Combining Motifs, CRE Activity, And Gene Expression Data Using ML Greatly Improves the Accuracy of Tissue-Specific TF Network Maps

Wooseok J. Jung[1], Sandeep Acharya[2], Daniel P. Ruskin[4], Shu Liao[1], Vaha Akbary Moghaddam[3], Zolboo Erdenebaatar[1], Michael R. Brent[1]

[1]Department of Computer Science and Engineering, Washington University, St Louis, MO, [2]Division of Computational and Data Sciences, Washington University, St Louis, MO, [3]Division of Statistical Genomics, Washington University School of Medicine, St Louis, MO, [4]College of Information, University of Maryland, College Park, MD

Abstract

Transcription factor (TF) network maps link TFs to their direct, functional gene targets whose transcription they regulate by binding cis-regulatory elements (CREs). Existing methods to reconstruct these networks typically rely either on TF motifs in CREs or gene expression data alone, limiting their accuracy. Motif data alone often fail to identify actual TF binding sites, while expression data cannot distinguish direct from indirect regulatory relationships. Additionally, accurate TF networks must be tissue-specific due to varied TF activities and expression patterns across tissues.

We introduce METANets (Motif Expression TF Association Networks), a novel supervised ensemble learning approach integrating TF motifs, TF binding locations, CRE activity, and gene expression data. Using XGBoost models, we predict TF binding in CREs based on TF motif and gene expression features, derived from linear (LASSO) and non-linear (BART) regression models trained on tissue-specific and aggregated RNA-seq data. This approach was applied to 36 human tissues from GTEx.

METANets significantly outperform existing motif-only and expression-only approaches, capturing more direct, functional TF targets. Evaluations against ChIP-seq binding data and gene ontology enrichment demonstrate METANets' superiority in identifying functional targets directly bound by TFs. Furthermore, tissue specificity assessed through tissue-specific expression quantitative trait loci (eQTLs) confirms METANets effectively capture tissue-specific regulation, performing comparably with other networks.

Our approach markedly improves TF network reconstruction by combining complementary data types, enhancing the accuracy and utility of tissue-specific transcriptional regulatory maps. METANets provide robust resources for researchers investigating TF-mediated regulation within human tissues.

Background

A transcription factor (TF) network map is a directed graph in which nodes represent TFs and genes, and edges link TFs to the genes they directly regulate. Numerous algorithms have been developed for mapping TF networks from a variety of data types, and these methods generally fall into two broad categories. The first category uses TF binding evidence, such as DNA sequence motifs [1] or TF ChIP-seq [2], often combined with promoter and enhancer annotations to define regulatory elements. Motif data describe locations where a TF could bind and TF ChIP-seq data reveal the genomic regions where a TF actually binds. However, binding evidence alone cannot confirm whether a TF functionally affects the transcription rates of its target genes. For instance, many TF-bound motifs have no apparent effect on gene transcription [3-5]. Furthermore, most motifs for a TF are not bound by the TF [6, 7], and many sites where a TF binds lack motifs [8, 9].

Another class of methods uses gene expression measurements under baseline conditions or after perturbations of TF expression [10-15]. While perturbation-response data can reveal the functional effects of TFs, they do not confirm direct binding. Moreover, expression-based methods are prone to capturing indirect effects, as correlations between TF and target gene (TG) expression may arise from transitive effects such as shared upstream regulators or mediated interactions [16-18].

While many regulatory relationships are shared across tissues, some TF-TG interactions are context-specific. TF network maps inferred from gene expression and binding data have been shown to vary significantly across cell types and tissues [19-22]. Our group has previously mapped TF networks in *Saccharomyces cerevisiae* and *Drosophila melanogaster* using both TF binding and gene expression data, successfully capturing direct and functional targets of TFs [11, 12]. However, these mapping efforts focused on a single, aggregate context and did not address regulatory variation across tissues, which limits their utility for understanding context-specific gene regulation in human tissues. Most widely used approaches for tissue-specific TF network mapping remain limited in their ability to capture direct and functional targets. For example, TF motif-based approaches using cis-regulatory element (CRE) activity levels across tissues can suggest tissue-specific binding potential, but do not confirm functional relationships [23], while correlation-based networks fail to account for the physical binding of the TF in the target gene's CREs [24, 25]. One notable exception is [20] which uses PANDA, combining TF motifs and protein-protein interactions (PPI) with tissue-specific expression data in a message-passing framework, to map 38 tissue-specific TF networks. Building on our prior work, we introduce METANet, a supervised machine learning approach to predict tissue-specific direct and functional TF-TG regulatory relationships. Unlike PANDA's unsupervised approach, METANet uses supervised learning which allows the model to learn directly from TF ChIP-seq-based binding

signals. Using tissue-specific gene expression data and TF network maps derived from [23] as predictor variables, METANet predicts the probability that a TF binds a target gene in TF-ChIP-seq data the model has not seen.

This paper makes four contributions. We provide a comprehensive resource of human tissue-specific TF network maps across 36 tissues, encompassing ~237 TFs and ~12,150 protein-coding genes. We evaluate the ability of our TF network maps to capture direct and functional targets of TFs using four evaluation metrics and compare our performance to the networks proposed by Sonawane et al. We also compare to the original networks from Marbach et al (which does not use gene expression data). Across 36 tissues, we show that METANet better captures both the direct and functional targets of TFs compared to benchmark network maps. We also found that the tissue-specificity of METANet is comparable to that of the network maps from Marbach et al and Sonawane et al. To explore potential applications of METANet, input it to FISHNET [26], a network-driven gene prioritization method, using transcriptome-wide association (TWAS) summary statistics for traits associated with cardiovascular risk from the Long Life Family Study (LLFS) cohort. METANet identified replicated gene-trait associations that were missed when other networks were passed as inputs to FISHNET. Furthermore, METANet identified regulatory mechanisms that may mediate the effect of known trait-associated variants.

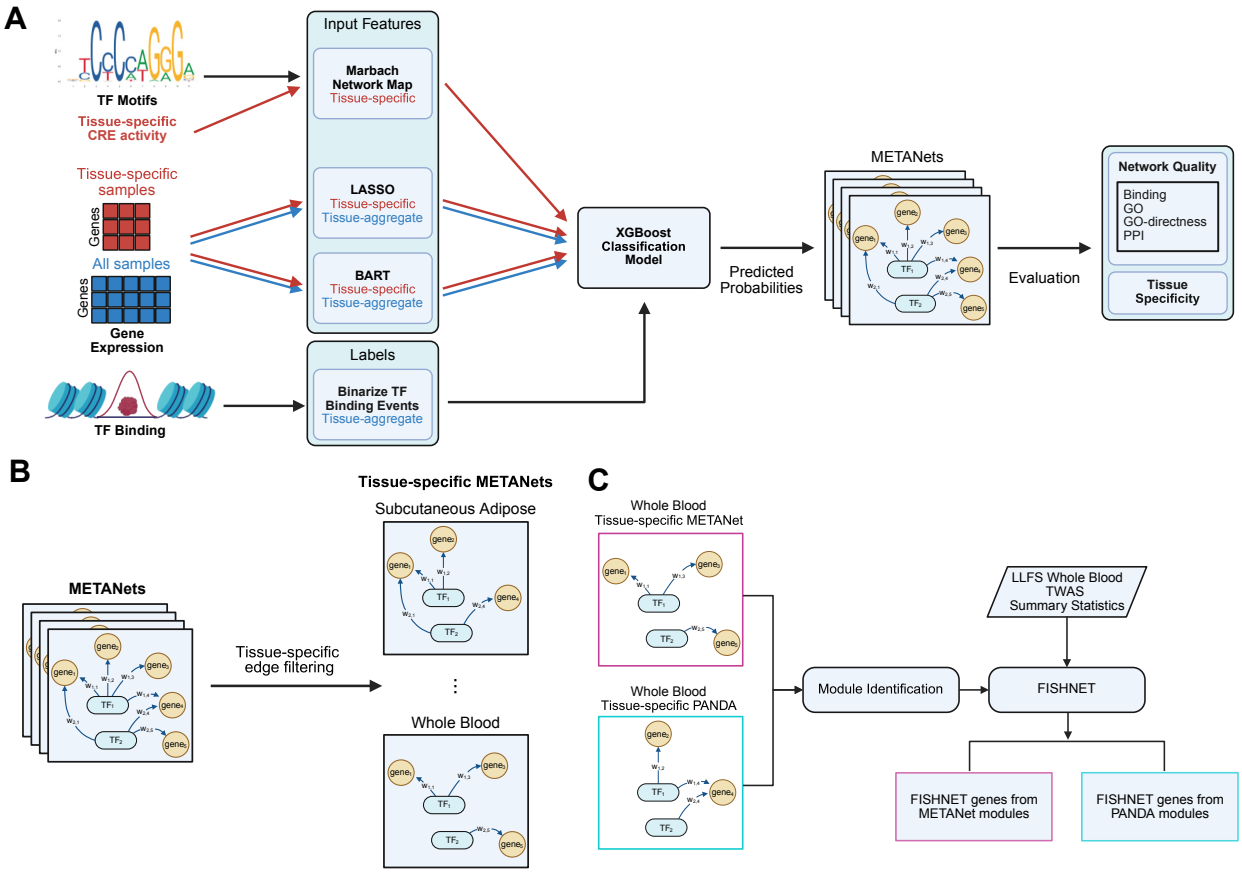

Fig. 1 METANets overview (**A**) For each tissue, METANets combines evidence scores from tissue-specific TF network maps from Marbach et al. 2016, tissue-specific and tissue-aggregate LASSO and BART regression models from GTEx expression data, and binarized TF binding labels using XGBoost. The predicted probabilities from the XGBoost model are used as edge scores in our tissue TF network maps. Red arrows indicate tissue-specific data while blue arrows indicate tissue-aggregate data. The predicted METANets are evaluated for network quality and tissue specificity. (**B**) Tissue-specific METANets are inferred from the predicted METANets by filtering for tissue-specific edges. (**C**) Modules identified from tissue-specific METANet and tissue-specific PANDA are used in FISHNET with TWAS summary statistics from the LLFS to identify FISHNET genes.

Results

Overview of TF Network Map Reconstruction

To map tissue-specific transcription factor (TF) regulatory networks, we developed METANets, a supervised learning framework that integrates TF motif data, gene expression-derived features, and TF binding evidence. For each of 36 human tissues, we trained an XGBoost model [27] to predict whether a TF binds within the cis-regulatory elements (CREs) of a gene, using ChIP-seq data from REMAP 2020 [28] as ground truth (Fig. 1). Each instance in the model corresponds to

a candidate TF-target gene (TG) pair and was represented by five biologically motivated features: (1) a motif-based regulatory score from the tissue-specific Marbach network [23], which combines TF motif occurrences with tissue-specific CRE activity; (2–5) four expression-derived features capturing both linear and nonlinear TF-TG relationships. The features were derived by regressing the expression of each gene on the expression of all TFs using both LASSO [29] and Bayesian Additive Regression Trees (BART) [30]. We extracted from the model for each gene measures of the relatedness of each TF's expression to the gene's expression (see Methods for details). Both regression algorithms were applied to both tissue-specific and tissue-aggregated RNA-seq data from GTEx [31] (Fig. 1). We used 10-fold nested cross-validation to train and evaluate each model. Predicted probabilities from test folds were used as edge scores in the METANet, linking TFs to their likely direct functional targets. In total, we mapped the TF networks of 36 human tissues, encompassing an average of 237 TFs and 12150 genes.

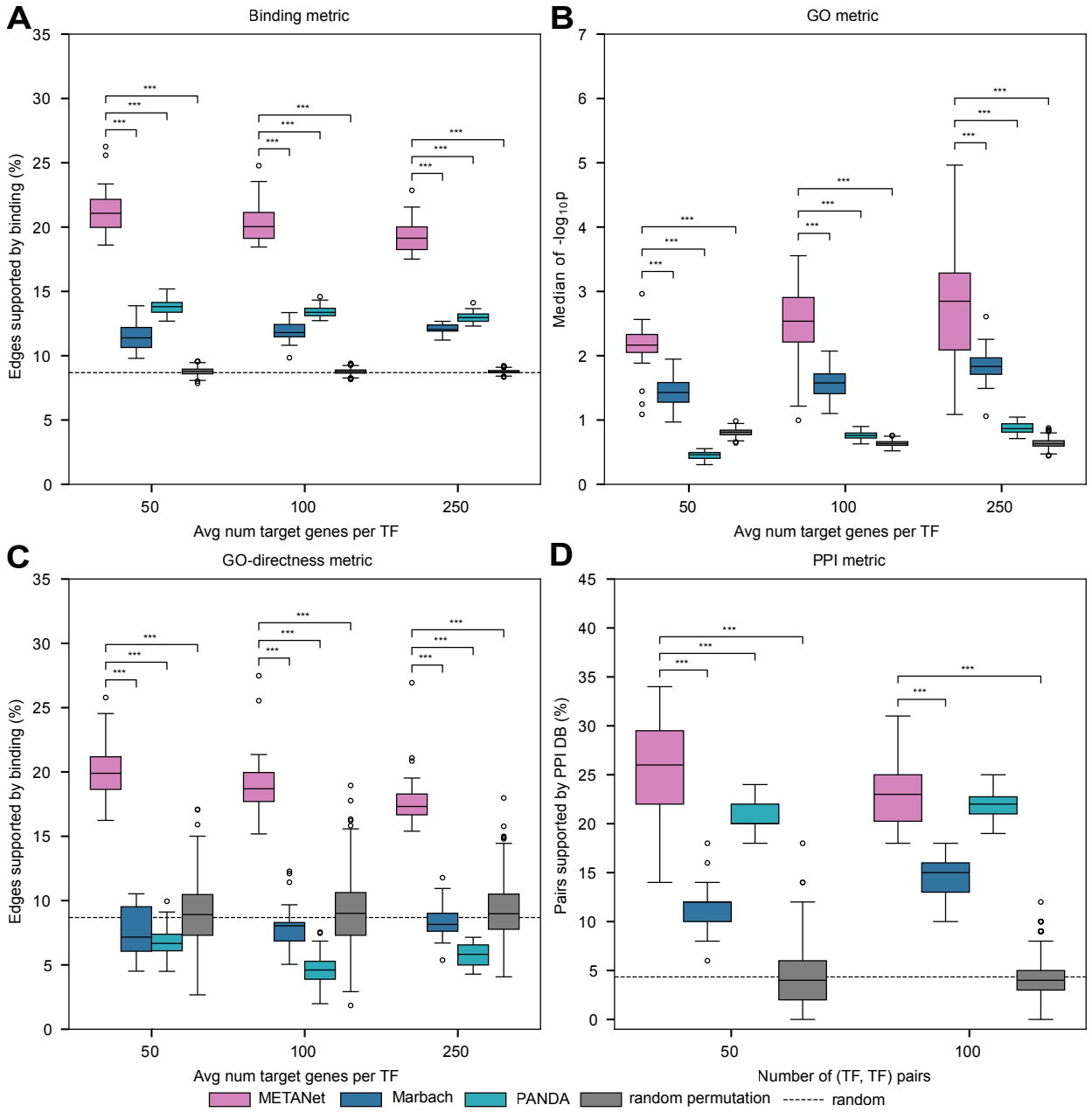

Fig. 2 Evaluation of network quality. Performance of METANets, Marbach networks, and PANDA networks on 26 tissues for which all three networks were available. Pink: METANets; Blue: Marbach network maps; Cyan: PANDA network maps; Gray: random expectation based on permutations of METANets; black dashed: random expectation. (**A**) Binding metric (defined in main text) calculated at different edge-score thresholds. METANets outperform all other network maps. Random expectation is the probability that a randomly selected TF-TG edge will be supported by binding data. (**B**) GO metric (defined in main text). METANets outperform all other network maps and the random expectation. Random permutation is the GO metric for 50 METANets in which edge scores were randomly permuted. (**C**) GO-directness metric (defined in main text). METANets significantly outperform all other network maps. Random expectations include the GO-directness evaluations of the GO results of the 50 permuted network maps generated in (B) and the binding metric random expectation. (**D**) PPI metric (defined in main text). METANets significantly outperform all other network maps. The random expectation is the probability that a randomly selected (TF, TF) pair will be supported by PPI data. Significance bars only drawn for significant paired *t*-tests with METANet. *$P \leq 0.05$, **$P \leq 0.01$, ***$P \leq 0.001$.

126 **METANets Capture Direct and Functional Targets of TFs**

127 We compared METANet against the motif-based Marbach networks, constructed using motif
128 occurrences and tissue-specific CRE activity levels [23], and PANDA networks from Sonawane et
129 al., built using motifs and expression via an unsupervised approach [20], across 26 common tissues.
130 For the PANDA networks, we used the fully connected PANDA output networks from Sonawane
131 et al. prior to filtering for tissue-specific edges. All networks we evaluated include continuous edge
132 scores, from which discrete networks can be created by thresholding and retaining only edges
133 above the chosen cutoff. To enable fair comparisons across tissues, we applied thresholds scaled
134 to the average number of targets per TF (e.g. top 20000 edges for 200 TFs implies 100 targets per
135 TF), and we examined how evaluation results changed as more, lower-scoring edges were included.

136 To assess network quality, we used four complementary metrics: Binding, Gene Ontology (GO),
137 GO-directness, and protein-protein interaction (PPI). To provide an empirical null expectation for
138 each metric, we generated 50 random networks by permuting the edge scores of METANets. These
139 random networks were also subjected to the same cutoff thresholds as other networks before
140 evaluation.

141 **Binding** The Binding metric is the fraction of edges supported by binding data showing that the
142 TF binds in one or more of the target's CREs, sourced from the FANTOM5 database [32, 33].
143 METANet significantly exceeded both Marbach and PANDA (paired *t*-tests, P<0.001 for each; Fig.
144 2A) and far exceeded the random expectation (*t*-test, P= 1.8×10^{-32}), indicating stronger
145 enrichment for directly bound targets.

146 **Gene Ontology (GO) enrichment** To compute the GO metric, we carried out over-representation
147 analysis of GO biological process annotations among the targets of each TF. Each TF's score was

148 the maximum –log P-value over all GO terms. If a TF's target set showed no significant enrichment
149 for any GO biological process, we assigned the TF a P-value of 1 (equivalent to –log P = 0). The
150 final network-level score was taken as the median of these scores across all TFs. METANet
151 outperformed Marbach and PANDA (both P<0.001; Fig. 2B) and surpassed the random
152 permutation (P<0.001; see Methods), showing greater functional coherence among the predicted
153 targets of each TF.

154 **GO-directness** The GO-directness metric is the fraction of targets annotated with each TF's most
155 significant GO term that are also supported by ChIP-seq binding. This metric penalizes networks
156 that achieve high GO scores by including many, functionally related, but indirect, targets.
157 METANet retained significantly more binding support than both Marbach and PANDA (P<0.001;
158 Fig. 2C) and the random permutations (P<0.001).

159 **Protein-protein interaction (PPI)** The PPI metric is the fraction of top-ranked (TF, TF) pairs,
160 based on target set similarity, that are supported by high-confidence physical protein-protein
161 interactions in the STRING database (confidence score ≥ 0.7) [34]. METANet outperformed
162 Marbach (P<0.001) and PANDA (P<0.05), recovering biologically corroborated physical TF-TF
163 interactions (Fig. 2D).

164 Together, these results show that METANets are enriched for direct and functionally coherent TF-
165 TG interactions relative to other networks.

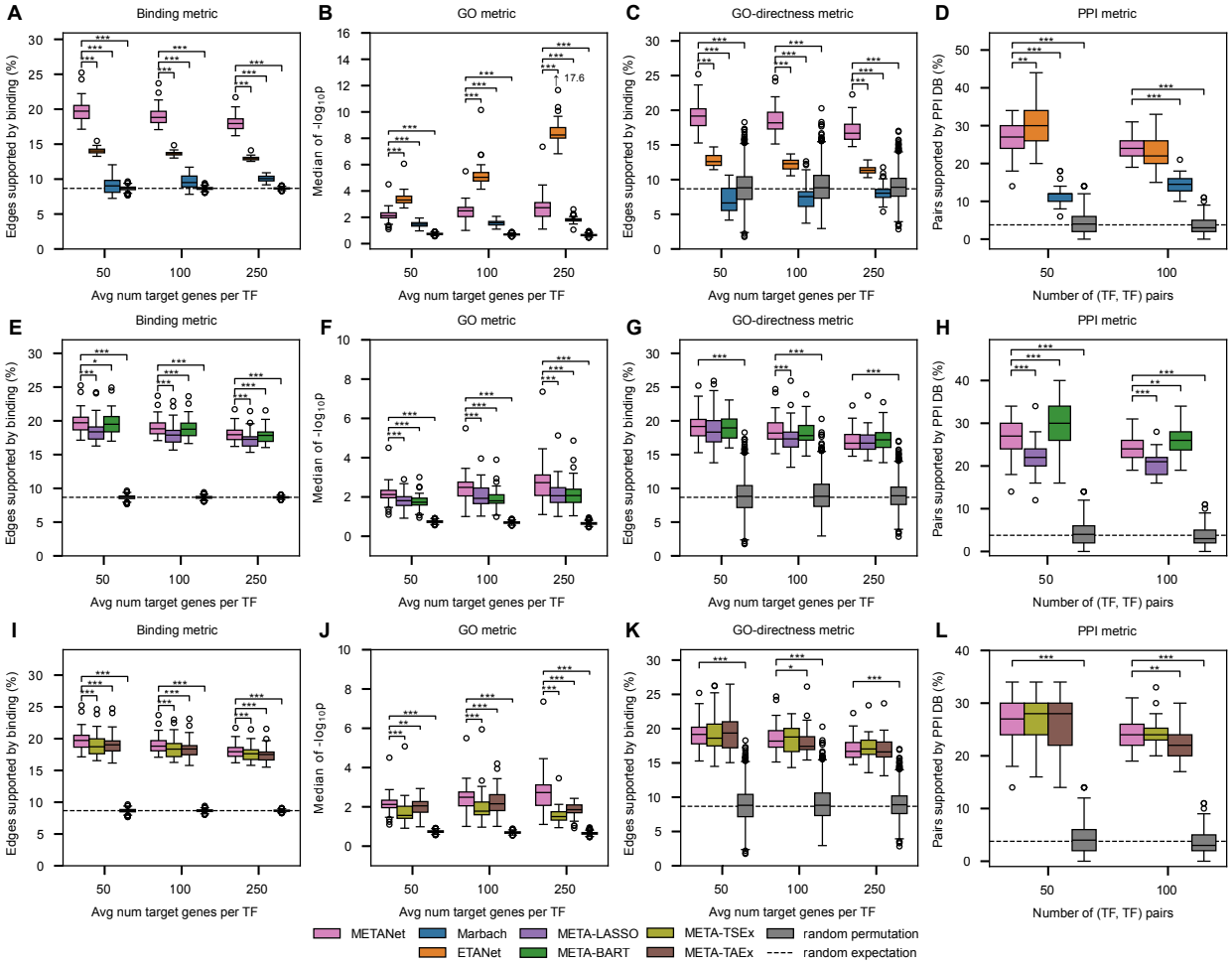

Fig. 3 Network quality evaluations against ablated networks. (**A-B**) Performance of METANets and the Marbach networks (motif-only) and ETANets (expression-only) on 36 tissues. (**A**) Binding metric. METANets significantly outperformed both ETANets and Marbach networks. (**B**) GO metric. ETANets outperformed METANets and Marbach networks. (**C**) GO-directness metric. METANets outperformed ETANets, predicting more accurate TF network maps of direct and functional TF-TG interactions. (**D**) PPI metric. ETANets outperformed Marbach networks in both thresholds but were comparable to METANets for top 100 (TF, TF) pairs. **(E-H)** Performance of METANets and ablated TF network maps META-LASSO and META-BART. (**E**) Binding metric. METANet and META-BART significantly outperform META-LASSO, while METANet and META-BART perform comparably. (**F**) GO metric. METANets outperforms both ablated TF network maps. (**G**) GO-directness metric. All three TF network maps significantly outperform the random expectation based on 50 permutations of METANets. METANets is comparable with both ablated network maps. (**H**) PPI metric. META-BART outperforms both METANets and META-LASSO. **(I-L)** Performance of METANets and ablated TF network maps META-TSEx and META-TAEx. (**I**) Binding metric. METANets significantly outperform both ablated TF network maps. (**J**) GO metric. METANets outperforms both ablated TF network maps. (**K**) GO directness metric. All three TF network maps significantly outperform the random permutation and METANets is comparable with both ablated network maps. (**L**) PPI metric. All three network maps perform similarly with and outperform the random permutation. Significance bars only drawn for significant paired *t*-tests with METANet. *$P \leq 0.05$, **$P \leq 0.01$, ***$P \leq 0.001$.

Combining Motif and Expression Features Improves Network Mapping

185 We next evaluated the contribution of TF motif- and gene expression-derived features by
186 comparing METANets to the Marbach networks, which are built using only motifs, and to
187 ETANets, which we built in the same way as METANets but not including the edge score from
188 Marbach networks, using only expression data across all 36 tissues. METANets significantly
189 outperformed both Marbach and ETANet in the Binding metric ($P<0.001$), indicating the value of
190 integrating both information sources (Fig. 3A). Interestingly, ETANet performed best in the GO
191 metric, suggesting that gene expression alone captures broader functional coherence (Fig. 3B).
192 However, METANets significantly outperformed ETANets in the GO-directness metric ($P<0.001$),
193 indicating that many of ETANets' additional functionally coherent targets are likely indirect (Fig.
194 3C). Thus, METANets provide a more accurate map of direct, functional TF-TG interactions than
195 both the Marbach networks and the PANDAs networks.

196 **Linear and Nonlinear Expression Features are Synergistic**

197 To assess whether combining linear (LASSO) and nonlinear (BART) expression-derived features
198 improved network quality, we compared METANet against ablated versions using only one feature
199 type: META-LASSO and META-BART. In the Binding metric, METANet and META-BART
200 outperformed META-LASSO. Compared to each other, METANet and META-BART performed
201 comparably (Fig. 3A). In the GO metric, METANet significantly outperformed both ablated
202 network maps (Fig. 3B), suggesting that leveraging both LASSO and BART features better
203 captures functionally coherent relationships. GO-directness was similar across models (Fig. 3D),
204 consistent with gains in functional targets without loss of direct targets.

205 **Combining Tissue-specific and Tissue-aggregate Expression Features is Helpful**

To assess the contribution of expression-derived feature specificity, we compared METANet to two ablated variants. META-TSEx retains the motif-based Marbach feature and tissue-specific expression features and excludes tissue-aggregate expression features (LASSO and BART trained on all samples pooled across tissues). Conversely, META-TAEx retains the Marbach feature and tissue-aggregate expression features, but not the tissue-specific expression features. In the Binding and GO metrics, METANet outperformed both ablated models (Fig. 3E, F). In the GO-directness metric, METANet significantly outperformed META-TAEx at one threshold (100 targets per TF) (Fig. 3G). Together, these results show that combining both tissue-specific and tissue-aggregate expression features enables METANets to identify substantially more functionally coherent targets of TFs without sacrificing direct binding support.

METANets are Tissue Specific and Comparable to Existing Methods

We evaluated tissue specificity using tissue-specific expression quantitative trait loci (eQTLs) from GTEx [35]. For each edge in each METANet, we queried whether at least one tissue-specific eQTL overlapped the TF's ChIP-seq peaks in the promoter or annotated enhancers of the target gene. We defined two complementary metrics:

1. eQTL support: Fraction of edges supported by at least one tissue-specific eQTL.
2. eQTL count: Sum, across all edges, of the number of tissue-specific eQTLs supporting each edge.

To assess a network's tissue-specificity, we performed a permutation test. First, we calculated an observed statistic S_{obs} by ranking all 36 networks within each tissue and summing the ranks of the "matching" pairs (e.g., the rank of the liver network in the liver tissue). A smaller sum indicates better overall specificity. Next, to generate a null distribution, we performed 10000 permutations.

228 In each permutation, we randomly shuffled the network labels before summing the ranks of the
229 newly assigned matching pairs to obtain the permutation statistic S_{perm}. This process simulates
230 the null hypothesis that there is no true correspondence between a network and its native tissue.
231 The empirical p-value was then calculated as the proportion of permutations with $S_{perm} \leq S_{obs}$.

Empirical test for tissue-specificity for 36 tissues			
eQTL support	Average number of targets per TF		
	50	100	250
METANet	0.498 (666)	**0.027 (597)**	**0.022 (613)**
Marbach	**0.003 (567)**	**$2x10^{-4}$ (554)**	**$6x10^{-4}$ (537)**
eQTL counts	Average number of targets per TF		
	50	100	250
METANet	0.106 (604)	**0.003 (535)**	**0.007 (563)**
Marbach	0.119 (611)	0.173 (623)	**$3.4x10^{-3}$ (537)**

232 **Table 1 Tissue specificity evaluation of METANet and Marbach against the null across 36 tissues.** Permutation
233 test p-values (and the sum of observed matching ranks) against the distribution of 10000 permutation sums.
234 METANets showed significant tissue specificity compared to the null in eQTL support at average 100 and 250 targets
235 per TF thresholds. The Marbach networks showed significant tissue specificity in eQTL support at all thresholds.
236 METANets showed significant tissue specificity in eQTL counts at average 50 and 100 targets per TF thresholds. The
237 Marbach networks show significant tissue specificity in eQTL counts at average 250 targets per TF only. See also
238 Additional file 1: Fig. S1.

Empirical test for tissue-specificity for 26 tissues			
eQTL support	Average number of targets per TF		
	50	100	250
METANet	0.241 (333)	0.096 (320)	0.143 (324)
PANDA	0.091 (319)	0.176 (330)	0.424 (346)
eQTL counts	Average number of targets per TF		
	50	100	250
METANet	0.297 (335)	**0.007 (284)**	0.129 (324)
PANDA	0.202 (329)	0.245 (331)	0.183 (328)

239 **Table 2 Tissue specificity evaluation of METANet and PANDA against the null across 26 tissues.** Permutation
240 test p-values (and the sum of observed matching ranks) against the distribution of 10000 permutation sums.
241 METANets and PANDA networks did not show significant tissue specificity in eQTL support. METANets were
242 significantly tissue specific in eQTL counts at the average 100 targets per TF threshold. PANDA networks remained
243 not significantly tissue specific in eQTL counts. See also Additional file 1: Fig. S1.

244 METANets are as tissue specific as the Marbach networks

245 We compared the tissue specificity of top-scoring edges in METANets and the Marbach networks.
246 METANets showed significant tissue specificity compared to the null in both eQTL support and
247 eQTL count (P<0.05) at higher average targets per TF thresholds (≥100 targets per TF) (Table 1).

The Marbach networks showed similar patterns, and the Wilcoxon signed-rank test revealed no significant difference in the distributions of matching tissue ranks between METANets and Marbach networks in tissue specificity (Table 3). Thus, METANets achieve comparable tissue specificity while outperforming the Marbach networks in all network quality metrics.

Wilcoxon signed-rank test vs. Marbach			
Average number of targets per TF	50	100	250
eQTL support	0.95 (294.0)	0.30 (266.5)	0.88 (289.0)
eQTL counts	0.24 (243.0)	0.84 (269.5)	0.43 (282.0)

Table 3 Tissue specificity evaluation against Marbach. Wilcoxon signed-rank test p-values (and W statistics) between matching tissue ranks of METANets and Marbach networks for 36 tissues. Across all thresholds of average number of targets per TF, METANets and Marbach networks were not significantly different.

Wilcoxon signed-rank test vs. PANDA			
Average number of targets per TF	50	100	250
eQTL support	0.57 (152.0)	0.78 (152.0)	0.87 (156.5)
eQTL counts	0.90 (145.5)	0.41 (132.0)	0.88 (144.5)

Table 4 Tissue specificity evaluation against PANDA. Wilcoxon signed-rank test p-values (and W statistics) between matching tissue ranks of METANets and PANDA networks for 26 tissues. Across all thresholds of average number of targets per TF, METANets and PANDA networks were not significantly different.

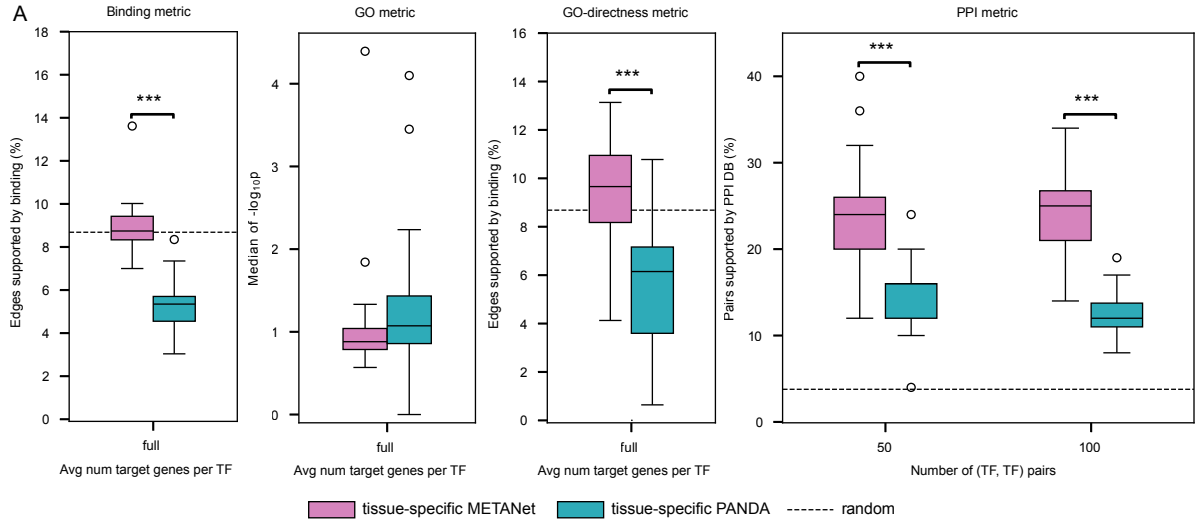

Fig. 4 Network quality evaluation of tissue-specific networks (A-D) Network quality evaluation of tissue-specific METANets and tissue-specific PANDA networks. (**A**) Binding metric. Tissue-specific METANets significantly outperformed tissue-specific PANDA networks (paired *t*-test, P<0.05). (**B**) GO metric. No significant difference between tissue-specific METANets and tissue-specific PANDA networks. (**C**) GO-directness metric. Tissue-specific METANets outperformed tissue-specific PANDA networks (P<0.001). (**D**) Tissue-specific METANets outperformed tissue-specific PANDA networks at thresholds of 50 and 100 (TF, TF) pairs. **P*≤0.05, ***P*≤0.01, ****P*≤0.001.

Empirical test for tissue-specificity for 26 tissues		
Network	eQTL support	eQTL count
Tissue-specific METANet	**0.034 (307)**	0.250 (338)

Tissue-specific PANDA	0.233 (332)	0.199 (341)

Table 5 Empirical test of tissue-specific METANet and tissue-specific PANDA against the null across 26 tissues. Permutation test p-values (and the sum of observed matching ranks) against the distribution of 10000 permutation sums. Tissue-specific METANets showed significant tissue specificity in eQTL support and tissue-specific PANDA networks were not significantly tissue-specific. Neither tissue-specific METANets nor tissue-specific PANDA networks were significantly tissue-specific in eQTL counts. See also Additional file 1: Fig. S1.

Wilcoxon signed-rank test vs. tissue-specific PANDA	
eQTL support	0.485 (125.0)
eQTL counts	0.881 (168.5)

Table 6 Tissue specificity evaluation against tissue-specific PANDA. Wilcoxon signed-rank test p-values (and W statistics) between matching tissue ranks of tissue-specific METANets and tissue-specific PANDA networks for 26 tissues. Full tissue-specific networks were evaluated, without thresholding for top-scoring edges. No significant difference between tissue-specific METANets and tissue-specific PANDA in both eQTL support and eQTL count.

METANets Comparable to PANDA in Tissue Specificity but Outperform in Network Quality

Against the PANDA networks (26 tissues), neither PANDA nor METANets showed significant tissue specificity in eQTL support (Table 2). METANets were significantly better than chance in eQTL count in one threshold (average 100 targets per TF), but PANDA was not. The Wilcoxon signed-rank test found no significant difference between METANets and PANDA in either metric, indicating comparable tissue specificity (Table 4).

It is important to note that PANDA was originally developed as a general framework for inferring regulatory networks by integrating motif, expression, and protein-protein interaction data [36]. In its original form, PANDA produces a fully connected, weighted TF-gene network, but does not explicitly model tissue-specificity. The concept of *tissue-specific* PANDA networks was introduced by Sonawane et al. (2017), who applied PANDA to 38 GTEx tissues and then post-processed the resulting networks with a tissue specificity filtering step. In this approach, each edge's weight in a given tissue was compared to its distribution across all tissues, and a normalized interquartile range (IQR) score was used to identify edges unusually strong in one tissue. Edges with high IQR scores were retained as "tissue specific."

289 To enable direct comparison with the tissue-specific filtering strategy of Sonawane et al., we
290 applied the same IQR-based approach to METANets. These filtered networks were evaluated
291 without thresholding for top-scoring edges, as the number of TFs, genes, and edges varied
292 drastically across tissues. Tissue-specific METANets showed significant eQTL support but not
293 eQTL counts, while tissue-specific PANDA networks showed no significant signal in either metric
294 (Table 5). Again, no significant difference was found between tissue-specific versions of
295 METANets and PANDA networks (Wilcoxon signed-rank $P > 0.05$) (Table 6).

296 Despite similar performance in tissue specificity, tissue-specific METANets outperformed tissue-
297 specific PANDA in the Binding and GO-directness metrics ($P<0.05$ and $P<0.001$, respectively),
298 while performing similarly in the GO metric (Fig. 4A-C). Tissue-specific METANets also
299 performed better in the PPI metric overall (Fig. 4D).

300 **Tissue-specific METANets are Useful for Genetic Analyses**

301 To evaluate the biological relevance of tissue-specific METANets, we applied FISHNET, a method
302 that uses prior biological knowledge to identify genes with suggestive association signals that are
303 more likely to replicate across independent datasets. FISHNET integrates gene-level P-values with
304 networks and functional annotations, based on the principle that spurious signals are randomly
305 scattered, while true signals tend to cluster in functionally coherent subnetwork. Combining
306 network topology, functional enrichment, and permutation-derived thresholds, FISHNET
307 identifies a focused set of exceptional candidates termed FISHNET genes.

308 We used the whole blood tissue-specific METANet as the reference network input to FISHNET.
309 This network was partitioned into modules using a network modularization algorithm based on
310 modularity optimization. As input, we provided gene-level P-values derived from measured TWAS

311 of 11 traits associated with cardiovascular risk in the Long Life Family Study (LLFS) cohort and
312 performed replication analysis in the Framingham Heart Study (FHS) cohort [37, 38]. Across the
313 11 traits, FISHNET identified 114 FISHNET genes of which 24 were replicated. Using the whole
314 blood tissue-specific PANDA network as input, we identified 0 FISHNET genes (see Additional
315 file 2: Tables S1 and S2).

316 A central biological story emerging from these results centers on *SREBF2*, a replicated FISHNET
317 gene negatively associated with HDL. Our findings support a model in which high HDL levels
318 may suppress *SREBF2* activity, whereas low HDL permits *SREBF2* activation and sustains
319 inflammatory signaling. *SREBF2* encodes SREBP2, a master transcription factor of sterol
320 metabolism, canonically activated in adipocytes under low-cholesterol conditions [39-41]. Recent
321 studies have shown that inflammatory stress can also activate SREBP2, enabling it to bind and
322 induce inflammatory and interferon-response genes [42-44]. In the whole blood tissue-specific
323 METANet, *SREBF2* appears in the same module as other negatively HDL-associated FISHNET
324 genes–*S100A9, S100A12, TLR8,* and *ITGB2*–all of which converge on pathways that activate the
325 NADPH oxidase 2 (NOX2) complex. *S100A9* and *S100A12* are calgranulins, a type of alarmin,
326 that can act as ligands for the receptor for advanced glycation end products (RAGE), triggering
327 assembly and activation of the NOX2 complex [45-47]. *TLR8* is a pattern recognition receptor that
328 primes NOX2 activation in human neutrophils [48]. *ITGB2* promotes leukocyte adhesion to the
329 endothelium and triggers "outside-in" signaling that further promotes NOX2 assembly and
330 activation [49]. Collectively, these genes link *SREBF2* to NOX2-based excessive production of
331 reactive oxygen species (ROS), which oxidize HDL and convert it into a pro-inflammatory form
332 (oxHDL) that contributes to atherosclerosis. Notably, *S100A9*, *S100A12*, and *TLR8* also activate
333 NF-κB, inducing cytokines (TNFα and IL-1β) known to activate *SREBF2* [50-53]. Together, these

334 findings suggest a positive feedback loop in which reduced HDL allows *SREBF2* activation, which
335 in turn promotes inflammatory cascades that further lower HDL, providing a mechanistic link
336 between cholesterol metabolism, immune activation, and chronic low-grade inflammation
337 characteristic of obesity.

338 **Discussion**

339 In this work, we developed METANets, a supervised ensemble framework that combines TF
340 motifs, cis-regulatory element (CRE) activity, and expression-derived features to construct tissue-
341 specific TF network maps across 36 human tissues. Our modeling approach relied on three
342 fundamental ideas: (1) using gene expression data (functional) to predict TF binding events (direct)
343 prioritizes direct, functional targets of TFs, (2) incorporating both tissue-aggregate and tissue-
344 specific expression features captures both the shared and tissue-specific regulatory edges, and (3)
345 using tree-based regression models to incorporate both linear and nonlinear expression patterns
346 improves model robustness. Our evaluations across multiple orthogonal metrics show that our
347 approach significantly outperforms motif-only, expression-only, and unsupervised integrative
348 methods in capturing direct and functionally coherent TF-TG relationships.

349 Gene expression-based models are adept at identifying functionally coherent gene sets, but the
350 expression of a TF and a potential target gene could be correlated for reasons other than the TF
351 regulating the gene directly: the TF might regulate indirectly (mediated by other proteins), or the
352 TF and the gene might simply be co-regulated by a third factor. Conversely, motif-based
353 approaches anchor predictions in physical binding potential but cannot confirm functional
354 relevance, as potential binding sites are frequently unoccupied [6, 7] and even occupied sites often
355 do not affect transcription [3-5, 54, 55]. By training our models to explicitly predict TF ChIP-seq

binding events, METANets leverage the strengths of both data types without a noticeable tradeoff. Our results demonstrate METANets consistently outperforming the motif-only Marbach networks and our expression-only ETANets in the GO-directness metric that jointly assess directness and functional coherence.

METANets also illustrate the power of incorporating diverse feature types to model the complexity of gene regulation. We found that regulatory relationships are not exclusively linear or non-linear; by combining features from both LASSO (linear) and BART (non-linear) regression models, METANets captured a more direct and functionally coherent set of TF targets. Similarly, we that found neither tissue-specific nor tissue-aggregate features clearly outperformed the other in isolation, but together they predicted more robust sets of TF targets. The METANets approach performed best in the individual Binding and GO metrics without sacrificing GO-directness as a tradeoff. The practical utility of this approach was highlighted in our downstream genetic analysis, where the whole blood METANet enabled FISHNET to identify replicated gene-trait associations, proposing a module linking *SREBF2* to HDL and chronic inflammation.

Our work presents a subtle methodological shift for the field of regulatory network inference. For years, the community has largely relied on two distinct categories of methods: those based on physical binding evidence like TF motifs, which often lack functional validation, and those based on gene expression, which are susceptible to confounding by indirect relationships. While unsupervised integration methods have attempted to bridge this gap, METANets presents a new alternative by successfully implementing a supervised learning framework. The core innovation is reframing the problem: instead of merely correlating different data types, we use functional data (gene expression) to predict direct physical interactions (TF binding).

This approach provides a methodological foundation for using the ever-expanding space of omics data integration. The principle of using one data modality as a ground-truth label to train predictive models on other, more widely available data types can be generalized beyond the specific application in this work. It makes TF network mapping more broadly applicable, as ChIP-seq data for every TF is not available, while RNA-seq data is abundant. Our approach can make use of ChIP-seq in predicting targets of TFs that have ChIP-Seq data, but it can also extract general patterns in gene expression data that are predictive of direct binding and use those patterns to infer targets for TFs that do not have ChIP-Seq data. We also presented a new evaluation of networks based on utility for applications rather than network characteristics.

Several limitations suggest avenues for further development. One key limitation is the lack of tissue-specific TF binding data in our framework. While tissue-specific expression data are readily available at scale, obtaining comparable TF binding data across tissues remains challenging. As a result, METANets relied on tissue-aggregated TF binding data. Emerging paired-seq technologies that jointly measure TF binding and gene expression in the same cell could make predictions of TF binding events more robust and the single-cell measurements can facilitate the direct mapping of granular cell-type-specific networks. Development of scalable methods to address the inherent sparsity of single-cell data will be crucial for realizing this goal.

Another opportunity lies in extending the METANets framework to additional omics data types. Chromatin accessibility (ATAC-seq) and chromatin conformation (Hi-C, HiChIP) represent rich sources of contextual information that could further refine predictions of TF-TG relationships. Beyond tissues, the same framework could be adapted to disease states, environmental exposures, or dynamic perturbations, producing condition-specific network maps that illuminate regulatory mechanisms underlying phenotypic variation.

401 In summary, METANets advance the state of regulatory network mapping by demonstrating that
402 supervised integrative approaches can accurately prioritize direct, functional TF-TG interactions.
403 Beyond their immediate utility, METANets highlight the potential of combining heterogeneous
404 data sources to produce context-dependent regulatory maps, setting the stage for increasingly
405 precise models of human gene regulation.

406 **Methods**

407 **Transcription factor (TF) binding data**

408 TF binding locations were determined using data from TF ChIP-seq (Chromatin
409 Immunoprecipitation followed by sequencing) experiments. We downloaded TF ChIP-seq data
410 from the REMAP database, which consists of 5,798 experiments for 1,135 TFs across over 600
411 tissues and cell lines [28]. TFs that did not meet the criteria defined by Lambert et al. [56] were
412 removed. Additionally, binding peaks with FDR q-value ≤ 0.01 were removed for each TF. After
413 these filtering steps, we retained 799 TFs with high-confidence binding peaks.

414 **Regulatory element annotations**

415 Promoter and enhancer annotations were obtained from the FANTOM5 database, which contains
416 data on 209,911 promoters and 41,456 enhancers for 17,778 protein-coding genes, along with
417 activity level measurements across 1,829 cell, tissue, and cell-line samples [32, 33].

418 **Mapping binding peaks to regulatory regions**

419 For each TF, all high-confidence (FDR q-value ≤ 0.01) binding peaks within FANTOM5 promoter
420 and enhancer regions were examined. In cases where multiple overlapping peaks were present
421 within a single regulatory element, the peak with the highest confidence score among the

422 overlapping regions was retained. A score was calculated for each regulatory element by summing
423 the confidence scores of the retained peaks, weighted by their length in base pairs (see Additional
424 file 1: Fig. S2). For each gene, a gene score was derived by aggregating the scores of all regulatory
425 elements annotated for the gene. Since TF binding events within each regulatory region represent
426 independent evidence of gene regulation, the aggregated score across regulatory elements reflects
427 the cumulative likelihood of the TF regulating the gene.

428 **Marbach networks**

429 We downloaded 394 cell-type- and tissue-specific networks from Marbach et al. [23]. We formed
430 36 tissue networks by merging closely related networks by taking the graph union (union of the
431 node and edge sets while retaining the maximum weight for each edge). We used the mapping file
432 for GTEx eQTL validation from Marbach et al. as the baseline reference for 13 GTEx tissues. We
433 manually curated the remaining 23 tissues using annotations from the BRENDA tissue ontology
434 [57] (see Additional file 2: Table S3).

435 **Tissue-specific gene expression data**

436 We downloaded bulk RNA-seq read counts and transcripts per million (TPM) data from the
437 Genotype-Tissue Expression (GTEx) consortium V8 release [58]. The GTEx project provides a
438 comprehensive catalogue of gene expression profiles across 54 human tissues using biospecimens
439 from approximately 1,000 postmortem donors. The GTEx read count data for protein-coding genes
440 was filtered to remove lowly expressed genes. Specifically, genes with ≤ 3 counts per million in
441 $\geq 98.5\%$ of samples were removed. Genes that failed to pass this filter were also removed from
442 the TPM data.

443 **Construction of binding labels**

The network mapping approach uses binary labels representing binding events for each TF-gene pair. We assigned a positive label to genes with nonzero binding scores for each TF to construct binary labels for TF-gene binding events. If more than 10% of genes had a positive score for a given TF, only the top 10% of highest-scoring genes were retained as positive labels. All remaining gene pairs were labeled as negative.

Construction of gene expression-based features

The gene expression-based features were constructed by training two regression algorithms, LASSO and Bayesian Additive Regression Trees (BART), to predict the expression level of each gene from the expression levels of all TFs [11-13, 59]. LASSO (Least Absolute Shrinkage and Selection Operator) is a linear regression method with L1 regularization. BART is a non-parametric sum-of-trees model that captures complex, non-linear relationships through an ensemble of regression trees. Features for a given TF-gene pair reflect the effect of the TF's expression level on the gene's predicted expression level in the trained models – for example, the feature extracted from a LASSO model is the learned coefficient of the TF's expression level. Both algorithms are run on the filtered tissue-specific gene expression TPM data and, separately, on the aggregated TPM data from all GTEx tissue samples, yielding 4 regression features for each TF-gene pair.

Training Machine Learning Algorithms

METANets were inferred by combining diverse features using the XGBoost model [60]. **Error! Reference source not found.**A depicts the inputs and outputs of the reconstruction of METANets. A separate XGBoost model was trained for each of the 36 tissues, where instances are TF-target gene (TG) edges for a specific tissue. This machine learning model employs gradient-boosted

466 decision trees for regression and classification tasks. The Python XGBClassifier implementation
467 (v2.1.1) was used, with the binding labels as the response variable and a combination of tissue-
468 specific and tissue-aggregate features as predictors. Tissue-specific features include TF-TG edges
469 from the Marbach networks and edges inferred from LASSO and BART. Tissue-aggregate features
470 include edges inferred from LASSO and BART using samples from all tissues.

471 The models were trained using 10-fold nested cross-validation, stratified based on the distribution
472 of binding labels. A new XGBoost model was trained for each outer fold, using the remaining 9
473 folds as training data and the single fold a test data. The training data within each outer fold was
474 further divided into 5 inner folds for hyperparameter tuning. Hyperparameter tuning was
475 conducted using the Tree of Parzen Estimators (TPE) from the Hyperopt package (v.0.2.7), which
476 models hyperparameters as nodes in a tree structure [61, 62]. The predicted probabilities that the
477 TF binds in the gene's CREs, in the test folds, are used as edge scores in our tissue-specific
478 METANets.

479 **Binding metric**

480 We calculated the percentage of top-scoring TF-TG edges supported by TF binding data at different
481 thresholds. Top-scoring edges were defined by sorting edges by the edge score in descending order.
482 Thresholds were scaled to the number of target genes per TF. This is an average, so different TFs
483 have different numbers of target genes at each threshold. A higher percentage of binding support
484 reflects greater enrichment of direct edges in the network map. The random expectation is the
485 probability that a randomly selected TF-TG edge is supported by the binding data.

486 **GO metric**

487 We performed GO enrichment analysis for each TF's targets using GO-Term-Finder v0.86 [63].
488 To focus on terms associated with specific biological processes, rather than extremely generic
489 terms like 'biosynthetic process', we excluded terms with number of annotated genes below 5 or
490 exceeding 300. Each TF's score was the maximum –log P-value over all GO terms. If a TF's target
491 set showed no significant enrichment for any GO biological process, we assigned the TF a P-value
492 of 1 (equivalent to –log P = 0). The final network-level score was taken as the median of these
493 scores across all TFs. We did these analyses for each threshold, as in the binding metric, but without
494 excluding TFs that did not have binding data available.

495 **GO-directness metric**

496 To determine whether improvements on the GO evaluation metric came at the cost of including
497 indirect edges, we calculated, for each TF, the percentages of binding support for edges that
498 matched the GO term with the highest minus log P-value. For each threshold, we reported the
499 average percentage of binding support across TFs. The random expectation is identical to that of
500 the Binding metric.

501 **PPI metric**

502 First, we calculated the Jaccard similarity between the sets of target genes of each (TF, TF) pair;
503 the Jaccard similarity is 1 when the target sets are the same and 0 when none of the targets in the
504 sets are shared. Next, we sorted (TF, TF) pairs by Jaccard similarity and calculated the percentage
505 of the top 100 pairs that are supported by physical interactions from the protein-protein interaction
506 (PPI) STRING database[34]. The database has 477,903 interactions predicted with high
507 confidence (SRTING score ≥ 0.7). A higher percentage of PPI support means that (TF, TF) pairs
508 are likely to work as a part of a known physical protein complex. We did these analyses for the

TF-TG edges at thresholds of 50 and 100 targets per TF. The random expectation is the probability that a randomly selected pair of TFs present in a tissue network is supported by the PPI database. This probability was computed separately for each tissue and then averaged across the 36 tissues.

Empirical Null for Network Quality

To establish an empirical null to evaluate network quality, we generated 50 permuted networks by randomly shuffling the fully connected METANet edge scores. Each permuted network was subjected to the same cutoff thresholds as the real networks and evaluated across all four metrics, providing a consistent null distribution.

Processing of PANDA Networks

Gene regulatory networks for 38 tissues were downloaded from Zenodo [64]. To enable direct comparison with our METANets, we retained the 26 tissues shared with the 36 METANet tissues. For each tissue, we used two versions of the PANDA networks: (1) the full networks containing weighted edges between all transcription factors (TFs) and genes, and (2) the tissue-specific networks provided in the same dataset, in which tissue-specific edges had already been annotated by the original authors. To maintain consistency, both versions of the PANDA networks were subjected to the same downstream processing steps: restricting to protein-coding gene targets, removing self-regulatory edges, and retaining only the subset of TFs common to both PANDA and METANets.

Tissue-Specific Edge Filtering

Tissue-specific regulatory edges were identified using the interquartile range (IQR)-based filtering approach originally introduced by Sonawane et al. [20]. For each TF-TG pair *(i,j)*, the edge weight

in tissue t is denoted $w_{ij}^{(t)}$. For each edge, we computed a normalized specificity score by comparing the tissue-specific weight to the distribution of weights across all tissues:

$$s_{ij}^{(t)} = \frac{w_{ij}^{(t)} - \mathrm{median}(w_{ij})}{\mathrm{IQR}(w_{ij})},$$

where $\mathrm{median}(w_{ij})$ and $\mathrm{IQR}(w_{ij})$ are calculated across all tissues for the same edge. Edges with $s_{ij}^{(t)} > 2$ were considered specific to tissue t. This filtering approach was applied to METANets to construct tissue-specific versions of our networks, directly comparable to the tissue-specific PANDA networks generated in Sonawane et al. [20].

Validation of Tissue Specificity using eQTL Data

We downloaded cis-eQTL data from the GTEx consortium V8 release [31]. Significant variant-gene associations were obtained for each of the 36 tissues. For each TF-target gene (TF-TG) edge in a network, we determined whether one or more of a tissue's significant variants overlapped the TF's ChIP-seq peaks within the promoter or annotated enhancers of the target gene. We defined two complementary metrics for a given network:

1. eQTL support: the fraction of edges supported by at least one tissue-specific eQTL.
2. eQTL count: the sum, across all edges, of the number of tissue-specific eQTLs supporting each edge.

To assess a network's tissue-specificity, we performed a permutation test. First, we calculated an observed statistic S_{obs} by ranking all 36 networks within each tissue and summing the ranks of the "matching" pairs (e.g., the rank of the liver network in the liver tissue). A smaller sum indicates better overall specificity. Next, to generate a null distribution, we performed 10000 permutations.

550 In each permutation, we randomly shuffled the network labels before summing the ranks of the
551 newly assigned matching pairs to obtain the permutation statistic S_{perm}. This process simulates
552 the null hypothesis that there is no true correspondence between a network and its native tissue.
553 The empirical p-value was then calculated as the proportion of permutations with $S_{perm} \leq S_{obs}$
554 (see Additional file 1: Fig. S4)

555 Direct comparisons between different networks were performed using the Wilcoxon signed-rank
556 test on the distributions of matching ranks.

557 **Gene-trait Association Analysis using FISHNET**

558 We applied FISHNET, a network-guided replication framework that integrates prior biological
559 knowledge with gene-level association statistics, to evaluate the biological relevance of tissue-
560 specific METANets [65]. FISHNET is designed to prioritize genes that exhibit suggestive
561 association signals but are more likely than expected by chance to replicate across independent
562 datasets.

563 The primary input to FISHNET was the whole blood tissue-specific METANet. For comparison,
564 we also applied FISHNET using the whole blood tissue-specific PANDA network constructed by
565 Sonawane et al. (2017). Each network was partitioned into modules using the and M1 modularity
566 optimization algorithm, as implemented in the MONET toolbox [66].

567 As input gene-level signals, we used transcriptome-wide association study (TWAS) p-values
568 derived from the Long Life Family Study (LLFS). We considered 11 traits associated with
569 cardiovascular risk, including body mass index (BMI), high-density lipoprotein (HDL), low-
570 density lipoprotein (LDL), and triglycerides, among others. Replication was assessed using gene-
571 level association statistics from the Framingham Heart Study (FHS) cohort, analyzed for the same

572 11 traits. Details on RNA-seq data processing and covariate adjustments have been previously
573 described by Acharya et al. [65].

574 **Declaration of interests**

575 The authors declare no competing interests.

576 **Acknowledgements**

577 We are grateful to the entire Long Life consortium, its participants, and its investigators, without
578 whom this work would not have been possible. This work was supported by grant AG063893 from
579 the National Institute on Aging (for collecting data from Long Life Family Study cohort) and by
580 grant GM141012 from the National Institute of General Medical Sciences (for development of TF
581 network mapping algorithms).

582 The Framingham Heart Study is conducted and supported by the National Heart, Lung, and Blood
583 Institute (NHLBI) in collaboration with Boston University (Contract No. N01-HC-25195,
584 HHSN268201500001I and 75N92019D00031). This manuscript was not prepared in collaboration
585 with investigators of the Framingham Heart Study and does not necessarily reflect the opinions or
586 views of the Framingham Heart Study, Boston University, or NHLBI.

587 **Data and code availability**

588 METANets, tissue-specific METANets, as well as other networks evaluated in this work are
589 available at https://doi.org/10.5281/zenodo.17309371. Code is available at
590 https://github.com/BrentLab/METANet. All data for modeling were downloaded from the publicly
591 available sources indicated in Methods.

References

1. Janky Rs, Verfaillie A, Imrichová H, Van de Sande B, Standaert L, Christiaens V, Hulselmans G, Herten K, Naval Sanchez M, Potier D: **iRegulon: from a gene list to a gene regulatory network using large motif and track collections**. *PLoS computational biology* 2014, **10**(7):e1003731.

2. Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan K-K, Cheng C, Mu XJ, Khurana E, Rozowsky J, Alexander R: **Architecture of the human regulatory network derived from ENCODE data**. *Nature* 2012, **489**(7414):91-100.

3. Slattery M, Zhou T, Yang L, Machado ACD, Gordân R, Rohs R: **Absence of a simple code: how transcription factors read the genome**. *Trends in biochemical sciences* 2014, **39**(9):381-399.

4. White MA, Myers CA, Corbo JC, Cohen BA: **Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks**. *Proceedings of the National Academy of Sciences* 2013, **110**(29):11952-11957.

5. Cusanovich DA, Pavlovic B, Pritchard JK, Gilad Y: **The functional consequences of variation in transcription factor binding**. *PLoS genetics* 2014, **10**(3):e1004226.

6. Wang J, Zhuang J, Iyer S, Lin X, Whitfield TW, Greven MC, Pierce BG, Dong X, Kundaje A, Cheng Y: **Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors**. *Genome research* 2012, **22**(9):1798-1812.

7. Dror I, Golan T, Levy C, Rohs R, Mandel-Gutfreund Y: **A widespread role of the motif environment in transcription factor binding across diverse protein families**. *Genome Research* 2015, **25**(9):1268-1280.

8. Worsley Hunt R, Wasserman WW: **Non-targeted transcription factors motifs are a systemic component of ChIP-seq datasets**. *Genome biology* 2014, **15**:1-16.

9. Horton CA, Alexandari AM, Hayes MG, Marklund E, Schaepe JM, Aditham AK, Shah N, Suzuki PH, Shrikumar A, Afek A: **Short tandem repeats bind transcription factors to tune eukaryotic gene expression**. *Science* 2023, **381**(6664):eadd1250.

10. Langfelder P, Horvath S: **WGCNA: an R package for weighted correlation network analysis**. *BMC bioinformatics* 2008, **9**:1-13.

11. Abid D, Brent MR: **NetProphet 3: a machine learning framework for transcription factor network mapping and multi-omics integration**. *Bioinformatics* 2023, **39**(2).

12. Kang Y, Liow HH, Maier EJ, Brent MR: **NetProphet 2.0: mapping transcription factor networks by exploiting scalable data resources**. *Bioinformatics* 2018, **34**(2):249-257.

628 13. Haynes BC, Maier EJ, Kramer MH, Wang PI, Brown H, Brent MR: **Mapping functional**
629 **transcription factor networks from gene expression data**. *Genome Res* 2013,
630 **23**(8):1319-1328.

631 14. Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P: **Inferring regulatory networks from**
632 **expression data using tree-based methods**. *PloS one* 2010, **5**(9):e12776.

633 15. Skok Gibbs C, Jackson CA, Saldi G-A, Tjärnberg A, Shah A, Watters A, De Veaux N,
634 Tchourine K, Yi R, Hamamsy T: **High-performance single-cell gene regulatory**
635 **network inference at scale: the Inferelator 3.0**. *Bioinformatics* 2022, **38**(9):2519-
636 2528.

637 16. Xiao N, Zhou A, Kempher ML, Zhou BY, Shi ZJ, Yuan M, Guo X, Wu L, Ning D, Van
638 Nostrand J *et al*: **Disentangling direct from indirect relationships in association**
639 **networks**. *Proc Natl Acad Sci U S A* 2022, **119**(2).

640 17. Feizi S, Marbach D, Medard M, Kellis M: **Network deconvolution as a general**
641 **method to distinguish direct dependencies in networks**. *Nat Biotechnol* 2013,
642 **31**(8):726-733.

643 18. Mercatelli D, Scalambra L, Triboli L, Ray F, Giorgi FM: **Gene regulatory network**
644 **inference resources: A practical overview**. *Biochim Biophys Acta Gene Regul Mech*
645 2020, **1863**(6):194430.

646 19. Melé M, Ferreira PG, Reverter F, DeLuca DS, Monlong J, Sammeth M, Young TR,
647 Goldmann JM, Pervouchine DD, Sullivan TJ: **The human transcriptome across**
648 **tissues and individuals**. *Science* 2015, **348**(6235):660-665.

649 20. Sonawane AR, Platig J, Fagny M, Chen C-Y, Paulson JN, Lopes-Ramos CM, DeMeo DL,
650 Quackenbush J, Glass K, Kuijjer ML: **Understanding Tissue-Specific Gene**
651 **Regulation**. *Cell Reports* 2017, **21**(4):1077-1088.

652 21. Gamazon ER, Segrè AV, Van De Bunt M, Wen X, Xi HS, Hormozdiari F, Ongen H,
653 Konkashbaev A, Derks EM, Aguet F: **Using an atlas of gene regulation across 44**
654 **human tissues to inform complex disease-and trait-associated variation**. *Nature*
655 *genetics* 2018, **50**(7):956-967.

656 22. Haigis KM, Cichowski K, Elledge SJ: **Tissue-specificity in cancer: The rule, not the**
657 **exception**. *Science* 2019, **363**(6432):1150-1151.

658 23. Marbach D, Lamparter D, Quon G, Kellis M, Kutalik Z, Bergmann S: **Tissue-specific**
659 **regulatory circuits reveal variable modular perturbations across complex**
660 **diseases**. *Nat Methods* 2016, **13**(4):366-370.

661 24. Saha A, Kim Y, Gewirtz AD, Jo B, Gao C, McDowell IC, Engelhardt BE, Battle A, Aguet
662 F, Ardlie KG: **Co-expression networks reveal the tissue-specific regulation of**
663 **transcription and splicing**. *Genome research* 2017, **27**(11):1843-1858.

25. Pierson E, Consortium G, Koller D, Battle A, Mostafavi S: **Sharing and specificity of co-expression networks across 35 human tissues**. *PLoS computational biology* 2015, **11**(5):e1004220.

26. Acharya S: **Thinking Beyond Association Tests: Using Mathematical Models to Identify Genes Influencing Exceptional Longevity**. *Ph.D.* United States -- Missouri: Washington University in St. Louis; 2025.

27. Chen T, Guestrin C: **Xgboost: A scalable tree boosting system**. In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining: 2016*. 785-794.

28. Cheneby J, Menetrier Z, Mestdagh M, Rosnet T, Douida A, Rhalloussi W, Bergon A, Lopez F, Ballester B: **ReMap 2020: a database of regulatory regions from an integrative analysis of Human and Arabidopsis DNA-binding sequencing experiments**. *Nucleic Acids Res* 2020, **48**(D1):D180-D188.

29. Tibshirani R: **Regression shrinkage and selection via the lasso**. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 1996, **58**(1):267-288.

30. Chipman HA, George EI, McCulloch RE: **BART: Bayesian additive regression trees**. *The Annals of Applied Statistics* 2010, **4**(1).

31. The GTEx Consortium: **The GTEx Consortium atlas of genetic regulatory effects across human tissues**. *Science* 2020, **369**(6509):1318-1330.

32. Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T *et al*: **An atlas of active enhancers across human cell types and tissues**. *Nature* 2014, **507**(7493):455-461.

33. Consortium F, the RP, Clst, Forrest AR, Kawaji H, Rehli M, Baillie JK, de Hoon MJ, Haberle V, Lassmann T *et al*: **A promoter-level mammalian expression atlas**. *Nature* 2014, **507**(7493):462-470.

34. Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, Santos A, Doncheva NT, Roth A, Bork P *et al*: **The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible**. *Nucleic Acids Research* 2017, **45**(D1):D362-D368.

35. Consortium G: **The GTEx Consortium atlas of genetic regulatory effects across human tissues**. *Science* 2020, **369**(6509):1318-1330.

36. Glass K, Huttenhower C, Quackenbush J, Yuan G-C: **Passing messages between biological networks to refine predicted interactions**. *PloS one* 2013, **8**(5):e64832.

37. KANNEL WB, FEINLEIB M, McNAMARA PM, GARRISON RJ, CASTELLI WP: **AN INVESTIGATION OF CORONARY HEART DISEASE IN FAMILIES: THE FRAMINGHAM OFFSPRING STUDY**. *American Journal of Epidemiology* 1979, **110**(3):281-290.

700 38. Splansky GL, Corey D, Yang Q, Atwood LD, Cupples LA, Benjamin EJ, D'Agostino Sr RB,
701 Fox CS, Larson MG, Murabito JM: **The third generation cohort of the National Heart,**
702 **Lung, and Blood Institute's Framingham Heart Study: design, recruitment, and**
703 **initial examination**. *American journal of epidemiology* 2007, **165**(11):1328-1335.

704 39. Jiang T, Wang Z, Proctor G, Moskowitz S, Liebman SE, Rogers T, Lucia MS, Li J, Levi M:
705 **Diet-induced obesity in C57BL/6J mice causes increased renal lipid**
706 **accumulation and glomerulosclerosis via a sterol regulatory element-binding**
707 **protein-1c-dependent pathway**. *Journal of biological chemistry* 2005,
708 **280**(37):32317-32325.

709 40. Bauer S, Wanninger J, Schmidhofer S, Weigert J, Neumeier M, Dorn C, Hellerbrand C,
710 Zimara N, Schäffler A, Aslanidis C: **Sterol regulatory element-binding protein 2**
711 **(SREBP2) activation after excess triglyceride storage induces chemerin in**
712 **hypertrophic adipocytes**. *Endocrinology* 2011, **152**(1):26-35.

713 41. Horton JD, Goldstein JL, Brown MS: **SREBPs: activators of the complete program of**
714 **cholesterol and fatty acid synthesis in the liver**. *The Journal of clinical investigation*
715 2002, **109**(9):1125-1131.

716 42. Zhao L, Chen Y, Tang R, Chen Y, Li Q, Gong J, Huang A, Varghese Z, Moorhead JF, Ruan
717 XZ: **Inflammatory stress exacerbates hepatic cholesterol accumulation via**
718 **increasing cholesterol uptake and de novo synthesis**. *Journal of gastroenterology*
719 *and hepatology* 2011, **26**(5):875-883.

720 43. Kusnadi A, Park SH, Yuan R, Pannellini T, Giannopoulou E, Oliver D, Lu T, Park-Min K-
721 H, Ivashkiv LB: **The cytokine TNF promotes transcription factor SREBP activity and**
722 **binding to inflammatory genes to activate macrophages and limit tissue repair**.
723 *Immunity* 2019, **51**(2):241-257. e249.

724 44. Fowler JWM, Boutagy NE, Zhang R, Horikami D, Whalen MB, Romanoski CE, Sessa
725 WC: **SREBP2 regulates the endothelial response to cytokines via direct**
726 **transcriptional activation of KLF6**. *Journal of Lipid Research* 2023, **64**(8).

727 45. Dong H, Zhang Y, Huang Y, Deng H: **Pathophysiology of RAGE in inflammatory**
728 **diseases**. *Frontiers in Immunology* 2022, **13**:931473.

729 46. Xiao X, Yang C, Qu S-L, Shao Y-D, Zhou C-Y, Chao R, Huang L, Zhang C: **S100 proteins**
730 **in atherosclerosis**. *Clinica chimica acta* 2020, **502**:293-304.

731 47. Bowman MAH, Schmidt AM: **S100/calgranulins EN-RAGEing the blood vessels:**
732 **implications for inflammatory responses and atherosclerosis**. *American journal*
733 *of cardiovascular disease* 2011, **1**(1):92.

734 48. Makni-Maalej K, Marzaioli V, Boussetta T, Belambri SA, Gougerot-Pocidalo M-A,
735 Hurtado-Nedelec M, Dang PM-C, El-Benna J: **TLR8, but not TLR7, induces the**
736 **priming of the NADPH oxidase activation in human neutrophils**. *Journal of*
737 *Leucocyte Biology* 2015, **97**(6):1081-1087.

738 49. Chen S-H, Oyarzabal EA, Hong J-S: **Critical role of the Mac1/NOX2 pathway in**
739 **mediating reactive microgliosis-generated chronic neuroinflammation and**
740 **progressive neurodegeneration**. *Current opinion in pharmacology* 2016, **26**:54-60.

741 50. Zhang X, Sun D, Zhou X, Zhang C, Yin Q, Chen L, Tang Y, Liu Y, Morozova-Roche LA:
742 **Proinflammatory S100A9 stimulates TLR4/NF-κB signaling pathways causing**
743 **enhanced phagocytic capacity of microglial cells**. *Immunology letters* 2023,
744 **255**:54-61.

745 51. Cluzeau T, McGraw KL, Irvine B, Masala E, Ades L, Basiorka AA, Maciejewski J,
746 Auberger P, Wei S, Fenaux P: **Pro-inflammatory proteins S100A9 and tumor**
747 **necrosis factor-α suppress erythropoietin elaboration in myelodysplastic**
748 **syndromes**. *haematologica* 2017, **102**(12):2015.

749 52. Yang Z, Yan WX, Cai H, Tedla N, Armishaw C, Di Girolamo N, Wang HW,
750 Hampartzoumian T, Simpson JL, Gibson PG: **S100A12 provokes mast cell activation:**
751 **a potential amplification pathway in asthma and innate immunity**. *Journal of*
752 *allergy and clinical immunology* 2007, **119**(1):106-114.

753 53. Qin J, Yao J, Cui G, Xiao H, Kim TW, Fraczek J, Wightman P, Sato S, Akira S, Puel A:
754 **TLR8-mediated NF-κB and JNK activation are TAK1-independent and MEKK3-**
755 **dependent**. *Journal of Biological Chemistry* 2006, **281**(30):21013-21021.

756 54. Kang Y, Patel NR, Shively C, Recio PS, Chen X, Wranik BJ, Kim G, McIsaac RS, Mitra R,
757 Brent MR: **Dual threshold optimization and network inference reveal convergent**
758 **evidence from TF binding locations and TF perturbation responses**. *Genome*
759 *research* 2020, **30**(3):459-471.

760 55. Mahendrawada L, Warfield L, Donczew R, Hahn S: **Low overlap of transcription**
761 **factor DNA binding and regulatory targets**. *Nature* 2025, **642**(8068):796-804.

762 56. Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, Chen X, Taipale J, Hughes
763 TR, Weirauch MT: **The Human Transcription Factors**. *Cell* 2018, **172**(4):650-665.

764 57. Gremse M, Chang A, Schomburg I, Grote A, Scheer M, Ebeling C, Schomburg D: **The**
765 **BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms**
766 **for enzyme sources**. *Nucleic Acids Res* 2011, **39**(Database issue):D507-513.

767 58. Aguet F, Anand S, Ardlie KG, Gabriel S, Getz GA, Graubert A, Hadley K, Handsaker RE,
768 Huang KH, Kashin S *et al*: **The GTEx Consortium atlas of genetic regulatory effects**
769 **across human tissues**. *Science* 2020, **369**(6509):1318-1330.

770 59. Chipman HA, George EI, McCulloch RE: **BART: BAYESIAN ADDITIVE REGRESSION**
771 **TREES**. *The Annals of Applied Statistics* 2010, **4**(1):266-298.

772 60. Chen T, Guestrin C: **XGBoost: A Scalable Tree Boosting System**. In: *Proceedings of*
773 *the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data*
774 *Mining; San Francisco, California, USA*. Association for Computing Machinery 2016:
775 785–794.

61. Bergstra J, Bardenet R, Bengio Y, Kégl B: **Algorithms for hyper-parameter optimization**. *Advances in neural information processing systems* 2011, **24**.

62. Bergstra J, Yamins D, Cox D: **Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures**. In: *International conference on machine learning: 2013*. PMLR: 115-123.

63. Boyle EI, Weng S, Gollub J, Jin H, Botstein D, Cherry JM, Sherlock G: **GO:: TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes**. *Bioinformatics* 2004, **20**(18):3710-3715.

64. Sonawane AR, John Platig, Maud Fagny, Cho-Yi Chen, Joseph N Paulson, Camila M Lopes-Ramos, Dawn L DeMeo, John Quackenbush, Kimberly Glass, and Marieke Lydia Kuijjer: **Gene regulatory networks for 38 human tissues**. In. Zenodo; 2017.

65. Acharya S, Moghaddam VA, Jung WJ, Kang YS, Liao S, Province MA, Brent MR: **FISHNET: A Network-based Tool for Analyzing Gene-level P-values to Identify Significant Genes Missed by Standard Methods**. *bioRxiv* 2025:2025.2001. 2029.635546.

66. Tomasoni M, Gómez S, Crawford J, Zhang W, Choobdar S, Marbach D, Bergmann S: **MONET: a toolbox integrating top-performing methods for network modularization**. *Bioinformatics* 2020, **36**(12):3920-3921.