communications chemistry

A Nature Portfolio journal

Article

https://doi.org/10.1038/s42004-026-01991-y

Simulation-guided chemical direct reprogramming informed by temporal cellular conversion processes at the single-cell level

Check for updates

Ryoku Ito[1,4], Momoko Hamano[1,4] ✉, Ryota Kawasaki[1], Hikaru Watanabe[1], Arisa Matsuo[1] & Yoshihiro Yamanishi[2,3] ✉

Direct reprogramming (DR) involves inducing cell differentiation by converting somatic cells directly into target cells without bypassing induced pluripotent stem cells. Small molecule–induced DR reduces tumorigenic risk relative to transcription factor–triggered DR. However, identifying small molecules for DR through experimental exploration alone remains challenging. This study presents a computational method, SuperDIRECTEUR, to predict small molecules for DR using single-cell temporal transcriptome data. The cellular conversion processes in DR were simulated by calculating the RNA velocity in each cell. These processes were then classified into multiple stages (i.e., primal, immature, and mature stages), and a variant of simulated annealing was employed to search for the combination of small molecules that mimic stage-specific gene expression patterns during DR. We applied SuperDIRECTEUR to identifying candidate small molecules for DR converting mouse embryonic fibroblasts to induced neurons, and were able to reproduce experimentally verified and functionally related molecules inducing the corresponding conversions in a temporal stage-specific manner. The target proteins of the predicted small molecules in the early and later stages were distinctively involved in biologically relevant pathways toward neuronal differentiation and maturation. The proposed method has potential for practical applications in regenerative medicine.

Induced pluripotent stem cells (iPSCs) can differentiate into various cell types with distinct functions, making them common candidates for establishing target cells in regenerative medicine[1,2]. Direct reprogramming (DR), the direct conversion of somatic cells into target cells without bypassing iPSCs, has gained notable attention. Unlike iPSC-based methods[3,4], DR reduces genomic damage from mutations and generates target cells more quickly by requiring fewer inductive factors. Mouse embryonic fibroblasts (MEFs) can be converted to induced neurons (iNs) using DR-based transcription factors (TFs), such as Ascl1, Lmx1a, and Nurr1[5]. However, TF-based DR relies on gene insertion via a viral vector to overexpress TFs, which may increase the tumorigenic risk[6].

Small molecule–induced DR is expected to mitigate the gene insertion-associated tumorigenic risk[7–9]. The use of small molecules enabled the conversion of skin-derived fibroblasts into cardiomyocytes[10] and neurons[11]. However, small molecule-based DR has been reported in only a few cell types, and the number of candidate small molecules required to induce DR tends to be very large. Thus, exploring small molecules for DR remains challenging. Nevertheless, identifying the optimal combination of small molecules for DR is difficult only with in vitro experiments due to high costs.

To facilitate small molecule-based DR, computational methods have been proposed to predict DR-inducing small molecules. Although the discovery of new combinations of small molecules encompassing biological pathways targeted by known DR-inducing small molecules has been reported, this approach cannot be applied to cell types lacking experimentally identified DR-inducing small molecules[12]. In addition, approaches using transcriptome data have been reported, including the regression-

[1]Department of Bioscience and Bioinformatics, Faculty of Computer Science and Systems Engineering, Kyushu Institute of Technology, Fukuoka, Japan. [2]Department of Complex Systems Science, Graduate School of Informatics, Nagoya University, Nagoya, Japan. [3]Division of Interdisciplinary Research and Development (R&D), Aichi Cancer Center Research Institute, Nagoya, Japan. [4]These authors contributed equally: Ryoku Ito, Momoko Hamano. ✉e-mail: momoko@bio.kyutech.ac.jp; yamanishi@i.nagoya-u.ac.jp

based method with biological pathways[13], network-based method with information on DR-inducing TFs[14], and combinatorial optimization method that mimics the gene expression patterns of DR-inducing TFs[15]. However, these previous studies either failed to incorporate temporal alterations in gene expression during DR or relied on bulk transcriptome data.

In this study, we present a novel computational method, SuperDIRECTEUR (Simulation-sUPERvised DIrect REprogramming by Chemicals with TranscriptomE Utilization for Regenerative Medicine), to predict the optimal combinations of small molecules for DR using single-cell transcriptome data. The cellular conversion processes in DR are simulated by calculating the RNA velocity in each cell. These processes are classified into multiple stages (i.e., primal, immature, and mature stages), and a variant of simulated annealing is performed to predict the combination of small molecules that mimic DR-specific gene expression patterns. We show the usefulness of SuperDIRECTEUR in the application to DR converting MEFs to iNs. The proposed method is expected to be useful for regenerative medicine applications.

Results

Overview of the proposed method, SuperDIRECTEUR, for predicting small molecules for DR

To consider the temporal cellular conversion process, we examined the RNA velocity for each cell in single-cell transcriptome data during DR and predicted candidate small molecules for DR. The proposed method, SuperDIRECTEUR, comprises five steps (Fig. 1).

Step 1 involves calculating the RNA velocity for each cellular state using single-cell transcriptome profiles during the DR conversion of MEFs to iNs (Fig. 1A).

Step 2 involves simulating the cellular conversion processes during DR by applying a Markov chain simulation based on RNA velocity. Here, the cellular conversion processes are classified into three stages (primal, immature, and mature) according to the gene expression patterns of the cellular states (Supplementary Fig. 1), followed by the construction of a stage-specific transcriptome matrix for each stage. Finally, the gene expression levels of the stage-specific transcriptome between two adjacent stages (i.e., primal stage and immature) are compared to identify differentially expressed genes (DEGs) and to construct cellular conversion-specific profiles (CSPs) (Fig. 1B).

Step 3 entails constructing a set of chemically induced gene expression profiles (CIPs) for small molecules, including approved drugs. To identify small molecules that mimic CSP in DR, we calculate the gene expression rank similarity (GRS) as the similarity of the gene expression patterns between CSP and CIP (Fig. 1C, right).

Step 4 involves preparing a set of CIPs for small molecules, including approved drugs. The DR score is calculated for each small molecule, where the DR score represents its potential to induce DR based on GRS, after which candidate small molecules with high DR scores are selected (Fig. 1D).

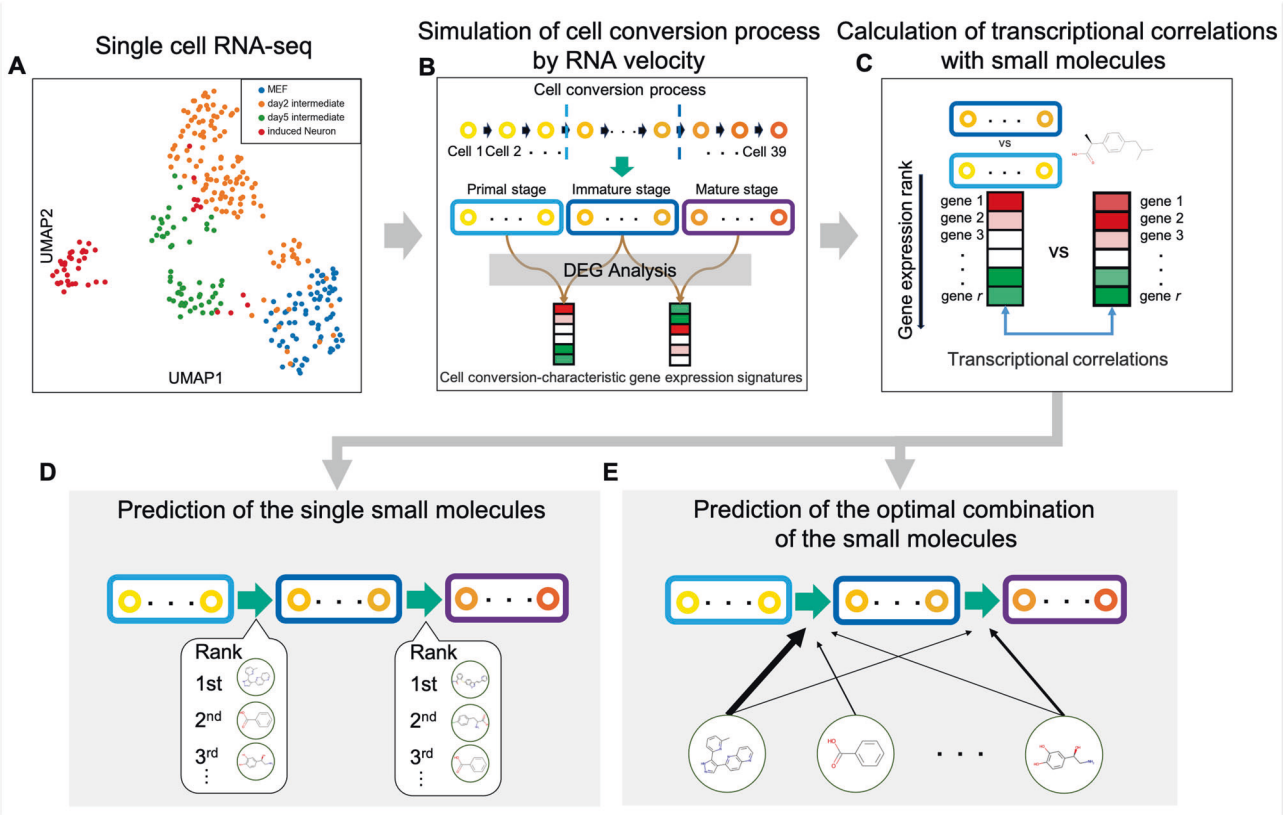

Fig. 1 | Overview of the proposed method for predicting small molecules that induce direct reprogramming (DR) using single-cell RNA-seq. A Gene expression patterns during the DR of mouse embryonic fibroblasts (MEFs) to induced neurons (iNs) were visualized using UMAP. Each dot indicates a cell, and the color of the dots represents the time points in the cellular conversion process during DR. **B** RNA velocity was calculated for each cell to simulate the cellular conversion process from MEFs to iNs. To consider temporal cellular conversion, cells were classified into three stages: primal, immature, and mature. To capture stage-specific transcriptomic profiles, DEGs showing significant expression between adjacent stages were identified. Cellular conversion-specific profiles (CSPs) were constructed using DEGs. **C** The CSP was compared with the small molecule–induced gene expression profile (CIP) by ranking the gene expression levels, and the gene expression rank similarity (GRS) was calculated for each small molecule and the combination of small molecules. **D** Small molecules capable of inducing cellular conversion at each temporal stage were predicted. The DR score was calculated for each small molecule, and high-scoring small molecules were selected as candidates. **E** Combinations of small molecules that would induce cellular conversion at each temporal stage were explored using a variant of simulated annealing. The combinations of small molecules with high DR scores were selected as candidates.

