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. Since single-cell RNA sequencing (scRNA-seq) made high-resolution transcriptomic view increasingly accessible, precise identification of gene regulatory network (GRN) describing cell types and cell states became achievable. However, extracting key regulatory elements, including gene regulatory pathways (GRPs), transcription factors (TFs), and targetomes, that accurately and completely reflects functionality changes in biological phenomena remains challenging. Herein, we describe AGEAS, an semi-supervised automated machine learning (AutoML) based genetic regulatory element extraction system that assesses importances of GRPs in resulting biological phenomena, such as cell type differentiation, physiological and pathological development, and reconstructs GRNs with extracted important GRPs for comprehensive inference. With several case studies in divergent research areas, we show that AGEAS can indeed extract informative regulatory elements and reconstruct networks to indicate regulatory changes in biological phenomena of interest.

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

Introduction

As high-resolution sequencing technologies become increasingly applicable and accessible, an efficient and robost analytical system capable of extracting key genetic features responsible for cell types and cell states difference with limited prior biological knowledge also become highly demanded. Several methodologies like SCENIC[1] already demonstrated informativeness and robustness in studying cellular phenotypes with GRN analysis. Regulons, collections of a TF and corresponding targetomes, can successfully address cell identities with more comprehensive information compared with differential expressing genes (DEGs).[1,2] Even though limited computational methods are capable of completely and precisely extracting key genetic regulatory elements in biological process of interest, Mogrify[3] and CellNet[4] have demonstrated GRN based methods can help to analyze cell type differentiation and to implement cell reprogramming. However, both methods may require large scale of additional background data to be applicable in other studies such as analyzing physiological or pathological development of selected cell type. The scREMOTE[5] published in 2022 is a potential method to extract generalized regulatory elements since limited background data is required with sequencing data representing sample classes. But prior knowledge on key TFs and marker genes is indispensable. A regulation trajectory of DEGs or regulons showing significant activity difference after GRN reconstruction would be able to reveal the developmental pathways as generalized approach. Due to general gene regulatory nature of TFs, numerous noisy signals are expected. Therefore, here we present AGEAS, an algorithm with feasibility in extracting key regulatory elements for biological phenomena of interest and can potentially promote further discoveries, even with scarce prior knowledge.

Several reasons motivated us to develop semi-supervised AutoML based method. Firstly, accurate labeling of regulatory elements' relatedness in biological phenomena would require extensive prior knowledge may be unavailable. While unsupervised clustering can potentially extract factors playing important roles, this approach may focus on specific aspects with regulatory elements showing similar patterns and result in losing comprehensiveness. As a result, we chose to apply biological phenomena associating sample classes, such as cell types and cell states which should be easily retrievable, as labels for classification models to predict with masked GRN as input and then interpret success predictions to extract key regulatory elements. Secondly,

it is also hard to guarantee comprehensiveness when developing a generalized classification model in differentiating GRNs, considering that capturing few significant differences would be sufficient for the model to reach outstanding accuracy. Hence, an extraction system with multiple independent classification models implemented with various algorithms should be helpful to recover from potential comprehensiveness loss. Thirdly, the process of interpreting all success prediction made by every classification model and integrating corresponding interpretation results may be computational expensive. Thus, applying an AutoML-based model and feature selection procedure that can effectively decrease interpretation cost would be necessary to increase feasibility of the method.

To evaluate the performance of AGEAS, we primarily applied it on studying somatic cell reprogramming process and further addressed performance in analyzing 3 other biological processes in cell subtype differentiation, physiological and pathological development.

Method

The basic principle of AGEAS is to find key regulatory elements associated with biological phenomena of interest through analyzing how well-performing classification models distinguish GRNs of sample with the phenomena from those 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 with GRPs as input factors. With heavily weighted GRPs and corresponding genes repeatedly obtained from interpretations of successful sample class predictions, GRNs could be formed and potentially play an important role in differentiating the studying sample classes. The overall workflow can be summarized in Figure 1. By default, four separate extractor units in workflow run in parallel after data preprocessing part in order to increase output stability considering the stochastic nature of AGEAS, and all extracted regulatory elements are used to form GRNs, which will later be combined into one atlas. The 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 to reconstruct corresponding pseudo-sample GRNs (psGRNs). For each sample class, gene expression matrices (GEMs) with the same class label are concatenated as one comprehensive expression matrix. With comprehensive GEMs, a meta-level GRN (meta-GRN) is reconstructed before psGRNs to provide generic guidance on reconstruction. In general, the workflow of this step can be summarized in 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, DEGs are considered as important factors of studying phenomena. Here we apply the Mann-Whitney U rank test (MWW) implemented by *SciPy*[6] to exclude genes with indistinguishable expression level distribution across GEMs of different classes. The p-value for rejecting null hypothesis, that expression profile underlying class 1 samples is the same as the expression profile underlying class 2 samples, is set to 0.05 by default. Furthermore, a $\log_2$ fold change (log2FC) filter is implemented in AGEAS. However, enabling the log2FC filter is not encouraged, considering that upstream TFs indirectly regulate key genes associated with phenomena of interest may not be significantly differential expressed. The lof2FC filter shall mostly be used to decrease meta-GRN's total degree in a compromising position caused by limited computational resources. After differential expression based filters, a standard deviation (σ) filter is applied to exclude genes with low expression level 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 based on prior knowledge of input GEMs, for example, the normalization method applied to the GEMs.

