

Elucidating and leveraging design principles towards realistic Boolean models of gene regulatory networks

By

Subbaroyan Ajay

LIFE10201904001

The Institute of Mathematical Sciences
Chennai

*A thesis submitted to the
Board of Studies in Life Sciences
In partial fulfillment of requirements
for the Degree of*

DOCTOR OF PHILOSOPHY

of

HOMI BHABHA NATIONAL INSTITUTE



July 2024

Homi Bhabha National Institute

Recommendations of the Viva Voce Committee

As members of the Viva Voce Committee, we certify that we have read the dissertation prepared by **Subbaroyan Ajay** entitled: "Elucidating and leveraging design principles towards realistic Boolean models of gene regulatory networks" and recommend that it may be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy.



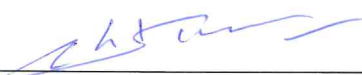
Chair - Prof. Sitabhra Sinha

Date: 3/7/24



Supervisor/Convener - Prof. Areejit Samal

Date: 3/7/24



Member 1 - Prof. Rahul Siddharthan

Date: 3/7/24



Member 2 - Prof. Satyavani Vemparala

Date: 3/7/24



Member 3 - Prof. Manikandan Narayanan

Date: 3/7/24



External Examiner - Prof. Amit Ghosh

Date: 3/7/24

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to HBNI. I hereby certify that I have read this dissertation prepared under my direction and recommend that it may be accepted as fulfilling the dissertation requirement.

Date: 3/7/24

Place: CHENNAI


Supervisor

Statement by Author

This dissertation has been submitted in partial fulfillment of requirements for an advanced degree at Homi Bhabha National Institute (HBNI) and is deposited in the Library to be made available to borrowers under rules of the HBNI.

Brief quotations from this dissertation are allowable without special permission, provided that accurate acknowledgement of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the Competent Authority of HBNI when in his or her judgement the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.

A handwritten signature in black ink, appearing to read 'Ajay Subbaroyan', with a stylized, flowing script.

Subbaroyan Ajay

Declaration

I, hereby declare that the investigation presented in this thesis has been carried out by me. The work is original and has not been submitted earlier as a whole or in part for a degree or diploma at this or any other Institution or University.

A handwritten signature in black ink, appearing to read 'Ajay Subbaroyan', written in a cursive style.

Subbaroyan Ajay

List of Publications included in the thesis

Journals

Published

1. *Leveraging developmental landscapes for model selection in Boolean gene regulatory networks*,
A. Subbaroyan, P. Sil, O.C. Martin^{*} and A. Samal^{*},
Briefings in Bioinformatics, 24(3):bbad160 (2023).
<https://doi.org/10.1093/bib/bbad160>
2. *Relative importance of composition structures and biologically meaningful logics in bipartite Boolean models of gene regulation*,
Y. Yadav[†], **A. Subbaroyan**[†], O.C. Martin^{*} and A. Samal^{*},
Scientific Reports, 12(1):18156 (2022).
<https://doi.org/10.1038/s41598-022-22654-7>
3. *Minimum complexity drives regulatory logic in Boolean models of living systems*,
A. Subbaroyan, O.C. Martin^{*} and A. Samal^{*},
PNAS Nexus, 1(1):pgac017 (2022).
<https://doi.org/10.1093/pnasnexus/pgac017>
4. *A preference for Link Operator Functions can drive Boolean biological networks towards critical dynamics*,
A. Subbaroyan, O.C. Martin^{*} and A. Samal^{*},
Journal of Biosciences, 47:17 (2022).
<https://doi.org/10.1007/s12038-022-00256-9>

[[†] Joint-first authors; ^{*} Corresponding author(s)]



List of Publications not included in the thesis

Journals

Published

1. *Biologically meaningful regulatory logic enhances the convergence rate in Boolean networks and bushiness of their state transition graph*,
P. Sil[†], **A. Subbaroyan**[†], S. Kulkarni, O.C. Martin^{*} and A. Samal^{*},
Briefings in Bioinformatics, 25(3):bbae150 (2024).
<https://doi.org/10.1093/bib/bbae150>
2. *Preponderance of generalized chain functions in reconstructed Boolean models of biological networks*,
S. Mitra[†], P. Sil[†], **A. Subbaroyan**[†], O.C. Martin^{*} and A. Samal^{*},
Scientific Reports, 14(1):6734 (2024).
<https://doi.org/10.1038/s41598-024-57086-y>
3. *ViCEKb: Vitiligo-linked Chemical Exposome Knowledgebase*,
N. Chivukula, K. Ramesh, **A. Subbaroyan**, A. K. Sahoo, G. B. Dhanakoti,
J. Ravichandran, A. Samal^{*},
Science of The Total Environment, 913:169711 (2023).
<https://doi.org/10.1016/j.scitotenv.2023.169711>

[[†] Joint-first authors; ^{*} Corresponding author(s)]



Oral or Poster presentations

1. Poster presentation titled *Leveraging developmental landscapes for model selection in Boolean gene regulatory networks* at Theoretical Approaches in Cancer Progression and Treatment, held in International Center for Theoretical Sciences (ICTS), Bengaluru from March 11-22, 2024.
2. Poster presentation titled *Leveraging developmental landscapes for model selection in Boolean gene regulatory networks* at Contemporary Perspectives in Computational Biology, held in The Institute of Mathematical Sciences (IMSc), Chennai from February 19-20, 2024.
3. Poster presentation titled *Leveraging developmental landscapes for model selection in Boolean gene regulatory networks* at Modelling and Tackling Complex Biological Systems, held in The Institute of Mathematical Sciences (IMSc), Chennai from October 13-14, 2023.
4. Poster presentation titled *Leveraging developmental landscapes for model selection in Boolean gene regulatory networks* at Perspectives in Nonlinear Dynamics (PNLD), held in Indian Institute of Technology Madras, Chennai from August 1-4, 2023.
5. Oral presentation titled *Elucidating and leveraging design principles towards realistic reconstruction of Boolean models of gene regulatory networks* at Network Biology Day, held in The Institute of Mathematical Sciences (IMSc), Chennai on July 20, 2023.
6. Poster presentation titled *Leveraging developmental landscapes for model selection in Boolean gene regulatory networks* at Physics of Cells and Tissues Symposium, held in Indian Institute of Science (IISc), Bengaluru from February 15-16, 2023.
7. Oral presentation titled *Simple regulatory logic appears to drive complex cellular decisions* at IMSc60 celebration, held in The Institute of Mathematical Sciences (IMSc), Chennai from January 2-5, 2023.

8. Poster presentation titled *Minimum complexity drives regulatory logic in Boolean models of living systems* at IMSc60 celebration, held in The Institute of Mathematical Sciences (IMSc), Chennai from January 2-5, 2023.
9. Online Oral presentation titled *Minimum complexity drives regulatory logic in Boolean models of living systems* at ICTP-ICTS Winter school on Quantitative Systems Biology, held in ICTP, Trieste, Italy from December 5-16, 2022.
10. Poster presentation titled *Minimum complexity drives regulatory logic in Boolean models of living systems* at Cellular Lineages and Development: from single cells to landscapes, held in Alappuzha, Kerala from November 1-4, 2022.
11. Online Oral presentation titled *Minimum complexity drives regulatory logic in Boolean models of living systems* at Sainsbury Laboratory Symposium held in Sainsbury Laboratory, Cambridge, UK from September 21-23, 2022.
12. Online Poster presentation titled *Minimum complexity drives regulatory logic in Boolean models of living systems* at Sainsbury Laboratory Symposium held in Sainsbury Laboratory, Cambridge, UK from September 21-23 2022.
13. Poster presentation titled *Minimum complexity drives regulatory logic in Boolean models of living systems* at HBNI Theme Meeting on Life Sciences held in RRCAT, Indore from September 7-10, 2022.

Research visits and seminars

1. Seminar titled *Living beings favour odd and simple regulatory logic*, held at the Indian Institute of Technology Madras, Chennai on October 12, 2022.

A handwritten signature in black ink, featuring a stylized 'S' and 'A' that are interconnected, with the name 'Subbaroyan' written in a cursive script below them.

Subbaroyan Ajay

This thesis is dedicated to
my sister and my parents

Acknowledgements

At the outset, I am fortunate to have been supervised by Prof. Areejit Samal. I have learnt much from Prof. Samal which I cannot express here. Prof. Samal's relentless dedication and hardwork, his scientific courage to explore new ideas and moral courage to face the realities of the world with an impartial eye have inspired me everyday. He has shown me that I can be much more than what I am. I am also grateful to Prof. Samal for the scientific exposure he has given me, in particular the opportunity to work with Prof. Olivier C. Martin. I am deeply indebted to Prof. Martin for the scientific mentorship, time and moral support I have received from him over the years.

I would like to thank present or former lab members Ajaya, Priyotosh and Yasharth. Ajaya has been with me during my darkest phases in IMSc and has given me a shoulder to lean on whenever I needed it the most. Thank you for everything Ajaya. My heartfelt gratitude to Priyotosh, with whom I have thoroughly enjoyed collaborating. I express my heartfelt thanks to Yasharth for his kindness and understanding.

I would also like to thank other present and former members that have been associated with the lab: Akshaya, Gaurav, Geetha, Gokul, Janani, Kavya, Kishan, Kundhanathan, Madhumita, Naman, Nikhil, Pavithra, Priyashree, Rishabh, Saumitra, Shanmuga Priya, Shashwat, Subham, Suchetana, Vinayak and Vivek. My special thanks to my collaborators Kundhanathan, Saumitra and Suchetana with whom I have enjoyed working. Special thanks to Nikhil and Shanmuga Priya for their fun and humor. My sincere thanks to Janani for moral support in difficult times.

I would like to thank my doctoral committee members Prof. Manikandan Narayanan, Prof. Rahul Siddharthan, Prof. Satyavani Vemparala, Prof. Sitabhra Sinha for their encouragement and suggestions on my work. I would like to thank

Prof. Manikandan Narayanan for giving me the opportunity to speak at IIT Madras, and Prof. Karthik Raman for providing me with valuable feedback in that talk. I would like to thank Prof. Sitabhra Sinha for his stimulating courses from which I benefited immensely.

I was fortunate to have presented and discussed my work to several faculty who visited IMSc and would like to thank them for their invaluable comments and suggestions: Prof. Anshu Bhardwaj, Prof. Dhiraj Kumar, Prof. Kamachi Mudali, Prof. Mitali Mukherjee, Prof. Nirav Bhatt, Dr. Nitin Saurabh, Prof. Prasun Mukherjee, Prof. Ramakrishna Ramaswamy, Prof. Sagar Pandit and Prof. Vinay Nandicoori.

I would like to thank several colleagues in IMSc for their company or scientific discussions at some juncture during the course of the PhD: Abhijit, Dhananjay, Gaurav, Garima, Iyyappan, Manav, Neelam, Ramit, Rashi, Rakesh, Sanjay, Sathish, Velmurugan. Special thanks to Garima for her guidance.

I thank IMSc and DAE for their funding and support during my PhD. I would like to thank all administrative staff members and computer committee members for doing the needful whenever necessary. My heartfelt thanks to everyone who works in the canteen, housekeeping, civil and electrical departments for making IMSc a wonderful place to work. A special thanks to security staff Mr. Raju Murthy for his kindness and care. I would also like to thank all the members of the IMSc football community.

I would like to thank faculty members with whom I have interacted with both online and offline in several conferences for their interest and feedback: Prof. David Kestler, Prof. Edwige Moyroud, Prof. Henrik Jönsson, Prof. Herbert Levine, Prof. Jacob Feldman, Prof. James Locke, Prof. Luciano Marcon, Prof. Matteo Rauzi, Prof. Mohit Kumar Jolly, Prof. Namrata Gundiah, Prof. Philipp Thomas, Prof. Sergiu Hart and Prof. Vidyanand Nanjundiah. Special thanks to Prof. Mohit Ku-

mar Jolly for inviting me to his conference in IISc and providing me the opportunity to engage in discussion with Prof. Herbert Levine and Prof. David Kestler. Special thanks to Prof. Sergiu Hart as well for his encouraging words and suggestion of a research problem.

I would like to thank colleagues whom I met in various conferences for engaging academic and non-academic discussions: Abhijeet, Abhigyan, Abhishek, Amitabh, Anton, Anubhav, Anushree, Aritra, Atchuta, Budhaditya, Darshan, Felix, Gilberto, Kishore, Mohammadreza, Prasanna, Sarthak, Sorique, Soutrick, Sucharita, Sunil, Supriya, Vinay and Yasir. Special thanks to Aritra and Kishore with whom I enjoyed several conversations.

I would also like to thank Prof. Gautam Menon and Prof. Sarika Jalan for mentoring me during internships at IMSc and IIT Indore respectively. I am indebted to my several of my teachers in Bachelors, Masters and beyond: Prof. Elankumaran, Mr. Nageswara Rao, Prof. Sudharsan and Prof. Sunil Kumar. In particular, I thank Mr. Nageswara Rao for teaching me the value of commitment and hard work. Thank you Minakshi ma'am, Pritesh sir and Pasha sir for instilling in me the love for biology, mathematics and physics respectively during my school days.

Lastly, I would like to thank my dear sister, Sneha, my father Subbaroyan Ramaswamy and my mother Padmavathi Subbaroyan for always being there and never losing hope in me. I would also like to thank my *athais* Balambal, Jayashree, Sarada and Suganthi, and my *athimbers* K. S. Krishnan, Lakshmanan and T. V. Ananthanarayanan for the love, care and affection they have showered on me in all these years. I would also like to thank all my cousins Anand, Archana, Arjun, Rahul, Gokul, Smitha, Srividya and Vidhya. I am also deeply grateful to my brother-in-law Aditya Subramanyam and his family for their love, kindness and wishes.

Subbaroyan Ajay

Contents

List of Figures	i
List of Tables	vii
Abstract	xi
1 Introduction	1
1.1 Boolean models of gene regulatory networks	3
1.2 Regulatory logic rules in Boolean biological networks	6
1.3 Relative stability as a model selection constraint	10
1.4 Thesis organization	14
2 Biologically meaningful Boolean functions: Description and properties	18
2.1 Boolean functions	19
2.1.1 Representations of BFs	19
2.1.2 Categorization of BFs based on their bias and isometries	22
2.2 Biologically meaningful types of BFs	24
2.2.1 Effective functions	24
2.2.2 Unate functions	25
2.2.3 Canalyzing functions	26
2.2.4 Nested canalyzing functions	27
2.2.5 Read-once functions	28
2.3 Characterizing the space of biologically meaningful BFs	30

2.4	Theoretical results on the properties of biologically meaningful types of BFs	36
2.4.1	Combining two independent BFs	37
2.4.2	Effective functions	38
2.4.3	Unate functions	39
2.4.4	Nested canalizing functions	40
2.4.5	Read-once functions	40
2.5	Discussion	46
3	Minimum complexity drives regulatory logic in Boolean models of living systems	48
3.1	Enrichments, relative enrichments and p -value tests	49
3.1.1	Enrichments and relative enrichments	49
3.1.2	Statistical significance tests	50
3.2	Enrichment of different types of BFs in reconstructed Boolean models of gene regulatory networks	51
3.2.1	Enrichment in types when comparing to the ensemble of random BFs	56
3.2.2	Relative enrichment in sub-types when comparing to the ensemble of random BFs	58
3.3	Complexity Measures	59
3.3.1	Minimal expressions and Boolean complexity	59
3.3.2	Average sensitivity of BFs	62
3.4	Enriched functions in biological data have minimum complexity . . .	64
3.4.1	Boolean complexity and average sensitivity are strongly correlated	64
3.4.2	NCFs have the minimum average sensitivity within their $k[P]$ set when P is odd	65

3.4.3	Good sets having an even number of vertices has Boolean complexity strictly less than k	69
3.5	Implications of using biologically meaningful BFs for Boolean network dynamics	72
3.5.1	Computing the distributions of network average sensitivities .	72
3.5.2	Estimating the overlap between the distributions of network average sensitivities for various types of BFs and the biological case	73
3.5.3	Ensembles generated with NCFs and RoFs have the maximum overlaps with biological case	74
3.6	Results from the repeat analyses after discarding the ineffective inputs to BFs in the reference biological dataset	75
3.7	Discussion	76
4	Leveraging developmental landscapes for model selection in Boolean models of gene regulatory networks	78
4.1	Relative stability and ordering of fixed points	79
4.1.1	Basin of Attraction	82
4.1.2	Steady State Probability	83
4.1.3	Mean First Passage Time	83
4.1.4	Basin Transition Rate	84
4.1.5	Stability Index	84
4.2	Constructing biologically plausible ensembles for model selection . . .	85
4.2.1	Two biological models and their ensembles of DGRNs	87
4.3	The five measures of relative stability are strongly correlated with each other	88
4.3.1	The relative stability measure for the BTR is identical to that of the BOA	89
4.4	Inferring cellular lineage trees using MFPT	93

