Prioritization of Candidate Genes Through Boolean Networks*

Clémence Réda[1][0000−0003−3238−0258] and Andrée Delahaye-Duriez[1,2,3][0000−0003−4324−7372]

[1] Univ. Paris Cité, Neurodiderot, Inserm, F-75019 Paris
[2] Univ. Paris 13, Sorbonne Paris Nord, UFR de santé, médecine et biologie humaine, F-93000 Bobigny
[3] Unité fonctionnelle de médecine génomique et génétique clinique, Hôpital Jean Verdier, AP-HP, F-93140 Bondy
{clemence.reda,andree.delahaye}@inserm.fr

Abstract. The *in silico* detection of master regulator genes is a popular approach to speed up drug development. These genes might be directly related to the onset of the disease, or may act on one pathway which counteracts the associated symptoms. Then, one could perhaps screen drugs to select chemical compounds targeting these genes. In prior works, the detection of these candidates was performed through the identification of the regulatory interactions between genes of interest for the disease. Indeed, system biology approaches have proven a useful tool to integrate transcriptomic data and predict transcriptional profiles under gene perturbations. However, for rare or tropical neglected diseases, building such a regulatory model can become a tedious and time-consuming task. In this work, we show how to build, in a reproducible and transparent fashion, a gene regulatory network using publicly available data. Then, we describe a method to identify master regulatory genes, which have an impact on the dynamics of the gene regulation in a specific disease-related transcriptional context. We showed that our novel method for the identification of master regulatory genes was consistent with network controllability measures, while targeting genes that were significantly enriched for epilepsy-related terms. Our pipeline allows for systematic and transparent Boolean network synthesis, and identification of master regulators, which might help tackle the issue of rare or tropical neglected diseases.

Keywords: master regulator prioritization · drug-resistant epilepsy · boolean network · influence maximization · machine learning application.

1 Introduction

We propose a novel generic method for the detection of master regulator genes, which can be applied to any disease, and relies on a dynamical interplay between

* Institut National pour la Santé et la Recherche Médicale (Inserm, France), the French Ministry of Higher Education and Research, Université Paris Cité, Université Sorbonne Paris Nord, and the French National Research Agency.

a gene regulatory network and gene expression data. We focus here, as a proof-of-concept, on an application to epilepsy.

Epilepsy actually encompasses various neurological diseases and syndromes, which can originate from brain injury or genetic background, that have in common a propensity to trigger chronic epileptic crises. Epileptic crises are characterized by a transitory abnormal neuron electric discharge, which might lead to unconsciousness, seizures, and/or body stiffness. Epilepsy is one of the most common neurological diseases worldwide, with around 50 million people living with this disease [69]. Moreover, more than 25% of epileptic patients are afflicted with drug-resistant epilepsy [26] –also called refractory epilepsy– that is, symptoms in those patients could not be managed by at least two different antiepileptic therapies. This shows the limits of conventional antiepileptic medication, which are often molecules with antiseizure effects, and emphasizes the need to look for novel therapeutic candidates. Epilepsy-related genes are shown to be usually mainly expressed in a specific brain region, called hippocampus [44], which is also affected by morphological changes linked to neuronal discharges in some epileptic patients [49]. The exact relationship between lesions in the hippocampus and epilepsy-associated symptoms is still unclear, but might be related to the fact that hippocampus is one of the most excitable parts of the brain [36]. Several animal models of epilepsy exist, including a mouse model where injection of pilocarpine induce symptoms similar to temporal lobe epilepsy [59], or another involving sodium channels, which are used to convey electric potentials (Dravet syndrome model, by knocking out gene *Scn1a* [33]).

In prior works, the identification of master regulators in gene networks has been a powerful method to detect novel genes of interest for a given disease. Master regulator genes are DNA sequences which might have a large, global influence on the expression of a group of genes in a specific pathway. For instance, *SESTRIN3* [31] and *CSF1R* [58] were prioritised as candidate antiepileptic drug targets using different systems-biology approaches dedicated to identifying master regulators of epilepsy-associated networks of gene expression. Such genes might be forcibly expressed or knocked out –*i.e.*, no more expressed– by molecules, which might be interesting antiepileptic drug candidates. Other approaches exploit the location of a given gene inside a gene regulatory network –the more central it is, the most regulatory it should be– or the concept of "network controllability" [41]. Yet, most of the cited approaches for the detection of interesting regulatory genes only leverage topological knowledge about the network, without considering the actual dynamics of the regulatory system. A notable exception is the work by [71], which aimed at finding master regulator genes related to rheumatoid arthritis based on expression data. Their approach combines a transcription factor (TF) co-regulatory network and gene expression in fibroblast-like synoviocytes in patients afflicted with rheumatoid arthritis. TF influence in these samples was assessed using the tool CoRegNet [48], which computes a score of influence for a given TF on a set of transcriptional profiles. This score is defined for any TF t, that activates a set $\mathcal{A}_t$ of genes and inhibits a set $\mathcal{I}_t$ of genes, and

for a given matrix of transcriptional profiles M [4] as follows

$$\text{Influence}(t) := \frac{\left(\frac{1}{|\mathcal{A}_t|}\sum_{a\in\mathcal{A}_t} M[a,:]\right) - \left(\frac{1}{|\mathcal{I}_t|}\sum_{i\in\mathcal{I}_t} M[i,:]\right)}{\sqrt{(s_{\mathcal{A}_t})^2/|\mathcal{A}_t| + (s_{\mathcal{I}_t})^2/|\mathcal{I}_t|}} \tag{1}$$

where $s_{\mathcal{A}_t}$ (resp., $s_{\mathcal{I}_t}$) is the standard deviation of expression levels of all genes in $\mathcal{A}_t$ (resp., $\mathcal{I}_t$) across all profiles in M. However, such a computation does not take into account downstream transcriptional cascades [8], that is, regulatory effects which trickle down the network, beyond the genes directly regulated by the TF. However, taking into account these regulatory cascades might allow to control for off-target genes, which are genes subject to non specific and involuntary changes, for which perturbation might lead to serious side effects [30]. In order to model these regulatory cascades, we are interested in Boolean networks, which model discrete gene regulatory interactions, for their increased interpretability. Indeed, in this type of network, the expression level of a gene is reduced to binary values (genes are either expressed, or not expressed), and is the product of a unique logical function, which takes inputs from the expression states of *direct* regulators of this gene [34, 64]. Yet building a Boolean network for a large number of genes is a painful task without automation.

In order to tackle these issues, we developed a fully automated pipeline to infer a Boolean network which models the regulatory interactions in a well-chosen cell line. Our method is based on gene perturbation experiments, and on the integration of supplementary biological information to further constraint our inference procedure. Transcriptional profiles are extracted from the LINCS L1000 database, which collects a large number of profiles for several cell types and genetic perturbations [62]. In our application, we focused on the gene module M30, which global expression was shown to be anticorrelated with various epileptic profiles and with the severity of epilepsy [18]. Using the Boolean network selected by this method, we ranked genes in M30 according to their ability to permanently modify the global expression of the network, and prioritized top genes. In favor of their important role in epilepsy-related biological processes, this set of candidate master regulators was significantly enriched in terms associated with epilepsy and neurodevelopmental issues with respect to the M30 module.

2 Methods

2.1 Reproductible inference of a cell-line specific Boolean network

This part of our work aims at designing a method which, given a subset of genes of interest, is able to retrieve a gene regulatory network on these genes that allows the prediction of transcriptomic profiles under perturbation. We consider the formalism of Boolean networks, introduced in [34, 64], which are popular models to describe gene regulations.

[4] In this notation, rows are genes, and columns are samples.

Boolean networks. A Boolean network is first characterized by a graph –that is, the network– which connects genes by their regulatory interactions. Such connections are enriched with the direction of the interaction, which distinguishes between regulator and regulated genes, and with the sign of this interaction, that is, whether the regulator inhibits or activates the expression of its target. Second, the dynamics of the system are described by logical functions, called "gene regulatory functions", where variable correspond to the binary expression level (or state) of genes in the network. A single function is assigned to each gene. The expression of a given variable is set to 1 if the associated gene is expressed, otherwise 0. For a given gene g, its associated formulæ contains in its premise the variables corresponding to *direct* regulators of g –*i.e.*, direct predecessors of the node in the network– and in its conclusion the variable associated with the expression level of g. Then, given the expression states of the regulators at a given time step, one can obtain the expression state of the considered gene at the next time step, by evaluating the corresponding formulæ. The network state (or configuration) is the concatenation of all gene expression states. The order of evaluation of regulatory functions to go from one network state to another is called "update step". From a Boolean network, one can build a *state-transition diagram*, where an edge goes from a given network state A to another network state B if and only if one can reach state B from state A in a single update step. One can read from this diagram attractor states, that is, self-looping nodes, which are defined as steady stable network configurations. That is, the application of the update step to this configuration will lead to itself. Attractor states are interesting because they are commonly related to observable biological phenotypes [7, 68]. This diagram also displays cycles of configurations, which correspond to unsteady stable configurations ; the application of the update step makes the system oscillate between a set of configurations in a cyclic way, and can also have a biological interpretation [65]. A state-transition diagram is associated with a given model *and* a type of update. Several types have been suggested in the literature [12], the most well-known ones being the synchronous and the asynchronous updates. In the former, all regulatory functions are evaluated in a single step, whereas in the asynchronous update, only one regulatory function is evaluated at one update step. Recently, [51] have introduced new dynamics for Boolean networks, which was shown to be flexible enough to represent (a)synchronous dynamics as well as multi-level formalisms, that is, beyond boolean values for gene expression.

Building a Boolean network from scratch. Our work focused on combining several data sources and methods for the design of an end-to-end pipeline for Boolean network synthesis, as represented in Figure 1. This network models the regulatory dynamics on a subset of genes, in the absence of external perturbation, in a well-chosen cell line ; for instance, the regulations between M30 genes in brain cell lines for our application to epilepsy. Contrary to the contemporaneous work of [45] applied to cancer, here we do not have any access to a generic model which could model any type of epilepsy to start with. Moreover, relying too much on prior epilepsy-oriented knowledge might lead us to find

already known gene candidates, whereas finding novel master regulators might help investigating refractory epilepsy. The big picture of this pipeline comprises of the following three main steps in chronological order, respectively denoted (A), (B) and (C) in Figure 1 :

(A) Data collection. Step **(A)** encompasses the collection and filtering of information from public, large databases : measurements of transcriptomic data are retrieved from the LINCS L1000 database [62] using careful filtering and quality control measures ; known unsigned, undirected protein pairwise regulatory interactions involving genes in M30 are obtained from the STRING database [63].