With candidate genes passing filters above, some gene pairs are formed and evaluated by the 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, a TF list will be retrieved from integrated *TRANSFAC*[7] dataset according to the provided species information. Utilizing genetic interaction database like *GTRD*[8] and *BioGRID*[9], AGEAS checks whether the binding ability of gene pair is confirmed or not. By default, if the TF recorded has at least one binding site within target gene's promoter range (-1000 to +100) by Chromatin Immunoprecipitation Sequencing (ChIP-seq) dataset retrieved from *GTRD*[8], the potential GRP gene pair will be passed to expression correlation assessment. For TFs not covered by interaction database, *GRNBoost2*[10]-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:

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

Here $GA()$ denotes *GRNBoost2*[10]-like algorithm; M_1 denotes concatenated class 1 samples GEM with genes passed filters; M_2 denotes concatenated class 2 samples GEM with genes passed filters; t denotes the TF having largest amount of recorded interaction in dataset; g denotes total amount of unique genes in all samples passed filters.

After potential GRP gene pairs are obtained, an expression correlation filter is used to exclude gene pairs with low covariance. To assess expression correlation, AGEAS applies Pearson's Correlation coefficient[11] (PCC), one of the widely adopted methods.[12] With default setting, gene pairs can reach absolute correlation coefficient of 0.2 and 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:

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

Here l is number of samples in a comprehensive GEM which can be expressed as $\left\{ x_j{}_{j=0}^{l} \right\}$; 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[13]-based algorithm 1 which performs grid search for well-performing classification models with varying model training resource and pruning aggressiveness.

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

The model selection algorithm requires five inputs (1)R, the maximum amount of training resource, equivalent to all available 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 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)[14] 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[15]				
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[16]				
Gradient Boosted Decision Trees (GBDT)	**18**	5(1)	13(4)	16
Implemented with scikit-learn[17]				
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 clarified 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.[18] 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[19] 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 separated from psGRNs to improve extraction comprehensiveness. Thus, in this step, AGEAS iteratively train classification 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 separated 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 scaled importance values.

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

Feature importance estimation

AGEAS applies concept of The Shapley value[20] 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*[21] 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 importance 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 importance 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 integrate the feature importance 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:

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

Algorithm	SHAP[21] 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*[16] can have feature importance approximated with average weight gain at each split involving the feature. Linear SVM implemented with *scikit-learn*[17] 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 importance should be approximated with Kernel Explainer.

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 separated 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 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 predictive power of utilizing extraction result of AGEAS, we first applied AGEAS to study somatic reprogramming to induced pluripotent stem cells (iPSC) achieved in mice by the year 2006,[22] milestone discovery of cell plasticity.[23] With σ thread set to 2.0 for constraining total GRP amount in psGRNs, public scRNA-seq based GEMs of embryonic stem cells (ESCs) and mouse embryonic fibroblasts (MEFs) shown in Table 4 were used as input. The extraction result can be summarized as TF regulons, and TFs having highest regulatory degree in all extracted GRPs are marked as top TFs (Figure 4a).

Among top TFs, Pou5f1 (Oct4) has more association with GRPs extracted as important differential regulatory element between MEF and ESC than any other TFs while forming largest regulon and having expression pattern significantly favoring ESC. Thus, we can infer Pou5f1 is key regulatory element associated with ESC identity. Extensive studies already addressed Pou5f1's important role in pluripotency maintenance.[24–26] Moreover, to implement cell reprogramming to iPSC, Nanog and Sox2 are also noteworthy since they are closely interacting with Pou5f1 in extracted network, having high regulatory degrees, and differently expressed in ESC. Multiple previous studies confirmed that all of Oct4, Nanog, and Sox2 are playing important role to induce cell conversion from somatic cell to iPSC.[22,27–30] Cebpb, Hdac2, and Smc3, extracted as top TFs frequently acting as regulatory source in extracted important GRPs and directly affecting Pou5f1 expression, were also reported with relevant functions.[31–33]

Therefore, we suppose investigating TFs acting as common regulatory sources in extracted important differential GRPs and correspond regulons can effectively study differences between cell types cell states. 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 describe in Method section above, we can assess AGEAS's performance in following scenarios with public data set (Table 4):

- **Dopaminergic neuron generation**:

  Analyzing difference between human iPSC-derived radial glial / neuronal co-culture, as neural progenitors,[34] and tyrosine hydroxylase (TH) expressing purified dopaminergic neurons (DANs), we address the applicability of AGEAS on cell subtype differentiation problems.