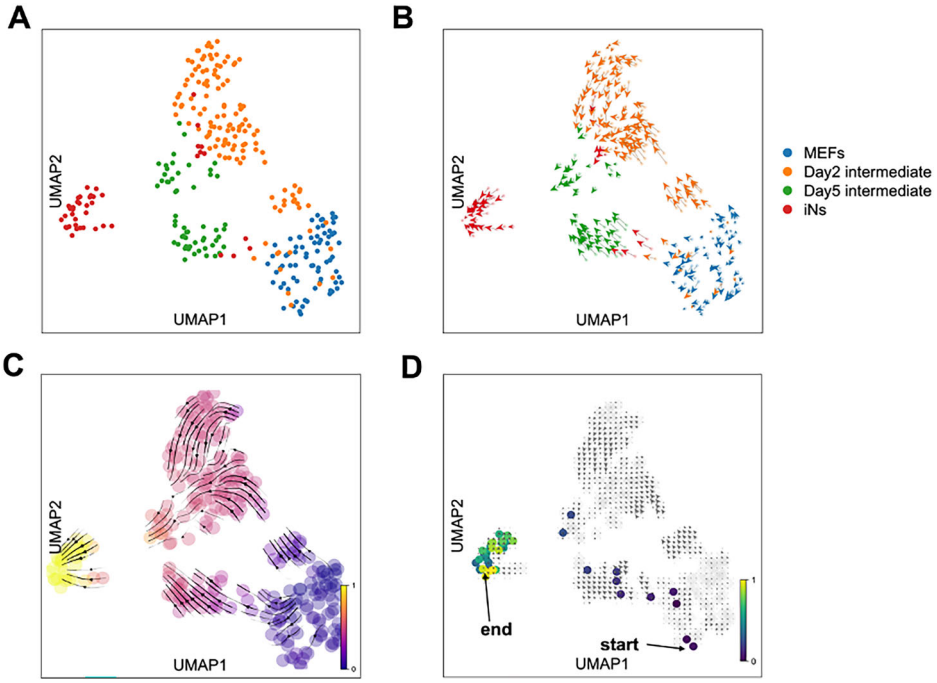

Fig. 2 | Simulation of the cellular conversion process during DR from fibroblasts into neurons. **A** Gene expression patterns during DR from MEFs to iNs were visualized using UMAP. Each dot indicates a cell, and the color of the dots represents the cell types during DR. Blue dots indicate MEFs, yellow dots indicate cells in the intermediate state after Ascl1 induction for 2 days, green dots indicate cells in the intermediate state after Ascl1 induction for 5 days, and red dots indicate iNs induced by Ascl1, Brn2, and Myt1l. **B** RNA velocities derived from the dynamic model of DR from MEFs to iNs were visualized using UMAP. The arrows indicate the estimated future states of each cell. **C** Temporal alterations in the cellular conversion process during DR were visualized using scVelo's latent time. The color of the dots represents the internal clock of the cells in the latent time, and the arrows indicate the estimated future states of each cell. **D** The cellular conversion process during DR was simulated based on the transition probabilities calculated between cells using velocities. The color of the dots represents the time point of the potential fate of the transitioning cells based on the similarity between cellular states.

Finally, Step 5 involves performing a variant of simulated annealing, a combinatorial optimization algorithm, to explore the optimal combination of small molecules with high DR scores (Fig. 1E).

More detailed workflows of the procedures are shown in Supplementary Figs. 2, 3 and 4.

Simulation of the cellular conversion process during DR

To simulate the cellular conversion process during DR from MEFs to iNs, we calculated the RNA velocity for each cell using single-cell transcriptome data, after which the simulated cellular conversion processes were constructed by applying a Markov chain simulation based on RNA velocity.

We first performed a trajectory analysis based on the RNA velocity for each cell derived from the "dynamical model" (Fig. 2). Each cell was pre-annotated with time points in the single-cell transcriptome data following a previous study[16]. Therefore, to represent the cells during DR, we grouped the cells into four representative cellular states: "MEFs," "Day 2 intermediate," "Day 5 intermediate," and "iNs" (Fig. 2A).

Visualizing the calculated RNA velocities revealed that the RNA velocity of each cellular state was directed from MEFs to iNs through Day 2 and Day 5 intermediates (Fig. 2B). Visualizing the temporal alterations based on the RNA velocity–estimated latent time showed that the latent time changed from MEF to iNs through Day 2 and Day 5 intermediates (Fig. 2C).

To investigate the cellular state alteration during DR, we constructed a transition probability matrix from the calculated RNA velocities and simulated the alterations in the cellular states using Markov chain simulations, which resulted in the acquisition of simulated cellular conversion processes comprising potential fates from 39 cellular states (Fig. 2D). Potential fates were identified along the trajectory of MEFs, Day 5 intermediate, and iNs (Supplementary Table 1). This simulation was performed based on the transition probabilities among the cellular states, and similar results were observed in multiple trials (Supplementary Fig. 5). Therefore, we used the simulation results below.

Biological interpretation of each cellular conversion process

The efficiency of small molecule–based DR is dependent on the timing of chemical treatment. Thus, the critical genes for DR induction are considered dependent on the cellular conversion process stages over time. The gene expression patterns of the cellular state at each time point during the cellular conversion processes were analyzed, grouping the cellular states into multiple stages (i.e., primal, immature, and mature stages). DEGs were identified between stages, and their biological functions were examined.

The cosine similarity of gene expression profiles among cellular states during the cellular conversion process was first calculated, leading to the classification of cellular states into three stages: primal (6 cells), immature (3 cells), and mature (30 cells). The experimental conditions for each cell selected by simulation were shown in Supplementary Table 1. Concatenating the gene expression profiles of the classified cellular states at each stage, a stage-specific transcriptome matrix was then constructed. Finally, after extracting the DEGs significantly expressed between the primal and immature stages in the stage-specific transcriptome matrix, a cellular conversion-specific profile represented as $CSP_{Imm/Pri}$ was constructed. The number of DEGs in $CSP_{Imm/Pri}$ was 633 (Fig. 3A left panel). Similarly, DEGs with significant expression alterations between the immature and mature stages in the stage-specific transcriptome matrix were extracted, and a cellular conversion-specific profile represented as $CSP_{Mat/Imm}$ was constructed. The number of DEGs constituting $CSP_{Mat/Imm}$ was 1810 (Fig. 3A right panel).

To understand the biological functions of cellular conversion from primal to immature stages, we performed pathway enrichment analysis on the DEGs specifically identified in $CSP_{Imm/Pri}$ (348 genes) (Fig. 3B). "TGF-beta signaling pathway," a pathway involved in early-stage reprogramming, was detected[17]. "Oxidative phosphorylation" was also detected. The fact that the metabolic processes of cells change from glycolysis to oxidative phosphorylation during DR from MEFs to iNs suggests DEG involvement in cell proliferation, differentiation, and metabolic processes.

Pathway enrichment analysis was also performed on the DEGs specifically identified in $CSP_{Mat/Imm}$ (1525 genes) (Fig. 3C). "MAPK signaling pathway," a pathway controlled during DR, "Hippo signaling pathway," a regulator of the proliferation and apoptosis of neurons, and "Axon guidance pathway," a pathway involved in the formation of nerve cells, were detected[8,18,19]. These results suggest DEG involvement in cell proliferation, differentiation, and metabolic processes.