4.4.1	Minimum Spanning Arborescence	93
4.4.2	Constructing a potential cellular lineage tree using the MFPT and MSA	94
4.4.3	Distribution of lineage trees computed using MFPT for vari- ous ensembles	95
4.5	Scaling up the computation of MFPT to reliably infer the relative stability of attractors in larger Boolean networks	96
4.5.1	Inferences drawn from MFPT are insensitive to changes in noise intensities	96
4.5.2	Stochastic approach to estimate the MFPT	98
4.5.3	Comparison of the stochastic approach to the exact method of computing MFPT	98
4.6	<i>Arabidopsis thaliana</i> root development: A case study	100
4.6.1	Relative orderings of the biological fixed points	102
4.6.2	A greedy algorithm for Boolean model selection	104
4.6.3	Greedy algorithm generates many models satisfying the ex- pected developmental landscape	107
4.7	Discussion	108
5	A preference for link operator functions can drive Boolean bio- logical networks towards critical dynamics	112
5.1	Link operator functions	113
5.2	Characterizing the space of LOFs	117
5.2.1	Relationship between the different types of LOFs	117
5.2.2	Cardinality of the different types of LOFs	119
5.3	Preponderance of LOFs in reconstructed Boolean models of gene reg- ulatory networks	120
5.4	LOFs as facilitators of Boolean model reconstruction and selection: Two case studies	123

5.5	Implications of using LOFs for Boolean network dynamics	125
5.6	Discussion	128
6	Relative importance of composition structures and biologically meaningful logics in bipartite Boolean models of gene regulation	130
6.1	Bipartite Boolean networks, composition structures and composed BFs	131
6.1.1	Bipartite Boolean networks	131
6.1.2	Composition structures	132
6.1.3	Composed BFs	135
6.2	Empirical evidence for the presence of composition structures	137
6.2.1	Quantifying the presence of protein complexes that can act as transcriptional regulators in Humans	137
6.2.2	Quantifying the presence of protein complexes that can act as transcriptional regulators in Yeast	138
6.2.3	TF binding regions and active enhancers	139
6.2.4	Composition structures arising through enhancers	142
6.3	Characterizing the space of composed BFs	143
6.3.1	Properties of composed BFs	143
6.3.2	Accounting for all the permutations of inputs of composed BFs	144
6.3.3	Overlap of composed BFs across various k -input composition structures	147
6.3.4	Comparing restriction levels: composition structures versus biologically meaningful types	147
6.4	Enrichments of composed BFs in reconstructed Boolean models of gene regulatory networks	149
6.5	Discussion	154
7	Summary and future outlook	157
7.1	Summary	157

7.2	Future outlook	161
A	Reference biological datasets of Boolean functions	164
A.1	Reference biological dataset compiling reconstructed discrete models of living systems used to quantify the preponderance of different biologically meaningful Boolean functions	164
A.2	Models in the Cell Collective database used for quantifying the preponderance of link operator functions	165
B	Procedure to generate random BFs belonging to various types of BFs	168
B.1	Randomized generation of biologically meaningful BFs used in Chapter 3	168
B.2	Randomized generation of biologically meaningful BFs used in Chapter 5	170
C	Additional Figures and Tables for Chapter 3	173
D	Additional Figures and Tables for Chapter 4	183
E	Additional Figures and Tables for Chapter 5	215
F	Additional Figures and Tables for Chapter 6	222
	References	242

List of Figures

1.1	A toy Boolean network and its associated state transition graph. . . .	4
1.2	Schematic overview of the broad objectives of the thesis and how the thesis addresses them.	11
2.1	Different representations of Boolean functions.	21
2.2	The number of biologically meaningful types of BFs for a given number of inputs $k \leq 5$	29
2.3	The fraction of biologically meaningful types of BFs among all BFs for a given number of inputs $k \leq 5$	29
2.4	Schematic figure depicting the four possible <i>sign</i> combinations for the $k = 3$ inputs to a node (gene) f	33
2.5	Frequency distribution of read-once functions (RoFs) across different bias P for functions with (a) $k = 4$, (b) $k = 5$, (c) $k = 6$, (d) $k = 7$, and (e) $k = 8$ inputs.	34
2.6	A schematic of the overlaps between different types of biologically meaningful BFs in the space of all 4-input BFs.	35
2.7	Flowchart describing our program to check whether a BF with $k \leq 10$ inputs entered by an user, is a read-once function (RoF).	45
3.1	Overlap of different types of BFs and their distribution in the reference biological dataset.	56
3.2	Dependence of the two complexity measures on the bias and associated 2D projections for all BFs with $k = 4$ inputs.	63

3.3	A Good set (GS) with P vertices where P is odd on a k -dimensional hypercube is equivalent to a NCF in $k[P]$ set ($k = 4, P = 13$).	67
3.4	A Good set (GS) with P vertices where P is odd on a k -dimensional hypercube is equivalent to a NCF in $k[P]$ set ($k = 4, P = 5$).	68
3.5	Good set (GS) for P vertices where P is even on a k -dimensional hypercube is equivalent to an IEF in that $k[P]$ set ($k = 4, P = 6$). . .	70
3.6	Distribution of the network average sensitivity when using the list of inputs from 88 biological models but enforcing different types of BFs to the nodes, namely effective functions (EF), effective and unate functions (EUF), canalyzing functions (CF), effective and canalyzing functions (ECF), nested canalyzing functions (NCF), read-once functions (RoF) and non-NCF RoFs.	73
4.1	Developmental landscape inspired by Waddington and relative stability measures.	80
4.2	Pearson correlation between pairs of relative stability measures in the ensemble $Root_{sc-NCF}$	88
4.3	Scatter plots displaying values of relative stability in the ensemble $Root_{sc-NCF}$	90
4.4	Pearson correlation between different pairs of relative stability measures for a given pair of fixed points for the ensemble $Root_{sc-NCF}$. . .	91
4.5	Scatter plots between the different pairs of relative stability measures for the pair of attractors 1 and 2 for the ensemble $Root_{sc-NCF}$	92
4.6	Pearson correlation between different pairs of relative stability measures for the ensemble $Root_{sc-EUF}$	93
4.7	Frequency distribution of the minimum spanning arborescences (MSAs) in the ensemble $Root_{sc-NCF}$	95
4.8	Pearson correlation between RS_{MFPT} values computed by exact methods for different pairs of noise values for the ensemble $Root_{sc-NCF}$. . .	96

4.9	Number and fraction of models which differ in at least one comparison of partial ordering of the different biological fixed points when considering two different noise values, in the ensemble $Root_{sc-NCF}$.	97
4.10	Correlation between the MFPT obtained via the exact method versus the proposed stochastic method using the ensemble $Root_{sc-NCF}$.	99
4.11	Barplot of the mean first passage time (MFPT) from one biological fixed point to another, computed via the stochastic method for a model taken from the ensemble $Root_{sc-NCF}$.	101
4.12	Hierarchy of fixed points based on basin sizes and minimum spanning arborescence (MSA) for the 2013, 2017 and 2020 root development models of <i>Arabidopsis thaliana</i> .	103
4.13	Workflow of our methodology for model selection, illustrated on the 2020 model of the Boolean GRN of <i>Arabidopsis thaliana</i> RSCN.	106
4.14	A flowchart of the workflow of the model selection procedure.	109
5.1	Schematic figure illustrating the various types of consistent LOFs.	114
5.2	The reduction in the size of the space of consistent LOFs in comparison to the space of all BFs with increasing number of inputs.	118
5.3	The fractions of the various types of consistent LOFs in the reference biological dataset.	121
5.4	Schematic figure of pancreas development and Epithelial-mesenchymal transition (EMT) GRNs along with their attractors	124
5.5	Network average sensitivity distribution of the various models in the reference biological dataset using various types of BFs	127
6.1	Boolean functions in unipartite versus bipartite network models of transcriptional gene regulation.	133
6.2	Non-trivial composition structures arising due to enhancers bound by multiple TFs.	140

6.3	Prevalence of active enhancers bound by multiple TFs in human cell lines.	141
6.4	Overlaps between the sets of BFs compatible with different composition structures at $k = 4$ and $k = 5$ inputs.	146
6.5	Abundance of composed BFs in reconstructed biological networks. . .	150
C.1	Overlap of different types of BFs and their distribution in the modified reference biological dataset.	174
C.2	Distribution of the network average sensitivity of 88 models after discarding the ineffective inputs.	175
D.1	Biological networks used to generate ensembles of Boolean models. . .	184
D.2	Pearson correlation between different pairs of relative stability measures for the ensembles $Root_{sc-NCF}^*$, $Panc_{sc-NCF}$ and $Panc_{sc-NCF}^*$. . .	185
D.3	Scatter plots displaying values of relative stability in the ensemble $Root_{sc-NCF}^*$	186
D.4	Scatter plots displaying values of relative stability in the ensemble $Panc_{sc-NCF}$	187
D.5	Scatter plots displaying values of relative stability in the ensemble $Panc_{sc-NCF}^*$	188
D.6	Pearson correlation between different pairs of relative stability measures for a given pair of fixed points for the ensemble $Root_{sc-NCF}^*$. . .	189
D.7	Pearson correlation between different pairs of relative stability measures for a given pair of fixed points for the ensemble $Panc_{sc-NCF}$. . .	190
D.8	Pearson correlation between different pairs of relative stability measures for a given pair of fixed points for the ensemble $Panc_{sc-NCF}^*$. . .	190
D.9	Scatter plots between the different pairs of relative stability measures for the pair of attractors 1 and 3 for the ensemble $Root_{sc-NCF}$	191

D.10	Scatter plots between the different pairs of relative stability measures for the pair of attractors 1 and 4 for the ensemble $Root_{sc-NCF}$	192
D.11	Scatter plots between the different pairs of relative stability measures for the pair of attractors 2 and 3 for the ensemble $Root_{sc-NCF}$	193
D.12	Scatter plots between the different pairs of relative stability measures for the pair of attractors 2 and 4 for the ensemble $Root_{sc-NCF}$	194
D.13	Scatter plots between the different pairs of relative stability measures for the pair of attractors 3 and 4 for the ensemble $Root_{sc-NCF}$	195
D.14	Pearson correlation between different pairs of relative stability mea- sures for the ensembles $Root_{sc-EUF}^*$, $Panc_{sc-EUF}$ and $Panc_{sc-EUF}^*$. .	196
D.15	Frequency distribution of the minimum spanning arborescences (MSAs) for the ensemble $Root_{sc-NCF}^*$	197
D.16	Frequency distribution of the minimum spanning arborescences (MSAs) for the ensemble $Panc_{sc-NCF}$	198
D.17	Frequency distribution of the minimum spanning arborescences (MSAs) for the ensemble $Panc_{sc-NCF}^*$	199
D.18	Pearson correlation between RS_{MFPT} values computed by ex- act methods for different pairs of noise values for the ensembles $Root_{sc-NCF}^*$, $Panc_{sc-NCF}$ and $Panc_{sc-NCF}^*$	199
D.19	Number and fraction of models which differ in at least one compar- ison of partial ordering of the different biological fixed points when considering two different noise values, in the ensemble $Root_{sc-NCF}^*$.	200
D.20	Number and fraction of models which differ in at least one compar- ison of partial ordering of the different biological fixed points when considering two different noise values, in the ensemble $Panc_{sc-NCF}$.	200
D.21	Number and fraction of models which differ in at least one compar- ison of partial ordering of the different biological fixed points when considering two different noise values, in the ensemble $Panc_{sc-NCF}^*$.	201

D.22	Correlation between the MFPT obtained via the exact method versus the proposed stochastic method using the ensemble $Root_{sc-NCF}^*$	202
D.23	Correlation between the MFPT obtained via the exact method versus the proposed stochastic method using the ensemble $Panc_{sc-NCF}$	203
D.24	Correlation between the MFPT obtained via the exact method versus the proposed stochastic method using the ensemble $Panc_{sc-NCF}^*$	204
D.25	Barplot of the mean first passage time (MFPT) from one biological fixed point to another, computed via stochastic methods for 3 models, each specific to one of 3 ensembles $Root_{sc-NCF}^*$, $Panc_{sc-NCF}$ and $Panc_{sc-NCF}^*$	205
D.26	2013 and 2017 GRNs of the <i>Arabidopsis thaliana</i> root development. .	206
D.27	2020 model of the <i>Arabidopsis thaliana</i> RSCN Boolean GRN and its fixed points with $AUX = 1$	207

List of Tables

2.1	The number of BFs belonging to the different types, at a given number of inputs $k \leq 5$	30
2.2	The fraction of BFs belonging to the different types, at a given number of inputs $k \leq 5$	31
2.3	Parity distribution (number) of biologically meaningful BFs for a given number of inputs k	32
2.4	Parity distribution (fraction) of biologically meaningful BFs for a given number of inputs k	33
2.5	The number of EUFs at a given number of inputs $k \leq 5$ for different combinations of activators and inhibitors.	36
2.6	The number and fraction of NCFs and RoFs among all BFs and fraction of NCFs within RoFs for a given number of inputs k	37
3.1	Number of different types of biologically meaningful BFs in the reference biological dataset.	52
3.2	Fraction of different types of biologically meaningful BFs in the reference biological dataset.	53
3.3	p -value tests for enrichments of the different types of BFs in the reference biological dataset.	54
3.4	Fractions of functions that are RoFs, non-NCF RoFs or NCFs, in the space of all 2^{2^k} BFs (f_0) and in the reference biological dataset (f_1). .	55
3.5	The relative enrichment ratios E_R for the RoFs and NCFs in the ensemble of odd bias BFs, EFs and UFs.	57

3.6	The relative enrichment ratio E_R of the NCFs in the CFs and RoFs.	58
3.7	Quantifying the fraction of models in different ensembles with network average sensitivities (s) lying outside the distribution of s for biological networks.	74
5.1	The different types of consistent LOFs.	117
5.2	Number of LOFs as a function of the number of activators (m), the number of inhibitors (n) and the total number of inputs (k)	119
5.3	The abundance of LOFs in the collection of BFs from reconstructed models of biological systems.	122
5.4	Model selection by using different types of BFs with and without the steady state constraints.	123
6.1	Comparison of the number and fraction of BFs allowed by different composition structures, with and without including all possible permutations of input variables.	144
6.2	Number of BFs in different composition structures that display biologically meaningful properties.	148
6.3	Relative enrichment of biologically meaningful BFs among composed BFs of different composition structures in the reference biological dataset.	151
6.4	Comparison between the enrichments of composed BFs and biologically meaningful BFs of minimum complexity in the reference biological dataset.	152
A.1	Various models from the Cell Collective database which have been considered for analysis in the study on LOFs.	166
C.1	Number of different types of biologically meaningful BFs in the modified reference biological dataset.	176

C.2	Fraction of different types of biologically meaningful BFs in the modified reference biological dataset.	177
C.3	p -value tests for statistical enrichments of the different types of BFs in the modified reference biological dataset.	178
C.4	Enrichment of the different types of BFs in the modified reference biological dataset.	179
C.5	Relative enrichment of the different types of BFs in the modified reference biological dataset.	180
C.6	Statistical test for relative enrichment of NCFs in CFs and RoFs in the modified reference biological dataset.	181
C.7	Quantifying the fraction of models in different ensembles with network average sensitivities (s) lying outside the distribution of s for biological networks (without ineffective inputs).	182
D.1	Correlation between the exact and stochastic methods to compute MFPT using the $Root_{sc-NCF}$ ensemble.	208
D.2	Correlation between the exact and stochastic methods to compute MFPT using the $Root_{sc-NCF}^*$ ensemble.	209
D.3	Correlation between the exact and stochastic methods to compute MFPT using the $Panc_{sc-NCF}$ ensemble.	210
D.4	Correlation between the exact and stochastic methods to compute MFPT using the $Panc_{sc-NCF}^*$ ensemble.	211
D.5	Boolean functions for the 2013 RSCN model in the BoolNet format. .	212
D.6	Boolean functions for the 2017 RAM model in the BoolNet format. .	213
D.7	Boolean functions for the 2020 RSCN model in the BoolNet format. .	214
E.1	Fraction of link operator functions (LOFs) among the complete space of Boolean functions (BFs) for a given number of inputs k	215

E.2	The abundance of link operator functions (LOFs) in the collection of Boolean functions (BFs) from reconstructed models of biological systems.	216
E.3	The number of Boolean functions (BFs) in the reference biological dataset for each input with only activators or only inhibitors, and those with at least one activator and one inhibitor.	217
E.4	The network average sensitivity of models where the numbers of activating and inhibiting inputs are as in the reconstructed biological networks but have particular types of Boolean functions namely, random effective functions (EFs), random effective and unate functions (EUFs), AND-NOT, OR-NOT, AND-pairs, OR-pairs.	218
E.5	Overlap of the network average sensitivity (s) distribution of various BFs with the outliers of the distribution of s of the biological models	221
F.1	A list of 169 complexes from the set of 1325 complexes in <i>H. sapiens</i> such that all the protein subunits of these complexes are transcription factors.	222
F.2	Classification of transcription factors in <i>H. sapiens</i>	231
F.3	A list of 17 complexes from the set of 617 complexes in <i>S. cerevisiae</i> such that all the protein subunits of these complexes are transcription factors.	234
F.4	Classification of transcription factors in <i>S. cerevisiae</i>	235
F.5	Comparison of the fractions of four types of biologically meaningful BFs and the fraction of the BFs allowed by the most restrictive composition structures for $k \leq 5$ inputs.	236
F.6	Fraction of BFs in different composition structures that display biologically meaningful properties.	236
F.7	Number and fraction of BFs with odd bias in different composition structures.	237

F.8	Enrichment of composed BFs in the reference biological dataset. . . .	237
F.9	p -values corresponding to the relative enrichment values of biologically meaningful BFs within different composition structures.	238
F.10	p -values for comparison between the enrichments of composed BFs and biologically meaningful BFs with minimum complexity in the reference biological dataset.	239
F.11	Comparison between the enrichments of composed BFs and biologically meaningful BFs in the reference biological dataset.	240
F.12	p -values for comparison between the enrichments of composed BFs and biologically meaningful BFs in the reference biological dataset. .	241

Abstract

Cells make decisions based on underlying gene regulatory networks (GRNs). GRNs may be modeled as a Boolean network (BN) in which nodes and directed edges represent genes or proteins and their interactions respectively. In BNs, a gene assumes a binary state and its temporal dynamics is governed by the state of its regulators via a *regulatory logic rule* (or logical update rule or Boolean function (BF)). The dynamics of BNs under synchronous update (in which all nodes are updated simultaneously) lead to fixed point attractors (which correspond to cellular phenotypes) or cyclic attractors.

