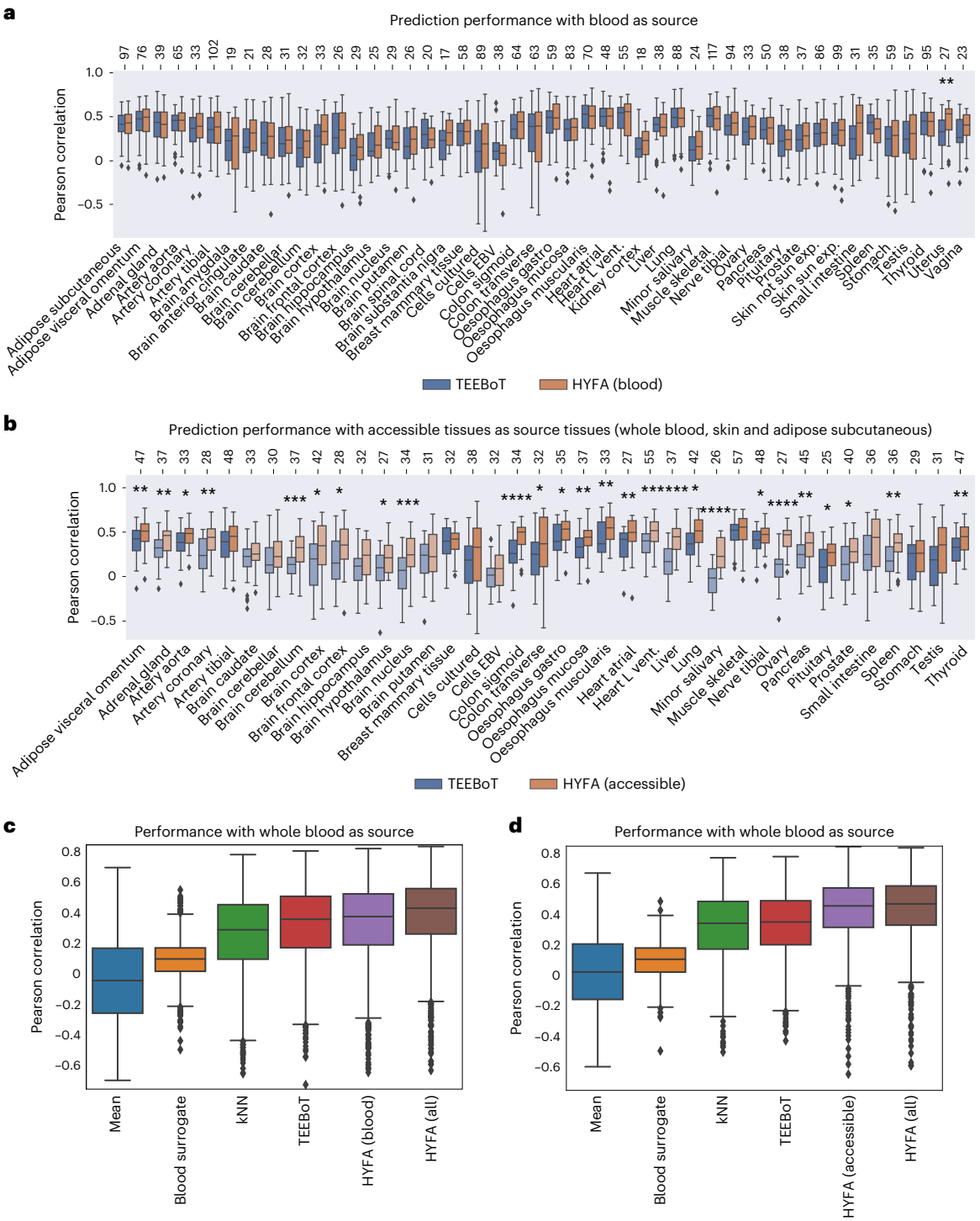


Fig. 3 | Performance comparison across gene expression imputation methods. a,**b**, Per-tissue comparison between HYFA and TEEBoT when using whole blood (**a**) and all accessible tissues (whole blood, skin sun exposed, skin not sun exposed and adipose subcutaneous) (**b**) as reference. HYFA achieved superior Pearson correlation in 25 out of 48 target tissues when a single tissue was used as reference (**a**) and all target tissues when multiple reference tissues were considered (**b**). For under-represented target tissues (fewer than 25 individuals with source and target tissues in the test set), we considered all the validation and test individuals (translucent bars). We employed two-sided Mann–Whitney–Wilcoxon tests to compute *P* values (*$1\times10^{-2} < P \leq 5\times10^{-2}$, **$1\times10^{-3} < P \leq 1\times10^{-2}$, ***$1\times10^{-4} < P \leq 1\times10^{-3}$, ****$P \leq 1\times10^{-4}$). The top axis indicates the total number *n* of independent individuals for every target tissue. **c**,**d**, Prediction performance from whole-blood gene expression ($n = 2{,}424$ samples from 167 GTEx donors) (**c**) and accessible tissues as reference ($n = 675$ samples from 167 test GTEx donors) (**d**). Mean imputation replaces missing values with the feature averages. Blood surrogate utilizes gene expression in whole blood as a proxy for the target tissue. *k*-nearest neighbours (kNN) imputes missing features with the average of measured values across the *k*-nearest observations ($k = 20$). TEEBoT projects reference gene expression into a low-dimensional space with principal component analysis (30 components), followed by linear regression to predict target values. HYFA (all) employs information from all collected tissues of the individual. Boxes show quartiles, centrelines correspond to the median and whiskers depict the distribution range (1.5 times the interquartile range). Outliers outside the whiskers are shown as distinct points.

https://doi.org/10.1038/s42256-023-00684-8

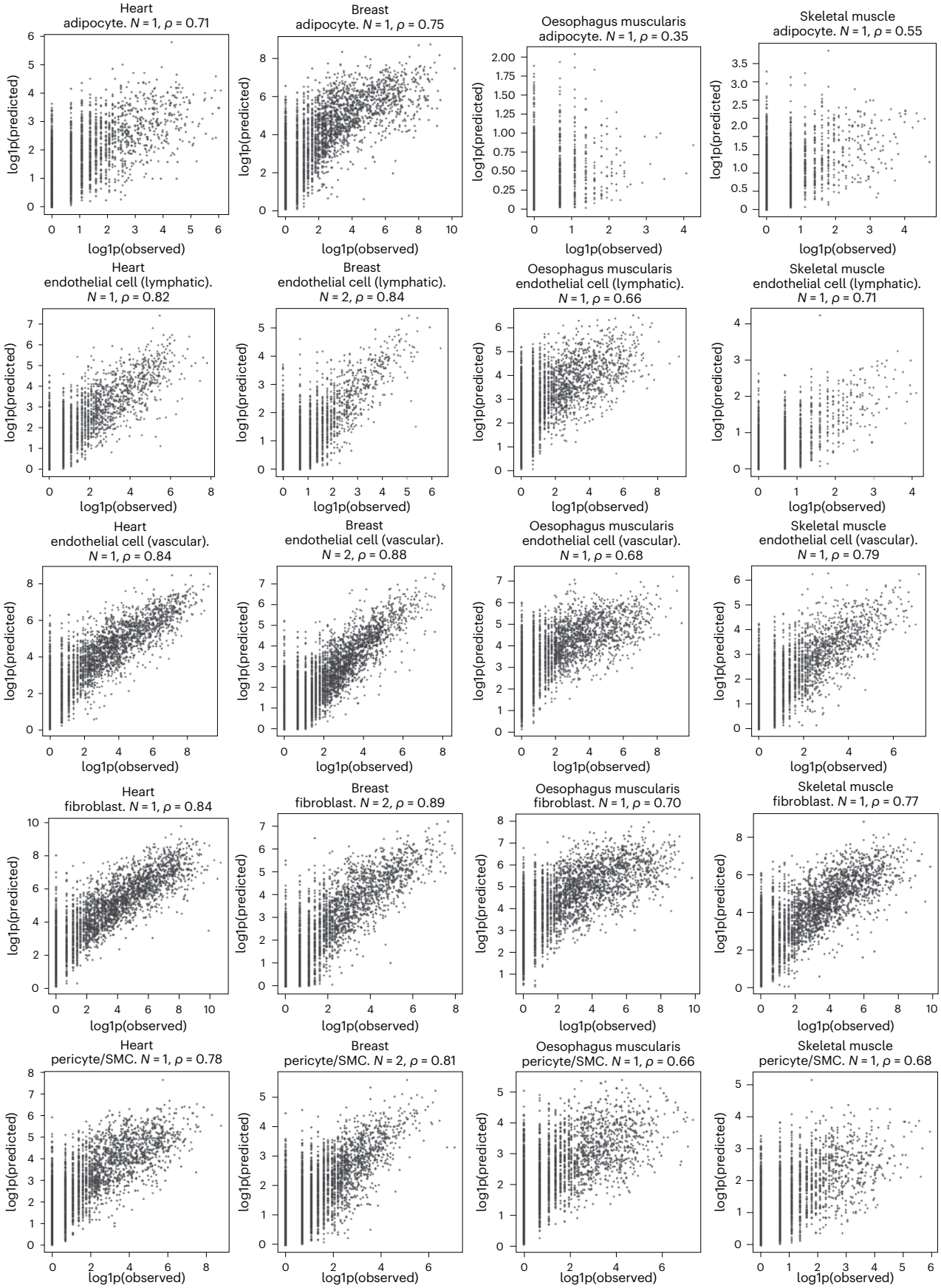


Fig. 4 | Prediction of cell-type signatures. HYFA imputes individual- and tissue-specific cell-type signatures from bulk multi-tissue gene expression. The scatter plots depict the Pearson correlation between the logarithmized ground truth and predicted signatures for N unseen individuals. To infer the signatures, we used the observed library size $l_i^{(k,q)}$ and number of cells $n_i^{(k,q)}$ (Methods). SMC, smooth muscle cell.

expression profiles. We evaluated performance using the observed (Fig. 4) and inferred library sizes (Supplementary Section K). To attenuate the small-data-size problem, we applied transfer learning on the model trained for the multi-tissue imputation task (Methods). We observed strong prediction performance (Pearson correlation ρ between log ground truth and log predicted signatures) for vascular endothelial cells (heart, $\rho = 0.84$; breast, $\rho = 0.88$; oesophagus muscularis, $\rho = 0.68$) and fibroblasts (heart, $\rho = 0.84$; breast, $\rho = 0.89$; oesophagus muscularis, $\rho = 0.70$). Strikingly, HYFA recovered the cell-type profiles of tissues that were never observed in the training set with high correlation (Fig. 4 and Supplementary Section K)—for example, skeletal muscle (vascular endothelial cells, $\rho = 0.79$; fibroblasts, $\rho = 0.77$; pericytes/smooth muscle cells, $\rho = 0.68$), demonstrating the benefits of the factorized tissue representations. Overall, our results highlight the potential of HYFA to impute unknown cell-type signatures even for tissues that were not considered in the original single-cell study. Additionally, our analyses point to promising downstream applications as single-cell RNA-seq datasets become larger in number of individuals (Supplementary Section N), including deconvolution and cell-type-specific eQTL mapping.

Multi-tissue imputation improves eQTL detection

The GTEx project has enabled the identification of numerous genetic associations with gene expression across a broad collection of tissues[2], also known as eQTLs[26]. However, eQTL datasets are characterized by small sample sizes, especially for difficult-to-acquire tissues and cell types, reducing the statistical power to detect eQTLs[27]. To address this problem, we employed HYFA to impute the transcript levels of every uncollected tissue for each individual in GTEx, yielding a complete gene expression dataset of 834 individuals and 49 tissues. We then performed eQTL mapping (Methods) on the original and imputed datasets and observed a substantial gain in the number of unique genes with detected eQTLs, the so-called eGenes (Fig. 5). Notably, this metric increased for tissues with low sample size (Spearman $\rho = -0.83$)—which are most likely to benefit from borrowing information across tissues with shared regulatory architecture. Kidney cortex displayed the largest gain in number of eGenes (from 215 to 12,557), while there was no increase observed for whole blood.

To assess the quality of the identified eQTLs from HYFA imputation, we conducted systematic replication analyses of (1) the whole-blood eQTL–eGene pairs, using the eQTLGen blood transcriptome dataset in more than 30,000 individuals[28], and (2) the frontal cortex eQTL–eGene pairs, using the PsychENCODE prefrontal cortex transcriptome dataset in 1,866 individuals[29]. For each tissue, we quantified the replication rate for eQTL–eGene pairs using the π_1 statistic[30]. Notably, we found a highly significant enrichment for low replication P values among the HYFA-derived eQTL–eGene pairs (Fig. 5), demonstrating strong reproducibility of the results. The replication rate π_1 was 0.80 for whole blood and 0.96 for frontal cortex. We also evaluated the extent to which the HYFA imputation could capture regulatory variants that directly modulate gene expression using experimentally validated causal variants from the Massively Parallel Reporter Assay dataset[31]. Notably, among the causal regulatory variants from this experimental assay, we found a highly significant enrichment for low P values among the HYFA-identified eQTLs in blood and in frontal cortex (Fig. 5). Thus, HYFA imputation enabled identification of biologically meaningful, replicable eQTL hits in the respective tissues. Our results generate a large catalogue of new tissue-specific eQTLs (Data availability), with potential to enhance our understanding of how regulatory variation mediates variation in complex traits, including disease susceptibility.