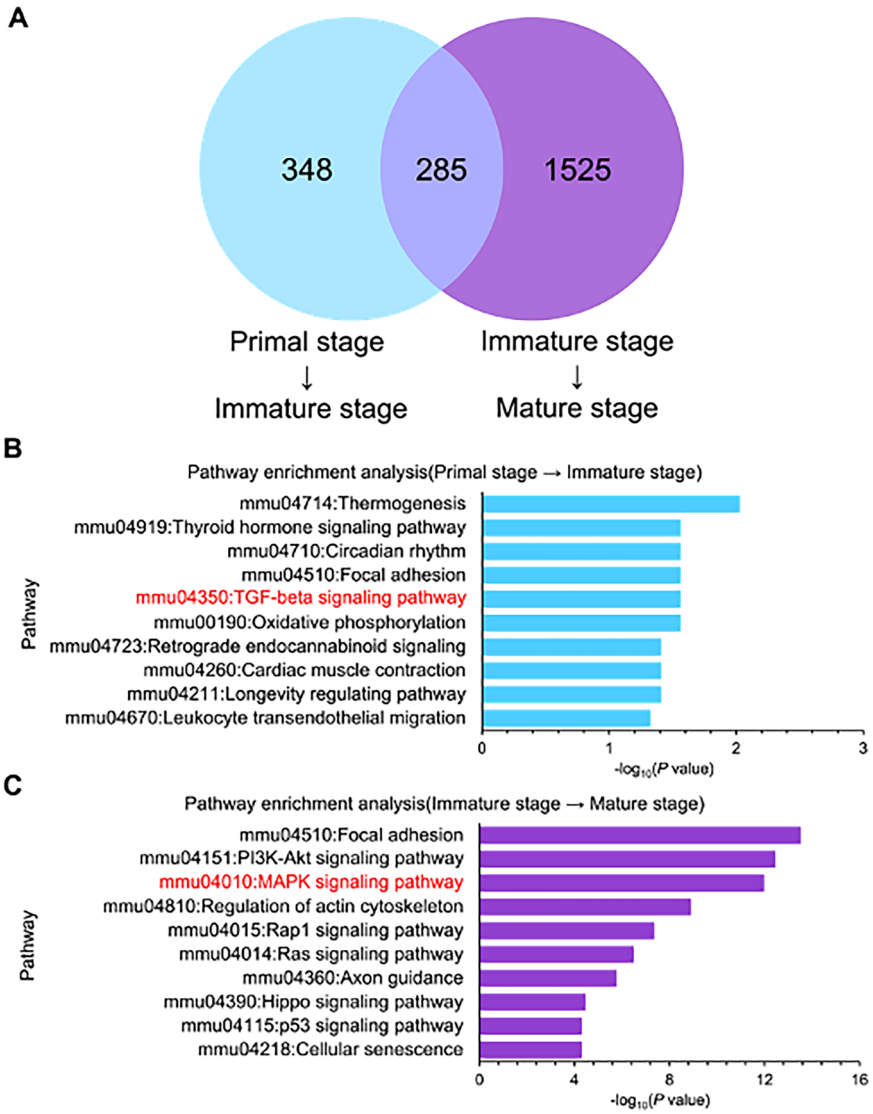

Fig. 3 | Detection of DEGs for cellular conversion. **A** Venn diagram showing the number of DEGs that were significantly expressed during cellular conversion from the primal stage to the immature stage and from the immature stage to the mature stage. **B** Top 10 pathways enriched with 348 DEGs identified between the primal and immature stages. Experimentally proven pathways regulated during DR are highlighted in red. **C** Top 10 pathways enriched with 1525 DEGs identified between the immature and mature stages. Experimentally proven pathways regulated during DR are highlighted in red.

Individual candidate small molecules for DR from MEFs to iNs

Calculating the DR score with $CSP_{Imm/Pri}$, $CSP_{Mat/Imm}$, and CIP, we searched for individual candidate small molecules capable of inducing conversion from the primal stage to the immature stage and from the immature stage to the mature stage, respectively (Table 1).

We predicted small molecules for cellular conversion from the primal stage to the immature stage based on their DR scores (Upper part of Table 1). Among the top 10 small molecules, geldanamycin, CHIR-99021, and SB-203580 corresponded to known DR-inducing small molecules that were experimentally proven to induce DR[11,20,21]. Dinaciclib, triamcinolone, and linifanib exhibited functions similar to known DR-inducing small molecules. Dinaciclib is functionally similar to the known DR-inducing small molecule pifithrin-α (PFT-α) in terms of regulating the same p53 signaling pathway. Linifanib regulates the MAPK signaling pathway, which plays a role in DR in neuronal cells[8,22]. Additionally, triamcinolone has a similar efficacy to anti-inflammatory agents such as dexamethasone that enhance DR.

Next, we examined the predicted small molecules for cellular conversion from the immature stage to the mature stage (Lower part of Table 1). Troglitazone, ZM336372, and LY-364947 corresponded to known DR-inducing small molecules. Linifanib, nifedipine, vorinostat, and masitinib had similar functions to known DR-inducing small molecules. Specifically, nifedipine, which regulates L-type calcium channels, is functionally similar to the known DR-inducing small molecule BayK 8644. Vorinostat and masitinib, which regulate the MAPK and Notch signaling pathways, were reported to induce DR in neuronal cells[8,23,24]. These results suggest the involvement of the predicted small molecules with CIPs that are similar $CSP_{Imm/Pri}$ and $CSP_{Mat/Immn}$ in the DR of MEFs to iNs.

Combination of candidate small molecules for DR from MEFs to iNs

DR induction with a single small molecule is often difficult but requires a combination of multiple small molecules. Therefore, we searched for the optimal combination of small molecules for cellular conversion from the primal stage to the immature stage and from the immature stage to the mature stage by performing a variant of simulated annealing, known as a combinatorial optimization method (Table 2).

Five small molecules for cellular conversion from the primal stage to the immature stage were predicted (Left part of Table 2). SB-431542 corresponded to known DR-inducing small molecules[25]. In addition, dinaciclib, ketoprofen, and dexibuprofen were functionally similar to known DR-inducing small molecules. We evaluated the ratio of known DR-inducing small molecules among the predicted small molecules using Fisher's exact test and confirmed that the known DR-inducing small molecules were significantly enriched in the predicted small molecules ($p = 0.00035054$). Ketoprofen and dexibuprofen were identified as small molecules with anti-inflammatory effects that promote DR.

Table 1 | The predicted small molecules for cellular conversion from the primal stage to the immature stage and immature stage to the mature stage

Primal stage → Immature stage			
Rank	**Small molecule**	**Function**	**DR Score**
1	**Geldanamycin**	**Hsp90 Inhibitor**	0.396
2	Dinaciclib*	CDK inhibitor*	0.391
3	Procaine	Anesthetic	0.370
4	**CHIR-99021**	**GSK-3 Inhibitor**	0.294
5	Triamcinolone*	Anti-inflammatory*	0.284
6	**SB-203580**	**MAPK inhibitor**	0.283
7	Linifanib*	Tyrosine kinase inhibitor*	0.274
8	Dactolisib	PI3K inhibitor	0.254
9	Mephenytoin	Anticonvulsant	0.248
10	Tetryzoline	Adrenergic receptor agonist	0.245
Immature stage → Mature stage			
Rank	**Small molecule**	**Function**	**DR Score**
1	**Troglitazone**	**Tyrosine kinase inhibitor**	2.37
2	Linifanib*	Tyrosine kinase inhibitor*	1.58
3	Nifedipine*	Calcium channel blocker*	1.54
4	**ZM-336372**	**Raf Inhibitor**	1.21
5	**LY-364947**	**TGFβ inhibitor**	1.12
6	Vorinostat*	HDAC inhibitor*	1.09
7	Batimastat	Matrix metalloproteinase inhibitor	1.08
8	Masitinib*	Receptor tyrosine kinase inhibitor*	1.08
9	Zolpidem	GABA-A receptor agonist	1.03
10	LY-2603618	Checkpoint kinase inhibitor	1.02

Experimentally proven small molecules capable of inducing DR are highlighted in bold. Small molecules that are functionally similar to known DR-inducing small molecules are highlighted with asterisks.

Table 2 | The predicted combination of small molecules for cellular conversion from the primal stage to the immature stage and from the immature stage to the mature stage

Primal stage → Immature stage		Immature stage → Mature stage	
Small molecule	**Function**	**Small molecule**	**Function**
Dinaciclib*	CDK inhibitor*	**Troglitazone**	**PPARγ agonist**
Mephenytoin	Anticonvulsant	LY-2603618	Checkpoint kinase inhibitor
Ketoprofen*	COX inhibitor*	Zolpidem	GABA-A receptor agonist
Dexibuprofen*	COX inhibitor*	**ZM-336372**	**Raf Inhibitor**
SB-431542	**TGFβ receptor kinase inhibitor**	Nifedipine*	Calcium channel blocker*
DR Score	0.754	DR Score	6.21

Experimentally proven small molecules capable of inducing DR are highlighted in bold. Small molecules that are functionally similar to known DR-inducing small molecules are highlighted with asterisks.

The predicted combinations of small molecules for cellular conversion from the immature stage to the mature stage included troglitazone and ZM-336372, which are known DR-inducing small molecules[26,27]. Nifedipine was identified as a small molecule with functions similar to those of known DR-inducing small molecules such as BayK 8644 (Right part of Table 2). Fisher's exact test confirmed a significant enrichment of known DR-inducing small molecules among the predicted candidates, based on the evaluation of their ratio ($p = 0.00035054$).

We also investigated the performance when the number of small molecules was altered. We predicted the combination of small molecules set to $k = 2, 3, 4$. As shown in Supplementary Table 2, even when the number of small molecules was reduced, known small molecules or small molecules with similar functions were reproduced to a certain extent.

Biological mechanism of the predicted combinations of small molecules for DR

To elucidate the molecular biological mechanisms regulated by the predicted combinations of small molecules for DR, we evaluated their biological functions and functional associations with the target proteins of small molecules based on the compound-protein interactome data (Fig. 4).

The five small molecules for cellular conversion from the primal stage to the immature stage had 89 target proteins, where the target proteins of individual small molecules were merged.

A subnetwork of the arachidonate lipoxygenase protein family—regulated by ketoprofen—was detected in the association network of the target proteins of small molecules in the predicted combination (Fig. 4A). During DR from fibroblasts to neurons, the metabolic pathway shifts from glycolysis to oxidative phosphorylation, and regulating arachidonate lipoxygenase helps maintain this switching of metabolic processes[1]. Bcl2 was also identified as a highly central node. Bcl2 is a protein that inhibits apoptosis and can reduce lipid peroxidation and promote DR into iNs[1,28]. Bcl2 expression increased during cellular conversion from the primal stage to the immature stage.

