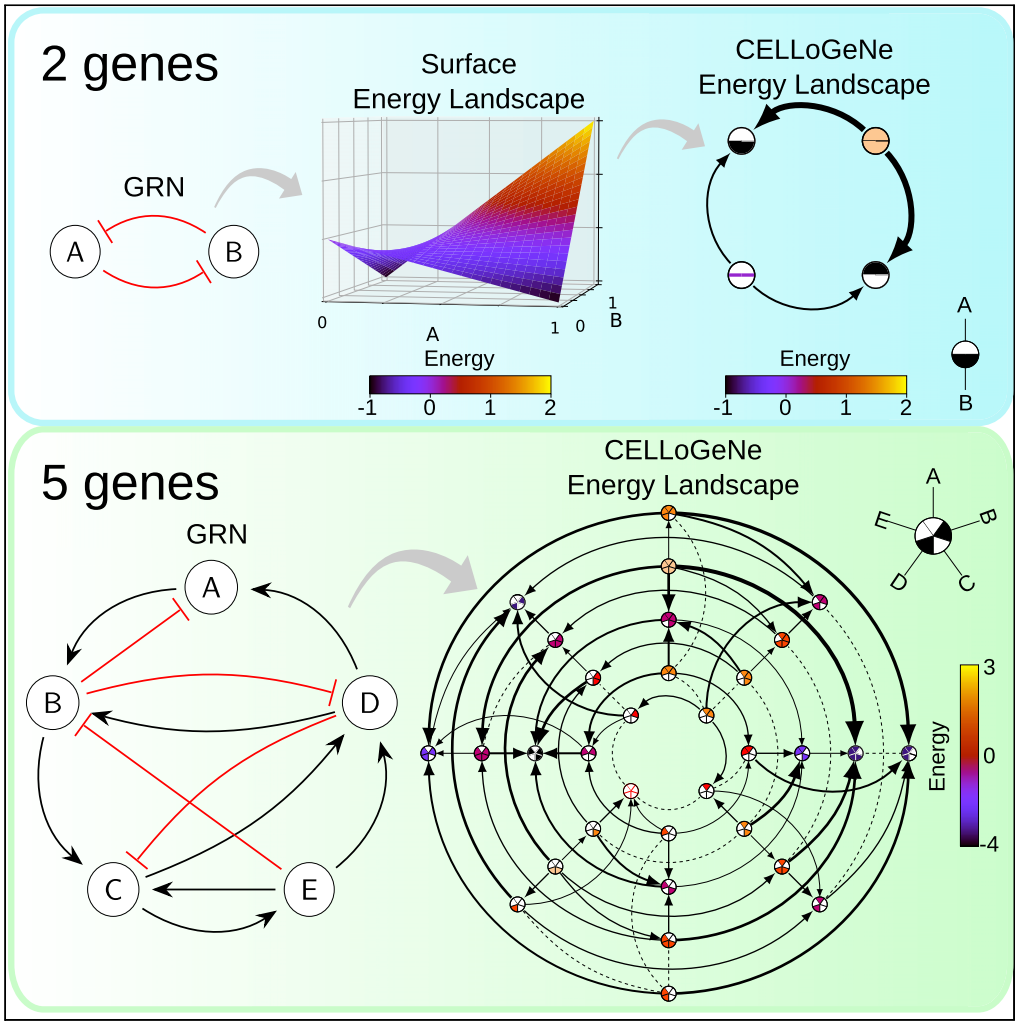
iScience

CellPress
OPEN ACCESS

Article

CELLoGeNe - An energy landscape framework for logical networks controlling cell decisions



Emil Andersson,
Mattias Sjö,
Keisuke Kaji,
Victor Olariu

victor.olariu@thep.lu.se

Highlights

CELLoGeNe – Computation of Energy Landscapes of Logical Gene Networks

Cell states as landscape attractors

Maintenance and acquisition of cell pluripotency applications

Single cell stochastic landscape navigation and visualization tool

Andersson et al., iScience 25, 104743
August 19, 2022 © 2022 The Authors.
https://doi.org/10.1016/j.isci.2022.104743

iScience

CellPress
OPEN ACCESS

Article

CELLoGeNe - An energy landscape framework for logical networks controlling cell decisions

Emil Andersson,[1] Mattias Sjö,[1] Keisuke Kaji,[2] and Victor Olariu[1,3,*]

SUMMARY

Experimental and computational efforts are constantly made to elucidate mechanisms controlling cell fate decisions during development and reprogramming. One powerful computational method is to consider cell commitment and reprogramming as movements in an energy landscape. Here, we develop Computation of Energy Landscapes of Logical Gene Networks (CELLoGeNe), which maps Boolean implementation of gene regulatory networks (GRNs) into energy landscapes. CELLoGeNe removes inadvertent symmetries in the energy landscapes normally arising from standard Boolean operators. Furthermore, CELLoGeNe provides tools to visualize and stochastically analyze the shapes of multi-dimensional energy landscapes corresponding to epigenetic landscapes for development and reprogramming. We demonstrate CELLoGeNe on two GRNs governing different aspects of induced pluripotent stem cells, identifying experimentally validated attractors and revealing potential reprogramming roadblocks. CELLoGeNe is a general framework that can be applied to various biological systems offering a broad picture of intracellular dynamics otherwise inaccessible with existing methods.

INTRODUCTION

The human body contains more than 200 different types of cells, which is a small number compared to the approximately 10^{13} cells in total (Alberts et al., 2008). The distinguishing feature between different cell types is which genes are activated, or *expressed*. In natural development, cells can only transition from pluripotent cells into more specialized cells, a process governed by changing gene expressions. However, experimental efforts have been made to reprogram cells, from specialized cells into pluripotent cells or other specialized cell types directly; for instance (Davis et al., 1987; Xie et al., 2004; Takahashi and Yamanaka, 2006; Takahashi et al., 2007; Zhou et al., 2008; Aasen et al., 2008; Kim et al., 2009; Loh et al., 2010; Szabo et al., 2010; Staerk et al., 2010; Vierbuchen et al., 2010; Ieda et al., 2010; Huang et al., 2011). Cell reprogramming is conducted by forcing expression of transcription factors (TFs); for recent overviews see (Aydin and Mazzoni, 2019; Wang et al., 2021). In particular, Takahashi and Yamanaka showed that it is possible to reprogram adult fibroblasts to induced pluripotent cells (iPSCs) by forcing the expression of just four TFs (Oct4, Klf4, Sox2, and c-Myc) in both mouse and human cells (Takahashi and Yamanaka, 2006; Takahashi et al., 2007). However, conversion rates are poor with many roadblocks in the reprogramming paths (O'Malley et al., 2013; Chantzoura et al., 2015). Many computational efforts have been made to elucidate the intricate gene regulatory networks (GRNs) governing cell decisions and reprogramming; for example (Reinitz et al., 1995; Gardner et al., 2000; Chen et al., 2000; Chickarmane et al., 2012; Dunn et al., 2014, 2019; Xu et al., 2014; Olariu et al., 2016,2017a,2017b).

The human genome is estimated to contain at least 30,000 genes (Roest Crollius et al., 2000), which either can be expressed at various levels or not expressed at all. Taking the whole genome with a continuous expression level into account when constructing a model for reprogramming is not feasible because of the vast space of possible gene expression patterns a cell can exhibit. Fortunately, there are quite a few simplifications that can be performed, which still yield good results. In a model, it is often possible to only consider a handful up to a few dozens of genes shown to play an important role in the reprogramming process, depending on the complexity of the proposed model. However, the gene expression space still has an infinite amount of points, leading to another common simplification, namely, binarizing the gene expression into either being OFF (0) or ON (1). This type of binary representation was pioneered and studied already in the '60s and '70s (Jacob and Monod, 1961; Sugita, 1963; Kauffman, 1969; Thomas,

[1]Computational Biology and Biological Physics, Department of Astronomy and Theoretical Physics, Lund University, Sölvegatan 14A, 221 00 Lund, Sweden

[2]Centre for Regenerative Medicine, University of Edinburgh, Edinburgh BioQuarter, 5 Little France Drive, Edinburgh EH16 4UU, UK

[3]Lead contact

*Correspondence: victor.olariu@thep.lu.se

https://doi.org/10.1016/j.isci.2022.104743

This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1973). With this simplification, the gene expression space is reduced to a finite space with 2^N states, where N is the number of genes used in the model. Such Boolean models have been previously explored, see (Mendoza et al., 1999; Fauré et al., 2006; Davidson, 2010; Peter et al., 2012).

Considering cell commitment and reprogramming as movements in an energy landscape is a powerful tool for analyzing decisions, which has been used in several previous studies (Bhattacharya et al., 2011; Wang et al., 2011; Zhou et al., 2012; Mojtahedi et al., 2016; Olariu et al., 2017a; Corson and Siggia, 2017; Sáez et al., 2021). Energy landscapes have their roots in a qualitative metaphor for cell fate decisions where Waddington envisaged pluripotent cells as marbles at the top of a hill, whereas cell differentiation was represented by them rolling down different available paths on the valleys and eventually stopping at a specialized state (Waddington, 1957). Energy landscapes are based on the same conceptual idea, but with a quantitative mapping from GRNs. Each theoretically possible cell state is assigned a specific energy value depending on the genes' activation statuses and how they are regulated according to the GRN. In physics, energy is a quantity that is as small as possible when a system is completely relaxed. Hence, low energies are awarded to cell states where the network is "comfortable", i.e., where the gene expressions and regulatory forces agree, and conversely for high energies. In this paradigm, cell states correspond to the minima in the landscape, i.e., attractors in the landscape where the metaphoric marbles come to a stop. Cell fate decisions are, hence, represented by the attractors' *basins of attraction*, which include all those states from which the marbles inevitably will roll down into the attractors; when a marble has entered a basin it cannot escape.

Here, we develop CELLoGeNe (Computation of Energy Landscapes of Logical Gene Networks), which maps Boolean implementation of GRNs into energy landscapes. Applying standard Boolean rules is accompanied by symmetries that may not be desired. CELLoGeNe solves this issue by implementing a three-state extension to Boolean logic where a gene's expression (ON or OFF) is decoupled from its effect on a target gene (positive (+1), negative (−1), or neutral (0), see STAR Methods-Logical representation and STAR Methods-Three-state logic.

An energy landscape is a function with as many dimensions as the number of genes (N) considered in the GRN. This leads to visualization challenges. CELLoGeNe provides a tool to plot multi-dimensional discrete energy landscapes (up to seven dimensions before the plot becomes too cluttered). This is achieved by considering each binary cell state as a corner on an N-dimensional hypercube, which is flattened out to two dimensions.

In order to analyze the strengths of the basins of attractions, we implemented a stochastic method that probes the shape of the energy landscape through weighted random walk. In essence, we release a large number of the metaphorical marbles at each cell state of the energy landscape, add a noise level, and record at which state they stop. This yields the probability of reaching the different attractors from each state of the landscape. By specifically analyzing how marbles may move from one attractor to another when noise is included, one effectively analyses cell development and reprogramming.

We applied CELLoGeNe to a GRN governing maintenance and self-renewal of pluripotency (Dunn et al., 2014). Here, we identify attractors that are experimentally validated (Dunn et al., 2014), establishing that energy landscapes computed by CELLoGeNe indeed can correspond to experimental findings. Thereafter, we applied CELLoGeNe to a GRN controlling reprogramming from mouse embryonic fibroblast (MEF) to iPSCs. The topology of the network was extracted from experimental and computational results published by our laboratories and others. We identified known cell states as attractors and a few additional attractors that predict the existence of potential bottlenecks in cell reprogramming.

CELLoGeNe is a general framework that can be applied to various biological systems controlled by a regulatory network. An important strength of CELLoGeNe is that it provides a broad picture of intracellular dynamics, as opposed to existing methods like solving systems of rate equations where it is hard to find all potential stable states. In fact, CELLoGeNe is not confined only to biology applications as it can be applied to any system that is governed by a logical regulatory network.

RESULTS

The state of the art of computationally studying cell reprogramming consists of two modeling strategies: (1) Dynamical systems approach – constructing rate equations for change in gene expression and numerically

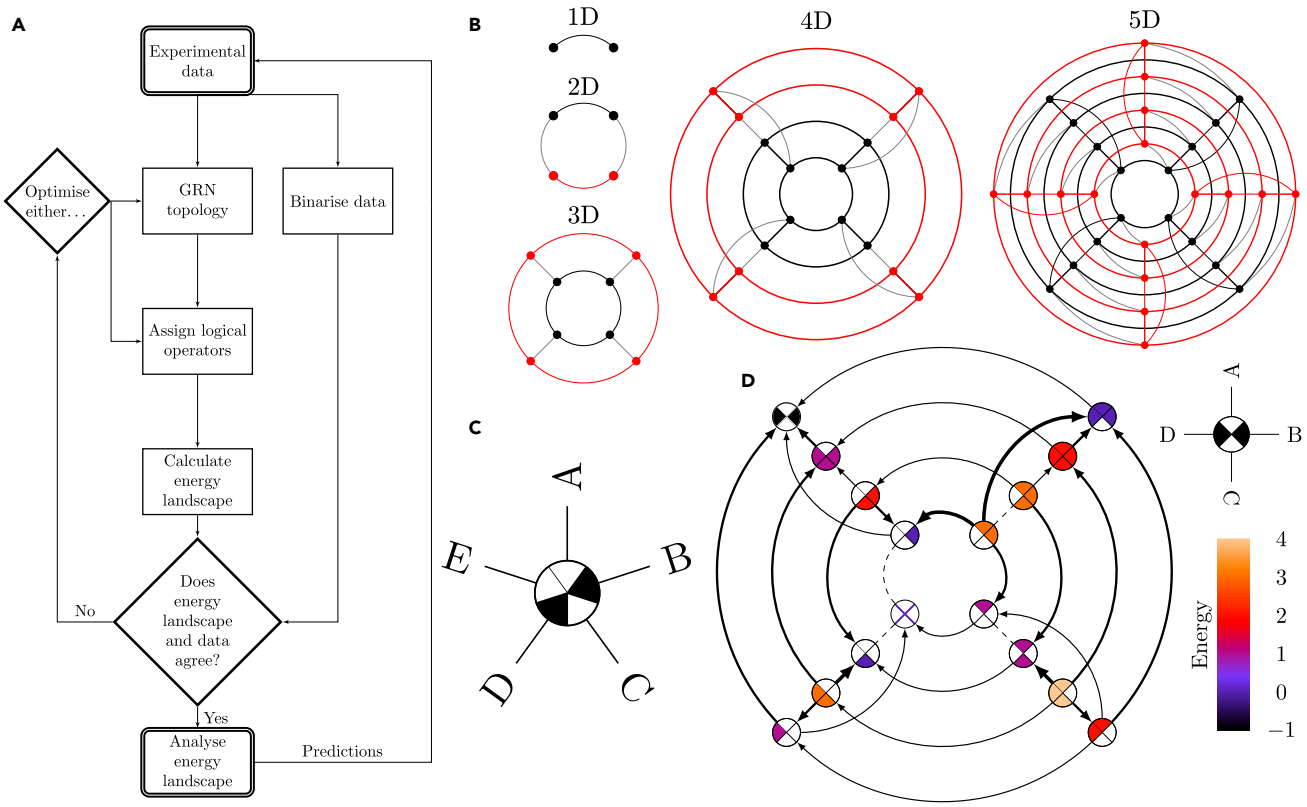


Figure 1. CELLoGeNe Workflow and Visualization of High-Dimensional Discrete Energy Landscapes

(A) Flowchart illustrating the main workflow of CELLoGeNe.

(B) Illustration of step-wise construction of an energy landscape in 1 to 5 dimensions, where the number of dimensions is the same as the number of genes in the corresponding GRN. A landscape of dimension N consists of two instances of the landscape of dimension $N - 1$, one black and one red, which are connected with the grey lines.

(C) A landscape node for a 5-dimensional energy landscape indicating the expression statuses of the 5 genes A to E, in this case A= OFF, B = ON, C = OFF, D = ON, E = OFF.

(D) Example of a full energy landscape plot for a synthetic 4-dimensional landscape. Each landscape node represents a possible state of the logical cell where the color of the node shows the energy value of that state, where the values are given by the color bar to the right. Each node is divided into four sectors, where the sectors correspond to the genes as indicated in the legend state above the color bar. Filled sectors correspond to expressed genes whereas white sectors correspond to unexpressed genes. The arrows connect neighboring states and point toward lower energies. The thickness of the arrows is proportional to the magnitude of the energy gradient. Dashed lines illustrate flat segments in the landscape, i.e., two neighboring states having the same energy value. The legend state to the right of the landscape indicates which sector corresponds to which gene and this is the same for all landscape nodes.

solve them finding stable states, (ii) Boolean approach – binarizing the GRN and logically updating the network states until a stable state is reached (Olariu and Peterson, 2019). These two approaches suffer from a drawback, namely, they do not provide an overview of all available stable states connected to a GRN. This drawback can be overcome by calculating free energy landscapes linked to the GRN. To this end, we developed CELLoGeNe, a method for translating Boolean models for GRNs into free energy landscapes. CELLoGeNe uses a three-valued logic which provides a better model of gene interactions when translated to the energy picture.