Brain–gut axis

The brain–gut axis is a bidirectional communication system of signalling pathways linking the central and enteric nervous systems. We investigated whether the transcriptomes of tissues from the gastrointestinal system are predictive of gene expression in brain tissues (Fig. 2 and Supplementary Section G). Overall, the top predicted genes were enriched in multiple signalling-related terms (for example cytokine receptor activity and interleukin-1 receptor activity), consistent with existing knowledge that gut microbes communicate with the central nervous system through signalling mechanisms[32]. Genes in the intersection were also notably enriched in the ciliary neurotrophic factor receptor activity, which plays an important role in neuron survival[33], enteric nervous system development[34] and body weight control[35].

HYFA-learned metagenes capture known biological pathways

A key feature of HYFA is that it reuses knowledge across tissues and metagenes, allowing exploitation of shared regulatory patterns. We explored whether HYFA's inductive biases encourage learning of biologically relevant metagenes. To determine the extent to which metagene factors relate to known biological pathways, we applied gene set enrichment analysis (GSEA)[36] to the gene loadings of HYFA's encoder (Methods). Similarly to ref. 37, for a given query gene set, we calculated the maximum running sum of enrichment scores by descending the sorted list of gene loadings for every metagene and factor. We then computed pathway enrichment P values through a permutation test and employed the Benjamini–Hochberg method to correct for multiple testing independently for every metagene factor.

In total, we identified 18,683 statistically significant enrichments (false discovery rate, FDR < 0.05) of KEGG biological processes[38] (320 gene sets; Fig. 6) across all HYFA metagenes ($n = 50$) and factors ($n = 98$). Among the enriched terms, 2,109 corresponded to signalling pathways and 1,300 to pathways of neurodegeneration. We observed considerable overlap between several metagenes in terms of biologically related pathways: for example, factor 95 of metagene 11 had the lowest FDR for both Alzheimer's disease (FDR < 0.001) and amyotrophic lateral sclerosis (FDR < 0.001) pathways. Enrichment analysis of TRRUST[39] transcription factors (TFs) further identified important regulators including GATA1 (known to regulate the development of red blood cells[40]), SPI1 (which controls haematopoietic cell fate[41]), CEBPs (which play an important role in the differentiation of a range of cell types and the control of tissue-specific gene expression[42,43]) and STAT1 (a member of the STAT protein family that drives the expression of many target genes[44]). We also observed that the learnt HYFA factors recapitulate synergistic effects among the enriched TFs (Supplementary Section H and Extended Data Fig. 6). For example, GATA1 and SPI1, which were simultaneously enriched in 7 factors (FDR < 0.05), functionally antagonize each other through physical interaction[45]. Similarly, IRF1 induces STAT1 activation via phosphorylation[44,46] and both TFs were enriched in 10 factors (FDR < 0.05), aligning with our enrichment analyses of GO biological process terms (Supplementary Section I and Extended Data Figs. 7 and 8). Altogether, our analyses suggest that HYFA-learned metagenes and factors are amenable to biological interpretation and capture information about known regulators of tissue-specific gene expression.

Discussion

Effective multi-tissue omics integration promises a system-wide view of human physiology, with potential to shed light on intra- and multi-tissue molecular phenomena. Such an approach challenges single-tissue and conventional integration techniques—often unable to model a variable number of tissues with sufficient statistical strength, necessitating the development of scalable, nonlinear and flexible methods. Here we developed HYFA, a parameter-efficient approach for joint multi-tissue and cell-type gene expression imputation, which imposes strong inductive biases to learn entity-independent relational semantics and demonstrates excellent imputation capabilities.

We performed extensive benchmarks on data from GTEx[2] (v8 and v9), the most comprehensive human transcriptome resource available, and evaluated imputation performance over a broad collection of tissues and cell types. In addition to standard transcriptome

https://doi.org/10.1038/s42256-023-00684-8

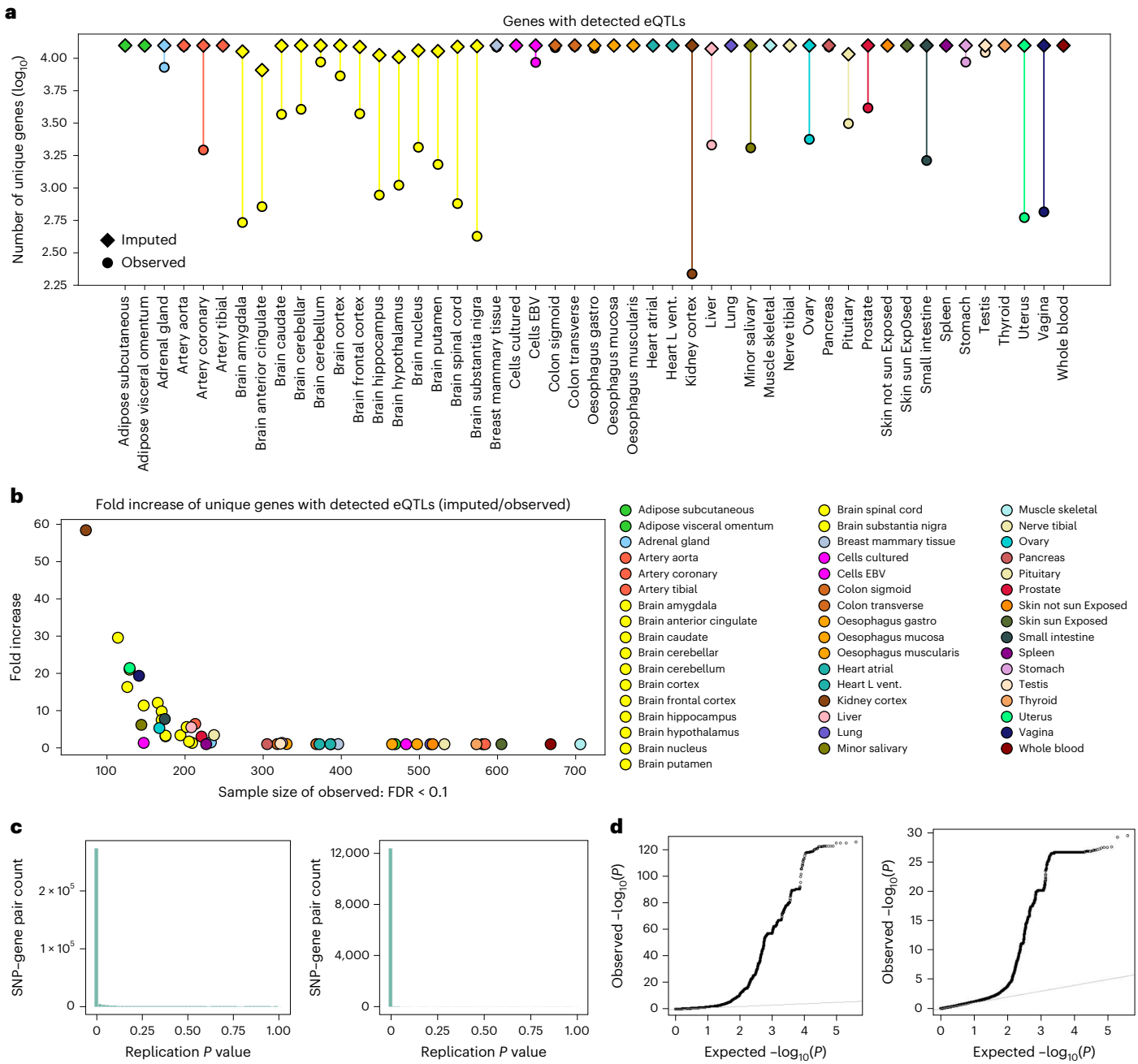


Fig. 5 | HYFA's imputed data improves eQTL discovery. a, Number of unique genes with detected eQTLs (FDR < 0.1) on observed (circle) and full (observed plus imputed; rhombus) GTEx data. Note logarithmic scale of *y* axis. The eQTLs were mapped using Matrix eQTL[55,70] assuming an additive genotype effect on gene expression. Matrix eQTL conducts a test for each single nucleotide polymorphism (SNP)–gene pair and makes adjustments for multiple comparisons by computing the Benjamini–Hochberg FDR[71]. **b**, Fold increase in number of unique genes with mapped eQTLs (*y* axis) versus observed sample size (*x* axis). **c**, Histogram of replication *P* values among the HYFA-identified cis-eQTLs for whole blood (left) and brain prefrontal cortex (right). For replication, we used the independent eQTLGen Consortium ($n > 30{,}000$; ref. 28) and PsychENCODE ($n = 1{,}866$; ref. 29) eQTL datasets, respectively. **d**, Quantile–quantile plot showing the causal variants' association with gene expression in blood (left) and brain frontal cortex (right) in the HYFA-derived dataset using experimentally validated causal variant data from application of the Massively Parallel Reporter Assay dataset[31]. All statistical tests were two sided. HYFA's imputed data substantially increase the number of identified associations with high replicability and strong enrichment of causal regulatory variants.

imputation approaches, we compared our method with TEEBoT[1], a linear method that predicts target gene expression from the principal components of the reference expression. In the single-tissue reference scenario, HYFA and TEEBoT attained comparable imputation performance, outperforming standard methods. In the multi-tissue reference scenario, HYFA consistently outperformed TEEBoT and standard approaches in all target tissues, demonstrating HYFA's capabilities to borrow nonlinear information across a variable number of tissues and exploit shared molecular patterns.

In addition to imputing tissue-level transcriptomics, we investigated the ability of HYFA to predict cell-type-level gene expression from multi-tissue bulk expression measurements. Through transfer learning, we trained HYFA to infer cell-type signatures from a cohort of single-nucleus RNA-seq[13] with matching GTEx-v8 donors. The inferred

https://doi.org/10.1038/s42256-023-00684-8

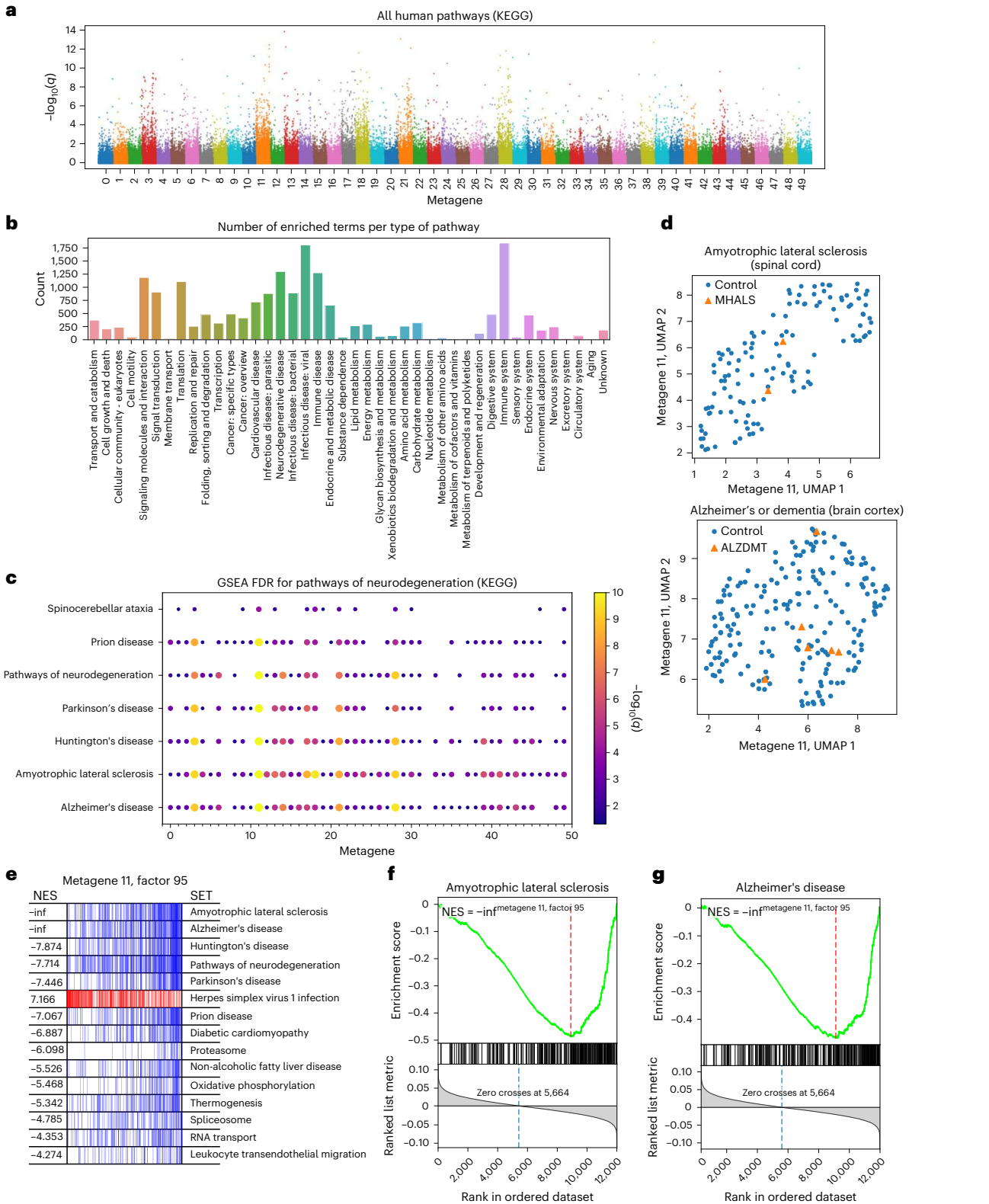


Fig. 6 | Pathway enrichment analysis of metagene factors. a, Manhattan plot of the GSEA results on the metagenes ($n = 50$) and factors ($n = 98$) learned by HYFA. The x axis represents metagenes (coloured bins) and each offset within the bin corresponds to a different factor. The y axis is the −log q value (FDR) from the GSEA permutation test, corrected for multiple testing via the Benjamini–Hochberg procedure. We identified 18,683 statistically significant enrichments (FDR < 0.05) of KEGG biological processes across all metagenes and factors. **b**, Total number of enriched terms for each type of pathway. **c**, FDR for pathways of neurodegeneration. For each pathway and metagene, we selected the factor with the lowest FDR and depicted statistically significant values (FDR < 0.05). Circle sizes are proportional to −log FDR values. Metagene 11 (factor 95) had the lowest FDR for both amyotrophic lateral sclerosis and Alzheimer's disease. **d**, UMAP of latent values of metagene 11 for all spinal cord (amyotrophic lateral sclerosis: orange) and brain cortex (Alzheimer's disease or dementia: orange) GTEx samples. **e**, Leading-edge subsets of top 15 enriched gene sets for factor 95 of metagene 11. NES, normalized enrichment score; SET, gene set. **f**,**g**, Enrichment plots for amyotrophic lateral sclerosis (**f**) and Alzheimer's disease gene sets (**g**).

cell-type signatures exhibited a strong correlation with the ground truth despite the low sample size, indicating that HYFA's latent representations are rich and amenable to knowledge transfer. Strikingly, HYFA also recovered cell-type profiles from tissues that were never observed at transfer time, pointing to HYFA's ability to leverage gene expression programs underlying cell-type identity[47] even in tissues that were not considered in the original study[13]. HYFA may also be used to impute the expression of disease-related genes in a tissue of interest (Supplementary Section J).