  From extracted TFs (Figure 4b), we hypothesize ISL1, ELF3, and PBX3 playing important role in DAN differentiation due to their high regulatory degree in extracted important GRPs. Previously, ISL1 was determined essential for differentiation of prethalamic DANs,[35] and ELF3 was found to be neuronal precursor cell marker associated with neural stem cell development,[36] consistent with it's high expression level in neuronal co-culture samples. Moreover, a DAN development study reported PBX3 being required for correct differentiation of neuroblast, product of radial glial cell, into midbrain DAN and survival of midbrain DAN.[37]

  Through stratification of all extracted GRPs, we only found HDAC2 and RBPJ as potential upstream regulatory elements of all ISL1, ELF3, and PBX3. Few reports demonstrated HDAC2 is associated with neurogenesis of radial glial cells.[38,39] Nevertheless, HDAC2's role in DAN differentiation is yet to be clarified. RBPJ was reported essential for DAN survival and affecting DAN development by regulating ASCL1 which is an essential factor in neurogenesis.[40,41] Original research where we retrieved scRNA-seq data also demonstrated ASCL1 as necessary factor regulating DA neurotransmitter selection.[34] Although AGEAS extracts ASCL1 as one of top TFs; however, no direct regulatory relationship between RBPJ and ASCL1 is identified.

- **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.

  As indicated by extracted differential GRPs, we primarily investigated Sp1 and Esr1 ($ER\alpha$) which also have top regulatory degrees and considerably expression changes (Figure 4c). Previous studies demonstrated that Sp1 promotes CM hypertrophy[42] and maturation of electrophysiology, Ca^{2+} handling.[43,44] Esr1 was found modulating myocardial development for postnatal cardiac growth.[45,46]

  Beside inter-regulation between Sp1 and Esr1, 3 other top TFs (Srf, Cebpb, Rest) and 2 TFs forming relatively small regulons (Gata4, Prox1), are expected to have reversible regulation with both Sp1 and Esr1 based on extraction network. All 5 TFs were found having CM maturation related functions. (Table 3)

Gene	Function
Srf	Regulate sarcomere genes and broadly impact almost every aspect of CM maturation[44]
Cebpb	Repress CM proliferation and hypertrophy[47,48]
Rest (Nrsf)	Repress CM hypertrophy; Prevent dilated cardiomyopathy[49]
Prox1	Repress CM hypertrophy; Prevent dilated cardiomyopathy[50]
Gata4	Promote CM proliferation and hypertrophy[51]

Table (3) TFs interacting with both Sp1 and Esr1 order by extracted regulon size.

- **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.

  Based on extracted important GRPs, we identified Mef2c, Jun, and Jund as potential representatives of key regulatory elements in HSC activation (Figure 4d). As TF extracted with most important GRPs, Mef2c has been well-documented as key regulator in HSC activation.[52–54] Extensive studies also found Jun and Jund, both being functional components of the AP1 TF complex, essential for fibrosis-related HSC activation and profibrogenic process.[55–58]

  Analyzing TF regulons of Mef2c, Jun, and Jund, we suppose Mef2c is acting as targetome of Jund and inter-regulatory element of Jun. While no regulatory relationship between Jund and Jun determined, 3 extracted top TFs (Lhx2, Nfib, Junb) can potentially influence expressions of both Jund and Jun. Since Junb is also functional component of AP1, we infer it associated with HSC activation in a similar manner. Furthermore, Lhx2 was found indispensable for quiescent HSCs.[59,60] Although, little literature addressed Nfib's role in HSC, Nfib was reported acting as anti-apoptotic gene for cell proliferation during CCl_4 induced liver damage.[61]

Conclusion

With result achieved so far, AGEAS shows robustness in extracting key regulatory elements for several biological researches. Key TFs inferred by extracted important GRPs and regulon analysis in all 4 *in silico* cases were confirmed informative to reveal main functionalities responsible for biological process of interest. Furthermore, the direct extractions of ISL1 in DAN differentiation and Jund in HSC activation as regulatory source in top amount of important GRPs can demonstrate that AGEAS is not focusing solely on TF regulon sizes in reconstructed meta-GRN nor expression differences between sample classes. Overall, we anticipate that AGEAS will become useful in providing insightful regulatory information and promoting pioneer researches in revealing complicate mechanism behind biological phenomena.