Stuart Kauffman conceived BNs in 1969, and modeled GRNs as *random* BNs due to paucity of experimental data. Advances in experimental techniques including omics approaches have fostered the reconstruction of *real* Boolean GRNs for several cellular processes in a wide range of species. It is now imperative to understand whether the regulatory logic rules in such models, which have remained largely unexplored, are just random or possess distinct features.

In this thesis, we systematically investigate the nature of real regulatory logic rules by first compiling a dataset of 2687 logic rules from 88 reconstructed discrete models, and then examining in that dataset, the preponderance of various known types of *biologically meaningful BFs*. Two types that are particularly preponderant in our dataset are read-once functions (RoFs) and nested canalizing functions (NCFs). We explain this by showing that RoFs and NCFs have the minimum complexity at a given number of inputs (k) and given bias (P) in terms the *Boolean complexity* and the *average sensitivity* respectively. Furthermore, we also explore the abundance and biological plausibility of more recently published types of logic rules, namely, link operator functions (LOFs) and composition structures respectively.

The voluminous biological data generated over the past three decades has driven both manual reconstruction of Boolean GRNs and advancement of automated meth-

ods to reconstruct Boolean GRNs. Even if the network structure of a GRN is kept fixed, there is generally a very large number of combinations of BFs (across all nodes) that can recover the same set of biological fixed points (or cellular phenotypes), and it is usually unclear how a certain model or subset of models are chosen as the biologically relevant ones during reconstruction.

In this thesis, we leverage the *relative stability* of cell states derived from its developmental landscape to develop a framework that performs model selection on an ensemble of models that are otherwise equally plausible in the cell states they recover and in the type of logic rules (based on the minimum complexity criteria) they employ. We demonstrate our model selection framework on the latest root development Boolean GRN of *Arabidopsis thaliana* and provide several *improved* models over the original one.

In sum, we elucidate design principles of regulatory logic in GRNs and leverage those principles to develop methods for realistic reconstruction of Boolean models of biological systems in this thesis.

Chapter 1

Introduction

Cells in a multicellular organism contain the same genome, yet they exhibit a variety of phenotypes. A cell's phenotype is an outcome of a set of intricate and coordinated action of its molecular constituents, and often involves an integration of multiple molecular cues from its local environment as well [1]. The cellular machinery is driven by key molecular players such as DNA, RNA, proteins and ligands, and the interactions between them such as DNA-protein, protein-protein, RNA-RNA, protein-ligand and DNA-ligand interactions. More specifically, proteins called transcription factors can regulate gene expression by binding to *cis*-regulatory elements on the DNA such as promoters and enhancers [2], proteins can activate (or inhibit) other proteins by binding to them and forming functional (or non-functional) complexes [3,4], RNAs such as microRNAs can regulate gene expression by binding to mRNA and degrading them [5], enzymes such as kinases can activate inactive enzymes by catalyzing their phosphorylation [6], and receptor proteins can transduce signals via conformational changes they undergo by binding to ligands such as hormones, peptides and other small molecules [6]. In order to get a global, systemic perspective of how such a large and diverse set of molecular processes control cellular phenotypes, biological networks have been reconstructed by piecing together many experimental datasets [7–10]. Biological networks include gene regulatory

networks (GRN), signal transduction networks and metabolic networks. These networks differ in the types of nodes and interactions but are all important for various fundamental cellular processes such as cell growth, cell division and cell differentiation. More explicitly, GRNs map the *cis*-regulatory control by transcription factors of all genes implicated in a particular process - a node in a GRN may represent the *cis*-regulatory region of a gene or a transcription factor that gene codes for depending on whether the node is a regulatee or a regulator respectively [8, 11]. Signal transduction networks capture the relay of information within the cell by integrating several signaling pathways and its nodes are proteins or ligands, and edges are protein-ligand or protein-protein interactions [6]. Metabolic networks comprise of all the biochemical reactions in a cell and are necessary for self-maintenance mechanisms such as homeostasis [12, 13]. However, it is important to note that in reality these networks are not mutually exclusive and they together determine cellular phenotypes - metabolic enzymes may act as transcription factors [14] and transcription factors act as a bridge between gene regulation and signaling [15]. The architecture of these biological networks have been shown to be far from random [7–10] and reveal several *design principles* such as modularity, robustness to perturbations and the presence of network motifs that cells exploit for performing tasks necessary for survival [16, 17]. Gene expression and protein activity change over time depending on the presence or absence of their regulators, and hence, phenotypic changes a cell undergoes may be understood by viewing its GRNs and signal transduction networks as dynamical systems [18]. Biological networks may be modeled using different mathematical frameworks with the degree of coarse graining being to some extent proportional to the network’s complexity in terms of the number of nodes and edges [19]. So, small genetic networks with few nodes are better modeled using ordinary differential equations followed by mid-sized gene networks with tens to hundreds of nodes being modeled using discrete states, and large-sized networks with over thousands of nodes can be studied using flows [19]. One of the simplest

treatment of GRNs as dynamical systems is via Boolean networks (BNs) as proposed by Stuart Kauffman [20, 21] and is the mathematical framework that this thesis is based on.

1.1 Boolean models of gene regulatory networks

A Boolean model of a GRN consists of nodes that correspond to genes or proteins and directed edges representing the interactions between genes and proteins. A node is a *cis*-regulatory region of a gene for its regulators (nodes with edges pointing towards the gene) and a transcription factor for its regulatees (nodes with edges pointing away from it). In a BN, nodes can assume two expression states, namely, *on* or *off*, akin to switches and are represented by Boolean variables that takes values 1 (on) or 0 (off) [20–22]. For a BN with N nodes, we denote by $x_i(t)$ the state of node i at time t , where $i \in \{1, 2, 3, \dots, N\}$ and $x_i \in \{0, 1\}$. We use a vector $\mathbf{X}(t)$ to denote the state of all variables $x_i(t)$ of the network. $\mathbf{X}(t)$ is the gene expression pattern of the network at time t . As each node can assume 2 states, there are 2^N possible gene expression patterns, defining the state space of the BN. Furthermore, a Boolean function (BF) or *regulatory logic rule* governs the dynamics at each node by performing logical operations on the inputs to that node and returning an output. The BF associated with a node captures the combinatorial regulatory logic at the promoter region of the associated gene by the transcription factors (TFs) that regulate it. More succinctly, $x_i(t+1) = f_i(x_i^1(t), x_i^2(t), \dots, x_i^k(t))$, where f_i is the BF that acts on the k inputs to node i , $x_i^m(t)$ ($m \in \{1, k\}$) and $x_i^m(t) \in \{x_1(t), x_2(t), \dots, x_N(t)\}$, to return an output $x_i(t+1)$. Figure 1.1(a) and (b) shows a 5 node BN and its BFs. In general, the network structure only determines the set of nodes (variables) that regulate the state of each node, and the sign of interaction if it is known. More explicitly, the network structure (see Figure 1.1(a)) generally does not determine the specific logical operators such as AND (\wedge) and OR (\vee) that arise in the combinatorial regulatory logic rule. The AND and OR

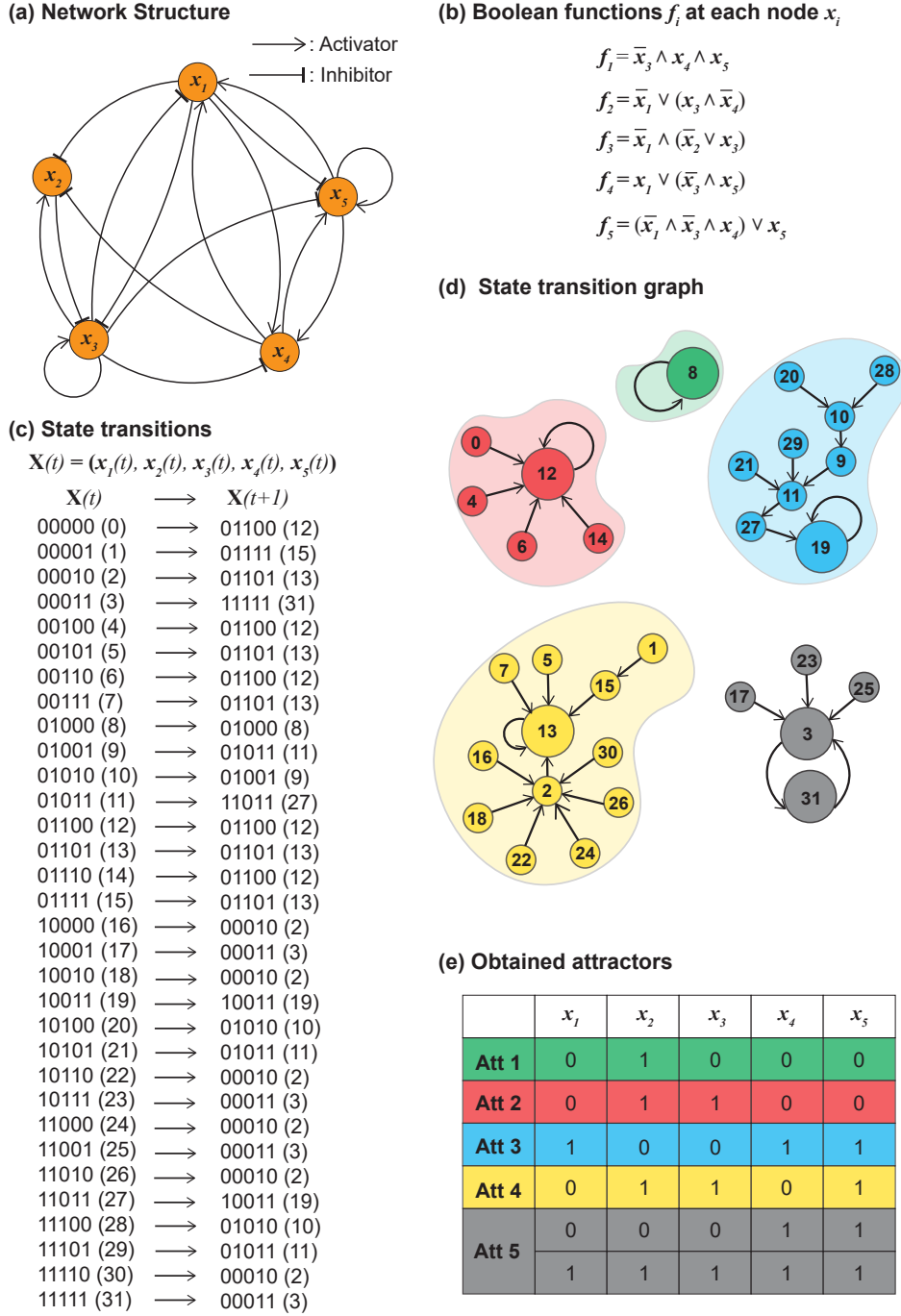


Figure 1.1: A toy Boolean network and its associated state transition graph. (a) The network structure of a toy Boolean model consisting of 5 genes and 16 edges. (b) Boolean function (BF) at each node specified as a Boolean expression of its input variables. \wedge corresponds to the AND operator, \vee corresponds to the OR operator and \bar{x}_i corresponds to a negated variable. (c) The state to state transitions for all $2^5 = 32$ states when synchronously updated using the BFs or logic rules provided in (b). (d) The state to state transitions in (c) lead to trajectories which either converge to fixed point attractors (as shown in colors green, blue, red and yellow) or cyclic attractors (as shown in grey). (e) The attractor states of this model. The cyclic attractor consists of 2 states shown in grey.

operators in BFs (see Figure 1.1(b)) are usually assigned by the modelers based on information about the regulation of a node by its inputs as obtained from published literature or experimental datasets. As will be elaborated in the next chapter, the number of possible BFs at any node in a BN is 2^{2^k} (without any restriction on the signs of the inputs). Note that Figure 1.1 is a toy model, and the BF at each node with k inputs was chosen arbitrarily from all possible BFs with k inputs that respect the given signs of the interactions.

Since the BFs $\mathbf{F} = (f_1, f_2, \dots, f_N)$ act on all nodes simultaneously, the update is said to be *synchronous* [21] and is expressed by the equation:

$$\mathbf{X}(t+1) = \mathbf{F}[\mathbf{X}(t)] \quad (1.1)$$

Throughout this thesis, each node is assigned exactly one BF and the BNs are updated synchronously. Starting from an initial state $\mathbf{X}(0)$, a state space trajectory can be traced out by the recursive application of Eq. (1.1). The collection of all such trajectories constitutes the *state transition graph* of the BN (see Figure 1.1(c) and (d) for state transitions and the state transition graph respectively). Under synchronous dynamics, a trajectory can meet with 2 possible fates. One, it reaches a state which on further update remains unaltered, in which case the trajectory is said to have converged to a *fixed point attractor* or simply, a *fixed point*. Two, it keeps cycling through a set of states, in which case the trajectory is said to have converged to a *cyclic attractor* or simply, a *cycle*. All the states which converge to an attractor (including the attractor itself), constitute its basin of attraction. We remark that the dynamics presented above is purely deterministic and generates a state transition graph with multiple disconnected components, each one corresponding to one attractor. Fixed point attractors represent emergent steady state expression patterns arising from the complex net of molecular interactions, and specify several cellular phenotypes including growth, differentiation, apoptosis,

quiescence and motility among others [23]. These fixed point attractors or cellular phenotypes may also be viewed as valleys on a potential landscape, where switching between phenotypes can then be viewed as biological phase transitions [23]. Experiments have revealed that cell fates are indeed high-dimensional attractor states [24]. Figure 1.1(e) shows the attractors; Att1, Att2, Att3 and Att4 are fixed points and Att5 is a cycle. We remark here that asynchronous update schemes where nodes are not updated simultaneously are also widely used in Boolean modeling and are argued to be more biologically realistic since different nodes may be regulated in different time scales [25]. However, it is important to note that fixed point attractors remain unchanged across different updating schemes though the states flowing to those attractors may differ with the update scheme [25].

Though gene expression is measured as continuous values of protein concentrations, yet such a coarse graining to binary states as in BNs can offer insights into the dynamics of the network [19]. One of the initial successes that demonstrated the power of this framework was by Albert and Othmer [26] where they predicted the general dynamical trajectory of the segment polarity gene network in *Drosophila melanogaster* solely on the basis of their reconstructed Boolean model. Since then, several Boolean models have been reconstructed that successfully replicated expected phenotypes [27, 28] and have also had implications in providing therapeutic interventions in diseased cell states [29, 30].