Overview of CELLoGeNe

The structural workflow of CELLoGeNe is illustrated in Figure 1A. In order to apply CELLoGeNe to a problem, experimental gene expression data and a GRN topology are needed. The data must contain the gene expressions for stable cell states which are binarized into gene expression ON (1) or OFF (0). These states should appear as attractors in the energy landscape in order for the energy landscape to be considered biologically plausible. The GRN may either be previously known, as they can be inferred from experiments (Bolouri and Davidson, 2002; Hecker et al., 2009), or hypothetical topologies can be tested through

CELLoGeNe and validated by experimental data. After the GRN topology is fixed, the logical combinations of operators combining multiple input signals at the GRN nodes are identified (Figure 1A, STAR Methods-Combining input signals with operators). This can either be done manually (by expert curation), or by letting CELLoGeNe test possible configurations, either exhaustively or stochastically (STAR Methods-Testing configurations of operators). After the operators have been assigned, the energy landscape is calculated by comparing the genes' binary expression values with their corresponding resulting input signals for each cell state (STAR Methods-The discrete energy). The minima in the energy landscape, i.e., the attractors, are compared to the binarized experimental data. If the binary expression profiles of the known cell states are not present as minima in the energy landscape, the search through possible combinations of logical operators is continued. However, if the search was exhaustive, then the GRN topology does not agree with the experimental data and should be further optimized. If the known cell states are present in the landscape, i.e., the data and energy landscape agree, it means that the landscape could be biologically plausible and can be further analyzed (Figure 1A). The analysis includes finding all minima in the energy landscape which provides predictions for cell states which could play a role as barriers in cell transitions between experimentally identified attractors. CELLoGeNe also provides a stochastic method for analyzing the basins of attraction for each landscape minimum (STAR Methods-Marble simulations).

Moreover, we developed a visualization tool, enabling us to depict energy landscapes with more than three dimensions (Figures 1B–1D). In essence, each cell state must be connected to all its neighboring cell states, where a neighboring state only differs in one gene expression value, i.e., a cell state has as many neighbors as genes in the GRN. Then it is only a matter to structure the states and their connections between neighbors in an ordered way, as shown in Figure 1B for 1 to 5 genes. Each cell state is represented by a node which is divided into as many sectors as the number of genes (Figure 1C). Each sector is either colored or white to represent if a sector's corresponding gene is ON or OFF. An example of a full energy landscape plot for a 4-gene system is shown in Figure 1D. Here, the color of a cell state indicates the energy value of that cell state, as mapped by the color bar to the right. Neighboring cell states are now connected by arrows where the arrow thickness is proportional to the energy difference between the states. If two neighboring cell states have the same energy, it is represented by a dashed line. The landscape plot is accompanied by a labeled cell state in the top right which shows which sector corresponds to which gene and it applies to all cell states. For full details on this visualization technique, see STAR Methods-Visualisation of high-dimensional energy landscapes.

Demonstration on a toy model

In the following section, we demonstrate CELLoGeNe by applying it to a toy model. We consider the GRN in Figure 2A, with five genes (nodes) *A* to *E* with (→) depicting activation and (⊣) repression. A five-gene GRN results in $2^5 = 32$ possible cell states. We use CELLoGeNe to compute an energy landscape. Given a configuration of operators, the energy of each cell state is calculated by adding together the contributions from each gene. Each gene in a cell state contributes with a low energy (-1) if its expression value agrees with the resulting input signal given by the logical operators and the input genes whereas a high energy ($+1$) is assigned in case of disagreement. In the case of no resulting input signal, we assign a neutral energy (0).

As outlined above, the two prerequisites for applying CELLoGeNe is to have a GRN and experimental data; however, because this is a toy model, we do not have any gene expression data. Hence, we introduced synthetic constraints by considering the states (*A* = OFF, *B* = ON, *C* = ON, *D* = OFF, *E* = ON) and (*A* = ON, *B* = OFF, *C* = ON, *D* = ON, *E* = ON) as attractors, depicted as blue and red in Figure 2C. Note that these states correspond to 10110 and 11101 respectively in the binary number representation (STAR Methods-Binary representation of genes). The next step is to assign logical operators to combine the input signals. For instance, gene *C* receives input both from gene *B* and *E* (Figure 2A), thus the dual signal must be combined with an operator into a single signal (STAR Methods-Combining input signals with operators). Different operators combine the effect of the input genes differently; therefore, the energy landscape depends on the operators used. For simplicity, we here only considered the two operators $\underline{\vee}$ and $\underline{\wedge}$ as defined in STAR Methods-Combining input signals with operators. CELLoGeNe constructed the energy landscape for each possible configuration of these two operators (STAR Methods-The discrete energy), and we picked one valid configuration which yielded an energy landscape fulfilling the synthetic constraints (Figure 2B). The full energy landscape is displayed in Figure 2D and the found attractors are summarized in Figure 2C. Each landscape node represents a cell state, and each cell state is denoted by a pie chart. If the

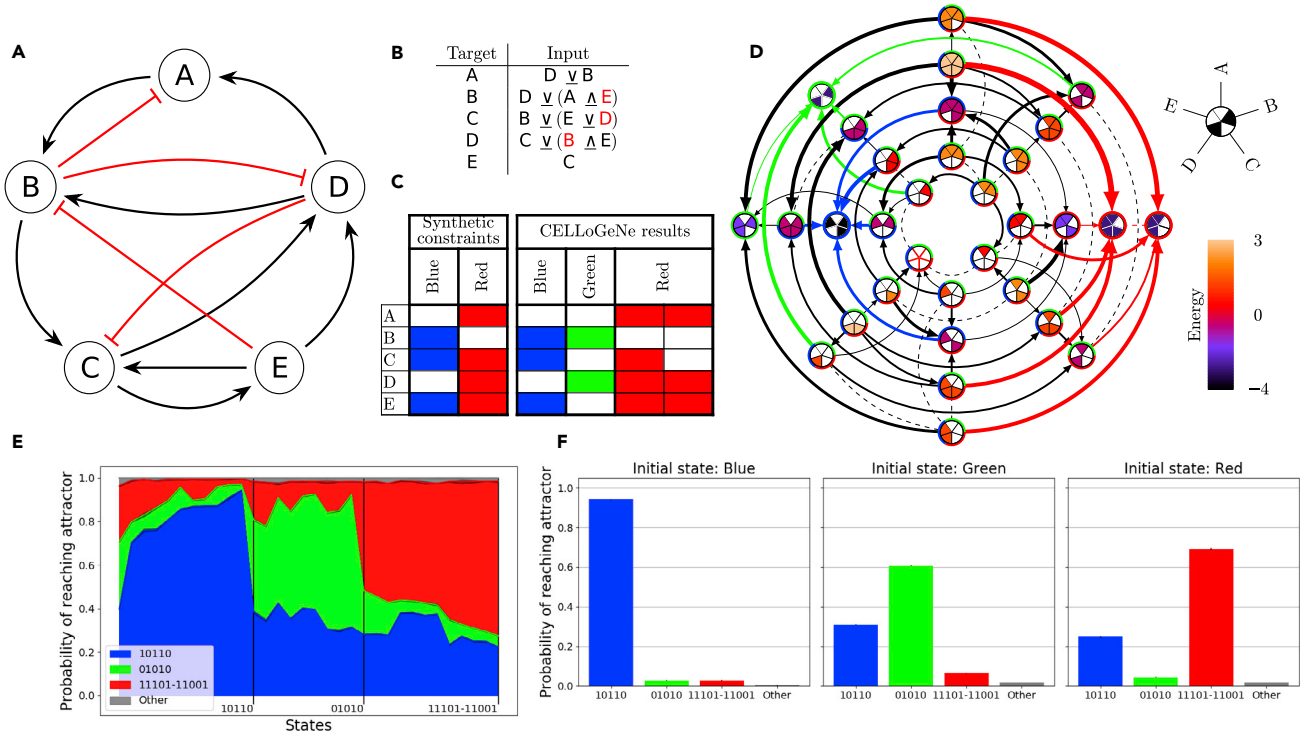


Figure 2. Applying CELLoGeNe on a Toy Model

(A) A GRN for a simple 5-gene toy model.

(B) Example of a configuration of logical operators yielding an energy landscape which fulfils a set of synthetic constraints.

(C) The synthetic constraints used and the resulting attractors for the energy landscape computed from the configuration in B.

(D) The energy landscape corresponding to the configuration of logical operators displayed in B. Each landscape node represents a possible state of the logical cell where the color of the node shows the energy value of that state, given by the color bar to the right. Each node is divided into five sectors, where each sector corresponds to the five genes as indicated by the legend above the color bar. Filled sectors correspond to expressed genes whereas white sectors correspond to unexpressed genes. The arrows connect neighboring states and point toward lower energies. The thickness of the arrows depends on the magnitude of the energy gradient. Dashed lines illustrate flat segments in the landscape, i.e., the two neighboring states having the same energy value. The arrows to the three attractors in the landscape are color-coded. The colored ring around each landscape node shows in which basins of attraction the node resides.

(E) Overview of the basin sizes and relative strengths given from marble simulations. The colored fields represent the probability of reaching the different attractors for all possible initial states. The initial states are displayed in a continuous manner on the horizontal axis. The plot is divided into three segments to highlight how large fraction of the initial states each of the attractors is the dominating attractor. The states are ordered with increasing probability of reaching the dominating attractor to the right. 10,000 marbles were initialized in each state with noise level $\beta = 1.0$. The simulations were repeated three times and the error band indicates the standard deviations.

(F) Simulated reprogramming with marble simulations using the same settings as in E. Each panel shows the probability distribution of reaching the different attractors when initialized in the blue, green, and red attractors respectively. The simulations were repeated three times and the error bars indicate the standard deviations.

piece of pie corresponding to, e.g., gene *A* (see the legend in the top right corner of Figure 2D) is filled with a color, it means that gene *A* is expressed. The color of the node reflects the energy value of that cell state, with energy levels given by the colour bar to the right in Figure 2D. Neighboring states are connected with arrows, where their thickness is proportional to the magnitude of the energy difference between those states. Dashed lines illustrate neighboring states with the same energy, i.e., flat areas of the energy landscape. For more details on the construction of landscape plots, see STAR Methods-Visualisation of high-dimensional energy landscapes.

In addition to the attractors given by the synthetic constraints (blue and red), we identified another attractor in the landscape (green). Because the green state is not previously known, CELLoGeNe predicts that a stable state with only genes *B* and *D* ON should exist for this system. In addition, the red state has a neighboring state with the same energy value, i.e., the state is a degenerate attractor. A degenerate attractor state suggests that the gene which can be either ON or OFF, in this case gene *C*, does not affect that

specific cell type. Because the blue attractor has lower energy than both the red and green attractors, it presumably is a stronger attractor. Note that energy -5 ($-N$ in general) means that all the genes' expression values in the cell state agree with their input signal and $+5$ ($+N$) means that all disagree.

The colored circle around each landscape node (Figure 2D) shows which attractors are possible to reach from the particular cell state by following the direction of the arrows or dashed lines, i.e., the basins of attraction. To investigate the basins' relative strengths, we developed a stochastic method to analyze the probability of reaching each attractor from each cell state (STAR Methods-Marble simulations). In principle, we let a metaphorical marble perform a weighted random walk in the landscape. At each step, the marble either rolls in one random direction, with a probability given by the difference in energy of the neighboring states, or does not move. One marble simulation corresponds to a single-cell dynamic fate. We introduced a noise level so that the marble, with a small probability, also can roll uphill. If the marble comes to a stop and stays in the same state for 3 updates, we consider the marble to have reached a final state. The stochasticity of marble simulations captures cell-to-cell heterogeneity. By repeating such simulations many times, the probability of reaching each attractor from a specific initial cell state is obtained. These stochastic simulations with noise can form a bulk accounting for population average experimental data allowing all attractors to be reached from each cell state, although, with different probabilities. For degenerate attractors, the probabilities of reaching each of the degenerate states are added. Note that the degenerative states should have the same probability of being reached because they have the same energy level. The result from these simulations is presented in Figure 2E shows the probability of reaching the different attractors from each initial state. In principle, each state could have been represented by a separate and labeled bar; however, because this kind of plot should be extendable to several thousands of cell states instead of 32, we chose to represent the cell states in a continuous and unlabelled manner. The plot is divided into three sections corresponding to different dominating attractors, e.g., the blue attractor is dominating in the left section. The initial cell states are ordered so that the probability of reaching the dominating attractor is increased to the right. This plot provides a qualitative illustration of the strength of each attractor. At the top of Figure 2E, there is a small grey band, which means that for a small fraction of simulations, the marble stopped at other states than any of the three attractors. This is because of noise in the simulations.

In cell reprogramming, a cell transition from one stable cell type to another. In our framework, this corresponds to the marble moving from one landscape attractor to another. The noise level perturbs the marble out from the attractor, which then either rolls back or away into another attractor (Figure 2F). Note that the three initial attractor states highlighted in Figure 2F are also present in Figure 2E. In this landscape, the blue attractor corresponds to a cell state which is more difficult to reprogram from. The green and red attractors represent cell states amenable to be reprogrammed towards the blue cell type.

Maintenance of naïve pluripotency

In this section, we apply CELLoGeNe to a GRN governing self-renewal and maintenance of naïve pluripotency in mouse embryonic stem (ES) cells. *In vitro*, mouse ES cells can self-renew indefinitely, without losing their differentiation capacity into all three germ layers, in certain culture conditions including medium containing leukemia inhibitory factor (LIF), inhibitor of glycogen synthase kinase 3 (CHIR99021, CH), and inhibitor of mitogen-activated protein (PD0325901, PD), where the latter two are together often referred to as two inhibitors (2i) (Ying et al., 2008). A minimal GRN controlling maintenance of mouse ES cells in four combinations of media *LIF + CH + PD*, *CH + PD*, *LIF + CH*, *LIF + PD* was put forward by (Dunn et al., 2014). Here, we apply CELLoGeNe to the published minimal GRN and use the available binarized expression data for each stable state for the four combinations of medium components (Figures 3A and 3C).

The GRN (Figure 3A) consists of 15 nodes: 12 TFs and 3 medium components acting as input nodes. In the CELLoGeNe framework, all of these are treated equally and are referred to as genes. In this case, there are $2^{15} = 32\,768$ possible cell states, which is considerably more than the 32 in the toy model. Furthermore, because most of the GRN nodes have multiple inputs, there are more than 28 trillion configurations of operators (the exact number of possible configurations is $28\,179\,280\,429\,056 \sim 10^{13}$) when using all 6 of the proposed operators (STAR Methods-Configurations of operators).

