AGEAS: Automated Machine Learning based Genetic Regulatory Element Extraction System

Jack Yu[1,2,*1,+], **Masayoshi Nakamoto**[2,+], **and Jiawang Tao**[1,*2,+]

[1]Center for Health Research, Guangzhou Institutes of Biomedicine and Health, Chinese Academy of Sciences, Guangzhou 510530, China
[2]Shenzhen Mozhou Technology Co., Ltd, Shenzhen, China
[*1]Correspondence: gyu17@alumni.jh.edu
[*2]Correspondence: tao_jiawang@gibh.ac.cn
[+]these authors contributed equally to this work

ABSTRACT

As rapid progress in sequencing technology since last decade, numerous mechanisms underlying cell functions and developmental processes have been revealed as complex regulations of gene expressions. While single-cell RNA sequencing (scRNA-Seq) made high-resolution transcriptomic view increasingly accessible, precise identification of gene regulatory network (GRN) describing cell states became achievable. However, extracting key regulatory elements, including gene regulatory pathways (GRPs), transcription factors (TFs), and downstream genes, accurately reflecting main functionality changes remain challenging. Herein, we describe AGEAS, an automated machine learning (AutoML) based genetic regulatory element extraction system that assesses importances of GRPs in differentiating sample classes, such as cell types and developmental stages. With several case studies in divergent research areas, we show that AGEAS can indeed extract informative regulatory elements and reconstruct networks referring phenotype or biological process of interest. Furthermore, wet lab experiment validations... Overall, AGEAS provides an ...

Availabilityand implementation
The AGEAS code is available at https://github.com/JackSSK/Ageas.

Introduction

TBD

Method

The basic principle of AGEAS is to find key regulatory elements associated with phenotype of interest through analyzing how well-performing classification models distinguish GRNs of sample with the phenotype from GRNs of samples without. To reconstruct sufficient GRNs for each class, RNA-Seq based expression data is segmented into subsets while each one is analogized as a pseudo-sample having discrete expression data. The pseudo-sample GRNs (psGRNs) are reconstructed accordingly; thus, classification models can be trained, evaluated, and interpreted having GRPs as input factors. With heavily weighted GRPs and associated genes repeatedly obtained from interpretations of success sample class predictions, GRNs could be formed and inferred playing important role in differentiating the studying sample classes. The overall workflow can be summarized as Figure 1. By default setting, four seperate extractor units in workflow run parallelly after data preprocessing part, and all extracted regulatory elements are used to form GRNs later combined into one atlas. Following sections describe each step of AGEAS in more depth.

Step 1:Data preprocessing

The main purpose of this step is to build pseudo-samples and reconstruct corresponding pseudo-sample GRNs (psGRNs). For each sample class, gene expression matrices (GEMs) labeled with same class label are concatenated as one comprehensive expression matrix. With comprehensive GEMs, a meta-level GRN (meta-GRN) is reconstructed in advance of psGRNs to provide generic guidance on reconstruction. In general, the workflow of this step can be represented as Figure 2.

Reconstruct meta-GRN

Firstly, genes included in the comprehensive GEMs are assessed and determined whether having potential to form informative GRPs with other genes. Commonly, differentially expressed genes (DEGs) would be considered as important factors of

studying feature. Here we apply the Mann-Whitney U rank test (MWW) implemented by *SciPy*[1] to exclude genes having indistinguishable expression level distribution across GEMs of different classes. The p-value for rejecting null hypothesis, that expression distribution underlying class 1 samples is the same as the expression distribution underlying class 2 samples, is set to 0.05 by default. Furthermore, a $\log_2$ fold change (log2FC) filter is also implemented in AGEAS. However, enabling the log2FC filter is not encouraged considering upstream TFs indirectly regulating key genes associated with feature of interest may not always have significant expression level difference between sample classes. The lof2FC filter shall mostly be used in order to decrease meta-GRN's GRP total amount in a compromising position caused by limited computational resources. After differential expression based filters, a standard deviation (σ) filter is applied to exclude genes either having low expression value or merely affected by dynamic expression status of other genes. By default, the σ threshold is set to 1.0, the lowest expression level in raw gene count matrix gained from RNA-Seq data. The threshold value should be adjusted corresponding to prior knowledge of input GEMs, for example of what normalization method was applied to the GEMs.

With candidate genes passed filters above, some gene pairs are formed and assessed potential of representing GRPs. To reduce overall computational complexity, a gene pair shall be formed with at least one TF which could be the regulatory source of GRP. If not further specified, TF list will be retrieved from integrated *TRANSFAC*[2] datasets according to the provided species information. Utilizing genetic interaction database like *GTRD*[3] and *BioGRID*[4], AGEAS checks whether the binding ability of gene pair is confirmed or not. By default, if TF being recorded to have binding site within target gene's promoter range (-1000 to +100) by Chromatin Immunoprecipitation Sequencing (ChIP-Seq) dataset retrieved from *GTRD*[3], the potential GRP gene pair will be passed to expression correlation assessment. For TFs not covered by interaction database, *GRNBoost2*[5]-like algorithm is initiated to predict potential regulatory target genes. The prediction importance threshold can either be set manually or automatically based on recorded interactions as:

$$Threshold = min(GA(M_1,t) \uplus GA(M_2,t), \frac{1}{g})$$

Here $GA()$ denotes *GRNBoost2*[5]-like algorithm; M_1 is concatenated class 1 samples GEM with genes passed filters; M_2 is concatenated class 2 samples GEM with genes passed filters; t is the TF having largest amount of recorded interaction in dataset; g is total amount of unique genes in all samples passed filters.

After potential GRP gene pairs obtained, a expression correlation filter is used to exclude gene pairs having low covariance which implies weak relationship. To assess expression correlation, AGEAS applies Pearson's Correlation coefficient[6] (PCC) whihc is one of the widely adopted methods.[7] With default setting, gene pairs can reach absolute correlation coefficient of 0.2 while keeping correspond p-value lower than 0.05 are included in meta-GRN as validated GRP.

Reconstruct psGRNs

The comprehensive GEMs is divided into few sample subsets with sliding window algorithm (SWA) to build pseudo-samples. With SWA, we can gain GEM of *i-th* pseudo-sample through:

$$SWA(i) = \left\{ x_j {}^{j+k}_{j=i*p} \right\}, j+k < l$$

Here l is number of samples in a comprehensive GEM which can be expressed as $\left\{ x_j {}^{l}_{j=0} \right\}$; k denotes window size; p denotes padding stride.

If sample amount is considerably low or imbalanced, customized window size and padding stride can be applied to generate sufficient amount of pseudo-samples for later classifier training and assessment processes. Utilizing meta-GRN, psGRNs are reconstructed with GEMs of pseudo-sample. Each GRP gene pair in meta-GRN is formed with expression data in pseudo-sample and filtered by same PCC filter in meta-GRN reconstruction process.

After all pseudo-samples have been used to reconstruct psGRN, every psGRN can be represented as matrix comprising GRPs' PCC values. The order of GRPs in each psGRN is also unified through adding GRPs included in other psGRNs with 0.0 PCC value.

Step 2: Classification model selection

Since AGEAS is not aiming to develop optimized model architectures for psGRN classification but gain insights from multiple models as divergent as possible, the main goal of this step is to select configurations of models capable to make correct psGRN classifications from provided configuration set. Regarding the fact that computational power is limited resource, portion of less efficient model configurations shall be pruned although interpreting success predictions of more classification models would lead to more comprehensive insight into sample class differences. Therefore, we apply a simple Hyperband[8]-based algorithm 1 which performs grid search for well-performing classification models with varying model training resource and pruning aggressiveness.

The model selection algorithm requires five inputs (1)R, the maximum amount of training resource, equivalent to all avaliable psGRNs (2)C, the total set of provided classification model configurations (3)I, the number of iterations for model pruning (default set as 3) (4)α_{max}, the maximum portion of R can be fed to single model (default set as 0.95) (5)k_{min}, the

Algorithm 1: Model selection algorithm

Input: R, C, I (default $I = 3$), α_{max}(default $\alpha_{max} = 0.9$), k_{min}(default $k_{min} = 0.5$)
$\alpha_{low} = \frac{1}{(2^I - 1)}$;
for $i \in \{0, 1, ..., I-1\}$ **do**
  if $i == I - 1$ **then**
    $\alpha = \alpha_{max}$;
  else
    $alpha = 2^i \alpha_{low}$;
  end
  $r = \alpha R$;
  $P = \emptyset$;
  for $c \in C$ **do**
    $p = run\_then\_evaluate(c, r, R)$;
    Append p to P;
  end
  $k = max(1 - \alpha, k_{min})$;
  $C = top\_configs(C, P, k)$;
end
Return: C

minimum portion of remaining model configurations will be kept by single pruning iteration (default set as 0.5). Furthermore, two functions are also required while need to be defined based on input model configurations:

- $run\_then\_evaluate(c, r, R)$: trains classification model initialized using configuration c for the allocated resource r, then returns prediction accuracy (ACC), the area under a receiver operating characteristic curve (AUROC)[9] score, and total cross-entropy loss (L_{CE}) calculated through predicting sample class for all psGRNs R.
- $top\_configs(C, P, k)$: takes a set of model configurations C with associated evaluation results P and returns configurations having ACC, AUROC score, or L_{CE} reaching top k portion.

By default, AGEAS initializes with 128 model configurations utilizing 9 integrated classification algorithms listed in Table 1.

Algorithm	**#λ**	**Categorical**	**Continuous**	**# Configs**
Implemented with Pytorch[10]				
Transformer	**14**	4	10	32
1D Convolutional Neural Network (1D-CNN)	**10**	2	8	32
Hybrid Convolutional Neural Network (Hybrid-CNN)	**10**	2	8	32
Gated Recurrent Unit (GRU)	**10**	4	6	4
Long Short-Term Memory (LSTM)	**11**	4	7	4
Standard Recurrent Neural Network (RNN)	**11**	5	6	4
Implemented with XGBoost[11]				
Gradient Boosted Decision Trees (GBDT)	**18**	5(1)	13(4)	16
Implemented with scikit-learn[12]				
Random Forests (RF)	**14**	5(1)	9(1)	2
Support Vector Machine (SVM)	**7**	4	3(3)	2

Table (1) Classification model algorithms integrated in AGEAS with correspond numbers of hyperparameter and preset model configurations. Categorical hyperparameters and continuous numerical hyperparameters are clarifed beside total number of hyperparameters (#λ). Conditional hyperparameters which are required for selected other hyperparameters are shown in brackets if there is any. # Configs indicates total amount of configurations AGEAS applies by default.

The general architecture designs of 1D-CNN and Hybrid-CNN are implemented referring to 1D-CNN and 2D-Hybrid-CNN applied in recent cancer type prediction study.[13] However, taking one convolution layer and adjacent max-pooling layer as a layer set, we implemented both CNN models with flexibilities on number of layer set, which is fixed as 1 in original research. An example of 1D-CNN with 2 convolution layer sets is illustrated in Figure 3.

For transformer models, considering psGRNs are already be represented by numerical data matrix while GRPs should barely have positional relationships in the matrix, the embedding layer and positional encoding layer designed for input data tokenization in standard architecture[14] are replaced with a single linear layer in AGEAS.

Step 3: Feature selection

With selected well-performing models, AGEAS already can start repetition of model training and interpretation in Step 4 to extract key GRPs. However, few uninformative GRPs in training psGRNs merely relied by any model to make classification could be pruned in advance for saving computational power. Furthermore, to prevent classification models focusing on a small group of GRPs regardless of training psGRNs, GRPs draw excessive attention shall also be seperated from psGRNs to improve extraction comprehensiveness. Thus, in this step, AGEAS iteratively train classificaion models with dynamic α_{max} portion of psGRNs and obtain feature importance scores as described in subsection below to find GRPs either scored extremely high or considerably low.

More specifically, at each iteration, GRPs having z-scores ranked as bottom b portion (default set as 0.1) are discarded. Also, an i-th ranked GRP will be seperated from psGRNs and passed to Step 5 directly if having z-score fulfilling the condition:

$$Z^{i}_{score} \geq max(Z^{thread}_{score}, \frac{Z^{i-1}_{score}}{3}, 3 \cdot IQR)$$

The Z^{thread}_{score} is input z-score threshold (default set as 3.0), and IQR stands for interquartile range calculated by the beginning of each iteration. With this criterion, AGEAS ensures only GRPs draw significantly more attention shall be selected by each iteration despite the data distribution of varying z-score scalled importance values.

By default, AGEAS iterates this feature selection step for 3 times.

Feature importance estimation

AGEAS applies concept of The Shapley value[15] for estimating importance of each input feature, equivalent to GRP of input psGRN, in any kind of classification model while making predictions. Specific Shapley value calculation or approximating methods are implemented with *SHAP*[16] and applied to different algorithms as shown in Table 2. Regarding the standard differences between feature importance estimating methods, we utilize *softmax* function to normalize feature importances and define the normalized importance calculation function as:

$$T(X) = softmax(\{F(x)\}, x \in X)$$

Here X is total set of all features, and $F(x)$ is the importance estimating method of individual classification model being interpreted. If the feature importances can be approached with internalized method $f(x)$, $F(x)$ is set as:

$$F(x) = \frac{f(x)}{\sum_{x' \in X} f(x')}$$

Otherwise, $F(x)$ is defined utilizing correctly classified input samples S, equivalent to psGRNs, with Shapley values ϕ of feature x when predicting sample s as class c_1 or class c_2:

$$F(x) = \sum_{s \in S} \frac{\left|\phi^{x}_{c_1,s}\right| + \left|\phi^{x}_{c_2,s}\right|}{2}$$

After all selected classification models M have been interpreted, we can intergrate the feature importances matrices weighted with corresponding models' L_{CE} to one matrix A as:

$$A = \left\{\sum_{m \in M} (1 - L^{m}_{CE}) T_m(x)\right\}, x \in X$$

Then, generalized importance values are obtained through z-score calculation and later sorted by descending order:

$$Z_{score} = \left\{\frac{a - \bar{A}}{\sigma_A}\right\}, a \in A$$

Step 4: Top GRP extraction

To extract GRPs can effectively define sample class differences, AGEAS iteratively initializes classification models with configurations gained from Step 2, trains them with randomly selected α_{max} portion of psGRNs scaled after Step 3, and interpret every models' correct predictions as mentioned in subsection above. At each iteration, AGEAS receives a z-score scaled feature importance matrix A and add up each GRP's score accordingly from matrix A' kept from previous iteration if there is one. Then, top a (default set as 100) ranked GRPs having z-score greater than 0.0 extracted from A are compared with GRPs extracted from A' by same setting. If less than d (default set as 0.05) portion of GRPs are distinct for GRP sets stratified from A and A', AGEAS will consider the GRP extration result from this iteration being consistent with previous one. Extraction iteration will terminate if either encountering n (default set as 3) continuous consistent result or running out of preset iteration number (default set as 10). All feature importance scores in matrix A from last iteration are divided by total extraction iteration number processed, and top a ranked GRPs are considered as key GRPs for sample class differentiation and passed to next step.

Step 5: Key network reconstruction

Analoging every GRP previously extracted or seperated with high z-score in Step 3 as a directional edge connecting two gene vertices, equivalent to a TF and a gene, in graph theory, AGEAS attempts to reconstruct a network graph representing regulatory

Algorithm	SHAP[16] Method
Transformer	Gradient Explainer
1D-CNN	Deep Explainer
Hybrid-CNN	Deep Explainer
GRU	Gradient Explainer
LSTM	Gradient Explainer
RNN	Gradient Explainer
GBDT*	Tree Explainer
RF	Tree Explainer
SVM*	Linear Explainer / Kernel Explainer

Table (2) Classifier algorithms with applicable Shapley value approximating methods. Algorithms marked with * have internalized feature importance estimating methods which will be applied with higher priority than Shapley value based methods. GBDT implemented with *XGBoost*[11] can have feature importance approximated with average weight gain at each split involving the feature. Linear SVM implemented with *scikit-learn*[12] can have the importance estimated with feature coefficient or using Linear Explainer. However, for SVM with kernel function, the feature coefficient estimation would be inappropriate, and feature importances should be approximated with Kernel Explainer.

differences between query sample classes. Since there is no guarantee on all GRP edges can be connected, some regulatory relationships between gene vertices could be missed. Hence, AGEAS utilizes meta-GRN gained from Step 1 to find GRPs which can further elucidate the regulatory relationships and adds the GRPs back to the network graph.

For a limited iteration time (default set as 1), AGEAS exhaustively search meta-GRN for TFs which can directly regulate any genes or TFs already covered in the network graph and add the returned TFs as new vertices. Next, any GRP in meta-GRN capable to connect two distinct vertices will be added if it is not covered yet. The network graph after expansion above represents the key genetic regulatory differences AGEAS extracted from input sample classes.

Results

To assess AGEAS's predictive power, we first applied AGEAS to study somatic reprogramming to induced pluripotent stem cells (iPSC) achieved in mice by the year 2006,[17] milestone discovery of cell plasticity.[18] With σ thread set to 2.0 for constraining total GRP amount in psGRNs, public scRNA-Seq based gene expression matrices (GEMs) of embryonic stem cells (ESCs) and mouse embryonic fibroblasts (MEFs) shown in Table 3 were used as input. The extraction result can be summarized as TF regulons, collections of a TF and corresponding direct regulatory targets, and TFs forming largest regulons with extracted GRPs are shown in Figure 4(a). Among TFs with sizeable regulons, Pou5f1 (Oct4), Nanog, and Sox2 are also most differentially expressed in MEF and ESC samples. Thus, we can infer these TFs as import factors closely related with cell type differences. Multiple previous studies confirmed that all of Oct4, Nanog, and Sox2 are playing important role in maintaing pluripotency[19] and inducing conversion from somatic cell to iPSC.[17,20–23]

To further address AGEAS's applicability and limitation, we applied AGEAS to three other scRNA-Seq based studies in distinct research areas. With all default settings and analytical procedure applied above, we can assess AGEAS's performance in following scenarios with public data set shown in Table 3:

- **Dopaminergic neuron generation**:

  Analyzing difference between human iPSC-derived radial glial / neuronal co-culture, as neural progenitors,[24] and tyrosine hydroxylase (TH) expressing purified dopaminergic (DA) neurons, we address the applicability of AGEAS on cell subtype differentiation problems.

  From extracted TFs shown in Figure 4(b), we hypothesize ISL1, PHOX2B, and ELF3, referring to their correspond regulon sizes and differential expression profiles, are regulating essential functions differentiating DA neurons from neural progenitors. Both ISL1 and PHOX2B were found related with DA phenotype in previous studies: ISL1 is essential for differentiation of prethalamic DA neurons,[25] and PHOX2B is connected with TH expression influencing DA neuron development and improving DA activity.[26,27] Moreover, ELF3 was found to be neuronal precursor cell marker associated with neural stem cell development,[28] consistent with it's high expression level in neuronal co-culture class.

  EBF1, which forms the largest regulon with extracted GRPs, was determined as key neuronal differentiation regulator in the ventral telencephalon, nature source of neuronal and glial cells.[29,30] Furthermore, EBF1 plays important role in mesodiencephalic DA neuron migration.[31] ASCL1, found as necessary factor regulating DA neurotransmitter selection by

original research where we retrieved scRNA-Seq data,[24] is also forming substantial regulon. Remarkably, within extracted network, ASCL1 appears to act as upstream regulator of several large TF regulons, including ISL1 and PHOX2B, in purified DA neurons.

- **Postnatal cardiomyocyte maturation**:

  To address the performance of AGEAS while analyzing cell physiological development, scRNA-Seq data of cardiomyocytes (CMs) in postnatal day 7 mice (P7) and day 28 mice (P28) are used as input for extracting key genetic regulatory elements in postnatal cardiomyocyte maturation.

  Following same regulon size and expression profile based analytical procedure, we primarily investigated three TFs shown in Figure 4(c): Sp1, Esr1 ($ER\alpha$), and Cebpb. Previous studies demonstrated that Sp1 promotes CM hypertrophy[32] and maturation of electrophysiology and Ca^{2+} handling.[33,34] Esr1 was found modulating myocardial development for postnatal cardiac growth,[35,36] and Cebpb's role as repressor of CM growth and proliferation while preventing pathological CM hypertrophy also demonstrated by several studeis.[37,38]

  Other TFs known playing important roles in CM maturation, such as Srf,[34,37] Nkx2-5,[37,39,40] and Yap1,[34,40,41] are also forming sizeable regulons in GRN atlas returned by AGEAS.

- **Hepatic stellate cell activation in CCl_4 induced liver fibrosis**:

  Here we address AGEAS's applicability on cell pathological development studies through analyzing hepatic stellate cells (HSCs) in mice liver administrated with chronic carbon tetrachloride (CCl_4) for 6 weeks and according portal fibroblasts (PF) simulating activated HSCs.

  Within TFs shown in Figure 4(d), we suppose Lhx2, Osr1, and Ebf1 are regulating HSC activation considering their correspond TF regulon sizes and expression changes. Previously, Lhx2 was found indispensable for quiescent HSCs.[42,43] Extensive studies found Osr1 related with pathological fibrogenic process in liver fibroblast.[44–46] However, little literature address Ebf1's role in liver fibrosis. One recent study reported the expression of Ebf1 increases concurrent with Acta2, marker gene of myofibroblast formation[47] which is essential for fibrosis, in wound fibroblast.[48] Also, Ebf1 was found associated with lung fibrosis.[49]

  Furthermore, Mef2c, forming largest TF regulon, has been well-documented as key regulator in HSC activation.[50–52]

To demonstrate empirically the predictive capabilities... Since Lhx2's expression pattern is strongly favoring HSC than PF, and previous study proved Lhx2 KO promotes liver fibrosis[42]... We suppose overexpression of Lhx2 could halt HSC as quiescent in order to prevent pathological fibrogenic process... In vitro experiment goes here...

Discussion

In present study, we showed AGEAS's applicability in analyzing scRNA-Seq derived data in cell reprogramming development, cell subtype differentiation . According to test cases so far, key TFs closely related with sample class difference appear to either form substantial regulon with extracted GRPs or high regulatory influence on other key genes. Without any previous knowledge on studying CSs, Gene Ontology (GO) enrichment analysis can potentially narrow down key cellular functions changed in different CSs.

However, larger scales of applicability test can further confirm AGEAS's capability on guiding frontier researches. While more computational resources become available, other classification algorithms and according interpretation methods will be added and tested with current version on larger scales.

With result achieved so far, we anticipate that AGEAS will become useful in providing insightful advices on not only learning differences between CSs and cell subtypes, but also finding key GFs to induce CS conversions unachievable so far.

Limitations... Fidelity of GRPs... Reconstruct GRN with ATAC-Seq by motif enrichment instead of ChIP-Seq... (especially considering technology like Share-Seq which can perform scRNA-Seq with scATAC-Seq in same cell simultaneously...) Apply AGEAS on tissue studies with spatial transcriptomics...

References

1. Virtanen, P. *et al.* SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods* **17**, 261–272, DOI: 10.1038/s41592-019-0686-2 (2020).
2. Matys, V. Transfac(r) and its module transcompel(r): Transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* **34**, DOI: 10.1093/nar/gkj143 (2006).
3. Kolmykov, S. *et al.* GTRD: an integrated view of transcription regulation. *Nucleic Acids Res.* **49**, D104–D111, DOI: 10.1093/nar/gkaa1057 (2020). https://academic.oup.com/nar/article-pdf/49/D1/D104/35364856/gkaa1057.pdf.

4. Stark, C. Biogrid: A general repository for interaction datasets. *Nucleic Acids Res.* **34**, DOI: 10.1093/nar/gkj109 (2006).
5. Moerman, T. *et al.* Grnboost2 and arboreto: Efficient and scalable inference of gene regulatory networks. *Bioinforma. (Oxford, England)* **35**, DOI: 10.1093/bioinformatics/bty916 (2018).
6. Song, L., Langfelder, P. & Horvath, S. Comparison of co-expression measures: Mutual information, correlation, and model based indices. *BMC Bioinforma.* **13**, DOI: 10.1186/1471-2105-13-328 (2012).
7. Liu, L.-y. D., Hsiao, Y.-C., Chen, H.-C., Yang, Y.-W. & Chang, M.-C. Construction of gene causal regulatory networks using microarray data with the coefficient of intrinsic dependence. *Bot. Stud.* **60**, DOI: 10.1186/s40529-019-0268-8 (2019).
8. Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A. & Talwalkar, A. Hyperband: A novel bandit-based approach to hyperparameter optimization (2018).
9. Hanley, J. A. & McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology* **143**, 29–36, DOI: 10.1148/radiology.143.1.7063747 (1982).
10. Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. In Wallach, H. *et al.* (eds.) *Advances in Neural Information Processing Systems 32*, 8024–8035 (Curran Associates, Inc., 2019).
11. Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794 (2016).
12. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
13. Mostavi, M., Chiu, Y.-C., Huang, Y. & Chen, Y. Convolutional neural network models for cancer type prediction based on gene expression - bmc medical genomics (2020).
14. Vaswani, A. *et al.* Attention is all you need (2017).
15. Roth, A. E. & Shapley, L. S. The shapley value : essays in honor of lloyd s. shapley. *Economica* **101**, 123 (1991).
16. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*, 4768–4777 (2017).
17. Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126**, 663–676, DOI: 10.1016/j.cell.2006.07.024 (2006).
18. Wang, H., Yang, Y., Liu, J. & Qian, L. Direct cell reprogramming: Approaches, mechanisms and progress. *Nat. Rev. Mol. Cell Biol.* **22**, 410–424, DOI: 10.1038/s41580-021-00335-z (2021).
19. Niwa, H. How is pluripotency determined and maintained? *Development* **134**, 635–646, DOI: 10.1242/dev.02787 (2007).
20. Wang, B. *et al.* Induction of pluripotent stem cells from mouse embryonic fibroblasts by jdp2-jhdm1b-mkk6-glis1-nanog-essrb-sall4. *Cell Reports* **27**, DOI: 10.1016/j.celrep.2019.05.068 (2019).
21. Shi, Y. *et al.* Induction of pluripotent stem cells from mouse embryonic fibroblasts by oct4 and klf4 with small-molecule compounds. *Cell Stem Cell* **3**, 568–574, DOI: 10.1016/j.stem.2008.10.004 (2008).
22. Yu, J. *et al.* Induced pluripotent stem cell lines derived from human somatic cells. *Science* **318**, 1917–1920, DOI: 10.1126/science.1151526 (2007).
23. Wang, Y. *et al.* Reprogramming of mouse and human somatic cells by high-performance engineered factors. *EMBO reports* **12**, 373–378, DOI: 10.1038/embor.2011.11 (2011).
24. Earley, A. M., Burbulla, L. F., Krainc, D. & Awatramani, R. Identification of ascl1 as a determinant for human ipsc-derived dopaminergic neurons. *Sci. Reports* **11**, DOI: 10.1038/s41598-021-01366-4 (2021).
25. Filippi, A., Jainok, C. & Driever, W. Analysis of transcriptional codes for zebrafish dopaminergic neurons reveals essential functions of arx and isl1 in prethalamic dopaminergic neuron development. *Dev. Biol.* **369**, 133–149, DOI: 10.1016/j.ydbio.2012.06.010 (2012).
26. Hoekstra, E. J., von Oerthel, L., van der Linden, A. J. & Smidt, M. P. Phox2b influences the development of a caudal dopaminergic subset. *PLoS ONE* **7**, DOI: 10.1371/journal.pone.0052118 (2012).
27. Fan, Y. *et al.* Transcription factors phox2a/2b upregulate expression of noradrenergic and dopaminergic phenotypes in aged rat brains. *Neurotox. Res.* **38**, 793–807, DOI: 10.1007/s12640-020-00250-9 (2020).
28. Tang, Y. *et al.* Elf a β-spectrin is a neuronal precursor cell marker in developing mammalian brain; structure and organization of the elf/β-g spectrin gene. *Oncogene* **21**, 5255–5267, DOI: 10.1038/sj.onc.1205548 (2002).
29. Garel, S., Marin, F., Grosschedl, R. & Charnay, P. Ebf1 controls early cell differentiation in the embryonic striatum. *Development* **126**, 5285–5294, DOI: 10.1242/dev.126.23.5285 (1999).

30. Faedo, A. *et al.* Differentiation of human telencephalic progenitor cells into msns by inducible expression of gsx2 and ebf1. *Proc. Natl. Acad. Sci.* **114**, DOI: 10.1073/pnas.1611473114 (2017).

31. Yin, M. *et al.* Ventral mesencephalon-enriched genes that regulate the development of dopaminergic neurons in vivo. *J. Neurosci.* **29**, 5170–5182, DOI: 10.1523/jneurosci.5569-08.2009 (2009).

32. Sun, S. *et al.* Dendropanax morbifera prevents cardiomyocyte hypertrophy by inhibiting the sp1/gata4 pathway. *The Am. J. Chin. Medicine* **46**, 1021–1044, DOI: 10.1142/s0192415x18500532 (2018).

33. Brady, M. Sp1 and sp3 transcription factors are required for trans-activation of the human serca2 promoter in cardiomyocytes. *Cardiovasc. Res.* **60**, 347–354, DOI: 10.1016/s0008-6363(03)00529-7 (2003).

34. Guo, Y. & Pu, W. T. Cardiomyocyte maturation. *Circ. Res.* **126**, 1086–1106, DOI: 10.1161/circresaha.119.315862 (2020).

35. Luo, T. & Kim, J. K. The role of estrogen and estrogen receptors on cardiomyocytes: An overview. *Can. J. Cardiol.* **32**, 1017–1025, DOI: 10.1016/j.cjca.2015.10.021 (2016).

36. Kararigas, G., Nguyen, B. T. & Jarry, H. Estrogen modulates cardiac growth through an estrogen receptor α-dependent mechanism in healthy ovariectomized mice. *Mol. Cell. Endocrinol.* **382**, 909–914, DOI: 10.1016/j.mce.2013.11.011 (2014).

37. Boström, P. *et al.* C/ebpb controls exercise-induced cardiac growth and protects against pathological cardiac remodeling. *Cell* **143**, 1072–1083, DOI: 10.1016/j.cell.2010.11.036 (2010).

38. Zou, J. *et al.* C/ebpb knockdown protects cardiomyocytes from hypertrophy via inhibition of p65-nfkb. *Mol. Cell. Endocrinol.* **390**, 18–25, DOI: 10.1016/j.mce.2014.03.007 (2014).

39. Serpooshan, V. *et al.* Nkx2.5+ cardiomyoblasts contribute to cardiomyogenesis in the neonatal heart. *Sci. Reports* **7**, DOI: 10.1038/s41598-017-12869-4 (2017).

40. Padula, S. L., Velayutham, N. & Yutzey, K. E. Transcriptional regulation of postnatal cardiomyocyte maturation and regeneration. *Int. J. Mol. Sci.* **22**, 3288, DOI: 10.3390/ijms22063288 (2021).

41. Hou, N. *et al.* Activation of yap1/taz signaling in ischemic heart disease and dilated cardiomyopathy. *Exp. Mol. Pathol.* **103**, 267–275, DOI: 10.1016/j.yexmp.2017.11.006 (2017).

42. Wandzioch, E., Kolterud, A., Jacobsson, M., Friedman, S. L. & Carlsson, L. Lhx2–/– mice develop liver fibrosis. *Proc. Natl. Acad. Sci.* **101**, 16549–16554, DOI: 10.1073/pnas.0404678101 (2004).

43. Kolterud, A., Wandzioch, E. & Carlsson, L. Lhx2 is expressed in the septum transversum mesenchyme that becomes an integral part of the liver and the formation of these cells is independent of functional lhx2. *Gene Expr. Patterns* **4**, 521–528, DOI: 10.1016/j.modgep.2004.03.001 (2004).

44. Zhou, Y. *et al.* Osr1 regulates hepatic inflammation and cell survival in the progression of non-alcoholic fatty liver disease. *Lab. Investig.* **101**, 477–489, DOI: 10.1038/s41374-020-00493-2 (2020).

45. Gupta, V., Gupta, I., Park, J., Bram, Y. & Schwartz, R. E. Hedgehog signaling demarcates a niche of fibrogenic peribiliary mesenchymal cells. *Gastroenterology* **159**, DOI: 10.1053/j.gastro.2020.03.075 (2020).

46. Matsuda, M. & Seki, E. The liver fibrosis niche: Novel insights into the interplay between fibrosis-composing mesenchymal cells, immune cells, endothelial cells, and extracellular matrix. *Food Chem. Toxicol.* **143**, 111556, DOI: 10.1016/j.fct.2020.111556 (2020).

47. Nagamoto, T., Eguchi, G. & Beebe, D. C. Alpha-smooth muscle actin expression in cultured lens epithelial cells (2000).

48. Guerrero-Juarez, C. F. *et al.* Single-cell analysis reveals fibroblast heterogeneity and myeloid-derived adipocyte progenitors in murine skin wounds. *Nat. Commun.* **10**, DOI: 10.1038/s41467-018-08247-x (2019).

49. Liu, X. *et al.* Categorization of lung mesenchymal cells in development and fibrosis. *iScience* **24**, 102551, DOI: 10.1016/j.isci.2021.102551 (2021).

50. Wang, X. *et al.* Regulation of hepatic stellate cell activation and growth by transcription factor myocyte enhancer factor 2. *Gastroenterology* **127**, 1174–1188, DOI: 10.1053/j.gastro.2004.07.007 (2004).

51. Zhang, W., Ping, J., Zhou, Y., Chen, G. & Xu, L. Salvianolic acid b inhibits activation of human primary hepatic stellate cells through downregulation of the myocyte enhancer factor 2 signaling pathway. *Front. Pharmacol.* **10**, DOI: 10.3389/fphar.2019.00322 (2019).

52. Zhang, Y.-B. *et al.* Hydroxysafflor yellow a attenuates carbon tetrachloride-induced hepatic fibrosis in rats by inhibiting erk5 signaling. *The Am. J. Chin. Medicine* **40**, 481–494, DOI: 10.1142/s0192415x12500371 (2012).

Author contributions statement

J.Y.: Methodology, Software, Writing- Original draft preparation, Project administration **M.N.**: Methodology, Software **J.T.**: Writing - Review & Editing, Project administration

Additional information

All scRNA-Seq datasets are retrieved from Gene Expression Omnibus(GEO) as described in the Table below.

Sample Class	Accession number
GSE103221	
MEF	GSM3629847
ESC	GSM3629848
GSE137720	
CCl_4 a6w hsc	GSM4085625
CCl_4 a6w pf	GSM4085627
GSE185275	
glial/neuronal co-culture	GSM5609927
purified DA neurons	GSM5609930
GSE156482	
p7 CM	GSM4732221
p28 CM	GSM4732225

Table (3)

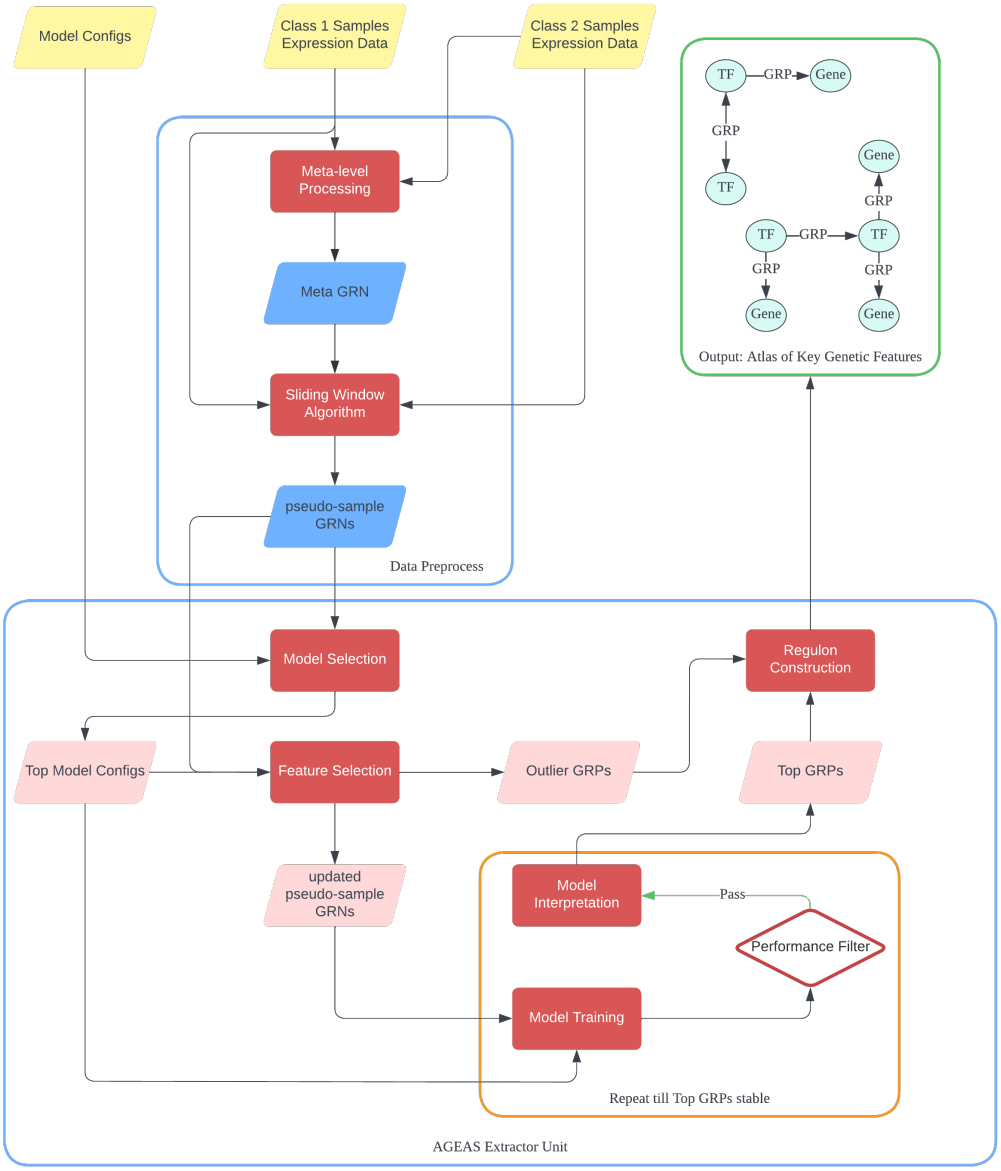

Figure (1) The overall workflow of AGEAS: ***(1)*** Reconstruct meta-level GRN (meta-GRN) with expression data of all samples. ***(2)*** Build pseudo-samples with sliding window algorithm and reconstruct GRNs with GRPs identified in meta-GRN accordingly. ***(3)*** Select best performing classifiers in predicting class labels of pseudo-sample GRNs (psGRNs). ***(4)*** Interpret how top models make classifications and gradually exclude GRPs with low weights or outlier-level high weights. ***(5)*** Repeatedly train classifiers with differnt set of psGRNs as training data to extract GRPs frequently ranked as top important features for decision. ***(6)*** Reconstruct GRNs with extracted GRPs and GRPs excluded as significant outliers.

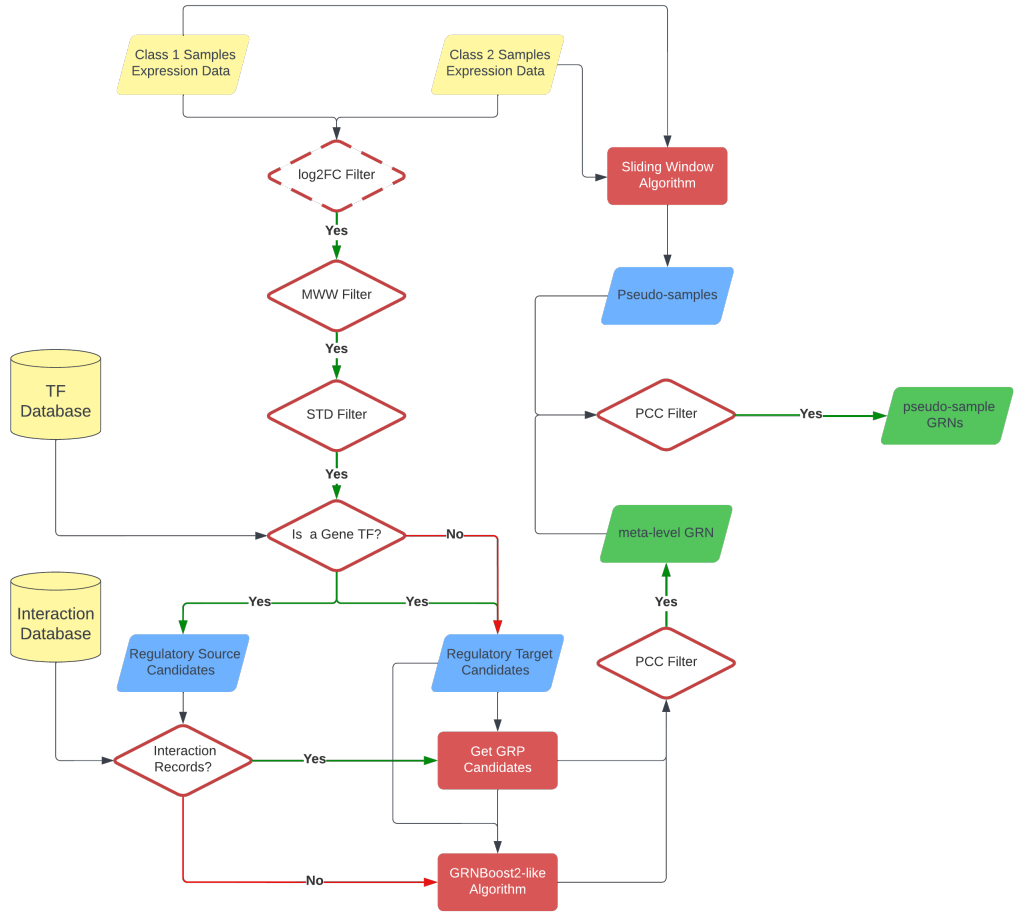

Figure (2) Workflow to reconstruct meta-GRN and psGRNs. ***(1)*** Filter genes from GEMs with log2FC filter(optional), MWW filter, and σ filter. ***(2)*** Find candidate GRP gene pairs from either interaction database or predictions made by *GRNBoost2*[5]-like algorithm. ***(3)*** Filter candidate GRP gene pairs with PCC filter and reconstruct meta-GRN with validated GRPs. ***(4)*** Generate pseudo-samples with SWA. ***(5)*** Utilize GRPs in meta-GRN as generic guidance to form candidate GRPs for pseudo-samples. ***(6)*** Filter every candidate GRP for all pseudo-samples with PCC filter and reconstruct psGRNs with validated GRPs.

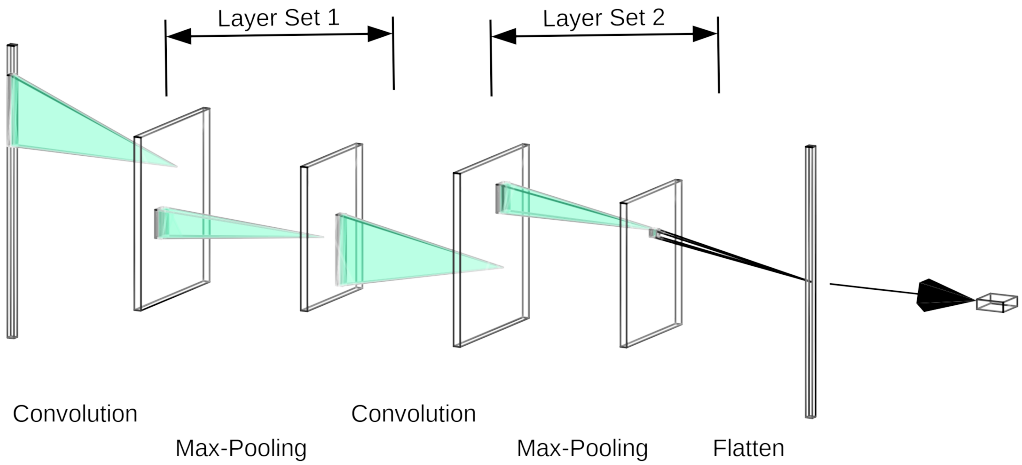

Figure (3) 1D-CNN with 2 convolution layer set.

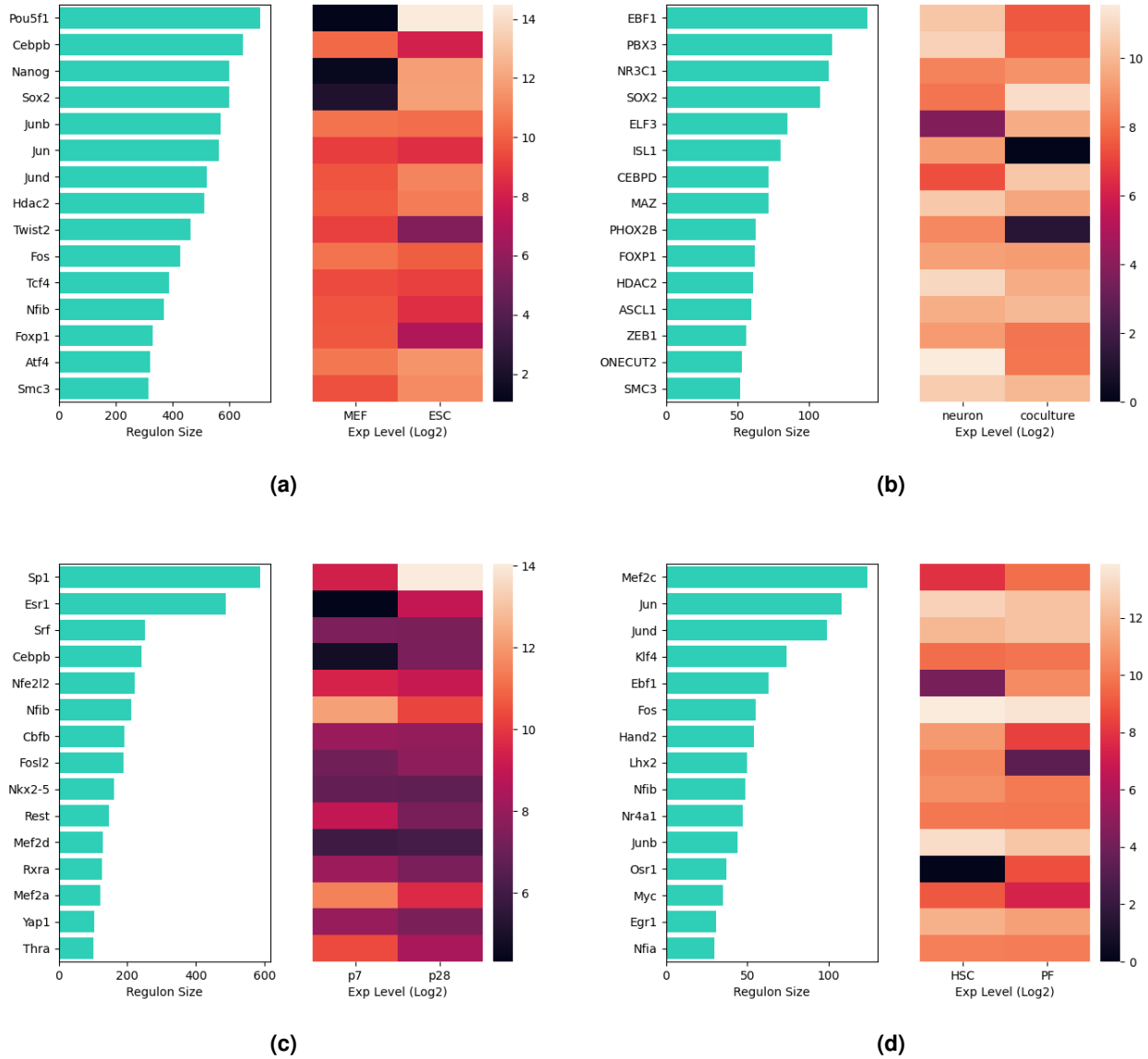

Figure (4) TFs forming largest regulons with GRPs extracted by AGEAS. Each TF has regulon size indicating amount of direct regulatory genes and log_2 gene expression levels in binary sample classes. ***(a)*** Mouse embryonic fibroblast vs. Embryonic stem cell ***(b)*** Purified dopaminergic neuron vs. Radial glial/neuronal co-culture ***(c)*** 7 days postnatal cardiomyocyte vs. 28 days postnatal cardiomyocyte ***(d)*** Hepatic stellate cell vs. Portal fibroblast (both after 6 weeks of CCl4 administration)