Discussion

Considering the rapid development of both sequencing technology and machine learning algorithms, we implemented AGEAS with modular design. In data preprocessing part, the meta-GRN reconstruction section can be replaced with other methods to further increase fidelity of regulatory relationships or integrating more information such as cis-regulatory elements (CREs) accessibility which can be obtained from Assay for Transposase-Accessible Chromatin Sequencing (ATAC-seq) data. Moreover, CRE accessibility information combining with motif enrichment analysis is proved capable to identity potential regulatory pathways replacing ChIP-seq based dataset AGEAS utilizing in present study.[5] Refine meta-GRN reconstruction method in consideration of having input data from simultaneous scRNA-seq and single cell ATAC-seq (scATAC-seq), such as SHARE-seq published in late 2020,[62] can potentially further improve applicability of AGEAS; however, accessibility of corresponding datasets will be essential for performance assessment and method development. Similarly, initial set of classificaion models is also subject to be changed responding to future computational studies.

Also, it is important to note that AGEAS can be applied with GEMs normalized with different methods on user's choice as well as derived from other sequencing technologies such as bulk RNA-seq or spatial sequencing in revealing tissue differences. Replacing transcriptiomic data with proteomic data may also be achievable to analysis protein interaction networks. However, larger scale of applicability tests shall be performed in advance.

Due to the stochastic nature of AGEAS, extraction results of AGEAS are not expected to stay identical for repeated applications with exactly same inputs. Main reason underlying the inconsistency is caused by how AGEAS attempt to obtain top performing models for later prediction interpretations. Repeated random subsets of psGRNs using in model training and selection are expected to vary for each application hence resulting in different performance of classification models and ranking results. As a result, the importance values of GRPs might differ and slightly affect top TFs extracted later.

Two possible workarounds of inconsistent issue would be: (i) manually set random seed for every process influencing subset selection. (ii) set multiple extractor unit and combine extraction results. In present study, we found n = 4 units yielding relatively stable results, but this value could vary based on application scenario.

References

1. Aibar, S. *et al.* Scenic: Single-cell regulatory network inference and clustering. *Nat. Methods* **14**, 1083–1086, DOI: 10.1038/nmeth.4463 (2017).
2. Van de Sande, B. *et al.* A scalable scenic workflow for single-cell gene regulatory network analysis. *Nat. Protoc.* **15**, 2247–2276, DOI: 10.1038/s41596-020-0336-2 (2020).
3. Rackham, O. J. *et al.* A predictive computational framework for direct reprogramming between human cell types. *Nat. Genet.* **48**, 331–335, DOI: 10.1038/ng.3487 (2016).
4. Cahan, P. *et al.* Cellnet: Network biology applied to stem cell engineering. *Cell* **158**, 903–915, DOI: 10.1016/j.cell.2014.07.020 (2014).
5. Tran, A., Yang, P., Yang, J. Y. & Ormerod, J. T. Scremote: Using multimodal single cell data to predict regulatory gene relationships and to build a computational cell reprogramming model. *NAR Genomics Bioinforma.* **4**, DOI: 10.1093/nargab/lqac023 (2022).
6. 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).
7. Matys, V. Transfac(r) and its module transcompel(r): Transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* **34**, DOI: 10.1093/nar/gkj143 (2006).
8. 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.
9. Stark, C. Biogrid: A general repository for interaction datasets. *Nucleic Acids Res.* **34**, DOI: 10.1093/nar/gkj109 (2006).
10. 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).
11. 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).
12. 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).
13. Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A. & Talwalkar, A. Hyperband: A novel bandit-based approach to hyperparameter optimization (2018).
14. 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).
15. 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).
16. 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).
17. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
18. 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).
19. Vaswani, A. *et al.* Attention is all you need (2017).
20. Roth, A. E. & Shapley, L. S. The shapley value : essays in honor of lloyd s. shapley. *Economica* **101**, 123 (1991).
21. 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).
22. 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).
23. 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).

24. Niwa, H. How is pluripotency determined and maintained? *Development* **134**, 635–646, DOI: 10.1242/dev.02787 (2007).

25. Shi, G. & Jin, Y. Role of oct4 in maintaining and regaining stem cell pluripotency. *Stem Cell Res. &; Ther.* **1**, DOI: 10.1186/scrt39 (2010).