With this amount of possible configurations, it is not computationally feasible to exhaustively test all configurations; hence, we performed exhaustive searches for cases when we considered only three sets of two

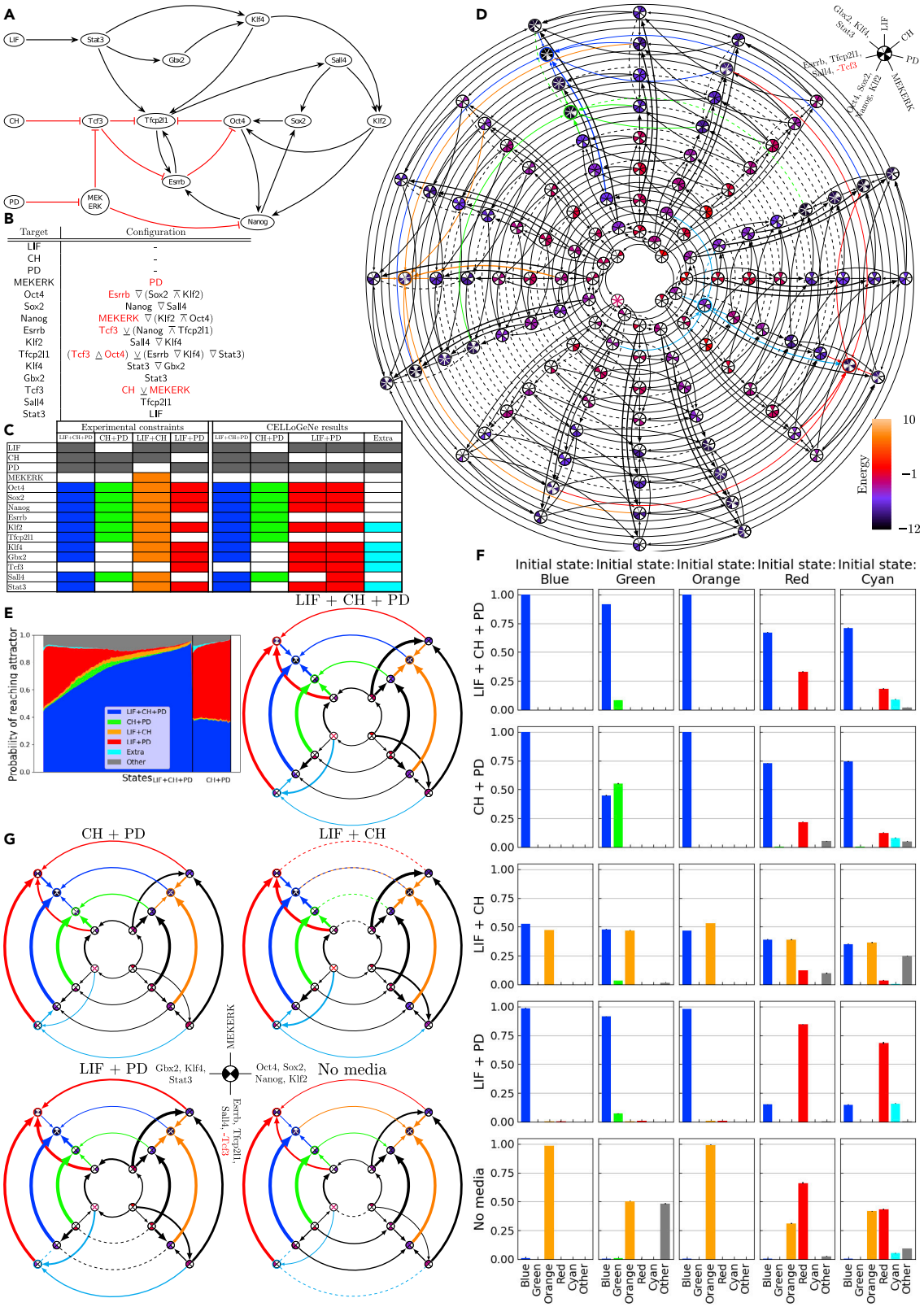

A
LIF
Stat3
Klf4
Gbx2
Sall4
CH
Tcf3
Tfcp2l1
Oct4
Sox2
Klf2
Esrrb
PD
MEK
ERK
Nanog
B
Target
Configuration
LIF -
CH -
PD -
MEKERK PD
Oct4 Esrrb ▽(Sox2 ⩓ Klf2)
Sox2 Nanog ▽Sall4
Nanog MEKERK ▽(Klf2 ⩓ Oct4)
Esrrb Tcf3 ⩒(Nanog ⩓ Tfcp2l1)
Sall4 ▽Klf4
Klf2 (Tcf3 △Oct4) ⩒(Esrrb ▽Klf4) ▽Stat3
Tfcp2l1 Stat3 ▽Gbx2
Klf4 Stat3
Gbx2 Tcf3 CH ⩒MEKERK
Sall4 Tfcp2l1
Stat3 LIF
C
Experimental constraints
CELLoGeNe results
LIF+CH+PD
CH+PD
LIF+CH
LIF+PD
Extra
D
Gbx2, Klf4, Stat3
LIF
CH
PD
MEKERK
Esrrb, Tfcp2l1, Sall4, -Tcf3
Oct4, Sox2, Nanog, Klf2
Energy
10
−1
−12
E
Probability of reaching attractor
States
LIF+CH+PD
CH+PD
LIF+CH
LIF+PD
Extra
Other
LIF + CH + PD
F
Initial state: Blue
Initial state: Green
Initial state: Orange
Initial state: Red
Initial state: Cyan
LIF + CH + PD
CH + PD
LIF + CH
LIF + PD
No media
Blue
Green
Orange
Red
Cyan
Other
G
CH + PD
LIF + CH
LIF + PD
Gbx2, Klf4, Stat3
MEKERK
Oct4, Sox2, Nanog, Klf2
No media
Esrrb, Tfcp2l1, Sall4, -Tcf3

Figure 3. Applying CELLoGeNe on a Network Governing Maintenance of Pluripotency

(A) GRN describing maintenance of pluripotency (Dunn et al., 2014). Black arrows represent activation, whereas blunted red arrows represent repression.

(B) Chosen configuration of logical operators.

(C) Experimental constraints (Dunn et al., 2014) and the resulting attractors from CELLoGeNe.

(D) Energy landscape where groups of genes have been collected into the same nodes (see legend in the top-right). Each landscape node represents a possible state of the logical cell where the color of the node shows the energy value of that state, where the values are given by the color bar to the right. Each node is divided into seven sectors, where the sectors correspond to the genes as indicated in the legend. Filled sectors correspond to expressed genes whereas white sectors correspond to unexpressed genes. The arrows connect neighboring states and point toward lower energies. The thickness of the arrows is proportional to the magnitude of the energy gradient. Dashed lines illustrate flat segments in the landscape, i.e., two neighboring states having the same energy value. The connections to the attractors of the landscapes are color-coded.

(E) Overview of the relative basin sizes and strengths given from stochastic marble simulations. The colored fields represent the probability of reaching the different attractors for all possible initial states. The initial states are displayed in a continuous manner on the horizontal axis. The plot is divided into two segments to highlight how large fraction of initial states each of the attractors is the dominating attractor. The states are ordered with increasing probability of reaching the dominating attractor to the right. 100 marbles per initial state were simulated with noise-level $\beta = 1.5$ and the simulations were repeated three times.

(F) Simulated differentiation corresponding to transferring a stable cell population to a different medium. Each panel shows the distribution of final attractors reached by marbles initialized from the different attractors. The marbles could not change media in these simulations. Noise-level $\beta = 1.5$ was used with 10 000 marbles per initial state and the simulations were repeated three times. The error bars indicate the standard deviations.

(G) The five disconnected energy landscapes illustrating differentiation of cells transferred between media.

operators $\{\underline{\vee}, \underline{\wedge}\}$, $\{\overline{\vee}, \overline{\wedge}\}$ and $\{\uparrow, \downarrow\}$. Moreover, we stochastically tried 10^6 configurations with all operators. No configurations from the three two-operator sets yielded energy landscapes where all four experimental constraints (Figure 3C) were fulfilled. In the stochastic search, the state in *LIF* + *CH* + *PD* was the most prevalent attractor occurring in 46 % of the 110^6 calculated energy landscapes (Figure S1A). Thus, the GRN seems extremely robust when it comes to maintaining pluripotency in *LIF* + *CH* + *PD*. Also the stable state in *LIF* + *CH* was one of the 10 most prevalent attractors, occurring in 23 % of the landscapes. In total, 163 valid configurations were found. Three of them (Figure S1B) were deemed more plausible than the others because they had lower degeneracy in the attractors. However, none of these configurations were biologically satisfactory when the operator combinations were scrutinized. Firstly, all three configurations required that both *CH* and *MEKERK* are ON in order to repress *Tcf3*. This is not reasonable because the presence of *PD* inhibits *MEKERK*, therefore leaving *CH* without any effect on *Tcf3*, which is against the known behavior of *CH*. Secondly, in some instances, repressive input is effectively used as activation. An example of this can be seen in the input for *Oct4* in Configuration 2: $Oct4 \leftarrow (Sox2 \quad \downarrow \quad Klf2)\overline{\wedge}Esrrb$. Here, *Oct4* can only receive an activation signal if all three input genes are ON, even the repressive *Esrrb*. There is no way to receive repression of *Oct4* from *Esrrb* with this configuration. Thirdly, in some cases, an inhibitor and an activator are combined with an $\underline{\wedge}$-operator, i.e., both the activator and inhibitor must be ON in order for the target gene to receive a repressive signal. An example of this can be seen in Configuration 3 for $Tfcp2l1 \leftarrow \left(\left(Stat3\overline{\vee}\left(Esrrb\underline{\wedge}Oct4\right)\right)\underline{\wedge}Klf4\right)\underline{\vee}Tcf3$, where both the activating *Esrrb* and repressive *Oct4* must be ON for *Tfcp2l1* to be repressed by *Oct4*, effectively turning *Esrrb* into a repressor.

To avoid unreasonable configurations, we manually curated a configuration (Figure 3B), where none of the repressors effectively worked as activators or needed an activator in order to repress its target, and where *CH* had the authority to inhibit *Tcf3* on its own. Although the curated configuration is not unique, the alternative configurations respecting all the constraints will produce similar landscapes qualitatively. The resulting attractors are presented in Figure 3C. Note that the experimentally stable state in *LIF* + *CH* (orange) is not an attractor in the energy landscape. Nonetheless, we do not consider this to be a problem. The GRN (Figure 3A) cannot produce the orange state as a single attractor while still following the reasoning about repressors above because it is very close to the stable state in *LIF* + *CH* + *PD* (blue). The only non-media gene that differ between the experimentally stable states blue and orange is *MEKERK* (*MEKERK* = OFF in blue and MEKERK = ON in orange). Thus, when *PD*= OFF, then *MEKERK* is free and can be either ON or OFF with the same energy, which corresponds to the same state as blue except in another medium. This is clearly illustrated in the energy landscape plot (Figure 3D) (in order to not get an incomprehensibly large plot, we merged groups of genes into shared dimensions as denoted by the legend to the right in Figure 3D). Here, it is clear that the orange state in *LIF* + *CH* and the blue state in *LIF* + *CH* + *PD* only differ

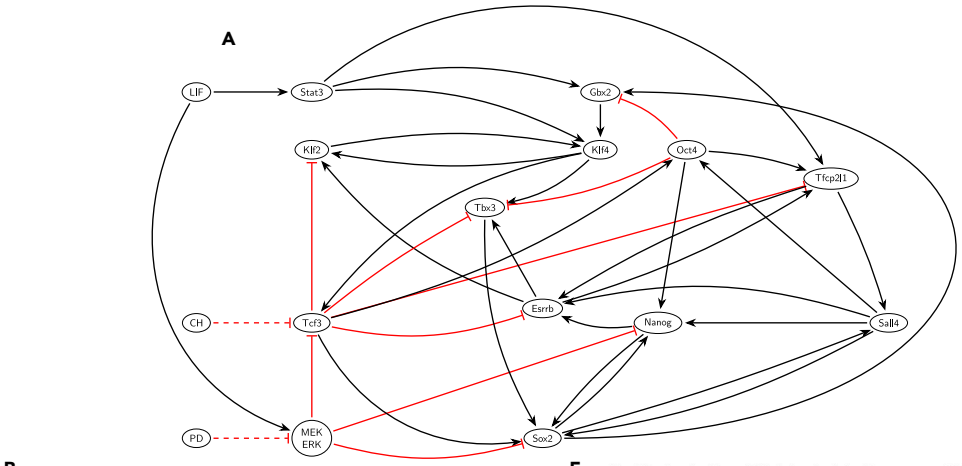
by two genes, and the landscape is flat for changing *MEKERK* (orange dashed line in the direction of 11 o'clock). Then changing medium from *PD*= OFF to *PD*= ON is favorable in energy but cannot happen spontaneously in experiments. Therefore, the transition between orange and blue states is not possible automatically because it implies a media change which cannot be conducted by the cell. In Figures 3F and 3G we instead simulated the scenario where a cell colony is manually transferred between different media.

The manually curated configuration predicts an additional stable state, labeled "Extra" with cyan color in Figure 3. This state is in *LIF* + *PD* medium but is not a pluripotent state. Only *Klf2* of the pluripotent genes (the pluripotent genes are here considered to be *Oct4*, *Sox2*, *Nanog*, and *Klf2*) is ON. However, this attractor is very weak as shown by the marble simulations (Figure 3E). The strongest attractor is the blue state in *LIF* + *CH* + *PD* media, which can be reached from all initial states and is the dominating attractor for approximately 80% of the states. The red attractor in *LIF* + *PD* is the dominating attractor for around 20% of the states. The orange *LIF* + *CH* state is fairly even, but with a low probability of being reached from all initial states. The green *CH* + *PD* attractor, on the other hand, is quite weak and can practically only be reached from about half of the states, suggesting that it is easier to maintain the pluripotency with *LIF* present. The fact that the blue attractor is found to be the strongest one is in good agreement with the results from (Dunn et al., 2014) where it was shown that ES cells cultured in *LIF* + *CH* + *PD* are remarkably robust. Over the whole domain, there is a 5 to 15% probability of reaching other states than the attractors. This probably is a sign that large regions of the energy landscapes are flat. This might correspond to spontaneous exit from pluripotency towards somatic or other unidentified cell fates.

Finally, we performed simulations representing transferring a cell colony cultured in one medium to another as done experimentally in (Dunn et al., 2014). This is done with the marble simulations, but with constraints prohibiting the marble to roll in a direction that changes the medium. The simulations take place on different, disconnected, parts of the energy landscape (Figure 3G), which can be thought of as an *in silico* version of putting cell cultures on a plate containing a specific medium *in vitro*. Simulations were performed for each of the 25 possible combinations of initial attractors and media (including absence of medium). The resulting attractor distributions are displayed in Figure 3F where the medium is the same in each row, and the initial state is the same in each column. This model simulation predicted that the blue state is the strongest attractor, even in other media than *LIF* + *CH* + *PD*. The reason for this can be seen in the energy landscapes with disconnected media (Figure 3G). Blue is the sole minimum in three of the media, and a degenerate minimum together with orange in *LIF* + *CH*. It is only in the landscape with no medium where blue is not a minimum, being replaced by orange.

Merging groups of genes for visualization purposes has the disadvantage that the resulting landscape plot is not an exact representation of the energy landscape as the neighboring nodes in the merged landscape are not neighboring cell states (neighboring cell states have Hamming distance 1, e.g. the states 10010 and 10011 in a 5-gene system). For this reason, the green state, for instance, does not appear as a minimum in the *CH* + *PD* landscape. From the green state, there is lower energy in the neighboring state where *Gbx2*, *Klf4* and *Stat3* are all ON; however, in the full landscape, a neighboring state only changes one gene's expression. This is also the reason why the cyan attractor has an arrow pointing towards the red attractor in Figure 3D; the cyan and red attractors are neighbors in the merged landscape, but not in the full. Even though these merged landscapes do not reveal connections between direct neighbors, they are still very useful for an illustrative purpose as they show the main directions in the landscape, which is important in the stochastic marble simulations.