(B) Data processing. Then, step **(B)** comprises of the processing of this information into appropriate inputs for the inference of Boolean networks. First, a set of binarized phenotypes is built, corresponding to profiles from single gene perturbations and their associated controls, from LINCS L1000. Then, a signed network of possible valid regulatory interactions is constructed from the protein-protein interactions in STRING, by filtering out and signing edges based on gene pairwise expression correlations computed on LINCS L1000 profiles.

(C) Network inference. Finally, step **(C)** starts by the inference of a set of Boolean networks which satisfy all the experimental and topological constraints given by the phenotypes and the signed network. The experimental constraints comprise of knockout or overexpression experiments, where the control phenotype is considered the initial condition, and the perturbed phenotype the final configuration reached after the gene perturbation. A Boolean network solution should satisfy all of these time-series constraints by only considering a set of regulatory interactions present in the signed network. The final step of the procedure is the selection of an optimal Boolean network among these solutions, according to its topology. This final inferred network is selected through a desirability function maximization [4], which depends on several topological measures.

Details in the implementation are available in Appendix A. Note that our method can also guess an appropriate gene subset associated with a given disease, although this issue did not arise in our application to epilepsy. If no gene set is provided, the method automatically retrieves genes from DisGeNet [54] using the disease Concept ID (CID) [20]. Table 6 in Appendix reports the values used to filter out genes from the DisGeNet database. The single network obtained at the end of step **(C)** is a dynamical system which can predict the behavior of gene expression under one or several gene perturbations, by considering the stable states (attractors and cycles) reachable from a given initial state under these perturbations. We now dwell on how to use this model to rank genes according to their regulatory influence on the remainder of the network.

2.2 Detection of master regulators in a specific disease-context

As mentioned in Section 1, when looking for therapeutic candidates, one might be interested in master regulators, that is, genes at the top of the gene regulation hierarchy, which change in expression induces the largest downstream gene expression change ; for instance, by encoding for a transcription factor which

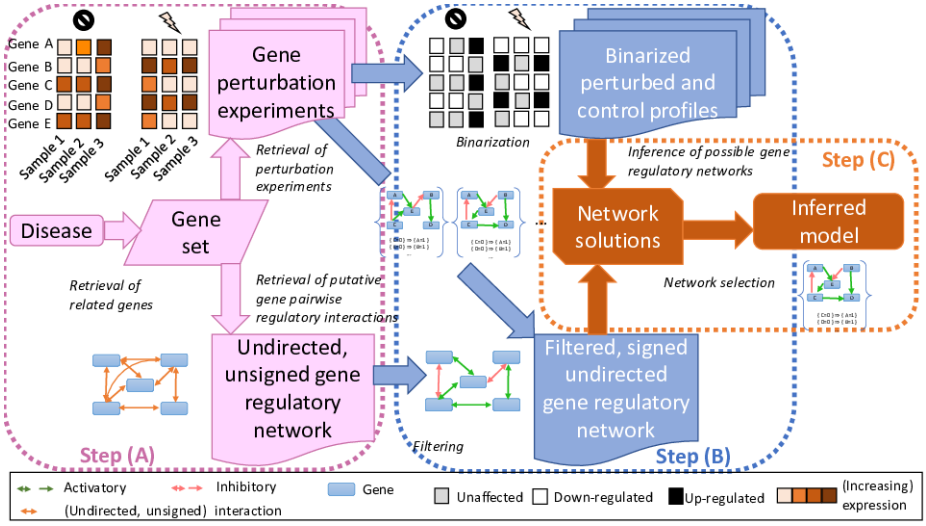

Fig. 1. Overview of the pipeline for the automated building of Boolean networks.

affects the transcription of other genes [43]. In practice, it is frequently quantified using the node (outgoing) degree and detection of hub nodes in the network. Many measures defining such a "centrality" property can be computed using Cytoscape [56], through modules NetworkAnalyzer [3] and CytoCtrlAnalyser [70] : for instance, control centrality [42], which has been recently used to identify regulations between *NFATC4* and Type 2 diabetes-associated genes [57]. However, these measures only use the topological information in the network, whereas our network inference pipeline allows –along with the identification of regulatory connections– the inference of gene regulatory functions, adding interesting dynamical information. This is why we designed a master regulator detection method which leverages this information to model regulatory cascades. In [25], a Machine Learning technique called "influence maximization" was exploited to identify key genes in a continuous model of the yeast regulatory network. In our work, we adapted influence maximization to Boolean networks. For long, online recommendation and advertising researchers have been interested in influence maximization [35], which aims at finding a node subset of fixed size which influences most the remainder of the network. In order to make this technique applicable to Boolean networks, we need to explicitly define the concept of influence on gene expression in these models. This influence is called "spread process", and is the quantity that propagates along the edges of the network from any node. We define influence in an iterative way ; first we consider a single gene and one initial network state. Then, we proceed to define multi-gene influence starting from an initial state. We eventually describe influence for several genes across a set of initial network states.

Genewise influence in a Boolean network. The most intuitive definition of influence –denoted in the remainder of the paper "spread value"– $I(n)$ of a given node n on the other nodes in a Boolean network with initial state i, would be that any perturbation of this node would "greatly change" the attractor states reachable from state i, compared to attractor states reachable from that state in the absence of perturbation. We define this great change, yielding a positive spread value, by the fact that those two sets of attractors have an empty intersection. Let us denote $\mathcal{A}(i, P)$ the set of attractor states reachable from state i, under the set of perturbations P. Set P contains pairs of gene names and their associated perturbation (either 0 for knockout, or 1 for overexpression). Let us also denote $\mathcal{O}$ the set of output genes, that we define here as the set of genes with a positive ingoing-degree, and a given similarity measure $\mathcal{S}$ between network states. Then, we define the spread value for node n, initial state i, and Boolean network $\mathcal{B}$ as

$$\mathrm{SV}_{\mathcal{B}}(\{n\}, i) = 1 - \max\left\{\mathcal{S}\left(a^1_{|\mathcal{O}}, a^2_{|\mathcal{O}}\right) : a^1, a^2 \in \mathcal{A}(i, \emptyset) \times \in \mathcal{A}(i, \{(n, \neg i[n])\})\right\} ,$$

where $a^1_{|\mathcal{O}}$ and $a^2_{|\mathcal{O}}$ are the respective restrictions of network states a^1 and a^2 to the set of output genes $\mathcal{O}$. The perturbation denoted by $(n, \neg i[n])$ means that gene n is perturbed in the opposite direction to its expression state $i[n]$ in i : for instance, if n is expressed in state i, then we consider knockouts of gene n. The restriction to output genes in $\mathcal{O}$ is actually important in order to have consistent results when considering isolated nodes.

Note that, if n does not have a determined expression state in initial state i, we set the associated perturbation set to $\emptyset$. This implies that some genes with individual spread value equal to 0 can either have no true influence on the network, or have no determined expression state in initial state s, which means that they are not measured during the generation of transcriptomic data. In the latter case, this most likely means that the gene is not expressed in the considered cell line(s).

Most importantly, this value $\mathrm{SV}_{\mathcal{B}}(\{n\}, i)$ is equal to 0 if and only if

$$\mathcal{A}(i, \emptyset) \cap \mathcal{A}(i, \{(n, \neg i[n])\}) \neq \emptyset ,$$

that is, if there is any attractor in common. However, if the intersection is empty, $\mathrm{SV}_{\mathcal{B}}(\{n\}, i)$ is not necessarily equal to 1, as reachable attractors might still be close to those obtained without any external perturbation.

Geneset influence in a Boolean network. When considering a set $\mathcal{N}$ of node instead of a single node n, the influence of $\mathcal{N}$ is the spread value computed over all attractors reachable under simultaneous perturbations of these nodes

$$\mathrm{SV}_{\mathcal{B}}(\mathcal{N}, i) = 1 - \max\left\{\mathcal{S}\left(a^1_{|\mathcal{O}}, a^2_{|\mathcal{O}}\right) : a^1, a^2 \in \mathcal{A}(i, \emptyset) \times \mathcal{A}(i, \{(n, \neg i[n]) : n \in \mathcal{N}\})\right\} .$$

Aggregation of values for several initial states. Finally, if we consider a whole set of initial states $\mathcal{I}$ and a gene set $\mathcal{N}$, the associated influence is defined

as the geometric mean of spread values across initial states

$$\mathrm{SV}_{\mathcal{B}}(\mathcal{N},\mathcal{I}) = \left(\prod_{i\in\mathcal{I}}(\mathrm{SV}_{\mathcal{B}}(\mathcal{N},i)+1)\right)^{1/|\mathcal{I}|} - 1 \ ,$$

where $|\mathcal{I}|$ is the number of initial states. Note that we need to correct for zeroes to order to avoid the collapse of this measure when one perturbation does not trigger a change in reachable attractors for one of the initial states, while keeping spread values between 0 and 1 for better interpretability.

Once the spread process is defined, we propose the following greedy influence maximization algorithm to prioritize master regulators.

Algorithm 1 Greedy influence maximization algorithm for Boolean networks

Input: $\mathcal{B}$ a Boolean network on node set V ; K the minimal number of simultaneous perturbations on the network ; $\mathcal{I}$ set of initial Boolean states
Initialize $\mathcal{N} = \emptyset$, $k = 0$
repeat
$k \leftarrow k + 1$
Adding to set $\mathcal{N}$ nodes that maximize the spread value

$$\mathcal{N} \leftarrow \mathcal{N} \cup N_k \ , \text{ where } N_k \leftarrow \arg\max_{n\in V\setminus\mathcal{N}} \mathrm{SV}_{\mathcal{B}}(\mathcal{N}\cup\{n\},\mathcal{I})$$

until

$$k = K \text{ or } \max_{n\in V\setminus\mathcal{N}} \mathrm{SV}_{\mathcal{B}}(\mathcal{N}\cup\{n\},\mathcal{I}) \leq \mathrm{SV}_{\mathcal{B}}(\mathcal{N},\mathcal{I})$$

Output: $\mathcal{N}$

Influence maximization algorithm on Boolean networks. We describe how to leverage spread values to identify master regulators in the network. Current literature on influence maximization [52], which is a NP-hard problem, relies on the fact that the spread function is submodular : roughly, as the considered subset increases, the difference in the value of this function due to adding another single element to the subset decreases. However, no such property can be assessed for the definition of spread defined in the previous paragraph. We then slightly adapted the greedy algorithm in [35] in Algorithm 1. This algorithm determines the set of nodes of minimal size K which are the most influent, where K is a predefined fixed value. It goes as follows : starting from an empty set of nodes $\mathcal{N}_0$, a fixed set of initial states $\mathcal{I}$, and a Boolean network $\mathcal{B}$, at each step $k \in \{1, 2, \ldots, K\}$, the algorithm selects the node $n \notin \mathcal{N}_k$ which maximizes spread value $\mathrm{SV}_{\mathcal{B}}(\mathcal{N}_k \cup \{n\}, \mathcal{I})$ and computes the set $\mathcal{N}_{k+1} = \{n\} \cup \mathcal{N}_k$. The algorithm stops at $k = K$, or at the first step k when the spread value $\mathrm{SV}_{\mathcal{B}}(\mathcal{N}_k, \mathcal{I})$ is no longer increasing, that is,

$$\max\{\mathrm{SV}_{\mathcal{B}}(\mathcal{N}_k \cup \{n\}, \mathcal{I}) : n \notin \mathcal{N}_k\} \leq \mathrm{SV}_{\mathcal{B}}(\mathcal{N}_k, \mathcal{I}) \ .$$

This condition is necessary to compensate for the fact that the function might not be submodular. If, at a given step k, several nodes maximize the spread value, they are all added to set $\mathcal{N}_{k+1}$. The iteratively built set $\mathcal{N}_K$ is then the set of possible K-sized gene subsets to simultaneously perturb on the network, such that the set of attractors reachable from initial set $\mathcal{I}$ is greatly modified. In this work, $K = 1$, that is, we only looked at individual contributions of genes to the changes, and we ranked gene n among genes in the network according to its spread value $\mathrm{SV}_{\mathcal{B}}(\{n\}, \mathcal{I})$.

Set of initial network states ($\mathcal{I}$). We consider transcriptomic profiles from human hippocampi afflicted with temporal lobe epilepsy (TLE) in [44] for the initial states, such that genes are ranked according to their influence in an epileptic context. Temporal lobe epilepsy is one of the most common forms of partial epilepsy, where seizures affect one part of the brain, and is often associated with cases of refractory epilepsy that cannot be surgically treated [28]. Details about the implementation and initial states are available in Appendix C. 207 genes out of 232 genes both from the M30 module and present in the network are mapped to expression levels in these states, which means that for these genes, we are sure that any spread value equal to 0 for any of these genes truly means that the gene has no influence over the remainder of the network.

Similarity between attractor states ($\mathcal{S}$). The definition of the spread process relies on a similarity function $\mathcal{S}$ defined between two network states, that was left to be defined. In our implementation, we wanted to compute the differences in the presence of ones *and* zeroes, which prevented us from directly using Jaccard's score. Based on previous surveys of the state-of-the-art on binary distances [15], we implemented a "normalized" ℓ_1-norm distance. That is, if a^1 and a^2 are the two binary vectors to compare (of size d), then the resulting similarity $\mathcal{S}(a^1, a^2)$ between a^1 and a^2 is :

$$\mathcal{S}(a^1, a^2) = 1 - \frac{1}{d} \sum_{i=1}^{d} |a^1[i] - a^2[i]| \; .$$

This expression is exactly the percentage of row-wise equal coefficients in a^1 and a^2, and yields 1 when $a^1 = a^2$, and 0 for $a^2 = (a^1 + 1) \equiv [2]$ (modulo 2). It penalizes in a symmetric way differences in 1's and 0's.

Genericity of the method. Note that this methodology, combining the synthesis of a Boolean network and influence maximization, can generically be applied to any disease. To adapt this pipeline to another disease, one needs to change the gene subset and the cell line(s) on which the network should be built, as well as the set of initial network states for the detection of master regulators. In the application to epilepsy, we considered the M30 gene module, and the two brain cell lines present in the LINCS L1000 database. Code for the synthesis of the Boolean network and the detection of master regulators is available at `https://github.com/clreda/PrioritizationMasterRegulators`.

3 Results

3.1 Networks obtained from the inference procedure

We discuss the network solutions resulting from step **(C)**.

The final network compiles several sources of regulation. The final network obtained at the end of step (C) is shown in Figure 2. In this figure, nodes are colored by their degree ; the darker the color, the higher the degree. Edges in Figure 2 are colored according to their source of evidence as reported by the STRING database [63]. One can notice that there are a lot of undirect gene-to-gene regulatory interactions in this network. This actually is not very surprising, since few gene pairwise interactions are experimentally tested compared to all possibly existing interactions. Moreover, our model does not aim at taking into account exclusively transcriptomic interactions, but possibly non-physical, post-transcriptomic effects.

The synthesis outputs similar network solutions. Now, we consider all 25 [5] network solutions generated at step **(C)**. We assess how far they are from each other, in terms of node degree distribution, edge numbers, redundancy in interactions, unicity of regulatory functions for each node across those solutions, and values of general topological parameter (GTP). GTP is a value comprised between 0 and 1 that is used to select the final network among the 25 ones (as further described in Subsection A.6 in Appendix) and characterizes the proximity of a network topology to a scale-free-like one. Table 1 shows distribution statistics about the values of GTP and the unicity of gene regulatory functions across solutions. Note that all solutions present similar topologies, with similar GTP scores quite close to 1, which matches what can be expected from biochemical interaction networks in non-fungi systems [10]. Moreover, except for less than 25% of the genes in the network, genes are assigned at most 3 different regulatory functions across all solutions, which shows that their function in the network is globally preserved. Figure 3 displays the boxplots of edge number and node degree distributions across solutions. These two plots show that, as mentioned before, the typical scale-free topology, with a few "hub nodes" with large degree and a large number of genes with few regulatory interactions, is present in all solutions. There are 74 interactions (that is, around $30 - 34\%$ of edges) which are present in at least 75% of the solutions, among which 25 are present in all of them. They are shown in Table 7 located at the Appendix. These numbers are confirmed by plotting the network comprising of all genepairwise interactions which are present in at least one solution, shown in Figure 4. All in all, the networks obtained just before the model selection step mostly seem similar, both functionally –at the level of regulatory functions– and topologically –considering the node degrees, the number of edges, and the GTP scores.

[5] That number was chosen for reasons related to computational cost and time. Note that in Appendix we discuss how adding 25 additional network solutions neither changes the final network, nor the conclusions made in this section.

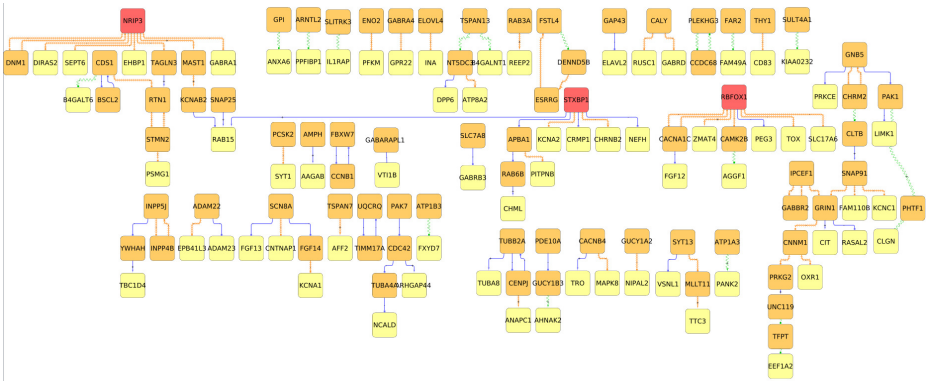

Fig. 2. Inferred network resulting from our pipeline applied to the M30 gene module. Tee-headed arrows represent inhibitory regulatory interactions, whereas triangle-headed arrows are activatory regulations. Edges are drawn according to their source of evidence (on the *undirected* interactions) as reported by the STRING database: contiguous arrows denote coexpression, solid line denote experimental proofs of interaction, and sinewave interactions are derived from text-mining procedures. Gene nodes are colored according to their out-degree: lightest color for genes with outdegree equal to 0, darkest for genes with an outdegree higher than 5. Isolated nodes are not shown.

Table 1. Distribution statistics on the number of *unique* regulatory functions (RFs) across solutions per gene, and on the value of the general topological parameter (GTP) used for network selection in step (C) of the inference procedure. All values are rounded up to the 3$^{\text{rd}}$ decimal place.

	Min.	25$^{\text{th}}$ quantile	Median	Mean	75$^{\text{th}}$ quantile	Max.
# RFs	1	1	2	2.202	3	11
GTP	0.796	0.798	0.800	0.800	0.800	0.802

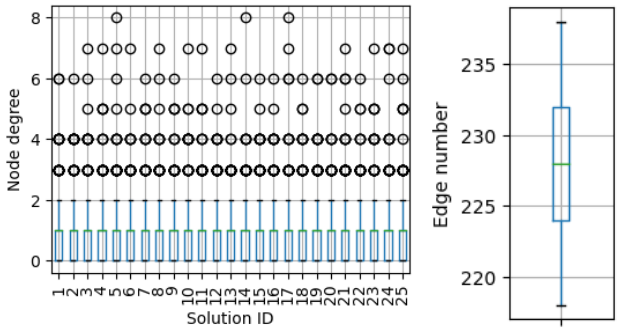

Fig. 3. *Left-hand plot*: Boxplots of node total degrees (ingoing and outgoing degree) per solution. The green lines represent median values. *Right-hand plot*: Boxplot of the number of edges across solutions (which all comprise of 232 M30 genes). Again, the green line represent the median value.

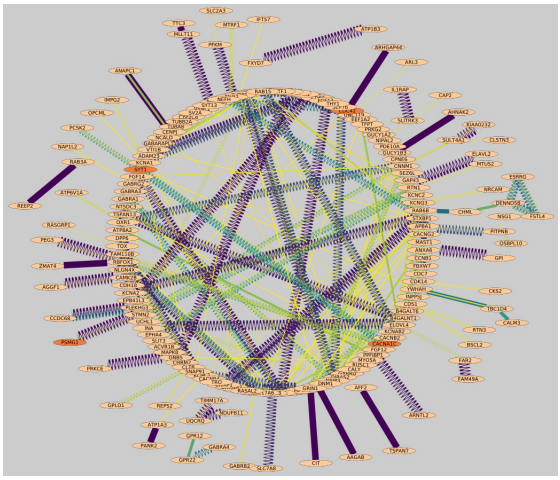

Fig. 4. Network comprising of all gene-to-gene interactions which are present in at least one solution. The darker and thicker an edge is, the more frequent it is across solutions. Sinewave edges are inhibitory interactions, whereas solid lines denote activatory interactions. Orange nodes correspond to the genes which are perturbed in the LINCS L1000 experimental profiles used for inference.

3.2 Recommended master regulator candidates

We now study the master regulator candidates ranked by spread value.

Spread values are correlated with network centrality and gene pathogenicity. We computed the correlation between spread values for our application to epilepsy, genewise Control Centrality [42] values computed with CytoCtrlAnalyser [70], and genewise outgoing degrees. The outgoing degree is the number of direct downstream targets, whereas Control Centrality is the number of nodes which are affected by a change in the considered node, based on the topology of the directed network. More specifically, in order to compute the Control Centrality for any gene g, at some time step t, (continuous) expression levels $x(t) \in \mathbb{R}^N$ are time-invariant and depend linearly on those at the previous time step $t-1$ $x(t-1) \in \mathbb{R}^N$, where N is the number of genes in the network

$$\frac{\partial x(t)}{\partial t} = Ax(t) + u_g(t) \ , \tag{2}$$

where A is the adjacency matrix in $\mathbb{R}^{N \times N}$ associated with the network, and $u_g(t) \in \mathbb{R}$ is the external signal imposed on node g at time t (either overexpression if it is positive, or knockout otherwise). In such a system, computing the number of nodes which can be controlled by gene g boils down to getting the rank of the so-called controllability matrix related to A and g, that is a function of powers of matrix A. This rank can be computed by solving a combinatorial optimization problem described in Equation (3) in [42]. Moreover, since the true nonzero values in A as well as $u_g(\cdot)$ are often unknown, Control Centrality

aims at quantifying *structural* controllability, independently from the values of nonzero coefficients in A and $u_g(\cdot)$. All in all, Control Centrality is a solid counterpart to our method. It does not take into account neither the set of regulatory functions nor the gene expression levels in patients, but models regulatory cascades through the differential equation in Equation 2. We also compared spread values to scores associated with the pathogenicity of genes :

- probability of loss of function intolerance (pLI) [38], which quantifies the intolerance to the loss of function of a given gene in patient and control cohorts.

- enhancer-domain score (EDS) [67], which studies the conservation of the regulatory domain around genes.

- residual variation intolerance score (RVIS) [53], which is related to the presence of functional genetic variation in patient exomes, and is anticorrelated with gene pathogenicity.

Finally, we computed "TF" influence scores [48] as well, which expression is reported in Equation 1. Figure 5 displays the correlation heatmap between these different measures. We observed that, contrary to (TF) influence values, spread values were consistent and strongly correlated with network controllability measures, that is Control Centrality and the outgoing degree. Moreover, spread values are more strongly correlated with gene pathogenicity-related measures pLI and (opposite of) RVIS. We tested whether the spread value was actually totally determined by the number of downstream (not necessarily direct) regulated genes. To do so, we performed a Spearman's ρ linear correlation test on the spread values and the number of downstream regulated targets for each gene. We confirmed that there is a strong, significative correlation between the two –which is expected, given the definition of the spread value– but that the spread value is not completely determined by this value ; that is, the associated statistic is not equal to 1 ($\rho = 0.82$, $p = 3.10^{-57}$).

Top genes for spread values are significatively enriched in disease-related terms. Moreover, from Figure 6, it can be noticed that there is a lot of discrepancy between pLI scores and spread values on M30 genes. Nonetheless, it should be noted that [72] warns against genes which are involved in recessive forms of diseases, while having a low pLI score. That is actually the case for gene *GNB5*, which has a central place in our network (shown in Figure 1) with spread value 0.024, pLI score close to 0, and is involved in a recessive form of epileptic encephalopathy [55]. Moreover, using the online tool WebGestalt [39], we performed a Over-Representation Analysis (ORA), in order to check if the shortlist of 14 genes with spread value greater than 0.01 was significantly enriched in epilepsy-associated terms, compared to the 232 genes present in the network. The disease terms were annotations from the DisGeNet database [54]. [6] Indeed, this shortlist is (weakly) significantly enriched in genes related to the term "Epileptic encephalopathy" at level 5% (odds ratio $OR = 7.5$, Benjamini-Hochberg (BH) [6]-adjusted $p \approx 0.038$), and more strongly enriched with (neuro)developmental issues, for instance, "Loss of developmental mile-

[6] Remember that, in the application to epilepsy, we *did not* use genes from DisGeNet, but the preselected set of genes M30.

stones" ($OR = 10.5$, BH-adjusted $p \approx 0.012$), as reported in Figure 5. Similar results can be observed on another family of gene annotations, GLAD4U [32], as shown in Figure 8 in Appendix. The considered shortlist of genes is shown on Figure 6. These enrichment results go beyond the fact that M30 is globally enriched in epilepsy-related *de novo* mutations compared to the *whole* measured genome in brain cell lines, as shown in [18] : what is shown is that, among genes *in the M30 module*, ranking by spread values still prioritized interesting genes.

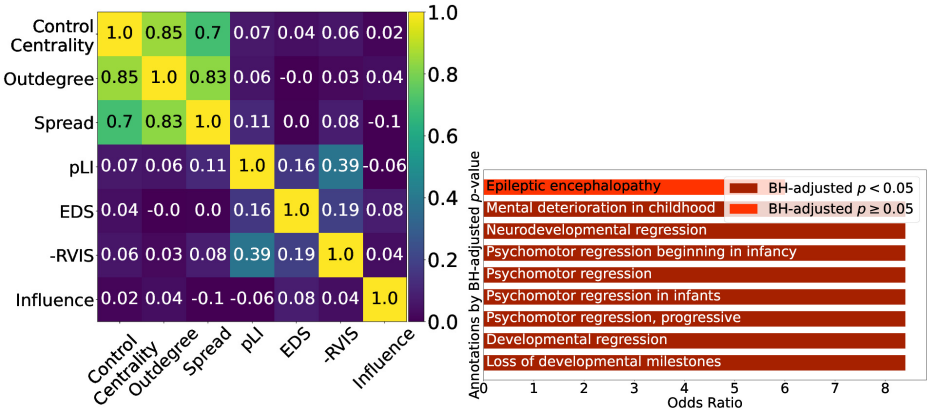

Fig. 5. *Left*: Spearman's ρ correlation heatmap between different gene measures either related to the influence of a node on a network, or to the genetic variations associated with pathogenicity. *Right*: Enrichment results from the ORA analysis. All reported adjusted p-values are lower than 20%.

Most top genes cause human epilepsies. Based on the results shown in Figure 5 and Table 2, a shorter list of candidate genes is selected. It comprises of genes with rather large spread value (greater than 0.01) and pLI score (greater than 0.9), *and* of genes with very large spread value (greater than 0.02). The last condition holds in order to avoid the previously mentioned shortcoming incurred by pLI scores. These candidate genes are *CACNA1A*, *RBFOX1*, *STXBP1*, *DNM1*, *NRIP3*, *SCN8A*, *CHRM2*, *GNB5*, *TUBB2A*, *PAK7*, and *GRIN1*, shown on Figure 6. Most of these candidates –except for *NRIP3*, which is notably mainly expressed in the hippocampus– have a relationship to epilepsy-related symptoms in humans shown in prior works [47, 37, 60, 2, 11, 55, 1, 17, 50], as expected due to their membership to the M30 module. Some of these genes may never been investigated in the research related to epilepsy, such as *NRIPS3*. For other genes, for instance, *STXBP1* and *GRIN1*, knockouts of orthologuous genes were associated with epileptic seizures in zebrafish [27].

Table 2. Distribution statistics (rounded up to the 5^{th} decimal place) of the spread values obtained for M30 genes present in the inferred network.

Minimum	25^{th} quantile	Median	Mean	75^{th} quantile	Maximum
0.0	0.0	0.0	0.00254	0.0	0.0556

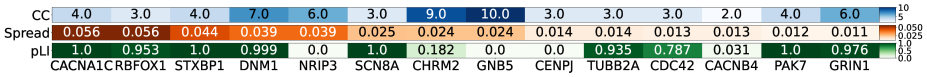

Fig. 6. Genes ranked by decreasing spread value, restricted to spread value greater than 0.01 (*center bar*), with their associated Control Centrality (CC) (*top bar*), and pLI scores (*bottom bar*).

4 Discussion

We introduced in this work two main contributions to *in silico* disease research. First, we designed a method for the automated inference of a gene regulatory network from scratch, starting from a subset of genes. This method carefully combines information from several public databases and methods, and infers a dynamical model of gene regulation adapted to specific cell lines. This method yields quite robust network solutions, and is easily reproducible. Then, we showed how to exploit this dynamical system to detect master regulator genes. We applied our methodology to investigate epilepsy, and to find novel candidate genes to hopefully tackle drug-resistant epilepsy. A list of candidate genes was prioritized, which perturbations greatly impact the whole network in an epileptic transcriptomic context. This methodology allows reproducible and transparent research, while reducing the amount of data needed as input, which is one of the main caveats of researching on rare or tropical neglected diseases.

Acknowledgements

This work was supported by the Institut National pour la Santé et la Recherche Médicale (Inserm, France), the French Ministry of Higher Education and Research (#ENS.X19RDTME-SACLAY19-22) (C.R.), Université Paris Cité, Université Sorbonne Paris Nord, the French National Research Agency (#ANR-19-CE23-0026-04) (C.R.), (#ANR-18-CE17-0009-01) (A.D.-D., C.R.), (#ANR-18-CE37-0002-03) (A.D.-D.,C.R.). The supporting bodies played no role in any aspect of study design, analysis, interpretation or decision to publish this data.

References

1. Al-Eitan, L.N., Al-Dalalah, I.M., Mustafa, M.M., Alghamdi, M.A., Elshammari, A.K., Khreisat, W.H., Al-Quasmi, M.N., Aljamal, H.A.: Genetic polymorphisms of cyp3a5, chrm2, and znf498 and their association with epilepsy susceptibility:

a pharmacogenetic and case–control study. Pharmacogenomics and personalized medicine **12**, 225 (2019)
2. Appenzeller, S., Balling, R., Barisic, N., Baulac, S., Caglayan, H., Craiu, D., De Jonghe, P., Depienne, C., Dimova, P., Djémié, T., et al.: De novo mutations in synaptic transmission genes including dnm1 cause epileptic encephalopathies. The American Journal of Human Genetics **95**(4), 360–370 (2014)
3. Assenov, Y., Ramírez, F., Schelhorn, S.E., Lengauer, T., Albrecht, M.: Computing topological parameters of biological networks. Bioinformatics **24**(2), 282–284 (2008)
4. Babichev, S., Durnyak, B., Senkivskyy, V., Sorochynskyi, O., Kliap, M., Khamula, O.: Technique of gene regulatory networks reconstruction based on aracne inference algorithm. In: IDDM. pp. 195–207 (2019)
5. Béal, J., Montagud, A., Traynard, P., Barillot, E., Calzone, L.: Personalization of logical models with multi-omics data allows clinical stratification of patients. Frontiers in physiology p. 1965 (2019)
6. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological) **57**(1), 289–300 (1995)
7. Bloomingdale, P., Nguyen, V.A., Niu, J., Mager, D.E.: Boolean network modeling in systems pharmacology. Journal of pharmacokinetics and pharmacodynamics **45**(1), 159–180 (2018)
8. Bolouri, H., Davidson, E.H.: Transcriptional regulatory cascades in development: initial rates, not steady state, determine network kinetics. Proceedings of the National Academy of Sciences **100**(16), 9371–9376 (2003)
9. Bravais, A.: Analyse mathématique sur les probabilités des erreurs de situation d'un point. Impr. Royale (1844)
10. Broido, A.D., Clauset, A.: Scale-free networks are rare. Nature communications **10**(1), 1–10 (2019)
11. Butler, K.M., da Silva, C., Shafir, Y., Weisfeld-Adams, J.D., Alexander, J.J., Hegde, M., Escayg, A.: De novo and inherited scn8a epilepsy mutations detected by gene panel analysis. Epilepsy research **129**, 17–25 (2017)
12. Chatain, T., Haar, S., Paulevé, L.: Boolean networks: beyond generalized asynchronicity. In: International Workshop on Cellular Automata and Discrete Complex Systems. pp. 29–42. Springer (2018)
13. Cheng, L., Li, L.: Systematic quality control analysis of lincs data. CPT: pharmacometrics & systems pharmacology **5**(11), 588–598 (2016)
14. Chevalier, S., Froidevaux, C., Paulevé, L., Zinovyev, A.: Synthesis of boolean networks from biological dynamical constraints using answer-set programming. In: 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI). pp. 34–41. IEEE (2019)
15. Choi, S.S., Cha, S.H., Tappert, C.C.: A survey of binary similarity and distance measures. Journal of systemics, cybernetics and informatics **8**(1), 43–48 (2010)
16. Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C.: Introduction to algorithms. MIT press (2009)
17. Cushion, T.D., Paciorkowski, A.R., Pilz, D.T., Mullins, J.G., Seltzer, L.E., Marion, R.W., Tuttle, E., Ghoneim, D., Christian, S.L., Chung, S.K., et al.: De novo mutations in the beta-tubulin gene tubb2a cause simplified gyral patterning and infantile-onset epilepsy. The American Journal of Human Genetics **94**(4), 634–641 (2014)

18. Delahaye-Duriez, A., Srivastava, P., Shkura, K., Langley, S.R., Laaniste, L., Moreno-Moral, A., Danis, B., Mazzuferi, M., Foerch, P., Gazina, E.V., et al.: Rare and common epilepsies converge on a shared gene regulatory network providing opportunities for novel antiepileptic drug discovery. Genome biology **17**(1), 1–18 (2016)
19. DisGeNet: Faq : Original data sources. https://www.disgenet.org/ (2022), accessed: [May 4, 2022]
20. Doğan, R.I., Leaman, R., Lu, Z.: Ncbi disease corpus: a resource for disease name recognition and concept normalization. Journal of biomedical informatics **47**, 1–10 (2014)
21. Dunn, S.J., Yordanov, B.: Automated reasoning for the synthesis and analysis of biological programs. In: Automated Reasoning for Systems Biology and Medicine, pp. 37–62. Springer (2019)
22. Eves, E.M., Tucker, M.S., Roback, J.D., Downen, M., Rosner, M.R., Wainer, B.H.: Immortal rat hippocampal cell lines exhibit neuronal and glial lineages and neurotrophin gene expression. Proceedings of the National Academy of Sciences **89**(10), 4373–4377 (1992)
23. Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A.K., Slichter, C.K., Miller, H.W., McElrath, M.J., Prlic, M., et al.: Mast: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell rna sequencing data. Genome biology **16**(1), 1–13 (2015)
24. Gebser, M., Kaminski, R., Kaufmann, B., Ostrowski, M., Schaub, T., Wanko, P.: Theory solving made easy with clingo 5. In: Technical Communications of the 32nd International Conference on Logic Programming (ICLP 2016). Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik (2016)
25. Gibbs, D.L., Shmulevich, I.: Solving the influence maximization problem reveals regulatory organization of the yeast cell cycle. PLoS computational biology **13**(6), e1005591 (2017)
26. González, F.L., Osorio, X.R., Rein, A.G.N., Martínez, M.C., Fernández, J.S., Haba, V.V., Pedraza, A.D., Cerdá, J.M.: Drug-resistant epilepsy: definition and treatment alternatives. Neurología (English Edition) **30**(7), 439–446 (2015)
27. Griffin, A., Carpenter, C., Liu, J., Paterno, R., Grone, B., Hamling, K., Moog, M., Dinday, M.T., Figueroa, F., Anvar, M., et al.: Phenotypic analysis of catastrophic childhood epilepsy genes. Communications biology **4**(1), 1–13 (2021)
28. Han, C.L., Hu, W., Stead, M., Zhang, T., Zhang, J.G., Worrell, G.A., Meng, F.G.: Electrical stimulation of hippocampus for the treatment of refractory temporal lobe epilepsy. Brain research bulletin **109**, 13–21 (2014)
29. Harrington, E.C.: The desirability function. Industrial quality control **21**(10), 494–498 (1965)
30. Huang, Y., Furuno, M., Arakawa, T., Takizawa, S., de Hoon, M., Suzuki, H., Arner, E.: A framework for identification of on-and off-target transcriptional responses to drug treatment. Scientific reports **9**(1), 1–9 (2019)
31. Johnson, M.R., Behmoaras, J., Bottolo, L., Krishnan, M.L., Pernhorst, K., Santoscoy, P.L.M., Rossetti, T., Speed, D., Srivastava, P.K., Chadeau-Hyam, M., et al.: Systems genetics identifies sestrin 3 as a regulator of a proconvulsant gene network in human epileptic hippocampus. Nature communications **6**(1), 1–11 (2015)
32. Jourquin, J., Duncan, D., Shi, Z., Zhang, B.: Glad4u: deriving and prioritizing gene lists from pubmed literature. BMC genomics **13**(8), 1–12 (2012)
33. Kalume, F., Westenbroek, R.E., Cheah, C.S., Frank, H.Y., Oakley, J.C., Scheuer, T., Catterall, W.A., et al.: Sudden unexpected death in a mouse model of dravet syndrome. The Journal of clinical investigation **123**(4), 1798–1808 (2013)

34. Kauffman, S.A.: Metabolic stability and epigenesis in randomly constructed genetic nets. Journal of theoretical biology **22**(3), 437–467 (1969)
35. Kempe, D., Kleinberg, J., Tardos, É.: Maximizing the spread of influence through a social network. In: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 137–146 (2003)
36. Kuruba, R., Hattiangady, B., Shetty, A.K.: Hippocampal neurogenesis and neural stem cells in temporal lobe epilepsy. Epilepsy & Behavior **14**(1), 65–73 (2009)
37. Lal, D., Trucks, H., Møller, R.S., Hjalgrim, H., Koeleman, B.P., de Kovel, C.G., Visscher, F., Weber, Y.G., Lerche, H., Becker, F., et al.: Rare exonic deletions of the rbfox1 gene increase risk of idiopathic generalized epilepsy. Epilepsia **54**(2), 265–271 (2013)
38. Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al.: Analysis of protein-coding genetic variation in 60,706 humans. Nature **536**(7616), 285–291 (2016)
39. Liao, Y., Wang, J., Jaehnig, E.J., Shi, Z., Zhang, B.: Webgestalt 2019: gene set analysis toolkit with revamped uis and apis. Nucleic acids research **47**(W1), W199–W205 (2019)
40. Lim, N., Pavlidis, P.: Evaluation of connectivity map shows limited reproducibility in drug repositioning. Scientific reports **11**(1), 1–14 (2021)
41. Liu, Y.Y., Slotine, J.J., Barabási, A.L.: Controllability of complex networks. nature **473**(7346), 167–173 (2011)
42. Liu, Y.Y., Slotine, J.J., Barabási, A.L.: Control centrality and hierarchical structure in complex networks (2012)
43. Mattick, J.S., Taft, R.J., Faulkner, G.J.: A global view of genomic information–moving beyond the gene and the master regulator. Trends in genetics **26**(1), 21–28 (2010)
44. Mirza, N., Appleton, R., Burn, S., du Plessis, D., Duncan, R., Farah, J.O., Feenstra, B., Hviid, A., Josan, V., Mohanraj, R., et al.: Genetic regulation of gene expression in the epileptic human hippocampus. Human molecular genetics **26**(9), 1759–1769 (2017)
45. Montagud, A., Béal, J., Tobalina, L., Traynard, P., Subramanian, V., Szalai, B., Alföldi, R., Puskás, L., Valencia, A., Barillot, E., et al.: Patient-specific boolean models of signalling networks guide personalised treatments. Elife **11**, e72626 (2022)
46. Mudunuri, U., Che, A., Yi, M., Stephens, R.M.: biodbnet: the biological database network. Bioinformatics **25**(4), 555–556 (2009)
47. Myers, C.T., McMahon, J.M., Schneider, A.L., Petrovski, S., Allen, A.S., Carvill, G.L., Zemel, M., Saykally, J.E., LaCroix, A.J., Heinzen, E.L., et al.: De novo mutations in slc1a2 and cacna1a are important causes of epileptic encephalopathies. The American Journal of Human Genetics **99**(2), 287–298 (2016)
48. Nicolle, R., Radvanyi, F., Elati, M.: Coregnet: reconstruction and integrated analysis of co-regulatory networks. Bioinformatics **31**(18), 3066–3068 (2015)
49. Ogren, J.A., Wilson, C.L., Bragin, A., Lin, J.J., Salamon, N., Dutton, R.A., Luders, E., Fields, T.A., Fried, I., Toga, A.W., et al.: Three-dimensional surface maps link local atrophy and fast ripples in human epileptic hippocampus. Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society **66**(6), 783–791 (2009)
50. Ohba, C., Shiina, M., Tohyama, J., Haginoya, K., Lerman-Sagie, T., Okamoto, N., Blumkin, L., Lev, D., Mukaida, S., Nozaki, F., et al.: Grin 1 mutations cause

encephalopathy with infantile-onset epilepsy, and hyperkinetic and stereotyped movement disorders. Epilepsia **56**(6), 841–848 (2015)
51. Paulevé, L., Kolčák, J., Chatain, T., Haar, S.: Reconciling qualitative, abstract, and scalable modeling of biological networks. Nature communications **11**(1), 1–7 (2020)
52. Perrault, P., Healey, J., Wen, Z., Valko, M.: Budgeted online influence maximization. In: International Conference on Machine Learning. pp. 7620–7631. PMLR (2020)
53. Petrovski, S., Wang, Q., Heinzen, E.L., Allen, A.S., Goldstein, D.B.: Genic intolerance to functional variation and the interpretation of personal genomes. PLoS genetics **9**(8), e1003709 (2013)
54. Piñero, J., Ramìrez-Anguita, J.M., Saüch-Pitarch, J., Ronzano, F., Centeno, E., Sanz, F., Furlong, L.I.: The disgenet knowledge platform for disease genomics: 2019 update. Nucleic acids research **48**(D1), D845–D855 (2020)
55. Poke, G., King, C., Muir, A., de Valles-Ibáñez, G., Germano, M., Moura de Souza, C.F., Fung, J., Chung, B., Fung, C.W., Mignot, C., et al.: The epileptology of gnb5 encephalopathy. Epilepsia **60**(11), e121–e127 (2019)
56. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., Ideker, T.: Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research **13**(11), 2498–2504 (2003)
57. Sharma, A., Halu, A., Decano, J.L., Padi, M., Liu, Y.Y., Prasad, R.B., Fadista, J., Santolini, M., Menche, J., Weiss, S.T., et al.: Controllability in an islet specific regulatory network identifies the transcriptional factor nfatc4, which regulates type 2 diabetes associated genes. NPJ systems biology and applications **4**(1), 1–11 (2018)
58. Srivastava, P.K., van Eyll, J., Godard, P., Mazzuferi, M., Delahaye-Duriez, A., Van Steenwinckel, J., Gressens, P., Danis, B., Vandenplas, C., Foerch, P., et al.: A systems-level framework for drug discovery identifies csf1r as an anti-epileptic drug target. Nature communications **9**(1), 1–15 (2018)
59. Srivastava, P.K., Bagnati, M., Delahaye-Duriez, A., Ko, J.H., Rotival, M., Langley, S.R., Shkura, K., Mazzuferi, M., Danis, B., van Eyll, J., et al.: Genome-wide analysis of differential rna editing in epilepsy. Genome research **27**(3), 440–450 (2017)
60. Stamberger, H., Nikanorova, M., Willemsen, M.H., Accorsi, P., Angriman, M., Baier, H., Benkel-Herrenbrueck, I., Benoit, V., Budetta, M., Caliebe, A., et al.: Stxbp1 encephalopathy: a neurodevelopmental disorder including epilepsy. Neurology **86**(10), 954–962 (2016)
61. Stoll, G., Caron, B., Viara, E., Dugourd, A., Zinovyev, A., Naldi, A., Kroemer, G., Barillot, E., Calzone, L.: Maboss 2.0: an environment for stochastic boolean modeling. Bioinformatics **33**(14), 2226–2228 (2017)
62. Subramanian, A., Narayan, R., Corsello, S.M., Peck, D.D., Natoli, T.E., Lu, X., Gould, J., Davis, J.F., Tubelli, A.A., Asiedu, J.K., et al.: A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. Cell **171**(6), 1437–1452 (2017)
63. Szklarczyk, D., Gable, A.L., Nastou, K.C., Lyon, D., Kirsch, R., Pyysalo, S., Doncheva, N.T., Legeay, M., Fang, T., Bork, P., et al.: The string database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic acids research **49**(D1), D605–D612 (2021)
64. Thomas, R.: Boolean formalization of genetic control circuits. Journal of theoretical biology **42**(3), 563–585 (1973)

65. Thomas, R., Thieffry, D., Kaufman, M.: Dynamical behaviour of biological regulatory networks—i. biological role of feedback loops and practical use of the concept of the loop-characteristic state. Bulletin of mathematical biology **57**(2), 247–276 (1995)
66. Von Mering, C., Jensen, L.J., Snel, B., Hooper, S.D., Krupp, M., Foglierini, M., Jouffre, N., Huynen, M.A., Bork, P.: String: known and predicted protein–protein associations, integrated and transferred across organisms. Nucleic acids research **33**(suppl_1), D433–D437 (2005)
67. Wang, X., Goldstein, D.B.: Enhancer domains predict gene pathogenicity and inform gene discovery in complex disease. The American Journal of Human Genetics **106**(2), 215–233 (2020)
68. Wery, M., Dameron, O., Nicolas, J., Remy, E., Siegel, A.: Formalizing and enriching phenotype signatures using boolean networks. Journal of Theoretical Biology **467**, 66–79 (2019)
69. (WHO), W.H.O.: Epilepsy. https://www.who.int/news-room/fact-sheets/detail/epilepsy (2022), accessed: [April 29, 2022]
70. Wu, L., Li, M., Wang, J., Wu, F.X.: Cytoctrlanalyser: a cytoscape app for biomolecular network controllability analysis. Bioinformatics **34**(8), 1428–1430 (2018)
71. Zerrouk, N., Miagoux, Q., Dispot, A., Elati, M., Niarakis, A.: Identification of putative master regulators in rheumatoid arthritis synovial fibroblasts using gene expression data and network inference. Scientific reports **10**(1), 1–13 (2020)
72. Ziegler, A., Colin, E., Goudenège, D., Bonneau, D.: A snapshot of some pli score pitfalls. Human mutation **40**(7), 839–841 (2019)

A Building the boolean network

This section of the Appendix further describes the procedure used to infer a Boolean network in our application to epilepsy.

A.1 Step (A) : Building a undirected unsigned graph

This step builds an undirected, unsigned network of putative gene-to-gene regulations. The gene module M30, as defined by [18], comprises of 320 genes which global expression anticorrelates with epileptic phenotypes. We retrieved all 320 genes of the M30 module from the additional file 1 in [18]. Undirected and unsigned protein-pairwise interactions involving two proteins encoded by M30 genes were extracted from the STRING database [63] for the human (NCBI taxon ID 9606). In order to perform the inference, it is necessary for computational reasons to restrict the set of edges to consider ; however, (weak) connectivity in the graph of interactions should also be preserved to fully exploit the dynamical constraints provided later on.

Considering the full network retrieved from the STRING database, we trimmed out isolated genes –*i.e.*, without any interactions with any gene, not even themselves. 318 genes out of 320 were retained after this procedure, with a total of $14,662$ edges. The STRING database also provides scores associated with each undirected edge named "combined scores", which are comprised between 0 and

$1,000$, and aggregate various scores related to the type of evidence supporting these edges [66]. This higher this score is, the more strongly supported the associated edge is. Provided a user-provided threshold η on this combined score,

- First, we built a protein-protein interaction (PPI) network by preserving all edges with a combined score greater than η ;

- Then, considering all edges which contain at least one gene that do not appear in the set of edges at the previous step, we sorted them in the order of decreasing combined score, and added them sequentially (adding simultaneously edges with the same score) to the network, until the number of weakly connected components is 1.

We tested (weak) connectivity by performing a Depth-First Search [16], which is a well-known procedure that explores all the nodes in a graph by favoring the exploration of child nodes instead of sibling nodes, until all nodes have been visited.

In order to select the threshold $\eta = 400$, we performed a gridsearch on $[\![100; 1,000]\!]$ with a step of 5, and selected the value which minimized the number of edges. This step is automatically performed the first time the inference procedure in the repository [7] is run, such that the user can use the threshold value η recommended by the grisearch. Choosing $\eta = 400$ allowed reducing the number of undirected edges from $14,662$ to $1,633$.

A.2 Step (A) : Gene perturbation experiments

At this step, we restricted the set of genes –subsequently, of interactions– to genes present in the database of transcriptional profiles LINCS L1000 [62]. To do so, we converted all gene identifiers in M30 into EntrezGene IDs using BioDBnet [46]. Then, we filtered out genes for which the EntrezGene ID was not present in LINCS L1000. After this step, 236 genes, out of 318, were retained. We selected experiments present in LINCS L1000 such that at least one gene from M30 has been perturbed in a genetic experiment (knockdown or overexpression, along with control samples) on a brain cell line. Unfortunately, there are no hippocampal neuron human (HN-h) cell lines in LINCS L1000; These cells are able to differenciate into neurons and glial cells as shown in the rat [22], which is why we assumed that the neural progenitor cell (NPC) line present in LINCS L1000 might be appropriate. Furthermore, among the genetic perturbation experiments listed in the database, we selected those which satisfy all following conditions

- which are associated to the highest metric `distil_ss` (as provided by the LINCS L1000) which is correlated with the number of significantly differentially expressed transcripts found in the differential analysis between the matching genetically treated and the control groups. In practice, this measure is correlated with the reproducibility of a drug signature [40].

- where there is at least two replicates from the same plate for the perturbed and control (of type `ctl_vector`) conditions.

[7] `https://github.com/clreda/PrioritizationMasterRegulators`

- which interference scale (computed as described in [13]) is positive. This ensures that the associated genetic perturbation experiment was successful. That means that a gene which has been perturbed by a knockdown (resp., an overexpression) has an expression lower (resp., greater) in treated profiles than in controls, compared to an appropriate housekeeping gene. The expression levels in these housekeeping genes should not dramatically change in both groups of profiles.

- where the associated experiment is either using shRNA (knockdown perturbation), cDNA, also known as knock-in (overexpression perturbation), or CRISPR (knockout perturbation).

- where the associated cell line is either SHSY5Y (neuroblastoma) or NPC (neural progenitor cells), which are the only brain cell lines in LINCS L1000.

The result of this step is a matrix of M30 genes by experimental profiles, which contains Level 3 LINCS L1000 data (normalized expression data for the whole genome) for each perturbation experiment. See Table 4 in Appendix for the list of experimental profiles retained in the application to epilepsy.

A.3 Step (B) : Binarization of experiments into binary profiles

Although there are known methods for the binarization of (single) RNA-seq data [23, 5], probably due to the fact Level 3 LINCS L1000 data is a combination of measured and inferred expression data, for different platforms (RNA-sequencing data for the most recent version, microarray for the first generated profiles), there were issues with the model fitting ; only a few genes were assigned a binary value 0 or 1 –the alternative being that they are not considered expressed "enough", according to the thresholds computed by these methods, to be assigned a state equal to 1, nor too weakly expressed to be assigned a state equal to 0). A data-driven method to tune the granularity of the binarization, adaptive to the selected perturbation expression data, was necessary in order to explicitly enforce a trade-off between a full reliance on undirected edges provided by the STRING database, and on experimental profiles from LINCS L1000.

Binarization. We designed an *ad hoc* binarization method to satisfy these constraints. This binarization was independently performed on each cell line. Gene expression data (as normalized RNA counts) was first quantile-normalized and clipped to the interval $[0, 1]$. Control samples (for the same cell line) were aggregated by considering the genewise median expression value. Given the threshold ζ, all genes with expression greater than $1 - \zeta$ were considered greatly expressed (with assigned state 1), whereas genes with expression lower than ζ were considered non-expressed (with assigned state 0). Genes which expression levels were in the interval $[\zeta, 1 - \zeta]$ had an undetermined expression state. Note that the quantile normalization is necessary, even though the initial expression data was normalized, in order to apply a same threshold ζ on all profiles. The higher ζ is, the more constrained the experiments are, as more genes have a determined expression state 0 or 1. Lower thresholds mean less constrained experiments, and a higher preference for the regulatory interactions filtered from the STRING database over expression data from LINCS L1000.

Using a bisection method in interval $[0; 0.5]$ with precision 0.0005, we identified $\zeta = 0.265$ as the maximum threshold such that the inference of Boolean networks satisfying these experimental constraints admits at least one solution. We recommend using this bisection method to determine the threshold ζ when using the pipeline with another dataset.

Background expression data. However, this method relies on having enough data to compute reliable statistics of expression for each gene, which is why, for each cell line, we automatically retrieved from LINCS L1000 a "background" expression matrix, which we concatenated to the set of profiles before binarization. After binarization, we removed samples associated with the background dataset. In order to collect the background expression matrix, we selected all experiments in the considered cell line, with type `pert_sh` (knockdown experiments), and we filtered out experiments with less than two replicates, with metric `distil_cc_q75` greater or equal to 0.2, and with metric `pct_self_rank_q25` lower than 0.05. Metrics `distil_cc_q75` and `pct_self_rank_q25` are two measures associated with experimental profiles which quantify the reproducibility based on the correlation between the same technical replicates (`distil_cc_q75`) and the diversity of profiles for a given experimental setting (`pct_self_rank_q25`). These rules correspond to the requirements for reproducible and distinct (so-called "gold") profiles according to LINCS L1000 documentation. Finally, we selected the same-plate replicates with the highest value of `distil_ss`. [8]

A.4 Step (B) : Implementation of topological constraints

The inference of a Boolean network relies on a set of admissible interactions and a set of time-series expression constraints. Indeed, solution networks only comprise of a subset of these admissible interactions, such that all constraints provided by the observations are satisfied.

In order to build the set of admissible interactions, we considered the PPI network extracted from the STRING database. Since these interactions are unsigned, we decided to reduce the number of possible interactions –and thus, the computational cost of the method– by using the gene perturbation expression matrix retrieved from LINCS L1000 (Table 4)

- First, a Pearson's r [9] gene correlation matrix was computed from these profiles, and raised to the power of β coefficient-wise, which allowed signing the interactions using pairwise correlation signs.

- Then, to preserve connectivity, we built the filtered signed undirected network similarly to what we previously did, using a threshold on the correlation values equal to $\tau = 0.4$.

β was chosen as it is known that raising an adjacency matrix A to the power of β yields coefficients $A[i, j]$ in position (i,j) equal to the number of paths (with eventually repeated edges) between node i and node j of length β. τ was chosen as a compromise between richness of the network (number of edges) and computational cost, by a bisection search in interval $[0.01; 1]$ with step 0.005,

[8] These measures are further described at `https://clue.io/connectopedia/glossary`.

which would be the way to go to apply our method to other datasets. After this procedure, we removed isolated genes in the network (that is, with both ingoing-degree and outgoing-degree equal to 0). After this step, 232 genes were left in the network, with 637×2 putative genepairwise interactions (one for each direction between two genes). We stress on the fact that preserving connectivity will be crucial for properly exploiting the experimental data, which is why we trim out isolated genes.

A.5 Step (B) : Implementation of experimental constraints

After building the set of admissible interactions, we turned to building dynamical constraints, that is, the binary expression states of genes in the network according to experimental profiles from LINCS L1000.

The experiments shown in Table 4 comprise of control and perturbed profiles in single gene perturbation experiments –by knockdown through shRNA in our application to epilepsy. First, these profiles were binarized using the binarization procedure described above. Then, for each knockdown experiment, we considered as initial condition the profile obtained from control samples, and as final condition the one obtained from perturbed samples, which is set as a (steady) attractor states.

In order to implement the new dynamics in [51], we used the Python package BoNeSiS [14], which infers by answer-set programming Boolean networks –both the set of regulatory interactions and regulatory functions– that satisfy the experimental constraints with a subset of admissible interactions. We use the procedure in BoNeSiS which randomizes the search for network solutions. Moreover, in order to avoid trivial solutions without interactions, we also implemented the constraint that the state where all genes were not expressed (*i.e.*, with expression state 0) cannot lead to any of the reported final attractor states. This constraint can be challenged, as one might assume that a network could end up in the state where all genes are turned off in a transient way, if there are some genes which are only regulated by inhibitors. However, in practice, the inference procedure without this constaint yields singularly trivial and poorly connected solutions (*i.e.*, most genes ending up without any regulators). We conjecture that it is linked to the procedure of answer-set programming, as similar methods, for instance Re:In [21], give the option of adding supplementary constraints about the presence of an activator for some genes.

A.6 Step (C) : Inference solutions and model selection

Inference of Boolean network solutions. In BoNeSiS, we asked for the enumeration of at most 1 solution to the set of topological and experimental constraints defined above. In the implementation of the most permissive semantics in [51] by [14], the size of the Boolean function specification can be upper-bounded by a prespecified value. In my application, I have used the maximum total (*i.e.*, ingoing and outgoing) degree of the underlying network, in order to

avoid spurious gene regulatory functions. Due to the intrinsic randomness stemming from the solver clingo [24], and the randomized search procedure used in BoNeSiS, we iterated this enumeration, such that we obtained 25 Boolean network solutions (among which 25 are unique in terms of regulatory functions).

Selection of an optimal model. In order to select a "representative" network consistent with what is known about the topology of biological networks, [4] compiled a list of network measures to maximize in biological networks, and computed a single scalar criterion value comprised in the interval $[0, 1]$ to maximize through the Harrington desirability index [29]. This value was called "general topological parameter". In practice, using the notation from [4], we considered the following weights

$$a_{\texttt{DS}} = 3, a_{\texttt{CL}} = 3, a_{\texttt{Centr}} = 3, \text{ and } a_{\texttt{GT}} = 1 \,,$$

where

- `DS` corresponds to the *network density*, that is, the ratio of the number of edges to the maximum number of possible connections between the nodes in the network (that is, if the network was fully connected) ; for a network of n nodes, this maximum number is equal to $(n-1)n/2$.

- `CL` corresponds to the *network clustering coefficient* which is the average of node-wise clustering coefficients. The clustering coefficient of a node is the ratio of the degree of the considered node and the maximum possible number of connections such that this node and its current neighbors form a clique (*i.e.*, form a fully connected graph).

- `Centr` corresponds to the *network centralization*, which is correlated with the similarity of the network to a graph with a star topology.

- `GT` corresponds to the *network heterogeneity*, which quantifies the nonuniformity of the node degrees across the network by computing the ratio between the standard deviation of the node degrees and the average degree across the network.

The higher the weights, the more importance is given to having a large associated coefficient. Finally, for every network solution N returned by BoNeSiS, we computed

$$\exp\left(\texttt{mean}\left\{-\exp\left(x \times a - 1\right) : (x, a) \in \mathcal{V}(N)\right\}\right) \,,$$

where $\mathcal{V}(N)$ is the set of pairs (value, weight) associated with each topological measure

$$\mathcal{V}(N) := \{(\texttt{DS}(N), a_{\texttt{DS}}), (\texttt{CL}(N), a_{\texttt{CL}}), (\texttt{Centr}(N), a_{\texttt{Centr}}), (\texttt{GT}(N), a_{\texttt{GT}})\} \,.$$

The final network was the one which maximized this quantity.

A.7 Robustness on a larger set of 50 solutions

Since the enumeration of solutions is still computationally expensive and time-consuming, we focused on a collection of 25 network solutions. However, in order

to assess the robustness of our inference procedure, we enumerated an additional set of 25 solutions, and reproduced the two plots shown in the main paper. Note that these 25 solutions were different from the first 25 ones, yielding a set of 50 unique solutions (in terms of gene regulatory functions). The selection of the optimal model run on these 50 models returned the same network shown in the main paper. Table 1 and Figure 3 allow us to conclude similarly to the main paper, that is, the networks obtained just before the network selection step are mostly functionally and topologically similar.

Table 3. Distribution statistics on the number of *unique* regulatory functions (RFs) across solutions per gene, and on the value of the general topological parameter (GTP) used for network selection in step (C) of the inference procedure. All values are rounded up to the 3$^{\text{rd}}$ decimal place. Applied on the set of 50 solutions.

	Min.	25$^{\text{th}}$ quantile	Median	Mean	75$^{\text{th}}$ quantile	Max.
# RFs	1	1	2	2.635	3	17
GTP	0.794	0.796	0.797	0.797	0.799	0.802

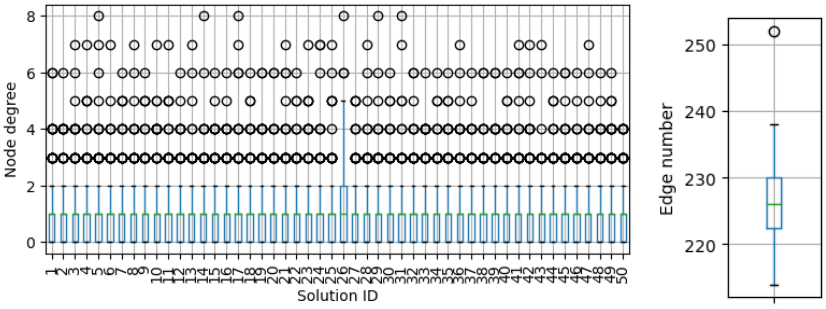

Fig. 7. *Left-hand plot*: Boxplots of node total degrees (ingoing and outgoing degree) per solution. The green lines represent median values. *Right-hand plot*: Boxplot of the number of edges across solutions (each solution comprise of 232 nodes). The green line represent the median value. Applied on the set of 50 solutions. Note that the first 25 boxplots match the plot in Figure 3.

B Tables

This section of the Appendix shows supplementary tabular data about the inference of the Boolean network.

B.1 Experimental profiles from LINCS L1000

Table 4. Experimental profiles retrieved from LINCS L1000 for the application to epilepsy, as annotated in LINCS L1000. * KD stands for knockdown. ** Time (in hours) of exposure to the perturbagen. # number of replicates

Profile brew identifier (suffix)	Cell line	Gene	Type	Time**	Dose	Nb#
KDB003_NPC_96H	NPC	PSMG1	KD	96	$1.5\mu L$	4
Samples {X1,2.A2,X2,X3.A2}						
Treated B6_DUO52HI53LO:K16						
Control B6_DUO52HI53LO:F13						
EKW001_SHSY5Y_120H	SHSY-5Y	SOD1	KD*	120	N/A	3
Samples {X1,X2,X3}						
Treated F1B3_DUO52HI53LO:J20						
Control F1B3_DUO52HI53LO:I05						
Treated F1B3_DUO52HI53LO:H17		SYT1	KD	120	N/A	3
Treated F1B3_DUO52HI53LO:I19		CACNA1C	KD	120	N/A	3
Treated F1B3_DUO52HI53LO:A03		CDC42	KD	120	N/A	3

B.2 Parameters

Table 5. Parameter values for the synthesis of Boolean networks (application to epilepsy).

	Definition	Value
η	Threshold for selecting edges from STRING [63]	400
ζ	Threshold for the binarization step	0.265
β	Power applied to the matrix of genepairwise correlations	1
τ	Threshold for filtering out edges in the putative network	0.4

Table 6. Threshold values (lower bounds on scores) for the retrieval of disease-associated genes from DisGeNet [54]. The full definitions of these indices are reported at this page. [19] EI : Evidence Index. DSI : Disease Specificity Index. DPI : Disease Pleiotropy Index.

	Score	EI	DSI	DPI	Source
Value	0	0	0.25	0	CURATED

C Implementation of the influence maximization algorithm

This section deals with supplementary data about the implementation of the influence maximization procedure.

C.1 Iteration of attractor states

In order to enumerate attractors under perturbations, we used PyMaBoSS [61]. We ran PyMaBoSS with $1,000$ trajectories, for reachable attractors within 50 time steps, and parameters `time_tick` $= 1$, `use_physrandgen` $= 0$. Unfortunately, this method does not guarantee the similarity of attractors from one iteration to another, but our tests showed that, although there is some noticeable change in the resulting spread values, it does not affect the final ranking on genes. We never had to deal with the case where no attractor state is retrieved with these parameter values.

C.2 Choice of initial states

In our application, we considered the integration of a disease-specific context by considering 48 hippocampi normalized transcriptional profiles of humans affected with Temporal Lobe Epilepsy (TLE) [44] (EMTAB 3123 on ArrayExpress). The main idea is that we specifically target genes which regulatory influence is high in a transcriptional context for epilepsy. We restricted these epileptic profiles to genes present in the network, and binarized the profiles according to the binarization procedure described in the first section, with corresponding threshold ζ equal to 0.5, so that all genes have a determined binary expression state.

D Additional results

This section shows additional results related to the inference of the Boolean network and the spread values.

Table 7. Regulatory interactions present in all of the 25 solutions. * strongest evidence source from the STRING database. "Association in databases" means associated in curated pathway databases.

Regulator	Regulated	Sign	Evidence source*
RBFOX1	PEG3	Inhibitory	Coexpression
SLITRK3	IL1RAP	Inhibitory	Text-mining
TSPAN7	AFF2	Activatory	Coexpression
UQCRQ	TIMM17A	Inhibitory	Coexpression
CENPJ	ANAPC1	Activatory	Coexpression
SYT13	MLLT11	Inhibitory	Coexpression
GUCY1B3	AHNAK2	Activatory	Text-mining
PLEKHG3	CCDC68	Inhibitory	Text-mining
MLLT11	TTC3	Activatory	Text-mining
SULT4A1	KIAA0232	Inhibitory	Text-mining
GRIN1	CIT	Activatory	Interaction
GPI	ANXA6	Inhibitory	Text-mining
RBFOX1	ZMAT4	Activatory	Coexpression
FAR2	FAM49A	Inhibitory	Text-mining
RAB3A	REEP2	Activatory	Coexpression
CAMK2B	AGGF1	Inhibitory	Association in databases
GAP43	ELAVL2	Inhibitory	Coexpression
ADAM22	EPB41L3	Inhibitory	Coexpression
CDC42	ARHGAP44	Activatory	Interaction
GNB5	PAK1	Inhibitory	Association in databases
STMN2	PSMG1	Inhibitory	Coexpression
AMPH	AAGAB	Activatory	Interaction
GNB5	PRKCE	Inhibitory	Association in databases
ATP1B3	FXYD7	Inhibitory	Association in databases
ATP1A3	PANK2	Activatory	Text-mining

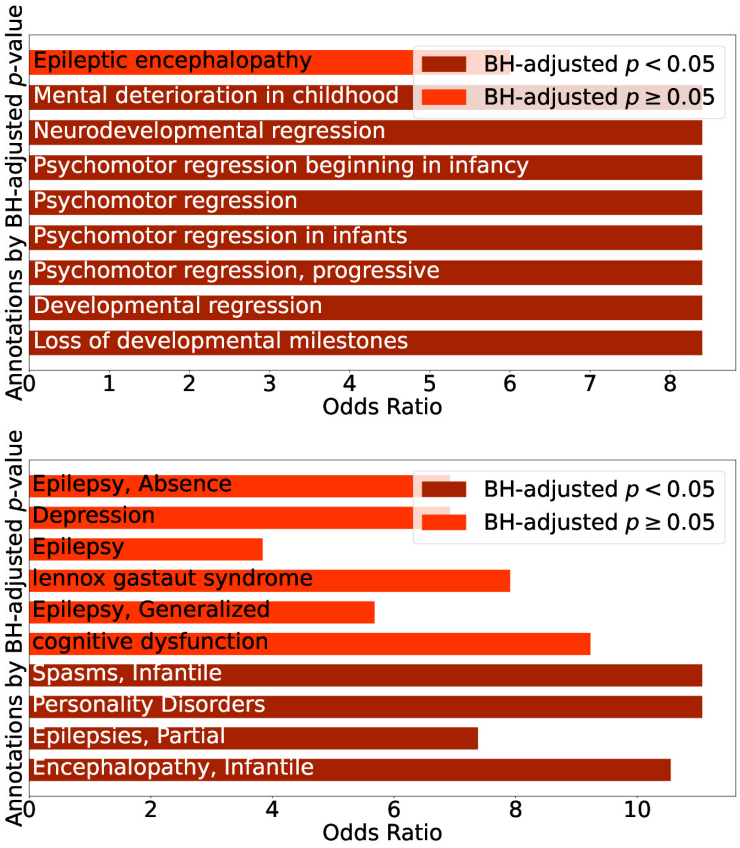

Fig. 8. Enrichment results from the ORA analysis on the filtered list of genes based on spread values from the DisGeNet annotations [54] (*top*) and GLAD4U [32] (*bottom*). The top-10 annotations (in increasing order of BH-adjusted p-value) are reported. All of these adjusted p-values are lower than 20%.