A subnetwork of the cyclin-dependent kinase protein family was detected. Preventing the cell cycle in the G1 phase in fibroblasts can promote efficient conversion from fibroblasts to neurons[29]. The expression levels of Cdk4 and Cdk6 were suppressed, where Cdk4 and Cdk6 were the target proteins associated with the transition from the G1 phase to the S phase in the cell cycle. This indicates that the expression of cell cycle-related genes is regulated during cellular conversion from the primal stage to the immature stage during DR. This suggests the biological relevance of the small molecules targeting these proteins. Overall, these results suggest that the predicted combination of small molecules suppresses cell proliferation, resulting in DR from the MEF into iNs.

Finally, we performed gene ontology (GO) and pathway enrichment analyses on the target proteins (Fig. 4C). As a result, some GO terms associated with "mRNA transcription" and "chromatin remodeling" were detected. "Hedgehog signaling pathway," which regulates DR, was also detected. These results suggest the involvement of the target proteins in early cellular conversion from the primal stage to the immature stage through transcription activation.

Our evaluation of the five small molecules for cellular conversion from the immature stage to the mature stage yielded 162 target proteins. In the association network of the target proteins, subnetworks of calcium voltage-gated channel subunits and gamma-aminobutyric acid (GABA) receptor subunits were detected (Fig. 4B). They are targets of nifedipine, one of the small molecules in the predicted combination. Calcium voltage-gated channel subunits facilitate the influx of Ca^{2+} into cells, generating Ca^{2+} signals that regulate various cellular functions, such as MAPK signaling[30]. In addition, GABA receptor subunits bind to the inhibitory neurotransmitter GABA, thereby suppressing its release[30,31]. Therefore, small molecules targeting calcium voltage-gated channel subunits and GABA receptor subunits may promote neuronal function. Moreover, the association network included Mapk14 and Braf, components of the MAPK signaling pathway regulated in DR. These results indicate that the predicted combination of small molecules may regulate the pathways that promote neuronal differentiation and development.

Similarly, GO and pathway enrichment analyses were performed on the target proteins (Fig. 4D). Some GO terms associated with neurogenesis were detected, such as "neuronal development" and "angiogenesis." Vascular endothelial growth factors (VEGFs) not only induce angiogenesis but

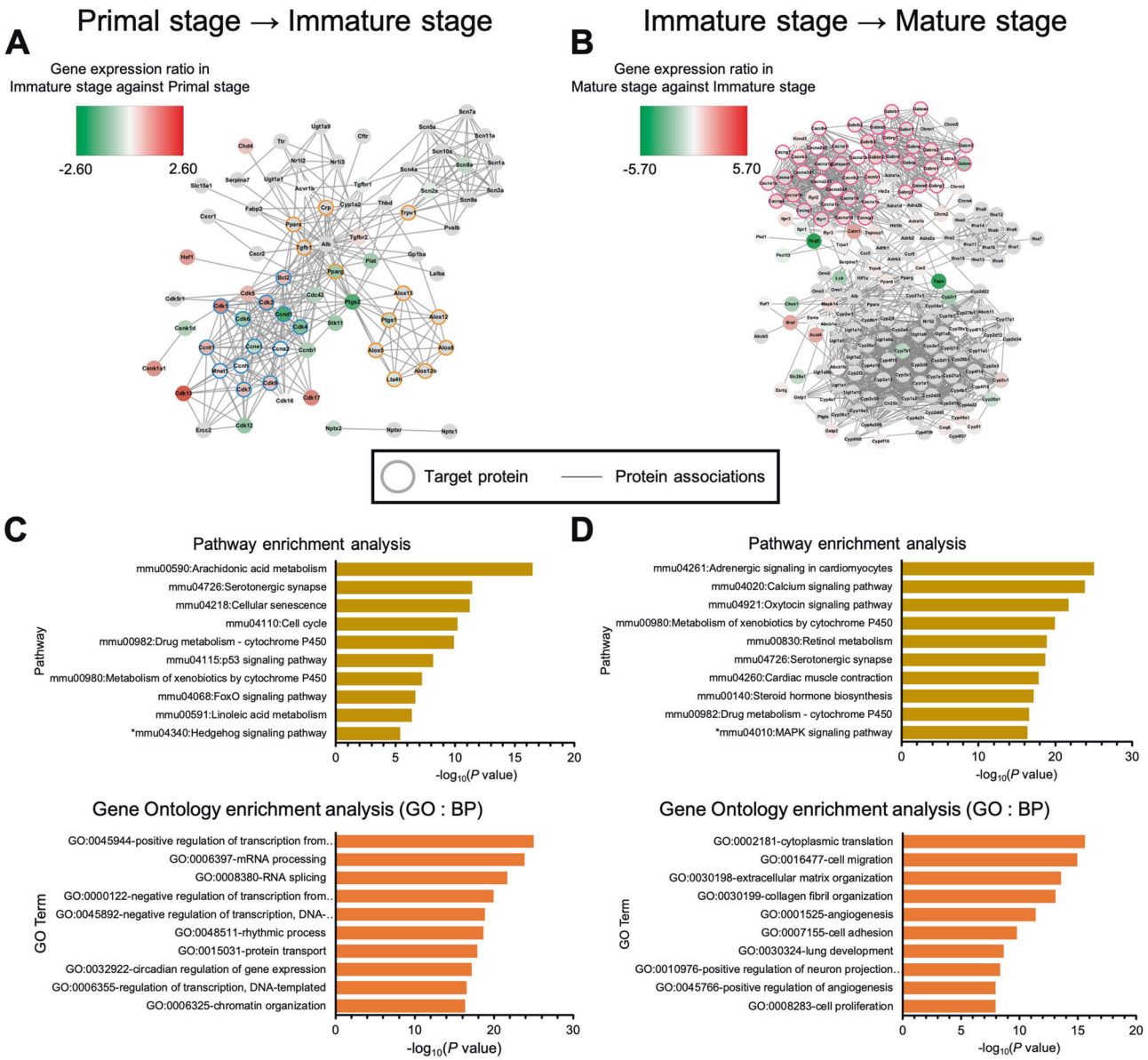

Fig. 4 | The molecular mechanisms of the target proteins of small molecules in the predicted combination for cellular conversion. Protein–protein association networks of target proteins of the predicted small molecules for cellular conversion from the primal to immature stages (**A**) and from the immature to mature stages (**B**). The expression ratio was depicted by the color of the node using the target cell-specific transcriptome profile. The red node indicates upregulation of the gene expression level, the green node indicates downregulation of the gene expression level, and the gray node indicates that the gene expression level was unknown. The colors of the node outlines indicate the detected subnetworks. **C** The top 10 GO terms and KEGG pathways of 89 target proteins of the predicted small molecules for cellular conversion from the primal to immature stages. Experimentally proven pathways regulated during DR are highlighted in red. **D** The top 10 GO terms and KEGG pathways of 162 target proteins of the predicted small molecules for cellular conversion from the immature to mature stages. Experimentally proven pathways regulated during DR are highlighted in red.

also promote neurogenesis[32,33]. The "MAPK signaling pathway" and "Retinol metabolism," which regulates neuronal proliferation and differentiation, were also detected[34,35]. These results suggest the involvement of the predicted combination of small molecules in the later stages of cellular conversion that enhances the process from the immature stage to the mature stage toward iN maturation.

Discussion

In this study, we developed a computational method, SuperDIRECTEUR, to simulate the cellular conversion process using single-cell transcriptome data acquired during DR. Our proposed method can predict the optimal combinations of small molecules for DR at each temporal stage of cellular conversion. The novelty of SuperDIRECTEUR lies in the construction of the simulated cellular conversion processes based on RNA velocity utilizing single-cell transcriptome data during DR. SuperDIRECTEUR can classify the cellular conversion processes into multiple stages and predict the optimal combinations of small molecules for each temporal stage of the cellular conversion. SuperDIRECTEUR is expected to contribute to regenerative medicine advancement.

A related previous study is a network-based approach with TFs[14]. In the previous work, single-cell transcriptome data from healthy adult tissues were used, whereas in this study, single-cell transcriptome data during DR were used. Another related previous study is a transcriptome-based approach with gene expression profiles of the source and target cells[15]. In the previous study, only a snapshot of gene expression patterns was considered, while in this study, the temporal gene expression patterns in the entire

cellular conversion process during DR from source cells to target cells were considered. SuperDIRECTEUR has several advantages over the previous method. First, it can predict small molecules by capturing dynamic cellular conversion states at the single-cell level. Second, it can divide the DR induction stage into several stages and predict the optimal small molecule combinations or individual small molecules for each stage. Third, it can predict small molecules based on their gene expression profiles that are dependent on various experimental conditions, rather than the average of gene expression profiles across different experimental conditions. Thus, our proposed method can capture the dynamic features of changes in gene expression throughout the cellular conversion process over time.

A unique feature of this study is that we emphasize the prediction of small molecules at each temporal stage of cellular conversion processes during DR. Even if iPSC or DR are used for producing target cells, the treatment of small molecules at appropriate time points is known to improve conversion efficiency[7,36]. The appropriate timing for the treatment of small molecules is cell type-specific. The number of stages for the targeted cellular conversion in our proposed method can be flexibly changed, depending on the target cell type. Our proposed method can be applied to DR induction in multiple stages, which contributes to the development of chemical reprogramming methods for various cell types. As another example, the result of applying the proposed method to cardiomyocytes is shown in Supplementary Information (Supplementary Table 3). In the application, we used single-cell RNA-seq data during DR induction for cardiomyocytes (GSE196125). Dynamical modeling predicted the transitions of 44 representative cells, whose cellular states could be divided into five stages. The results of small molecule predictions for the four patterns between the five stages are shown in Supplementary Table 3. The prediction results included at least one known DR-inducing small molecule or small molecules with similar functions at each stage. In the future, this method can be further extended to single-cell RNA-seq data on DR induction in other cell types when these datasets become publicly available.