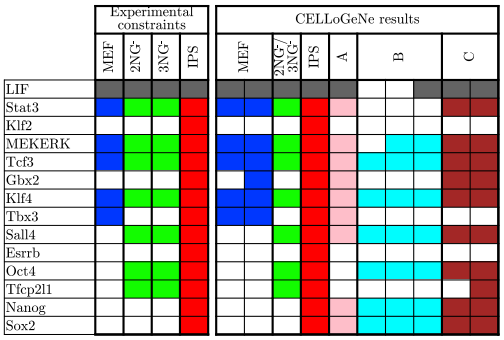
The disconnected landscapes (Figure 3G) explain the attractor distributions. For instance, the green attractor is stronger for the green initial state in *CH* + *PD* than in *LIF* + *CH* + *PD*. Comparing the corresponding landscapes, there is a larger energy difference between the green and blue state in *LIF* + *CH* + *PD* than in *CH* + *PD*, illustrated both by the color scale and the thickness of the arrow, yielding higher probability for cells to move from the green state to the blue. The situation is similar for the red and blue states in *LIF* + *PD* compared to the other media. The orange and blue states are degenerate in *LIF* + *CH* (Figure 3G - blue and orange dotted line), which corresponds to the third row of Figure 3F where reaching the blue and orange states has the same probability, with a slight bias towards the initial state. The orange state is the strongest attractor state when simulating the no-medium scenario. Experimentally, only *MEKERK* should be ON in the absence of media (Dunn et al., 2014). However, the presence of the orange state as an



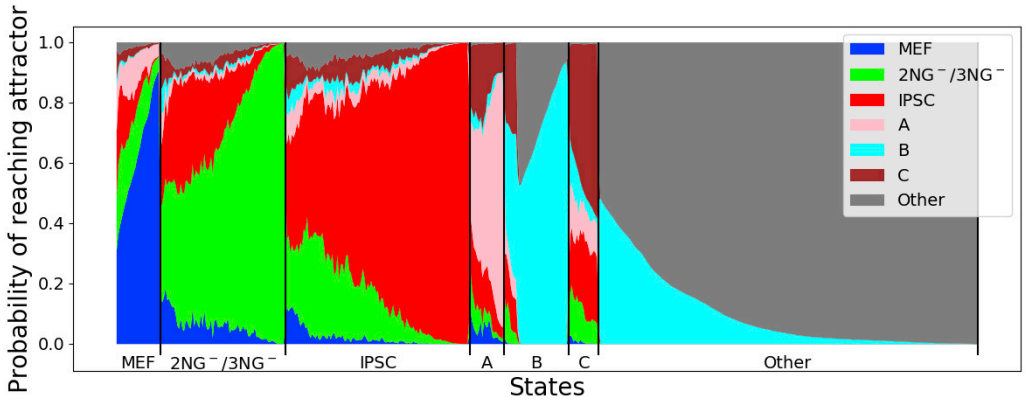
B

Target	Configuration
LIF	-
Stat3	LIF
Klf2	(Klf4 ⩓ Esrrb) ▽ Tcf3
MEKERK	LIF
Tcf3	Klf4 ▽ MEKERK
Gbx2	(Stat3 ⩓ Sox2) ▽ Oct4
Klf4	Stat3 ▽ (Klf2 ⩓ Gbx2)
Tbx3	Klf4 ⩒ ((Tcf3 △ Oct4) ▽ Esrrb)
Sall4	Sox2 ▽ Tfcp2l1
Esrrb	((Sall4 ⩓ Tfcp2l1) ⩓ Nanog) ▽ Tcf3
Oct4	Sall4 ⩓ Tcf3
Tfcp2l1	(Stat3 ⩓ (Oct4 ▽ Esrrb)) ▽ Tcf3
Nanog	((Oct4 ▽ Sall4) ⩓ Sox2) ▽ MEKERK
Sox2	(Tcf3 ⩓ (Nanog ▽ (Tbx3 ⩓ Sall4))) ▽ MEKERK

C



D



E

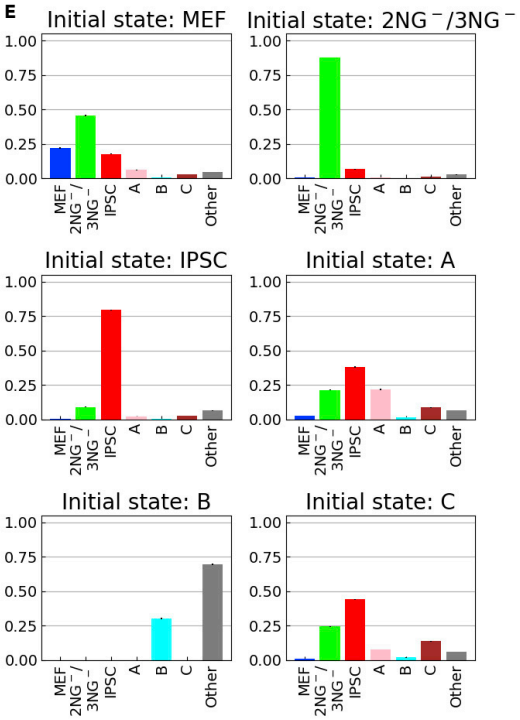


Figure 4. Applying CELLoGeNe on a Network Controlling Reprogramming from MEF to iPSC

(A) GRN describing reprogramming of fibroblasts to iPSC. Black arrows represent activation, while blunted red arrows represent repression.

(B) Chosen configuration of logical operators.

(C) Experimental constraints (O'Malley et al., 2013) and attractors of the energy landscape.

(D) Overview of the relative basin sizes and strengths given from stochastic marble simulations. The colored fields represent the probability of reaching the different attractors for all possible initial states. The initial states are displayed in a continuous manner on the horizontal axis. The plot is divided into segments to highlight how large fraction of the initial states each of the attractors is the dominating attractor. The states are ordered with increasing probability of reaching the dominating attractor to the right. 1000 marbles per initial state were simulated with noise-level $\beta = 1.5$ and the simulations were repeated three times.

(E) Simulated reprogramming with each of the found attractors as initial states. The height of the bars represents the probability of ending up in an attractor. The marble simulations were run with $\beta = 0.7$ and 10,000 marbles per initial state and were repeated three times. The error bars indicate the standard deviations.

attractor in the absence of media can be explained. In CELLoGeNe, when no medium components are present, both *MEKERK* and *Stat3* receive a neutral input signal. Then, it is clearly beneficial for *MEKERK* to be *ON* and *Tcf3* to be OFF because that agrees with *MEKERK* ⊣ *Tcf3*. If all other genes are *OFF* (i.e., the cell state with only *MEKERK*= ON), then only *Tcf3* and *Nanog* receive input signal and the cell state has energy − 2. The local landscape area around this state is largely flat because many of the genes have no input signal. Turning on *Stat3* would in itself not yield lower energy, however, the state with all downstream targets from *Stat3* turned ON would be beneficial, which corresponds to the orange state. It should be noted that for the green initial state within no medium, CELLoGeNe identified a large fraction of "Other" states. This is probably because of the green state already having many genes OFF, thus being close to the large flat region around only *MEKERK*= ON.

When applied to a regulatory circuitry controlling self-renewal, CELLoGeNe uncovers the energy landscape containing attractors that correspond to experimentally observed stem cell states under various media. We managed to compress a multitude of experimental results into one framework, which provides further details such as: (1) Assigned probabilities to possible stem cell states; (2) Identified most likely destinations when perturbing a cell in a specific medium; (3) Provided plausible explanations for leakage from pluripotency.

Reprogramming MEF to iPSC

A natural application of CELLoGeNe is analysis of cell reprogramming systems. Here, we apply CELLoGeNe to a GRN governing the reprogramming of mouse embryonic fibroblasts (MEF) to induced pluripotent stem cells (iPSCs). Reprogramming experimental protocols have low conversion rates because of roadblocks preventing efficient reprogramming (O'Malley et al., 2013; Chantzoura et al., 2015). Using CELLoGeNe, we can analyze the full energy landscape and find unknown potential bottlenecks. We first applied CELLoGeNe to a GRN which has been updated from governing self-renewal and reprogramming in Dunn et al. (2019). However, it was not possible to find any configuration of logical operators that fitted our data. We used experimental data from four stable states MEF, $2NG^-$, $3NG^-$, and iPSC, where cells in the $2NG^-$ and $3NG^-$ states tend to stay in the same states without progressing towards iPSCs (O'Malley et al., 2013). After binarizing the data we found that $2NG^-$ and $3NG^-$ correspond to the same cell state (Figure 4C). Because CELLoGene applied to the GRN in (Dunn et al., 2019) could not find the stable states from our data, we augmented the GRN with interactions found in literature concerning gene expression and perturbation experiments of reprogramming systems, see Figure 4A and Table 1. Armed with our resulting network, we succeeded in finding configurations yielding valid energy landscapes.

Our GRN (Figure 4A and Table 1) consists of 14 nodes, 13 TFs and 1 medium component, *LIF*, yielding $2^{14} = 16\,384$ possible states. Since only *LIF* was used in our experiments (O'Malley et al., 2013), *CH* and *PD* inputs were not considered. Because of many input signals for each gene, it becomes computationally unfeasible to exhaustively test all configurations because of the prohibitively large number (the exact number is 26 142 282 979 403 407 520 956 416 ($\sim 10^{25}$)) of possible configurations when using all 6 operators. Hence, we explored the configuration space stochastically, testing 10^6 configurations, where a total of 2 valid configurations were found (Figure S2B).

We found that the iPSC attractor is one of the 10 most prevalent minima in the 10^6 calculated energy landscapes (Figure S2A), occurring in 7.1% of the landscapes (as reference, the most prevalent minimum occurs in approximately 16% of the energy landscapes). Thus, the GRN is robust in its ability to reprogram to iPSCs

Table 1. Components and sources for the gene regulatory network governing reprogramming from MEF to iPSC.

From	To	Effect	Reference(s)
LIF	Stat3	activate	Dunn et al. (2019); Niwa et al. (2009)
Klf4	Klf2	activate	Dunn et al. (2014, 2019)
Esrrb	Klf2	activate	Xu et al. (2013, 2014); Yeo et al. (2014); Chen et al. (2008); Dunn et al. (2019)
Tcf3	Klf2	repress	Qiu et al. (2015)
LIF	MEKERK	activate	Graf et al. (2011)
MEKERK	Tcf3	repress	Dunn et al. (2019)
Klf4	Tcf3	activate	Zhang et al. (2013); O'Malley et al. (2013)
Stat3	Gbx2	activate	Tai and Ying (2013); Dunn et al. (2019)
Sox2	Gbx2	activate	Xu et al. (2013, 2014)
Oct4	Gbx2	repress	Xu et al. (2013, 2014); Dunn et al. (2019)
Stat3	Klf4	activate	Dahéron et al. (2004); Bourillot and Savatier (2010); Dunn et al. (2019)
Klf2	Klf4	activate	Dunn et al. (2019)
Gbx2	Klf4	activate	Wang et al. (2017); Xu et al. (2013, 2014)
Klf4	Tbx3	activate	Dunn et al. (2019)
Tcf3	Tbx3	repress	Tam et al. (2008); Dunn et al. (2019)
Oct4	Tbx3	repress	Tam et al. (2008); Xu et al. (2013, 2014)
Esrrb	Tbx3	activate	Adachi et al. (2018)
Tfcp2l1	Sall4	activate	Tanimura et al. (2013); Dunn et al. (2014, 2019)
Sox2	Sall4	activate	Dunn et al. (2019)
Tfcp2l1	Esrrb	activate	Wang et al. (2019); Dunn et al. (2014, 2019)
Nanog	Esrrb	activate	Festuccia et al. (2012); Heurtier et al. (2019); Dunn et al. (2019)
Sall4	Esrrb	activate	Tatetsu et al. (2016); Dunn et al. (2019)
Tcf3	Esrrb	repress	Martello et al. (2012); Dunn et al. (2019)
Sall4	Oct4	activate	Tanimura et al. (2013); Ho et al. (2013); Dunn et al. (2019)
Tcf3	Oct4	activate	Ho et al. (2013); Dunn et al. (2019)
Stat3	Tfcp2l1	activate	Martello et al. (2013); Dunn et al. (2019)
Esrrb	Tfcp2l1	activate	Wang et al. (2019); Dunn et al. (2019)
Tcf3	Tfcp2l1	repress	Qiu et al. (2015); Martello et al. (2012); Ye et al. (2017); Dunn et al. (2019)
Oct4	Tfcp2l1	activate	Martello et al. (2013); Stirparo et al. (2021)
Oct4	Nanog	activate	Rodda et al. (2005); Boyer et al. (2005); Loh et al. (2006); Olariu et al. (2016); Dunn et al. (2019)
Sall4	Nanog	activate	Zhang et al. (2006); Yang et al. (2008); Lim et al. (2008); Tatetsu et al. (2016); Dunn et al. (2019)
Sox2	Nanog	activate	Rodda et al. (2005); Masui et al. (2007); Chen et al. (2008); Dunn et al. (2019)
MEKERK	Nanog	repress	Ying et al. (2008); Martello et al. (2013); Hamilton and Brickman (2014); Dunn et al. (2019)
Tbx3	Sox2	activate	Ivanova et al. (2006); Esmailpour and Huang (2012); Dunn et al. (2019)

(*Continued on next page*)

Table 1. *Continued*

From	To	Effect	Reference(s)
Nanog	Sox2	activate	Rodda et al. (2005); Boyer et al. (2005); Chickarmane et al. (2006, 2012); Dunn et al. (2019)
Sall4	Sox2	activate	Tanimura et al. (2013); Tatetsu et al. (2016); Dunn et al. (2019)
Tcf3	Sox2	activate	Yi et al. (2008); Dunn et al. (2019)
MEKERK	Sox2	repress	Hamilton and Brickman (2014); Dunn et al. (2019)

with respect to different configurations of operators. The MEF and $2NG^{-}/3NG^{-}$ attractors, on the other hand, only occur in 0.2% and 0.7% of the landscapes, respectively. Thus, the combined probability of randomly finding an energy landscape containing all three attractors is approximately one per million. This agrees with finding two valid configurations in the stochastic search. It is not surprising that MEF has a low probability of occurring because our GRN governs transitions from the MEF state. However, the CELLoGeNe energy landscape needs to exhibit an attractor corresponding to MEF cells as this is the starting cell state in our experiments. This leads to important constraints on resulting operator configurations. We used a similar strategy as in the previous result section, constructing a manual configuration which we deemed to be biologically plausible (Figure 4B). This configuration yielded the required attractors, where MEF is a two-state degenerate attractor whereas $2NG^{-}/3NG^{-}$ and iPS are both single state attractors (Figure 4C). Also three additional attractors were found, which we refer to as A, B and C (Figure 4C).

We performed marble simulations from each possible state without changing *LIF* from its initial value in order to probe the relative strengths and sizes of the basins of attraction (Figure 4D). We observed dominating attractors corresponding to MEF, $2NG^{-}/3NG^{-}$ and iPSC states, with the iPSC attractor being the largest whereas MEF having the smallest dominating region. We also observed smaller regions where attractors A, B and C are the dominating ones. This reveals that all attractors' basins have a region where they are the strongest, as opposed to the landscape for maintenance of pluripotency (Figure 3E). The fact that all of the three additional attractors A, B and C have regions where they dominate indicates that they all can act as bottlenecks. Figure 4D also shows that from around 40% of the energy landscape states, i.e., the section furthest to the right, it is close to impossible to reach any of the experimentally known attractors. Attractor B and "Other" states are the only reachable attractors from this region. This region corresponds to those states where *LIF* is not present demonstrating that it is not possible to reach and maintain pluripotency without the adequate medium.

We conducted cell reprogramming simulations starting from the landscape's six attractors (Figure 4E). When the cells were initialized in MEF, approximately a quarter of the cells stayed in MEF, barely half transitioned to $2NG^{-}/3NG^{-}$ and a fifth transitioned directly to iPSC. The remaining cells ended up in A, C or other. Most of the cells initialized in $2NG^{-}/3NG^{-}$ stayed there while the fraction of cells transitioning to iPSC was increased with a higher noise level (Figure S3). When the cells were initialized in iPSC, most of the cells stayed in iPSC and a small fraction transitioned to $2NG^{-}/3NG^{-}$. These three simulation results recapitulate experimental results observed in (O'Malley et al., 2013). The relative height of each bar varies with noise level (Figure S3), however, the general behavior remains the same.

We performed a deeper analysis of the potential bottlenecks, by comparing the gene expressions for attractor states. Attractor A has the same genes ON as MEF, except three additional expressed genes: *Sall4*, *Nanog* and *Sox2* (Figure 4C). Therefore, attractor A seems to be a roadblock when transitioning from MEF towards iPSC as more genes are turned ON. If the genes are turned ON in the wrong order, the cell risks getting stuck in attractor A. This could potentially happen if *Sall4* is activated which could turn ON*Sox2* and in turn activate *Nanog*, by following the logic of the operator configuration of the GRN (Figures 4B and 4A). If, instead, the positive feedback loop with *Oct4*, *Tfcp2l1* and *Sall4* is activated, then *Oct4* could turn OFF*Tbx3* and the cells get stuck in $2NG^{-}/3NG^{-}$. If a few additional genes are also activated (*Gbx2*, *Nanog* and *Sox2*), then the cells get stuck in attractor C. Hence, it seems like attractor A mainly is a roadblock between MEF and iPSC, just like $2NG^{-}/3NG^{-}$, whereas attractor C seems to be a roadblock between $2NG^{-}/3NG^{-}$ and iPSC. Attractor B is not often reached from MEF or any of the other

attractors because it requires to turn OFF*Stat*3, which is directly activated by *LIF*. However, if attractor B is reached, it seems to be impossible for the cells to come back to the iPSC path.

When applied to a GRN governing reprogramming from MEF to iPSC, CELLoGeNe provides an overview of the energy landscape controlling cell fate decisions. With this broad overview, we identified known cell states as well as potential cell reprogramming bottlenecks. Marble simulations where we consider all possible states to be initial states gave a detailed overview of the possible roadblocks for reprogramming to iPS cells. This *in silico* analysis represents a powerful tool because a huge number of reprogramming starting points can be considered, which is virtually impossible experimentally. When we conducted simulations corresponding to performed reprogramming experiments, we obtained results which recapitulated experimental observations. Moreover, deeper analysis shed light on mechanisms that could lead to cells being trapped at the attractors corresponding to the observed and newly predicted reprogramming roadblocks.