26. PAN, G. J., CHANG, Z. Y., SCHÖLER, H. R. & PEI, D. Stem cell pluripotency and transcription factor oct4. *Cell Res.* **12**, 321–329, DOI: 10.1038/sj.cr.7290134 (2002).

27. 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).

28. 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).

29. Yu, J. *et al.* Induced pluripotent stem cell lines derived from human somatic cells. *Science* **318**, 1917–1920, DOI: 10.1126/science.1151526 (2007).

30. 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).

31. Chronis, C. *et al.* Cooperative binding of transcription factors orchestrates reprogramming. *Cell* **168**, DOI: 10.1016/j.cell.2016.12.016 (2017).

32. Jamaladdin, S. *et al.* Histone deacetylase (hdac) 1 and 2 are essential for accurate cell division and the pluripotency of embryonic stem cells. *Proc. Natl. Acad. Sci.* **111**, 9840–9845, DOI: 10.1073/pnas.1321330111 (2014).

33. Nitzsche, A. *et al.* Rad21 cooperates with pluripotency transcription factors in the maintenance of embryonic stem cell identity. *PLoS ONE* **6**, DOI: 10.1371/journal.pone.0019470 (2011).

34. 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).

35. 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).

36. 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).

37. Villaescusa, J. C. *et al.* A pbx1 transcriptional network controls dopaminergic neuron development and is impaired in parkinson's disease. *The EMBO J.* **35**, 1963–1978, DOI: 10.15252/embj.201593725 (2016).

38. Jawerka, M. *et al.* The specific role of histone deacetylase 2 in adult neurogenesis. *Neuron Glia Biol.* **6**, 93–107, DOI: 10.1017/s1740925x10000049 (2010).

39. Tang, T. *et al.* Hdac1 and hdac2 regulate intermediate progenitor positioning to safeguard neocortical development. *Neuron* **101**, DOI: 10.1016/j.neuron.2019.01.007 (2019).

40. Shi, M. *et al.* Notch-rbpj signaling is required for the development of noradrenergic neurons in mouse locus coeruleus. *J. Cell Sci.* DOI: 10.1242/jcs.102152 (2012).

41. Toritsuka, M. *et al.* Regulation of striatal dopamine responsiveness by notch/rbp-j signaling. *Transl. Psychiatry* **7**, DOI: 10.1038/tp.2017.21 (2017).

42. 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).

43. 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).

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

45. 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).

46. 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).

47. 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).

48. 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).

49. Kuwahara, K. Nrsf regulates the fetal cardiac gene program and maintains normal cardiac structure and function. *The EMBO J.* **22**, 6310–6321, DOI: 10.1093/emboj/cdg601 (2003).

50. Petchey, L. K. *et al.* Loss of prox1 in striated muscle causes slow to fast skeletal muscle fiber conversion and dilated cardiomyopathy. *Proc. Natl. Acad. Sci.* **111**, 9515–9520, DOI: 10.1073/pnas.1406191111 (2014).

51. 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).

52. 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).

53. 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).

54. 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).

55. Smart, D. E. *et al.* Jund is a profibrogenic transcription factor regulated by jun n-terminal kinase-independent phosphorylation. *Hepatology* **44**, 1432–1440, DOI: 10.1002/hep.21436 (2006).

56. Smart, D. E. *et al.* Jund regulates transcription of the tissue inhibitor of metalloproteinases-1 and interleukin-6 genes in activated hepatic stellate cells. *J. Biol. Chem.* **276**, 24414–24421, DOI: 10.1074/jbc.m101840200 (2001).

57. Yoshida, K. *et al.* Transforming growth factor-β and platelet-derived growth factor signal via c-jun n-terminal kinase-dependent smad2/3 phosphorylation in rat hepatic stellate cells after acute liver injury. *The Am. J. Pathol.* **166**, 1029–1039, DOI: 10.1016/s0002-9440(10)62324-3 (2005).

58. Mann, D. A. Transcriptional regulation of hepatic stellate cell activation. *Gut* **50**, 891–896, DOI: 10.1136/gut.50.6.891 (2002).

59. 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).

60. 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).

61. Roy, S. *et al.* Mir-1224 inhibits cell proliferation in acute liver failure by targeting the antiapoptotic gene nfib. *J. Hepatol.* **67**, 966–978, DOI: 10.1016/j.jhep.2017.06.007 (2017).

62. Ma, S. *et al.* Chromatin potential identified by shared single-cell profiling of rna and chromatin. *Cell* **183**, DOI: 10.1016/j.cell.2020.09.056 (2020).