1.2 Regulatory logic rules in Boolean biological networks

Regulatory logic rules in BNs drive its dynamics from one time instant to the next. In GRNs, the logical rule at a gene encodes the combinatorial regulation at its promoter region by its regulators such as transcription factors (TFs). A classic instance is the transcription regulation of the lac operon in *Escherichia coli*. In the presence of

lac repressor protein (LacR), the RNA polymerase cannot bind to the transcription start site, thereby preventing the transcription of the operon [31]. However, presence of the cAMP receptor protein (CRP) in the absence of LacR facilitates the recruitment of RNA polymerase, leading to the transcription of the lac operon. Though we have abstracted away several intermediate molecular processes in the above example, the overall transcriptional logic rule governing the expression of the lac operon can be encapsulated by the Boolean formula “CRP AND NOT LacR” [31]. In signaling pathways, where protein-protein interactions mediate signal transduction, BFs encode whether a protein is in an active or inactive state. A simple example is the activation of cyclin dependent kinase (cdk). cdk is active only in the absence of the cyclin inhibitor (CDI) such as p53 and in the presence of cyclin (cyc) such as cyclin D1 [23]. The logical rule governing the activity of cdk can be expressed as “cyc AND NOT CDI”. A more complicated scenario arises in the signaling network of epidermal growth factor (EGF) and neuregulin-1 (NRG1). Here, ErbB1 and ErbB3 (multiple receptor kinases) form a heterodimer ErbB13 which undergoes increased phosphorylation after binding with EGF or NRG1, which in turn activates further downstream signaling processes. However, since ErbB1 prefers binding to ErbB2 over other ErbB receptors, the presence of ErbB2 prevents the formation of ErbB13 complex by binding to ErbB1. The logical rule associated with ErbB13 is then “(EGF AND ErbB1 AND ErbB3 AND NOT ErbB2) OR (NRG1 AND ErbB1 AND ErbB3 AND NOT ErbB2)” [32]. In eukaryotes, genomic regions such as enhancers to which multiple TFs may bind also participate in the combinatorial regulation of gene expression, thereby increasing the complexity of gene regulation [2]. With a large-scale collection of such regulatory logic rules that govern molecular decisions, it may be possible to systematically investigate and uncover their associated design principles.

Stuart Kauffman had proposed *canalyzing* functions (CFs) as a type of regulatory logic that several biological systems employ in molecular decision making [22]. The

regulatory logic associated with a gene is said to be a CF if the state of at least one of the regulators of a gene determines its expression state regardless of the state of the other regulators. Interestingly, Kauffman found that the fraction of CFs in all BFs decreases with an increasing number of inputs [22]. Put differently, imposing the property of canalization on BFs restricts the space of all BFs to a smaller biologically relevant space of BFs. Since the proposition of CFs, other properties that BFs are expected to possess based on biological observations have also been introduced [33–35]. In 2003, Kauffman *et al.* [36] showed in a dataset of 139 Boolean rules of transcriptional regulation (published by Harris *et al.* [37]), that 133 of them belonged to a subset of CFs, namely, nested canalizing functions (NCFs) - an observation for which the explanation is not immediately apparent. Note that if the regulatory logic was random, we would expect that most of these rules would not be NCFs. The mathematical properties of the NCFs has been an active area of research since then [38–40]. Over the past two decades, several types of regulatory logic rules that are potentially biologically meaningful have been introduced in the literature such as chain functions [41], post-classes [42], link operator functions (LOFs) [35] and composed BFs arising from composition structures [43].

At the time of inception of BNs, hardly any data on combinatorial gene regulation was available, and so GRNs were modeled using random Boolean networks (RBNs). In RBNs, each node has a fixed number of incoming edges from randomly chosen nodes, and each node is assigned a BF that is drawn from all possible k -input BFs under some probability distribution [21]. The advent of sequencing technologies and its coupling with techniques such as chromatin immunoprecipitation assays has revolutionized the study of transcriptional control of cellular processes [44]. This has propelled manual and computational reconstruction of numerous BNs that model a diverse range of cellular phenotypes across a wide range of organisms so much so that repositories such as the Cell Collective [45] have been developed to archive these models. Now, large-scale meta analysis of Boolean models in this repository and

others are being undertaken to uncover design principles that underlie regulation of gene expression [46,47]. In that context, most studies have so far focused on how the topology of the BN affects its dynamics. Though the topology is extremely important, it is insufficient to provide a holistic understanding of the dynamics without addressing the question of the types of regulatory logic rules used therein. Indeed, though a large number of topologies of GRNs are possible theoretically, yet only a subset of them actually occur in nature. The number of k -input BFs is 2^{2^k} , a space that is extremely large and is of the order of 10^9 even for 5-input BFs.

Objective 1: Analogous to network architecture of biological networks being far from random, are real regulatory logic rules also far from random?

In this thesis, we address this first objective by first collating the various types BFs in the Boolean modeling literature that are biologically meaningful based on the properties they possess such as *effectiveness* [33,48], *unateness* [34], canalyzation and nested canalyzation [22], leading to effective functions (EFs), unate functions (UFs), CFs and NCFs respectively. We quantify their fraction in the space of all BFs for different number of inputs, and computationally check which types overlap and to what extent. We backup several of our computational observations with theoretical proofs. We also propose as a potential candidate of biologically meaningful type, the read-once functions (RoFs) that have only been explored in computer science so far. Our results are reported in the publication [49]. We undertake a similar exercise with the more recently proposed types of BFs, namely, the LOFs and the composition structures, and report our results in the publications [50] and [51] respectively. More explicitly, we quantify their fraction in the space of all BFs for various inputs, and computationally check and quantify the overlaps between these types of BFs with the other biologically meaningful types. Next, we assess the preponderance of different types of biologically meaningful BFs in a reference dataset of Boolean regulatory logic rules that have been extracted from a large number of manually reconstructed

Boolean models and provide a complexity-based explanation of why certain types are more preponderant compared to other types. Our central result is that regulatory logic rules in reconstructed Boolean models are *minimally complex* [49]. The above mentioned points are summarized in the top 4 panels of Figure 1.2. These results are reported in the publication [49]. We also quantify the preponderance of the more recent LOFs and composed BFs in large datasets of regulatory logic rules and report our observations in the publications [50] and [51] respectively. The implications of using numerous types of BFs for a given network structure, for the network dynamics is also explored [49, 50]. Lastly, using ChIP-seq data, we investigate the biological plausibility of composition structures in transcriptional gene regulation [51].

1.3 Relative stability as a model selection constraint

Manually reconstructed BNs of GRNs have successfully modeled gene expression patterns of various cell fates [52, 53]. Yet, such models are only one of several possible ones that can recover those cell fates. In fact, this is true even if one fixes the network structure of the GRN and models differ only in the choice of logic rules across the nodes [54]. So, a question arises which model(s) is (are) the most biologically relevant in the space of models that are equally plausible at the level of the biological cell fates they recover? Indeed, it is extremely cumbersome to obtain experimentally validated regulatory logic rule at each node of the GRN as it requires several perturbation experiments of various combinations of genes [8]. Therefore modelers usually infer regulatory logic at a gene by gathering experimental observations pertaining to its regulation from published literature. But this imparts some subjective bias (conscious or unconscious) in the choice of regulatory logic rules in these Boolean models. To overcome such biases, several efforts have been undertaken to automate the reconstruction of BNs [55, 56] by incorporating several

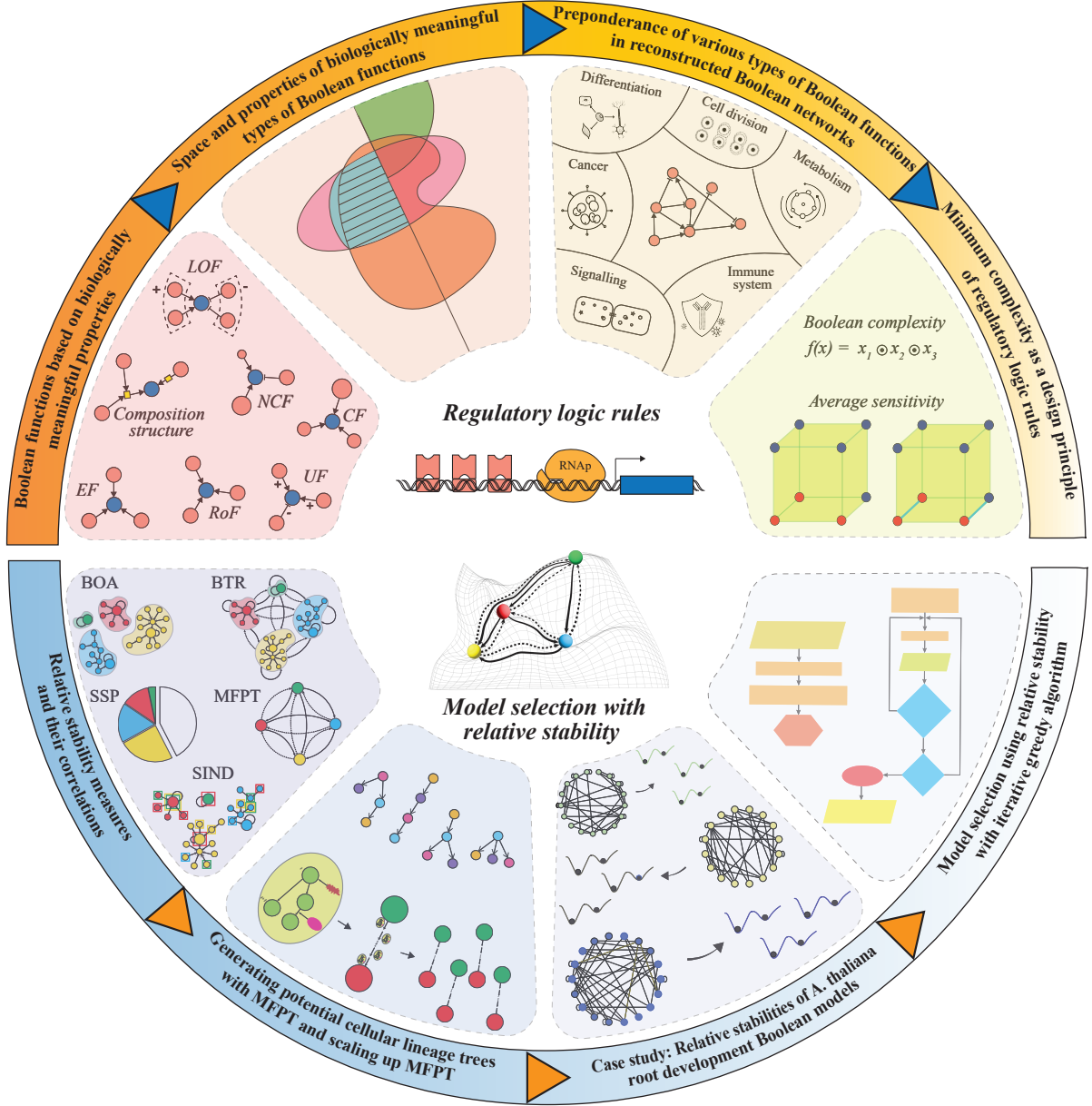


Figure 1.2: Schematic overview of the broad objectives of the thesis and how the thesis addresses them. The four panels in the top half of the figure provides an overview of our approach to tackle the first objective, namely, whether regulatory logic rules in reconstructed BNs is random or not. The four panels in the bottom half of the figure provide an overview of our approach to tackle the second objective, namely, developing a model selection framework that uses relative stability as a constraint.

constraints derived from experimental observations and datasets. Some of these constraints have been imposed at the level of the logic rule, some at the level of the state transitions trajectories, some at the level of the attractor and its basin, or a some combination of these. In particular, constraints that place restrictions on the emergent dynamics such as state transition trajectories inferred from time series data [56], relative stability between cell states [54, 57], *reachability* between states [58], and attractor multiplicity [59] enable a *reverse engineering* approach to infer the regulatory logic rules at the nodes.

We provide here a brief overview of some automated model inference and model selection methods that have been developed over the past two decades. One of the earliest algorithms to infer BNs is REVEAL - it uses a mutual information based approach to show that a few state transition pairs are sufficient to infer Boolean GRNs [60]. Other algorithms for Boolean GRN inference include ones that start with data for various inputs such as time series data from gene expression datasets [55, 56, 61, 62], manually constructed prior knowledge networks based on literature evidence for interactions [63] or partial information about the network and rules [59], and return a Boolean GRN. In many model inference frameworks, the choice of regulatory logic rules are restricted to certain types. For instance, Martin *et al.* [56] restrict to an AND-NOT type of regulatory logic rule which is a subset of the NCFs, Maucher *et al.* [64] allow for effective and unate functions, Zhou *et al.* [54] impose both the NCF and the unateness condition based on interaction signs, Ghaffarizadeh *et al.* [65] restrict the BFs to the more general class of NCFs, Biosketches [59] force BFs to be effective, unate and canalyzing, and in ATEN [66] BFs are a type of LOF. Other biological constraints based on the emergent dynamics such as attractor landscape, that is, the phenotypic constraints, the attractor multiplicity and basin of attraction are also being used [58, 59].

Though all these methodologies lead to models that respect many biological constraints, yet, an important constraint based on the hierarchies of cell states on the

developmental landscape of developmental GRNs (DGRNs), have been overlooked. Such a constraint has been quantified via the relative stability of a pair of cell states [54, 57] - the propensity to transition from one cell state to another. Incorporation of such a constraint into model selection workflows eliminates models that may be equally plausible in the type of logic rule they employ and the biological attractors they recover but do not conform to the hierarchies of cell states on the developmental landscape. Lastly, we re-emphasize that constraining Boolean models using relative stability provides a reverse engineering approach to infer regulatory logic rules at nodes.

Objective 2: Development of a model selection workflow that incorporates relative stability of cell states as a criteria for model selection

In this thesis, we address the second objective by leveraging the design principle based on hierarchies between cell states on the developmental landscape and minimum complexity design principle for regulatory logic rules. We first show that various measures of relative stability [54, 67] in the literature are strongly correlated. Using the mean first passage time (MFPT) as our relative stability measure we propose a method to construct a potential cellular lineage tree. In order to scale the MFPT to larger BNs, we propose its stochastic counterpart. With this methodology, we take as a case study, the developmental landscape of 3 Boolean models of *Arabidopsis thaliana* root development and find that the latest one (a 2020 model) does not respect the biologically expected hierarchy of cell states based on their relative stabilities. Therefore we develop an iterative greedy algorithm that searches for models which satisfy the expected hierarchy of cell states. Our methodology thus provides new tools that can enable reconstruction of more realistic and accurate Boolean models of developmental GRNs. The above mentioned points are compactly illustrated in the bottom 4 panels of Figure 1.2. The results of this work are reported in the publication [68].

1.4 Thesis organization

The remaining chapters of this thesis are organized as follows:

Chapter 2 presents a delineation of various types of BFs in the Boolean modeling literature that are considered to be biologically meaningful based on properties that gene regulation is expected to possess. To understand these biologically meaningful BFs from a mathematical standpoint, we first describe different representations of BFs, namely, the truth table, Boolean expression and Boolean hypercube and present the $k[P]$ classification of BFs as introduced by Feldman [69,70] based on the number of inputs (k) and their *bias* (P). Such a classification provides a bias-based perspective on biologically meaningful BFs that has hitherto remained unexplored in Boolean models. Following this, we provide formal definitions of these different types of biologically meaningful BFs, namely, effective functions (EFs), unate functions (UFs), canalyzing functions (CFs) and nested canalyzing functions (NCFs), and introduce for the first time, the read-once functions (RoFs) as a potential candidate for a biologically meaningful type of BF. Next, we characterize the space of biologically meaningful functions by computationally quantifying the fraction occupied by the biologically meaningful types in the space of all BFs for different number of inputs and also their overlaps with one another. Lastly, we provide theoretical proofs to explain several observations regarding the overlaps of these biologically meaningful types obtained via computational means. **The work reported in this chapter is contained in the published manuscript [49].**

Chapter 3 presents a systematic study of the preponderance of various biologically meaningful types of BFs in a dataset of regulatory logic rules from reconstructed Boolean models of biological systems and a complexity-based explanation of why certain types are more preponderant than others. So far, the properties of random BNs have been investigated extensively as models of regulation in biological systems. However, the BFs specifying the associated logical update rules in recon-

structured Boolean models of biological systems are not be expected to be random. To address the question of which types of BFs are preponderant in real Boolean models of GRNs, we first extract 2,687 regulatory logic rules from 88 published discrete biological Boolean models. A surprising feature is that most of the BFs in our dataset have odd bias, that is they produce ON outputs for a total number of input combinations that is odd. We explain this observation, along with the enrichment of RoFs and its NCF subset, in terms of two complexity measures: Boolean complexity [69] based on string lengths in formal logic, which is yet unexplored in biological contexts, and the so-called average sensitivity [71]. RoFs have the minimum Boolean complexity in a $k[P]$ set for all odd biases P . Furthermore, using a half-century old proof on hypercubes [72], we show that NCFs have the minimum average sensitivity in a $k[P]$ set with odd bias P (in addition to having minimum Boolean complexity in the same set). These results reveal the importance of minimum complexity in the regulatory logic of biological networks. **The work reported in this chapter is contained in the published manuscript [49].**

Chapter 4 presents a framework for model selection of Boolean DGRNs by leveraging design principles based on regulatory logic rules and on the hierarchies of cell states on the developmental landscape. During the reconstruction of Boolean DGRNs, even if the network structure is fixed, there is generally a large number of combinations of BFs that will reproduce the different cell fates (biological attractors). Here we leverage the developmental landscape to enable model selection on such ensembles (generated by restricting the regulatory logic rules to certain biologically meaningful types) using the relative stability of the attractors. First, we show that previously proposed measures of relative stability [54,67] are strongly correlated and we stress the usefulness of the one that captures best the cell state transitions via the mean first passage time (MFPT) as it also allows the construction of a cellular lineage tree. A property of great computational importance is the insensitivity of the stability measures to changes in noise intensities, allowing us to use stochastic

approaches to estimate the MFPT and thereby scale up its computation to large networks. With this methodology, we revisit different Boolean models of *Arabidopsis thaliana* root development published in 2013 [73], 2017 [74] and 2020 [75], and show that the most recent one does not respect the biologically expected hierarchy of cell states based on relative stabilities. Lastly, we develop an iterative greedy algorithm that searches for models which satisfy the expected hierarchy of cell states, starting from one that does not satisfy some of those constraints, and find that its application to the root development model yields many models that meet this expectation. Our methodology thus provides new tools that can enable reconstruction of more realistic and accurate Boolean models of DGRNs. **The work reported in this chapter is contained in the published manuscript [68].**

Chapter 5 presents a characterization of various types of link operator functions (LOFs), their abundance in a dataset of regulatory logic rules obtained from reconstructed Boolean models of biological systems and how they can drive the dynamics of Boolean GRNs towards criticality. First, we define the LOFs as proposed by Zobolas *et al.* [35] and consider for further study the four types that are biologically consistent. We then theoretically enumerate the number of LOFs for different number of inputs and show that among all BFs, and even within the effective and unate functions (EUFs), the biologically consistent LOFs form a tiny subset. Of these different types of LOFs, namely, AND-NOT, OR-NOT, AND-pairs and OR-pairs, we find that the AND-NOT LOFs are particularly abundant among BFs extracted from reconstructed Boolean models of biological systems. By leveraging these facts, namely, the tiny representation of LOFs in the space of EUFs and their presence in a reference biological dataset, we show that the space of acceptable models can be shrunk considerably by applying steady-state constraints to BFs, followed by the choice of biologically consistent LOFs which satisfy those constraints. Finally, we demonstrate that among a wide range of BFs, the LOFs drive biological network dynamics towards criticality. **The work reported in this chapter is contained**

in the published manuscript [50].

Chapter 6 presents an investigation of the plausibility of composition structures as a mathematical framework that can capture biological gene regulation. BNs are coarse-grained to an extent that they abstract away molecular specificities of gene regulation. Alternatively, bipartite BN models of gene regulation explicitly distinguish genes from TFs. In such bipartite models, multiple TFs may simultaneously contribute to gene regulation by forming heteromeric complexes, thereby giving rise to composition structures [43]. Since bipartite Boolean models are relatively recent, an empirical investigation of their biological plausibility is lacking. In this chapter, we estimate the prevalence of composition structures arising through heteromeric complexes in humans and yeast. Moreover, we present an alternate mechanism by which composition structures may arise, that is, as a result of multiple TFs binding to *cis*-regulatory regions and also provide empirical support for this mechanism. Next, we quantify the extent of restriction on the space of BFs imposed by composition structures versus that imposed by biologically meaningful BFs. We find that though composition structures can severely restrict the number of BFs, the two types of minimally complex BFs, NCFs and RoFs, are comparatively more restrictive. Finally, we find that composed BFs arising from composition structures are highly enriched in reconstructed Boolean models of biological systems, but this enrichment most likely stems from the enrichment of NCFs and RoFs. **The work reported in this chapter is contained in the published manuscript [51].**

Chapter 7 concludes this thesis with a brief summary of the research reported across different chapters. The chapter also discusses the future prospects of this thesis.

Chapter 2

Biologically meaningful Boolean functions: Description and properties

Extensive studies of biological networks made possible by recent advances in large-scale data acquisition have revealed that their topological structure is very far from random [7, 10, 17, 76]. However, hardly any efforts (except a few such as [37]) have gone into systematically understanding how far Boolean functions (BFs) encoding the associated regulatory logic rule in nodes of Boolean networks (BNs), are from being *random*. A first step in that direction would be to collate the various types of BFs that are potentially *biologically meaningful* from the existing literature, and characterize their properties. Several biologically meaningful functions have been introduced in the literature, some of which include: effective functions (EFs) [33, 48], unate functions (UFs) [34, 77], canalyzing functions (CFs) [22, 78] and nested canalyzing functions (NCFs) [36, 38–40, 79]. Though each of these types of BFs have been extensively studied individually, a holistic understanding of how these different types are related to each other remains unexplored.

In this chapter, we systematically study the properties of biologically meaningful BFs, namely, EFs, UF, CFs and NCFs. We also introduce as a potential type of biologically meaningful regulatory logic, the *read-once functions* (RoFs) that are well studied in computer science but have not been explored in the context of gene regulation. We first quantify the fraction of functions of the biologically meaningful BFs in the space of all BFs, following which we quantify the overlaps or intersections of these different types of biologically meaningful BFs. Our observations based on our computational exploration of the overlaps between the different spaces brought forth several questions some of which include: (i) Are all ineffective functions even biased? (ii) Are all odd biased functions effective? (iii) Are all UFs effective? (iv) Are all RoFs odd biased? We prove several of our computational observations and state them as properties in Section 2.4. Finally, we leverage many of our obtained properties to design an *RoF checker* algorithm that checks whether a given BF is a RoF. **The work reported in this chapter is contained in the published manuscript [49].**

2.1 Boolean functions

BFs govern the temporal dynamics of the BN and capture the combinatorial regulatory logic at the promoter region of the gene. Since a major part of this thesis is devoted to understanding the nature and properties of regulatory logic, we provide in this section various representations and properties of BFs, and a classification scheme based on the isometries of BFs. All of these properties lay the foundation for the results on biologically meaningful BFs presented in the subsequent chapters.

2.1.1 Representations of BFs

Let $f = f(x_1, x_2, \dots, x_k)$ be a BF of k input variables where each variable $x_i \in \{0, 1\}$. A BF maps 2^k different possibilities for the k input variables to output values 0 or

1, i.e., $f : \{0, 1\}^k \mapsto \{0, 1\}$.

Truth table and associated ordered binary vector

A BF f with k inputs can be expressed in the form of a truth table with $k + 1$ columns and 2^k rows. The first k columns correspond to input variables and the last column to the output values (see Figure 2.1(a)). Each of the 2^k rows correspond to a possible state of the set of k input variables. In our convention the last entry of each row gives the output value for the corresponding state of the input variables (see Figure 2.1(a)). Thus, a BF can be stored as a binary vector of size 2^k , where each element of the vector corresponds to the output value of the corresponding row of the truth table (see Figure 2.1(b)). A BF can also be encoded as the integer which is the decimal equivalent of the binary vector of size 2^k . Again, there are two conventions to encode the truth table as an integer: (i) where the output bits corresponding to rows 0 and $2^k - 1$ of the truth table respectively are the most significant and least significant bits respectively, and (ii) where the output bits corresponding to rows 0 and $2^k - 1$ of the truth table respectively are the least significant and most significant bits respectively. Throughout this thesis, we will use the encoding convention (i). Note that it is necessary to fix the ordering of the input variables in the input columns of the truth table (from left to right). Since the output value for each of the 2^k different states of the k input variables can take either of two values, 0 or 1, the number of possible BFs f with k inputs is 2^{2^k} . Thus, the number of possible BFs f blows up quickly with increasing k [22], e.g., there are over 10^9 BFs with $k = 5$ inputs.

Boolean Expression

Alternatively, a BF f can instead be represented as an algebraic expression (see Figure 2.1(c)) constructed with the k input variables which are combined via the logical operators AND (\cdot or \wedge), OR ($+$ or \vee) and NOT ($'$ or $-$). For example, the AND function and OR function of 2 input variables, x_1 and x_2 , are given by the

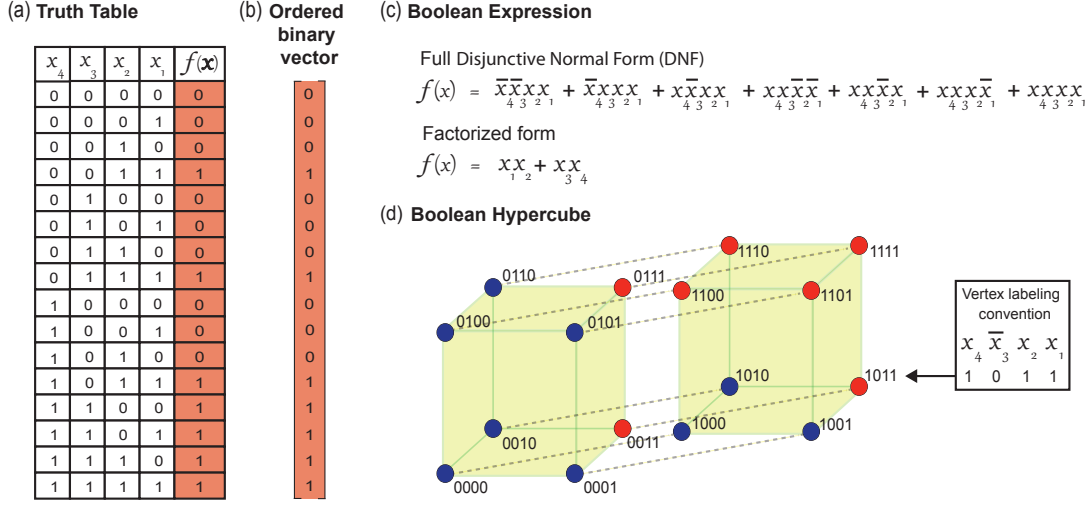


Figure 2.1: Different representations of Boolean functions.

Boolean expressions, $x_1 \wedge x_2$ (or $x_1 \cdot x_2$, or x_1x_2) and $x_1 \vee x_2$ (or $x_1 + x_2$), respectively. In this work, the $+$ symbol in a Boolean expression does not correspond to working modulo 2, instead $(1+1)$ has the value 1, not 0. The term *literal* refers to a Boolean variable (e.g., x_i) or its complement (e.g., $\overline{x_i}$). Throughout this work, BF and *Boolean variables* refer to a *logical update rule* and to *inputs*, respectively, of nodes in a BN.

Colored Boolean Hypercube

A visually illustrative representation of a BF is obtained by coloring a Boolean hypercube. A k -dimensional hypercube (k -cube) is composed of vertices and edges where each vertex is labelled by a string of k bits, and is connected to vertices with labels that differ from its label in exactly one bit. Two vertices connected by an edge are called *neighbors*. A k -cube thus has 2^k vertices, with each vertex having k neighbors. The total number of edges in a k -cube is $(k \times 2^k)/2 = k2^{k-1}$ (division by 2 removes the doubly counted edges). A BF may thus be represented by a k -cube in which each vertex is labeled by the input combination $x_kx_{k-1}x_{k-2} \dots x_2x_1$ ($x_i \in \{0,1\}$) and is colored with an output bit (0 or 1) (see Figure 2.1(d)).

2.1.2 Categorization of BFs based on their bias and isometries

In his work, Feldman [69, 70] classified BFs based on the number of inputs (or variables, D , in his notation), and on the unnormalized bias P . In the Boolean modeling literature, the *bias* p of a BF refers to the *probability* for it to take the value 1 [22, 71]. The unnormalized bias P , which we will refer to simply as the *bias* henceforth, is the number of 1s in the output vector of the BF. In other words, bias is simply the Hamming weight of the BF. If a BF has odd or even bias P , the *parity* of the BF is said to be odd or even, respectively. Note that the bias is identical to the parity of the BF, and is consistent with the standard notion of parity used in computer science literature [80]. However, it is important to note that parity bits - bits that are appended to a binary string to obtain an even or odd parity string - are not used in this thesis. Such bits are typically used in telecommunication systems for error-detection. For notational consistency, we will use k instead of D hereafter and will denote the set of all BFs for a given number of inputs k and bias P as $k[P]$. It is easy to see that the number of $k[P]$ sets for a given k is $2^k + 1$.

In addition, within any given $k[P]$ set, Feldman [69, 70] introduced partitions using equivalence classes based on isomorphisms. Two BFs f and g are defined as *isomorphic* if they are identical up to permutations and negations of any of their input variables. In terms of the Boolean expression, permutation of a pair of variables implies the exchange of those two variables in that expression and the negation of a variable is the changing the literal from a positive one to a negative one or vice versa. Similarly, in terms of the truth table, permutation implies the permuting the columns of the truth table and then reordering the rows (which include the output of the BF) in lexicographical order to bring the table into its canonical form, and negation of an input x_i of a BF f can be obtained by swapping the values of the initial table's *output* column in all pairs of rows that differ only by the value of x_i . For example, the BF $f = x_1 \wedge (x_2 \vee x_3)$ is isomorphic to the BF $g =$

$x_2 \wedge (x_1 \vee \bar{x}_3)$. The code to perform these operations of permutation and negation of any BF are provided in https://github.com/asamallab/MCBF/tree/main/BF_codes. These transformations define equivalence classes of BFs within the $k[P]$ set. For each class, we can choose the representative BF where the first occurrence of each variable arises both sequentially (with indices $1, 2, 3, \dots$) and as a positive literal. Note that in some cases both literals of a given variable may be present in the Boolean expression (e.g., $(x_1 \wedge \bar{x}_2) \vee (x_2 \wedge \bar{x}_1)$), in which case that expression is taken as the representative Boolean expression. Note also that every function in $k[P]$ has a *complementary* function in $k[2^k - P]$ which can be obtained via complementation of the corresponding Boolean expression (see e.g., [70]). In terms of the truth table, this is equivalent to complementing the values in its output column. Visually, a BF in a $k[P]$ set can be thought of as a k -cube wherein any P vertices are colored red (corresponding to output value 1) and the remaining $2^k - P$ vertices are colored blue (corresponding to output value 0) (see Figure 2.1(d)). Note that on the Boolean hypercube, isomorphic elements of any arrangement of P red vertices (i.e., 1s) can be generated by rotation (i.e., permutation) and reflection (i.e., negation) about any axis (i.e., input variable) of the Boolean hypercube (i.e., function). For any given assignment of 1s on P vertices of the k -cube, the total number of edges stemming from these P vertices is kP . Of these, some edges end at one of the other $P - 1$ vertices with value 1; we refer to this set of edges as E_{11} . Similarly, we denote by E_{01} the remaining edges, ending at any of the $2^k - P$ other vertices having the value 0. We then have $E_{01} + 2E_{11} = kP$ [72].

In sum, the introduction of the set $k[P]$ and the classification scheme proposed by Feldman [69, 70] efficiently encapsulates the 2^{2^k} possible BFs with k inputs in terms of significantly fewer representative quantities and expressions. Interestingly, Reichhardt and Bassler [81], using concepts borrowed from isomer chemistry and group theory, have shown how to count the number of distinct representative non-isomorphic BFs in each $k[P]$ set.

2.2 Biologically meaningful types of BFs

The number of BFs that can be assigned to a node with k inputs is 2^{2^k} (see Section 2.1.1). Clearly, this number explodes with growing number of inputs and it is thus imperative to focus on those subsets of BFs which possess biologically meaningful properties [33]. In this section, we compute the number and fraction of these biologically meaningful types of BF in the space of all BFs and within other such types. Using the observations from those computations we identify several patterns that we prove and provide as properties of different types of BF. In particular, we illustrate the utility of these simple properties along with those presented in the previous section to design algorithms that can generate RoFs and check whether a given BF is a RoF.