In post-imputation analysis, we studied whether the imputed data improve eQTL discovery. We employed HYFA to impute the gene expression levels of every uncollected tissue in GTEx-v8, yielding a complete dataset, and performed eQTL mapping. Compared with the original dataset, we observed a substantial gain in number of genes with detected eQTLs, with kidney cortex showing the largest gain. The increase was highest for tissues with low sample sizes, which are the ones expected to benefit the most from knowledge sharing across tissues. Notably, HYFA's detected eQTLs with their target eGenes could be replicated using independent, single-tissue transcriptome datasets that focus on depth, including the blood eQTLGen[28] and the brain frontal cortex PsychENCODE[29] datasets. Moreover, we found a substantial enrichment for experimentally validated causal variants from the Massively Parallel Reporter Assay[31] dataset. Our results uncover a large number of previously undetected tissue-specific eQTLs and highlight the ability of HYFA to exploit shared regulatory information across tissues.

Finally, HYFA can provide insights on coordinated gene regulation and expression mechanisms across tissues. We analysed to what extent tissues from the gastrointestinal system are informative about gene expression in brain tissues—an important question that may shed light on the biology of the brain–gut axis—and identified enriched biological processes and molecular functions. Through GSEA[36], we observed, among the HYFA-learned metagenes, a substantial number of enriched pathways, TFs and known regulators of biological processes, opening the door to biological interpretations. Future work might also seek to impose stronger inductive bias to ensure that metagenes are identifiable and robust to batch effects.

We believe that HYFA, as a versatile graph representation learning framework, provides a novel methodology for effective integration of large-scale multi-tissue biorepositories. The hypergraph factorization framework is flexible (it supports k-uniform hypergraphs of arbitrary node types) and may find application beyond computational genomics.

Methods

Problem formulation

Suppose we have a transcriptomics dataset of N individuals/donors, T tissues and G genes. For each individual $i \in \{1, \ldots, N\}$, let $\mathbf{X}_i \in \mathbb{R}^{T\times G}$ be the gene expression values in T tissues and define the donor's demographic information by $\mathbf{u}_i \in \mathbb{R}^C$, where C is the number of covariates. Denote by $\mathbf{x}_i^{(k)}$ the kth entry of $\mathbf{X}_i$, corresponding to the expression values of donor i measured in tissue k. For a given donor i, let $\mathcal{T}(i)$ represent the collection of tissues with measured expression values. These sets might vary across individuals. Let $\tilde{\mathbf{X}}_i \in (\mathbb{R} \cup \{*\})^{T\times G}$ be the measured gene expression values, where * denotes unobserved, so that $\tilde{\mathbf{x}}_i^{(k)} = \mathbf{x}_i^{(k)}$ if $k \in \mathcal{T}(i)$ and $\tilde{\mathbf{x}}_i^{(k)} = *$ otherwise. Our goal is to infer the uncollected values in $\tilde{\mathbf{X}}_i$ by modelling the distribution $p(\mathbf{X} = \mathbf{X}_i | \tilde{\mathbf{X}} = \tilde{\mathbf{X}}_i, \mathbf{U} = \mathbf{u}_i)$.

Multi-tissue model

An important challenge of modelling multi-tissue gene expression is that a different set of tissues might be collected for each individual. Moreover, the data dimensionality scales rapidly with the total number of tissues and genes. To address these problems, we represent the data in a hypergraph and develop a parameter-efficient neural network that operates on this hypergraph. Throughout, we make use of the concept of metagenes[14,15]. Each metagene characterizes certain gene expression patterns and is defined as a linear combination of multiple genes[14,15].

Hypergraph representation

We represent the data in a hypergraph consisting of three types of node: donor, tissue and metagene nodes.

Mathematically, we define a hypergraph $\mathcal{G} = \{\mathcal{V}_\mathrm{d} \cup \mathcal{V}_\mathrm{m} \cup \mathcal{V}_\mathrm{t}, \mathcal{E}\}$, where $\mathcal{V}_\mathrm{d}$ is a set of donor nodes, $\mathcal{V}_\mathrm{m}$ is a set of metagene nodes, $\mathcal{V}_\mathrm{t}$ is a set of tissue nodes and $\mathcal{E}$ is a set of multi-attributed hyperedges. Each hyperedge connects an individual i with a metagene j and a tissue k if $k \in \mathcal{T}(i)$, where $\mathcal{T}(i)$ are the collected tissues of individual i. The set of all hyperedges is defined as $\mathcal{E} = \{(i,j,k,\mathbf{e}_{ij}^{(k)}) | (i,j,k) \in \mathcal{V}_\mathrm{d} \times \mathcal{V}_\mathrm{m} \times \mathcal{V}_\mathrm{t}, k \in \mathcal{T}(i)\}$, where $\mathbf{e}_{ij}^{(k)}$ are hyperedge attributes that describe characteristics of the interacting nodes, that is features of metagene j in tissue k for individual i.

The hypergraph allows represention of data in a flexible way, generalizing the bipartite graph representation from ref. 48. On the one hand, using a single metagene results in a bipartite graph where each edge connects an individual i with a tissue k. In this case, the edge attributes $\mathbf{e}_{i1}^{(k)}$ are derived from the gene expression $\mathbf{x}_i^{(k)}$ of individual i in tissue k. On the other hand, using multiple metagenes leads to a hypergraph where each individual i is connected to tissue k through multiple hyperedges. For example, it is possible to construct a hypergraph where genes and metagenes are related by a one-to-one correspondence, with hyperedge attributes $\mathbf{e}_{ij}^{(k)}$ derived directly from expression $x_{ij}^{(k)}$. The number of metagenes thus controls a spectrum of hypergraph representations and, as we shall see, can help alleviate the inherent oversquashing problem of graph neural networks.

Message-passing neural network

Given the hypergraph representation of the multi-tissue transcriptomics dataset, we now present a parameter-efficient graph neural network to learn donor, metagene and tissue embeddings, and infer the expression values of the unmeasured tissues. We start by computing hyperedge attributes from the multi-tissue expression data. Then, we initialize the embeddings of all nodes in the hypergraph, construct the message-passing neural network and define an inference model that builds on the latent node representations obtained via message passing.

Computing hyperedge attributes. We first reduce the dimensionality of the measured transcriptomics values. For every individual i and measured tissue k, we project the corresponding gene expression values $\mathbf{x}_i^{(k)}$ into low-dimensional metagene representations $\mathbf{e}_{ij}^{(k)}$:

$$\mathbf{e}_{ij}^{(k)} = \mathrm{ReLU}\left(\mathbf{W}_j \mathbf{x}_i^{(k)}\right) \quad \forall j \in 1, \ldots, M \tag{1}$$

where M, the number of metagenes, is a user-definable hyperparameter and $\mathbf{W}_j \forall j \in 1, \ldots, M$ are learnable parameters. In addition to characterizing groups of functionally similar genes, employing metagenes reduces the number of messages being aggregated for each node, addressing the oversquashing problem of graph neural networks (Supplementary Section B).

Initial node embeddings. We initialize the node features of the individual $\mathcal{V}_\mathrm{p}$, metagene $\mathcal{V}_\mathrm{m}$ and tissue $\mathcal{V}_\mathrm{t}$ partitions with learnable parameters and available information. For metagene and tissue nodes, we use learnable embeddings as initial node values. The idea is that these weights, which will be approximated through gradient descent, should summarize relevant properties of each metagene and tissue. We initialize the node features of each individual with the available demographic information $\mathbf{u}_i$ of each individual i (we use age and sex). We encode sex as a binary value and age as a float normalized by 100 (for example, age 30 is encoded as 0.30). Importantly, this formulation allows transfer learning between sets of distinct donors.

Message-passing layer. We develop a custom graph neural network layer to compute latent donor embeddings by passing messages along the hypergraph. At each layer of the graph neural network, we perform

message passing to iteratively refine the individual node embeddings. We do not update the tissue and metagene embeddings during message passing, in a similar vein to knowledge graph embeddings[49], because their node embeddings already consist of learnable weights that are updated through gradient descent. Sending messages to these nodes would also introduce a dependence between individual nodes and tissue and metagene features (and, by transitivity, dependences between individuals). However, if we foresee that unseen entities will be present in testing (for example, new tissue types), our approach can be extended by initializing their node features with constant values and introducing node-type-specific message-passing equations.

Mathematically, let $\{\mathbf{h}_1^{\mathrm{d}},\ldots,\mathbf{h}_N^{\mathrm{d}}\}$, $\{\mathbf{h}_1^{\mathrm{m}},\ldots,\mathbf{h}_M^{\mathrm{m}}\}$ and $\{\mathbf{h}_1^{\mathrm{t}},\ldots,\mathbf{h}_T^{\mathrm{t}}\}$ be the donor, metagene and tissue node embeddings, respectively. At each layer of the graph neural network, we compute refined individual embeddings $\{\hat{\mathbf{h}}_1^{\mathrm{d}},\ldots,\hat{\mathbf{h}}_N^{\mathrm{d}}\}$ as follows:

$$\hat{\mathbf{h}}_i^{\mathrm{d}} = \phi_{\mathrm{h}}\left(\mathbf{h}_i^{\mathrm{d}}, \mathbf{m}_i\right), \quad \mathbf{m}_i = \sum_{j=1}^{M}\sum_{k\in\mathcal{T}(i)} \phi_{\mathrm{a}}\left(\mathbf{h}_j^{\mathrm{m}}, \mathbf{h}_k^{\mathrm{t}}, \mathbf{m}_{ijk}\right),$$
$$\mathbf{m}_{ijk} = \phi_{\mathrm{e}}\left(\mathbf{h}_i^{\mathrm{d}}, \mathbf{h}_j^{\mathrm{m}}, \mathbf{h}_k^{\mathrm{t}}, \mathbf{e}_{ij}^{(k)}\right), \tag{2}$$

where the functions ϕ_{e} and ϕ_{h} are edge and node operations that we model as MLPs, and ϕ_{a} is a function that determines the aggregation behaviour. In its simplest form, choosing $\phi_{\mathrm{a}}\left(\mathbf{h}_j^{\mathrm{m}}, \mathbf{h}_k^{\mathrm{t}}, \mathbf{m}_{ijk}\right) = \frac{1}{M|\mathcal{T}(i)|}\mathbf{m}_{ijk}$ results in average aggregation. We analyse the time complexity of the message-passing layer in Supplementary Section A. Optionally, we can stack several message-passing layers to increase the expressivity of the model.

The architecture is flexible and may be extended as follows.

- Incorporation of information about the individual embeddings $\mathbf{h}_i^{\mathrm{d}}$ into the aggregation mechanism ϕ_{a}.
- Incorporation of target tissue embeddings $\mathbf{h}_u^{\mathrm{t}}$, for a given target tissue u, into the aggregation mechanism ϕ_{a}.
- Update hyperedge attributes $\mathbf{e}_{ij}^{(k)}$ at every layer.