SuperDIRECTEUR is based on single-cell transcriptome data to characterize the cellular conversion process during DR. However, a limitation of SuperDIRECTEUR is its reliance solely on transcriptome data as input despite the potential of incorporating various types of single-cell omics data. Epigenetic controls such as DNA methylation and histone modification regulate gene expression. Thus, SuperDIRECTEUR could be improved by using ChIP-seq data that quantify TF binding sequences and histone modification regions at the single-cell level and ATAC-seq data that quantify chromatin accessibility information at the single-cell level.

Materials and methods

Single-cell transcriptome data during DR from MEFs to iNs

Single-cell transcriptome data were obtained from the Gene Expression Omnibus (GEO) database. Single-cell transcriptome data of 405 cells undergoing DR from MEFs to iNs, achieved by overexpressing Ascl1, Brn2, and Myt1l, were obtained from the GSE67310 dataset[16].

Processing of single-cell transcriptome data

Raw reads were preprocessed with seqkit (v2.4.0), FASTQC (v0.11), and trim-galore (v0.6.2), followed by sequence alignment using hisat2 (v2.2.0). BAM files were then generated with SAMtools (v2.8.0). Only trim-galore was run with a quality threshold of 30, while the default parameters were used for all other tools. All reads were mapped to the mouse reference genome (GRCm39/mm39). Additionally, the velocyto.py command-line tool was used with the run-smartseq2 command to generate loom files from the BAM files[16].

Each cell was preannotated by the experimenter with cell-type information. In this study, 106 cells, annotated as "iNs cells derived from MEFs 20 days upon Ascl1 induction," which exhibited muscle-like characteristics, were excluded from the downstream analysis, leaving 299 cells. These 299 cells were annotated as follows: 73 cells labeled as "Fibroblast culture from embryonic (E14.5) mouse tail tips" were designated as MEFs; a total of 128 cells annotated as "iNs cells derived from clonal Ascl1-inducible MEFs 2 days upon Ascl1 induction" and "iNs cells derived from MEFs 2 days upon Ascl1 induction" were grouped under "Day 2 intermediate"; 55 cells labeled as "iNs cells derived from MEFs 5 days upon Ascl1 induction" were designated as "Day 5 intermediate"; and 43 cells labeled as "iNs cells derived from MEFs 22 days upon induction of Ascl1, Brn2, and Mytl1" were designated as "iNs."

The preprocessing was further performed using the API from scVelo[37]. Filtering genes with counts of 10 or more across all cells resulted in 14,090 genes. Transcriptome data for each cell were normalized, followed by a logarithmic transformation of expression levels. The transcriptome profile of cell c was defined as $\boldsymbol{x}(c) = \left(x_1, x_2, \ldots, x_q\right)^{\boldsymbol{T}}$, where q is the number of genes ($q = 14,090$). Next, the transcriptome profiles for all cells $c_1, c_2, \ldots, c_S$ were defined as $\boldsymbol{x}(c_1), \boldsymbol{x}(c_2), \ldots, \boldsymbol{x}(c_S)$, where S denotes the number of cells ($S = 299$). The DR-induction gene expression matrix was defined as $\boldsymbol{X} = (\boldsymbol{x}(c_1), \boldsymbol{x}(c_2), \ldots, \boldsymbol{x}(c_S))$. Here, the rows in $\boldsymbol{X}$ correspond to genes (q =14,090) and the columns in $\boldsymbol{X}$ correspond to cellular states ($S = 299$). $\boldsymbol{X}$ is a matrix of size 14,090 × 299.

Visualization of single-cell transcriptome data

From all the genes in the DR-induction gene expression matrix $\boldsymbol{X}$, 2000 genes with the highest variability in expression across cells were selected, and the genes were visualized using uniform manifold approximation and projection (UMAP). For UMAP-based visualization, the scvelo.tl.umap method was employed, with the parameter set to min_dist = 0.5.

Construction of stage-specific transcriptome matrices

From all genes in the DR-induction gene expression matrix $\boldsymbol{X}$, 2000 genes with the highest variability in expression across cells were selected, and the RNA velocity for each cell was calculated using the "scvelo.tl.velocity" method. We applied the dynamic model in scVelo to this calculation. Subsequently, the transition probabilities between cells were computed using the "scvelo.utils.get_transition_matrix" method with default settings, resulting in the construction of the transition probability matrix T. The matrix T is an S×S square matrix (where $S = 299$ in this study), where each element represents the probability $P_{a,b}$ of a cell $c_{a\in S}$ transitioning to a cell $c_{b\in S}$, rows correspond to the cell before the transition, and columns correspond to the cell after the transition.

A Markov chain simulation was performed using the "scvelo.utils.get_cell_transition" method with a seed value of 1, starting from a randomly selected cell in the MEF cluster as the initial state. The simulation was applied to the transition probability matrix T, allowing transitions (including self-transitions) based on the probabilities indicated in the matrix. Cells were transitioned based on the transition probability in the transition probability matrix T from the row (corresponding to the cell in the current state) to the column (corresponding to the cell in the next state). In these simulations, if a cell is determined to have no probability of transitioning to a neighboring cell, it is excluded from the list of candidates for future transitions. After the simulation, cells that transitioned back to themselves were excluded from the analysis.

Among all cells $c_1, c_2, \ldots, c_S$, a subset of s cells ($s < S$) that were probabilistically transitioned during the simulation was selected to represent the cellular states characteristic of DR induction. By reiterating this process, we selected 39 cells that were likely to undergo future transitions ($s = 39$). The sequential cellular states of the s cells were denoted as $c_1^{(trans)}, c_2^{(trans)}, \ldots, c_s^{(trans)} (s = 39)$. The gene expression profile corresponding to each cellular state $c^{(trans)}$ was defined as $\boldsymbol{x}(c^{(trans)}) = [x_1(c^{(trans)}), x_2(c^{(trans)}), \ldots, x_q(c^{(trans)})]^{\mathrm{T}}$. The cell conversion processes during DR are represented by $c_1^{(trans)}, c_2^{(trans)}, \ldots, c_s^{(trans)}$, and they were further divided into multiple stages following the procedure described below.

Step 1. For the gene expression profile $\boldsymbol{x}(c_1^{(trans)})$ of the cellular state $c_1^{(trans)}$, the cosine similarities of the gene expression profiles with

subsequent cellular states $\boldsymbol{x}(c_2^{(trans)}), \ldots, \boldsymbol{x}(c_s^{(trans)})$ were calculated as follows:

$$\cos\left(\boldsymbol{x}\left(c_1^{(trans)}\right), \boldsymbol{x}\left(c_t^{(trans)}\right)\right) (t = 2, \ldots, s)$$

The maximum value of t that satisfies $\cos(\boldsymbol{x}(c_1^{(trans)}), \boldsymbol{x}(c_t^{(trans)})) \geq 0.50$ was defined as s_{pri} (in this study $s_{pri} = 6$). The set of cellular states $c_1^{(trans)}, .., c_{s_{pri}}^{(trans)}$ was collectively referred to as the primary stage of the cell conversion process.

Furthermore, the gene expression profiles corresponding to all cellular states in the primal stage, $c_1^{(trans)}, .., c_{s_{pri}}^{(trans)}$, were combined to form the stage-specific gene expression matrix $\boldsymbol{X}^{(Pri)} = \left[\boldsymbol{x}(c_1^{(trans)}), \boldsymbol{x}\left(c_2^{(trans)}\right), \ldots, \boldsymbol{x}\left(c_{s_{pri}}^{(trans)}\right)\right]$. Here, the rows in $\boldsymbol{X}^{(Pri)}$ correspond to genes (q =14,090) and the columns in $\boldsymbol{X}^{(Pri)}$ correspond to cellular states in the primal stage ($s_{pri} = 6$). $\boldsymbol{X}^{(Pri)}$ is a matrix of size 14,090 × 6.

Step 2. Similar to Step 1, for the gene expression profile $\boldsymbol{x}(c_{s_{pri}+1}^{(trans)})$ of the cellular state $c_{s_{pri}+1}^{(trans)}$, the cosine similarities of the gene expression profiles with the subsequent cellular states $\boldsymbol{x}(c_{s_{pri}+1}^{(trans)}), \ldots, \boldsymbol{x}(c_{s_{pri}+s_{imm}}^{(trans)})$ were calculated as follows:

$$\cos\left(\boldsymbol{x}\left(c_{s_{pri}+1}^{(trans)}\right), \boldsymbol{x}\left(c_t^{(trans)}\right)\right) \left(s_{pri} + 1 < t \leq s\right).$$

The maximum value of t that satisfies $\cos(\boldsymbol{x}(c_{s_{pri}+1}^{(trans)}), \boldsymbol{x}(c_t^{(trans)})) \geq 0.50$ was defined as $s_{pri} + s_{imm}$ (in this study $s_{pri} + s_{imm} = 9$). The set of cellular states $c_{s_{pri}+1}^{(trans)}, \ldots, c_{s_{pri}+s_{imm}}^{(trans)}$ was collectively referred to as the immature stage. Additionally, the gene expression profiles of cellular states in the immature stage were represented as $\boldsymbol{x}(c_{s_{pri}+1}^{(trans)}), \ldots, \boldsymbol{x}(c_{s_{pri}+s_{imm}}^{(trans)})$, and they were combined to form the stage-specific gene expression matrix $\boldsymbol{X}^{(Imm)} = \left[\boldsymbol{x}\left(c_{s_{pri}+1}^{(trans)}\right), \ldots, \boldsymbol{x}\left(c_{s_{pri}+s_{imm}}^{(trans)}\right)\right]$. Here, the rows in $\boldsymbol{X}^{(Imm)}$ correspond to genes (q =14,090) and the columns in $\boldsymbol{X}^{(Imm)}$ correspond to cellular states in the immature stage ($s_{imm} = 3$). $\boldsymbol{X}^{(Imm)}$ is a matrix of size 14,090 × 3.

