# Predicting bacterial fitness in Mycobacterium tuberculosis with transcriptional regulatory network-informed interpretable machine learning

**Ethan Bustad[1], Edson Petry[2‡], Oliver Gu[2‡], Braden T. Griebel[1,3], Tige R. Rustad, David R. Sherman[4], Jason H. Yang[2,5*], Shuyi Ma[1,3,6,7*]**

[1] Center for Global Infectious Disease Research, Seattle Children's Research Institute, Seattle WA, USA

[2] Center for Emerging and Re-emerging Pathogens, Rutgers New Jersey Medical School, Newark NJ, USA

[3] Department of Chemical Engineering, University of Washington, Seattle WA, USA

[4] Department of Microbiology, University of Washington, Seattle WA, USA

[5] Department of Microbiology, Biochemistry, & Molecular Genetics, Rutgers New Jersey Medical School, Newark NJ, USA

[6] Department of Pediatrics, University of Washington, Seattle WA, USA

[7] Pathobiology Graduate Program, Department of Global Health, University of Washington, Seattle WA, USA

[‡] These authors contributed equally.

*** Correspondence:**
Corresponding Authors
Shuyi Ma, shuyi.ma@seattlechildrens.org
Jason H. Yang, jason.y@rutgers.edu

**Keywords: *Mycobacterium tuberculosis*, transcriptional regulation, transcriptional regulatory network inference, network modeling, interpretable machine learning, growth regulation, stress adaptation. (Min.5-Max. 8)**

## Abstract

*Mycobacterium tuberculosis* (Mtb) is the causative agent of tuberculosis disease, the greatest source of global mortality by a bacterial pathogen. Mtb adapts and responds to diverse stresses such as antibiotics by inducing transcriptional stress-response regulatory programs. Understanding how and when these mycobacterial regulatory programs are activated could enable novel treatment strategies for potentiating the efficacy of new and existing drugs. Here we sought to define and analyze Mtb regulatory programs that modulate bacterial fitness. We assembled a large Mtb RNA expression compendium and applied these to infer a comprehensive Mtb transcriptional regulatory network and compute condition-specific transcription factor activity profiles. We utilized transcriptomic and functional genomics data to train an interpretable machine learning model that can predict Mtb fitness from transcription factor activity profiles. We demonstrated that this transcription factor activity-based model can successfully predict Mtb growth arrest and growth resumption under hypoxia and reaeration using only RNA-seq expression data as a starting point. These integrative network modeling and machine learning analyses thus enable the prediction of mycobacterial fitness under different environmental and genetic contexts. We envision these models can potentially inform the future design of prognostic assays and therapeutic intervention that can cripple Mtb growth and survival to cure tuberculosis disease.

## 1. Introduction

*Mycobacterium tuberculosis* (Mtb) remains a supremely successful pathogen, sickening 10.6 million people and killing over 1 million people worldwide each year [1]. An important factor for Mtb's success is its ability to adapt to a broad range of host-associated and treatment-associated stresses. The mechanisms underlying how Mtb dynamically regulates its growth and physiology in response to stress response remains incompletely understood. Characterizing the gene regulatory activities of transcription factors (TFs) under different environmental or stress conditions could help inform interventions that modulate Mtb growth and survival to cure tuberculosis disease.

Several groups have previously performed analyses to characterize Mtb's transcriptional regulatory network (TRN) using experimental and computational approaches [2; 3; 4; 5; 6; 7; 8; 9]. These efforts have largely relied on two strategies: 1) detailed profiling of the molecular impact of individual transcription factors (TFs) with recombinant induction and disruption strains, and/or 2) statistically informed TRN inference using data from large transcriptome compendia.

In principle, TRNs can be empirically assembled from measurements of TF-DNA binding activities and gene expression profiles from conditions with known individual TF perturbations. These data would enable the inference of direct regulatory interactions between TFs and their putative target genes, which exhibit altered transcriptional expression in response to TF perturbations and provide evidence of TF binding events proximal to a gene. To leverage this strategy, we previously engineered a library of Mtb recombinant TF induction (TFI) strains [2; 6], from which we profiled transcriptomes in 208 TFI strains by microarray analyses (GSE59086, [6; 10]) and detected ~16,000 ChIP-seq binding events for 154 TFs (~80% of all Mtb TFs) and 2,843 genes (~70% of all Mtb genes) [3; 10]. These detailed ChIP-seq and transcriptional profiles have yielded important insights into the regulatory programs active during Mtb broth culture. However, these experiments possessed several technical limitations. For example, our microarray profiling efforts were unable to measure changes in expression for 1,190 genes (~30% of Mtb genes) [6], and our ChIP-seq profiling efforts were unable to detect TF binding associated with 1,040 genes (~26% of Mtb genes) [3]. Moreover, the existing profiles have focused specifically on regulatory behavior of the Mtb laboratory strain H37Rv in log-phase growth in 7H9 media. Consequently, condition-specific interactions relevant to other environments or Mtb strains were not captured. Thus, despite such efforts, significant gaps remain in the ability to identify TF-gene regulatory interactions directly and comprehensively by only experimental activities.

Bioinformatic network inference approaches that utilize expression compendia comprising transcriptome responses under diverse biological conditions are a useful complementary strategy to recombinant strain profiling. These statistically informed approaches enable assessment of regulatory interactions across the multitude of conditions present in a transcriptome compendium. However, these computational network inference strategies are constrained by two limitations. First, large and biologically diverse gene expression data are needed to fuel identification of high-confidence statistical associations between TFs and putative target genes [11]. To meet this need, compendia of expression data may be curated from public microarray [4; 10] or RNA-seq [7; 12; 13] data. Second, statistical learning network inference algorithms differ in the assumptions made on the training data and on the interpretation of TF-gene associations. These assumptions are often biologically inaccurate. We previously performed such analyses and were able to only infer 598 clusters of coregulated gene expression for 3,922 genes [4]. Others recently performed similar analyses and inferred either 80 clusters for 3,906 genes [7] or 560 co-regulated gene modules for 3,912 genes [5]. These models have successfully revealed novel regulatory interactions impacting Mtb stress adaptation, but none of these regulatory models may be precisely interpreted as TF regulatory programs (as they only capture a fraction of Mtb's 214 TFs) and none can be used to directly estimate TF activities (i.e., the extent of regulation that each TF exerts on its regulated target genes, TFAs, [14]) under different experimental conditions. TRN inference efforts in other microbes, including the DREAM5 challenge for *E. coli* and *S. aureus* [15], have found that robust TRNs may be assembled by aggregating the regulatory relationships inferred by different statistical algorithms. We hypothesized that implementing a similar "wisdom of crowds" approach to aggregate complementary TRNs inferred via different statistical approaches would yield a more comprehensive and higher quality Mtb TRN.

Here we assembled a biologically diverse and batch corrected Mtb RNA-seq gene expression compendium. We integrated this RNA-seq compendium with the perturbative TFI microarray dataset to infer a comprehensive Mtb transcriptional regulatory network that included all 214 TFs and all 4,027 genes present in our RNA-seq expression compendium. We used this TRN to estimate TFA profiles corresponding to individual RNA expression profiles. We used the TFAs calculated from our RNA-seq compendium to train an interpretable machine learning regression model that could predict growth phenotypes previously measured in TF-induced strains [16]. We demonstrated that this regression model can accurately predict Mtb fitness under stressful environmental conditions such as hypoxia.

## 2. Methods

### 2.1 TFI microarray expression compendium assembly and normalization

Microarray expression data corresponding to TFI strains were downloaded from GEO (GSE59086). Groups were assigned to each sample by the identity of each strain. The Rv2160A gene fully encompasses the Rv2160c gene, so the Rv2160A and Rv2160c samples were combined into a single Rv2160 TFI strain group. This resulted in 208 TFI strain groups. These 208 strain groups included Rv0560, Rv3164c, and Rv3692 which were considered hypothetical TFs in TFI strain construction [6], but later determined to not be true Mtb TFs [10]. However, for the purpose of the analyses presented here, each of these 208 strains will be referred to as TFs. Smooth quantile normalization [17] was performed using *PySNAIL* [18] using the assigned group definitions.

### 2.2 RNA-seq expression compendium assembly, quality control, and normalization

The NCBI Sequence Read Archive (SRA) was queried with "*Mycobacterium tuberculosis*" for RNA expression samples containing raw FASTQ sequencing reads. 3,506 FASTQ sequencing reads were downloaded and combined with FASTQ sequencing reads from 398 unpublished RNA-seq profiles generated by our labs. We aligned these sequencing reads against the NC_000962.3 Mtb H37Rv reference genome using Bowtie 2 [19]. Read counts were compiled using *featureCounts* [20]. Samples with fewer than 400,000 total gene counts and samples duplicated in our preliminary compendium were excluded from further analysis. Sequencing counts between samples were normalized by transcripts per kilobase million (TPM). Group definitions were manually added to represent unique experimental conditions from each set of experiments; biological replicates for each experimental condition were given the same group definitions. Smooth quantile normalization [17] was performed using *PySNAIL* [18] using the assigned group definitions. Quality data, adapter and quality trimming statistics, and alignment and counts metrics were compiled and assessed using *MultiQC* [21].

### 2.3 UMAP visualization and cluster estimation

RNA expression compendia and TFAs were visualized by Uniform Manifold Approximation & Projection (UMAP) [22]. Clusters were estimated by *DBSCAN* [23]. The ε hyperparameter was optimized for each dataset by varying ε across 50 logarithmically distributed values from 0.1 to 10 and selecting the value of the elbow of the ε vs. Number of Outliers plot. This selection delivers the minimum number of clusters that maximizes inclusion of samples without overfitting the data (**Supplementary Figure S1**). UMAP and DBSCAN analyses were performed in Python using their implementations in *umap-learn* and *scikit-learn* [24].

### 2.4 Regulatory network inference methods

We implemented an ensemble of network inference methods by starting with a selection of methods featured in the DREAM5 challenge [15]. These methods were selected based on diversity in underlying statistical approach, predictive performance reported in the DREAM5 study, and the availability of a working implementation. Our initial selection consisted of ARACNe [25; 26], CLR [27], and GENIE3 [28]. We chose an ARACNe implementation that employs adaptive partitioning for more efficient processing [25; 26]. We used an R implementation of CLR available on CRAN from the *parmigene* package [29]. We used an R implementation of GENIE3 available on BioConductor [30]. To supplement these methods, we incorporated two other more recent advances in network inference approaches: cMonkey2 [31; 32] and iModulon [33]. We used a docker image containing a Python implementation of cMonkey2, available at https://hub.docker.com/r/weiju/cmonkey2. For iModulon, our desired output was different from the output of this algorithm implemented by the original authors. We thus made a custom implementation, borrowing heavily from https://github.com/SBRG/pymodulon and https://github.com/SBRG/iModulonMiner, in Python. In addition, we also chose to implement a regression strategy using Elastic Net regression, a more advanced technique than was used in DREAM5. Elastic Net is a regularization method that takes advantage of the unique properties of both the lasso (used extensively in DREAM5) and ridge regression [34]. Elastic Net performs better than lasso or ridge regression when predictors may be correlated and under-determined [35]. We modeled each gene individually on the expression of all the transcription factors, and used the resulting coefficients to both select significant relationships and score those relationships; this implementation was done in Python using *scikit-learn* [24]. Descriptions of each of these inference methods are provided in **Supplementary Table 3**.

Each method was wrapped to produce a ranked list of putative TF regulator-target gene relationships in order of the inferred strength of the regulatory relationship, from strongest to weakest. Execution was done using docker images (https://hub.docker.com/repositories/malabcgidr?search=network-inference). Auto-regulatory (self-targeting) relationships were excluded. Method hyperparameters were chosen to match either original publications or the DREAM5 challenge when possible. Execution for each method and optimization of their corresponding hyperparameters was validated by testing against the evaluation scripts provided in the supplemental material of [15; 32].

A network was generated for each combination of the two datasets (RNA-seq and TFI microarray) and 6 inference methods, yielding 12 total constituent networks.

### 2.5 Inferred network truncation and aggregation

The constituent networks were large, as many of the network inference methods did not require a cutoff threshold and did not perform multiple testing correction; the union of all inferred edges constituted over 90% of the possible Mtb regulatory space (where 100% would be every TF harboring a regulatory association with every Mtb gene). We therefore truncated each inferred network to incorporate the unique perspective of each model without aggregating too many low-confidence relationships. This was done by comparison with an independent validation set, comprising a presumed unbiased sampling of the true population of regulatory relationships in Mtb. This validation set was used to identify the extent of true positives in each network.

The validation data set was gleaned from Sanz et al., Material S1 [8]. The original list was filtered for relationships whose supporting evidence included at least one high-confidence physical methodology, namely values 4-9: LacZ-promoter fusion, GFP-promoter fusion, proteomic studies, electrophoretic mobility shift assays (EMSA), one hybrid reporter system, and chip-on-chip. This yielded a set of 433 high-confidence regulator-target relationships, including 51 regulators and 160 total target genes, that had little to no dependence on the transcriptional information used to build the constituent networks.

A cutoff threshold was chosen for each network by binning the ranks of validation hits into 32 bins and truncating the network at the first bin where the number of hits fell below the expected level of random overlap per bin. This level was calculated to equal the mean of a hypergeometric distribution, with a population size equal to the total regulatory space of Mtb, a set of true positive regulatory interactions identified by the Sanz validation set [8], and draws equal to the size of the inferred network, taken without replacement. This shrunk each network to an average of about 10% of its original size (3-28%) (**Supplementary Figure 2**). Three of the constituent networks displayed insufficient enrichment against the validation dataset: ARACNe/TFI, cMonkey2/TFI, and iModulon/TFI. Upon executing a Fisher's exact test to determine the chance of a random network achieving the same enrichment, these three failed to pass a strict cutoff of 0.0001. They were thus excluded from further aggregation.

The remaining truncated networks were then aggregated together, first into two combined networks, one for each underlying input transcriptome dataset (RNA-seq compendium and TFI microarray profile). Aggregation was performed by rank average as described in the DREAM5 challenge [15]. Repeating the enrichment analysis performed above, it was determined that the TFI aggregate would benefit from additional truncation and was thus truncated using the same threshold strategy described in the previous paragraph, whereas the RNA-seq network was already sufficiently enriched. These two networks were then aggregated together again by rank average, yielding one final aggregate network.

All these networks were validated against the Sanz et al. data set using the Matthews Correlation Coefficient (MCC), as described previously [36; 37] (**Supplementary Figure 3**).

### 2.6 Principal Component Analysis

Principal component analysis (PCA) was performed on the inferred networks (after truncation), the dataset-level aggregate networks, and the overall aggregate network, using the 16,792-

dimensional space represented by the ranks of edges shared across at least 3 of the inferred networks. Any relevant edges not included in a given network were assigned a rank of 16,792, the size of the space.

## 2.7 Regulatory directionality

The types of the regulatory connections (whether the TF up- or down-regulates the associated gene) were explored using a combination of the regression models and measured TFI gene expression values. Two elastic net models and two unpenalized linear models were used to infer direction of regulation based on the sign of the regression coefficients, one of each for each dataset (RNA-seq compendium and TFI microarray profile). We supplemented these regression associations with the directionality of significant differential gene expression (i.e. upregulated vs. downregulated expression) measured from the TFI microarray dataset. Linear models were fit in Python with the *statsmodels* package. Coefficients with an FDR < 0.05 were selected as evidence. Elastic net models with an $R^2$ of less than 0.8 were excluded; coefficients that were included by the remaining models were selected as evidence. TFI differential expression from the microarray dataset was filtered using an FDR < 0.05 and requiring at least 2-fold change in either direction. Elastic net models and TFI differential expression were considered strong evidence, whereas the unpenalized linear models were considered weak evidence. A flow chart depicting how the information from these models and differential expression analyses were used to define up vs. down regulation is shown in **Supplementary Figure 5**.

## 2.8 Comparing inferred networks against independent reference information

Additional orthogonal datasets were incorporated to corroborate the networks. All generated networks were tested against a set of published ChIP-seq binding relationships gleaned from Minch, et al. [3]. We took the intersection of their sets of statistically significant peaks (Supplementary Data 1 from [3]) and peaks in a canonical promoter region (Supplementary Data 3 from [3]) to yield 5,178 relationships, including 129 regulators and 2,271 total targets. The MCC was then calculated against this data set for each network.

Gene ontology enrichment analysis was then performed to ascertain the extent to which TF targeting could be used to gauge biological function within each group [38; 39]. For each TF, each set of genes that our network identified as upregulated, downregulated, or regulated in both directions by the regulator was analyzed for GO enrichment at an FDR < 0.05. All identified GO annotations that had a child annotation also identified for a given TF were removed for the sake of simplicity (**Supplementary Table 5C**). Results were filtered to regulators receiving at least 3 significant GO enrichments for further manual inspection and analysis (**Supplementary Tables 5A, 5B**), and those TFs with an annotated name and considered to have a testably specific functional role listed in the Mycobrowser annotation [40] were juxtaposed for network validation (**Table 1**). GO analysis was performed in Python using the *goatools* package [41]. Gene ontology data was taken from the 2024-06-17 release of go-basic.obo from the Gene Ontology knowledgebase [42] (https://purl.obolibrary.org/obo/go/releases/2024-06-17/go-basic.obo), and mappings to Mtb genes were taken from the European Bioinformatics Institute GOA project, release 20240805 (https://ftp.ebi.ac.uk/pub/databases/GO/goa/proteomes/30.M_tuberculosis_ATCC_25618.goa).

## 2.9 Calculating transcription factor activity profiles from network component analysis

Transcription factor activities for each expression profile was computed using Robust Network Component Analysis (ROBNCA) [43]. ROBNCA was implemented in Python, using code adapted from https://github.com/CovertLab/WholeCellEcoliRelease/tree/00cf7738cb8379c14d65ef632b2156bdf7c23434/reconstruction/ecoli/scripts/nca [44].

### 2.10 Associating network activity with bacterial fitness

We built a model associating mycobacterial growth with TF activity, as inferred from measured gene expression data. The GSE59086 microarray dataset was again used as a broad measure of TFI conditions, with relative growth data for 194 matching TFI conditions added from Ma et al., 2021, Table S1 as training data [16]. Expression levels in the form of log-2 fold-change were transformed into putative TFAs using the control strengths calculated via NCA from the aggregate network and RNA-seq compendium. A gradient boosted machine (GBM) model was trained to regress growth on TFAs, using a grid search cross-validation scheme to optimize hyperparameters based on bounds derived from [34], using the number of estimators to reward better performing models. The number of estimators was then optimized with a simple grid search. The model was implemented in Python using the *lightgbm* package [45; 46].

### 2.11 Hypoxia time-course experiment

Wildtype H37Rv (ATCC 27294) and H37Rv transformed with a control anhydrotetracycline (ATc)-inducible expression vector (H37Rv::pEXCF-empty, which does not induce recombinant gene expression) were cultured under in Middlebrook 7H9 with the oleic acid, bovine albumin, dextrose, and catalase (OADC) supplement (Difco) and with 0.05% Tween 80 at 37°C. H37Rv::pEXCF-empty was grown with the addition of 50 µg/ml hygromycin B to maintain the plasmid and induced with 100ng/mL ATc one day prior to onset of hypoxia. For hypoxia, strains were cultured in oxygen-limited conditions (1% aerobic $O_2$ tension) for 7 days, followed by reaeration on day 7-12, initiated by transferring cultures into continuously rolled bottles with 5:1 head space ratio using methods described previously [2; 47; 48; 49]. Bacterial survival and growth were enumerated by plating for colony forming units (CFU) on Middlebrook 7H10 solid media plates using standard microbiological methods.

Transcriptomes were generated by RNA-seq from bacterial cultures sampled from the aforementioned conditions using methods described previously [50]. Briefly, bacterial pellets suspended in TRIzol were transferred to a tube containing Lysing Matrix B (QBiogene) and vigorously shaken in a homogenizer. The mixture was centrifuged, and RNA was extracted from the supernatant with chloroform, followed by RNA precipitation by isopropanol and high-salt solution (0.8 M Na citrate, 1.2 M NaCl). Total RNA was purified using a RNeasy kit following the manufacturer's recommendations (Qiagen). rRNA was depleted from samples using the RiboZero rRNA removal (bacteria) magnetic kit (Illumina Inc., San Diego, CA). Illumina sequencing libraries were prepared from the resulting samples using the NEBNext Ultra RNA Library Prep kit for Illumina (New England Biolabs, Ipswich, MA) according to the manufacturer's instructions, and using the AMPure XP reagent (Agencourt Bioscience Corporation, Beverly, MA) for size selection and cleanup of adaptor-ligated DNA. We used the NEBNext Multiplex Oligos for Illumina (Dual Index Primers Set 1) to barcode the libraries to enable sample multiplexing per sequencing run. The prepared libraries were quantified using the Kapa quantitative PCR (qPCR) quantification kit and sequenced at the University of Washington Northwest Genomics Center with the Illumina NextSeq 500 High Output v2 kit (Illumina Inc., San Diego, CA). The sequencing run generated an average of 75 million base-pair paired-end raw read counts per library. Read alignment and gene expression estimation was carried out using a custom processing pipeline in R that harnesses the Bowtie 2 utilities [19; 51], which is publicly accessible at https://github.com/robertdouglasmorrison/DuffyTools, and https://github.com/robertdouglasmorrison/DuffyNGS.

Gene expression data were transformed from log-2 fold-change to putative TFAs using the control strengths calculated via NCA above and run through the GBM model to predict relative fitness level of the Mtb culture as it progressed through the hypoxia time-course.

### 2.12 False discovery rate correction

False discovery rate correction was performed using the two-stage Benjamini-Krieger-Yekutieli method [52].

## 3. Results

### 3.1 Generation of a large and biologically diverse Mtb gene expression compendium for TRN inference

Our previous attempts at TRN characterization utilized microarray expression profiles from recombinant TFI strains as perturbative training data (GSE59086, [6]). However, while this dataset enabled detailed characterization of transcriptional regulation of Mtb physiology during log-phase broth culture, it possessed poor biological diversity. UMAP and DBSCAN analyses reveal that expression profiles from these 698 microarray experiments and 208 TFI conditions only yielded 16 clusters of expression profiles (**Figure 1A**). This poor diversity likely arises from the original experimental design for these data, in which each TFI strain was grown to log-phase in albumin-dextrose-catalase (ADC)-supplemented 7H9 media before isolating RNA. UMAP and DBSCAN analyses suggested that this TFI microarray dataset alone would be insufficient for predicting TFAs corresponding to diverse experimental conditions. Moreover, microarray technologies have poor sensitivity and dynamic range for quantifying gene expression [53]. We found that 101 genes in this dataset did not possess expression measurements greater than 10 counts, indicating poor detection or poor evidence for expression in these experiments (**Figure 1B**). In addition, the median absolute deviation (MAD) was small (< 1) for nearly all genes, indicating the ability to detect gene expression changes across conditions was limited. These analyses collectively motivated the need to assemble a new RNA expression compendium.

We therefore collected samples from the NCBI Sequence Read Archive (SRA) and our own labs, aligned, filtered, normalized, and batch corrected by smooth quantile normalization [17; 18] (see **Methods** for details). Batch correction is an important pre-processing step for unifying data from different sources that is frequently overlooked in Mtb RNA expression compendium analyses [4; 7; 12; 13]. After performing these pre-processing steps, our final compendium comprised 3,496 RNA-seq samples from 1,288 experimental conditions (**Supplementary Table 1**). Expression counts for the RNA-seq compendium can be queried at https://tfnetwork.streamlit.app/.

UMAP and DBSCAN analyses of the batch corrected RNA-seq expression compendium validated its biological diversity (**Figure 1C-D, Supplementary Table 2**). We identified 142 unique expression clusters. This RNA-seq transcriptome compendium exhibited significantly greater dynamic range and variation in gene expression than in the TFI microarray dataset (**Figure 1D**). Of note, genes with high variation (high MAD) were mostly well-characterized stress response genes (e.g., Rv2031c *(hspX)*, Rv2626c *(hrp1)*, and Rv2623 *(TB31.7)*), with Rv2007c *(fdxA)* having higher variation than the commonly studied Rv3133c *(devR)* stress response regulator. These are consistent with expectation, as most stress response genes would be expected to only be induced in the presence of their specific stressor.

## 3.2 Inferred transcriptional regulatory network interactions enrich for shared functional processes

Network inference studies in other bacteria have shown that combining regulatory interactions from multiple different inference algorithms results in a TRN that outperform networks generated by a single method [15]. To more comprehensively characterize Mtb regulatory interactions, we applied a "wisdom of crowds" ensemble inference approach. We first applied a collection of regulatory network inference tools to generate TRN models using individual methods (see **Methods**). These tools were selected because they have been shown to be sensitive to distinct types of regulatory relationships in other bacteria [15] or they have previously been successfully applied to infer regulatory relationships in Mtb [4; 5; 7]. To further diversify the regulatory relationships inferred from these approaches, we applied these tools to both our assembled RNA-seq compendium as well as the TFI microarray dataset. Collectively, these inference activities yielded 12 networks that describe 779,213 unique interactions between 214 regulators and 4,029 target genes. We truncated these networks using a benchmark dataset of high confidence regulatory interactions with biochemical evidence that was curated by Sanz et al. [8] (see **Methods**). We used this high confidence regulatory interaction dataset to inform pruning of low-confidence regulatory relationships inferred from each of the individual inference methods (**Supplementary Figure 2**), yielding a shorter, more high-confidence network for each method. Principal component analysis of these networks revealed substantial diversity in the regulatory interactions identified between the different approaches applied to the two source datasets (**Figure 2B**).

We rank-aggregated the resulting 12 networks to consolidate regulatory relationships across the individual inference methods. The resulting aggregate network has 68,226 regulatory interactions that connect 214 transcriptional regulators with 4,027 target genes. Of these interactions, 37,236 are associated with transcriptional activation across conditions, 15,820 interactions are associated with transcriptional repression across conditions, 1,496 relationships are predicted to be either activating or repressing, depending on the environmental condition, and 11,766 regulatory relationships have an undetermined regulatory directionality (**Supplementary Table 4**). These interactions represent both direct, biophysical regulatory events as well as indirect regulatory relationships mediated by downstream regulators. These interactions also represent the union of regulatory relationships that are active in at least a subset of all the different environmental conditions profiled in our assembled source RNA-seq compendium and TF induction profiling datasets. Notably, not all these regulatory relationships will be active under all environmental conditions. The distribution of regulatory interactions per TF largely follows a power law distribution consistent with the scale free networks found to represent transcriptional regulation in other bacteria (**Supplementary Figure 4**). We found a deviation between the distribution of our aggregate network and the expected power law distribution for regulators with relatively few target genes. This is likely due to the inclusion of indirect regulatory relationships and relationships that are active under some but not all environmental conditions. The networks can be viewed at https://tfnetwork.streamlit.app/, and the TF-gene interactions are described in **Supplementary Table 4**.

To validate the connectivity of our aggregate network, we benchmarked it against experimentally profiled TF binding data we previously profiled by ChIP-seq in the TFI strains under log-phase broth culture [3]. To assemble a high-confidence regulatory association dataset, we included only significant ChIP-seq peaks associated with TF binding in the promoter region of target genes. We evaluated overlap between this high-confidence ChIP-seq regulatory interaction dataset and our inferred regulatory networks with the Matthews correlation coefficient (MCC). We find that most of the inferred networks that we generated had significant MCCs, and that the aggregate network outperforms the majority of inferred networks using individual methods (**Supplementary Figure 3**), whilst still retaining a large number of regulatory relationships (most of the better performing individual inference networks have relatively few regulatory interactions).

We also assessed the extent to which the regulatory relationships captured by our aggregate network preserved biological functional relationships between the regulating TFs and the target genes. For TFs with clear literature characterization of its function, we found a high degree of correspondence with the gene ontologies and annotated functions of its regulated target genes (**Table 1, Supplementary Table 5**). For example, Rv3574 (*kstR*) is a TF that has been linked to regulating cholesterol metabolism [54], and the target genes associated with *kstR* in our aggregate network also have gene ontology annotations linked to cholesterol metabolism (**Table 1**). Additionally, toxin-antitoxin target genes were enriched for growth regulation, highlighting that the regulatory relationships captured by the aggregate network include indirect regulatory relations. Collectively, this suggests the significant ontology and functional annotation enrichments made for genes and TFs that are currently poorly annotated represent testable hypotheses for function – this is one of the major advances from the aggregate network.

### 3.3 Network component analyses reveal per-sample Mtb TF activities under different conditions

Understanding when TFs are actively exerting their regulatory influence on their target genes can reveal mechanistic insights into bacterial physiology and stress response. Network component analysis (NCA) is an efficient way of estimating these TFA profiles from expression data by using a TRN to perform matrix decomposition [14]. Robust NCA (ROBNCA) is a variant of NCA that improves the performance of NCA calculations on noisy data with outlier measurements [43]. We applied ROBNCA to estimate TFAs corresponding to each sample in our TFI microarray and RNA-seq compendium.

To first determine and validate the ROBNCA TFA estimation approach on our data, we performed ROBNCA on the TFI microarray data using the aggregate network inferred only from the TFI data, as well as on 10 randomized networks to be used as negative controls. We hypothesized that if the estimated TFAs represent true TF activities, with high TFAs indicating strong net activator activity and low TFAs indicating strong net repressor activity, then the percentile ranks of TFAs for highly expressed TFs should be either very high or very low in their corresponding TFI strains. On the other hand, if the ROBNCA-calculated TFAs were spurious, then the TFA percentile ranks should be statistically indistinguishable from the TFA percentile ranks from randomized networks.

For each of the 208 TFI strains within the microarray expression dataset, we averaged the TFAs for all TFs across their biological replicates. We rank ordered TFs by their activities for each TFI strain, calculated the rank percentile activity of the induced TF for each TFI strain, and analyzed the distribution of these percentiles (**Figure 3A**). For the TFI microarray network, 31 TFs were ranked in the highest or lowest 15% of TFA ranks (greater than 1 standard deviation from the mean), implying that these TFs were the dominant regulators active in their respective TFI strain profiling condition. Interestingly, 91 TFs had TFAs in the middle 30% from 35-65%. These TFs were fairly uniformly distributed suggesting their related transcriptional programs were likely cross-regulated by other TFs.

Importantly, this suggested that induction of TF expression alone may be insufficient for fully inducing some transcriptional programs, thus supporting the use of TFAs over untransformed gene expression for downstream analysis.

We performed similar calculations for each of the randomized networks **(Supplementary Figure 6**) and averaged the TFA rank percentiles for all TFs from each randomized network (**Figure 3B**). We found that there were significantly fewer TFAs in the highest or lowest 15% of TFA ranks in these randomized networks than the TFAs calculated from the TFI expression dataset (p = 1.66e-49, z-test [55]). Similarly, there were significantly more TFAs in the middle 30% (p = 1.66e-49, z-test [55]). These differences between the ROBNCA-calculated TFA percentile distributions between TFI and randomized networks indicated that the TFAs estimated by ROBNCA were not spurious and likely reported on true biological condition-specific activities.

We next applied ROBNCA to our RNA-seq compendium using the TRN inferred from the RNA-seq compendium. UMAP and DBSCAN analyses revealed that the level of biological diversity of ROBNCA-predicted TFAs was similar to the diversity within the expression compendium, with 112 clusters of TFAs across the 3,496 samples (versus 142 for untransformed expression; **Figure 3C**). Amongst the TFs with the highest level of median activity were the essential nitric oxide-sensing Rv3219 (*whiB1*), histone-like protein Rv2986c (*hupB*), and sigma factor Rv2703 (*sigA*) (**Figure 3D**). Each of these would be expected to be constitutively active in live Mtb cells. Also consistent with expectation, the well-characterized stress response regulators Rv3133c (*devR*), Rv1994c (*cmtR*), Rv0827c (*kmtR*) and two-component system regulators Rv0602c (*tcrA*) and Rv0981 (*mprA*) were amongst the TFs with the highest TFA MAD.

Interestingly, the distribution of TFAs appeared different from the distribution of TF expression levels measured for each RNA-seq sample across the compendium **(Figure 3E)**. We tested the correlation of expression level vs. activity for each TF across the entire compendium and found that expression and activity were only moderately correlated across the dataset (Pearson's r = 0.48 ± 0.16 median ± MAD) **(Figure 3F)**. 31 TFs were strongly correlated (|Pearson's r| ≥ 0.7), 66 TFs were moderately correlated (0.7 > |r| ≥ 0.5), and 61 TFs were weakly correlated (0.5 > |r| ≥ 0.3). Relatedly, both median and MAD expression and activity were only weakly correlated across all TFs (median: r = 0.43; MAD: r = 0.32). These analyses further support our observation that TF expression level is not the sole determinant for TFAs for most TFs. Rather, expression and activity convey two distinct but complementary insights into transcriptional regulation, highlighting the importance of accounting for network interactions when investigating transcriptional regulation. In particular, we posit that TFs with weak correlation between expression and activity may require allosteric or other post-translational modification to trigger activation of transcriptional regulation. This hypothesis can be tested in future studies.

## 3.4 Transcription factor activity profiles can predict condition-specific bacterial fitness

Because transcriptional regulation plays important roles in coordinating Mtb growth adaptations under stress, we asked whether our regulatory network models could be used to predict fitness consequences of TF regulatory activities. To test this hypothesis, we utilized gradient boosting machine learning to construct an interpretable TFA regression model designed to predict the fitness of each TFI strain during log-phase culture based on each strain's calculated TFA profiles. We trained this model using the TFAs computed by ROBNCA from the RNA-seq compendium, paired with TFI fitness measurements that we previously collected in a Transcriptional Regulator Induced

Phenotype (TRIP) screen [16]. This TFA–fitness regression model was able to explain 87% of the observed variation of growth between the TFI strains in the TRIP screen (**Supplementary Figure 7**).

To determine if this TFA–fitness regression model could predict changes in Mtb fitness or growth from new data that were not used to train the model (e.g., under differing experimental conditions), we generated fitness predictions with our model using transcriptomes that we profiled from Mtb cells undergoing hypoxia and reaeration stress. From the TFA profiles calculated for cells exposed to hypoxia, the TFA–fitness regression model predicted a significant decrease in growth that persisted for each of the timepoints profiled under hypoxia (**Figure 4A, Supplementary Figure 8**). From the TFA profiles calculated for cells under reaeration, the model predicted a rebound in Mtb growth comparable to growth levels experimentally measured during log-phase culture. The kinetics of the shifts in growth predicted by the TFA–fitness regression model aligned well with the experimental measurements of Mtb bacteriostasis in hypoxia, followed by growth during reaeration (**Figure 4A, Supplementary Figure 8**). Importantly, the experimental growth data from the hypoxia-reaeration time course aligned better with the predictions from the TFA regression model than from an analogous regression model trained from TF expression data alone (**Supplementary Figure 10**). These results further support our premise that TFAs more effectively capture condition-specific transcriptional regulation than TF expression alone and implies that the activation and regulation of transcriptional programs under hypoxia and reaeration may involve allosteric or other post-transcriptional mechanisms.