Aggregation mechanism. In practice, the proposed hypergraph neural network suffers from a bottleneck. In the aggregation step, the number of messages being aggregated is $M|\mathcal{T}(i)|$ for each individual i. In the worst case, when all genes are used as metagenes (that is, $M = G$; it is estimated that humans have around $G \approx 25{,}000$ protein-coding genes), this leads to serious oversquashing—large amounts of information are compressed into fixed-length vectors[50]. Fortunately, choosing a small number of metagenes reduces the dimensionality of the original transcriptomics values, which in turn alleviates the oversquashing and scalability problems. We perform an ablation study on the number of metagenes and message-passing architectures in Supplementary Section B. To further attenuate oversquashing, we propose an attention-based aggregation mechanism ϕ_{a} that weighs metagenes according to their relevance in each tissue:

$$\phi_{\mathrm{a}}\left(\mathbf{h}_j^{\mathrm{m}}, \mathbf{h}_k^{\mathrm{t}}, \mathbf{m}_{ijk}\right) = \alpha_{jk}\mathbf{m}_{ijk}, \quad \alpha_{jk} = \frac{\exp\left[e\left(\mathbf{h}_j^{\mathrm{m}}, \mathbf{h}_k^{\mathrm{t}}\right)\right]}{\sum_v \exp\left[e\left(\mathbf{h}_v^{\mathrm{m}}, \mathbf{h}_k^{\mathrm{t}}\right)\right]},$$
$$e\left(\mathbf{h}_j^{\mathrm{m}}, \mathbf{h}_k^{\mathrm{t}}\right) = \mathbf{a}^{\mathrm{T}}\,\mathrm{LeakyReLU}\left(\mathbf{W}\left[\mathbf{h}_j^{\mathrm{m}} || \mathbf{h}_k^{\mathrm{t}}\right]\right),$$

where || is the concatenation operation and $\mathbf{a}$ and $\mathbf{W}$ are learnable parameters. The proposed attention mechanism, which closely follows the neighbour aggregation method of graph attention networks[51,52], computes dynamic weighting coefficients that prioritize messages originating from important metagenes. Optionally, we can leverage multiple heads[53] to learn multiple modes of interaction and increase the expressivity of the model.

Hypergraph model. The hypergraph model, which we define as f, computes latent individual embeddings $\hat{\mathbf{h}}_i^{\mathrm{d}}$ from incomplete multi-tissue expression profiles as $\hat{\mathbf{h}}_i^{\mathrm{d}} = f(\tilde{\mathbf{X}}_i, \mathbf{u}_i)$.

Downstream imputation tasks

The resulting donor representations $\hat{\mathbf{h}}_i^{\mathrm{d}}$ summarize information about a variable number of tissue types collected for donor i, in addition to demographic information. We leverage these embeddings for two downstream tasks: inference of gene expression in uncollected tissues and prediction of cell-type signatures.

Inference of gene expression in uncollected tissues

Prediction of the transcriptomic measurements $\hat{\mathbf{x}}_i^{(k)}$ of a tissue k (for example, uncollected) is achieved by first recovering the latent metagene values $\hat{\mathbf{e}}_{ij}^{(k)}$ for all metagenes $j \in 1,\ldots,M$, a hyperedge-level prediction task, and then decoding the gene expression values from the predicted metagene representations $\hat{\mathbf{e}}_{ij}^{(k)}$ with an appropriate probabilistic model.

Prediction of hyperedge attributes. To predict the latent metagene attributes $\hat{\mathbf{e}}_{ij}^{(k)}$ for all $j \in 1,\ldots,M$, we employ an MLP that operates on the factorized metagene $\mathbf{h}_j^{\mathrm{m}}$ and tissue representations $\mathbf{h}_k^{\mathrm{t}}$ as well as the latent variables $\hat{\mathbf{h}}_i^{\mathrm{d}}$ of individual i:

$$\hat{\mathbf{e}}_{ij}^{(k)} = \mathrm{MLP}\left(\hat{\mathbf{h}}_i^{\mathrm{d}}, \mathbf{h}_j^{\mathrm{m}}, \mathbf{h}_k^{\mathrm{t}}\right),$$

where the MLP is shared for all combinations of metagenes, individuals and tissues.

Negative-binomial imputation model. For raw count data, we use a negative-binomial likelihood. To decode the gene expression values for a tissue k of individual i, we define the probabilistic model $p(\mathbf{x}_i^{(k)}|\hat{\mathbf{h}}_i^{\mathrm{d}}, \mathbf{u}_i, k)$:

$$p\left(\mathbf{x}_i^{(k)}|\hat{\mathbf{h}}_i^{\mathrm{d}}, \mathbf{u}_i, k\right) = \prod_j^G p\left(x_{ij}^{(k)}|\hat{\mathbf{h}}_i^{\mathrm{d}}, \mathbf{u}_i, j, k\right),$$
$$p\left(x_{ij}^{(k)}|\hat{\mathbf{h}}_i^{\mathrm{d}}, \mathbf{u}_i, j, k\right) = \mathrm{NB}\left(x_{ij}^{(k)}; \mu_{ij}^{(k)}, \theta_{ij}^{(k)}\right),$$

where NB is a negative-binomial distribution. The mean $\mu_{ij}^{(k)}$ and dispersion $\theta_{ij}^{(k)}$ parameters of this distribution are computed as follows:

$$\boldsymbol{\mu}_i^{(k)} = l_i^{(k)}\mathbf{s}_i^{(k)}, \quad \mathbf{s}_i^{(k)} = \mathrm{softmax}\left(\mathbf{W}_s\hat{\mathbf{e}}_i^{(k)} + \mathbf{b}_s\right),$$
$$\boldsymbol{\theta}_i^{(k)} = \exp\left(\mathbf{W}_\theta\hat{\mathbf{e}}_i^{(k)} + \mathbf{b}_\theta\right), \quad \hat{\mathbf{e}}_i^{(k)} = \mathrm{MLP}\left(\Big\|_{j=1}^{M}\hat{\mathbf{e}}_{ij}^{(k)}\right),$$

where $\mathbf{s}_i^{(k)}$ are mean gene-wise proportions, $\mathbf{W}_s$, $\mathbf{W}_\theta$, $\mathbf{b}_s$ and $\mathbf{b}_\theta$ are learnable parameters and $l_i^{(k)}$ is the library size, which is modelled with a log-normal distribution

$$\log l_i^{(k)} \sim \mathcal{N}\left(l_i^{(k)}; v_i^{(k)}, \omega_i^{(k)}\right), \quad v_i^{(k)} = \mathbf{W}_v\hat{\mathbf{e}}_i^{(k)} + \mathbf{b}_v, \quad \omega_i^{(k)} = \exp\left(\mathbf{W}_\omega\hat{\mathbf{e}}_i^{(k)} + \mathbf{b}_\omega\right),$$

where $\mathbf{W}_v$, $\mathbf{W}_\omega$, $\mathbf{b}_v$ and $\mathbf{b}_\omega$ are learnable parameters. Optionally, we can use the observed library size.

Gaussian imputation model. For normalized gene expression data (that is, inverse normal transformed data), we use the Gaussian likelihood

$$p\left(\mathbf{x}_i^{(k)}|\hat{\mathbf{h}}_i^{\mathrm{d}}, \mathbf{u}_i, k\right) = \prod_j^G p\left(x_{ij}^{(k)}|\hat{\mathbf{h}}_i^{\mathrm{d}}, \mathbf{u}_i, j, k\right),$$
$$p\left(x_{ij}^{(k)}|\hat{\mathbf{h}}_i^{\mathrm{d}}, \mathbf{u}_i, j, k\right) = \mathcal{N}\left(x_{ij}^{(k)}; \mu_{ij}^{(k)}, \sigma^{2(k)}_{ij}\right),$$

where the mean $\mu_{ij}^{(k)}$ and s.d. $\sigma_{ij}^{(k)}$ are computed as follows:

$$\boldsymbol{\mu}_i^{(k)} = \mathbf{W}_\mu\hat{\mathbf{e}}_i^{(k)} + \mathbf{b}_\mu, \quad \boldsymbol{\sigma}_i^{(k)} = \mathrm{softplus}\left(\mathbf{W}_\sigma\hat{\mathbf{e}}_i^{(k)} + \mathbf{b}_\sigma\right),$$
$$\hat{\mathbf{e}}_i^{(k)} = \mathrm{MLP}\left(\Big\|_{j=1}^{M}\hat{\mathbf{e}}_{ij}^{(k)}\right),$$

$\mathbf{W}_\mu$, $\mathbf{W}_\sigma$, $\mathbf{b}_\mu$ and $\mathbf{b}_\sigma$ are learnable parameters and softplus$(x) = \log[1+\exp(x)]$.

Optimization. We optimize the model to maximize the imputation performance on a dynamic subset of observed tissues, that is, tissues that are masked out in training, similarly to ref. 54. For each individual i, we randomly select a subset $\mathcal{C} \subset \mathcal{T}(i)$ of pseudo-observed tissues and treat the remaining tissues $\mathcal{U} = \mathcal{T}(i) - \mathcal{C}$ as unobserved (pseudo-missing). We then compute the individual embeddings $\hat{\mathbf{h}}_i^{\mathrm{d}}$ using the gene expression of pseudo-observed tissues $\mathcal{C}$ and minimize the loss:

$$\mathcal{L}(\tilde{\mathbf{X}}_i, \mathbf{u}_i, \mathcal{C}, \mathcal{U}) = -\frac{1}{|\mathcal{U}|}\sum_{k\in\mathcal{U}} \log p\left(\mathbf{x}_i^{(k)} | \hat{\mathbf{h}}_i^{\mathrm{d}}, \mathbf{u}_i, k\right),$$

which corresponds to the average negative log likelihood across pseudo-missing tissues. Importantly, the pseudo-mask mechanism generates different sets of pseudo-missing tissues for each individual, effectively enlarging the number of training examples and regularizing our model. We summarize the training algorithm in Supplementary Section D.

Inference of gene expression from uncollected tissues. At test time, we infer the gene expression values $\hat{\mathbf{x}}_i^{(v)}$ of an uncollected tissue v from a given donor i via the mean, that is $\hat{\mathbf{x}}_i^{(v)} = \boldsymbol{\mu}_i^{(v)}$. Alternatively, we can draw random samples from the conditional predictive distribution $p(\mathbf{x}_i^{(k)} | \hat{\mathbf{h}}_i^{\mathrm{d}}, \mathbf{u}_i, k)$.

Prediction of cell-type signatures

We next consider the problem of imputing cell-type signatures in a tissue of interest. We define a cell-type signature as the sum of gene expression profiles across cells of a given cell type in a certain tissue. Formally, let $\mathbf{x}_i^{(k,q)}$ be the gene expression signature of cell type q in a tissue of interest k of individual i. Our goal is to infer $\mathbf{x}_i^{(k,q)}$ from the multi-tissue gene expression measurements $\tilde{\mathbf{X}}_i$. To achieve this, we first compute the hyperedge features of a hypergraph consisting of four-node hyperedges and then infer the corresponding signatures with a zero-inflated model.

Prediction of hyperedge attributes. We consider a hypergraph where each hyperedge groups an individual, a tissue, a metagene and a cell-type node. For all metagenes $j \in 1, \ldots, M$, we compute latent hyperedge attributes $\hat{\mathbf{e}}_{ij}^{(k,q)}$ for a cell type q in a tissue of interest k of individual i as follows:

$$\hat{\mathbf{e}}_{ij}^{(k,q)} = \mathrm{MLP}\left(\hat{\mathbf{h}}_i^{\mathrm{d}}, \mathbf{h}_j^{\mathrm{m}}, \mathbf{h}_k^{\mathrm{t}}, \mathbf{h}_q^{\mathrm{c}}\right),$$

where $\mathbf{h}_q^{\mathrm{c}}$ are parameters specific to each unique cell type q and the MLP is shared for all combinations of metagenes, individuals, tissues and cell types.

Zero-inflated model. We employ the following probabilistic model:

$$p\left(\mathbf{x}_i^{(k,q)} | \hat{\mathbf{h}}_i^{\mathrm{d}}, \mathbf{u}_i, k, q\right) = \prod_j^G p\left(x_{ij}^{(k,q)} | \hat{\mathbf{h}}_i^{\mathrm{d}}, \mathbf{u}_i, j, k, q\right),$$

$$p\left(\mathbf{x}_i^{(k,q)} | \hat{\mathbf{h}}_i^{\mathrm{d}}, \mathbf{u}_i, j, k, q\right) = \mathrm{ZINB}\left(x_{ij}^{(k,q)}; \mu_{ij}^{(k,q)}, \theta_{ij}^{(k,q)}, \pi_{ij}^{(k,q)}\right),$$

where ZINB is a zero-inflated negative-binomial distribution. The mean $\mu_{ij}^{(k,q)}$, dispersion $\theta_{ij}^{(k,q)}$ and dropout probability $\pi_{ij}^{(k,q)}$ parameters are computed as

$$\boldsymbol{\mu}_i^{(k,q)} = n_i^{(k,q)} l_i^{(k,q)} \,\mathrm{softmax}\left(\mathbf{W}_s \hat{\mathbf{e}}_i^{(k,q)} + \mathbf{b}_s\right),$$

$$\boldsymbol{\theta}_i^{(k,q)} = \exp\left(\mathbf{W}_\theta \hat{\mathbf{e}}_i^{(k,q)} + \mathbf{b}_\theta\right), \quad \boldsymbol{\pi}_i^{(k,q)} = \sigma\left(\mathbf{W}_\pi \hat{\mathbf{e}}_i^{(k,q)} + \mathbf{b}_\pi\right),$$

where $\mathbf{W}_s$, $\mathrm{W}_{\boldsymbol{\theta}}$, $\mathbf{W}_\pi$, $\mathbf{b}_s$, $\mathbf{b}_\theta$ and $\mathbf{b}_\pi$ are learnable parameters, $n_i^{(k,q)}$ is the number of cells in the signature and $l_i^{(k,q)}$ is their average library size. In training, we set $n_i^{(k,q)}$ to match the ground-truth number of cells. At test time, the number of cells $n_i^{(k,q)}$ is user definable. We model $l_i^{(k,q)}$ with a log-normal distribution

$$\log l_i^{(k,q)} \sim \mathcal{N}\left(l_i^{(k,q)}; v_i^{(k,q)}, \omega_i^{(k,q)}\right), \quad v_i^{(k,q)} = \mathbf{W}_v \hat{\mathbf{e}}_i^{(k,q)} + \mathbf{b}_v,$$

$$\omega_i^{(k,q)} = \exp\left(\mathbf{W}_\omega \hat{\mathbf{e}}_i^{(k,q)} + \mathbf{b}_\omega\right).$$

Optionally, we can use the observed library size.