Author contributions statement

J.Y.: Methodology, Software, Writing- Original draft preparation, Project administration **M.N.**: Methodology, Software **J.T.**: Methodology, Writing- Original draft preparation, 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 DANs	GSM5609930
GSE156482	
p7 CM	GSM4732221
p28 CM	GSM4732225

Table (4)

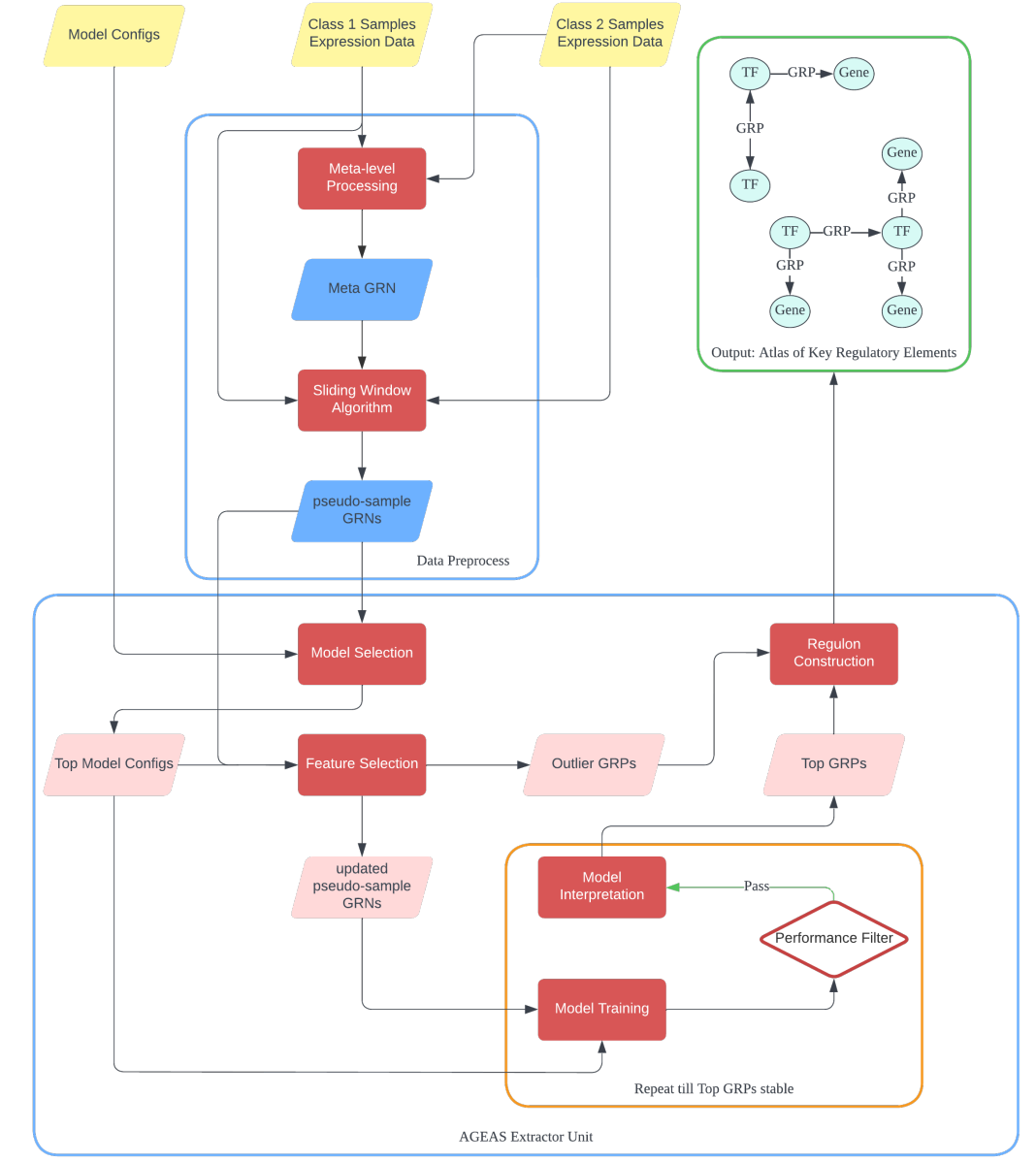

Figure (1) The overall workflow of AGEAS: ***(1)*** Reconstruct meta-level GRN (meta-GRN) with expression data of all samples. ***(2)*** Build pseudo-samples with sliding window algorithm and reconstruct GRNs with GRPs identified in meta-GRN accordingly. ***(3)*** Select best performing classifiers in 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 different 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.