Step 3. Similar to both Steps 1 and 2, for the gene expression profile $\boldsymbol{x}(c_{s_{pri}+s_{imm}+1}^{(trans)})$ of the cellular state $c_{s_{pri}+s_{imm}+1}^{(trans)}$, the cosine similarities of the gene expression profiles with the remaining cellular states $\boldsymbol{x}(c_{s_{pri}+s_{imm}+1}^{(trans)}), \ldots, \boldsymbol{x}(c_s^{(trans)})$ were calculated as follows:

$$\cos\left(\boldsymbol{x}\left(c_{s_{pri}+s_{imm}+1}^{(trans)}\right), \boldsymbol{x}\left(c_t^{(trans)}\right)\right) \left(s_{pri} + s_{imm} + 1 < t \leq s\right).$$

The maximum value of t that satisfies $\cos(\boldsymbol{x}(c_{s_{pri}+s_{imm}+1}^{(trans)}), \boldsymbol{x}(c_t^{(trans)})) \geq 0.50$ was defined as $s_{pri} + s_{imm} + s_{mat}$ (in this study $s_{pri} + s_{imm} + s_{mat} = 39$). The set of cellular states $c_{s_{pri}+s_{imm}+1}^{(trans)}, \ldots, c_{s_{pri}+s_{imm}+s_{mat}}^{(trans)}$ was collectively referred to as the mature stage. Additionally, the gene expression profiles of cellular states in the immature stage were represented as $\boldsymbol{x}(c_{s_{pri}+s_{imm}+1}^{(trans)}), \ldots, \boldsymbol{x}(c_{s_{pri}+s_{imm}+s_{mat}}^{(trans)})$, and they were combined to form the stage-specific gene expression matrix $\boldsymbol{X}^{(Mat)} = \left[\boldsymbol{x}\left(c_{s_{pri}+s_{imm}+1}^{(trans)}\right), \ldots, \boldsymbol{x}\left(c_{s_{pri}+s_{imm}+s_{mat}}^{(trans)}\right)\right]$. Here, the rows in $\boldsymbol{X}^{(Mat)}$ correspond to genes (q =14,090) and the columns in $\boldsymbol{X}^{(Mat)}$ correspond to cellular states in mature stage ($s_{mat} = 30$). $\boldsymbol{X}^{(Mat)}$ is a matrix of size 14,090 × 30.

Construction of chemically induced profiles (CIPs)

Gene expression profiles induced by 21,870 small molecules in human cell lines were obtained from the LINCS project[38], and a set of CIPs was constructed. Each CIP was represented as $\boldsymbol{y} = (y_1, y_2, \ldots, y_{q_{(LINCS)}})^{\mathrm{T}}$, where $q_{(LINCS)}$ represents the number of genes. Each small molecule was identified by InChIKey (https://iupac.org/who-weare/divisions/divisiondetails/inchi/). A total of 1022 drugs registered in the Kyoto Encyclopedia of Genes and Genomes) (KEGG) (https://www.genome.jp/kegg/kegg_ja.html) database were acquired[39]. Among the small molecules, 67 have been reported to induce DR in previous studies.

Construction of cellular conversion-specific profiles (CSPs)

Using the stage-specific gene expression matrices $\boldsymbol{X}^{(Pri)}, \boldsymbol{X}^{(Imm)}, \boldsymbol{X}^{(Mat)}$, which represent multiple stages in the cellular conversion process during DR, we identified genes that exhibited significant upregulation or downregulation between two adjacent stages. These genes are referred to as DEGs. The expression patterns between the primal and immature stages were compared to identify DEGs using the limma package in R with a significance threshold set at $p < 0.05$. As a result, the number of DEGs between $\boldsymbol{X}^{(Pri)}$ and $\boldsymbol{X}^{(Imm)}$ was 633 genes. In addition, gene expression profiles were averaged along the cell axis for $\boldsymbol{X}^{(Pri)}$ and $\boldsymbol{X}^{(Imm)}$, resulting in the stage-specific mean profiles as $\bar{\boldsymbol{x}}_k^{(Pri)} = \sum_{k=1}^{s_{pri}} \boldsymbol{X}_k^{(Pri)}$ and $\bar{\boldsymbol{x}}_k^{(Imm)} = \sum_{k=1}^{s_{imm}} \boldsymbol{X}_k^{(Imm)}$. The differential gene expression profile $\boldsymbol{z}_{Imm/Pri}$ was constructed from the log2-fold change of DEGs between $\bar{\boldsymbol{x}}^{(Pri)}$ and $\bar{\boldsymbol{x}}^{(Imm)}$. Each element of $\boldsymbol{z}_{Imm/Pri}$ was calculated as follows:

$$z_j = \log_2\left(\frac{\bar{x}_j^{(Imm)}}{\bar{x}_j^{(Pri)}}\right) \left(j = 1, 2, \ldots, q_{Imm/Pri}\right),$$

where $q_{Imm/Pri} = 633$.

Similarly, the gene expression patterns between the immature and mature stages were compared to identify the DEGs. In this case, the comparison between $\boldsymbol{X}^{(Imm)}$ and $\boldsymbol{X}^{(Mat)}$ produced 1810 DEGs. These genes were also considered to be regulated between the two adjacent stages. Stage-specific mean profiles were constructed for $\boldsymbol{X}^{(Imm)}$ and $\boldsymbol{X}^{(Mat)}$ by averaging the gene expression values along the cell axis as $\bar{\boldsymbol{x}}_k^{(Imm)} = \sum_{k=1}^{s_{imm}} \boldsymbol{X}_k^{(Imm)}$ and $\bar{\boldsymbol{x}}_k^{(Mat)} = \sum_{k=1}^{s_{mat}} \boldsymbol{X}_k^{(Mat)}$. From the log2-fold change of DEGs between $\bar{\boldsymbol{x}}^{(Imm)}$ and $\bar{\boldsymbol{x}}^{(Mat)}$, a differential gene expression profile $\boldsymbol{z}_{Mat/Imm}$ was constructed. Each element in $\boldsymbol{z}_{Mat/Imm}$ was calculated as follows:

$$z_j = \log_2\left(\frac{\bar{x}_j^{(Mat)}}{\bar{x}_j^{(Imm)}}\right) \left(j = 1, 2, \ldots, q_{Mat/Imm}\right),$$

where $q_{Mat/Imm} = 1810$.

Gene expression profiles constructed from the log2-fold changes in DEGs regulated between stages were referred to as CSPs. Specifically, $\boldsymbol{z}_{Imm/Pri}$ and $\boldsymbol{z}_{Mat/Imm}$ were designated as $\mathrm{CSP}_{Imm/Pri}$ and $\mathrm{CSP}_{Mat/Imm}$, respectively.

Calculation of the gene expression rank similarity (GRS) between CSP and CIP

We introduced a similarity index based on gene expression rankings between CSP and CIP to establish a metric for identifying candidate small molecules that induce DR. CSP and CIP in this study were originally quantified using different platforms, making a direct comparison of gene expression patterns between the two inappropriate. Therefore, to enable statistical comparison of expression levels, genes were ranked relative to one another within the gene expression profiles rather than using real-valued data. Then, we selected candidate small molecules that could potentially reproduce the gene expression patterns observed during cellular conversion.

To calculate the GRS between the mouse-derived $\mathrm{CSP}_{Imm/Pri}$, represented by $\boldsymbol{z}_{Imm/Pri}$, and the human-derived CIP, represented by $\boldsymbol{y}$, we first converted the genes in each profile to their common gene orthologs. As a result, the number of common gene orthologs between $\boldsymbol{z}_{Imm/Pri}$ and $\boldsymbol{y}$,

denoted as $r_{Imm/Pri}$, was 67 genes. We then defined $CSP_{Imm/Pri}$ based on the common gene orthologs as $\boldsymbol{z}'_{Imm/Pri} = (z'_1, z'_2, \ldots, z'_{r_{Imm/Pri}})^{\mathrm{T}}$, and CIP based on the common gene orthologs as $\boldsymbol{y}'_{Imm/Pri} = (y'_1, y'_2, \ldots, y'_{r_{Imm/Pri}})^{\mathrm{T}}$.

Similarly, we converted the genes in the mouse-derived $CSP_{Mat/Imm}$, $\boldsymbol{z}_{Mat/Imm}$, and the human-derived CIP, $\boldsymbol{y}$, to their common gene orthologs. The number of common gene orthologs between $\boldsymbol{z}_{Mat/Imm}$ and $\boldsymbol{y}$, denoted as $r_{Mat/Imm}$, was 56. Consequently, we defined $CSP_{Mat/Imm}$ with the common gene orthologs as $\boldsymbol{z}'_{Mat/Imm} = (z'_1, z'_2, \ldots, z'_{r_{Mat/Imm}})^{\mathrm{T}}$, and CIP with the common gene orthologs as $\boldsymbol{y}'_{Mat/Imm} = (y'_1, y'_2, \ldots, y'_{r_{Mat/Imm}})^{\mathrm{T}}$.

Next, for each CIP profile $\boldsymbol{y}'_{Imm/Pri}$ and $\boldsymbol{y}'_{Mat/Imm}$, we sorted the gene expression ratios in descending order to generate gene ranking lists. A similar procedure was applied to $\boldsymbol{z}'_{Imm/Pri}$ and $\boldsymbol{z}'_{Mat/Imm}$, based on common gene orthologs. The DrInsight package (version 0.1.2)[40] was used to identify candidate small molecules that mimic the gene expression patterns of $\boldsymbol{z}'_{Imm/Pri}$ and $\boldsymbol{z}'_{Mat/Imm}$. Specifically, we looked for genes showing consistent expression patterns, such as genes with increased expression in both CSP and CIP or decreased expression in both.

The importance of each small molecule was then calculated by assessing how closely it replicated the gene expression patterns in $\boldsymbol{z}'_{Imm/Pri}$ and $\boldsymbol{z}'_{Mat/Imm}$.

Finally, the Kolmogorov–Smirnov test was performed to evaluate differences between the distribution of outlier sum statistics for the selected small molecules and those for other small molecules, resulting in a p-value, which was used to define the GRS as follows:

$$\mathrm{GRS}(\boldsymbol{z}', \boldsymbol{y}') = -\log_{10}(p\text{-}value)$$

Here, the GRS between $\boldsymbol{z}'_{Imm/Pri}$ and $\boldsymbol{y}'_{Imm/Pri}$ was denoted as $GRS_{Imm/Pri}$ and the GRS between $\boldsymbol{z}'_{Mat/Imm}$ and $\boldsymbol{y}'_{Mat/Imm}$ was denoted as $GRS_{Mat/Imm}$.

Exploration of individual DR-inducing small molecules based on gene expression pattern similarity

This study explored small molecules that induce DR from a set of 1089 compounds (including 1022 pharmaceutical drugs and 67 known compounds reported to induce DR). The following scores were defined to identify small molecules that mimic the gene expression patterns of $\boldsymbol{z}'_{Imm/Pri}$ and $\boldsymbol{z}'_{Mat/Imm}$:

$$\mathrm{DR}_{Imm/Pri}(\boldsymbol{z}'_{Imm/Pri}, \boldsymbol{y}'_{Imm/Pri}) = \frac{s_{pri} + s_{imm}}{s} * \mathrm{GRS}_{Imm/Pri}(\boldsymbol{z}'_{Imm/Pri}, \boldsymbol{y}'_{Imm/Pri}) * \frac{b_{Imm/Pri}}{r_{Imm/Pri}}$$

$$\mathrm{DR}_{Mat/Imm}(\boldsymbol{z}'_{Mat/Imm}, \boldsymbol{y}'_{Mat/Imm}) = \frac{s_{imm} + s_{mat}}{s} * \mathrm{GRS}_{Mat/Imm}(\boldsymbol{z}'_{Mat/Imm}, \boldsymbol{y}'_{Mat/Imm}) * \frac{b_{Mat/Imm}}{r_{Mat/Imm}}$$

$b_{Imm/Pri}$ and $b_{Mat/Imm}$ represent the number of genes in $CSP_{Imm/Pri}$ and $CSP_{Mat/Imm}$, respectively, which meet the following conditions:

$$b_{Imm/Pri} = \sum_{l=1}^{r_{Imm/Pri}} I\{z'_l \cdot y'_l > 0\}$$

$$b_{Mat/Imm} = \sum_{l=1}^{r_{Mat/Imm}} I\{z'_l \cdot y'_l > 0\}$$

where I is an indicator function that returns 1 if the event is true and returns 0 otherwise.

Prediction of combinations of DR-inducing small molecules using an optimization algorithm

In this study, we predicted effective combinations of small molecules capable of inducing DR from a set of 1089 compounds comprising 1022 drugs and 67 compounds already known to induce DR. To identify the optimal combinations of small molecules, we developed an optimization algorithm based on simulated annealing[41,42]. To explore new DR-inducing small molecules from the candidate set of n (n = 1089) small molecules, the $I_k = \{i_1, \ldots, i_k\}(1 \le k \le n)$ index, a vector comprising the selected small molecules $\boldsymbol{v}(I_k) := (\boldsymbol{v}(I_k)_i)_{i=1,\ldots,n} \in \{0,1\}^n$ with an n-dimensional binary vector, was defined as follows:

$$\boldsymbol{v}(I_k)_i = \begin{cases} 1, & \text{if } i \in I_k \\ 0, & \text{otherwise} \end{cases},$$

where the selected small molecules were set to 1 and the unselected small molecules were set to 0.

Considering the experimental costs and the possibility of clinical application, a limited number of small molecules is desired. To restrict the number of small molecules to be selected, the sub-objective function f_{number} was defined as follows:

$$f_{\mathrm{number}}(I_k) = \begin{cases} 1 & \text{if } \sum_{i \in I} \boldsymbol{v}(I_k)_i \le h \\ -\frac{\left(\sum_{i \in I} \boldsymbol{v}(I_k)_i\right)}{T^2} & \text{otherwise,} \end{cases}$$

where h is the maximum number of small molecules that should be included in each combination (h was set to 5 in this study). f_{number} works as a function that returns a value of 1 or <1 if the number of small molecules is below or above a specified threshold (h), respectively. As the number of selected small molecules increases, the value of f_{number} decreases.

Next, to evaluate the similarity between CIP and CSP, the sub-objective function f_{express} was defined using $\mathrm{DR}_{Imm/Pri}$ or $\mathrm{DR}_{Mat/Imm}$ as follows:

$$f_{\mathrm{express}}(I_k) = \mathrm{DR}_{Imm/Pri}\left(\boldsymbol{z}'_{Imm/Pri}, \boldsymbol{y}'_{Imm/Pri}\right) \text{or } \mathrm{DR}_{Mat/Imm}(\boldsymbol{z}'_{Mat/Imm}, \boldsymbol{y}'_{Mat/Imm})$$

f_{express} works as a function that returns a higher value if the gene expression pattern of the small molecule perturbation (CIP) resembles the gene expression pattern of the cell transformation state (CSP). Otherwise, it returns a smaller value if the CIP does not resemble the CSP.

Finally, to identify the optimal combination of small molecules for DR, the following combinational optimization problem is solved:

$$\mathrm{maximize} f_{\mathrm{total}}(I_k) = f_{\mathrm{express}}(I_k) + f_{\mathrm{number}}(I_k)$$

In this study, simulated annealing, which is one of the local search methods in meta-heuristics, was applied to acquire a solution to the combinatorial optimization problem. We propose the following algorithm:

Step 1: Arbitrarily choose k satisfying $1 \le k \le n$ and randomly select a set I_k of k small molecules as the initial state. In other words, generate an n-dimensional 0–1 vector $\boldsymbol{v}(I_k)$ with k elements of 1 and $(n-k)$ elements of 0. Then, calculate the objective function $f_{\mathrm{total}}(I_k)$.

Step 2: Randomly select whether to increase the number of small molecules by one to generate a neighborhood solution or to decrease the number of small molecules by one to generate a neighborhood solution.

(i) In the case of increasing small molecules by one:

Randomly select one of the small molecules that have not yet been selected and add it to the candidate combinations. In other words, randomly select one of the elements of $\boldsymbol{v}(I_k)$ that is 0 and change it to 1 and set the n-dimensional binary vector as the neighborhood solution $\boldsymbol{v}(I_k^*)$.

(ii) In the case of decreasing small molecules by one:

Randomly select one of the small molecules from the already selected small molecules and remove it from the candidate combination. In other words, randomly select one of the elements of $\boldsymbol{v}(I_k)$ that is 1 and change it to 0 and set the n-dimensional binary vector as the neighborhood solution $\boldsymbol{v}(I_k^*)$. Then, the objective function $f\left(I_k^*\right)$ is calculated.

Step 3: If $f_{\mathrm{total}}\left(I_k^*\right) > f_{\mathrm{total}}(I_k)$, set $I_k := I_k^*$ with probability 1. Otherwise, $I_k := I_k^*$ with the probability $\exp(-(f^*_{\mathrm{total}} - f_{\mathrm{total}}) \cdot \mathrm{iteration} / n)$.

Step 4: If a predetermined fixed time has elapsed, let I_k be the final solution. Otherwise, return to Step 2.

Here, the number of iterations was within 200,000. We defined the highest objective function score as the prediction score. Note that, in the original annealing method, the end temperature was set as the end condition of the algorithm, but the calculation time was set as the end condition of the algorithm in this study.

Interactions between small molecules and proteins

To acquire a list of target proteins regulated by DR-inducing small molecules, we collected a set of interactions between small molecules and proteins. These interaction data were acquired from public databases such as ChEMBL[43], MATADOR[44], DrugBank[45], BindingDB[46], KEGG DRUG, the Psychoactive Drug Screening Program Ki (PDSP-Ki)[39,47], and the Therapeutic Target Database[48]. This dataset included 1089 small molecules and 2217 proteins.

Evaluation of the biological functions of the target protein groups of small molecules

GO and KEGG pathway enrichment analyses were performed to elucidate the biological functions of the target proteins of the predicted small molecules. The Database for Annotation, Visualization and Integrated Discovery (DAVID) was used for GO analysis of the target proteins of the predicted small molecules. The top three GO terms in the annotation clusters that ranked in the Functional Annotation Clustering function with statistical significance ($p < 0.05$) were extracted. The enrichment p-values of all extracted GO terms were calculated using DAVID.

Visualization of the PPA network of the target proteins of DR-inducing small molecules

To elucidate the association network among the target proteins of the predicted small molecules, the protein–protein association (PPA) network was constructed from the molecular association network data stored in the Search Tool for the Retrieval of Interacting Genes (STRING) Database (https://www.string-db.org/)[49]. In STRING, PPAs are based on evidence such as experiments, databases, co-expression, neighborhood, gene fusion, and co-occurrence. We used a dataset of 12,684,354 PPAs of *Mus Musculus* in the STRING database, involving 21,840 proteins. We extracted the PPA network involving the target proteins of the predicted small molecules and visualized it with Cytoscape[50].

Furthermore, to investigate which groups of proteins were densely connected within the PPA network, we used MCODE[51], a clustering algorithm in Cytoscape, to detect subnetworks. All parameters were set to default values during the subnetwork detection process.

Statistical evaluation of predicted small molecules for DR

To evaluate the proportion of known DR-inducing small molecules or small molecules with similar functions to known DR-inducing ones against the number of all small molecules in the predicted combination, the statistical significance was evaluated using Fisher's exact probability test.

Data availability

Data sharing not applicable to this article as no datasets were generated during the current study.

Code availability

The source code for the prediction method reported in this manuscript is deposited in Zenodo (https://doi.org/10.5281/zenodo.18909775) and is freely available under a CC 4.0 license.

Received: 20 June 2025; Accepted: 16 March 2026;
Published online: 18 May 2026

References

1. Xu, J., Du, Y. & Deng, H. Direct lineage reprogramming: Strategies, mechanisms, and applications. *Cell Stem Cell* **16**, 119–134 (2015)
2. Wang, H., Yang, Y., Liu, J. & Qian, L. Direct cell reprogramming: approaches, mechanisms and progress. *Nat. Rev. Mol. Cell Biol.* **22**, 410 (2021).
3. Yoshihara, M. et al. Hotspots of de novo point mutations in induced pluripotent stem cells. *Cell Rep.* **21**, 308–315 (2017).
4. Heilker, R., Traub, S., Reinhardt, P., Schöler, H. R. & Sterneckert, J. iPS cell derived neuronal cells for drug discovery. *Trends Pharm. Sci.* **35**, 510–519 (2014).
5. Caiazzo, M. et al. Direct generation of functional dopaminergic neurons from mouse and human fibroblasts. *Nature* **476**, 224–227 (2011).
6. Isomi, M. et al. Direct reprogramming with Sendai virus vectors repaired infarct hearts at the chronic stage. *Biochem Biophys. Res Commun.* **560**, 87–92 (2021).
7. Xie, X., Fu, Y. & Liu, J. Chemical reprogramming and transdifferentiation. *Curr. Opin. Genet Dev.* **46**, 104–113 (2017).
8. Qin, H., Zhao, A. & Fu, X. Small molecules for reprogramming and transdifferentiation. *Cell. Mol. Life Sci.* **74**, 3553–3575 (2017).
9. Federation, A. J., Bradner, J. E. & Meissner, A. The use of small molecules in somatic-cell reprogramming. *Trends Cell Biol.* **24**, 179–187 (2014).
10. Fu, Y. et al. Direct reprogramming of mouse fibroblasts into cardiomyocytes with chemical cocktails. *Cell Res* **25**, 1013–1024 (2015).
11. Hu, W. et al. Direct conversion of normal and Alzheimer's disease human fibroblasts into neuronal cells by small molecules. *Cell Stem Cell* **17**, 204–212 (2015).
12. Nakamura, T. et al. Small compound-based direct cell conversion with combinatorial optimization of pathway regulations. *Bioinformatics* **38**, II99–II105 (2022).
13. Napolitano, F. et al. Automatic identification of small molecules that promote cell conversion and reprogramming. *Stem Cell Rep.* **16**, 1381–1390 (2021).
14. Zheng, M. et al. A single cell-based computational platform to identify chemical compounds targeting desired sets of transcription factors for cellular conversion. *Stem Cell Rep.* **18**, 131–144 (2023).
15. Hamano, M. et al. DIRECTEUR: transcriptome-based prediction of small molecules that replace transcription factors for direct cell conversion. *Bioinformatics* **25**, btae048 (2024).
16. Treutlein, B. et al. Dissecting direct reprogramming from fibroblast to neuron using single-cell RNA-seq. *Nature* **534**, 391 (2016).
17. Chen, G., Guo, Y., Li, C., Li, S. & Wan, X. Small Molecules that Promote Self-Renewal of Stem Cells and Somatic Cell Reprogramming. *Stem Cell Rev. Rep.* **16**, 511–523 (2020).
18. Cheng, J., Wang, S., Dong, Y. & Yuan, Z. The Role and Regulatory Mechanism of Hippo Signaling Components in the Neuronal System. *Front. Immunol.* **11**, 281 (2020).
19. Janesick, A., Wu, S. C. & Blumberg, B. Retinoic acid signaling and neuronal differentiation. *Cell. Mol. Life Sci.* **72**, 1559–1576 (2015).
20. Jogula, S. et al. Geldanamycin-inspired compounds induce direct trans-differentiation of human mesenchymal stem cells to neurons. *Eur. J. Med Chem.* **135**, 110–116 (2017).
21. Graichen, R. et al. Enhanced cardiomyogenesis of human embryonic stem cells by a small molecular inhibitor of p38 MAPK. *Differentiation* **76**, 357–370 (2008).
22. Mikkelsen, T. S. et al. Dissecting direct reprogramming through integrative genomic analysis. *Nature* **454**, 49–55 (2008).
23. Cheng, L. et al. Generation of neural progenitor cells by chemical cocktails and hypoxia. *Cell Res.* **24**, 665–679 (2014).
24. Meissner, A. et al. Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **454**, 766–770 (2008).

25. 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 (2008).
26. Zhao, Y. et al. High-efficiency reprogramming of fibroblasts into cardiomyocytes requires suppression of pro-fibrotic signalling. *Nat. Commun.* **6**, 1–15 (2015).
27. Xu, H., Wang, Y., He, Z., Yang, H. & Gao, W. Q. Direct conversion of mouse fibroblasts to GABAergic neurons with combined medium without the introduction of transcription factors or miRNAs. *Cell Cycle* **14**, 2451–2460 (2015).
28. Gascón, S. et al. Identification and Successful Negotiation of a Metabolic Checkpoint in Direct Neuronal Reprogramming. *Cell Stem Cell* **18**, 396–409 (2016).
29. Jiang, H. et al. Cell cycle and p53 gate the direct conversion of human fibroblasts to dopaminergic neurons. *Nat. Commun.* **6**, 1–14 (2015).
30. Huang, Z. J. & Scheiffele, P. GABA and neuroligin signaling: Linking synaptic activity and adhesion in inhibitory synapse development. *Curr. Opin. Neurobiol.* **18**, 77–83 (2008).
31. Huang, Z. J. Activity-dependent development of inhibitory synapses and innervation pattern: role of GABA signalling and beyond. *J. Physiol.* **587**, 1881–1888 (2009).
32. Greenberg, D. A. & Jin, K. From angiogenesis to neuropathology. *Nat. 2005 438:7070* **438**, 954–959 (2005).
33. Crouch, E. E. et al. Ensembles of endothelial and mural cells promote angiogenesis in prenatal human brain. *Cell* **185**, 3753–3769.e18 (2022).
34. Mahony, S. et al. Ligand-dependent dynamics of retinoic acid receptor binding during early neurogenesis. *Genome Biol.* **12**, R2 (2011).
35. Chen, X. et al. A fast chemical reprogramming system promotes cell identity transition through a diapause-like state. *Nat. Cell Biol.* **25**, 1146–1156 (2023).
36. Bergen, V., Lange, M., Peidli, S., Wolf, F. A. & Theis, F. J. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat. Biotechnol.* **38**, 1408–1414 (2020).
37. Subramanian, A. et al. A next generation connectivity map: L1000 platform and the first 1,000,000. *Profiles Cell* **171**, 1437–1452.e17 (2017).
38. 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 (2010).
39. Chan, J., Wang, X., Turner, J. A., Baldwin, N. E. & Gu, J. Breaking the paradigm: Dr Insight empowers signature-free, enhanced drug repurposing. *Bioinformatics* **35**, 2818–2826 (2019).
40. Kirkpatrick, S., Gelatt, C. D. & Vecchi, M. P. Optimization by simulated annealing. *Science* **220**, 671–680 (1983).
41. Černý, V. Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm. *J. Optim. Theory Appl.* **45**, 41–51 (1985).
42. Gaulton, A. et al. ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* 40, (2012).
43. Günther, S. et al. SuperTarget and matador: Resources for exploring drug-target relationships. *Nucleic Acids Res.* **36**, D919–D922 (2008).
44. Knox, C. et al. DrugBank 3.0: A comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res*. **39**, D1035–D1041 (2011).
45. Liu, T., Lin, Y., Wen, X., Jorissen, R. N. & Gilson, M. K. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.* **35**, D198–D201 (2007).
46. Roth, B. L., Lopez, E., Patel, S. & Kroeze, W. K. The multiplicity of serotonin receptors: Uselessly diverse molecules or an embarrassment of riches? *Neuroscientist* **6**, 252–262 (2000).
47. Qin, C. et al. Therapeutic target database update 2014: A resource for targeted therapeutics. *Nucleic Acids Res.* **42**, D1118–D1123 (2014).
48. Szklarczyk, D. et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res*. **43**, D447–D452 (2015).
49. Shannon, P. et al. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
50. Bader, G. D. & Hogue, C. W. V. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* **4**, 2 (2003).

Acknowledgements

This work was supported by JSPS KAKENHI [18K19930, 20H05797, 21K18327, and 24K21338], The Okawa Foundation for Information and Telecommunications, and JST ACT-X Grant Number JPMJAX24LD.

Author contributions

R.I., M.H., R.K., H.W. and A.M. analyzed data. R.I. prepared the manuscript. M.H. and Y.Y. contributed to supervision of the study and writing of the manuscript. All authors discussed the results and commented on the paper at all stages.

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/s42004-026-01991-y.

Correspondence and requests for materials should be addressed to Momoko Hamano or Yoshihiro Yamanishi.

Peer review information Communications Chemistry thanks Fang Bai and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.