Optimization. Single-cell transcriptomic studies typically measure single-cell gene expression for a limited number of individuals, tissues and cell types, so aggregating single-cell profiles per individual, tissue and cell type often results in small sample sizes. To address this challenge, we apply transfer learning by pretraining f on the multi-tissue imputation task and then fine-tuning the parameters of the signature inference module on the cell-type signature profiles. Concretely, we minimize the loss:

$$\mathcal{L}\left(\mathbf{x}_i^{(k,q)}, \tilde{\mathbf{X}}_i, \mathbf{u}_i, k, q\right) = -\log p\left(\mathbf{x}_i^{(k,q)} | \hat{\mathbf{h}}_i^{\mathrm{d}}, \mathbf{u}_i, k, q\right),$$

which corresponds to the negative log likelihood of the observed cell-type signatures.

Inference of uncollected gene expression. To infer the signature of a cell type q in a certain tissue v of interest, we first compute the latent individual embeddings $\hat{\mathbf{h}}_i^{\mathrm{d}}$ from the multi-tissue profiles $\tilde{\mathbf{X}}_i$ and then compute the mean of the distribution $p(\mathbf{x}_i^{(k,q)} | \hat{\mathbf{h}}_i^{\mathrm{d}}, \mathbf{u}_i, k, q)$ as $\boldsymbol{\mu}_i^{(k,q)}(1 - \pi_i^{(k,q)})$. Alternatively, we can draw random samples from that distribution.

eQTL mapping

The breadth of tissues in the GTEx-v8 collection enabled us to comprehensively evaluate the extent to which eQTL discovery could be improved through the HYFA-imputed transcriptome data. We mapped eQTLs that act in cis to the target gene (cis-eQTLs), using all single nucleotide polymorphisms within ±1 megabase pairs of the transcription start site of each gene. For the imputed and the original (incomplete) datasets, we considered single nucleotide polymorphisms significantly associated with gene expression, at FDR ≤ 0.10. We applied the same GTEx eQTL mapping pipeline, as previously described[55], to the imputed and original datasets to quantify the gain in eQTL discovery from the HYFA-imputed dataset.

Pathway enrichment analysis

Similarly to ref. 37, we employed GSEA[36] to relate HYFA's metagene factors to known biological pathways. This is advantageous to over-representation analysis, which requires selecting an arbitrary cutoff to select enriched genes. GSEA, instead, computes a running sum of enrichment scores by descending a sorted gene list[36,37].

We applied GSEA to the gene loadings in HYFA's encoder. Specifically, let $\mathbf{W}_j \in \mathbb{R}^{F\times G}$ be the gene loadings for metagene j, where F is the number of factors (that is number of hyperedge attributes) and G is the number of genes (equation (1)). For every factor in $\mathbf{W}_j$, we employed blitzGSEA[56] to calculate the running sum of enrichment scores by descending the gene list sorted by the factor's gene loadings. The enrichment score for a query gene set is the maximum difference between $p_{\mathrm{hit}}(\mathcal{S}, i)$ and $p_{\mathrm{miss}}(\mathcal{S}, i)$ (ref. 37), where $p_{\mathrm{hit}}(\mathcal{S}, i)$ is the proportion of genes in $\mathcal{S}$ weighted by their gene loadings up to gene index i in the sorted list[37]. We then calculated pathway enrichment P values through a permutation test (with $n = 100$ trials) by randomly shuffling the gene list. We employed the Benjamini–Hochberg method to correct for multiple testing.

GTEx bulk and single-nucleus RNA-seq data processing

The GTEx dataset is a public resource that has generated a broad collection of gene expression data collected from a diverse set of human tissues[2]. We downloaded the data from the GTEx portal (Data availability). After the processing step, the GTEx-v8 dataset consisted of 15,197 samples (49 tissues, 834 donors) and 12,557 genes. The dataset was randomly split into 500 training, 167 validation and 167 testing donors. Each donor had an average of 18.22 collected tissues. The processing steps are described below.

Normalized bulk transcriptomics (GTEx-v8). Following the GTEx eQTL discovery pipeline (https://github.com/broadinstitute/gtex-pipeline/tree/master/qtl), we processed the data as follows.

1. Discard under-represented tissues ($n = 5$), namely bladder, cervix (ectocervix, endocervix), fallopian tube and kidney (medulla).
2. Select set of overlapping protein-coding genes across all tissues.
3. Discard donors with only one collected tissue ($n = 4$).
4. Select genes on the basis of expression thresholds of ≥0.1 transcripts per kilobase million in ≥20% of samples and ≥6 reads (unnormalized) in ≥20% of samples.
5. Normalize read counts across samples using the trimmed mean of M values method[57].
6. Apply inverse normal transformation to the expression values for each gene.

Cell-type signatures from a paired snRNA-seq dataset (GTEx-v9). We downloaded paired snRNA-seq data for 16 GTEx individuals[13] (Data availability) collected in eight GTEx tissues, namely skeletal muscle, breast, oesophagus (mucosa, muscularis), heart, lung, prostate and skin. We split these individuals into training, validation and testing donors according to the GTEx-v8 split. We processed the data as follows.

1. Select set of overlapping genes between bulk RNA-seq (GTEx-v9) and paired snRNA-seq dataset[13].
2. Select top 3,000 variable genes using the Scanpy function scanpy.pp.highly_variable_genes with flavour setting seurat_v3 (refs. 58,59).
3. Discard under-represented cell types occurring in fewer than 10 tissue–individual combinations.
4. Aggregate (that is sum) read counts by individual, tissue and (broad) cell type. This resulted in a dataset of 226 unique signatures, of which 135 belong to matching GTEx-v8 individuals.

Implementation and reproducibility

We report the selected hyperparameters in Supplementary Section B. HYFA is implemented in Python[60]. Our framework and implementation are flexible (that is, we support k-uniform hypergraphs), may be integrated in other bioinformatics pipelines and may be useful for other applications in different domains. We used PyTorch[61] to implement the model and Scanpy[58] to process the gene expression data. We performed hyperparameter optimization with wandb[62]. We employed blitzGSEA[56] for pathway enrichment analysis. We also used NumPy[63], scikit-learn[64], pandas[65], matplotlib[66], seaborn[67] and statannotations[68]. Figure 1 was created with BioRender.com.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The datasets analysed for this study, including bulk RNA-seq[2] and snRNA-seq[13], can be found in the GTEx portal: https://gtexportal.org/. We deposited our processed GTEx-v8 data here: https://figshare.com/articles/dataset/Processed_GTEx_v8_data/22650763. A detailed summary of the GTEx samples and donor information can be found at https://gtexportal.org/home/tissueSummaryPage. We downloaded MSK SPECTRUM data from https://cellxgene.cziscience.com/collections/4796c91c-9d8f-4692-be43-347b1727f9d8. We downloaded RNAseqDB data from https://github.com/mskcc/RNAseqDB. The full catalogue of HYFA-derived eQTLs is downloadable at https://doi.org/10.5281/zenodo.6815784.

Code availability

HYFA is publicly available at https://github.com/rvinas/HYFA (ref. 69) (https://doi.org/10.5281/zenodo.7863458).

References

1. Basu, M., Wang, K., Ruppin, E. & Hannenhalli, S. Predicting tissue-specific gene expression from whole blood transcriptome. *Sci. Adv.* **7**, eabd6991 (2021).
2. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
3. Yang, X. et al. High-throughput transcriptome profiling in drug and biomarker discovery. *Front. Genet.* **11**, 19 (2020).
4. Xu, C. et al. Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Mol. Syst. Biol.* **17**, e9620 (2021).
5. Hoon, D. S. et al. Molecular markers in blood as surrogate prognostic indicators of melanoma recurrence. *Cancer Res.* **60**, 2253–2257 (2000).
6. Cai, C. et al. Is human blood a good surrogate for brain tissue in transcriptional studies?. *BMC Genom.* **11**, 589 (2010).
7. Istas, G. et al. Identification of differentially methylated *BRCA1* and *CRISP2* DNA regions as blood surrogate markers for cardiovascular disease. *Sci. Rep.* **7**, 5120 (2017).
8. Gamazon, E. R. et al. Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. *Nat. Genet.* **50**, 956–967 (2018).
9. Kim, K. et al. Clinically accurate diagnosis of Alzheimer's disease via multiplexed sensing of core biomarkers in human plasma. *Nat. Commun.* **11**, 119 (2020).
10. Zhou, D. et al. A unified framework for joint-tissue transcriptome-wide association and Mendelian randomization analysis. *Nat. Genet.* **52**, 1239–1246 (2020).
11. Wang, J. et al. Imputing gene expression in uncollected tissues within and beyond GTEx. *Am. J. Hum. Genet.* **98**, 697–708 (2016).
12. Sul, J. H., Han, B., Ye, C., Choi, T. & Eskin, E. Effectively identifying eQTLs from multiple tissues by combining mixed model and meta-analytic approaches. *PLoS Genet.* **9**, e1003491 (2013).
13. Eraslan, G. et al. Single-nucleus cross-tissue molecular reference maps toward understanding disease gene function. *Science* https://doi.org/10.1126/science.abl4290 (2022).
14. Brunet, J.-P., Tamayo, P., Golub, T. R. & Mesirov, J. P. Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl Acad. Sci. USA* **101**, 4164–4169 (2004).
15. Raychaudhuri, S., Stuart, J. M. & Altman, R. B. Principal components analysis to summarize microarray experiments: application to sporulation time series. In *Biocomputing 2000* (eds Altman, B. et al.) 455–466 (World Scientific, 1999).
16. Svensson, V., Gayoso, A., Yosef, N. & Pachter, L. Interpretable factor models of single-cell RNA-seq via variational autoencoders. *Bioinformatics* **36**, 3418–3421 (2020).
17. Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O. & Dahl, G. E. Neural message passing for quantum chemistry. *Proc. Mach. Learning Res.* **70**, 1263-1272 (2017).