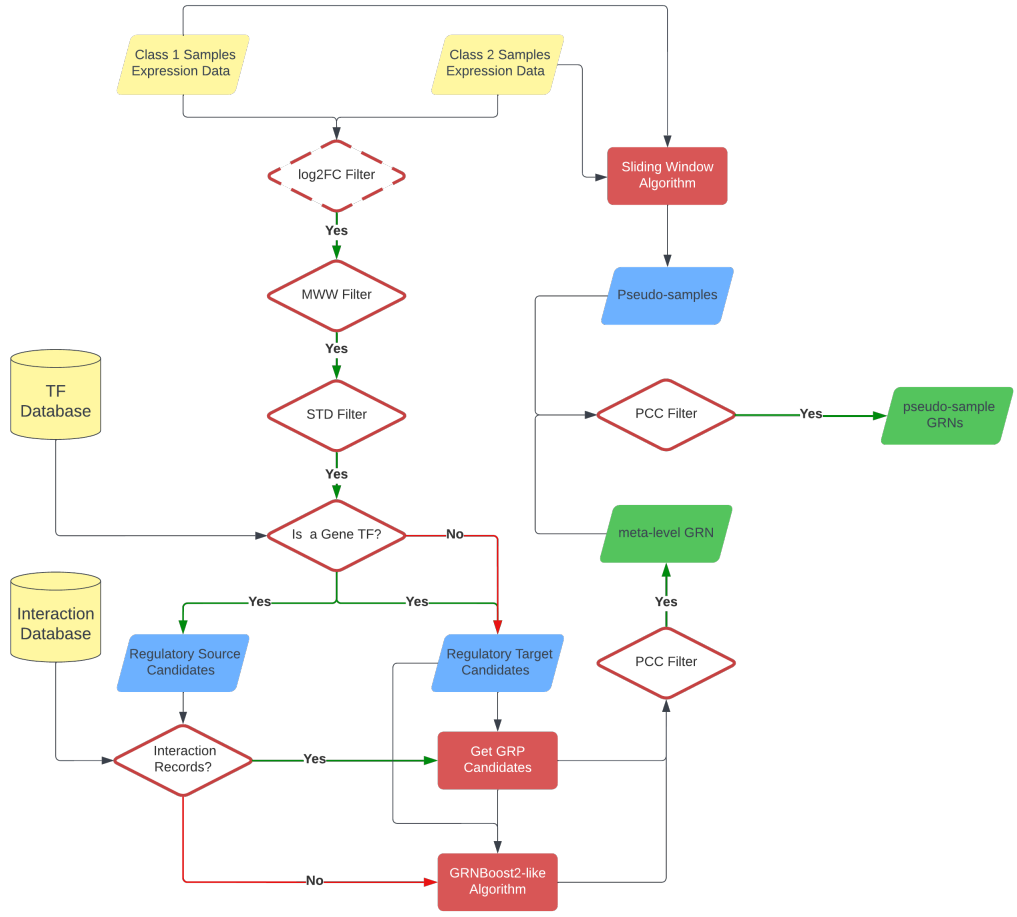

Figure (2) Workflow to reconstruct meta-GRN and psGRNs. ***(1)*** Filter genes from GEMs with log2FC filter(optional), MWW filter, and σ filter. ***(2)*** Find candidate GRP gene pairs from either interaction database or predictions made by *GRNBoost2*[10]-like algorithm. ***(3)*** Filter candidate GRP 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.

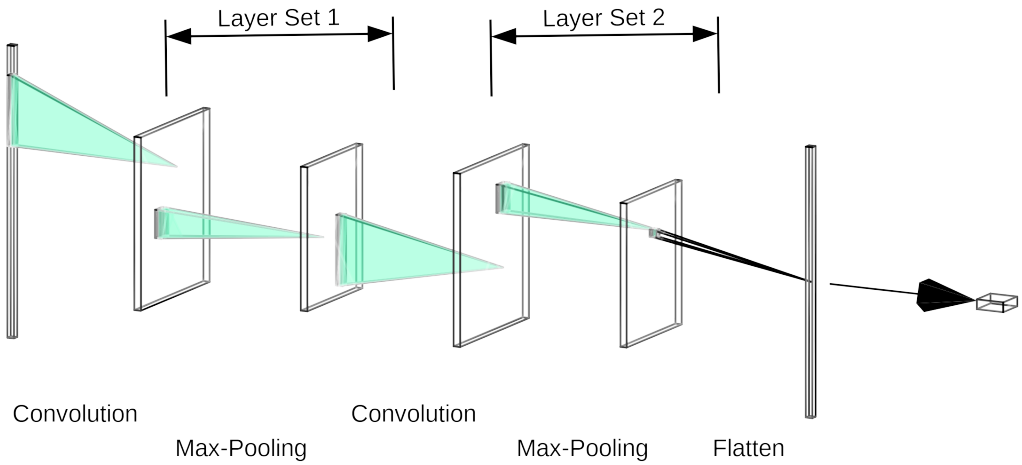

Figure (3) 1D-CNN with 2 convolution layer set.