DISCUSSION

In this study, we developed a novel framework, CELLoGeNe, which calculates energy landscapes for GRNs governing important processes like cell commitment, pluripotency maintenance and reprogramming. CELLoGeNe contains tools to analyse the resulting energy landscapes revealing cell states and reprogramming roadblocks through attractor identification and basins analysis. Getting access to the full energy landscape offers an overview of all stable states of the biological system considered. This is not achievable with the standard methods of solving rate equations or updating a Boolean network. Considering a three-state logic, undesired symmetries are avoided compared to standard Boolean logic. CELLoGeNe is equipped with a tool to visualize multidimensional energy landscapes, giving direct insight into how the energy changes with gene expression variation. Cell lineage commitment and reprogramming to pluripotency and its maintenance were investigated through the analysis of the basins of attraction surrounding stable cell states. We used a stochastic tool which provided a measure of the relative strengths of the attractors and their basin's extent, which was linked to the impact of reprogramming bottlenecks.

We applied CELLoGeNe to two systems describing aspects of maintaining and acquiring cell pluripotency. We analyzed a GRN controlling self-renewal of pluripotency and found attractors in the energy landscape corresponding to experimentally observed stem cell states under various media. CELLoGeNe enabled us to assign probabilities of finding the cells in a certain state. These probabilities were calculated by performing perturbation simulations of stable cell states in a specific medium and consequently identifying the most likely final cell state. When we applied CELLoGeNe to the core network in Dunn et al. (2014) we identified more attractors than the ones experimentally and computationally presented in Dunn et al. (2014). The discovery of these attractors offers a plausible explanation of the observed leakage from pluripotency (Chambers et al., 2007; Chickarmane et al., 2012; Marucci, 2017). CELLoGeNe was applied to a circuitry governing reprogramming from MEF to iPSC (Takahashi and Yamanaka, 2006; Takahashi et al., 2007). Cell reprogramming has a low efficiency and several experimental groups tried to improve this using various strategies: (1) Identifying new reprogramming factors (Buganim et al., 2012; Shu et al., 2013); (2) Fine-tuning the levels of overexpressed TFs (Papapetrou et al., 2009; Radzisheuskaya et al., 2013); (3) varying the order of introducing the factors used in reprogramming protocols (Ho et al., 2013; Olariu et al., 2017a); (4) Monitoring reprogramming progress in a step-wise manner identifying bottlenecks (O'Malley et al., 2013; Chantzoura et al., 2015). In this study, we used CELLoGeNe for identifying new reprogramming bottlenecks, which can further be combined with experiments leading to improving reprogramming efficiency. To this end, we constructed a GRN controlling cell reprogramming and uncovered a corresponding energy landscape containing the MEF- and iPSC-states as well as the experimentally identified $2NG^{-}/3NG^{-}$ roadblocks (O'Malley et al., 2013). The landscape contains three extra attractors which could correspond to potential roadblocks. The possibility for the cells to get stuck in these newly uncovered states links to observed inefficient conversion between MEF and iPSC. Analyzing the gene expression defining these extra attractors sheds light on possible mechanisms leading to conversion inefficiency.

CELLoGeNe can be fused with an experimental CRISPR/Cas9-mediated genome-wide knockout screen in reprogramming. The experiments could predict genes that can act as barriers to cell reprogramming. CELLoGeNe can be applied to GRNs containing the core reprogramming circuit augmented with the newly predicted genes. This would reveal the mechanisms through which the experimentally identified barriers act on preventing a successful cell reprogramming. Possible outcomes of adding the barrier genes to

the circuit consist of the emergence of new attractors that can act as a reprogramming roadblock or change of sizes of basins of attraction of existing attractors corresponding to experimentally identified roadblocks.

CELLoGeNe contains a tool that is capable of constructing a continuous energy landscape by interpolation, a feature which has not been used in this study. With a continuous energy landscape, it would be possible to construct an algorithm to map out optimal reprogramming paths which minimize the risk and prevents getting stuck at roadblocks. In this context, dynamical systems with cyclic attractors (Nordick et al., 2022) could be represented as well as cell divisions. Furthermore, CELLoGeNe is a general tool which can be used on any developmental biological system. A strong candidate for CELLoGeNe applications is the T-cell development, as it provides an excellent model system for studying lineage commitment from a multipotent progenitor in general because this biological system has been deeper experimentally and computationally analyzed. In fact, even though CELLoGeNe was specifically developed for analyzing cell development and reprogramming systems, it can be applied to any system governed by a regulatory interaction network with states that can be binarized. It should be noted that multiple operator configurations can produce different landscapes with different attractors that all fit the data. In such cases, it is recommended to define a score function based on details of the modelled system. Moreover, CELLoGeNe users should be aware that eventual newly identified attractors are predictions that need to be validated. To successfully apply CELLoGeNe to a non-biologically system, its network must have binary nodes affecting each other's state either positively, negatively or neutrally.

We have developed a powerful framework for computing energy landscapes for GRNs enabling us to overview all possible stable cell states, which are inaccessible with other computational means. We applied CELLoGeNe to two biological systems achieving a better understanding of maintenance of pluripotency and offering a plausible explanation for spontaneous exit from multipotent stem cell fates. We also confirmed observed roadblocks and identified new potential cell reprogramming barriers, moreover, revealing their action mechanisms.

Limitations of the study

One limitation of the study comes from the binarization of the gene expression data. This is a crucial step because the binary expression patterns of the stable cell states defines the constraints imposed on the energy landscapes. If the binarization were to be done differently, other energy landscapes with other attractors could be obtained. However, the fact that valid energy landscapes were found for the studied GRNs given the used constraints indicates that the binarization was satisfactory.

Another limitation comes from the discrete representation of the landscape which is also dependent on the number of nodes in the gene regulatory network. This is not ideal for simulating stochastic process like cell division while navigating the landscape. This could be circumvented by developing a method for simulating cell movements in the continuous versions of the landscapes.

In addition, the vast combinations of operator configurations give a considerable amount of freedom to the energy landscapes where many different landscapes could satisfy the same constraints. This means that it is problematic to know which landscapes best correspond to the biological system of interest. This can be handled by using prior knowledge about the gene interactions, e.g., the gene activators follow an AND logic. Furthermore, a ranking score can be constructed based on the landscapes qualities and how they agree with what is expected from the biological system in question. It should also be noted that the issue of degrees of freedom is not unique for the energy landscape framework; for a dynamical system model the freedom would instead lie in choosing the parameters of the rate equations.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials and availability
  - Data and code availability
- METHOD DETAILS

- Binary representation of genes
- Logical representation
- Combining input signals with operators
- The discrete energy
- The continuous energy
- Three-state logic
- Configurations of operators
- Testing configurations of operators
- Exhaustive search
- Stochastic search
- Marble simulations
- Visualisation of high-dimensional energy landscapes

- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2022.104743.

ACKNOWLEDGMENTS

The authors would like to thank Carsten Peterson and Bo Söderberg for discussions at various stages of the project. VO gratefully acknowledges the support of the USNational Institutes of Health (USPHS-grantR01HL119102) and Crafoordska Stiftelsen.

AUTHOR CONTRIBUTION

E.A., M.S., and V.O. designed the study. E.A. and M.S. built the CELLoGeNe software platform. E.A conducted the biological applications and computational analysis. K.K. provided, expertly curated and binarized the experimental data. K.K. explicated and conceptualized the use of experimental data. E.A., M.S., and V.O. wrote the manuscript. All authors provided inputs and comments on the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: April 4, 2022
Revised: June 1, 2022
Accepted: July 5, 2022
Published: August 19, 2022

REFERENCES

Aasen, T., Raya, A., Barrero, M.J., Garreta, E., Consiglio, A., Gonzalez, F., Vassena, R., Bilić, J., Pekarik, V., Tiscornia, G., et al. (2008). Efficient and rapid generation of induced pluripotent stem cells from human keratinocytes. Nat. Biotechnol. *26*, 1276–1284. https://doi.org/10.1038/nbt.1503.

Adachi, K., Kopp, W., Wu, G., Heising, S., Greber, B., Stehling, M., Araúzo-Bravo, M.J., Boerno, S.T., Timmermann, B., Vingron, M., et al. (2018). Esrrb unlocks silenced enhancers for reprogramming to naive pluripotency. Cell stem cell *23*, 266–275. https://doi.org/10.1016/j.stem.2018.05.020.

Alberts, B., Johnson, A., Lewis, J., Morgan, D., Raff, M., Roberts, K., Walter, P., Wilson, J., and Hunt, T. (2008). Molecular Biology of the Cell –, 5th ed. (Garland Science, Taylor & Francis Group).

Aydin, B., and Mazzoni, E.O. (2019). Cell reprogramming: the many roads to success. Annu. Rev. Cell Dev. Biol. *35*, 433–452. https://doi.org/10.1146/annurev-cellbio-100818-125127.

Bhattacharya, S., Zhang, Q., and Andersen, M.E. (2011). A deterministic map of Waddington's epigenetic landscape for cell fate specification. BMC Syst. Biol. *5*, 1–12. https://doi.org/10.1186/1752-0509-5-85.

Bolouri, H., and Davidson, E.H. (2002). Modeling transcriptional regulatory networks. Bioessays *24*, 1118–1129. https://doi.org/10.1002/bies.10189.

Bourillot, P.-Y., and Savatier, P. (2010). Krüppel-like transcription factors and control of pluripotenc. BMC biology *8*, 1–3. https://doi.org/10.1186/1741-7007-8-125.

Boyer, L.A., Lee, T.I., Cole, M.F., Johnstone, S.E., Levine, S.S., Zucker, J.P., Guenther, M.G., Kumar, R.M., Murray, H.L., Jenner, R.G., et al. (2005). Core transcriptional regulatory circuitry in human embryonic stem cells. cell *122*, 947–956. https://doi.org/10.1016/j.cell.2005.08.020.

Buganim, Y., Faddah, D.A., Cheng, A.W., Itskovich, E., Markoulaki, S., Ganz, K., Klemm, S.L., van Oudenaarden, A., and Jaenisch, R. (2012). Single-cell expression analyses during cellular reprogramming reveal an early stochastic and a late hierarchic phase. Cell *150*, 1209–1222. https://doi.org/10.1016/j.cell.2012.08.023.

Chambers, I., Silva, J., Colby, D., Nichols, J., Nijmeijer, B., Robertson, M., Vrana, J., Jones, K., Grotewold, L., and Smith, A. (2007). Nanog safeguards pluripotency and mediates germline development. Nature *450*, 1230–1234. https://doi.org/10.1038/nature06403.

Chantzoura, E., Skylaki, S., Menendez, S., Kim, S.-I., Johnsson, A., Linnarsson, S., Woltjen, K., Chambers, I., and Kaji, K. (2015). Reprogramming roadblocks are system dependent. Stem Cell Rep. *5*, 350–364. https://doi.org/10.1016/j.stemcr.2015.07.007.

Chen, K.C., Csikasz-Nagy, A., Gyorffy, B., Val, J., Novak, B., and Tyson, J.J. (2000). Kinetic analysis of a molecular model of the budding yeast cell cycle. Mol. Biol. Cell *11*, 369–391. https://doi.org/10.1091/mbc.11.1.369.

Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V.B., Wong, E., Orlov, Y.L., Zhang, W., Jiang, J., et al. (2008). Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell *133*, 1106–1117. https://doi.org/10.1016/j.cell.2008.04.043.

Chickarmane, V., Olariu, V., and Peterson, C. (2012). Probing the role of stochasticity in a model of the embryonic stem cell–heterogeneous gene expression and reprogramming efficiency. BMC Syst. Biol. *6*, 1–12. https://doi.org/10.1186/1752-0509-6-98.

Chickarmane, V., Troein, C., Nuber, U.A., Sauro, H.M., and Peterson, C. (2006). Transcriptional dynamics of the embryonic stem cell switch. PLoS computational biology *2*, e123. https://doi.org/10.1371/journal.pcbi.0020123.

Corson, F., and Siggia, E.D. (2017). Gene-free methodology for cell fate dynamics during development. Elife *6*, e30743. https://doi.org/10.7554/eLife.30743.

Dahéron, L., Opitz, S.L., Zaehres, H., Lensch, W.M., Andrews, P.W., Itskovitz-Eldor, J., and Daley, G.Q. (2004). LIF/STAT3 signaling fails tomaintain self-renewal of human embryonic stem cells. Stem cells *22*, 770–778. https://doi.org/10.1634/stemcells.22-5-770.

Davidson, E.H. (2010). Emerging properties of animal gene regulatory networks. Nature *468*, 911–920. https://doi.org/10.1038/nature09645.

Davis, R.L., Weintraub, H., and Lassar, A.B. (1987). Expression of a single transfected cDNA converts fibroblasts to myoblasts. Cell *51*, 987–1000. https://doi.org/10.1016/0092-8674(87)90585-X.

Dunn, S.-J., Li, M.A., Carbognin, E., Smith, A., and Martello, G. (2019). A common molecular logic determines embryonic stem cell self-renewal and reprogramming. EMBO J. *38*, e100003. https://doi.org/10.15252/embj.2018100003.

Dunn, S.-J., Martello, G., Yordanov, B., Emmott, S., and Smith, A.G. (2014). Defining an essential transcription factor program for naive pluripotency. Science *344*, 1156–1160. https://doi.org/10.1126/science.1248882.

Esmailpour, T., and Huang, T. (2012). TBX3 Promotes Human Embryonic Stem Cell Proliferation and Neuroepithelial Differentiation in a Differentiation Stage-dependent Manner. Stem cells *30*, 2152–2163. https://doi.org/10.1002/stem.1187.

Fauré, A., Naldi, A., Chaouiya, C., and Thieffry, D. (2006). Dynamical analysis of a generic Boolean model for the control of the mammalian cell cycle. Bioinformatics *22*, e124–e131. https://doi.org/10.1093/bioinformatics/btl210.

Festuccia, N., Osorno, R., Halbritter, F., Karwacki-Neisius, V., Navarro, P., Colby, D., Wong, F., Yates, A., Tomlinson, S.R., and Chambers, I. (2012). Esrrb is a direct Nanog target gene that can substitute for Nanog function in pluripotent cells. Cell stem cell *11*, 477–490. https://doi.org/10.1016/j.stem.2012.08.002.

Gardner, T.S., Cantor, C.R., and Collins, J.J. (2000). Construction of a genetic toggle switch in Escherichia coli. Nature *403*, 339–342. https://doi.org/10.1038/35002131.

Graf, U., Casanova, E.A., and Cinelli, P. (2011). The role of the leukemia inhibitory factor (LIF)—pathway in derivation and maintenance of murinepluripotent stem cells. Genes *2*, 280–297. https://doi.org/10.3390/genes2010280.

Hamilton, W.B., and Brickman, J.M. (2014). Erk signaling suppresses embryonic stem cell self-renewal to specify endoderm. Cell reports *9*, 2056–2070. https://doi.org/10.1016/j.celrep.2014.11.032.

Harris, C.R., Millman, K.J., Van Der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., et al. (2020). Array programming with NumPy. Nature *585*, 357–362. https://doi.org/10.1038/s41586-020-2649-2.

Hecker, M., Lambeck, S., Toepfer, S., Van Someren, E., and Guthke, R. (2009). Gene regulatory network inference: data integration in dynamic models–a review. Biosystems *96*, 86–103. https://doi.org/10.1016/j.biosystems.2008.12.004.

Heurtier, V., Owens, N., Gonzalez, I., Mueller, F., Proux, C., Mornico, D., Clerc, P., Dubois, A., and Navarro, P. (2019). The molecular logic of Nanog-induced self-renewal in mouse embryonic stem cells. Nature communications *10*, 1–15. https://doi.org/10.1038/s41467-019-09041-z.

Ho, R., Papp, B., Hoffman, J.A., Merrill, B.J., and Plath, K. (2013). Stage-specific regulation of reprogramming to induced pluripotent stem cells by Wnt signaling and T cell factor proteins. Cell Rep. *3*, 2113–2126. https://doi.org/10.1016/j.celrep.2013.05.015.

Huang, P., He, Z., Ji, S., Sun, H., Xiang, D., Liu, C., Hu, Y., Wang, X., and Hui, L. (2011). Induction of functional hepatocyte-like cells from mouse fibroblasts by defined factors. nature *475*, 386–389. https://doi.org/10.1038/nature10116.

Hunter, J.D. (2007). Matplotlib: a 2D graphics environment. Comput. Sci. Eng. *9*, 90–95. https://doi.org/10.1109/MCSE.2007.55.