18. Roenneberg, T. & Merrow, M. The circadian clock and human health. *Curr. Biol.* **26**, R432–R443 (2016).
19. Davière, J.-M. & Achard, P. Organ communication: cytokinins on the move. *Nat. Plants* **3**, 17116 (2017).
20. Bodine, S. C. et al. An American Physiological Society cross-journal Call for Papers on "Inter-Organ Communication in Homeostasis and Disease". *Am. J. Physiol. Lung Cell Mol. Physiol.* https://doi.org/10.1152/ajplung.00209.2021 (2021).
21. McInnes et al. UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.* https://doi.org/10.21105/joss.00861 (2018).
22. Ray, S. et al. Classification and prediction of clinical Alzheimer's diagnosis based on plasma signaling proteins. *Nat. Med.* **13**, 1359–1362 (2007).
23. Lage, K. et al. A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. *Proc. Natl Acad. Sci. USA* **105**, 20870–20875 (2008).
24. Lanoiselée, H.-M. et al. *APP*, *PSEN1*, and *PSEN2* mutations in early-onset Alzheimer disease: a genetic screening study of familial and sporadic cases. *PLoS Med.* **14**, e1002270 (2017).
25. Bekris, L. M., Yu, C.-E., Bird, T. D. & Tsuang, D. W. Genetics of Alzheimer disease. *J. Geriatr. Psychiatry Neurol.* **23**, 213–227 (2010).
26. Nica, A. C. & Dermitzakis, E. T. Expression quantitative trait loci: present and future. *Phil. Trans. R. Soc. B* **368**, 20120362 (2013).
27. Rockman, M. V. & Kruglyak, L. Genetics of global gene expression. *Nat. Rev. Genet.* **7**, 862–872 (2006).
28. Võsa, U. et al. Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat. Genet.* **53**, 1300–1310 (2021).
29. Wang, D. et al. Comprehensive functional genomic resource and integrative model for the human brain. *Science* **362**, eaat8464 (2018).
30. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA* **100**, 9440–9445 (2003).
31. Tewhey, R. et al. Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. *Cell* **165**, 1519–1529 (2016).
32. Martin, C. R., Osadchiy, V., Kalani, A. & Mayer, E. A. The brain–gut–microbiome axis. *Cell. Mol. Gastroenterol. Hepatol.* **6**, 133–148 (2018).
33. Davis, S. et al. The receptor for ciliary neurotrophic factor. *Science* **253**, 59–63 (1991).
34. Liu, S. Neurotrophic factors in enteric physiology and pathophysiology. *Neurogastroenterol. Motil.* **30**, e13446 (2018).
35. Xu, B. & Xie, X. Neurotrophic factor control of satiety and body weight. *Nat. Rev. Neurosci.* **17**, 282–292 (2016).
36. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
37. Zhao, Y., Cai, H., Zhang, Z., Tang, J. & Li, Y. Learning interpretable cellular and gene signature embeddings from single-cell transcriptomic data. *Nat. Commun.* **12**, 5261 (2021).
38. Kanehisa, M. et al. KEGG for linking genomes to life and the environment. *Nucleic Acids Res.* **36**, D480–D484 (2007).
39. Han, H. et al. TRRUST: a reference database of human transcriptional regulatory interactions. *Sci. Rep.* **5**, 11432 (2015).
40. Pevny, L. et al. Erythroid differentiation in chimaeric mice blocked by a targeted mutation in the gene for transcription factor GATA-1. *Nature* **349**, 257–260 (1991).
41. Sharrocks, A. D. The ETS-domain transcription factor family. *Nat. Rev. Mol. Cell Biol.* **2**, 827–837 (2001).
42. Wedel, A. & Lömsziegler-Heitbrock, H. The C/EBP family of transcription factors. *Immunobiology* **193**, 171–185 (1995).
43. Nerlov, C. The C/EBP family of transcription factors: a paradigm for interaction between gene expression and proliferation control. *Trends Cell Biol.* **17**, 318–324 (2007).
44. Ramana, C. V., Chatterjee-Kishore, M., Nguyen, H. & Stark, G. R. Complex roles of Stat1 in regulating gene expression. *Oncogene* **19**, 2619–2627 (2000).
45. Nerlov, C., Querfurth, E., Kulessa, H. & Graf, T. GATA-1 interacts with the myeloid PU.1 transcription factor and represses PU.1-dependent transcription. *Blood* **95**, 2543–2551 (2000).
46. Zenke, K., Muroi, M. & Tanamoto, K.-i IRF1 supports DNA binding of STAT1 by promoting its phosphorylation. *Immunol. Cell Biol.* **96**, 1095–1103 (2018).
47. Kotliar, D. et al. Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-seq. *eLife* **8**, e43803 (2019).
48. You, J., Ma, X., Ding, D., Kochenderfer, M. & Leskovec, J. Handling missing data with graph representation learning. In *NIPS'20: Proc. 34th International Conference on Neural Information Processing Systems* (eds Larochelle, H. et al.) 19075–19087 (Curran, 2020).
49. Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J. & Yakhnenko, O. Translating embeddings for modeling multi-relational data. In *NIPS'13: Proc. 26th International Conference on Neural Information Processing Systems* Vol. 26 (eds Burges, C. J. C. et al.) 2787–2795 (Curran, 2013).
50. Alon, U. & Yahav, E. On the bottleneck of graph neural networks and its practical implications. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2006.05205 (2021).
51. Brody, S., Alon, U. & Yahav, E. How attentive are graph attention networks? Preprint at *arXiv* https://doi.org/10.48550/arXiv.2105.14491 (2022).
52. Veličković, P. et al. Graph attention networks. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1710.10903 (2018).
53. Vaswani, A. et al. Attention is all you need. In *NIPS'17: Proc. 31st Conference on Neural Information Processing Systems (NIPS 2017)* Vol. 30 (eds Guyon, I. et al.) 6000–6010 (Curran, 2017).
54. Viñas, R., Azevedo, T., Gamazon, E. R. & Lió, P. Deep learning enables fast and accurate imputation of gene expression. *Front. Genet.* **12**, 624128 (2021).
55. GTEx Consortium. The genotype–tissue expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
56. Lachmann, A., Xie, Z. & Ma'ayan, A. blitzGSEA: efficient computation of gene set enrichment analysis through gamma distribution approximation. *Bioinformatics* **38**, 2356–2357 (2022).
57. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**, R25 (2010).
58. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
59. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902 (2019).
60. van Rossum, G. & Drake, F. L. Jr. *Python Reference Manual* (Centrum voor Wiskunde en Informatica Amsterdam, 1995).
61. Paszke, A. et al. PyTorch: an imperative style, high-performance deep learning library. In *NIPS'19: Proc. 33rd International Conference on Neural Information Processing Systems* (eds Wallach, H. et al.) 8024–8035 (Curran, 2019).
62. Biewald, L. Experiment tracking with Weights and Biases https://www.wandb.com/ (2020).
63. Harris, C. R. et al. Array programming with NumPy. *Nature* **585**, 357–362 (2020).
64. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learning Res.* **12**, 2825–2830 (2011).

65. McKinney, W. Data structures for statistical computing in Python. In *Proc. Ninth Python in Science Conference* (eds van der Walt, S. & Millman, J.) 56–61 (SciPy, 2010).
66. Hunter, J. D. Matplotlib: a 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).
67. Waskom, M. L. seaborn: statistical data visualization. *J. Open Source Softw.* **6**, 3021 (2021).
68. Charlier, F. et al. Statannotations. *Zenodo* https://doi.org/10.5281/zenodo.7213391 (2022).
69. Viñas, R., Joshi, C. & Gamazon Lab. rvinas/HYFA: v0.1.0. *Zenodo* https://doi.org/10.5281/zenodo.7863459 (2023).
70. Shabalin, A. A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**, 1353–1358 (2012).
71. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300 (1995).

Acknowledgements

We thank the reviewers for their constructive comments. We thank T. Azevedo, P. Barbiero, D. Buterez, I. Duta, E. Gómez de Lope, J. Lux, A. Margeloiu, J. Moss, P. Scherer and N. Simidjievski for useful feedback and discussions. The project leading to these results has received funding from Fundación Rafael del Pino (R.V.). C.K.J. was supported by the A*STAR Singapore National Science Scholarship (PhD). P. Liò was supported by FOREUM project "Start" and the EU project GO-DS21 (Gene Overdosage and Comorbidities During the Early Lifetime in Down Syndrome). E.R.G. acknowledges support from the following National Institutes of Health (NIH) grants: Genomic Innovator Award R35HG010718, NHGRI R01HG011138, NIMH R01MH126459 and NIA AG068026. We thank Vanderbilt's Advanced Computing Center for Research and Education (ACCRE) for infrastructure support.

Author contributions

R.V., E.R.G. and P. Liò conceived the study. R.V. developed and implemented the framework, with contributions from C.K.J. and D.G. C.K.J. and R.V. optimized the method and C.K.J. performed the ablation studies. P. Lin and E.R.G. performed the eQTL mapping analyses. C.K.J., R.V. and D.G. studied the scalability of the method. R.V. performed all other experiments and analyses. E.R.G., B.D. and P. Liò supervised the study. R.V. and E.R.G. wrote the manuscript with input from all other authors. All authors approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s42256-023-00684-8.

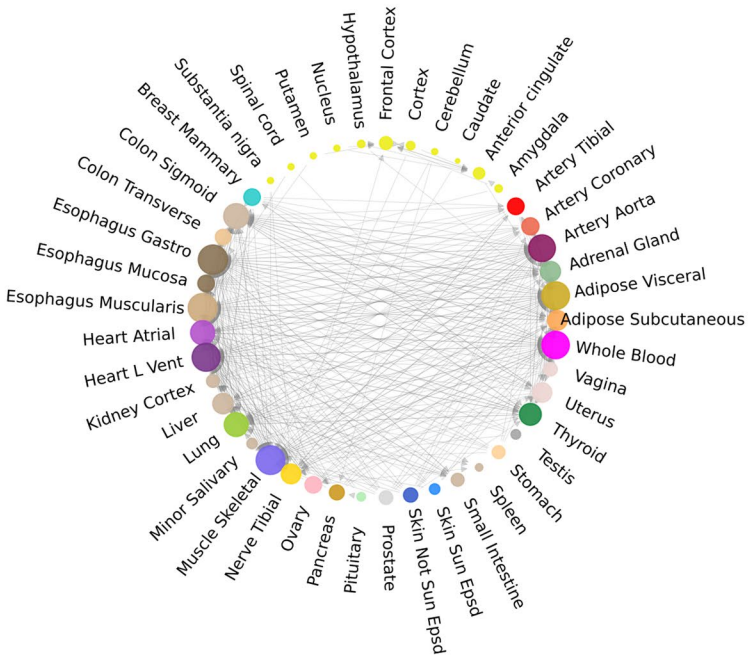
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-023-00684-8.

Correspondence and requests for materials should be addressed to Bianca Dumitrascu, Eric R. Gamazon or Pietro Liò.

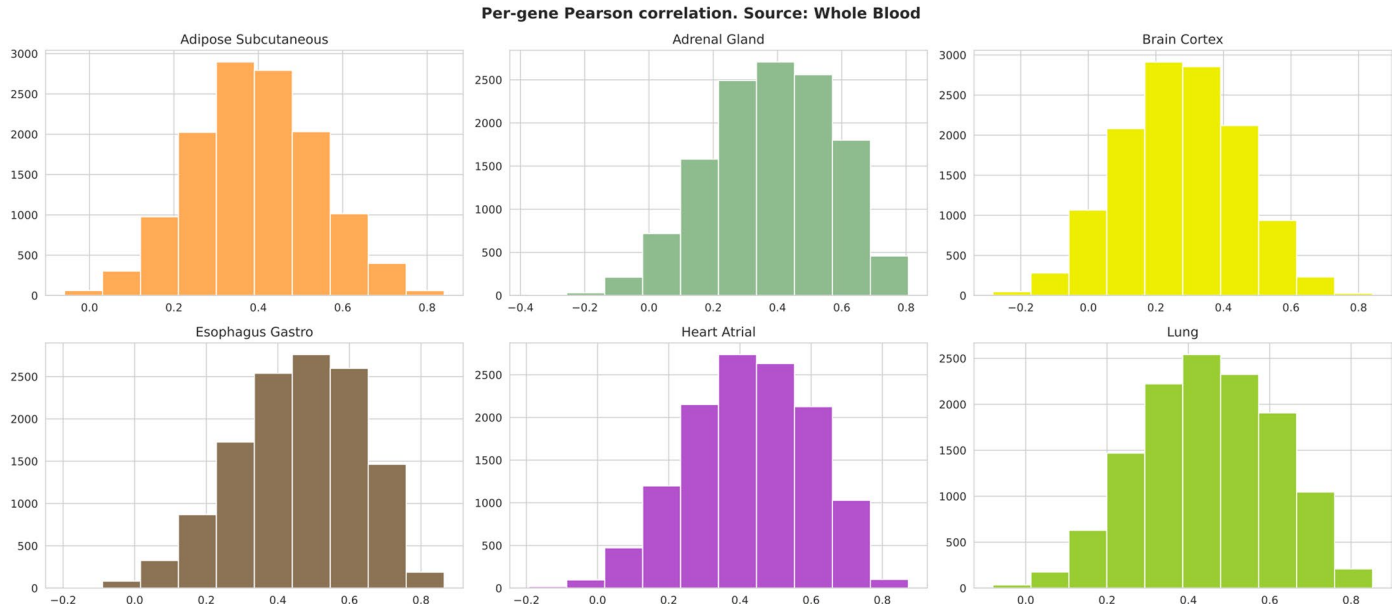
Peer review information *Nature Machine Intelligence* thanks Matthias Heinig and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

https://doi.org/10.1038/s42256-023-00684-8

a

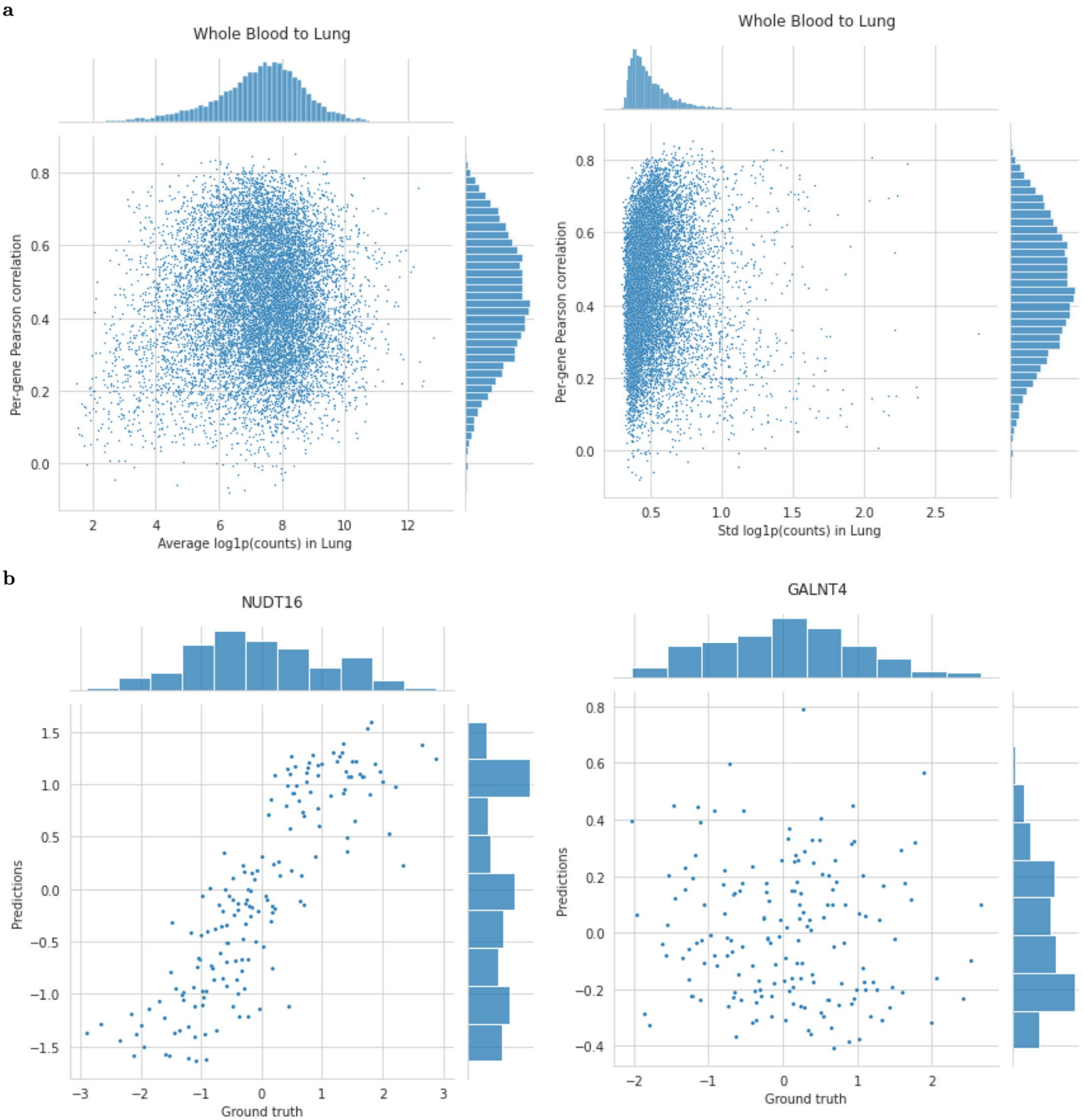


b



Extended Data Fig. 1 | Summary of per-gene prediction scores. (a) Network of tissues depicting the predictability of target tissues with HYFA using the average per-gene Pearson ρ correlation coefficients. Edges from reference to target tissues indicate an average per-gene $\rho > 0.4$. The dimension of each node is proportional to its degree. (b) Distribution of per-gene Pearson correlation coefficients in 6 target tissues (source tissue: whole blood). We attribute the unimodality of the distributions to the fact that the data was inverse Normal transformed (Methods).