Because the TFA–fitness regression model is openly interpretable, we examined which TFAs most strongly predicted the fitness changes under hypoxia and reaeration. We found that our TFA–fitness regression model predicts that growth restriction during hypoxia is primarily driven by the activities of 7 TFs whose TFA profiles changed significantly during hypoxia (**Figure 4B**). Importantly, each of these TFs have direct or indirect links to hypoxia in the literature (**Supplementary Figure 9**, **Supplementary Table 7**), thus further validating these model predictions and the use of TFAs as a lens into condition-specific stress response biology.

## 4. Discussion

Understanding the molecular drivers of phenotypic changes in an organism is a fundamental goal of biological research. In this study, we applied machine learning approaches to construct an interpretable TFA–fitness regression model that can utilize Mtb TRNs to predict experimentally measured changes in Mtb growth state in diverse environmental conditions. Our models build upon existing experimental profiling and network inference modeling efforts to characterize Mtb transcriptional regulation by integrating the data and algorithms developed in these prior studies [2; 3; 4; 5; 6; 7; 14; 15; 43]. Moreover, by integrating Mtb fitness profiling data from TRIP, our models have also enabled direct prediction of growth/survival phenotypic outcomes from condition-specific gene expression data inputs.

Our “wisdom of crowds” approach for inferring transcriptional regulatory interactions yielded significant enrichment of known regulatory relationships while also expanding the scope of represented experimental conditions. Our resulting TRN is substantially larger than the networks inferred by individual algorithms, while enriched for experimentally validated interactions. This highlights the utility of ensemble inference algorithms, as has been previously shown for regulatory network inference in other bacteria [15].

Importantly, our results demonstrate how network models can generate hypotheses on gene function in at least two complementary ways. First, we show by gene ontology enrichment analysis that there is significant correlation between the annotated function of a TF's target genes and the condition-specific regulatory function of the TF. It is important to note that the regulatory interactions identified by our aggregate TRN includes both direct regulatory interactions involving physical interactions between a TF and its target gene as well as indirect associations mediated by other factors. Both direct and indirect regulatory associations are important for coordinating changes in bacterial physiology [56], so it is expected that both types of interactions share annotated ontologies. Because ~25% of Mtb genes lack functional annotation [57], we think the regulatory relationships identified in our TRN can aid basic microbiological efforts in investigating Mtb gene function by generating hypotheses for the functions of these poorly characterized or unknown genes (**Supplementary Table 5)**.

Second, we show that TFA regression models can be trained to link condition-specific TFAs with TF fitness in log-phase broth culture to predict Mtb fitness under stress. Notably, we show that our TFA regression model was able to predict Mtb growth and bacteriostasis under hypoxia and reaeration – environmental conditions not used in training the TFA regression model. Our results biologically suggest that TFAs are a useful determinant of condition-specific changes in bacterial growth, and that the estimated TFA is more predictive of growth phenotypes that TF expression alone. This is consistent with expectation as Mtb uses transcriptional regulation to orchestrate behavioral adaptations to varying environments, including in growth phenotypes. Our modeling also enables inspection of which TFAs are driving the predicted bacterial fitness outcomes. This can inform the generation of hypotheses on the mechanisms underlying how TFs and their corresponding transcriptional programs are activated (e.g., via allosteric mechanisms and/or network interactions). Our TRN and TFA–fitness models could potentially inform the identification of regulatory mechanisms mediating Mtb response and adaptation to other clinically relevant stress conditions where gene expression profiling data are available. The TFs and target genes highlighted by these models may potentially represent future intervention targets aimed at modulating Mtb fitness in a therapeutically beneficial way. In light of the growing crisis of antimicrobial resistance [58] and multi- and extensively-drug-resistant tuberculosis [59], we think our approach will be important for curing tuberculosis disease [60].

More broadly, our work here demonstrates how network models can be utilized for biologically meaningful interpretable machine learning applications. A fundamental challenge in current machine learning activities is the difficulty in understanding how a trained machine learning model makes predictions [61; 62]. We previously demonstrated that machine learning regression models can be used to elucidate metabolic mechanisms underlying antibiotic lethality in *E. coli* [63], as well as to predict multidrug interaction outcomes in Mtb [50]. Our study here analogously extends this approach by training a regression model on TFAs estimated from TRN analyses to predict changes in Mtb growth state. The advantage of this strategy over other contemporary machine learning approaches is the direct utilization of prior knowledge encompassed by biological network models, which directly enable the generation of hypotheses for mechanisms linking network interactions to cell phenotypes. These hypotheses can then be experimentally tested [50; 63] and used as the basis for further mechanistic study [64] and investigation of translational potential.

Looking forward, we envision that this approach and our TFA regression model can be useful for several facets of tuberculosis research. We demonstrated that our model can be used to predict changes in Mtb growth state under environmental stress, which may inform the design of growth state assays under conditions where standard microbiological tools are not feasible. There is increasing appreciation that Mtb drug susceptibility is regulated by its environment [65; 66]. Our TFA–fitness regression model can be used to elucidate the molecular mechanisms underlying these phenotypes. Moreover, functional genetic datasets are becoming increasingly available using different technologies [16; 67; 68; 69; 70; 71; 72]. These data can be applied to train next-generation TFA–fitness regression models with improved predictive power. Finally, detailed characterizations of Mtb clinical strains are now providing significant insights into the how mutations or other forms of genomic diversity regulate drug susceptibility in human patients [72; 73; 74; 75]. We envision the TRN and TFA-fitness regression framework established here can be extended not only to study the mechanistic basis for differences between drug susceptibility amongst clinical isolates, but also to anticipate the drug susceptibility of new clinical isolates as they become curated.

## Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Author Contributions

E.B.: Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review, editing; E.P.: Data curation, Formal analysis, Visualization, Writing – review, editing; O.G.: Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – review, editing; B.T.G.: Data curation, Software, Visualization, Writing – review, editing; T.R.R.: Investigation, Resources, Methodology, Writing – review, editing; D.R.S.: Investigation, Resources, Methodology, Funding acquisition, Supervision, Writing – review, editing; J.H.Y.: Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Visualization, Validation, Formal Analysis, Writing – original draft, Writing – review, editing; S.M.: Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Visualization, Formal Analysis, Writing – original draft, Writing – review, editing.

## Funding

This work was supported by the Agilent Early Career Professor Award and National Institutes of Health grants R00-GM118907, U19-AI11276 (J.H.Y.); U19-AI62598 and R01-AI146194, (J.H.Y., D.R.S. and S.M.); R01-AI150826 and U19-AI135976 to (D.R.S. and S.M.); and DP2-AI164249 (S.M.).

## Acknowledgments

We thank Research Scientific Computing at Seattle Children's Research Institute for providing HPC resources that have contributed to this investigation. We thank Avi Shah for helpful discussions, and Jessica Assadi and Robert Morrison for technical assistance.

## Figures

**Figure 1**: A biologically diverse Mtb RNA expression compendium. (A) UMAP visualization of biological diversity in the TFI microarray data. TFI data were batch corrected by smooth quantile normalization before computing the UMAP. Density-based spatial clustering (DBSCAN) was performed on the UMAP to identify clusters of samples with similar gene expression. UMAP and DBSCAN analyses revealed 16 total expression clusters in the TFI dataset. (B) Median vs. median absolute deviation (MAD) plot of expression for each gene across the TFI dataset. Each point represents a gene. Median expression and MAD were calculated for each gene across the 698 samples. Colors reveal point density (yellow: high density, blue: low density). (C) UMAP visualization of samples from the normalized and batch corrected RNA-seq compendium determined by gene expression. UMAP and DBSCAN analyses reveal 142 clusters of samples with similar gene expression. (D) Median vs MAD plot of expression for each gene across the RNA-seq compendium.

**Figure 2**: Overview of aggregate network. **(A)** PCA was performed on each of the generated networks. The networks inferred from the RNASeq compendium (triangle symbols) cluster to the right, whereas the networks inferred from the recombinant TF induction transcriptomes (x symbols) fall to the left. The dataset-level aggregates each cluster loosely with the same-dataset constituent networks at the horizontal extremes, whereas the overall aggregate falls near the centroid of all networks. **(B)** Performance of each inferred and aggregate network, calculated against a set of TF–target gene relationships defined by a ChIP-Seq DNA-binding investigation of recombinant TFI strains [3], as measured by Matthews correlation coefficient (MCC). MCC quantifies the level of correlation between the two sets, with higher values indicating more correspondence. Blue bars depict the MCC for aggregate networks; the other colors depict the MCC for the individual inferred networks. Hatched bars indicate networks that were excluded from aggregation. The horizontal dashed line represents the 95th percentile MCC performance of 1000 randomly generated networks. Note that the excluded iModulon/TF induction network scores relatively highly by this metric, likely because of its size (~7k edges, versus an average of ~180k). See Methods for information about the exclusion criteria.

**Figure 3**: Compendium-wide transcription factor activities. (A) Distribution of the TFA rank percentiles for each induced TF in each strain from the TFI microarray dataset. ROBNCA was applied to the TFI microarray dataset using the network specifically inferred from the TFI dataset. For each sample, rank percentiles were computed for each TFA. TFAs were averaged across biological replicates for each TFI strain. Histogram depicts the percentile rank for TFAs corresponding to the over-expressed gene in each TFI strain. (B) Averaged distribution of TFA percentile ranks from ROBNCA using 10 randomized networks (**Supplementary Figure 6**). (C) UMAP visualization of samples from the normalized and batch corrected RNA-seq compendium as determined by TFA. UMAP and DBSCAN analyses reveal 112 clusters of samples with similar TFAs. (D) Median vs. MAD plot of activity for each TF across the RNA-seq compendium. (E) Median vs. MAD plot of expression for each TF across the RNA-seq compendium. (F) Distribution of Pearson's correlation coefficients between expression and activity for each TF across the RNA-seq compendium.

**Figure 4**: Machine learning model insights into Mtb growth through a hypoxic time-course. **(A)** *Top*: When Mtb grown for two days in log phase was subjected to hypoxic conditions (starting from day 0), the bacteria stopped growing for the duration of the imposed hypoxia, as indicated by the stable CFU between day 0 and day 7. When the culture was reintroduced to oxygen ("Reaeration", starting from day 7), the bacteria resumed growth, as indicated by significantly higher CFU after day 8.

*Bottom*: Our GBM model predicted a decrease in growth over the course of the period of hypoxia, and an increase in growth again upon reaeration, based only on transcriptional data measured over the course of the experiment. Each point represents an RNA-seq timepoint. **(B)** The GBM model can be interrogated to determine the primary drivers of the phenotype it predicts; when comparing the most impactful TFAs in hypoxic conditions (days 2-7) versus those in reliably reaerated conditions (days 9-12), 7 TFs were predicted to be particularly influential to the reduced growth in hypoxia versus reaeration, each contributing at least 5% of the total absolute impact predicted by the model. Shown here is the mean TFA change for each of the impactful TFs across days 2-7; other TFAs show no net activity change overall (see Methods for details on TFA change calculation).

## Tables

Table 1. Network regulators: annotation versus gene set enrichment analysis of inferred regulon.

| Regulator | Name | Mycobrowser gene product and function information | Inferred Regulon GO Annots. (FDR <0.05) | |
|---|---|---|---|---|
| | | | # | Summary |
| **Rv0353** | hspR | Probable MerR family heat shock protein transcriptional repressor. Involved in repression of heat shock proteins. Binds to three inverted repeats in the promoter region of the DnaK operon. Induced by heat shock. | 3 | heat response |
| **Rv1657** | argR | Probable arginine repressor (AHRC). Regulates arginine biosynthesis genes. | 4 | cobalamin synthesis; UMP synthesis; C-N bond formation |
| **Rv2215** | dlaT | Dihydrolipoamide acyltransferase, component of pyruvate dehydrogenase. Involved in TCA cycle; converts pyruvate to acetyl-CoA and $CO_2$. Also involved in defense against oxidative stress. | 51 | TCA cycle, respiration, downregulation of virulence factors |
| **Rv2359** | zur | Probable zinc uptake regulation protein. Acts as a global negative controlling element, with $Zn^{2+}$ binds operator of repressed genes. | 8 | downregulating translation, iron import |
| **Rv2374c** | hrcA | Probable heat shock protein transcriptional repressor. Involved in repression of class I heat shock proteins. Prevents heat-shock induction of these operons. | 17 | transcription and translation |
| **Rv2610c** | pimA | Alpha-mannosyltransferase. Involved in the first mannosylation step in phosphatidylinositol mannoside biosynthesis (transfer of mannose residues onto PI). | 64 | amino acid and nucleobase synth., respiration, growth/proliferation |
| **Rv2720** | lexA | Repressor. Represses genes involved in nucleotide excision repair and SOS response. Binds 14-bp palindromic sequence. | 10 | DNA binding, repair, cleavage |
| **Rv3301c** | phoY1 | Probable transcriptional regulatory protein PhoU-homolog 1. Involved in regulation of active transport of inorganic phosphate across the membrane. | 18 | ETC, oxidative phosphorylation |
| **Rv3417c** | groEL1 | 60 kDa chaperonin 1 (protein CPN60-1). Prevents misfolding, promotes refolding and proper assembly of unfolded polypeptides generated under stress conditions. | 15 | stress response |
| **Rv3574** | kstR | Transcriptional regulatory protein (probably TetR-family). Involved in transcriptional mechanism. Predicted to control regulon involved in lipid metabolism. | 22 | cholesterol, lipid, and carbon metabolism |
| **Rv0599c** | vapB27 | Possible antitoxin. | 13 | growth regulation, toxin sequestration, RNase |
| **Rv0608** | vapB28 | Possible antitoxin. | 12 | growth regulation, toxin sequestration, RNase |
| **Rv0623** | vapB30 | Possible antitoxin. | 12 | growth regulation, toxin sequestration, RNase |

| | | | | |
|---|---|---|---|---|
| **Rv1560** | vapB11 | Possible antitoxin. | 6 | growth regulation |
| **Rv1740** | vapB34 | Possible antitoxin. | 7 | growth regulation |
| **Rv1960c** | parD1 | Possible antitoxin. | 18 | growth regulation, toxin sequestration, RNase |
| **Rv2009** | vapB15 | Antitoxin. | 13 | growth regulation, RNase |
| **Rv2595** | vapB40 | Possible antitoxin. | 8 | growth regulation, toxin sequestration |
| **Rv2760c** | vapB42 | Possible antitoxin. | 4 | growth regulation, DNA repair |

## Supplementary Material

1 **Supplementary Figure 1 UMAP.** Hyperparameter optimization was performed on UMAPs from the (A) TFI microarray compendium, (B) RNA-seq compendium, or (C) TFAs calculated from the RNA-seq compendium. ε was varied from 0.1 to 10 on a logarithmic scale and numbers of clusters (left), numbers of outliers (center), and maximum cluster size (right) were computed for each ε. ε was selected from the elbow of the outliers plot (ε = 0.281 for TFI data, 0.309 for RNA-seq compendium and estimated TFAs).

**2** **Supplementary Figure 2 Inferred network validation.** Distribution of the ranks, in each network, of edges shared with the validation dataset from Sanz et al., 2011, [8] from each network. Each histogram is divided into 32 bins. Horizontal dashed lines represent the expected number of random matches between each network and the validation dataset. Truncation was performed on these networks at the first bin where the count dropped below the dashed line (see **Methods**). Panels with hashed backgrounds (B, F, and L) represent networks that were excluded from the aggregation due to insufficient enrichment.

3 **Supplementary Figure 3. Inferred network performance.** Performance of each inferred and aggregate network, calculated against a set of TF–target gene relationships identified by Sanz et al., 2011 [8] (see **Methods**), as measured by Matthews correlation coefficient (MCC). MCC quantifies the level of correlation between the two independent sets of relationships. Higher values indicate greater correlation. The blue bars depict the MCC for the dataset-level and overall aggregates. Other colors are used to depict the MCC for the individually inferred networks. Hatched bars indicate the networks that were excluded from aggregation. The horizontal dashed line represents the 95th percentile MCC performance of 1,000 randomly generated networks. See Methods for exclusion criteria.

4 **Supplementary Figure 4 TRN properties.** Out-degree distribution of TF-gene interactions (edges) from the overall aggregate network. This distribution significantly differs from a power law distribution on the left side of the plot, likely because the network includes indirect interactions. These will deflate counts of low-degree TFs (nodes) and inflate counts of higher-degree nodes.

5 **Supplementary Figure 5 Assignment of activating vs repressing regulatory interactions.** Flow chart depicting the logic used to assign directionality to regulatory relationships. Abbreviations used are defined in the legend in the bottom left.

6 **Supplementary Figure 6 TFA rank percentiles for randomized networks**. TRNs were randomized 10 times. For each random network, ROBNCA was used to compute TFAs for the TFI dataset. Rank percentiles were assigned to each TFA for each TFI microarray profile and averaged across replicates for each TFI strain. Plotted are TFA rank percentile distributions for all over-expressed TFs corresponding to their respective TFI strain from each randomized network.

7 **Supplementary Figure 7. TFA-fitness regression model performance.** (A) Fitness values predicted by the gradient boosted machine (GBM) model versus the experimentally measured values supplied to the model upon training. The line of best fit depicts the relationship between predicted and measured values. The slope of this line is slightly less than 1, indicating that the regression model modestly underestimates relative fitness changes. The model achieved a

coefficient of determination ($R^2$) of 0.87 against its training set, indicating that the model can explain 87% of the variation in fitness from the TRIP screen. (B) Residuals of the model predictions versus measured values form a roughly normal distribution, indicating a lack of bias and overall reliable predictive ability.

8 **Supplementary Figure 8 TFA hypoxia prediction.** Our TFA-fitness regression model predicted a decrease in growth over the course of the period of hypoxia, and an increase in growth again upon reaeration, based only on transcriptional data measured over the course of the experiment (each point represents an RNA-seq timepoint), in both the empty plasmid strain (blue) and wild-type H37Rv (orange).

**Supplementary Figure 9 Hypoxia-responsive TFAs.** The TFA-fitness regression model can be interrogated to determine drivers of hypoxia by comparing the most impactful TFAs under hypoxia (days 2-7) versus reaeration (days 9-12). 7 TFs were most important for predicting reduced growth under hypoxia versus reaeration. Each contributes at least 5% to total model predictions. Depicted is the mean change in TFA for each of the impactful TFs across days 2-7 (orange) versus days 9-12 (cyan). Other TFAs show negligible changes in activity across hypoxia or (see Methods for details on calculations for changes in TFA).

9 **Supplementary Figure 10 TF expression hypoxia prediction.** Hypoxia and reaeration fitness changes predicted by a GBM model trained using only TF expression data instead of TFAs.

10 **Supplementary Table 1 Expression data from the TFI microarray dataset.** Batch correction group assignments for each sample in the TFI microarray dataset. Smooth quantile normalized and microarray expression for all genes and all samples in the TFI microarray dataset. Median and MAD expression for each gene. Group assignments were used by the PySNAIL smooth quantile normalization algorithm for batch correction [18].

11 **Supplementary Table 2 Expression data from the RNA-seq expression compendium.** Batch correction group assignments for each sample in the RNA-seq compendium. Group assignments were used by the PySNAIL smooth quantile normalization algorithm for batch correction [18]. Median and MAD expression for each gene.

12 **Supplementary Table 3 Network inference methods.** Description of transcriptional regulatory network inference methods.

13 **Supplementary Table 4. Aggregate network directionality of regulation.** Summary of the assignments of activating (up) vs. repressing (down) regulatory interactions for all TF-gene regulatory interactions in the aggregate transcriptional regulatory network (TRN).

14 **Supplementary Table 5 TF Gene Ontology assignments.** GO enrichment for each transcriptional program regulated by each TF inferred by our aggregate TRN. (A) Annotated functions and a summary of GO enrichments found for targets from selected TFs. All TFs with at least 3 significant GO enrichment terms and a non-locus gene name in Mycobrowser [40]. 45 TFs meet these criteria. These data validate the accuracy of our network, as one would expect an accurate regulatory network to have target sets significantly enriched for the known functions of each TF. (B) Remaining TFs with at least 3 significant GO enrichments assigned by our analysis but without an annotated gene name (36 additional TFs). These data represent

predictions for potentially novel TF functions. (C) All GO enrichments identified by our analysis were corrected for FDR with a cutoff of 0.05.

15 **Supplementary Table 6 Transcription Factor Activities.** Median and MAD expression and activity for each TF in the RNA-seq compendium. Pearson correlation coefficient between TF expression and TFA for each TF across all samples in the RNA-seq compendium.

16 **Supplementary Table 7** Overview of the top 7 most important TFAs for predicting fitness under hypoxia as identified by our TFA regression model, validated by published evidence for mechanistic activation under hypoxia [3; 6; 76; 77; 78; 79; 80; 81; 82; 83; 84; 85; 86].

## Data Availability Statement

The transcriptome datasets analyzed for this study can be found in the supplemental material and at https://tfnetwork.streamlit.app. The code and software implementations associated with this study can be found at https://github.com/Ma-Lab-Seattle-Childrens-CGIDR/Mtb-TFA-fitness-regression and https://hub.docker.com/repositories/malabcgidr.

## References

[1] WHO, Global tuberculosis report 2023, 2023.

[2] J.E. Galagan, K. Minch, M. Peterson, A. Lyubetskaya, E. Azizi, L. Sweet, A. Gomes, T. Rustad, G. Dolganov, I. Glotova, T. Abeel, C. Mahwinney, A.D. Kennedy, R. Allard, W. Brabant, A. Krueger, S. Jaini, B. Honda, W.H. Yu, M.J. Hickey, J. Zucker, C. Garay, B. Weiner, P. Sisk, C. Stolte, J.K. Winkler, Y. Van de Peer, P. Iazzetti, D. Camacho, J. Dreyfuss, Y. Liu, A. Dorhoi, H.J. Mollenkopf, P. Drogaris, J. Lamontagne, Y. Zhou, J. Piquenot, S.T. Park, S. Raman, S.H. Kaufmann, R.P. Mohney, D. Chelsky, D.B. Moody, D.R. Sherman, and G.K. Schoolnik, The Mycobacterium tuberculosis regulatory network and hypoxia. Nature 499 (2013) 178-83.

[3] K.J. Minch, T.R. Rustad, E.J. Peterson, J. Winkler, D.J. Reiss, S. Ma, M. Hickey, W. Brabant, B. Morrison, S. Turkarslan, C. Mawhinney, J.E. Galagan, N.D. Price, N.S. Baliga, and D.R. Sherman, The DNA-binding network of Mycobacterium tuberculosis. Nat Commun 6 (2015) 5829.

[4] E.J. Peterson, D.J. Reiss, S. Turkarslan, K.J. Minch, T. Rustad, C.L. Plaisier, W.J. Longabaugh, D.R. Sherman, and N.S. Baliga, A high-resolution network model for global gene regulation in Mycobacterium tuberculosis. Nucleic Acids Res 42 (2014) 11291-303.

[5] E.J.R. Peterson, A.N. Brooks, D.J. Reiss, A. Kaur, J. Do, M. Pan, W.J. Wu, R. Morrison, V. Srinivas, W. Carter, M.L. Arrieta-Ortiz, R.A. Ruiz, A. Bhatt, and N.S. Baliga, MtrA modulates Mycobacterium tuberculosis cell division in host microenvironments to mediate intrinsic resistance and drug tolerance. Cell Rep 42 (2023) 112875.

[6] T.R. Rustad, K.J. Minch, S. Ma, J.K. Winkler, S. Hobbs, M. Hickey, W. Brabant, S. Turkarslan, N.D. Price, N.S. Baliga, and D.R. Sherman, Mapping and manipulating the Mycobacterium tuberculosis transcriptome using a transcription factor overexpression-derived regulatory network. Genome Biol 15 (2014) 502.

[7] R. Yoo, K. Rychel, S. Poudel, T. Al-Bulushi, Y. Yuan, S. Chauhan, C. Lamoureux, B.O. Palsson, and A. Sastry, Machine Learning of All Mycobacterium tuberculosis H37Rv RNA-seq Data Reveals a Structured Interplay between Metabolism, Stress Response, and Infection. mSphere 7 (2022) e0003322.

[8] J. Sanz, J. Navarro, A. Arbues, C. Martin, P.C. Marijuan, and Y. Moreno, The transcriptional regulatory network of Mycobacterium tuberculosis. PLoS One 6 (2011) e22178.

[9] G. Balazsi, A.P. Heath, L. Shi, and M.L. Gennaro, The temporal response of the Mycobacterium tuberculosis gene regulatory network during growth arrest. Mol Syst Biol 4 (2008) 225.

[10] S. Turkarslan, E.J.R. Peterson, T.R. Rustad, K.J. Minch, D.J. Reiss, R. Morrison, S. Ma, N.D. Price, D.R. Sherman, and N.S. Baliga, A comprehensive map of genome-wide gene regulation in Mycobacterium tuberculosis. Scientific Data 2 (2015).

[11] J.M. Escorcia-Rodriguez, E. Gaytan-Nunez, E.M. Hernandez-Benitez, A. Zorro-Aranda, M.A. Tello-Palencia, and J.A. Freyre-Gonzalez, Improving gene regulatory network inference and assessment: The importance of using network structure. Front Genet 14 (2023) 1143382.

[12] H. Poonawala, Y. Zhang, S. Kuchibhotla, A.G. Green, D.M. Cirillo, F. Di Marco, A. Spitlaeri, P. Miotto, and M.R. Farhat, Transcriptomic responses to antibiotic exposure in Mycobacterium tuberculosis. Antimicrob Agents Chemother 68 (2024) e0118523.

[13] C. Bei, J. Zhu, P.H. Culviner, M. Gan, E.J. Rubin, S.M. Fortune, Q. Gao, and Q. Liu, Genetically encoded transcriptional plasticity underlies stress adaptation in Mycobacterium tuberculosis. Nat Commun 15 (2024) 3088.

[14] J.C. Liao, R. Boscolo, Y.L. Yang, L.M. Tran, C. Sabatti, and V.P. Roychowdhury, Network component analysis: reconstruction of regulatory signals in biological systems. Proc Natl Acad Sci U S A 100 (2003) 15522-7.

[15] D. Marbach, J.C. Costello, R. Kuffner, N.M. Vega, R.J. Prill, D.M. Camacho, K.R. Allison, M. Kellis, J.J. Collins, and G. Stolovitzky, Wisdom of crowds for robust gene network inference. Nat Methods 9 (2012) 796-804.

[16] S. Ma, R. Morrison, S.J. Hobbs, V. Soni, J. Farrow-Johnson, A. Frando, N. Fleck, C. Grundner, K.Y. Rhee, T.R. Rustad, and D.R. Sherman, Transcriptional regulator-induced phenotype screen reveals drug potentiators in Mycobacterium tuberculosis. Nat Microbiol 6 (2021) 44-50.

[17] S.C. Hicks, K. Okrah, J.N. Paulson, J. Quackenbush, R.A. Irizarry, and H.C. Bravo, Smooth quantile normalization. Biostatistics 19 (2018) 185-198.

[18] P.H. Hsieh, C.M. Lopes-Ramos, M. Zucknick, G.K. Sandve, K. Glass, and M.L. Kuijjer, Adjustment of spurious correlations in co-expression measurements from RNA-Sequencing data. Bioinformatics 39 (2023).

[19] B. Langmead, and S.L. Salzberg, Fast gapped-read alignment with Bowtie 2. Nat Methods 9 (2012) 357-9.

[20] Y. Liao, G.K. Smyth, and W. Shi, featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30 (2014) 923-30.

[21] P. Ewels, M. Magnusson, S. Lundin, and M. Kaller, MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics 32 (2016) 3047-8.

[22] L. McInnes, J. Healy, and J. Melville, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv 1802.03426 (2020).

[23] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, Knowledge Discovery and Data Mining, 1996.

[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, Scikit-learn: Machine Learning in Python. J Mach Learn Res 12 (2011) 2825-2830.

[25] A.A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Dalla Favera, and A. Califano, ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics 7 Suppl 1 (2006) S7.

[26] A. Lachmann, F.M. Giorgi, G. Lopez, and A. Califano, ARACNe-AP: gene network reverse engineering through adaptive partitioning inference of mutual information. Bioinformatics 32 (2016) 2233-5.

[27] J.J. Faith, B. Hayete, J.T. Thaden, I. Mogno, J. Wierzbowski, G. Cottarel, S. Kasif, J.J. Collins, and T.S. Gardner, Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. PLoS Biol 5 (2007) e8.

[28] V.A. Huynh-Thu, A. Irrthum, L. Wehenkel, and P. Geurts, Inferring regulatory networks from expression data using tree-based methods. PLoS One 5 (2010).

[29] G. Sales, and C. Romualdi, parmigene--a parallel R package for mutual information estimation and gene network reconstruction. Bioinformatics 27 (2011) 1876-7.

[30] S. Aibar, C.B. Gonzalez-Blas, T. Moerman, V.A. Huynh-Thu, H. Imrichova, G. Hulselmans, F. Rambow, J.C. Marine, P. Geurts, J. Aerts, J. van den Oord, Z.K. Atak, J. Wouters, and S. Aerts, SCENIC: single-cell regulatory network inference and clustering. Nat Methods 14 (2017) 1083-1086.

[31] D.J. Reiss, N.S. Baliga, and R. Bonneau, Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks. BMC Bioinformatics 7 (2006) 280.

[32] D.J. Reiss, C.L. Plaisier, W.J. Wu, and N.S. Baliga, cMonkey2: Automated, systematic, integrated detection of co-regulated gene modules for any organism. Nucleic Acids Res 43 (2015) e87.

[33] A.V. Sastry, Y. Gao, R. Szubin, Y. Hefner, S. Xu, D. Kim, K.S. Choudhary, L. Yang, Z.A. King, and B.O. Palsson, The Escherichia coli transcriptome mostly consists of independently regulated modules. Nat Commun 10 (2019) 5536.

[34] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference and Prediction, Second Edition, Springer, 2008.

[35] H. Zou, and T. Hastie, Regularization and Variable Selection Via the Elastic Net. Journal of the Royal Statistical Society Series B: Statistical Methodology 67 (2005) 301-320.

[36] D. Chicco, Ten quick tips for machine learning in computational biology. BioData Min 10 (2017) 35.

[37] B.W. Matthews, Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et biophysica acta 405 (1975) 442-51.

[38] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, and G. Sherlock, Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25 (2000) 25-9.

[39] C. Gene Ontology, S.A. Aleksander, J. Balhoff, S. Carbon, J.M. Cherry, H.J. Drabkin, D. Ebert, M. Feuermann, P. Gaudet, N.L. Harris, D.P. Hill, R. Lee, H. Mi, S. Moxon, C.J. Mungall, A. Muruganugan, T. Mushayahama, P.W. Sternberg, P.D. Thomas, K. Van Auken, J. Ramsey, D.A. Siegele, R.L. Chisholm, P. Fey, M.C. Aspromonte, M.V. Nugnes, F. Quaglia, S. Tosatto, M. Giglio, S. Nadendla, G. Antonazzo, H. Attrill, G. Dos Santos, S. Marygold, V.

Strelets, C.J. Tabone, J. Thurmond, P. Zhou, S.H. Ahmed, P. Asanitthong, D. Luna Buitrago, M.N. Erdol, M.C. Gage, M. Ali Kadhum, K.Y.C. Li, M. Long, A. Michalak, A. Pesala, A. Pritazahra, S.C.C. Saverimuttu, R. Su, K.E. Thurlow, R.C. Lovering, C. Logie, S. Oliferenko, J. Blake, K. Christie, L. Corbani, M.E. Dolan, H.J. Drabkin, D.P. Hill, L. Ni, D. Sitnikov, C. Smith, A. Cuzick, J. Seager, L. Cooper, J. Elser, P. Jaiswal, P. Gupta, P. Jaiswal, S. Naithani, M. Lera-Ramirez, K. Rutherford, V. Wood, J.L. De Pons, M.R. Dwinell, G.T. Hayman, M.L. Kaldunski, A.E. Kwitek, S.J.F. Laulederkind, M.A. Tutaj, M. Vedi, S.J. Wang, P. D'Eustachio, L. Aimo, K. Axelsen, A. Bridge, N. Hyka-Nouspikel, A. Morgat, S.A. Aleksander, J.M. Cherry, S.R. Engel, K. Karra, S.R. Miyasato, R.S. Nash, M.S. Skrzypek, S. Weng, E.D. Wong, E. Bakker, et al., The Gene Ontology knowledgebase in 2023. Genetics 224 (2023).

[40] A. Kapopoulou, J.M. Lew, and S.T. Cole, The MycoBrowser portal: a comprehensive and manually annotated resource for mycobacterial genomes. Tuberculosis (Edinb) 91 (2011) 8-13.

[41] D.V. Klopfenstein, L. Zhang, B.S. Pedersen, F. Ramirez, A. Warwick Vesztrocy, A. Naldi, C.J. Mungall, J.M. Yunes, O. Botvinnik, M. Weigel, W. Dampier, C. Dessimoz, P. Flick, and H. Tang, GOATOOLS: A Python library for Gene Ontology analyses. Sci Rep 8 (2018) 10872.

[42] S. Carbon, and C. Mungall, Gene Ontology Data Archive (2024-06-17) [Data set]. Zenodo (2024).

[43] A. Noor, A. Ahmad, E. Serpedin, M. Nounou, and H. Nounou, ROBNCA: robust network component analysis for recovering transcription factor activities. Bioinformatics 29 (2013) 2410-8.

[44] T.A. Ahn-Horst, L.S. Mille, G. Sun, J.H. Morrison, and M.W. Covert, An expanded whole-cell model of E. coli links cellular physiology with mechanisms of growth rate control. NPJ Syst Biol Appl 8 (2022) 30.

[45] Y. Shi, G. Ke, Z. Chen, S. Zheng, and T.-Y. Liu, Quantized Training of Gradient Boosting Decision Trees. in: S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, (Eds.), 2022, pp. 18822--18833.

[46] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, LightGBM: A Highly Efficient Gradient Boosting Decision Tree. in: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, (Eds.), 2017.

[47] A.M. Sherrid, T.R. Rustad, G.A. Cangelosi, and D.R. Sherman, Characterization of a Clp protease gene regulator and the reaeration response in Mycobacterium tuberculosis. PLoS One 5 (2010) e11622.

[48] D.R. Sherman, M. Voskuil, D. Schnappinger, R. Liao, M.I. Harrell, and G.K. Schoolnik, Regulation of the Mycobacterium tuberculosis hypoxic response gene encoding alpha -crystallin. Proc Natl Acad Sci U S A 98 (2001) 7534-9.

[49] Y. Yuan, D.D. Crane, R.M. Simpson, Y.Q. Zhu, M.J. Hickey, D.R. Sherman, and C.E. Barry, 3rd, The 16-kDa alpha-crystallin (Acr) protein of Mycobacterium tuberculosis is required for growth in macrophages. Proc Natl Acad Sci U S A 95 (1998) 9578-83.

[50] S. Ma, S. Jaipalli, J. Larkins-Ford, J. Lohmiller, B.B. Aldridge, D.R. Sherman, and S. Chandrasekaran, Transcriptomic Signatures Predict Regulators of Drug Synergy and Clinical Regimen Efficacy against Tuberculosis. mBio 10 (2019).

[51] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and S. Genome Project Data Processing, The Sequence Alignment/Map format and SAMtools. Bioinformatics 25 (2009) 2078-9.

[52] Y. Benjamini, A.M. Krieger, and D. Yekutieli, Adaptive linear step-up procedures that control the false discovery rate. Biometrika 93 (2006) 491-507.

[53] Z. Wang, M. Gerstein, and M. Snyder, RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10 (2009) 57-63.

[54] S.L. Kendall, P. Burgess, R. Balhana, M. Withers, A. Ten Bokum, J.S. Lott, C. Gao, I. Uhia-Castro, and N.G. Stoker, Cholesterol utilization in mycobacteria is controlled by two TetR-type transcriptional regulators: kstR and kstR2. Microbiology (Reading) 156 (2010) 1362-1371.

[55] R. Sprinthall, Basic Statistical Analysis, Pearson Education, 2011.

[56] U. Alon, Network motifs: theory and experimental approaches. Nat Rev Genet 8 (2007) 450-61.

[57] S.J. Modlin, A. Elghraoui, D. Gunasekaran, A.M. Zlotnicki, N.A. Dillon, N. Dhillon, N. Kuo, C. Robinhold, C.K. Chan, A.D. Baughn, and F. Valafar, Structure-Aware Mycobacterium tuberculosis Functional Annotation Uncloaks Resistance, Metabolic, and Virulence Genes. mSystems 6 (2021) e0067321.

[58] G.B.D.A.R. Collaborators, Global burden of bacterial antimicrobial resistance 1990-2021: a systematic analysis with forecasts to 2050. Lancet (2024).

[59] M. Farhat, H. Cox, M. Ghanem, C.M. Denkinger, C. Rodrigues, M.S. Abd El Aziz, H. Enkh-Amgalan, D. Vambe, C. Ugarte-Gil, J. Furin, and M. Pai, Drug-resistant tuberculosis: a persistent global health concern. Nat Rev Microbiol 22 (2024) 617-635.

[60] M.N. Anahtar, J.H. Yang, and S. Kanjilal, Applications of Machine Learning to the Problem of Antimicrobial Resistance: an Emerging Model for Translational Research. Journal of clinical microbiology 59 (2021) e0126020.

[61] S. Lobentanzer, P. Rodriguez-Mier, S. Bauer, and J. Saez-Rodriguez, Molecular causality in the advent of foundation models. Mol Syst Biol 20 (2024) 848-858.

[62] V. Chen, M. Yang, W. Cui, J.S. Kim, A. Talwalkar, and J. Ma, Applying interpretable machine learning in computational biology-pitfalls, recommendations and opportunities for new developments. Nat Methods 21 (2024) 1454-1461.

[63] J.H. Yang, S.N. Wright, M. Hamblin, D. McCloskey, M.A. Alcantar, L. Schrübbers, A.J. Lopatkin, S. Satish, A. Nili, B.O. Palsson, G.C. Walker, and J.J. Collins, A White-Box Machine Learning Approach for Revealing Antibiotic Mechanisms of Action. Cell 177 (2019) 1649-1661.e9.

[64] A.J. Lopatkin, and J.H. Yang, Digital Insights Into Nucleotide Metabolism and Antibiotic Treatment Failure. Front Digit Health 3 (2021).

[65] J. Larkins-Ford, Y.N. Degefu, N. Van, A. Sokolov, and B.B. Aldridge, Design principles to assemble drug combinations for effective tuberculosis therapy using interpretable pairwise drug response measurements. Cell Rep Med 3 (2022) 100737.

[66] J. Larkins-Ford, T. Greenstein, N. Van, Y.N. Degefu, M.C. Olson, A. Sokolov, and B.B. Aldridge, Systematic measurement of combination-drug landscapes to predict in vivo treatment outcomes for tuberculosis. Cell Syst 12 (2021) 1046-1063 e7.

[67] M.A. DeJesus, E.R. Gerrick, W. Xu, S.W. Park, J.E. Long, C.C. Boutte, E.J. Rubin, D. Schnappinger, S. Ehrt, S.M. Fortune, C.M. Sassetti, and T.R. Ioerger, Comprehensive Essentiality Analysis of the Mycobacterium tuberculosis Genome via Saturating Transposon Mutagenesis. MBio 8 (2017).

[68] B. Bosch, M.A. DeJesus, N.C. Poulton, W. Zhang, C.A. Engelhart, A. Zaveri, S. Lavalette, N. Ruecker, C. Trujillo, J.B. Wallach, S. Li, S. Ehrt, B.T. Chait, D. Schnappinger, and J.M. Rock, Genome-wide gene expression tuning reveals diverse vulnerabilities of M. tuberculosis. Cell 184 (2021) 4579-4592 e24.

[69] S. Li, N.C. Poulton, J.S. Chang, Z.A. Azadian, M.A. DeJesus, N. Ruecker, M.D. Zimmerman, K.A. Eckartt, B. Bosch, C.A. Engelhart, D.F. Sullivan, M. Gengenbacher, V.A. Dartois, D. Schnappinger, and J.M. Rock, CRISPRi chemical genetics and comparative genomics identify genes mediating drug potency in Mycobacterium tuberculosis. Nat Microbiol 7 (2022) 766-779.

[70] W. Xu, M.A. DeJesus, N. Rucker, C.A. Engelhart, M.G. Wright, C. Healy, K. Lin, R. Wang, S.W. Park, T.R. Ioerger, D. Schnappinger, and S. Ehrt, Chemical Genetic Interaction Profiling Reveals Determinants of Intrinsic Antibiotic Resistance in Mycobacterium tuberculosis. Antimicrob Agents Chemother 61 (2017).

[71] P.O. Oluoch, E.-I. Koh, M.K. Proulx, C.J. Reames, K.G. Papavinasasundaram, K.C. Murphy, M.D. Zimmerman, V. Dartois, and C.M. Sassetti, Chemical genetic interactions elucidate pathways controlling tuberculosis antibiotic efficacy during infection. bioRxiv (2024) 2024.09.04.609063.

[72] A.F. Carey, J.M. Rock, I.V. Krieger, M.R. Chase, M. Fernandez-Suarez, S. Gagneux, J.C. Sacchettini, T.R. Ioerger, and S.M. Fortune, TnSeq of Mycobacterium tuberculosis clinical isolates reveals strain-specific antibiotic liabilities. PLoS Pathog 14 (2018) e1006939.

[73] N.D. Hicks, J. Yang, X. Zhang, B. Zhao, Y.H. Grad, L. Liu, X. Ou, Z. Chang, H. Xia, Y. Zhou, S. Wang, J. Dong, L. Sun, Y. Zhu, Y. Zhao, Q. Jin, and S.M. Fortune, Clinically prevalent mutations in Mycobacterium tuberculosis alter propionate metabolism and mediate multidrug tolerance. Nat Microbiol 3 (2018) 1032-1042.

[74] S. Stanley, C.N. Spaulding, Q. Liu, M.R. Chase, D.T.M. Ha, P.V.K. Thai, N.H. Lan, D.D.A. Thu, N.L. Quang, J. Brown, N.D. Hicks, X. Wang, M. Marin, N.C. Howard, A.J. Vickers, W.M. Karpinski, M.C. Chao, M.R. Farhat, M. Caws, S.J. Dunstan, N.T.T. Thuong, and S.M. Fortune, Identification of bacterial determinants of tuberculosis infection and treatment outcomes: a phenogenomic analysis of clinical strains. Lancet Microbe 5 (2024) e570-e580.

[75] C.C. The, A data compendium associating the genomes of 12,289 Mycobacterium tuberculosis isolates with quantitative resistance phenotypes to 13 antibiotics. PLoS Biol 20 (2022) e3001721.

[76] J.E. Cronan, The Escherichia coli FadR transcription factor: Too much of a good thing? Mol Microbiol 115 (2021) 1080-1085.

[77] M.A. Forrellad, M.V. Bianco, F.C. Blanco, J. Nunez, L.I. Klepp, C.L. Vazquez, L. Santangelo Mde, R.V. Rocha, M. Soria, P. Golby, M.G. Gutierrez, and F. Bigi, Study of the in vivo role of Mce2R, the transcriptional regulator of mce2 operon in Mycobacterium tuberculosis. BMC microbiology 13 (2013) 200.

[78] V. Gopinath, S. Raghunandanan, R.L. Gomez, L. Jose, A. Surendran, R. Ramachandran, A.R. Pushparajan, S. Mundayoor, A. Jaleel, and R.A. Kumar, Profiling the Proteome of

Mycobacterium tuberculosis during Dormancy and Reactivation. Mol Cell Proteomics 14 (2015) 2160-76.

[79] C. Larsson, B. Luna, N.C. Ammerman, M. Maiga, N. Agarwal, and W.R. Bishai, Gene expression of Mycobacterium tuberculosis putative transcription factors whiB1-7 in redox environments. PLoS One 7 (2012) e37516.

[80] J. Li, X. Wang, W. Gong, C. Niu, and M. Zhang, Crystallization and preliminary X-ray analysis of Rv1674c from Mycobacterium tuberculosis. Acta Crystallogr F Struct Biol Commun 71 (2015) 354-7.

[81] R. Manganelli, L. Cioetto-Mazzabo, G. Segafreddo, F. Boldrin, D. Sorze, M. Conflitti, A. Serafini, and R. Provvedi, SigE: A master regulator of Mycobacterium tuberculosis. Front Microbiol 14 (2023) 1075143.

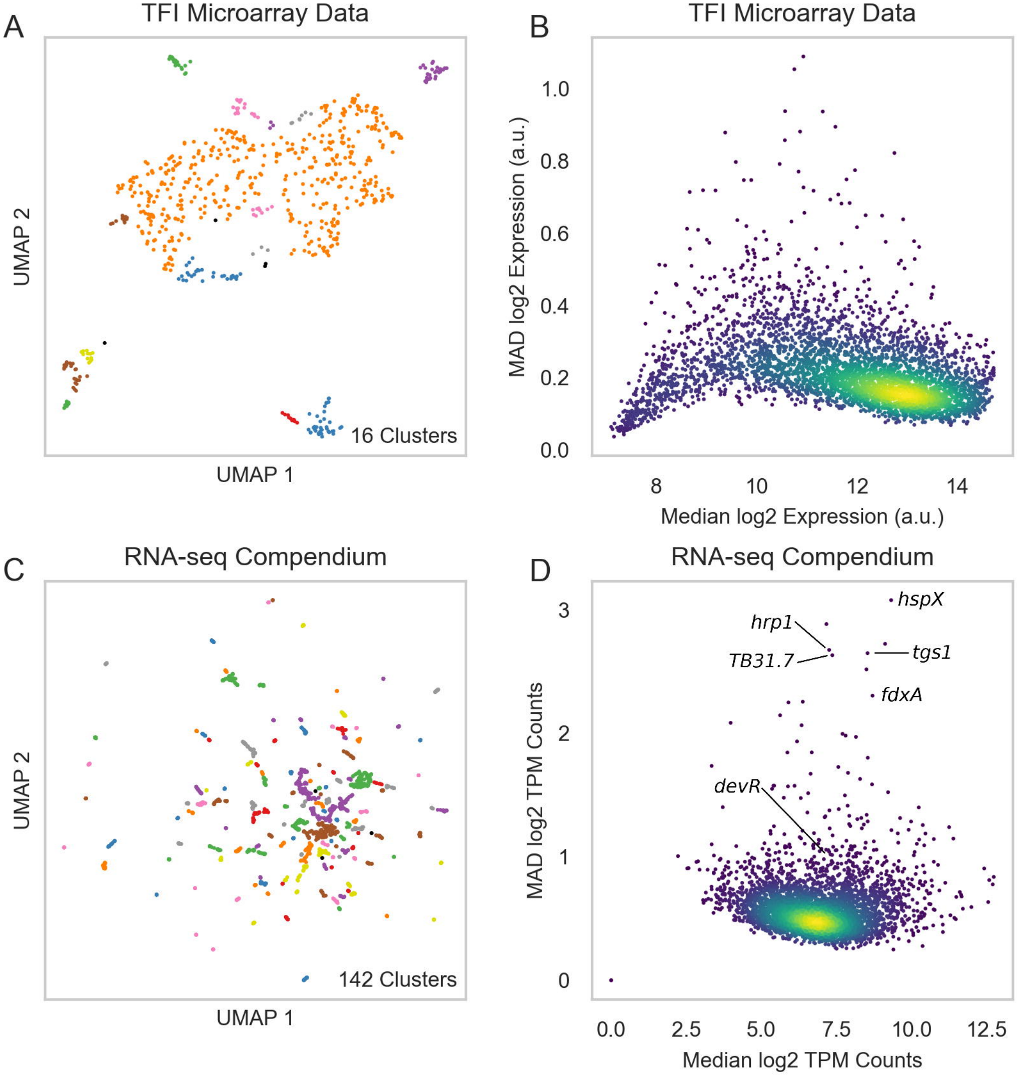
[82] S. Mehra, and D. Kaushal, Functional genomics reveals extended roles of the Mycobacterium tuberculosis stress response factor sigmaH. J Bacteriol 191 (2009) 3965-80.

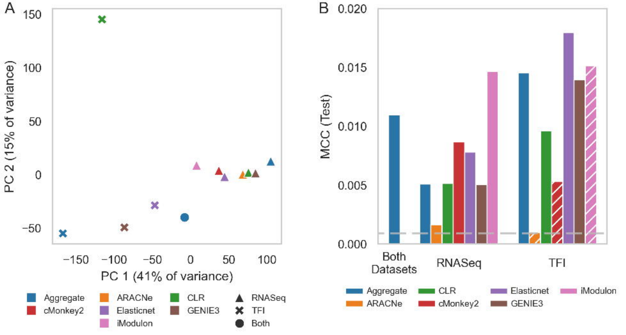
[83] B. Ramos, S.V. Gordon, and M.V. Cunha, Revisiting the expression signature of pks15/1 unveils regulatory patterns controlling phenolphtiocerol and phenolglycolipid production in pathogenic mycobacteria. PLoS One 15 (2020) e0229700.

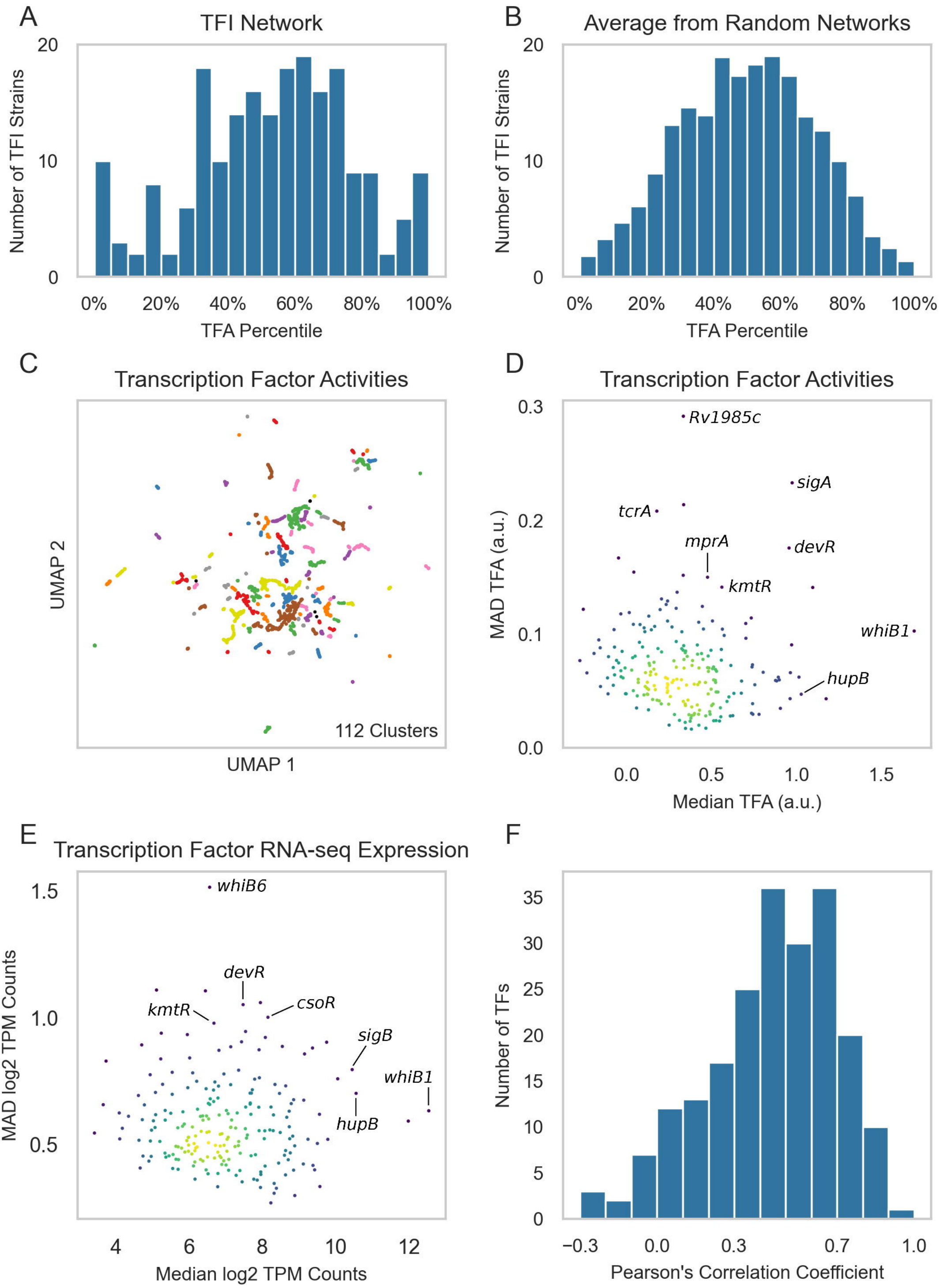
[84] S. Yousuf, R.K. Angara, A. Roy, S.K. Gupta, R. Misra, and A. Ranjan, Mce2R/Rv0586 of Mycobacterium tuberculosis is the functional homologue of FadR(E. coli). Microbiology (Reading) 164 (2018) 1133-1145.

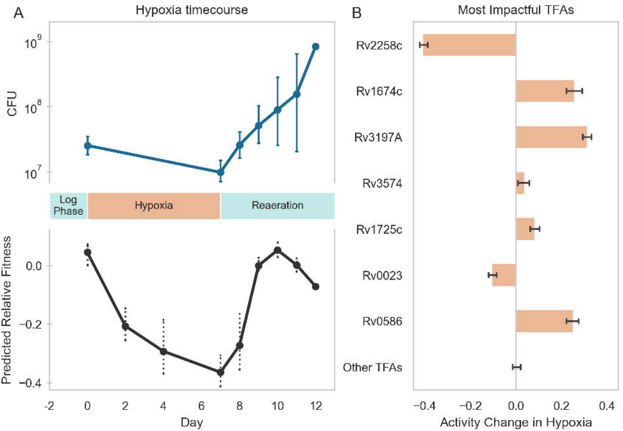
[85] T.R. Rustad, M.I. Harrell, R. Liao, and D.R. Sherman, The enduring hypoxic response of Mycobacterium tuberculosis. PLoS One 3 (2008) e1502.

[86] K.R. Nicholson, R.M. Cronin, A.R. Menon, M.K. Jennisch, D.M. Tobin, and P.A. Champion, The EspN transcription factor is an infection-dependent regulator of the ESX-1 system in M. marinum. bioRxiv (2023).


A
TFI Microarray Data
UMAP 2
UMAP 1
16 Clusters
B
TFI Microarray Data
MAD log2 Expression (a.u.)
0.0
0.2
0.4
0.6
0.8
1.0
8
10
12
14
Median log2 Expression (a.u.)
C
RNA-seq Compendium
UMAP 2
UMAP 1
142 Clusters
D
RNA-seq Compendium
MAD log2 TPM Counts
0
1
2
3
0.0
2.5
5.0
7.5
10.0
12.5
Median log2 TPM Counts
hspX
hrp1
TB31.7
tgs1
fdxA
devR


A
PC 2 (15% of variance)
150
100
50
0
−50
−150
−100
−50
0
50
100
PC 1 (41% of variance)
Aggregate
cMonkey2
ARACNe
Elasticnet
iModulon
CLR
GENIE3
RNASeq
TFI
Both
B
MCC (Test)
0.020
0.015
0.010
0.005
0.000
Both Datasets
RNASeq
TFI
Aggregate
ARACNe
CLR
cMonkey2
Elasticnet
GENIE3
iModulon


A
TFI Network
Number of TFI Strains
0
10
20
0%
20%
40%
60%
80%
100%
TFA Percentile
B
Average from Random Networks
Number of TFI Strains
0
10
20
0%
20%
40%
60%
80%
100%
TFA Percentile
C
Transcription Factor Activities
UMAP 2
UMAP 1
112 Clusters
D
Transcription Factor Activities
MAD TFA (a.u.)
0.0
0.1
0.2
0.3
Median TFA (a.u.)
0.0
0.5
1.0
1.5
Rv1985c
sigA
tcrA
mprA
devR
kmtR
whiB1
hupB
E
Transcription Factor RNA-seq Expression
MAD log2 TPM Counts
0.5
1.0
1.5
Median log2 TPM Counts
4
6
8
10
12
whiB6
devR
kmtR
csoR
sigB
whiB1
hupB
F
Number of TFs
0
5
10
15
20
25
30
35
−0.3
0.0
0.3
0.7
1.0
Pearson's Correlation Coefficient


A
Hypoxia timecourse
CFU
10^9
10^8
10^7
Log Phase
Hypoxia
Reaeration
Predicted Relative Fitness
0.0
−0.2
−0.4
-2
0
2
4
6
8
10
12
Day
B
Most Impactful TFAs
Rv2258c
Rv1674c
Rv3197A
Rv3574
Rv1725c
Rv0023
Rv0586
Other TFAs
−0.4
−0.2
0.0
0.2
0.4
Activity Change in Hypoxia