2.2.1 Effective functions

A BF may possess inputs which are mute or *ineffective* in the following sense: altering the binary state of an ineffective input, while keeping the state of other inputs unchanged, never alters the output value of the BF. Note that this must be true for all possible combinations of state of other inputs. Biologically, a regulatory element can be considered to be an *effective* regulator of the expression of a gene if and only if there exists some input condition wherein the modulation of the regulator alters the expression of the gene. If such a condition does not exist, then that regulator (or input) can be considered to be ineffective as it plays no role in regulating the gene under consideration. It follows that all inputs (or regulators) of a biologically meaningful BF should be effective [33]. A BF f with k inputs is an EF if and only if:

$$\forall i \in \{1, 2, \dots, k\}, \exists \mathbf{x} \in \{0, 1\}^k \text{ with } x_i = 0, f(\mathbf{x}) \neq f(\mathbf{x} + \mathbf{e}_i). \quad (2.1)$$

Here $\mathbf{e}_i \in \{0, 1\}^k$ denotes the unit vector associated to the component of index i . Simply put, all the inputs of an EF are effective.

As an example of an EF, consider the BF that encodes the activation of cyclin dependent kinase (cdk) presented in Chapter 1, Section 1.2 - cdk is activated only in the

presence of cyclin (cyc) and in the absence of cyclin inhibitor (CDI). This is captured via the Boolean expression “cyc AND NOT CDI” [23]. To assess the effectiveness of the input cyclin, we consider two cases: one in which CDI is present and the other in which CDI is absent. When CDI is present, the presence or absence of cyclin does not alter the output, namely, the activation state of cdk. But when CDI is absent, the presence of cyclin leads to the activation of cdk whereas the absence of cyclin does not activate cdk. The change of the output obtained by flipping the input state of cyclin (while keeping CDI fixed) defines the effectiveness of the input cyclin. Similarly, one can also argue that CDI is an effective input. Therefore, the BF for cdk activation is an EF.

2.2.2 Unate functions

A regulatory element may act as an activator or an inhibitor of the expression of a target gene. This information on the nature of the interaction between the regulator and its target gene can also be incorporated into the BF. The BFs that can account for this *sign* of regulatory interactions are known as *unate functions* (UFs) [34]. A BF f with k inputs is said to be increasing monotone (activating) in an input i (or variable x_i) if:

$$\forall \mathbf{x} \in \{0, 1\}^k \text{ with } x_i = 0, f(\mathbf{x}) \leq f(\mathbf{x} + \mathbf{e}_i), \quad (2.2)$$

and decreasing monotone (inhibiting) in an input i (or variable x_i) if:

$$\forall \mathbf{x} \in \{0, 1\}^k \text{ with } x_i = 0, f(\mathbf{x}) \geq f(\mathbf{x} + \mathbf{e}_i). \quad (2.3)$$

A BF f with k inputs is said to be a UF, if each input $i = 1, 2, \dots, k$ is increasing monotone (activating) or decreasing monotone (inhibiting). A BF f with k inputs is said to be a positive (respectively negative) UF, if every input $i = 1, 2, \dots, k$ is increasing monotone (respectively decreasing monotone) [34, 82].

As an example of an UF, consider again the activation of cdk presented in Chapter 1, Section 1.2 - cdk is activated only in the presence of cyc and in the absence of CDI. This is captured via the Boolean expression “cyc AND NOT CDI” [23]. To assess the

unateness of the input *cyc*, we consider two cases: one in which CDI is present and the other in which CDI is absent. When CDI is present, the change of the state of cyclin from its absence to its presence (change of its state from 0 to 1) does not alter the output, namely, the activation state of *cdk*. But when CDI is absent, the change of cyclin from its absence to its presence (change of its state from 0 to 1) changes the state of *cdk* from being inactive to being active. Thus, the increasing monotonicity condition is satisfied for the case when CDI is both present and absent. Therefore, *cdk* is an increasing monotone input. A similar argument can be made for CDI, which satisfies the decreasing monotone property. Such a BF is therefore a UF as both its inputs satisfy the monotonicity property.

2.2.3 Canalyzing functions

A BF f with k inputs is said to be canalyzing in an input i (or variable x_i) if and only if:

$$\begin{aligned} f(x_1, x_2, \dots, x_{i-1}, x_i = a, x_{i+1}, \dots, x_k) &= b, \\ \forall x_j, j \neq i, i \in \{1, 2, \dots, n\} \end{aligned} \tag{2.4}$$

In the above equation, a and b can take values 0 or 1, a is the canalyzing input value and b is the canalyzed value for input i . A BF f is a *canalyzing function* (CF) if at least one of its k inputs satisfies the canalyzing property [22]. Furthermore, for a CF, the number of distinct inputs that satisfy the canalyzing property (in a hierarchical fashion) is known as its *canalyzing depth* [83]. CFs have been grouped based on the number of canalyzing inputs [83]. For a CF with k inputs, the canalyzing depth can have an integer value in the range $\{1, 2, \dots, k\}$.

As an example of a CF, consider again the activation of *cdk* presented in Chapter 1, Section 1.2 - *cdk* is activated only in the presence of *cyc* and in the absence of CDI. This is captured via the Boolean expression “*cyc* AND NOT CDI” [23]. In the absence of *cyc*, *cdk* cannot be activated either in the presence or absence of the other inputs (in this case, CDI). Therefore *cyc* is a canalyzing input. Similarly, in the presence of CDI, *cdk* cannot be activated either in the presence or absence of the other inputs (in this case, *cyc*). Therefore CDI is also a canalyzing input. Since at least one input of this BF is

canalyzing, the BF is said to be a CF.

2.2.4 Nested canalyzing functions

Nested canalyzing functions (NCFs) have been previously studied in several works [36, 38, 39, 79]. A BF f with k inputs is *nested canalyzing* with respect to a permutation σ on its inputs $\{1, 2, \dots, k\}$ if:

$$f(\mathbf{x}) = \begin{cases} b_1 & \text{if } x_{\sigma(1)} = a_1, \\ b_2 & \text{if } x_{\sigma(1)} \neq a_1, x_{\sigma(2)} = a_2, \\ b_3 & \text{if } x_{\sigma(1)} \neq a_1, x_{\sigma(2)} \neq a_2, x_{\sigma(3)} = a_3, \\ \vdots & \\ b_k & \text{if } x_{\sigma(1)} \neq a_1, x_{\sigma(2)} \neq a_2, \dots, x_{\sigma(k)} = a_k, \\ \bar{b}_k & \text{if } x_{\sigma(1)} \neq a_1, x_{\sigma(2)} \neq a_2, \dots, x_{\sigma(k)} = \bar{a}_k. \end{cases} \quad (2.5)$$

In the above equation, a_1, a_2, \dots, a_k are the canalyzing input values and b_1, b_2, \dots, b_k are the canalyzed output values for input variables $x_{\sigma(1)}, x_{\sigma(2)}, \dots, x_{\sigma(k)}$ in the permutation σ of the k inputs. Here, \bar{a}_k and \bar{b}_k are the complements of the Boolean values a_k and b_k , respectively. We remark that Szallasi and Liang [79] had called these functions *hierarchically canalyzing* and subsequently, Kauffman [36] called them *nested canalyzing*.

As an example of a NCF, consider the Boolean expression for the formation of the heteromeric complex ErbB13 [32] presented in Chapter 1, Section 1.2 - “(EGF AND ErbB1 AND ErbB3 AND NOT ErbB2) OR (NRG1 AND ErbB1 AND ErbB3 AND NOT ErbB2)”. Applying the laws of Boolean algebra, this expression simplifies to “ErbB1 AND ErbB3 AND NOT ErbB2 AND (EGF OR NRG1)”. Here, ErbB1 and ErbB3 and ErbB2 are canalyzing inputs since the ErbB13 complex cannot be formed either in the absence of ErbB1 or in the absence of ErbB3 or in the presence of ErbB2. Thus, ErbB1 has a canalyzing input value of 0, similarly ErbB3 has a canalyzing input value of 0 and ErbB2 has a canalyzing input value of 1. Now, in the presence of ErbB1 and ErbB3 and in the absence of ErbB2, EGF and NRG1 act as canalyzing inputs since the presence of either

of these protein is sufficient to lead to the formation of the heteromeric complex ErbB13. This example illustrates how nested canalization may be achieved mechanistically.

2.2.5 Read-once functions

A BF of k variables is a *read-once function* (RoF) if it can be represented by a Boolean expression, using the operations of conjunction, disjunction and negation, in which every variable appears exactly once [84]. RoFs are also known as fanout-free functions in the computer science literature [85]. Mathematically, a k -input BF f is a RoF if, after stripping of all parentheses, there exists a permutation σ on $\{1, 2, \dots, k\}$ such that

$$f(\mathbf{x}) = X_{\sigma(1)} \odot X_{\sigma(2)} \odot X_{\sigma(3)} \dots \odot X_{\sigma(k)} \quad (2.6)$$

where $X_{\sigma(i)} \in \{x_{\sigma(i)}, \bar{x}_{\sigma(i)}\}$ and $\odot \in \{\wedge, \vee\}$. There are no restrictions on the placement of the parentheses between the variables. Here \wedge and \vee are the AND and OR operators respectively. For instance, the expressions for the RoFs, $x_1 \wedge x_2 \wedge (x_3 \vee x_4)$ and $x_1 \wedge (x_2 \vee x_3 \vee x_4)$, have the same 4 variables but different placement of parentheses and different positions of the AND or OR operators.

As an example of a RoF, consider again the Boolean expression for the formation of the heteromeric complex ErbB13 [32] presented in Chapter 1, Section 1.2 - “(EGF AND ErbB1 AND ErbB3 AND NOT ErbB2) OR (NRG1 AND ErbB1 AND ErbB3 AND NOT ErbB2)”. Applying the laws of Boolean algebra to this expression, we get the simplified expression “ErbB1 AND ErbB3 AND NOT ErbB2 AND (EGF OR NRG1)”. In this simplified expression, each literal (associated with each protein), appears exactly once. This feature of the BF defines the property of being read-once.

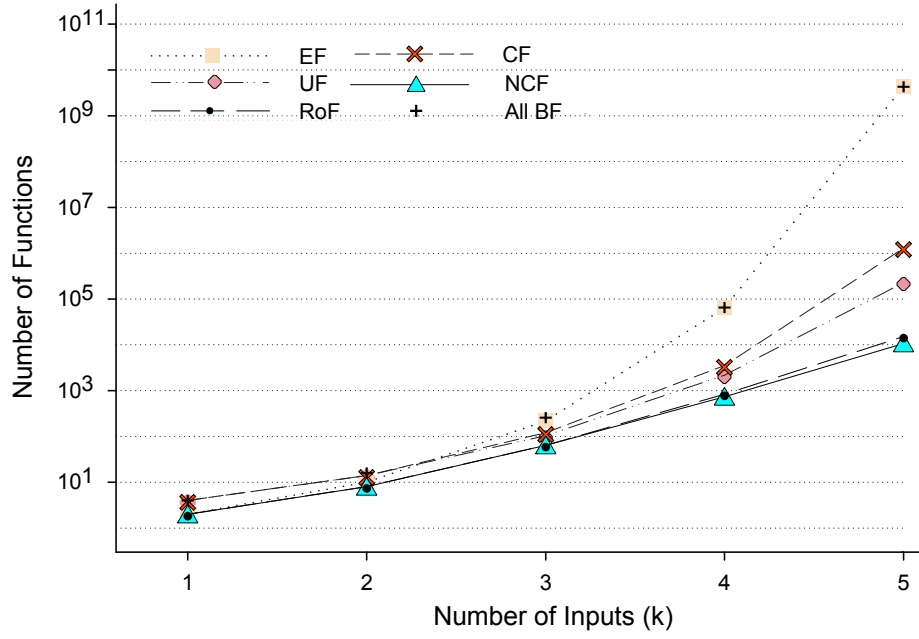


Figure 2.2: The number of biologically meaningful types of BF's for a given number of inputs $k \leq 5$.

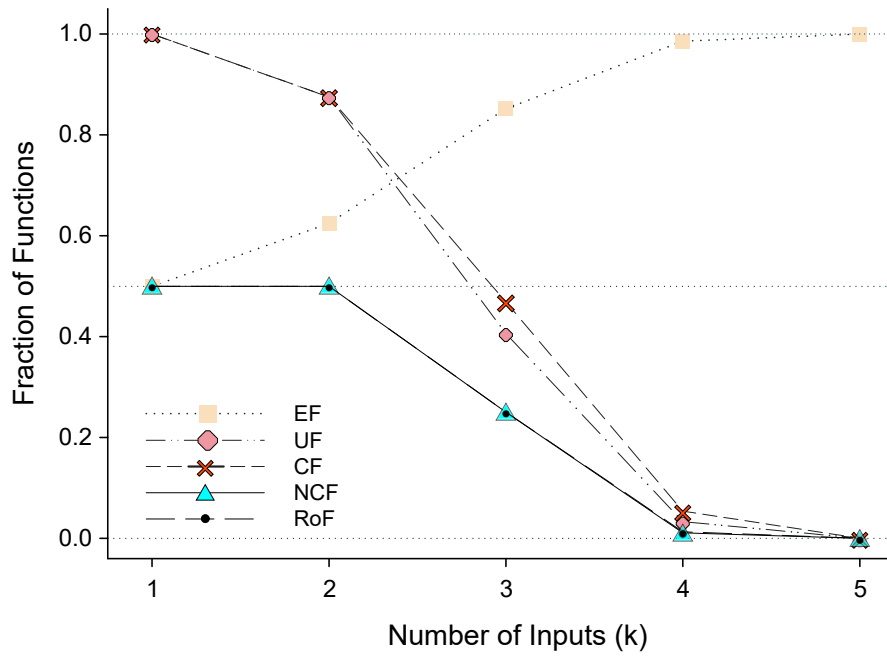


Figure 2.3: The fraction of biologically meaningful types of BF's among all BF's for a given number of inputs $k \leq 5$.

Table 2.1: The number of BFs belonging to the different types, at a given number of inputs $k \leq 5$. Here, EF corresponds to effective functions, UF to unate functions (all sign combinations), CF to canalyzing functions, EUF to effective and unate functions, ECF to effective and canalyzing functions, UCF to unate and canalyzing functions, EUCF to effective, unate and canalyzing functions. In addition, the table lists the total number of BFs for each k .

k	Types of BFs							
	All	EF	UF	CF	EUF	ECF	UCF	EUCF
1	4	2	4	4	2	2	4	2
2	16	10	14	14	8	8	14	8
3	256	218	104	120	72	88	96	64
4	65536	64594	2170	3514	1824	3104	1178	864
5	4294967296	4294642034	230540	1292276	220608	1275784	36796	31744

2.3 Characterizing the space of biologically meaningful BFs

Now we can systematically explore the relationships between the aforementioned types of biologically meaningful BFs. To the best of our knowledge, such a combined delineation of the different types of biologically meaningful BFs in the space of all 2^{2^k} BFs has not been carried out previously. Exhaustive enumeration of BFs for low values of k led us to conjecture some properties of these BFs for which we provide analytical proofs. The number of BFs in each of these types increases with increasing k as is expected (see Figure 2.2 and Table 2.1).

Computational enumeration up to $k \leq 5$, shows that the fraction of EFs in the space of all k -input BFs increases with increasing k (see Figure 2.3). In contrast, the fraction of UFs and CFs decreases with increasing k and tend to 0 (see Figure 2.3 and Table 2.2). The proportions of even bias functions (and consequently odd bias functions) within the sets EFs, UFs and CFs and also in their intersections at $k \leq 5$ appear to tend to 0.5 for increasing k (see Tables 2.3 and 2.4). Note that for a given number of inputs but various combinations of activators and inhibitors (see Figure 2.4), the proportion of even bias functions (and consequently odd bias) is constant for effective and unate functions (EUFs)

Table 2.2: The fraction of BFs belonging to the different types, at a given number of inputs $k \leq 5$. Here, EF corresponds to effective functions, UF to unate functions (all sign combinations), CF to canalyzing functions, EUF to effective and unate functions, ECF to effective and canalyzing functions, UCF to unate and canalyzing functions, EUCF to effective, unate and canalyzing functions.

k	Types of BFs						
	EF	UF	CF	EUF	ECF	UCF	EUCF
1	0.500	1.000	1.000	0.500	0.500	1.000	0.500
2	0.625	0.875	0.875	0.500	0.500	0.875	0.500
3	0.852	0.406	0.469	0.281	0.344	0.375	0.250
4	0.986	0.033	0.054	0.028	0.047	0.018	0.013
5	1.000	5.37×10^{-5}	3.01×10^{-4}	5.14×10^{-5}	2.97×10^{-4}	8.57×10^{-6}	7.39×10^{-6}

(see Table 2.5). Note that the bias is identical to the parity of the BF, and is consistent with the standard notion of parity used in computer science literature [80]. However, it is important to note that parity bits, bits that are appended to a binary string to obtain an even or odd parity string and are frequently used in telecommunication systems for error-detection in transmitted information, are not used in this thesis. Furthermore, computational enumeration up to $k \leq 10$, shows that the fraction of RoFs, NCFs and non-NCF RoFs among all BFs with k inputs decreases and tends to 0 with increasing k (see Figure 2.3 and Table 2.6). We also find that the fraction of NCFs that are RoFs decreases with increasing number of inputs (see Table 2.6). It is also feasible to perform such enumerations separately for the different possible values of the bias P . In Figure 2.5, we show the distribution of RoFs, NCFs and non-NCF RoFs for different biases, for the inputs $k = 4, 5, 6, 7$ and 8.

Figure 2.6 gives an overview of the space of biologically meaningful BFs across all 4-input BFs and serves as a visual guide to the overlaps between the different types of BFs. The space of all BFs can be divided into two equal parts based on the parity (odd and even) of the bias. Interestingly, all *ineffective functions* (IEFs) (BFs with at least one ineffective input) lie in the even bias half. This raises the question as to whether all IEFs have even bias. We theoretically prove that this is indeed the case (see Section 2.4). The

Table 2.3: Parity distribution (number) of biologically meaningful BFs for a given number of inputs k . The number of even parity functions of a particular type of BF is calculated. Here, EF corresponds to effective functions, UF to unate functions (all sign combinations), CF to canalizing functions, EUF to effective and unate functions, ECF to effective and canalizing functions, UCF to unate and canalizing functions, EUCF to effective, unate and canalizing functions. Note that entries labeled “-” are those which could not be computed due to inadequate computational resources.

k	Number of even parity functions							
	All	EF	UF	CF	EUF	ECF	UCF	EUCF
1	2	0	2	2	0	0	2	0
2	8	2	6	6	0	0	6	0
3	128	90	40	56	8	24	32	0
4	32768	31826	922	1754	576	1344	442	128
5	2147483648	—	110348	646132	100416	629640	15932	10880

UFs, which allow for all possible numbers of activators and inhibitors, are rather evenly distributed across even and odd biases and have some overlap with the IEF set (see Figure 2.6). Indeed, not all UFs are EFs (see Section 2.4). The CFs, like the UFs, are almost equally distributed across even and odd biases and overlap with the IEFs, EFs and UFs (see Figure 2.6).

Next, NCFs and RoFs lie in the odd bias half (see Figure 2.6). This warrants the conjecture that all NCFs and RoFs have odd bias, and we show that this is indeed the case (see Section 2.4). Moving to the NCFs, we see in Figure 2.6 that NCFs lie within the space of RoFs (see Section 2.4). NCFs are also a strict subset of RoFs when $k \geq 4$.

The following section provides theoretical proof of several computational observations mentioned in this section. Properties 2.4.1 - 2.4.3 pertain to combining two independent BFs, Property 2.4.4 pertains to EFs, Properties 2.4.5 - 2.4.7 pertain to UFs, Properties 2.4.8 - 2.4.10 pertain to NCFs and Properties 2.4.11 - 2.4.16 pertain to RoFs.

Table 2.4: Parity distribution (fraction) of biologically meaningful BFs for a given number of inputs k . The fraction of even parity functions of a particular type of BF is calculated with respect to the total number of functions of that BF type. Here, EF corresponds to effective functions, UF to unate functions (all sign combinations), CF to canalizing functions, EUF to effective and unate functions, ECF to effective and canalizing functions, UCF to unate and canalizing functions, EUCF to effective, unate and canalizing functions. Note that entries labeled “-” are those which could not be computed due to inadequate computational resources.

k	Fraction of even parity functions							
	All	EF	UF	CF	EUF	ECF	UCF	EUCF
1	0.5	0	0.5	0.5	0	0	0.5	0
2	0.5	0.2	0.429	0.429	0	0	0.429	0
3	0.5	0.413	0.385	0.467	0.111	0.273	0.333	0
4	0.5	0.493	0.425	0.499	0.316	0.433	0.375	0.148
5	0.5	—	0.479	0.5	0.455	0.494	0.433	0.343

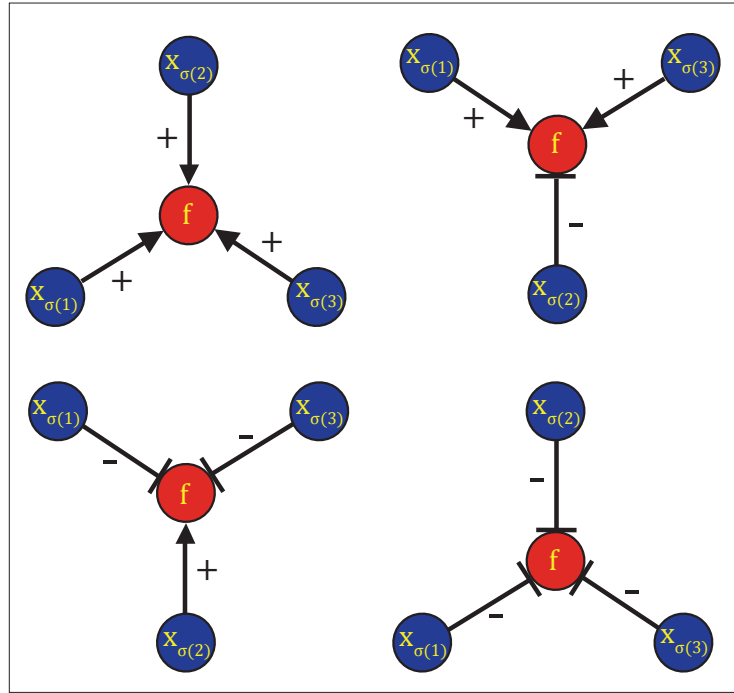


Figure 2.4: Schematic figure depicting the four possible *sign* combinations for the $k = 3$ inputs to a node (gene) f . Each input or regulatory interaction to the node f can be an activator or inhibitor and are shown as + or -, respectively. The 3 inputs to the node f are labeled by variables $x_{\sigma(1)}$, $x_{\sigma(2)}$ or $x_{\sigma(3)}$, without repetition of any label. σ is the permutation of the set $\{1, 2, 3\}$, and $\sigma(i)$ represents the i^{th} element of a given permutation.

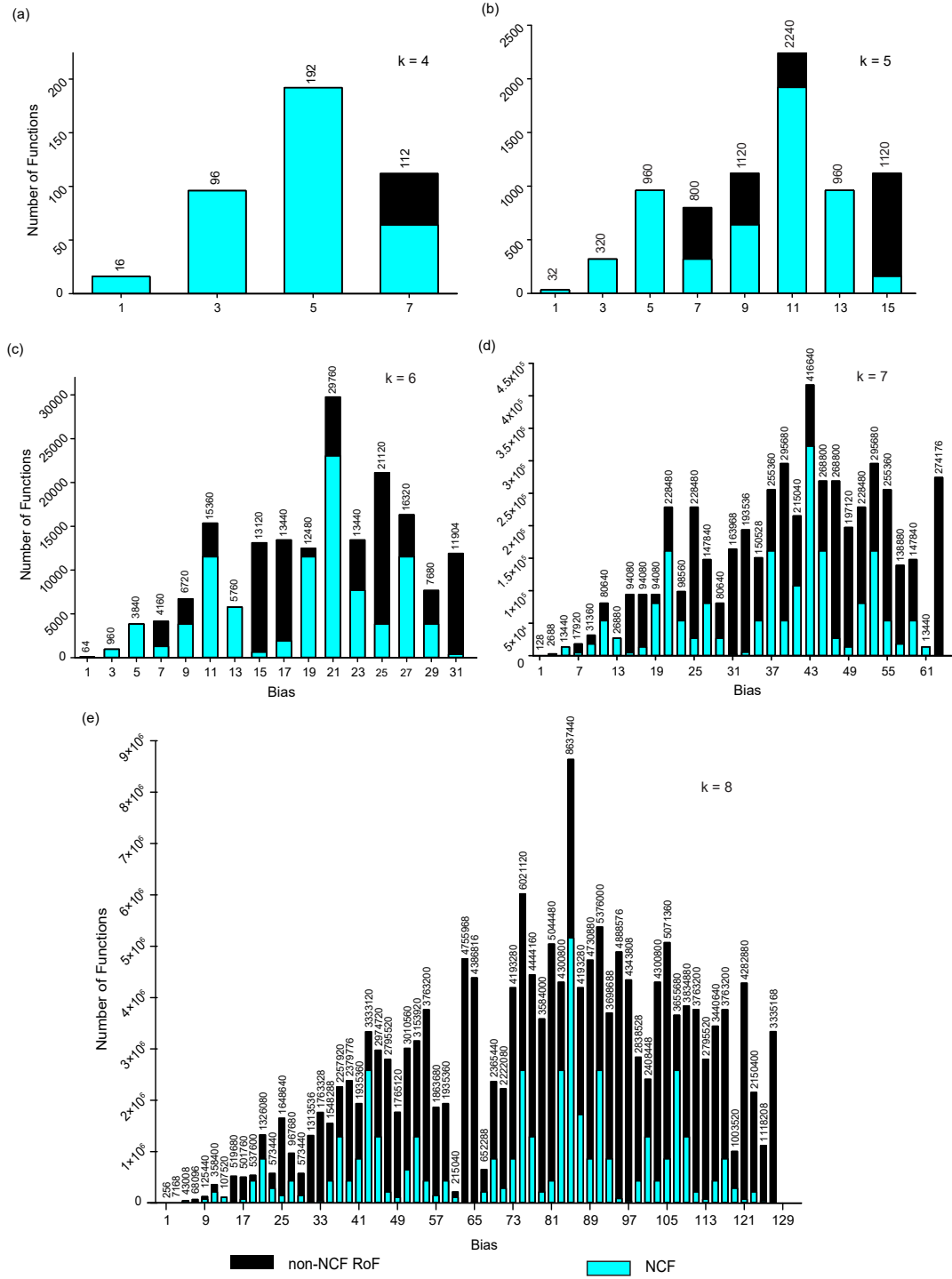


Figure 2.5: Frequency distribution of read-once functions (RoFs) across different bias P for functions with (a) $k = 4$, (b) $k = 5$, (c) $k = 6$, (d) $k = 7$, and (e) $k = 8$ inputs. For each bar, we display the number of RoFs with that bias value. The figure also gives the frequency distribution of nested canceling functions (NCFs), which are a subset of RoFs. Due to the *complementarity* property of BFs, the distribution is symmetric about the bias value 2^{k-1} for a given k , and therefore, we display only the first half of the distribution, from 0 to 2^{k-1} in each case.

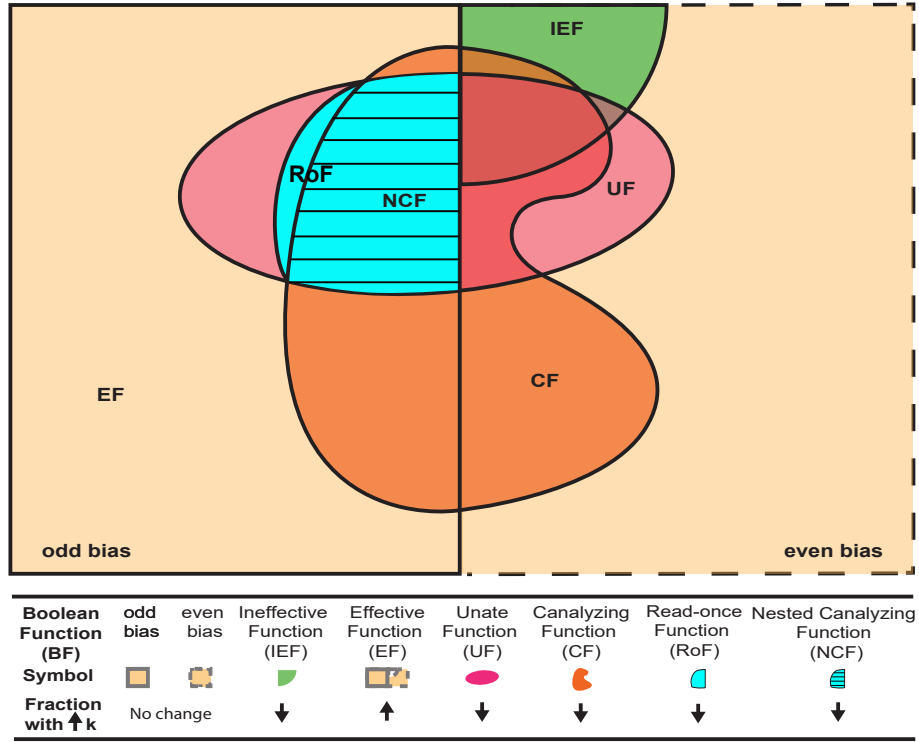


Figure 2.6: A schematic of the overlaps between different types of biologically meaningful BFs in the space of all 4-input BFs. This figure is not drawn to scale but the sizes of the sets corresponding to different types of BFs and their intersections respect the order of the actual values. The legend gives the correspondence between shapes with specific color and the different types of BFs. Ordering the different types of BFs with 4 inputs (which are *not* mutually exclusive) based on their sizes in a descending order gives: $EF > \text{Odd bias} = \text{Even bias} > CF > UF > RoF > NCF$. The up (or down) arrows in the legend depict the increase (or decrease) in the fraction of BFs that belong to a specific type as k increases (see Table 2.2 for the exact numbers).

Table 2.5: The number of EUFs at a given number of inputs $k \leq 5$ for different combinations of activators and inhibitors. The table also gives the fraction of EUFs that have even bias for different sign combinations. These numbers have been obtained via exhaustive computational enumeration of EUFs.

k	Activators	Inhibitors	EUFs		
			Total	Even bias	Fraction with Even bias
1	1	0	1	0	0
	0	1	1	0	0
2	2	0	2	0	0
	1	1	4	0	0
	0	2	2	0	0
3	3	0	9	1	0.111
	2	1	27	3	0.111
	1	2	27	3	0.111
	0	3	9	1	0.111
4	4	0	114	36	0.316
	3	1	456	144	0.316
	2	2	684	216	0.316
	1	3	456	144	0.316
	0	4	114	36	0.316
5	5	0	6894	3138	0.455
	4	1	34470	15690	0.455
	3	2	68940	31380	0.455
	2	3	68940	31380	0.455
	1	4	34470	15690	0.455
	0	5	6894	3138	0.455

2.4 Theoretical results on the properties of biologically meaningful types of BFs

In this section, we present the properties for EFs, UF, NCFs and RoFs, some of which were stated in the previous section as arising from computational observations. These properties have several implications in generating these different types of BFs and designing checks to find out if a BF belongs to any of the biologically meaningful types of BFs.

Table 2.6: The number and fraction of NCFs and RoFs among all BFs and fraction of NCFs within RoFs for a given number of inputs k . This table also lists the fraction of NCFs among RoFs for a given number of inputs.

k	Number of		Fraction of		NCFs/RoFs
	NCFs	RoFs	NCFs	RoFs	
1	2	2	0.500	0.500	1.000
2	8	8	0.500	0.500	1.000
3	64	64	0.250	0.250	1.000
4	736	832	0.011	0.013	0.885
5	10624	15104	2.47×10^{-6}	3.52×10^{-6}	0.703
6	183936	352256	9.97×10^{-15}	1.91×10^{-14}	0.522
7	3715072	10037248	1.09×10^{-32}	2.95×10^{-32}	0.370
8	85755372	337936384	7.41×10^{-70}	2.92×10^{-69}	0.254
9	2226939904	13126565888	1.66×10^{-145}	9.79×10^{-145}	0.170
10	64255903744	577818263552	3.57×10^{-298}	3.21×10^{-297}	0.111

2.4.1 Combining two independent BFs

Consider two *independent* BFs f_1 and f_2 with k_1 and k_2 inputs and bias P_1 and P_2 , respectively. Here, the two BFs are independent in the sense that they have no input variables in common. The truth tables for the BFs f_1 and f_2 have 2^{k_1} and 2^{k_2} rows, respectively. A simple way to combine the two BFs is via the AND or OR logical operators. We use the notation $f = f_1 \odot f_2$ where \odot is either the AND (\wedge) or OR (\vee) operator. The procedure to generate f with $2^{k_1+k_2}$ rows in its truth table, by combining f_1 and f_2 , can be expressed compactly as follows:

Algorithm 1 Algorithm to combine two independent BFs f_1 and f_2

```

1:  $f[r] = 0, r \in [1, 2^{k_1+k_2}]$ 
2:  $r \leftarrow 1$ 
3: for  $i \leftarrow 1$  to  $2^{k_1}$  do
4:   for  $j \leftarrow 1$  to  $2^{k_2}$  do
5:      $f[r] = f_1[i] \odot f_2[j]$ 
6:      $r \leftarrow r + 1$ 
7:   end for
8: end for

```

In Line 1 of the above algorithm, we initialize a vector f with $2^{k_1+k_2}$ elements to store the output values in each row of the truth table for $f = f_1 \odot f_2$. In Line 5 of the algorithm,

the output value for the i^{th} row of f_1 is combined with that of the j^{th} row of f_2 to give the output value for the r^{th} row of f . For example, if BFs f_1 and f_2 with 1 input and 2 inputs, respectively, have output vectors $(1, 0)$ and $(1, 1, 1, 0)$, respectively, then the output vector for the combined BF $f = f_1 \wedge f_2$ with 3 inputs is $(1, 1, 1, 0, 0, 0, 0, 0)$.