https://doi.org/10.1038/s42256-023-00684-8

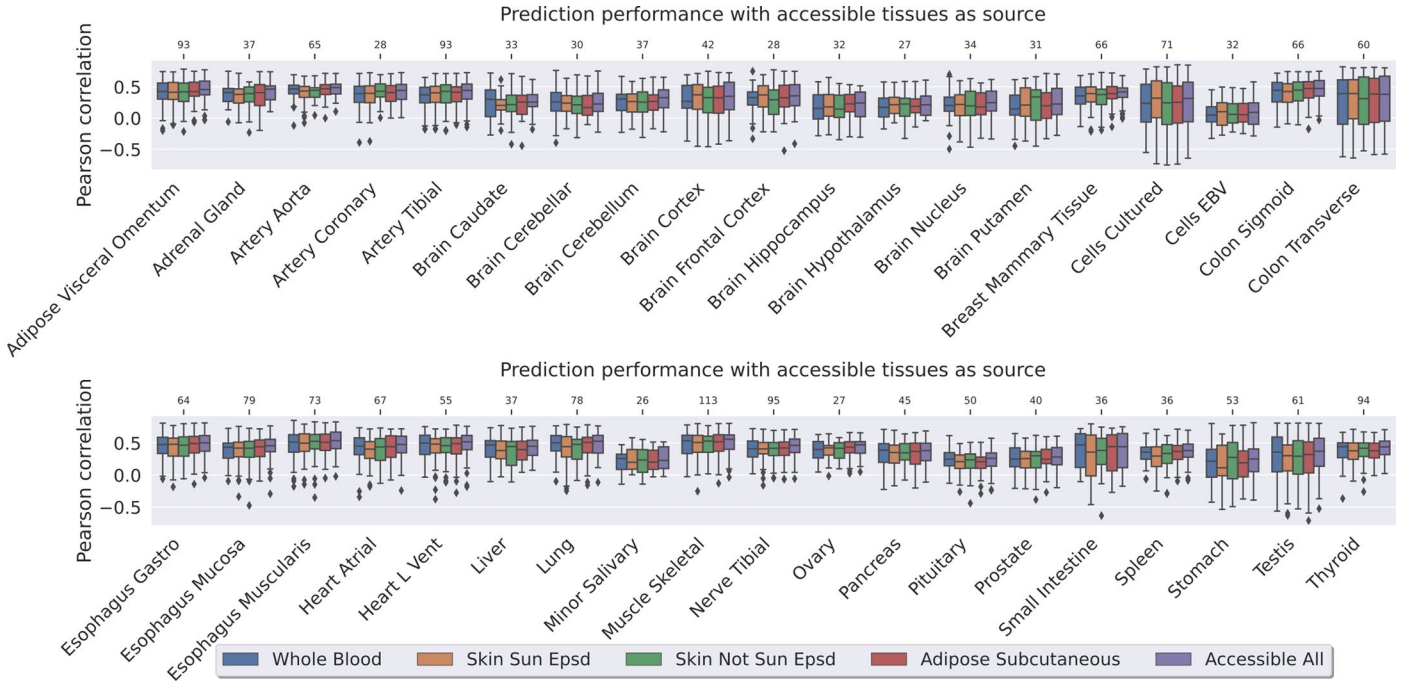


Extended Data Fig. 2 | Whole blood to lung predictions for unseen individuals. (a) Average and standard deviation of per-gene expression in lung versus prediction performance (Pearson correlation between predicted and ground truth expression; whole blood to lung). The per-gene predictions were uncorrelated with the averages and variances of the per-gene expression in the target tissue (average: $\rho = 0.07$, variance: $\rho = 0.06$). (b) Best and worst predicted lung genes (*NUDT16*: $\rho = 0.85$; *GALNT4*: $\rho = -0.08$; n=166).

https://doi.org/10.1038/s42256-023-00684-8

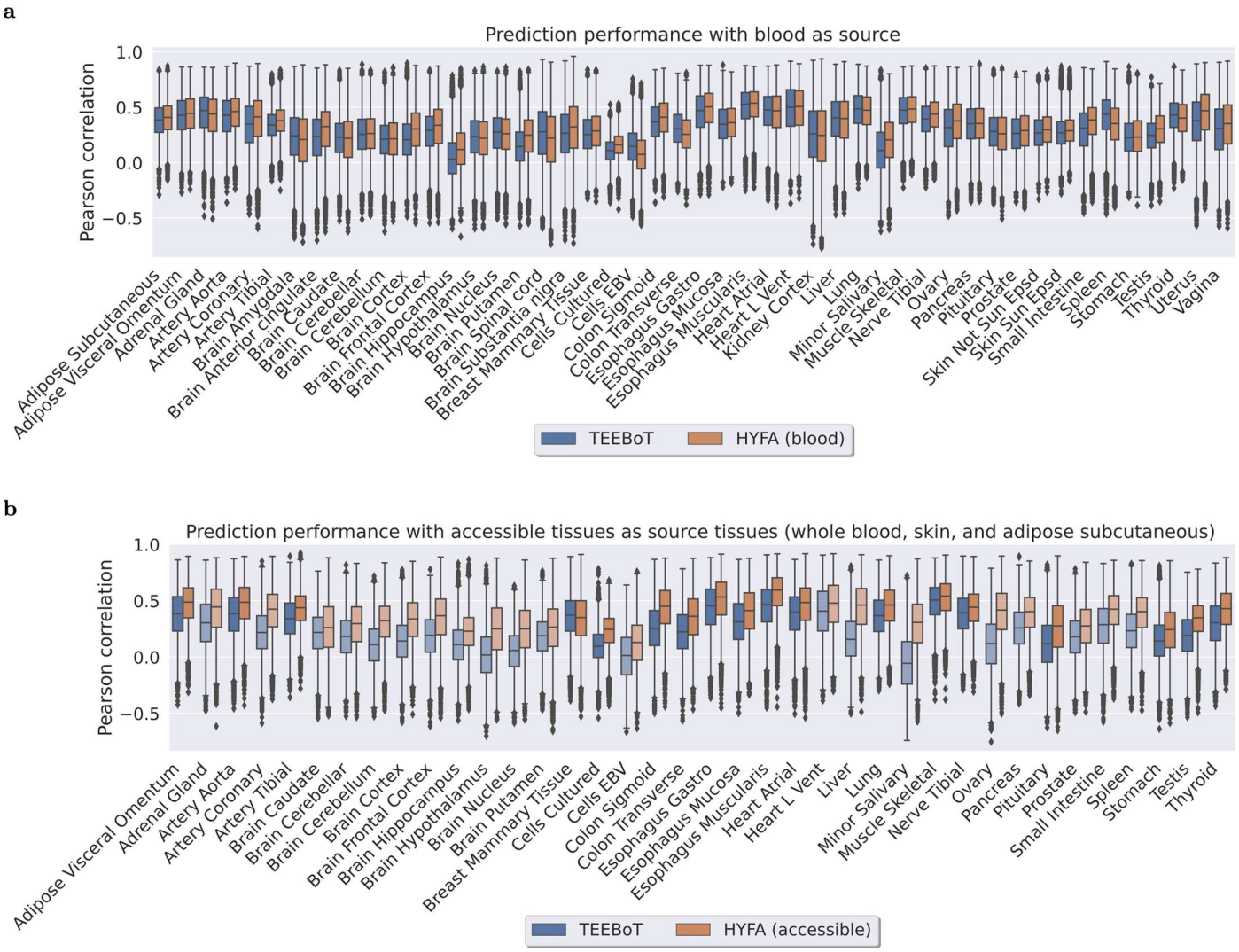
Extended Data Fig. 3 | Top predicted Alzheimer's disease-relevant genes in multiple brain regions, with whole blood as reference tissue. (a) Pearson correlation coefficient of top 20 predicted genes from the Alzheimer's disease pathway (KEGG), ranked by average correlation. (b, c, d) Average per-gene expression (x-axis) versus prediction performance (Pearson correlation between predicted and ground truth expression) in (b) cerebellum, (c) cortex, and (d) hippocampus. HYFA exhibits strong prediction performance for several Alzheimer's disease-relevant genes including *APOE* (cortex ρ=0.536, cerebellum: ρ=0.502), *APP* (cortex ρ=0.524), *PSEN1* (cerebellum: ρ=0.459), and *PSEN2* (cortex: ρ=0.590, cerebellum: ρ=0.559, hippocampus: ρ=0.403). In cerebellum, *PSEN1* (ρ=0.459), *PSEN2* (ρ=0.559), and *APOE* (ρ=0.502) attained above expected performances (average ρ=0.448). *APP* (ρ=0.524), *PSEN2* (ρ=0.590), and *APOE* (ρ=0.536) surpassed the expected correlation in cortex (average ρ=0.443).

https://doi.org/10.1038/s42256-023-00684-8



Extended Data Fig. 4 | Prediction scores for different accessible tissues as reference. For each target tissue, we predicted the expression values based on accessible tissues (whole blood, skin sun exposed, skin not sun exposed, and adipose subcutaneous). We report the Pearson correlation coefficient between the predicted values and the actual gene expression values. For any given target tissue, we used the same set of individuals to evaluate performance, namely individuals in the validation and test sets with collected gene expression measurements in all the corresponding tissues. Target tissues represented by less than 25 test individuals were discarded. HYFA attains the best performance in 32 out of 38 tissues when all accessible tissues are simultaneously used as reference. Boxes show quartiles, centerlines correspond to the median, and whiskers depict the distribution range (1.5 times the interquartile range). Outliers outside of the whiskers are shown as distinct points. The top axis indicates the total number of samples for every target tissue.

https://doi.org/10.1038/s42256-023-00684-8

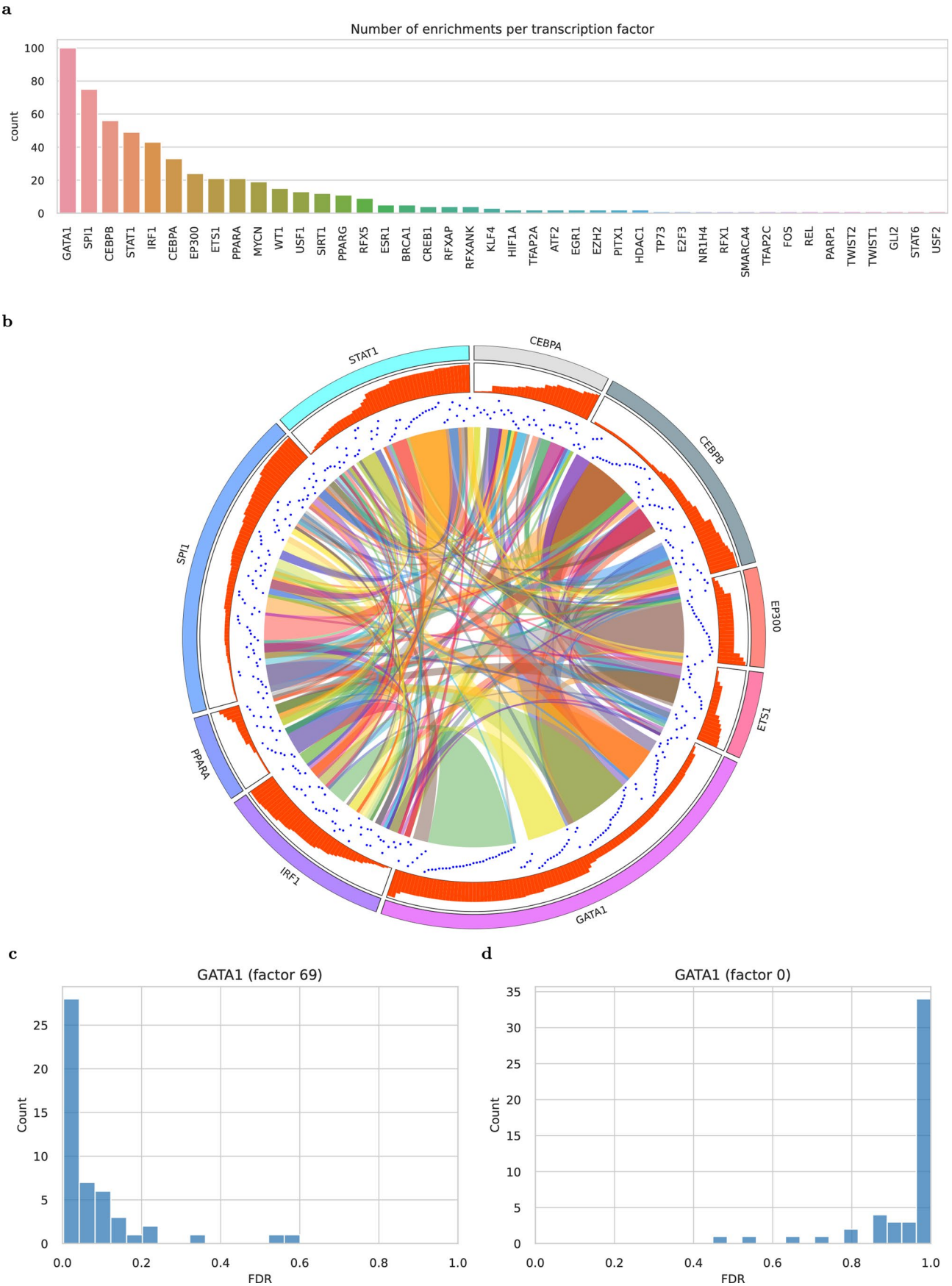


Extended Data Fig. 5 | See next page for caption.