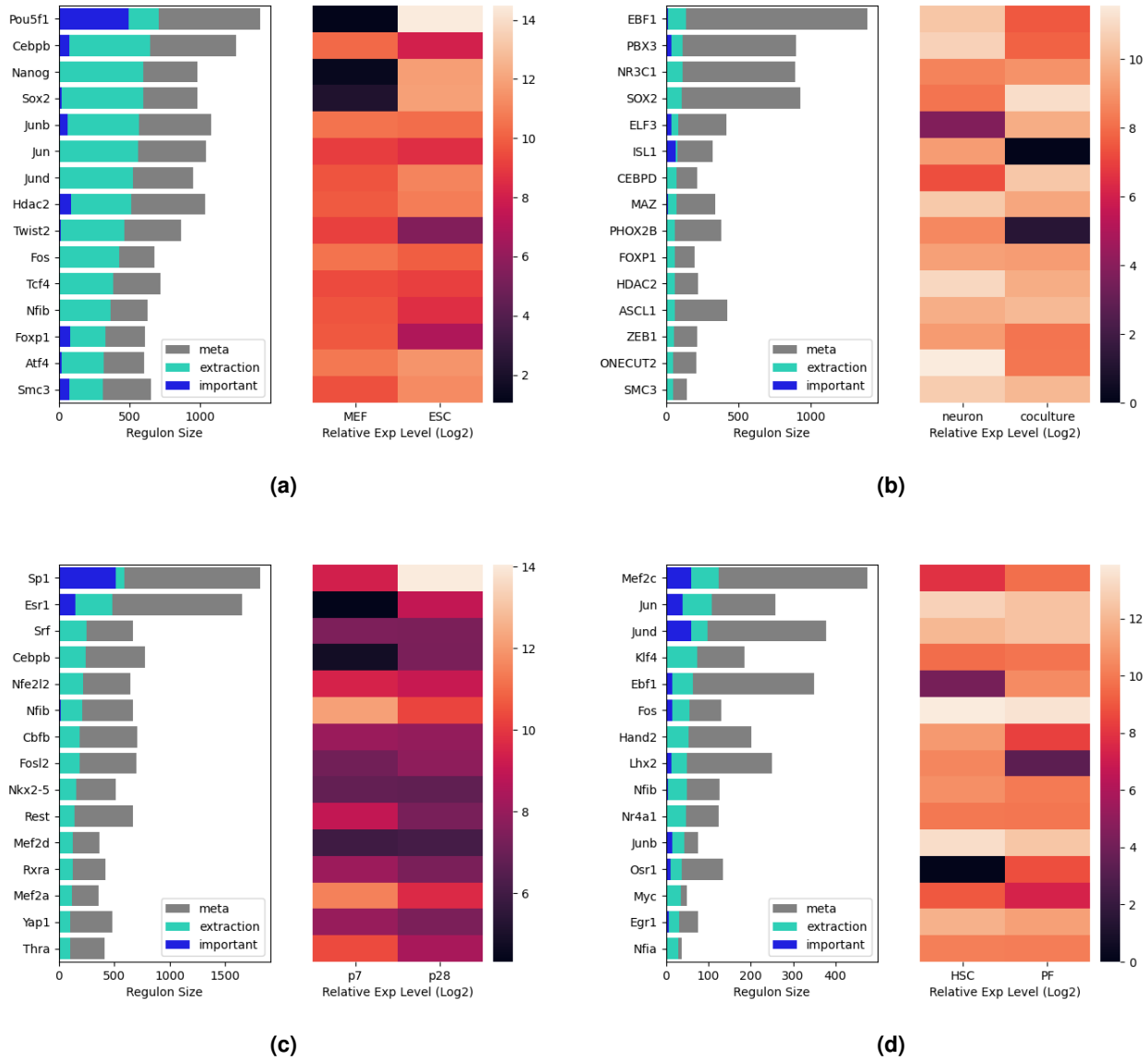

Figure (4) Top 15 TFs ranked by regulatory degrees in all GRPs extracted by AGEAS. ***(a)*** Mouse embryonic fibroblast vs. Embryonic stem cell ***(b)*** Purified dopaminergic neuron vs. Radial glial/neuronal co-culture ***(c)*** 7 days postnatal cardiomyocyte vs. 28 days postnatal cardiomyocyte ***(d)*** Hepatic stellate cell vs. Portal fibroblast (both after 6 weeks of CCl4 administration)