Ieda, M., Fu, J.-D., Delgado-Olguin, P., Vedantham, V., Hayashi, Y., Bruneau, B.G., and Srivastava, D. (2010). Direct reprogramming of fibroblasts into functional cardiomyocytes by defined factors. Cell *142*, 375–386. https://doi.org/10.1016/j.cell.2010.07.002.

Ivanova, N., Dobrin, R., Lu, R., Kotenko, I., Levorse, J., DeCoste, C., Schafer, X., Lun, Y., and Lemischka, I.R. (2006). issecting self-renewal instem cells with RNA interference. Nature *442*, 533–538. https://doi.org/10.1038/nature04915.

Jacob, F., and Monod, J. (1961). On the regulation of gene activity. In Cold Spring Harbor Symp. Quant. Biol. (Cold Spring Harbor Laboratory Press), pp. 193–211. https://doi.org/10.1101/SQB.1961.026.01.024.

Kauffman, S. (1969). Metabolic stability and epigenesis in randomly constructed genetic nets. J. Theor. Biol. *22*, 437–467. https://doi.org/10.1016/0022-5193(69)90015-0.

Kim, J.B., Sebastiano, V., Wu, G., Araúzo-Bravo, M.J., Sasse, P., Gentile, L., Ko, K., Ruau, D., Ehrich, M., van den Boom, D., et al. (2009). Oct4-induced pluripotency in adult neural stem cells. cell *136*, 411–419. https://doi.org/10.1016/j.cell.2009.01.023.

Lim, C.Y., Tam, W.-L., Zhang, J., Ang, H.S., Jia, H., Lipovich, L., Ng, H.-H., Wei, C.-L., Sung, W.K., Robson, P., et al. (2008). Sall4 regulatesdistinct transcription circuitries in different blastocyst-derived stem cell lineage. Cell stem cell *3*, 543–554. https://doi.org/10.1016/j.stem.2008.08.004.

Loh, Y.-H., Hartung, O., Li, H., Guo, C., Sahalie, J.M., Manos, P.D., Urbach, A., Heffner, G.C., Grskovic, M., Vigneault, F., et al. (2010). Reprogramming of T cells from human peripheral blood. Cell Stem Cell *7*, 15–19. https://doi.org/10.1016/j.stem.2010.06.004.

Loh, Y.-H., Wu, Q., Chew, J.-L., Vega, V.B., Zhang, W., Chen, X., Bourque, G., George, J., Leong, B., Liu, J., et al. (2006). The Oct4 and Nanogtranscription network regulates pluripotency in mouse embryonic stem cells. Nature genetics *38*, 431–440. https://doi.org/10.1038/ng1760.

Martello, G., Bertone, P., and Smith, A. (2013). Identification of the missing pluripotency mediator downstream of leukaemia inhibitory factor. The EMBO journal *32*, 2561–2574. https://doi.org/10.1038/emboj.2013.177.

Martello, G., Sugimoto, T., Diamanti, E., Joshi, A., Hannah, R., Ohtsuka, S., Göttgens, B., Niwa, H., and Smith, A. (2012). Esrrb is a pivotal targetof the Gsk3/Tcf3 axis regulating embryonic stem cell self-renewal. Cell stem cell *11*, 491–504. https://doi.org/10.1016/j.stem.2012.06.008.

Marucci, L. (2017). Nanog dynamics in mouse embryonic stem cells: results from systems biology approaches. Stem Cells Int. *2017*, 7160419. https://doi.org/10.1155/2017/7160419.

Masui, S., Nakatake, Y., Toyooka, Y., Shimosato, D., Yagi, R., Takahashi, K., Okochi, H., Okuda, A., Matoba, R., Sharov, A.A., et al. (2007). Pluripotency governed by Sox2 via regulation of Oct3/4 expression in mouse embryonic stem cells. Nature cell biology *9*, 625–635. https://doi.org/10.1038/ncb1589.

Mendoza, L., Thieffry, D., and Alvarez-Buylla, E.R. (1999). Genetic control of flower morphogenesis in Arabidopsis thaliana: a logical analysis. Bioinformatics *15*, 593–606. https://doi.org/10.1093/bioinformatics/15.7.593.

Mojtahedi, M., Skupin, A., Zhou, J., Castaño, I.G., Leong-Quong, R.Y.Y., Chang, H., Trachana, K., Giuliani, A., and Huang, S. (2016). Cell fate decision as high-dimensional critical state transition. PLoS Biol. *14*, e2000640. https://doi.org/10.1371/journal.pbio.2000640.

Niwa, H., Ogawa, K., Shimosato, D., and Adachi, K. (2009). A parallel circuit of lif signalling pathways maintains pluripotency of mouse es cells. Nature *460*, 118–122. https://doi.org/10.1038/nature08113.

Nordick, B., Yu, P.Y., Liao, G., and Hong, T. (2022). Nonmodular oscillator and switch based on RNA decay drive regeneration of multimodal gene expression. Nucleic Acids Res. *50*, 3693–3708. https://doi.org/10.1093/nar/gkac217.

O'Malley, J., Skylaki, S., Iwabuchi, K.A., Chantzoura, E., Ruetz, T., Johnsson, A.,

Tomlinson, S.R., Linnarsson, S., and Kaji, K. (2013). High-resolution analysis with novel cell-surface markers identifies routes to iPS cells. Nature *499*, 88–91. https://doi.org/10.1038/nature12243.

Olariu, V., and Peterson, C. (2019). Kinetic models of hematopoietic differentiation. Wiley Interdiscip. Rev. Syst. Biol. Med. *11*, e1424. https://doi.org/10.1002/wsbm.1424.

Olariu, V., Lövkvist, C., and Sneppen, K. (2016). Nanog, Oct4 and Tet1 interplay in establishing pluripotency. Sci. Rep. *6*, 1–11. 25438. https://doi.org/10.1038/srep25438.

Olariu, V., Manesso, E., and Peterson, C. (2017a). A deterministic method for estimating free energy genetic network landscapes with applications to cell commitment and reprogramming paths. R. Soc. Open Sci. *4*, 160765. https://doi.org/10.1098/rsos.160765.

Olariu, V., Nilsson, J., Jönsson, H., and Peterson, C. (2017b). Different reprogramming propensities in plants and mammals: are small variations in the core network wirings responsible? PLoS One *12*, e0175251. https://doi.org/10.1371/journal.pone.0175251.

Papapetrou, E.P., Tomishima, M.J., Chambers, S.M., Mica, Y., Reed, E., Menon, J., Tabar, V., Mo, Q., Studer, L., and Sadelain, M. (2009). Stoichiometric and temporal requirements of Oct4, Sox2, Klf4, and c-Myc expression for efficient human iPSC induction and differentiation. Proc. Natl. Acad. Sci. USA *106*, 12759–12764. Stoichiometric and temporal requirements of Oct4.

Peter, I.S., Faure, E., and Davidson, E.H. (2012). Predictive computation of genomic logic processing functions in embryonic development. Proc. Natl. Acad. Sci. USA *109*, 16434–16442. https://doi.org/10.1073/pnas.1207852109.

Qiu, D., Ye, S., Ruiz, B., Zhou, X., Liu, D., Zhang, Q., and Ying, Q.-L. (2015). Klf2 and Tfcp2l1, two Wnt/β-catenin targets, act synergistically toinduce and maintain naive pluripotency. Stem cell reports *5*, 314–322. https://doi.org/10.1016/j.stemcr.2015.07.014.

Radzisheuskaya, A., Chia, G.L.B., Dos Santos, R.L., Theunissen, T.W., Castro, L.F.C., Nichols, J., and Silva, J.C.R. (2013). A defined Oct4 level governs cell state transitions of pluripotency entry and differentiation into all embryonic lineages. Nat. Cell Biol. *15*, 579–590. https://doi.org/10.1038/ncb2742.

Reinitz, J., Mjolsness, E., and Sharp, D.H. (1995). Model for cooperative control of positional information in Drosophila by bicoid and maternal hunchback. J. Exp. Zool. *271*, 47–56. https://doi.org/10.1002/jez.1402710106.

Rodda, D.J., Chew, J.-L., Lim, L.-H., Loh, Y.-H., Wang, B., Ng, H.-H., and Robson, P. (2005). Transcriptional regulation of nanog by OCT4 and SOX2. Journal of Biological Chemistry *280*, 24731–24737. https://doi.org/10.1074/jbc.M502573200.

Roest Crollius, H., Jaillon, O., Bernot, A., Dasilva, C., Bouneau, L., Fischer, C., Fizames, C., Wincker, P., Brottier, P., Quétier, F., et al. (2000). Estimate of human gene number provided by genome-wide analysis using Tetraodon nigroviridis DNA sequence. Nat. Genet. *25*, 235–238. https://doi.org/10.1038/76118.

Sáez, M., Blassberg, R., Camacho-Aguilar, E., Siggia, E.D., Rand, D.A., and Briscoe, J. (2021). Statistically derived geometrical landscapes capture principles of decision-making dynamics during cell fate transitions. Cell Syst. *13*, 12–28.e3. https://doi.org/10.1016/j.cels.2021.08.013.

Shu, J., Wu, C., Wu, Y., Li, Z., Shao, S., Zhao, W., Tang, X., Yang, H., Shen, L., Zuo, X., et al. (2013). Induction of pluripotency in mouse somatic cells with lineage specifiers. Cell *153*, 963–975. https://doi.org/10.1016/j.cell.2013.05.001.

Staerk, J., Dawlaty, M.M., Gao, Q., Maetzel, D., Hanna, J., Sommer, C.A., Mostoslavsky, G., and Jaenisch, R. (2010). Reprogramming of peripheral blood cells to induced pluripotent stem cells. Cell Stem Cell *7*, 20–24. https://doi.org/10.1016/j.stem.2010.06.002.

Stirparo, G.G., Kurowski, A., Yanagida, A., Bates, L.E., Strawbridge, S.E., Hladkou, S., Stuart, H.T., Boroviak, T.E., Silva, J.C., and Nichols, J. (2021). OCT4 induces embryonic pluripotency via STAT3 signaling and metabolic mechanisms. Proceedings of the National Academy of Sciences *118*. https://doi.org/10.1073/pnas.2008890118.

Sugita, M. (1963). Functional analysis of chemical systems in vivo using a logical circuit equivalent. II. The idea of a molecular automaton. J. Theor. Biol. *4*, 179–192. https://doi.org/10.1016/0022-5193(63)90027-4.

Szabo, E., Rampalli, S., Risueño, R.M., Schnerch, A., Mitchell, R., Fiebig-Comyn, A., Levadoux-Martin, M., and Bhatia, M. (2010). Direct conversion of human fibroblasts to multilineage blood progenitors. Nature *468*, 521–526. https://doi.org/10.1038/nature09591.

Takahashi, K., and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. cell *126*, 663–676. https://doi.org/10.1016/j.cell.2006.07.024.

Tai, C.-I., and Ying, Q.-L. (2013). Gbx2, a LIF/Stat3 target, promotes reprogramming to and retention of the pluripotent ground stat. Journal of cell science *126*, 1039–1098. https://doi.org/10.1242/jcs.118273.

Takahashi, K., Tanabe, K., Ohnuki, M., Narita, M., Ichisaka, T., Tomoda, K., and Yamanaka, S. (2007). Induction of pluripotent stem cells from adult human fibroblasts by defined factors. cell *131*, 861–872. https://doi.org/10.1016/j.cell.2007.11.019.

Tantau, T. (2020). The TikZ and pgf packages. Manual for version 3.1. 5b. inf. téc 8.

Tam, W.-L., Lim, C.Y., Han, J., Zhang, J., Ang, Y.-S., Ng, H.-H., Yang, H., and Lim, B. (2008). T-cell factor 3 regulates embryonic stem cellpluripotency and self-renewal by the transcriptional control of multiple lineage pathways. Stem cells *26*, 2019–2031. https://doi.org/10.1634/stemcells.2007-1115.

Tanimura, N., Saito, M., Ebisuya, M., Nishida, E., and Ishikawa, F. (2013). Stemness-related factor Sall4 interacts with transcription factorsOct-3/4 and Sox2 and occupies Oct-Sox elements in mouse embryonic stem cells. Journal of Biological Chemistry *288*, 5027–5038. https://doi.org/10.1074/jbc.M112.411173.

Tatetsu, H., Kong, N.R., Chong, G., Amabile, G., Tenen, D.G., and Chai, L. (2016). SALL4, the missing link between stem cells, development and cancer. Gene *584*, 111–119. https://doi.org/10.1016/j.gene.2016.02.019.

Thomas, R. (1973). Boolean formalization of genetic control circuits. J. Theor. Biol. *42*, 563–585. https://doi.org/10.1016/0022-5193(73)90247-6.

Vierbuchen, T., Ostermeier, A., Pang, Z.P., Kokubu, Y., Südhof, T.C., and Wernig, M. (2010). Direct conversion of fibroblasts to functional neurons by defined factors. Nature *463*, 1035–1041. https://doi.org/10.1038/nature08797.

Waddington, C. (1957). The Strategy of the Genes (George Allen & Unwin).

Wang, M., Tang, L., Liu, D., Ying, Q.-L., and Ye, S. (2017). The transcription factor Gbx2 induces expression of Kruppel-like factor 4 to maintainand induce naïve pluripotency of embryonic stem cells. Journal of Biological Chemistry *292* (41), 17121–17128. https://doi.org/10.1074/jbc.M117.803254.

Wang, X., Wang, X., Zhang, S., Sun, H., Li, S., Ding, H., You, Y., Zhang, X., and Ye, S.-D. (2019). The transcription factor TFCP2L1 inducesexpression of distinct target genes and promotes self-renewal of mouse and human embryonic stem cells. Journal of Biological Chemistry *294*, 6007–6016. https://doi.org/10.1074/jbc.RA118.006341.

Wang, H., Yang, Y., Liu, J., and Qian, L. (2021). Direct cell reprogramming: approaches, mechanisms and progress. Nat. Rev. Mol. Cell Biol. *22*, 410–424. https://doi.org/10.1038/s41580-021-00335-z.

Wang, J., Zhang, K., Xu, L., and Wang, E. (2011). Quantifying the Waddington landscape and biological paths for development and differentiation. Proc. Natl. Acad. Sci. USA *108*, 8257–8262. https://doi.org/10.1073/pnas.1017017108.

Xie, H., Ye, M., Feng, R., and Graf, T. (2004). Stepwise reprogramming of B cells into macrophages. Cell *117*, 663–676. https://doi.org/10.1016/S0092-8674(04)00419-2.

Xu, H., Ang, Y.-S., Sevilla, A., Lemischka, I.R., and Ma'ayan, A. (2014). Construction and validation of a regulatory network for pluripotency and self-renewal of mouse embryonic stem cells. PLoS Comput. Biol. *10*, e1003777. https://doi.org/10.1371/journal.pcbi.1003777.

Xu, H., Baroukh, C., Dannenfelser, R., Chen, E.Y., Tan, C.M., Kou, Y., Kim, Y.E., Lemischka, I.R., and Ma'ayan, A. (2013). ESCAPE: databasefor integrating high-content published data collected from human and mouse embryonic stem cells. Database. https://doi.org/10.1093/database/bat045.

Yang, J., Chai, L., Fowles, T.C., Alipio, Z., Xu, D., Fink, L.M., Ward, D.C., and Ma, Y. (2008). Genome-wide analysis reveals Sall4 to bea major regulator of pluripotency in murine-embryonic stem cells. Proceedings of the National Academy

of Sciences *105*, 19756–19761. https://doi.org/10.1073/pnas.0809321105.

Ye, S., Zhang, T., Tong, C., Zhou, X., He, K., Ban, Q., Liu, D., and Ying, Q.-L. (2017). Depletion of Tcf3 and Lef1 maintains mouse embryonicstem cell self-renewal. Biology open *6*, 511–517. https://doi.org/10.1242/bio.022426.

Yeo, J.-C., Jiang, J., Tan, Z.-Y., Yim, G.-R., Ng, J.-H., Göke, J., Kraus, P., Liang, H., Gonzales, K.A.U., Chong, H.-C., et al. (2014). Klf2 is anessential factor that sustains ground state pluripotenc. Cell stem cel *14*, 864–872. https://doi.org/10.1016/j.stem.2014.04.015.

Yi, F., Pereira, L., and Merrill, B.J. (2008). Tcf3 functions as a steady-state limiter of transcriptional programs of mouse embryonic stem cell self-renewal. Stem cells *26*, 1951–1960. https://doi.org/10.1634/stemcells.2008-0229.

Ying, Q.-L., Wray, J., Nichols, J., Batlle-Morera, L., Doble, B., Woodgett, J., Cohen, P., and Smith, A. (2008). The ground state of embryonic stem cell self-renewal. nature *453*, 519–523. https://doi.org/10.1038/nature06968.