https://doi.org/10.1038/s42256-023-00684-8

Extended Data Fig. 5 | Performance comparison across gene expression imputation methods with per-gene metrics (n=12,557 genes). (a, b) Per-tissue comparison between HYFA and TEEBoT when using (a) whole-blood and (b) all accessible tissues (whole blood, skin sun exposed, skin not sun exposed, and adipose subcutaneous) as reference. We discarded target tissues represented by less than 25 test individuals. HYFA achieved superior Pearson correlation in (a) 25 out of 48 target tissues when a single tissue was used as reference and (b) all target tissues when multiple reference tissues were considered. For underrepresented target tissues (less than 25 individuals with source and target tissues in the test set), we considered all the validation and test individuals (translucent bars). (c, d) Prediction performance from (c) whole-blood gene expression and (d) accessible tissues as reference. Boxes show quartiles and whiskers depict the distribution range (1.5 times the interquartile range). Mean imputation replaces missing values with the feature averages. Blood surrogate utilises gene expression in whole blood as a proxy for the target tissue. k-Nearest Neighbours (kNN) imputes missing features with the average of measured values across the k nearest observations (k=20). TEEBoT projects reference gene expression into a low- dimensional space with principal component analysis (PCA; 30 components), followed by linear regression to predict target values. HYFA (all) employs information from all collected tissues. Boxes show quartiles, centerlines correspond to the median, and whiskers depict the distribution range (1.5 times the interquartile range). Outliers outside of the whiskers are shown as distinct points.

https://doi.org/10.1038/s42256-023-00684-8

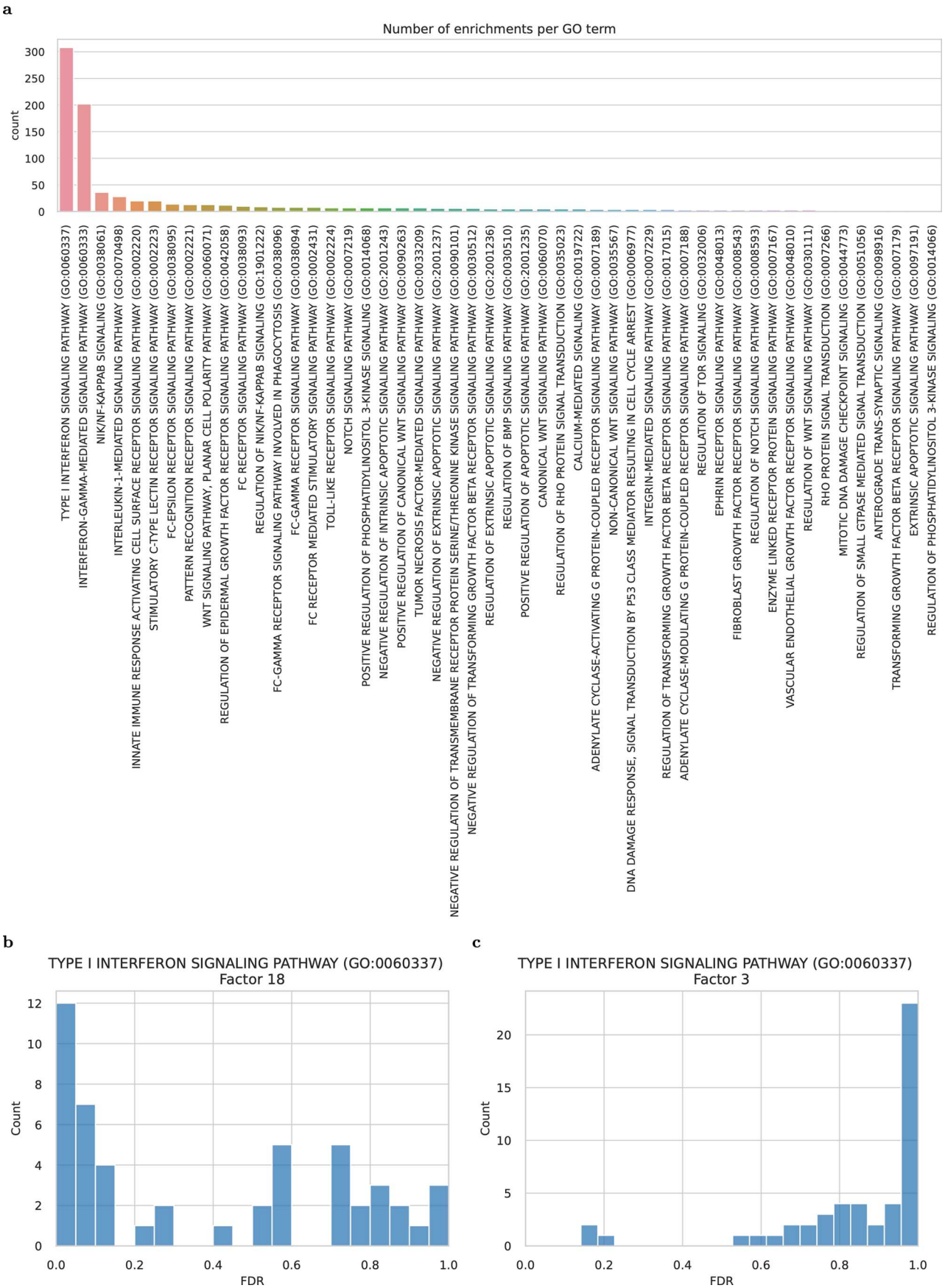


Extended Data Fig. 6 | See next page for caption.

https://doi.org/10.1038/s42256-023-00684-8

Extended Data Fig. 6 | Transcription factor (TF) enrichment analysis of metagene factors. For every metagene (n=50) and factor (n=98), we performed Gene Set Enrichment Analysis using the corresponding gene loadings of HYFA's encoder (Methods) and TF gene sets from the TRRUST database of transcription factors (Enrichr library: TRRUST_Transcription_Factors_2019). (a) Top enriched TFs, ranked by the total number of metagene factors in which the TFs were enriched (FDR < 0.05). (b) Circos plot of the top 9 enriched TFs (outer layer). The angular size is proportional to the number of enrichments. The second layer (bar plot) depicts the factor IDs where the TF was enriched, ranging from 0 (lowest bar) to 98 (highest bar). The third layer shows the corresponding metagene IDs (blue dots) of the enriched metagene factors, increasing monotonically within the same factor. The edges in the middle connect TFs whenever they are both enriched in the same factor (FDR < 0.05). (c, d) Distribution of the GATA1 false discovery rates in factor 69 (FDR < 0.05 in 28/50 metagenes) and an arbitrary factor (enriched in 0/50 metagenes).

https://doi.org/10.1038/s42256-023-00684-8



Extended Data Fig. 7 | See next page for caption.

https://doi.org/10.1038/s42256-023-00684-8

Extended Data Fig. 7 | GO Biological Process enrichment analysis of metagene factors. For every metagene (n=50) and factor (n=98), we performed Gene Set Enrichment Analysis using the corresponding gene loadings of HYFA's encoder (Methods) and Gene Ontology gene sets (GO Biological Process, version of 2021) (Enrichr library: GO_Biological_Process_2021). (a) Top enriched signaling GO terms, ranked by the total number of metagene-factors in which the terms were enriched (FDR < 0.05). (b, c) FDR distribution of the Type-I Interferon signaling pathway in factor 18 (FDR < 0.05 in 12/50 metagenes) and an arbitrary factor (enriched in 0/50 metagenes).

https://doi.org/10.1038/s42256-023-00684-8

Extended Data Fig. 8 | GO Biological Process FDRs for signaling pathways. GO Biological Process enrichment analysis of metagene factors. For every pathway and factor, we selected the metagene with lowest FDR and depicted statistically significant values (FDR < 0.05). Point sizes are inversely proportional to the FDR values. Type I interferons (IFNs), a family of cytokines that activate a variety of signaling cascades, were the most enriched. We also detected the simultaneous enrichment of interferon IRF1 and STAT1 (a member of the STAT protein family that drives the expression of many target genes) in 10 factors (FDR < 0.05; Extended Data Figure 6b), consistent with these results.

nature portfolio

Corresponding author(s): Eric R. Gamazon, Bianca Dumitrascu, Pietro Liò

Last updated by author(s): May 4, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection

We processed the GTEx gene expression data following the standard GTEx eQTL discovery pipeline. Please see: https://github. com/broadinstitute/gtex-pipeline/tree/master/qtl

Data analysis

We used PyTorch [Paszke et al., 2019] to implement the model and scanpy [Wolf et al., 2018] to process the gene expression data. We performed hyperparameter optimisation with wandb [Biewald, 2020]. We employed blitzGSEA [Lachmann et al., 2022] for the pathway enrichment analysis. We also used NumPy [Harris et al., 2020], scikit-learn [Pedregosa et al., 2011], pandas [Wes McKinney, 2010], matplotlib [Hunter, 2007], and seaborn [Waskom, 2021]. The library versions are specified in our Github repository (see requirements.txt file at https://github.com/rvinas/HYFA/blob/main/requirements.txt)

Package version numbers:

anndata==0.7.8
blitzgsea @ git+https://github.com/MaayanLab/blitzgsea.git@9752dd7c5f12d4c935cfeb944f3b045dc23ac2fb
biopython==1.79
bioservices==1.8.4
gseapy==0.10.8
h5py==3.6.0
ipykernel==6.7.0
matplotlib==3.5.1
matplotlib-venn==0.11.6
missingpy==0.2.0

networkx==2.7.1
numpy==1.21.5
pandas==1.4.1
PyYAML==6.0
scanpy==1.8.2
scikit-learn==1.0.2
scikit-misc==0.1.3
scipy==1.7.3
seaborn==0.11.2
statsmodels==0.13.2
supervenn==0.4.1
torch==1.8.0
torch-scatter==2.0.9
torch-sparse==0.6.12
tqdm==4.63.0
umap-learn==0.5.2
wandb==0.12.9

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The datasets analysed for this study, including bulk RNA-seq Consortium [2020] and snRNA-seq Eraslan et al. [2021], can be found in the GTEx portal: https://gtexportal.org/. A detailed summary of the GTEx samples and donor information can be found at: https://gtexportal.org/home/tissueSummaryPage. We downloaded MSK SPECTRUM data from https://cellxgene.cziscience.com/collections/4796c91c-9d8f-4692-be43-347b1727f9d8. The full catalog of HYFA-derived eQTLs is downloadable at: https://doi.org/10.5281/zenodo.6815784. We downloaded RNASeqDB data from https://github.com/mskcc/RNAseqDB.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	We refer the readers to the website https://gtexportal.org/home/tissueSummaryPage for a detailed description of the demographics of the GTEx dataset.
Population characteristics	We refer the readers to the website https://gtexportal.org/home/tissueSummaryPage for a detailed description of the demographics of the GTEx dataset.
Recruitment	We are using publicly available data.
Ethics oversight	We are using publicly available data.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We did not collect any data. We used the GTEx dataset (https://www.gtexportal.org/home/). We report the total number of samples as well as donors in the manuscript. We adopted the GTEx gene expression resource as our primary dataset because it is the largest multi-tissue transcriptome dataset available (Science 2020). This dataset was sufficient for the main analysis because the measured gene expression spans 49 different tissues and several tissues were collected for every individual.

Data exclusions	We discarded observations from underrepresented tissues (n=5) due to small sample size. These tissues are: namely bladder, cervix (ectocervix, endocervix), fallopian tube, and kidney (medulla).
Replication	Our imputation results can be replicated by rerunning our publicly available code (https://github.com/rvinas/HYFA/).
Randomization	This is not relevant to our study because we do not have test and control groups. In terms of model performance evaluation, we randomly split donors into three sets: train, validation, and test. The model was trained on samples from the train set and hyperparameters were optimised on the validation set. We used the test set exclusively to report the performance results.
Blinding	This is not relevant to our study because we did not collect any data (i.e. we used data collected by the GTEx Consortium). The GTEx gene expression data was already de-identified by the GTEx Consortium and thus we did not have to anonymise the data ourselves. We did not release nor reveal any sensitive information in the manuscript. The GTEx Consortium data is available for download from the GTEx portal (https://www.gtexportal.org/home/datasets).

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging