

ARTICLE OPEN



Deep neural network prediction of genome-wide transcriptome signatures – beyond the Black-box

Rasmus Magnusson ^{1,2✉}, Jesper N. Tegnér ^{3,4,5} and Mika Gustafsson ¹

Prediction algorithms for protein or gene structures, including transcription factor binding from sequence information, have been transformative in understanding gene regulation. Here we ask whether human transcriptomic profiles can be predicted solely from the expression of transcription factors (TFs). We find that the expression of 1600 TFs can explain >95% of the variance in 25,000 genes. Using the light-up technique to inspect the trained NN, we find an over-representation of known TF-gene regulations. Furthermore, the learned prediction network has a hierarchical organization. A smaller set of around 125 core TFs could explain close to 80% of the variance. Interestingly, reducing the number of TFs below 500 induces a rapid decline in prediction performance. Next, we evaluated the prediction model using transcriptional data from 22 human diseases. The TFs were sufficient to predict the dysregulation of the target genes ($\rho = 0.61$, $P < 10^{-216}$). By inspecting the model, key causative TFs could be extracted for subsequent validation using disease-associated genetic variants. We demonstrate a methodology for constructing an interpretable neural network predictor, where analyses of the predictors identified key TFs that were inducing transcriptional changes during disease.

npj Systems Biology and Applications (2022)8:9; <https://doi.org/10.1038/s41540-022-00218-9>

INTRODUCTION

Bridging the gap between genome sequences and phenotypes is a core challenge in genomics and personalized medicine. To this end, it is essential to characterize intermediate levels, such as cells, tissues, and organs, using a suite of molecular technologies. For example, genetic variants associated with diseases exert their effects through the modulation of these intermediate levels. The transcribed mRNA expression is one of the most accessible and important windows into the cell's regulatory machinery and changes in tissues and organs. Therefore, the analysis of mRNA expression is crucial for the study of diseases¹. Specifically, the elucidation of gene regulatory mechanisms is central since gene regulatory networks maintain cellular identity and mediate interactions between genetic variants and the environment of humans.

To reverse-engineer gene regulatory mechanisms, large amounts of RNA expression data have been generated from experimental model systems, including cell-lines and tissues from humans under different conditions. Among others, the recount2², the Genotype-Tissue Expression (GTEx) project³, and the ARCHS4 database⁴ have all made great amounts of data available. Bioinformatics analysis has been instrumental in clustering genes to make sense of such data and augmenting the power for hypothesizing putative genes involved in diseases⁵. Enrichment and pathway analysis increase the resolution by suggesting groups of genes or specific pathways associated with the observed changes in gene expression. Since transcription factors (TFs) are critical for the regulatory control of genes, a massive body of bioinformatics tools targets TF binding sites' predictions, suggesting key drivers behind pathways, groups of genes, or clusters⁶. While useful for descriptive purposes, such as associating such differentially expressed genes to many diseases, it is

challenging to gain functional and mechanistic insight into the regulatory machinery from such lists.

To advance beyond lists, clusters, and enrichment analysis, a complementary strategy, referred to as network science, instead targets the study of interactions between molecular entities, genotypes, and phenotypes^{7,8}. For example, gene regulation effectively acts via a network of interacting genes⁹. Notably, genes that interact with dysregulated genes without being differentially expressed themselves are often overlooked in differential expression studies¹⁰. Consequently, these networks are challenging to extract from data¹¹. The wisdom of the crowd strategy has turned out to be useful while not satisfactory¹². The limited progress originates from the fact that the inverse problem of inferring interactions from observations is statistically under-constrained. Moreover, these approaches have all struggled with the complex and non-linear dynamics that shape gene regulation, containing several saturation effects and abundant negative and positive feed-backs. These non-linearities impede most of the available correlation-based methods used to study gene expression¹³.

Recent progress in machine learning has fueled interest in whether such methods could facilitate the discovery and analysis of biological networks^{14,15}. Pioneering applications of deep neural networks (DNNs) in genomics include prediction of TF binding sites¹⁶ and the effects of non-coding genetic variants¹⁷. At the core of these techniques is the ability to capture non-linear relationships. The use of DNNs requires substantial amounts of data, which is now feasible due to the collection of massive amounts of genetic and RNA-seq data into easily accessible databases⁴. Beyond detection of features, recent use of DNN, such as deep autoencoders applied to transcription data, compresses gene expression data into a latent space. The original data can then be reconstructed from the latent space representation^{18,19},

¹Bioinformatics, Department of Physics, Chemistry and Biology, Linköping University, Linköping, Sweden. ²School of Bioscience, Systems Biology Research Center, University of Skövde, Skövde, Sweden. ³Biological and Environmental Sciences and Engineering Division, Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal 23955–6900, Saudi Arabia. ⁴Unit of Computational Medicine, Department of Medicine, Solna, Center for Molecular Medicine, Karolinska Institutet, Stockholm, Sweden. ⁵Science for Life Laboratory, Solna, Sweden. ✉email: rasmus.magnusson@his.se

DNNs have also been applied to understand the regulation of mRNA expression. Deep convolutional neural networks could predict 60–80% of human RNA abundance variation from the genomic sequence alone^{20,21}. While being the first important step towards predicting mRNA levels, the regulatory transcription factors were not separated from the remaining transcriptome, making a biological interpretation and translation to diseases challenging.

Here we develop a methodology that goes beyond producing lists of differentially expressed genes, but not so far as the yet intractable reconstruction of a complete gene regulatory network. Instead, we target the regulation exerted by the transcription factors and ask whether training a DNN on gene expression data could learn a predictive TF network. The expression of all other genes could be explained by the combinatorial control induced by the TFs. Importantly, we constraint the training such that the resulting predictive model is interpretable. We refer to this as a methodology for advancing beyond black-box machine learning models, which in turn is a first step towards what could be referred to as white-box fully interpretable models. We find that such models can indeed predict the expression of genes based on TFs and that the predicted relationships between TFs and their target genes largely overlap with known TF bindings. We apply and evaluate this predictive model using human disease

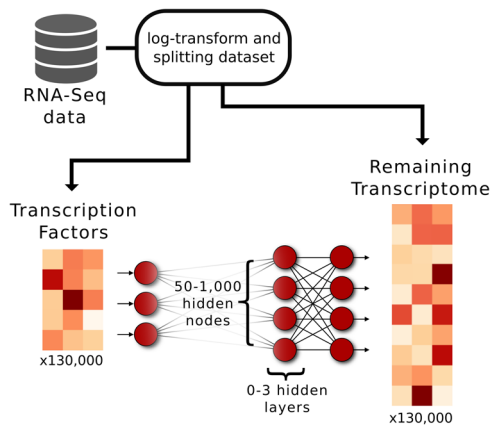


Fig. 1 Design of the artificial deep neural networks. The networks were made to predict the expression of 25,861 genes based on 1,625 gene regulators, i.e. transcription factors (TFs). We used more than 100,000 randomly drawn RNA-seq samples from the ARCHS4 database to train the models. Moreover, we designed 15 DNNs of 1–3 hidden layers, and one shallow neural network without any hidden layers. For the DNNs, each hidden layer consisted of either 50, 100, 250, 500, or 1000 hidden nodes using the exponential linear unit, ELU as activation function.

transcriptomes, thus opening the door for a mechanistic and interpretable machine learning analysis of the human gene regulation system.

RESULTS

First, we explore the influence of different neural network architectures, such as the number of layers and hidden nodes, on predictive performance. Next, we ask how we can inspect the trained network and disentangle the different predictors' contributions (TFs). Using a light-up network analysis technique, we identify a core set of TFs, including some key regulators. The final section of the results asks whether the prediction model could analyze disease-derived transcriptional data.

Accurate and robust prediction of the expression level of target genes using deep neural transcription factor networks

To optimize the applied Deep Neural Network (DNN) performance, we compared 15 DNN architectures consisting of 1–3 intermediate layers, each with a depth of 50–1000 hidden nodes. As a reference, we used a shallow NN without any intermediate layer (Fig. 1). The rationale is to identify the most compact architecture, measured by depth and width, capable of predicting most of the target gene expression with sufficient accuracy. Here, TFs were used exclusively as input and non-TFs as output target genes. We extracted TFs using the compendium provided by Lambert et al. 2018²², which lists TFs based on several sources, including popular TF databases such as TRANSFAC, JASPAR, and HT-SELEX. We trained models using the ARCHS4 database, using more than 100,000 randomly drawn RNA-seq samples to train and evaluate the models⁴. The performance was evaluated using the gene-specific coefficient of determination (presented as $1 - R^2$) on test data. We observed median $1 - R^2$ values between 0.12 and 0.03 depending on the model (Fig. 2a, b). Similarly, the mean $1 - R^2$ values of all DNNs were in the range of 0.11–0.07, whereas the shallow model had a notably worse performance ($1 - R^2 = 0.25$). A gene-specific list of the R^2 -coefficients, together with the corresponding Spearman correlations, can be found in Supplemental Table 1. Since the RNA-seq gene counts contain several intercept terms, such as sequencing depth and mean gene expressions, we tested the $1 - R^2$ when the input data were randomly permuted and found the mean $1 - R^2$ to be on the range of (0.79–0.67). Thus, we observed an increase in the ability to predict gene expression compared to 80% of the explained variance when predicting mRNA abundance solely from the DNA sequence, as Zrimec et al. 2020²¹ reported.

To test the trained models' generalizability, we applied them to predict expression from the independent Cancer Cell Line Encyclopedia resource²³, which contains mRNA profiles of 934

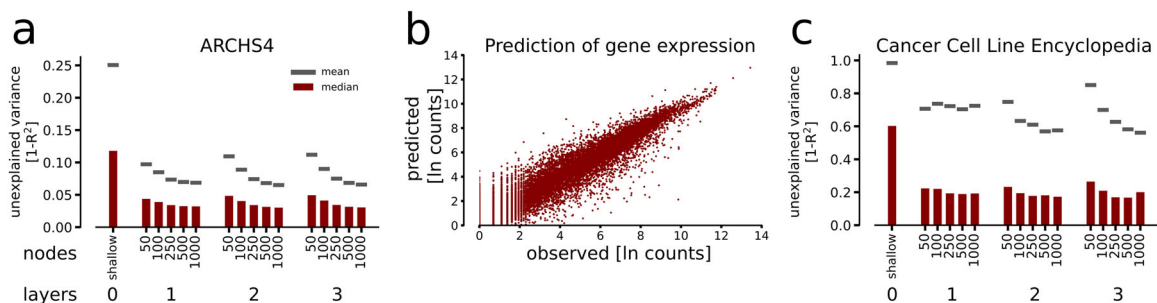


Fig. 2 Gene expression prediction performance. **a** We applied the DNNs and the shallow model to previously unseen gene expression profiles randomly selected from the ARCHS4 database, and calculated the coefficient of determination, R^2 , for each gene. Shown are the median (red bar) and mean (gray line) of the R^2 values for each model. **b** The typical prediction of an experiment is shown for the DNN with two hidden layers and 250 hidden nodes in each layer. **c** We applied the DNNs to predict gene expression values from 934 human tumor samples from the Cancer Cell Line Encyclopedia. Shown are the abilities to predict this data set for the respective models, following the same layout as in **a**.

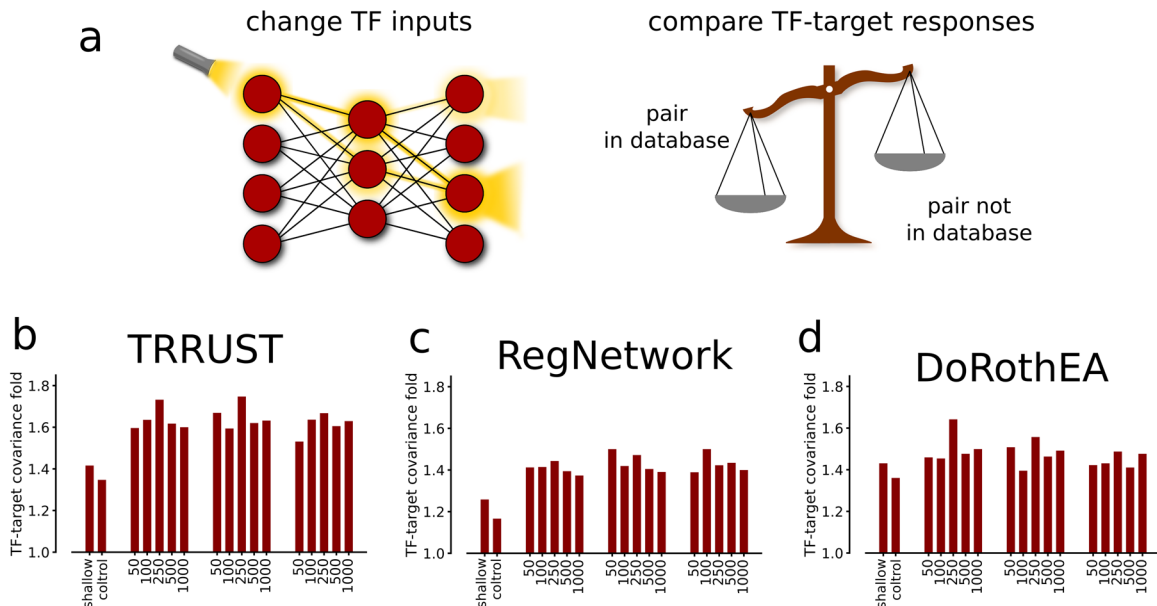


Fig. 3 Node light-up reveals known TF-target associations. **a** By applying a light-up analysis, i.e. by changing the input values of each TF independently and subsequently observing the changes on the output layer, we were able to estimate how the TF-to-target mappings corresponded to TF bindings known from literature. We defined the TF-target covariance fold as the median value of the light-ups for the TF-target regulations found in a database divided by the respective backgrounds. **b** The light-up enrichments for the TF-target associations found in the TRRUST database are shown. Note that the expected value representing no biological relevance between TF-target mapping is 1 and that light-up values are compared to absolute Spearman rank correlation values, labeled 'control'. We performed the same analysis with RegNetwork (**c**), and the interactions annotated with top confidence in DoRothEA (**d**).

human tumor cell lines. In this out-of-sample prediction, the DNNs showed a high median $1 - R^2$ between 0.26 and 0.17, whereas the shallow model only reached a median $R^2 = 0.60$. The mean $1 - R^2$ values were lower than the medians for all models (0.85–0.56), which we speculate is due to the inclusive list of target genes, such that genes with low variance deflate the mean $1 - R^2$. As a comparison, the top 25,000 genes had a mean $1 - R^2$ of < 0.25 in all deep models. With the input data randomly permuted, we observed the mean $1 - R^2$ in the range of (1.41–1.06). The non-linear DNNs could capture perturbations almost three-fold better than the shallow NN, suggesting their potential usefulness even in cancer medicine studies. We continued studying the predictive power of our approach on expression in individual cell types, and found a the learning to have captured a broad representation of human tissues (Supplemental Material 1). With these results, we concluded that the DNNs could faithfully predict the majority of the human transcriptome given the expression levels of ~ 1600 TFs, both in healthy and disease-affected states.

The node light-up technique revealed enrichment of validated TF-target associations within the prediction networks

Here we ask whether the trained networks are interpretable, that is, whether the learned TF target associations are biologically relevant or not. From a biological standpoint, one could compare directly with known physical DNA bindings between TFs and their target genes. However, due to the nonlinear dependencies embedded in a DNN, such an analysis is not trivial²⁴. For example, even simple DNNs can have millions of parameters, and all input values can potentially impact every output value. To approximate the learned dependencies between TFs and target genes, we, therefore, reasoned that in the predicted network, the TF expression perturbations would propagate most effectively to relevant target genes. In other words, such an analysis captures the effective gene dependencies, linear or non-linear. To this end, we used a light-up node analysis, following the implementation in¹⁸. In other words, the numerical value of each input node,

corresponding to a unique TF, was independently perturbed to either half or double of the mean gene expression. The other TF expression values were clamped to their average values, while the responses on the output layer were ranked by the response to such a TF change (Fig. 3a). To validate the DNNs, we compared the light-up values to previously known TF-target bindings. Specifically, we tested whether the light-up responses between such TF-target pairs were significantly higher than the pairs not annotated as interactions.

To this end, we used four sources of TF-target interactions. Specifically, the TF-specific DoRothEA database ($n_{TFs} = 94$, $n_{targets} = 2240$, $n_{edges} = 4498$), the ReMap dataset ($n_{TFs} = 296$, $n_{targets} = 15,196$, $n_{edges} = 269,757$), and the TRRUST database, which has a larger TF coverage ($n_{TFs} = 481$, $n_{targets} = 1,961$, $n_{edges} = 6,576$). The DoRothEA is based on manually curated TF-target interactions, and ChIP-seq derived measurements²⁵, the ReMap interactions are derived only from ChIP-seq measurements²⁶. The TRRUST database is an extraction of small-scale experimental analyses of TF regulations²⁷. We also included the RegNetwork database²⁸, which is a compilation of 25 commonly used TF-target binding databases and thus includes numerous additional interactions of somewhat less confidence ($n_{TFs} = 645$, $n_{targets} = 14,696$, $n_{edges} = 99,488$). We also noted a limited overlap between these three databases. For example, 927 interactions were shared between the TRRUST and DoRothEA sets, 1262 between DoRothEA and RegNetwork, 645 between ReMap and TRRUST, and 751 between RegNetwork and TRRUST. Strikingly, the light-up responses between known TF-target associations were significantly higher (26–75%) for all manually curated databases and DNNs (Fig. 3b–d). The more inclusive ReMap also displayed higher light-up values but at lower levels (5–8%). Thus, all models performed better than what was expected under the null hypothesis (60 of 60, binomial $P < 10^{-19}$). While the curated databases' results did not point to a single best model architecture, we observed the highest overall performance for the DNN with two hidden layers using 250 nodes in each layer. The highest enrichment rank was obtained

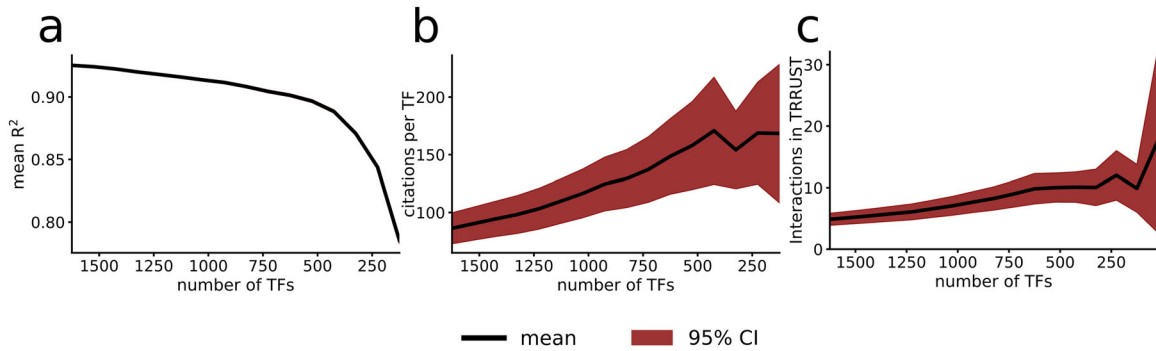


Fig. 4 Functional Hierarchies in Model Input. **a** We found the explanatory power on test data to decrease as the number of TFs were removed from the model. Particularly, we found around 500 TFs to carry a greater predictive power, as the explanatory power after that point more rapidly decreased. **b** We measured the median number of scientific studies at each iteration of our backward-selection algorithm, and found that as the number of input TFs decreased, the more studied were those that were left. **c** The TFs that were identified as more important also had more recorded interactions in the TRRUST database.

from TRRUST, while DoRothEA yielded the second-highest, and RegNetwork the lowest rank.

Since TFs can act as both inhibitors and initiators of transcription, we asked the extent to which the DNN light-up analyses also captured the directionality and sign of the TF-target interactions. Therefore, we compared the light-ups with known interactions, directions, and signs of TRRUST and DoRothEA. Of note, RegNetwork does not contain any annotation of interaction signs. Again, all tested DNNs showed significant overlaps with the interaction signs of both databases, with accuracies ranging between [0.62–0.67] ($P \in [10^{-120}, 10^{-58}]$) for DoRothEA, and [0.56–0.58] ($P \in [10^{-22}, 10^{-11}]$) for TRRUST. We provide the average precision- and receiver operating characteristic scores, with the corresponding figures, in the Supplemental Material 2. We defined accuracy as the percentage of correctly estimated signs of interaction. Again, the shallow model performed significantly worse and could only predict interaction signs compared to DoRothEA (accuracy = 0.57, $P < 10^{-20}$), while the comparison to TRRUST gave insignificant results.

In general, we find that the number of hidden layers or units for the DNNs has only a limited impact on the performance. All DNN models showed comparable performance across evaluative analysis, expression predictions of experiments from the ARCHS4 database, the cancer expression predictions, and the light-up comparison to TF-target databases. In practice, 250 hidden nodes in two hidden layers appeared to be a sufficient model size. Notably, the shallow NN never reached a satisfactory performance. Thus, the largest gain in explanatory power and overlap with existing databases came in our hands from adding at least one intermediate layer, thereby allowing for non-linear transformations. While these non-linearities turned out to be essential for the performance, they did not prevent us from inspecting the predictive network and extract relevant and validated biological knowledge using the light-up techniques.

Algorithmic extraction of a core set of validated regulator TFs from the DNN

We next searched for a minimal subset of key TFs required for predicting the target gene expressions. For this task, we implemented a backward-selection algorithm to stepwise remove TFs from the input layer based on their explanatory power (Methods). We observed highly consistent orders of TF-removal between the models, again suggesting that the TF-target relationship is robust to different DNN architectures. Indeed, the median correlation between different models, estimated from which step the TFs were removed, was 0.71 (geometric mean $P < 10^{-364}$). This suggests a high consistency between the independently trained models (Supplemental Material 3). This observation,

taken together with the light-up analysis outcome, suggests robustness of our results in that specifics of a DNN have only a minor influence on the results. Therefore, we used the model with two hidden layers and 250 hidden nodes in the analysis' remaining parts. The step at which each gene was removed is presented in Supplemental Table 2.

As expected, the explanatory power gradually decreases when removing predictors, here TFs (Fig. 4a). However, the shape of this loss of explanatory power exhibited two distinguishing features. First, the ability to predict mRNA expression remained relatively unperturbed even as most TFs were removed from the set. Indeed, for $n_{TFs} = 125$, the mean $1 - R^2 = 0.21$ compared to that of the full model of 1,625 TFs, which measured 0.07. Second, there appeared to be a stratification of TFs based on how important they are to explain the system, leading to a larger reduction in explanatory power towards the procedures' late stage.

To test whether our predicted core set of TFs could correspond to known, important regulatory TFs, we estimated how often or much they have been studied in the literature domain. First, we analyzed the average number of scientific publications per input TF, as annotated by PubMed²⁹, and found that the top predicted TFs were significantly more frequently studied (Fig. 4b). For example, the median number of publications was 51 for the top 125 TFs, compared to 24 for all TFs (Mann-Whitney U test $P < 8.2 \times 10^{-9}$). Second, we analyzed the light-up conformity to the TRRUST database, but this time for each step in the backward-selection algorithm. We found that the full model performed equally well or better at next-to-all steps in the backward-selection algorithm. This observation indicates that the model predominantly discovered known TF-target interactions regardless of model input size. Third, the number of annotated interactions per TF was significantly larger for the top-ranked TFs in the literature-based database TRRUST (9.9 for top 125 compared to 4.9, $P < 2.4 \times 10^{-6}$), suggesting that the top explanatory TFs were associated with more well-known regulatory bindings. We next analyzed the functional role of the removed TFs and performed enrichment analysis of their predicted downstream targets from our light-up technique (Methods). For each set of 100 removed TFs using the backward-selection procedure, we associated their top 500 target genes, which were then subjected to a KEGG pathway³⁰ enrichment analysis. In brief, we found that the top 100 TFs (i.e., the last removed TFs) were mostly enriched for cellular metabolism pathways, followed by immune-pathways for the TFs with rank 100–200 (Supplemental Table 3). Narrowing down further, we analyzed the 25 TFs preserved in our very last step in and a Pubmed query using the search term “master regulator” together with each of the TFs and found that 13 were co-localized with that search term in 358 unique articles, which supports the

notion that these are true upstream regulators. Thus, despite having abundant input variables, the DNN discovered known core TFs associated with central and well-annotated pathways.

Latent DNN space shows enrichment of functionally related and disease-associated genes

In addition to assessing the biological relevance of the specific interactions discovered by the predictive DNN, one may ask whether the predictor could be informative in a disease context. It has been demonstrated that compressing mRNA expression data in deep autoencoders can provide low-dimensional representations exposing complex characteristics of the input data domain. Importantly, from such a representation, one can readily extract sets of functionally related genes, known as modules, which in turn can be used to study diseases¹⁸. To further increase the resolution beyond the analysis of modules, here we aimed to use the TF-target gene interactions within the two 250 variables measuring intermediate latent layers.

Specifically, we tested whether the genes associated with these hidden variables shared cellular functionality. For this purpose, we again used a light-up response to associate genes to each hidden node independently. We annotated the top 500 responding genes to each node light-up by this procedure and performed a Reactome pathway enrichment analysis for each node-set. This procedure identified 175 unique Reactome pathways that significantly overlapped with at least one hidden node (using a Bonferroni correction of 0.05). In detail, 102 hidden nodes of the first and 162 of the second layer were associated with at least one Reactome pathway. These results suggested at least a subset of the hidden nodes represented different cellular pathways. We evaluated the putative agglomeration of disease-related genes in the hidden-layer light-up responses. To this end, using each hidden node, we performed a genome-wide association study (GWAS) enrichment analysis between annotations in the NHGRI-EBI GWAS catalog³¹ and the 500 genes with the highest light-up associations. We found 37 of the 153 diseases associated with one or several nodes, with 63 nodes in the first and 62 nodes in the second layer having at least one association. Second, we cross-checked these results by comparing the light-ups with the DisGeNET database³², which contains broader profiles of gene-disease associations. We tested against genes grouped in 26 disease-type categories and found 22 of these categories enriched with at least one hidden node. At least one significant overlap with a disease category was detected in 107 nodes in the first and 138 in the second layer. Thus, disease-genes appeared to co-occur in hidden-node light-ups in the DNN. This result suggests that the remaining genes found in such disease modules could be relevant in analyzing and interpreting disease-related biomarkers and mechanisms.

DNN analysis gives insights to human disease mechanisms involved in gene dysregulation

We finally aimed to test the clinical relevance of the DNN by using it to study disease-related changes in the expression of target genes, given the corresponding modifications of TF levels. To this end, we applied the DNN to independent RNA-seq data from the Expression Atlas³³ using differential expression patterns for diseases from 27 different studies containing 69 expression fold profiles. We tested if the DNN could predict the differential expression of target genes given the fold profiles of the TFs. Prediction quality was measured as the correlation between the predicted and observed fold ranking of significantly differentially expressed target genes for each disease. In other words, we set the TFs to their reference expression levels and applied the fold changes of each respective disease. We next calculated the correlation between the observed and predicted fold changes at the output layer. Notably, we observed highly significant correlations between these predictions and observations, with a median Spearman rank correlation of 0.61 (median $P < 10^{-216}$). This result established that disease mechanisms of dysregulation from TFs

could be faithfully propagated to the target level. Yet, as a correlation does not imply causation, we assessed whether the DNN could be used to disentangle which TFs drive the target dysregulation. To rank the impact of a predicted TFs, we replaced dysregulated genes with reference expression values independently for each TF. By this, we could evaluate the impact on the output layer and therefore use the change in correlation between predicted and observed target dysregulation (Fig. 5) as a basis for the ranking. Next, we matched these rankings to known genetic variants from genome-wide association studies (GWAS) in the 22 applicable cases and measured the area under the receiver operating characteristics curve (AUROC). The TF rankings in 10 out of 22 diseases significantly matched the TF GWAS annotation (false discovery rate = 0.11, 10/22 binomial test $P < 3.6 * 10^{-8}$). Furthermore, we found 20 of 22 diseases to have an AUROC larger than 0.5, i.e., the value expected under the null hypothesis (binomial test $P < 6.0 * 10^{-8}$).

Finally, we asked if TFs had to be differentially expressed to carry predictive power over disease-affected target gene regulation. This question is particularly important since causative disease-related changes are not necessarily manifested through dysregulations that are large enough to be detected in multiple-testing corrected statistical tests of expression changes. Notably, we found our approach of causative predictions on a TF level to also give significant results when only applied to TFs that themselves were not significantly dysregulated, with seven diseases remaining significant (Fig. 5, $P < 6.8 * 10^{-5}$). In other words, the removal of differential expression of TFs could predict disease-relevant TFs, even if the change in mRNA levels was modest. This suggests that our approach is generally applicable to find disease-causative elements at the TF-level, beyond what a more conventional RNA-seq analysis of gene expression detects. In summary, the TF-to-target DNN allowed a simple representation to describe the observed differential expression between patients of many different complex diseases and control in terms of TFs, which strikingly also correctly associated a truly enriched fraction of disease-associated TFs to several diseases.

DISCUSSION

We have presented a biologically interpretable, general machine learning method for predicting transcriptomic signatures, including disease signatures. Our learned models predict the expression of genes from the expression of transcription factors (TFs). The predicted relationships between TFs and their target genes largely overlap with known TF bindings. Hence, our DNN method goes beyond classic descriptive bioinformatic techniques such as clustering and enrichment analysis. Importantly, we do not address the yet intractable problem of complete deconvolution of the entire cellular interactome. Instead, our method does extract a core TF component out of such complex regulatory architecture. Therefore, the presented methodology sets the stage for the first step towards mechanistic and interpretable data-driven machine learning analysis of the human gene regulation system.

The early success of machine learning techniques targeted fundamental open bioinformatics problems such as predicting binding sites of TFs¹⁶ or functional prediction of non-coding RNA in a disease context¹⁷. The recent extension has focused on questions such as if one data type can be predicted from another datatype. Predicting gene expression from the DNA sequence or prediction of the 3D genomic structure from open chromatin data are just two recent examples. Tasaki et al.³⁴ used deep convolutional neural networks to predict differential expression based on mRNA features and the promoter regions of the genes. The L1000 project deals directly with the prediction of gene expression from a smaller subset of genes, where measurements of 1000 landmark genes are used to infer 80% of the transcriptome³⁵ computationally. Yet, here as in the other examples, these impressive modeling advances are difficult to translate into biological knowledge. As in different machine learning areas, these systems are useful predictors but act in

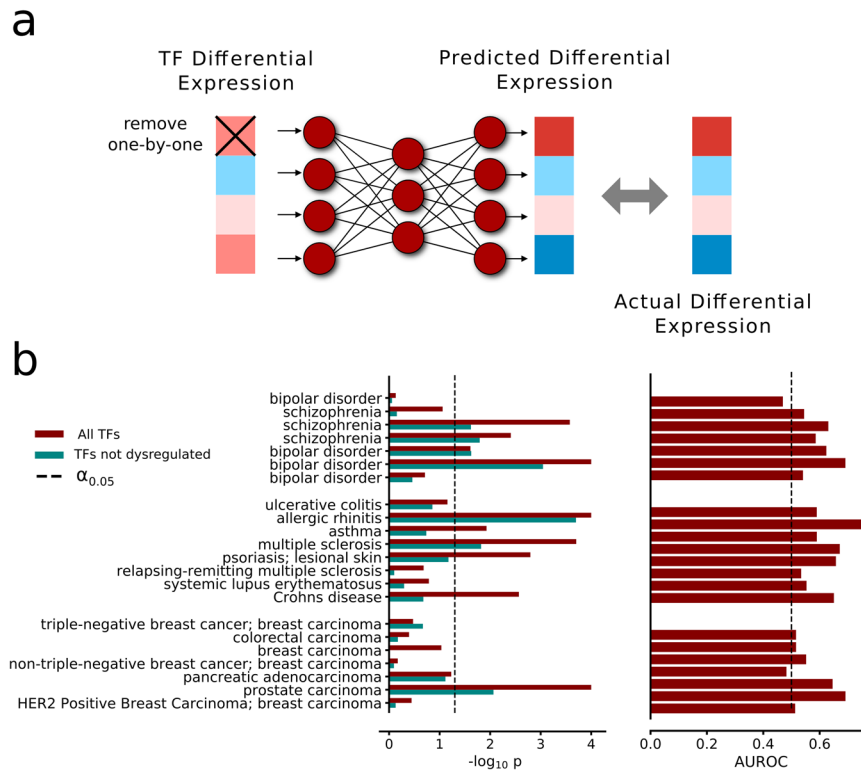


Fig. 5 Predicting gene dysregulations in disease using a DNN. **a** We tested if the DNN with two hidden layers of 250 hidden nodes each could be used to predict causative changes in disease states. We did this by analyzing gene expression changes from known diseases, as available in the Expression Atlas repository. By applying the disease changes to the transcription factor input layer, we could observe how these changes projected down to the target genes. Next, we removed the disease-fold changes of each TF independently and observed the changes in correlation between predicted and observed dysregulation of the target genes. Thus, we could rank the TFs on predicted causative disease changes on the target genes. **b** TF rankings significantly overlapped with GWAS in 10 of 22 diseases. ($-\log_{10} P$ values shown as red bars). The test was repeated with significantly differentially expressed TFs removed from the set, leaving 7 TF rankings to overlap with GWAS. ($-\log_{10} P$ values shown as teal bars). The corresponding area under the curves for the TF rankings (all TFs) are shown to the right, with 20 of 22 diseases having an AUROC greater than the expected, as generated under the null hypothesis.

practice as black-box systems. Related to this challenge is that even if functionally relevant representations can be identified, they can readily suffer from study-biases of identified prevalent genes. The remaining core challenges in using machine learning techniques include the interpretability of the model, large data requirement, and how to learn biologically meaningful representations within the machine learning model. A black-box model does not lend itself to interpretable and meaningful representations, potentially making the model more susceptible to adversarial attacks^{36,37}. Recently, it has become increasingly clear that deep neural networks (DNNs) have the potential to identify biologically meaningful molecular representations directly from data^{18,38}, and to revolutionize medicine^{39,40}. It is therefore of critical importance for the field to develop techniques supporting biological interpretation and insights from such predictive models.

Our main contribution is to design a constrained machine learning approach such that the predictor is interpretable from a biological standpoint. Using two hidden layers with 250 hidden nodes each was sufficient to capture well-known TF-target pairs. These results suggest that the learned representation has significant architectural overlap compared to a proper cellular control system, which is essential for the good generalizability of DNNs. By developing a back-tracking algorithm, we could uncover a minimal core set of 125 TFs sufficient to account for $\sim 80\%$ of the transcriptomics signature. Interestingly, these TFs represent the TFs being studied most frequently, which makes further studies linking prior structural information from other data types a logical next step to increase generalizability.

Our focus on TFs originates from the observation that TFs have been at the forefront in analyzing cellular reprogramming and converting cell-types^{41,42}. Our findings suggest that the presented DNN methodology could be useful as a general method for predictive but yet interpretable studies. In particular to pinpoint key candidate TFs for cellular reprogramming using large transcriptomics data⁴³. One of the most common approaches to analyzing disease mechanisms is studying gene expression changes between healthy and disease-affected individuals. Traditionally, a suite of bioinformatics descriptive mining techniques is applied to extract putative candidates of interest⁵. Nevertheless, it has proven challenging to pinpoint molecular mechanisms with high precision using such data-mining approaches. Consider a scenario where the impact of a perturbation on the gene regulatory system is of interest. Such cases are a common end-goal of analyzing transcriptomics and pivotal to understanding mechanisms such as drug perturbations or cancer impacts⁴⁴. Using our light-up analyses could help predict specific regulatory interactions and their effect on the transcriptome. This could be useful both for cellular reprogramming experiments as well as in the analysis of diseases. In contrast, a machine learning model using only gene sequences would not readily consider such changes.

Lastly, since our predictor's architecture has a biological interpretation, it could be used as a first approximation – like a blueprint – of the regulatory networks controlling the cell-identity and filtering effects of genetic variants. This problem has been at the forefront in systems biology since sequencing the human genome^{12,45}. Yet, despite two decades of brilliant work on reverse-

engineering gene regulatory networks from data, it remains an open challenge⁴⁶. The combinatorial complexity of such a network exceeds the amount and quality of available data given the current suite of models¹². As network predictions vary between methods, new tools have been developed to control the abundance of false interactions⁴⁷, illustrating that the problem is still outstanding. Yet, if robust methods, such as our proposed DNN technique, could elucidate the TF part of such a network, we could potentially approach the problem in a step-wise manner. Hence, in summary, a TF-centric reverse-engineering technique could therefore be a stepping stone for renewed systems biology efforts in elucidating the cellular regulatory machinery at scale.

MATERIALS AND METHODS

Data processing

We trained the models on gene expression data from ARCHS4, a database of >130,000 human RNA-seq gene expression measurements from GEO and SRA. We first separated the data into two sets, one of the genes annotated as TFs and one containing the rest of the genes, which were assumed to be regulated downstream of the TFs. We defined a TF as genes identified by Lambert et al.²², *The Human Transcription Factors*, which lists TFs based on TF databases such as TRANSFAC, JASPAR, and HT-SELEX. We defined the target genes, such as all the remaining genes, excluding pseudo-genes.

Next, we divided the data into 100 comma-separated files to be randomly accessed during the subsequent model training. Moreover, we removed 1200 gene expression profiles from the training set to use as validation of the model predictions. We normalized the data by applying the natural logarithm to the expression counts, annotated as x in Eq. (1).

$$x_{normalized} = \ln(x + 1) \quad (1)$$

Model design

We next aimed to predict the expression levels of the target genes using the TF levels. To this end, we designed the models to be feed-forward, fully connected neural networks. We built the models to have one input node for each TF, totaling 1525 input nodes and one output node for each target gene in the data, totaling 25,861 output nodes. Moreover, we opted to use the exponential linear unit (ELU) activation function on all nodes and across all layers. We used the Adam algorithm to minimize the mean squared error, with a learning rate of 0.001, parameter $\beta_{t1} = 0.9$, $\beta_{t2} = 0.999$, and decay of 0.01. The models were trained in the Keras package for Python 3, where we used a batch size of 50 experiments and with a validation split of 0.1. The code is available at https://github.com/rasma774/tf_target_white_box_dl.

Model light-up analysis

NNs are complex, non-linear models, and mapping input to output is not trivial. Here, we aimed to extract TF-target relationships TF-by-TF via comparing the model output between I) the output when the mean TF expression is given as input and II) when each TF has a doubled and halved expression. The rationale behind this approach is that targets that are dependent on a particular TF will have a greater response when the input value of this TF is altered, as compared to unrelated TFs. We next compared these responses to databases of known TF-target interactions, such that for each target, we divided the median of the light-up values for TFs that were known to regulate the target with the median of the rest of the TFs. In other words, we normalized the median light-up value with respect to the background. This metric is referred to as the TF-target covariance-fold in the manuscript.

The backwards-selection algorithm

We used a backwards-selection algorithm to identify the core set of TFs, i.e., the minimal set of TFs that could predict the target genes. The algorithm operated according to the following three steps. First, each input node, corresponding to one TF, was independently set to zero, and the corresponding $1 - R^2$ values were calculated. Second, the 100 TFs with the lowest explanatory power, identified as the ones where the $1 - R^2$ changed the least, were removed from the input layer of the DNN. Thirdly, we retrained the new and smaller model to explain the rest of the gene

expressions. For each iteration, we tested the model on the same test data from the ARCHS4-database as in the first validation experiment.

Disease analyses

To analyze diseases, we searched the Expression Atlas³³ to download all data according to the following criteria. I) The data was to come from a study carried out in human material. II) We only considered RNA-seq experiments. III) The data sets had the term 'disease' listed as an experimental variable and be of the 'differential' type. This query yielded data from a total of 27 studies, which together contained 69 expression fold profiles. Of these 69 comparisons, 56 were between a disease-affected and a healthy state, as opposed to between two disease-states, and we continued with these 56 studies. The 56 studies contain fold-changes between the healthy and disease-affected states, and by adding these changes to the mean expression values from the ARCHS4 database we could predict the fold changes on the target gene level. Furthermore, we chose only to study the correlation between the genes that were differentially expressed at a false discovery rate of 0.05.

The predictions of TFs causative of disease were made by adding the fold changes to the TFs as described above, followed by removing the dysregulation of each TF independently. This resulted in 1625 predicted changes in correlation between predicted and measured target expressions as compared to that from the full TF profile. We ranked the TFs on change in correlations and calculated the area under the receiver operating characteristics curve, i.e., the AUROC, for this ranking. A true positive identification was defined as the TF being associated with the respective disease, as manually curated from the GWAS catalog³¹. We next Monte Carlo sampled 10,000 random TF permutations and estimated the P value from the random AUROC distribution.

DATA AVAILABILITY

The training data were downloaded from the ARCHS4 database, annotated as "gene level" (accessed at <https://maayanlab.cloud/archs4/download.html>) and used to train the models. The curated disease data files can be accessed at https://github.com/rasma774/tf_target_white_box_dl.

CODE AVAILABILITY

The code used to generate the results, along with the DNN with 250 hidden nodes in two layers, can be found at https://github.com/rasma774/tf_target_white_box_dl.

Received: 24 June 2021; Accepted: 24 January 2022;

Published online: 23 February 2022

REFERENCES

1. Conesa, A. et al. A survey of best practices for RNA-seq data analysis. *Genome Biol.* **17**, 13 (2016).
2. L. Collado-Torres, et al. Reproducible RNA-seq analysis using recount2. *Nat. Biotechnol.* **35**, 319–321 (2017).
3. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
4. Lachmann, A. et al. Massive mining of publicly available RNA-seq data from human and mouse. *Nat. Commun.* **9**, 1366 (2018).
5. Hofmann-Apitius, M. et al. Bioinformatics Mining and Modeling Methods for the Identification of Disease Mechanisms in Neurodegenerative Disorders. *Int. J. Mol. Sci.* **16**, 29179–29206 (2015).
6. Andersson, R. & Sandelin, A. Determinants of enhancer and promoter activities of regulatory elements. *Nat. Rev. Genet.* **21**, 71–87 (2020).
7. Liu, C. et al. Computational network biology: Data, models, and applications. *Phys. Rep.* **846**, 1–66 (2020).
8. Diaz, J. E. et al. The transcriptomic response of cells to a drug combination is more than the sum of the responses to the monotherapies. *eLife* **9**, 1–62 (2020).
9. Ding, J. & Bar-Joseph, Z. Analysis of time-series regulatory networks. *Curr. Opin. Sys. Biol.* **21**, 16–24 (2020).
10. Bruhn, S., et al. A generally applicable translational strategy identifies S100A4 as a candidate gene in allergy. *Sci. Transl. Med.* **6**, 218ra4 (2014).
11. Castro, D. M., de Veaux, N. R., Miraldi, E. R. & Bonneau R. Multi-study inference of regulatory networks for more accurate models of gene regulation. *PLoS Comput. Biol.* **15**, e1006591 (2019).

12. Marbach, D. et al. Wisdom of crowds for robust gene network inference. *Nat. Methods*. 2012.
13. Saint-Antoine, M. M. & Singh, A. Network inference in systems biology: recent developments, challenges, and applications. *Curr. Opin. Biotechnol.* **63**, 89–98 (2020).
14. Camacho, D. M., Collins, K. M., Powers, R. K., Costello, J. C. & Collins, J. J. *Cell*. **14**, 1581–1592 (2018).
15. Jin, S., Zeng, X., Xia, F., Huang, W. & Liu, X. Application of deep learning methods in biological networks. *Briefings in Bioinformatics*. 1477–4054 (2020).
16. Alipanahi, B., Delong, A., Weirauch, M. T. & Frey, B. J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* <https://doi.org/10.1038/nbt.3300> (2015).
17. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence mode. *Nature Methods*. <https://doi.org/10.1038/nmeth.3547> (2015).
18. Dwivedi, S. K., Tjärnberg, A., Tegnér, J. & Gustafsson, M. Deriving disease modules from the compressed transcriptional space embedded in a deep autoencoder. *Nat. Commun.* <https://doi.org/10.1038/s41467-020-14666-6> (2020).
19. Grønbech, C. H. et al. scVAE: variational auto-encoders for single-cell gene expression data. *Bioinformatics* **36**, 4415–4422 (2020).
20. Agarwal, V. & Shendure, J. Predicting mRNA Abundance Directly from Genomic Sequence Using Deep Convolutional Neural Networks. *Cell Rep.* <https://doi.org/10.1016/j.celrep.2020.107663> (2020).
21. Zrimec, J. et al. Deep learning suggests that gene expression is encoded in all parts of a co-evolving interacting gene regulatory structure. *Nat. Commun.* **11**, 20411723 (2020).
22. Lambert, S. A. et al. The human transcription factors. *Cell*. **172**, 650–665. <https://doi.org/10.1016/j.cell.2018.01.029> (2018).
23. Barretina, J. et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*. <https://doi.org/10.1038/nature11003> (2012).
24. J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson. Understanding Neural Networks Through Deep Visualization. Preprint at <http://arxiv.org/abs/1506.06579> (2015).
25. Garcia-Alonso, L., Holland, C. H., Ibrahim, M. M., Turei, D. & Saez-Rodriguez, J. Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome Res.* **29**, 1363–1375 (2019).
26. Chéneby, J. et al. ReMap 2020: A database of regulatory regions from an integrative analysis of Human and Arabidopsis DNA-binding sequencing experiments. *Nucleic Acids Res.* **48**, D180–D188 (2020).
27. Han, H. et al. TRRUST: a reference database of human transcriptional regulatory interactions. *Sci Rep.* **12**, 11432 (2015).
28. Liu, Z. P., Wu, C., Miao, H. & Wu, H. RegNetwork: An integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. *Database* **2015**, 1–12 (2015).
29. Maglott, D., Stel, J., Pruitt, K. D. & Tatusova, T. Entrez gene: Gene-centered information at NCBI. *Nucleic Acids Res.* **35**, D26–31 (2011)
30. Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M. & Hirakawa, M. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.* **38**, D355–D360 (2009).
31. Buniello, A. et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).
32. Piñero, J. et al. DisGeNET: A comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res.* **45**, D833–D839 (2017).
33. Papatheodorou, I. et al. Expression Atlas update: From tissues to single cells. *Nucleic Acids Res.* **48**, D77–D83 (2020). no. D11ssn: 13624962.
34. Tasaki, S., Gaiteri, C., Mostafavi, S. & Wang, Y. Deep learning decodes the principles of differential gene expression. *Nature Machine Intelligence*. <https://doi.org/10.1038/s42256-0200201-6>. (2020).
35. Subramanian, A. et al. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell*. **171**, 1437–1452.e17.
36. Roscher, R., Bohn, B., Duarte, M. F. & Garcke, J. Explainable Machine Learning for Scientific Insights and Discoveries. *IEEE Access*. <https://doi.org/10.1109/ACCESS.2020.2976199> (2020).
37. Yu, M. K. et al. Visible Machine Learning for Biomedicine. *Cell* **173**, 1562–1565 (2018).
38. Naito, T. et al. A deep learning method for HLA imputation and trans-ethnic MHC fine-mapping of type 1 diabetes. *Nat. Commun.* **12**, 1639 (2021).
39. Cutillo, C. M. et al. Machine intelligence in healthcare—perspectives on trustworthiness, explainability, usability, and transparency. *NPJ Digit Med.* **3**, 47 (2020).
40. Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* **25**, 44–56 (2019). JanEpub 2019.
41. Dotson, G. A., Ryan, C. W., Chen, C., Muir, L., & Rajapakse, I. Cellular reprogramming: Mathematics meets medicine. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*. **13**, e1515 (2021).
42. Gascón, S., Masserdotti, G., Russo, G. L. & Götz, M. Direct Neuronal Reprogramming: Achievements, Hurdles, and New Roads to Success. *Cell. Stem Cell.* **21**, 18–34 (2017).
43. Rackham, O. J. et al. A predictive computational framework for direct reprogramming between human cell types. *Nat. Genet.* **48**, 331–335 (2016).
44. Wei, Q., Qian, Y., Yu, J. & Wong, C. C. Metabolic rewiring in the promotion of cancer metastasis: mechanisms and therapeutic implications. *Oncogene* **39**, 6139–6156 (2020).
45. Tan, K., Tegnér, J. & Ravasi, T. Integrated approaches to uncovering transcription regulatory networks in mammalian cells. *Genomics* **91**, 219–231 (2008).
46. Lagani, V., Triantafyllou, S., Ball, G., Tegnér, J. & Tsamardinos, I. Probabilistic Computational Causal Discovery for Systems Biology. In *Studies in Mechanobiology, Tissue Engineering and Biomaterials*. https://doi.org/10.1007/978-3-319-21296-8_3 (2016).
47. Magnusson, R. & Gustafsson, M. LiPLike: Towards gene regulatory network predictions of high certainty. *Bioinformatics*. **36**, 2522–2529. <https://doi.org/10.1093/bioinformatics/btz950> (2020).

ACKNOWLEDGEMENTS

This work was supported by the Swedish Research Council (grant 2019-04193(M.G.)), the Swedish foundation for strategic research (grant SB16-0095(R.M., M.G.)), the Center for Industrial IT (CENIIT)(R.M., M.G.), the Systems Biology Research Centre at University of Skövde under grants from the Knowledge Foundation (grant 20200014) (R.M.), and the King Abdullah University of Science and Technology (KAUST) (J.N.T). Computational resources were granted by Swedish National Infrastructure for Computing (SNIC; SNIC 2020/5-177, and LiU-2019-25). The authors would like to thank Dr. Robert Lehman for his helpful suggestions to improve this manuscript.

AUTHOR CONTRIBUTIONS

R.M. conceived the presented idea. All authors jointly designed the study and R.M. performed the computations. All authors jointly discussed the results and wrote the manuscript.

FUNDING

Open access funding provided by University of Skövde.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41540-022-00218-9>.

Correspondence and requests for materials should be addressed to Rasmus Magnusson.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022