Zhang, X., Peterson, K.A., Liu, X.S., McMahon, A.P., and Ohba, S. (2013). Gene regulatory networks mediating canonical Wnt signal-directed control of pluripotency and differentiation in embryo stem cells. Stem cells *31*, 2667–2679. https://doi.org/10.1002/stem.1371.

Zhang, J., Tam, W.-L., Tong, G.Q., Wu, Q., Chan, H.-Y., Soh, B.-S., Lou, Y., Yang, J., Ma, Y., Chai, L., et al. (2006). Sall4 modulates embryonicstem cell pluripotency and early embryonic development by the transcriptional regulation of Pou5f1. Nature cell biology *8*, 1114–1123. https://doi.org/10.1038/ncb1481.

Zhou, J.X., Aliyu, M.D.S., Aurell, E., and Huang, S. (2012). Quasi-potential landscape in complex multi-stable systems. J. R. Soc. Interface *9*, 3539–3553. https://doi.org/10.1098/rsif.2012.0434.

Zhou, Q., Brown, J., Kanarek, A., Rajagopal, J., and Melton, D.A. (2008). In vivo reprogramming of adult pancreatic exocrine cells to β-cells. nature *455*, 627–632. https://doi.org/10.1038/nature07314.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
Binarised data of maintenance and self-renewal in mouse embryonic stem cells	(Dunn et al., 2014)	N/A
Raw data of reprogramming from mouse embryonic fibroblasts to induced pluripotent stem cells	(O'Malley et al., 2013)	N/A
Software and Algorithms		
Python 3.7	Python Software Foundation	https://www.python.org/
NumPy 3.4.1	(Harris et al., 2020)	https://numpy.org/
Matplotlib 3.4.3	(Hunter, 2007)	https://matplotlib.org/
TikZ 3.1.5b	(Tantau, 2020)	https://www.ctan.org/pkg/pgf
CELLoGeNe	This paper	https://github.com/Emil-cbbp/CELLoGeNe.git

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Victor Olariu (victor.olariu@thep.lu.se).

Materials and availability

This study did not generate new unique reagents.

Data and code availability

- All original code making up the CELLoGeNe software is publicly available as of the data of publication at https://github.com/Emil-cbbp/CELLoGeNe.git.
- This paper analyses existing, publicly available data and are listed in the key resources table.
- Any additional information required to reanalyse the data reported in this paper is available from the lead contactupon request.

METHOD DETAILS

Binary representation of genes

The expression level of a gene can be normalised and represented with a number in the continuous interval $[0, 1]$. As an approximation, the gene expression can also be binarised into either being expressed (ON) or not expressed (OFF), which we will call a *logical gene*. Logical genes, thus, exist in the discrete space $\mathbb{B} = \{0, 1\}$. The N genes present in a GRN, i.e. the genes of interest, constitute the logical cell and exist in the space $\mathbb{B}^n = \{0, 1\}^n$, which we call the *expression* space. Since every gene can be either ON or OFF, the logical cell has 2^n states. Each state of a logical cell can be described by a vector $\boldsymbol{s} = (g_0, g_1, \ldots, g_{n-1})$ of length N where each component represent a gene's expression $g_i \in \mathbb{B}$. Another useful representation is to convert the vector $\boldsymbol{s}$ into a binary number s. We construct s by letting g_i represent the i^{th} bit of the binary number. Formally, this is equivalent to $s = \sum_{i=0}^{n} 2^i g_i$. For instance, consider the state A= OFF, B=ON, C=ON, D=OFF and E= ON of the toy-network in Figure 2A. The vector representation of this state is $\boldsymbol{s} = (0, 1, 1, 0, 1)$which is equivalent to $s = 10110_2 = 22_{10}$ in base 2 and 10 respectively. Thus, each state $\boldsymbol{s}$ of the logical cell can be indexed uniquely by the integer s between 0 and $2^n - 1$.

Logical representation

With traditional Boolean logic, the activation A → B is equivalent to $B = A$, while the repression A ⊣ B is equivalent to $B =$ NOT A. However, as motivated in STAR Methods-Three-state logic, this imposes an unwanted symmetry. Thus, we introduce a new three-state logic to describe network motifs.

The *expression* of the genes are kept in the $\mathbb{B}$, however, we introduce a new space, the *effect* space $\mathbb{E} = \{-1, 0, 1\}$: negative, neutral and positive effect, where activation and repression are mapped differently. The effect an input gene's expression has on a target gene is described with the mapping $\mathbb{B} \mapsto \mathbb{E}$ depending on the context: 0 always maps to 0 ($0 \mapsto 0$), while $1 \mapsto 1$ when the input gene is an activator and $1 \mapsto -1$ when it is an inhibitor. The interpretations of the elements in the effect space are quite straightforward. If the effect is neutral, the expression of the target gene is not affected. If the effect is positive, the expression will become 1 if it was 0, or stay 1 if it was already ON. Similarly for negative effect, the expression will become 0. This can be be encapsulated with the forcing function

$$f : \mathbb{E} \mapsto \mathbb{B}, \quad f(e) = \begin{cases} 1 & \text{if} \quad e = 1 \\ 0 & \text{if} \quad e = -1 \end{cases}, \qquad \text{(Equation 1)}$$

where $e \in \mathbb{E}$ is the effect and $f(0)$ is left undefined as 0 imposes no forcing.

Combining input signals with operators

When multiple genes act as input for a target gene, their effect must be combined into a single input signal. In more mathematical terms, operators that map $\mathbb{E} \times \mathbb{E} \mapsto \mathbb{E}$ are required. Combining the gene effects with logical operators is not uniquely the only possible way to combine signals. One other possibility would be to weigh together the input signals into a resulting effect in a similar fashion to artificial neural networks, with weights still needing to be optimised. This is, however, not considered in this study.

A binary operator in ordinary Boolean logic has $2^2 = 4$ possible inputs, and is uniquely defined by the set of outputs it assigns to these; therefore, there are $2^4 = 16$ possible Boolean operators. Similarly, an operator in three-state logic has $3^2 = 9$ possible inputs, giving $3^9 = 19683$ possible operators. The vast majority of this prohibitively large number of operators are quite useless, so selection is needed.

We take $q(i, j)$ to be the result when the operands are i and j, and define three properties that can be demanded of any reasonable operator:

- Idempotence — it is reasonable to assume that an interaction where all inputs are equal yields that same value as the output. Requires $q(i, i) = i$.
- Commutativity — there is no reason why the combination of genes should not be symmetric, and the lack of this property complicates the arithmetic. Requires $q(i, j) = q(j, i)$.
- Associativity — also a basic arithmetic property. Requires $q(i, q(j, k)) = q(q(i, j), k)$.

Placing these constraints on the standard Boolean operators leaves only AND and OR, which are usually the only ones that are included in Boolean models (along with unary NOT). On three-valued operators, the first two constraints leave only three free parameters of the original nine (for instance, $q(0, -1)$, $q(0, +1)$ and $q(+1, -1)$), on which the consequences of associativity can be worked out. For clarity, we can write q as a matrix $\mathbf{Q}$ such that $q(i-1, j-1) = \mathbf{Q}_{ij}$. With all three constraints, the result is a set of only nine operators, three of which are quite trivial:

$$\begin{bmatrix} -1 & -1 & -1 \\ -1 & 0 & -1 \\ -1 & -1 & +1 \end{bmatrix}, \quad \begin{bmatrix} -1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & +1 \end{bmatrix}, \quad \begin{bmatrix} -1 & +1 & +1 \\ +1 & 0 & +1 \\ +1 & +1 & +1 \end{bmatrix}. \qquad \text{(Equation 2)}$$

Except when required otherwise by idempotence, these just return a fixed value. The middle one is actually a rather straightforward generalisation of a two-valued Boolean operator: it acts as AND on $\{0, +1\}$ and $\{0, -1\}$, and maps the additional input combination $\{\pm 1, \mp 1\}$ to 0. The similarity to AND is shared with two more interesting operators:

$$\overline{\wedge} : \begin{bmatrix} -1 & 0 & +1 \\ 0 & 0 & 0 \\ +1 & 0 & +1 \end{bmatrix} \quad \text{and} \quad \underline{\wedge} : \begin{bmatrix} -1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & +1 \end{bmatrix}. \qquad \text{(Equation 3)}$$

We have chosen symbols based on that for AND, $\wedge$, with a bar above or below to symbolise mapping the different-sign inputs to $+1$ or -1, respectively. There exists a similar pair of operators for OR:

$$\overline{\vee} : \begin{bmatrix} -1 & -1 & +1 \\ -1 & 0 & +1 \\ +1 & +1 & +1 \end{bmatrix} \quad \text{and} \quad \underline{\vee} : \begin{bmatrix} -1 & -1 & -1 \\ -1 & 0 & +1 \\ -1 & +1 & +1 \end{bmatrix}. \qquad \text{(Equation 4)}$$

No associative analogue exists that maps different-sign inputs to 0. The two final operators are

$$\uparrow : \begin{bmatrix} -1 & 0 & +1 \\ 0 & 0 & +1 \\ +1 & +1 & +1 \end{bmatrix} \quad \text{and} \quad \downarrow : \begin{bmatrix} -1 & -1 & -1 \\ -1 & 0 & 0 \\ -1 & 0 & +1 \end{bmatrix}. \qquad \text{(Equation 5)}$$

They return the maximum and minimum input value, respectively, and act as hybrids of AND and OR.

Either one of the pairs $\{\underline{\wedge}, \underline{\vee}\}$, $\{\overline{\wedge}, \overline{\vee}\}$ or $\{\downarrow, \uparrow\}$ provides an operator set that is fairly balanced with respect to the frequency of each output value. For greater flexibility, but also greater complexity, the full set of all six operators can be used.

By removing the requirement of idempotence, potentially useful operators such as XOR can be added for Boolean logic. The corresponding modification gives 63 three-valued operators, none of which seems to be a good analogue.

The discrete energy

Our goal is to build a stochastic model which shares the general behaviour of a deterministic model built on the same network. To do so in a straightforward way, we assign an energy L to each gene of a state whenever its expression status equals the result of its logical function, and an energy H whenever it does not. The energy contributions for each gene are added to form the total energy of a state. This is repeated for each state to form the energy of the entire network. If $L < H$, we thereby reward states that "obey the logic" with a favourable energy.

In more mathematical details, for each state $\boldsymbol{s}$, a gene g with expression value $a_g \in \mathbb{B}$ is chosen. The other $n-1$ genes serve as input to g and form the input space $\boldsymbol{i}_g \in \mathbb{B}^{n-1}$. To an input state $\boldsymbol{i}_g$, there is a corresponding resulting effect $F(\boldsymbol{i}_g) \in \mathbb{E}$which is obtained by applying the combination operators, i.e. evaluating a logical function. The energy for state $\boldsymbol{s}, \mathcal{T}(\boldsymbol{s})$, is calculated by adding the energy contributions from each gene $\tau(a_g, i_g)$

$$\mathcal{T}(\boldsymbol{s}) = \sum_g \tau(a_g, i_g), \qquad \text{(Equation 6)}$$

where

$$\tau(a_g, i_g) = \begin{cases} L & \text{if} \quad a_g = f(F(i_g)) \\ H & \text{if} \quad a_g \neq f(F(i_g)) \\ \dfrac{L+H}{2} & \text{if} \quad F(i_g) = 0 \end{cases} \qquad \text{(Equation 7)}$$

and f is defined in Equation (1). Thus, the complete discrete energy landscape for the logical cell is represented by the vector $\mathcal{T}$ where each component $\mathcal{T}_s$ is the energy $\mathcal{T}(\boldsymbol{s})$ of corresponding state $\boldsymbol{s}$. With $L = -1, H = 1$, a state with energy value $\mathcal{T}(\boldsymbol{s}) = -n$ means that all gene expressions a_g agree with their corresponding forcing functions $F(f(\boldsymbol{i}_g))$ and vice versa for energy value $\mathcal{T}(\boldsymbol{s}) = n$.

The continuous energy

The discrete energy is a function $\mathcal{T} : \{0,1\}^n \to \mathbb{R}$, which we can interpolate into a continuous energy $E : [0,1]^n \to \mathbb{R}$, which then allows each gene to have any degree of expression. Let $\boldsymbol{s}_i$ be the expression of the i^{th} gene in the network, and $\boldsymbol{s} \in [0,1]^n$ be the vector of all $\boldsymbol{s}_i$. We will now write down the most general

form of $E(\boldsymbol{s})$ that is first-order in all s_i; that is, for each i, $E(\boldsymbol{s}) = A_i + B_i s_i$ where A_i, B_i are first-order functions of $\boldsymbol{s}_j, j \neq i$. This straightforwardly gives the general expression

$$E(\boldsymbol{s}) = \omega + \sum_i \boldsymbol{s}_i \omega_i + \sum_{i<j} \boldsymbol{s}_i \boldsymbol{s}_j \omega_{ij} + \sum_{i<j<k} \boldsymbol{s}_i \boldsymbol{s}_j \boldsymbol{s}_k \omega_{ijk} + \ldots + \boldsymbol{s}_1 \boldsymbol{s}_2 \cdots \boldsymbol{s}_n \omega_{12\ldots n}, \quad \text{(Equation 8)}$$

where all sums run between 1 and N.

We now claim that the coefficients $\omega_{ij\ldots}$ are uniquely determined by $\mathcal{T}$. We define the vector $\boldsymbol{\omega}_a = \omega_{ij\ldots}$, where $b = 2^i + 2^j + \ldots$ (i.e. the binary representation of b has 1's only at positions $i, j, \ldots$), and the vector $\mathcal{T}_a = \mathcal{T}(\boldsymbol{s})$ where $a = \sum 2^i \boldsymbol{s}_i$ (i.e. $\boldsymbol{s}$ forms the binary representation of a). With these definitions, there exists a $2^n \times 2^n$ matrix $\mathbf{M}$ such that Equation (8) reduces to

$$\mathbf{M}\boldsymbol{\omega} = \mathcal{T}. \quad \text{(Equation 9)}$$

It can be proven (as shown below) that $\mathbf{M}$ is invertible and has the beautiful Sierpiński-triangle-like structure

$$\mathbf{M}_{(n+1)}^{\pm 1} = \begin{bmatrix} \mathbf{M}_{(n)}^{\pm 1} & 0_{(n)} \\ \pm \mathbf{M}_{(n)}^{\pm 1} & \mathbf{M}_{(n)}^{\pm 1} \end{bmatrix}, \quad \mathbf{M}_{(1)}^{\pm 1} = \begin{bmatrix} 1 & 0 \\ \pm 1 & 1 \end{bmatrix}, \quad \text{(Equation 10)}$$

where $\mathbf{M}_{(n)}$ is the $2^n \times 2^n n$-gene matrix, and $0_{(n)}$ is the zero matrix of the same size. This proves that the energy function in Equation (8) is uniquely determined by $T(\boldsymbol{s})$, with

$$\boldsymbol{\omega} = \mathbf{M}^{-1}\mathcal{T} \quad \text{(Equation 11)}$$

providing the transition from discrete to continuous energy.

There is a final benefit to this energy form. When it is paired with the entropy function

$$S(\boldsymbol{s}) = -\sum_{i=1}^{n} [\boldsymbol{s}_i \log \boldsymbol{s}_i + (1 - \boldsymbol{s}_i)\log(1 - \boldsymbol{s}_i)], \quad \text{(Equation 12)}$$

which contains no cross-terms or additional free parameters, we get a free energy function $F(\boldsymbol{s}) = E(\boldsymbol{s}) - TS(\boldsymbol{s})$ (where T is the temperature) that generates sigmoid gain functions very closely approximating the Hill gain functions conventionally used in transcriptional dynamics (Olariu et al., 2017a). Thus, our model bridges the gap between stochastic and Boolean treatments.

We will now prove that the matrix $\mathbf{M}$ is invertible and we will derive its form. From Equation (9) and the definitions of $\boldsymbol{\omega}$ and $\mathcal{T}$, we find that $\mathbf{M}_{ab} = \boldsymbol{s}_i \boldsymbol{s}_j \cdots$ where $a = \sum 2^i \boldsymbol{s}_i$ (i.e. $\mathcal{T}_a = \mathcal{T}(\boldsymbol{s})$) and $\boldsymbol{\omega}_b = \omega_{ij\cdots}$. In other words, $\mathbf{M}_{ab}$ is 1 if the binary representation of b has a 1 wherever the binary representation of a has a 1 (and possibly in more places), and 0 otherwise. We can bring this into the language of set theory by identifying each number as the set of positions in which its binary representation has a 1. This representation has a close relationship with bitwise operations on binary numbers. In a C-like language, $a \cup b$ corresponds to (a | b), $a \cap b$ to (a & b), $\varnothing$ to 0, and $a \subseteq b$ to (a & b) == a. Thus, $\mathbf{M}_{ab}$ is 1 if and only if $b \subseteq a$.

From this, the expression for $\mathbf{M}_{(1)}$ given in Equation (10) should follow straightforwardly. To then prove the expression for $\mathbf{M}_{(n+1)}$, subdivide the 2^{n+1}-bit number a like $a = a_l \cup a_h$, where a_l contains its low n bits (i.e. $a_l = a \cap 2^n - 1$) and a_h its high bit (i.e., $a_h = a \cap 2^n$). Subdivide b similarly. Then $\mathbf{M}_{(n+1)}$ can be split into four $2^n \times 2^n$ quadrants:

$$\mathbf{M}_{(n+1)} \sim \begin{bmatrix} (a_h = 0, b_h = 0) & (a_h = 0, b_h = 2^n) \\ (a_h = 2^n, b_h = 0) & (a_h = 2^n, b_h = 2^n) \end{bmatrix}. \quad \text{(Equation 13)}$$

In the upper right quadrant, $a \not\subseteq b$ regardless of a_l and b_l, since a lacks the high bit b has. Therefore, that quadrant is the $2^n \times 2^n$ zero matrix. In the other quadrants, $a \subseteq b$ is equivalent to $a_l \subseteq b_l$, so each of those quadrants is just a copy of $\mathbf{M}_{(n)}$. By induction, this proves the form of $\mathbf{M}_{(n+1)}$ given in Equation (10).

The form of $\mathbf{M}_{(n+1)}^{-1}$ easily follows by applying the standard inversion formula for block matrices to $\mathbf{M}_{(n+1)}$. It can also be seen by noting that $\mathbf{M}_{(n+1)} = \mathbf{M}_{(1)} \otimes \mathbf{M}_{(n)}$ ($\otimes$ being the Kronecker product) and applying the Kronecker product inversion formula, giving $\mathbf{M}_{(n+1)}^{-1} = \mathbf{M}_{(1)}^{-1} \otimes \mathbf{M}_{(n)}^{-1}$.

Three-state logic

In this section, we will give a motivation for the introduction of the three-state logic. We will start by considering the conventional Boolean approach and then extend it stepwise.

A basic Boolean model, which uses the operators AND, OR to build functions and NOT to represent repression, yields less than satisfactory energy functions when our method is applied. This can be demonstrated using the simple network B ➝A, which assigns A the function $A = B$. With $L = -1$, $H = +1$, and $\boldsymbol{s} = (A \quad B)$, we get

$$\mathcal{T} = (-1 \quad 1 \quad 1 \quad -1) \Rightarrow \boldsymbol{\omega} = (-1 \quad 2 \quad 2 \quad -4) \Rightarrow E(\boldsymbol{s}) = -(2B-1)(2A-1). \tag{Equation 14}$$

For B ⊣ A, which is represented by the function $A = \text{NOT } B$, we get the negative of this energy. The energy functions are entirely symmetric in B and A, even though the networks are directed.

The source of this unwanted symmetry is that B affects A equally much whether it is expressed or not. A perhaps more reasonable model is that a gene is only affected when its input is expressed; a gene with an inactive input is "free", which we represent by giving it a third, "neutral" energy value between L and H (here, 0), regardless of the gene's state. With these modifications, we get for B ➝A

$$\mathcal{T} = (0 \quad 0 \quad 1 \quad -1) \Rightarrow \boldsymbol{\omega} = (0 \quad 0 \quad 1 \quad -2) \Rightarrow E(\boldsymbol{s}) = B(1-2A), \tag{Equation 15}$$

which properly reflects the directedness of the network. It also has an intuitive interpretation: the energy decreases when A is expressed, and the amount by which it does is proportional to B, i.e. how strongly A is activated. However, for B ⊣ A, we get

$$\mathcal{T} = (1 \quad -1 \quad 0 \quad 0) \Rightarrow \boldsymbol{\omega} = (1 \quad -2 \quad -1 \quad 2) \Rightarrow E(\boldsymbol{s}) = (B-1)(2A-1), \tag{Equation 16}$$

which is drastically different. It is directed, but goes against the above reasoning: B only affects A when it is not expressed.

To solve this, we stop using NOT altogether, and instead switch to the three-valued logic of TRUE (1), FALSE (0) and "negative TRUE" (−1), the latter of which represents a repressing input, as described in STAR Methods-Logical representation. A logical function whose result is −1 is obeyed if the targeted gene is not expressed, while 1 and 0 work as before. With the three-valued logic, B ➝A gives the same energy as Equation (15), while B ⊣ A gives the negative of it, which is equally reasonable.

The three-valued logic results in presumably better energy functions, but it also incurs increased complexity, sinceAND and OR need to be replaced with the three-valued operators as described in STAR Methods-Combining input signals with operators.

Configurations of operators

When multiple genes act as input for a target gene, their effects need to be combined into a single input signal with an operator, as described in STAR Methods-Combining input signals with operators. Which operator to use may be given from literature or experimental studies, but most often, it is not known which operator to use. Given that a real GRN most often has many nodes with several inputs, the network has many different configurations of operators where each configuration corresponds to a potentially unique energy landscape. The total number of configurations of a network can be factorised into the product of the number of configurations there are for each target gene.

If there are p possible operators and a target gene has input from k genes, we denote the number of configurations for that gene $N(p,k)$. For only one input gene, $N(p,1) = 1$ trivially, since no inputs need to be combined. For only two input genes, $N(p,2) = p$, since the only thing that can be changed is the operator that combines the two inputs. The rest is a matter of inductive reasoning, yielding the full recursive expression

$$N(p,k) = N(p,k-1)\cdot(3(p-1)+1) = N(p,k-1)\cdot(3p-2), \tag{Equation 17}$$

with the base cases given above. This has the closed form

$$N(p,k) = p\cdot(3p-2)^{k-2}, k>1. \tag{Equation 18}$$

In Table S1 some sample values for $N(p, k)$ are presented, clearly demonstrating why we wish to limit the number of operators. Note that the values in the table are for one single target gene of the network only; to get the total number of the configurations of the network, the correct factors must be multiplied.

Testing configurations of operators

If more than one logical operator is used to combine gene input signals, there exist multiple configurations of operators. There are two main strategies on how to test different configurations: exhaustively test every combination, or randomly test a subset of combinations. For small GRNs or when few operators are used, an exhaustive search is suitable but, as understood from the previous section, sometimes the number of possible configurations is so large that a stochastic search must be used instead.

Exhaustive search

Given a set of genes and a set of operators, there is a large number of possible ways to combine these. All possible configurations are not distinct, thanks to commutativity and associativity. For instance, with {*A*, *B*, *C*} and { $\underline{\wedge}$, $\underline{\vee}$ }, there are eight configurations:

$$A \mathbin{\underline{\wedge}} B \mathbin{\underline{\wedge}} C, \quad \left(A \mathbin{\underline{\wedge}} B\right) \mathbin{\underline{\vee}} C, \quad \left(A \mathbin{\underline{\vee}} B\right) \mathbin{\underline{\wedge}} C, \quad A \mathbin{\underline{\vee}} B \mathbin{\underline{\vee}} C,$$

and all permutations of the genes in the middle two cases.

We now seek an optimal way to enumerate all non-redundant configurations, so that we can determine which one produces valid energy landscapes. Considering the example $(A \mathbin{\underline{\vee}} B) \mathbin{\underline{\wedge}} C$, we note that if we change the first operator to $\underline{\wedge}$, both $(A \mathbin{\underline{\wedge}} B)$ and $(\ldots) \mathbin{\underline{\wedge}} C$ need to be recalculated. But if we change the second instead, we can reuse the result in parentheses. With this and other considerations, we can create specifications for an efficient method:

1. The outermost operators should be changed more often and operators inside parentheses should be changed only when all outer-level versions have been exhausted.
2. The same should apply to moving genes around within the parenthesis structure: more deeply nested genes should be accessed only after all permutations of outer genes have been used.
3. Changes should not be performed when commutativity and associativity make them superfluous.

Such a method is devised below.

A convenient representation of configurations that allows for dealing with associativity and commutativity to avoid redundancy is binary trees: the leaves are genes, and the non-leaves nodes are operators that combine their children. A *n*-gene compound can therefore be represented as a tree with *n* leaves, as illustrated in Figure S4A. If each node "knows" its logical function, this representation enables reuse of functions in accordance with specifications 1 and 2. Each target gene in a GRN requires its own binary tree. For the trivial case where a gene only receives input signal from one other gene, the tree simply consists of one leaf. If a gene does not receive any input signal, e.g. a medium component such as *LIF*, its corresponding tree is empty.

There is a recursive way to move around the parentheses in a tree-based manner that is well in line with the specifications. Since commutativity makes mirror images of a tree equivalent, we can without loss of generality only consider the cases where there are more leaves on the left side of the tree. The extreme case of this is when all operators are placed on the left edge of the tree, with only leaves on the right. This maximally left-heavy tree is the starting configuration of the method.

A demonstration of how it works is shown in Figure S4B on a minimal tree; after three permutations, it is back at the starting configuration. If node II has children of their own, the permutation procedure is recursively applied to it; this is shown in Figure S4C. After a single such permutation, it is possible to make a fresh re-permutation around node I. When that is done, node II is allowed to take another step, and so on. At any level in the tree, a node only permits its child to permute one step when it has finished its own cycle of permutations.

This last rule, which we shall call the parent-first order, ensures two things. Firstly, as soon as the method reaches a node for which both children are leaves, it is guaranteed to be finished, since all levels above it have passed through all possible configurations. Secondly, as shown in Figure S4C, a single permutation of the child replaces node *B* with a child of node *C*. It is then moved around by the parent, which puts it available for the parent's parent (if present), etc. Applied inductively, this property guarantees that every leaf and every possible subtree moves around all possible positions in the tree, and thus the tree goes through all possible configurations regardless of its size.

The process of permuting a tree can be readily extended to include operator changes as well; the full procedure is given in Figure S5. The algorithm becomes rather convoluted by the efforts to maximise function reuse and avoid redundant configurations—there are necessarily two associativity-safeguards labelled "same op. as left child?" in the figure—but stays implementation-friendly.

Operators are changed first because changing an operator is the cheapest possible change: it only requires the node's own function to change (plus those of its parents), while a permutation also requires that of its left child to change. Also, a permutation changes which leaves are present in the child's subtree, so the function needs to change its list of inputs in addition to its output; hence the "Update left child's input" step.

Running this algorithm on the root of a tree until it hits FINISHED goes through all possible non-degenerate combinations of genes and operators. We note that larger trees are somewhat more costly to work with due to the larger number of deeply nested changes, so we should keep the trees sorted by ascending size when applying the algorithm.

Some further complications are straightforward to account for, should they be present for some reason:

- If the operators are non-commutative, insert "Update function with tree mirrored" after "Update function" in Figure S5
- If the operators are non-associative, always answer "No" to "Same op. as left child?"

Stochastic search

Performing a stochastic search of the configuration space is substantially simpler than the exhaustive algorithm outlined above. For a given set of operators, a random binary tree representation (Figure S4A) is generated for each target gene in the GRN and the energy landscape is calculated. This is repeated until the desired number of configurations has been tested.

Marble simulations

In order to measure the relative strengths of the attractors' basins of attraction, we simulate cells performing weighted random walk in the energy landscape, analogous to letting marbles roll through the landscape until a minimum is reached. In each update, a cell can transition to a neighbouring state, i.e. changing one gene's expression, or remain in the same state. The updates are repeated until a stopping criterion is met. By initialising a large instances of cells in every cell state and counting how many times the different attractors are reached as the final state, the probability of reaching each attractor from each cell state is obtained. By comparing the sizes of the regions of the landscape from which the attractors are reachable, and the probability of reaching each attractor from each cell state, the basins of attraction, and their relative strength, is mapped out.

The possible transitions for a cell state are weighted with the Boltzmann factor. A state s has energy E_s and the probability to transition to any of its N neighbours, or not make a transition, depends on the energy difference between the states (Figure S6). We define the transition probability from state s to μ as

$$p_{\mu \leftarrow s} = \frac{e^{-\left(E_\mu - E_s\right)\beta}}{\sum_{k \in \{1,\ldots,n,s\}} e^{-\left(E_k - E_s\right)\beta}}, \quad \mu \in \{1, \ldots, n, s\}, \qquad \text{(Equation 19)}$$

where β is the temperature parameter or noise level. By decreasing β, the noise level is increased and the probability to transition to a state with a higher energy increases. The transition to state μ is chosen as the smallest μ fulfilling

$$\sum_{i=1}^{\mu} p_i > r, \qquad \text{(Equation 20)}$$

where $r \in [0, 1)$ is a uniform randomnumber. The cell state is updated until it has stayed in the same state during three consecutive updates, then it is considered to have reached a stable state.

Visualisation of high-dimensional energy landscapes

The energy landscapes treated here are not particularly visualisation-friendly, since they exist in a space of high dimensionality. With traditional plotting methods, it is hard to depict spaces in more than three dimensions. However, by utilising the fact that an N-gene binary GRN has 2^N states, and each state has N neighbouring states (with neighbouring state we mean states where only one bit changes, i.e. states with Hamming distance equal to 1), we can construct somewhat reasonable representations of energy landscapes with up to seven dimensions; beyond seven dimensions the plot becomes incomprehensibly large and cluttered. A known object which has 2^N states and N connections in Ndimensions is a N-dimensional hypercube. By mapping the corners of the energy function $\mathcal{T} \in \mathbb{B}^N$ to the corners of a hypercube, it just becomes a matter of plotting hypercubes onto two dimensions.

A 2-dimensional hypercube is simply a square. Thus, we let the four corners of a square represent the four states of a two-gene GRN (Figure 1B). We plot the square with rounded edges for future convenience. A three-dimensional hypercube is a regular cube. In essence, a cube is just two squares vertically stacked with edges connecting corresponding corners. Hence, we can plot this in a flat version by arranging one square outside of another and connecting corresponding corners (Figure 1B). By the same principle, to represent a four-gene GRN, we can create a four-dimensional flat hypercube by connecting two flattened cubes (Figure 1B). By continuing this arrangement in a clever way, we can create flat hypercubes up to seven dimensions before they start becoming incomprehensibly cluttered (see Figure 3D for an exemplary energy landscape in seven dimensions). In principle, this plotting technique could, of course, be extended to arbitrarily many dimensions.

Now, we have a structure for the landscape plots in place, where each cell state maps to a corner on a hypercube. The next step is to add information. To clearly distinguish the states, we let each node in the graph be represented by a pie chart, or sector, where each piece of pie represents a gene. A gene is ON if its corresponding sector is filled, and *OFF* if it is empty. For example, the state (A = OFF, B = ON, C = OFF, D = ON, E = OFF) is depicted as shown in Figure 1C. Here, each slice is specifically marked with a label indicating the corresponding gene. A labelled state like this is always accompanying an energy landscape plot as a legend applied to all the states. The energy value of each state is represented by a colour which maps to the energy via a colour bar. The final property of the basic version of the landscape plots is that every edge describes the energy difference between the connected states. The edge is drawn as an arrow, pointing towards the state with lower energy, and its width is proportional to the magnitude of the energy difference. Neighbouring states with the same energy are connected with a dashed line. Thus, the direction and thickness of the arrows show the relative probabilities of transitions between cell states. Transitions between states with the same energy are also possible. An example of a complete energy landscape is displayed in Figure 1D.

With this plotting technique, it is, as stated above, possible to represent up to seven-gene networks before the plots get prohibitively large. However, there are ways to circumvent this obstacle. One possible trick is to only plot hyperplanes of interest. Then, one chooses up to seven genes of interest to be displayed, and the remaining genes are fixed. Another trick is to merge several genes that behave similarly in the system of interest into one single dimension. If this can be done with several groups of genes, this has the possibility to reduce the dimensionality substantially. The downside of this trick is that states that are not neighbours in the full energy landscape will appear as neighbours in the reduced landscape. However, this can still be useful in some circumstances.

More details can be added for further clarity. Attractor states and their edges can be colour coded. For full landscapes, the basins of attraction may be illustrated by a coloured ring around the landscape node. The

ring is partitioned and colour coded according to which attractors' basins the state is part of. This requires that the attractors are colour coded as well. These additional details are utilised in Figure 2D.

QUANTIFICATION AND STATISTICAL ANALYSIS

All quantification and analyses were performed as described in the method details section of the STAR Methods. Data are presented as mean ± standard deviation.