Property 2.4.1. Let f_1 and f_2 have bias P_1 and P_2 . Then $f_1 \wedge f_2$ (hereafter denoted as f_{AND}) has bias equal to $P_1 P_2$.

Proof: For every occurrence of 0 in the output vector of f_1 , the output vector of f_{AND} will also be 0. For every occurrence of 1 in the output vector of f_1 , there will be P_2 occurrences of 1 in the output vector of f_{AND} . Thus, for P_1 occurrences of 1 in the output vector of f_1 , there will be $P_1 P_2$ occurrences of 1 in the output vector of f_{AND} .

Property 2.4.2. Let us denote $f_1 \vee f_2$ by f_{OR} , then f_{OR} has bias equal to $2^{k_1} P_2 + 2^{k_2} P_1 - P_1 P_2$.

Proof: For every occurrence of 0 in the output vector of f_1 , there will be P_2 occurrences of 1 in the output vector of f_{OR} . Thus, the contribution to the 1's in the output vector of f_{OR} from 0's in the output vector of f_1 is $(2^{k_1} - P_1)P_2$. For every occurrence of 1 in the output vector of f_1 , there will be 2^{k_2} occurrences of 1's in the output vector of f_{OR} . Thus, the contribution to the 1's in the output vector of f_{OR} from 1's in the output vector of f_1 is $2^{k_2} P_1$. In total, the number of 1's in the output vector of f_{OR} is equal to $(2^{k_1} - P_1)P_2 + 2^{k_2} P_1 = 2^{k_1} P_2 + 2^{k_2} P_1 - P_1 P_2$.

Property 2.4.3. Given the bias parities (even or odd) of f_1 and f_2 , the two previous results show that both f_{AND} and f_{OR} have *odd* parity (i.e., their bias is odd), if and only if P_1 and P_2 are both odd.

2.4.2 Effective functions

Property 2.4.4. The bias P of a BF with m ineffective inputs is a multiple of 2^m .

Proof: Consider the truth table of a BF f and assume the input variable x_i is ineffective. Then each row with $x_i = 0$ can be uniquely paired with the corresponding row having $x_i = 1$ where all other variables are unchanged. Since the output is the same in both of

these lines, one has either no 1s or two 1s in the output. Summing over all of the truth table (rows coming in pairs) will thus lead to an even number of 1s. The result for m ineffective inputs is then obtained by recurrence.

Corollary: It immediately follows that a BF with an odd bias P is effective.

2.4.3 Unate functions

Property 2.4.5. A UF can be represented by an expression in disjunctive normal form (DNF) in which all occurrences of any specific input variable (more precisely, literal) are either negated (i.e., negative input) or non-negated (i.e., positive input) [34, 82].

Property 2.4.6. If u_1 and u_2 are UFs with k_1 and k_2 independent input variables, respectively, then the combined BF $u = u_1 \odot u_2$ is also unate.

Proof: Consider two UFs u_1 and u_2 . For convenience, let us denote their DNF expressions by the same symbols u_1 and u_2 . Since each input variable in u_1 and u_2 is either a positive (x_i) or a negative (\bar{x}_i) literal (see Property 2.4.5), the combined expression $u = u_1 \odot u_2$ composed of $(k_1 + k_2)$ distinct input variables (due to independence) will also have each variable occur as only its positive or negative literal. This implies that the combined function u is a UF. As an example, consider the UFs $u_1 = (x_1 \vee \bar{x}_2)$ and $u_2 = (x_3 \wedge \bar{x}_4) \vee x_5$. The combined BF $u = u_1 \odot u_2$ under the AND operation is simply $u = (x_1 \vee \bar{x}_2) \wedge ((x_3 \wedge \bar{x}_4) \vee x_5)$. Since each literal appears in u only in its positive or its negative form in the expression for u , the combined BF is UF.

Property 2.4.7. If an input i of a UF u acts as both an activator and an inhibitor, then input i is ineffective.

Proof: An input i acts as both an activator and an inhibitor if and only if the input i satisfies the equality condition in Eqs. (2.2) and (2.3), respectively, for all pairs of rows for which the all other input variables $j \neq i$ are kept fixed. However, this is precisely equivalent to the condition for an input i to be ineffective.

2.4.4 Nested canalyzing functions

NCFs can be expressed as a Boolean expression as given by the equation:

$$f(\mathbf{x}) = X_{\sigma(1)} \odot (X_{\sigma(2)} \odot (X_{\sigma(3)} \odot \dots (X_{\sigma(k-1)} \odot X_{\sigma(k)}))) \quad (2.7)$$

where σ is a permutation on the inputs $\{1, 2, \dots, k\}$, $X_{\sigma(i)} \in \{x_{\sigma(i)}, \bar{x}_{\sigma(i)}\}$ and $\odot \in \{\wedge, \vee\}$.

Also, a k -input NCF has a canalyzing depth of k .

Property 2.4.8. NCFs have odd bias.

Proof: Consider the base case of a NCF with 1 input with the representative expression $NCF(1) = x_1$. Clearly, the Boolean expressions $f = x_1$ and $f = \bar{x}_1$ refer to the BFs with output vector $(0, 1)$ or $(1, 0)$, both of which have odd bias. Next, let us hypothesize that all NCFs with k inputs, i.e., $NCF(k)$, have odd bias P . We can then proceed by induction. A NCF with $k+1$ inputs is given by $NCF(k+1) = x_{k+1} \odot NCF(k)$ by definition (see Eq. (2.7)). Since the two independent BFs x_{k+1} and $NCF(k)$ have odd bias, using Property 2.4.3, the combined BF $NCF(k+1)$ will also have an odd bias. Note that a different proof for this property of NCFs was provided by Nikolajewa *et al.* [86].

Property 2.4.9. NCFs are EFs.

Proof: Since BFs with odd bias are EFs, using Properties 2.4.4 and 2.4.8, NCFs are also EFs.

Property 2.4.10. NCFs are UFs [34].

Proof: Following Aracena [34], since each variable or literal in the expression for a NCF (see Eq. (2.7)) appears exactly once, it follows that each variable is fixed to either its positive or negative form in the function's canonical NCF form. Thus, NCFs are UFs using Property 2.4.5.

2.4.5 Read-once functions

Property 2.4.11. Generation of representative RoFs

The following is a recursive scheme to generate all RoFs with k inputs, i.e., $RoF(k)$,

starting from RoFs with 1 input ($RoF(1)$). To do so, one can use the fact that the parentheses in the logical expression of a function in $RoF(k)$ define two sub-parts separated by an AND or an OR operator. Such a decomposition splits the k variables into two sets, and thus, any function in $RoF(k)$ can be decomposed into at least one of the following types:

$$\begin{aligned}
& RoF(k-1) \odot RoF(1) \\
& RoF(k-2) \odot RoF(2) \\
& RoF(k-3) \odot RoF(3) \\
& \vdots \\
& RoF(k - (k/2)) \odot RoF(k - (k/2)) \\
& \text{[for } k \text{ even]} \\
& \text{or} \\
& RoF(k - ((k-1)/2)) \odot RoF(k - ((k+1)/2)) \\
& \text{[for } k \text{ odd]}
\end{aligned}$$

where \odot corresponds to the AND (\wedge) or OR (\vee) operator. Such a decomposition allows one to enumerate all elements of $RoF(k)$ recursively.

The above algorithm does not only return the representative RoFs, sometimes it will return permutations thereof. To retain only the representative RoFs, we iteratively walk through the produced list and keep an element only if it is not equivalent to a previous element under permutation of the variables.

Property 2.4.12. RoFs have odd bias.

Proof: Consider the base case of a RoF with 1 input. Clearly, the BFs in $RoF(1)$ have output vector $(0, 1)$ or $(1, 0)$, both of which have odd bias. Next, let us hypothesize that BFs in $RoF(j) \forall j \in \{1, k\}$ have odd bias. We now refer to Property 2.4.3, whereby the combination of two BFs with odd bias results in a BF with odd bias. Next by induction, the RoF with $k+1$ inputs is given by $RoF(k+1) = RoF((k+1) - j) \odot RoF(j)$ for all

$j \in [1, (k+1)/2]$ for odd k , or $j \in [1, k/2]$ for even k (see Property 2.4.11). Since the two functions $RoF((k+1)-j)$ and $RoF(j)$ have odd bias, using Property 2.4.3, the function $RoF(k+1)$ will also have odd bias.

Property 2.4.13. RoFs are EFs.

Proof: From Property 2.4.12, RoFs have odd bias. From Property 2.4.4, BFs with odd bias are EFs. Thus, RoFs are EFs.

Property 2.4.14. RoFs are UFs.

Proof: Since each variable or literal in the expression for a RoF (see Eq. (2.6)) appears exactly once, it follows that each variable is fixed to either its positive or negative form in the RoF logical expression. Thus, RoFs are UFs according to Property 2.4.5.

Property 2.4.15. For any k , NCFs are a subset of RoFs.

Proof: Comparing the expression for NCFs (Eq. (2.7)) with the expression for RoFs (Eq. (2.6)), it is evident that NCFs form a subset of RoFs. Simply stated, all NCFs are RoFs but all RoFs need not be NCFs. Henceforth, we refer to the subset of the RoFs which are not NCFs as the *non-NCF RoFs*. To the best of our knowledge, RoFs (excluding NCFs) have not been considered in the biological literature.

Property 2.4.16. RoFs with bias P equal to 1, 3 and 5 are NCFs.

Proof: First consider the case where the bias P is 1. The DNF of a BF with k inputs and bias P equal to 1 has just one term, the conjunction of k literals, and thus, the function is a NCF (see Eq. (2.7)).

Next, we show that it is impossible to have a RoF with $k > 2$ (respectively, $k > 3$) and bias $P = 3$ (respectively $P = 5$) by combining RoFs with the OR operator. Let P_{OR} be the bias of $RoF_{OR}(k) = RoF(k_1) \vee RoF(k_2)$, where $k = k_1 + k_2$. Further, let P_1 and P_2 be the biases of $RoF(k_1)$ and $RoF(k_2)$, respectively. From Property 2.4.2, we have $P_{OR} = 2^{k_1}P_2 + 2^{k_2}P_1 - P_1P_2$, which is a positive monotonic function of P_1 and P_2 (for a fixed k_1 and k_2). The minimum value of P_{OR} is thus obtained at $P_1 = 1, P_2 = 1$. Thus, $\min(P_{OR}) = 2^{k_1} + 2^{k_2} - 1$. If k is both greater than 2 (or 3), it can be easily confirmed

that $\min(P_{\text{OR}}) > 3$ (respectively, $\min(P_{\text{OR}}) > 5$). Thus, the bias of RoF_{OR} for $k > 2$ (respectively, $k > 3$) cannot be 3 (respectively, 5).

Thus, it follows that a RoF with $k > 2$ (respectively, $k > 3$) and bias 3 (respectively, 5) can be generated only by combining RoFs with the AND operator. Let P_{AND} be the bias of $\text{RoF}_{\text{AND}}(k) = \text{RoF}(k_1) \wedge \text{RoF}(k_2)$, where $k = k_1 + k_2$. From Property 2.4.1, $P_{\text{AND}} = P_1 P_2$. Since 3 (respectively, 5) is prime, a $\text{RoF}_{\text{AND}}(k)$ with bias 3 (respectively, 5) can be generated only by combining two RoFs, $\text{RoF}(k_1)$ and $\text{RoF}(k_2)$, with biases 1 and 3 (respectively, 1 and 5). Let $\text{RoF}(k_2)$ have bias 3 (respectively, 5). By decomposition, $\text{RoF}(k_2)$ would in turn have to be generated by combining two RoFs, $\text{RoF}(k_{2,1})$ and $\text{RoF}(k_{2,2})$, with biases 1 and 3 (respectively, 1 and 5), and so on. Proceeding in this manner, we will be left with a *nested* RoF, with exactly one term having bias 3 (respectively, 5), and all other RoFs in the *nested* expression having bias 1. In other words, for bias 3 *nested* RoF would be of the form: $x_1 \wedge x_2 \wedge x_3 \wedge \dots \wedge x_{k-2} \wedge (x_{k-1} \vee x_k)$, and for bias 5 would be of the form: $x_1 \wedge x_2 \wedge x_3 \wedge \dots \wedge x_{k-3} \wedge (x_{k-2} \vee (x_{k-1} \wedge x_k))$, both of which are NCFs (see Eq. (2.7)).

Note that for $k = 1$, there are no BFs with bias P equal to 3. To complete the proof, consider the cases $k = 2$ and $k = 3$. For $k = 2$, RoFs with bias $P = 1$ are NCFs, hence its complement (with bias $P = 3$), is also a NCF. For $k = 3$, RoFs with bias $P = 5$ are NCFs since it is the complement of bias $P = 3$ which we showed to be NCFs for all values of k .

For the sake of compactness, we represent RoFs which are equivalent up to isomorphisms (i.e., permutations of indices and complementation of input variables) via a single representative BF or expression. Furthermore, we classify RoFs into $k[P]$ sets based on the number of inputs k and bias P [69]. In other words, we capture the complete set of RoFs in different $k[P]$ sets via representative RoFs wherein each representative RoF captures all RoFs that are equivalent up to isomorphisms. For example, among BFs with $k = 4$ inputs, there are 10 representative RoFs up to isomorphisms which are:

RoF	Expression	$k[P]$
f_1	$x_1 \wedge x_2 \wedge x_3 \wedge x_4$	4[1]
f_2	$x_1 \wedge x_2 \wedge (x_3 \vee x_4)$	4[3]
f_3	$x_1 \wedge (x_2 \vee (x_3 \wedge x_4))$	4[5]
f_4	$x_1 \wedge (x_2 \vee x_3 \vee x_4)$	4[7]
f_5	$(x_1 \wedge x_2) \vee (x_3 \wedge x_4)$	4[7]
f_6	$x_1 \vee (x_2 \wedge x_3 \wedge x_4)$	4[9]
f_7	$(x_1 \vee x_2) \wedge (x_3 \vee x_4)$	4[9]
f_8	$x_1 \vee (x_2 \wedge (x_3 \vee x_4))$	4[11]
f_9	$x_1 \vee x_2 \vee (x_3 \wedge x_4)$	4[13]
f_{10}	$x_1 \vee x_2 \vee x_3 \vee x_4$	4[15]

Among the above-mentioned 10 RoFs with $k = 4$, $f_1, f_2, f_3, f_4, f_6, f_8, f_9$ and f_{10} are also NCFs.

RoF checker

To check whether a BF is a RoF, we make use of the various properties of RoFs. To begin with, we generate a representative RoF for each equivalence class, going up to $k = 10$ inputs using the Property 2.4.11. We store the truth table, bias and the average sensitivity of each representative RoF in computer memory so that it can be used as a lookup table. Next, we implement the procedure shown in the flowchart (see Figure 2.7). This program takes as input a BF via its truth table representation; the bias of the BF is determined. The program then proceeds by performing successive tests, from quite simple to more complex, as follows. If the bias is even then the BF is not a RoF (see Property 2.4.12). Since NCFs are a subset of RoFs, we check whether the BF is a NCF as that is relatively simple computationally (just successively determine the canalyzing input variables). If the function is not a NCF, we check whether the input BF is a UF since RoFs are unate. If the BF is unate, we calculate average sensitivity. Then we use the lookup table to extract all of the representative RoFs having that bias and average sensitivity. Recall that all elements in an equivalence class have the same bias and average sensitivity. In case no representative RoF matches, then the BF is not a RoF. Assuming that there

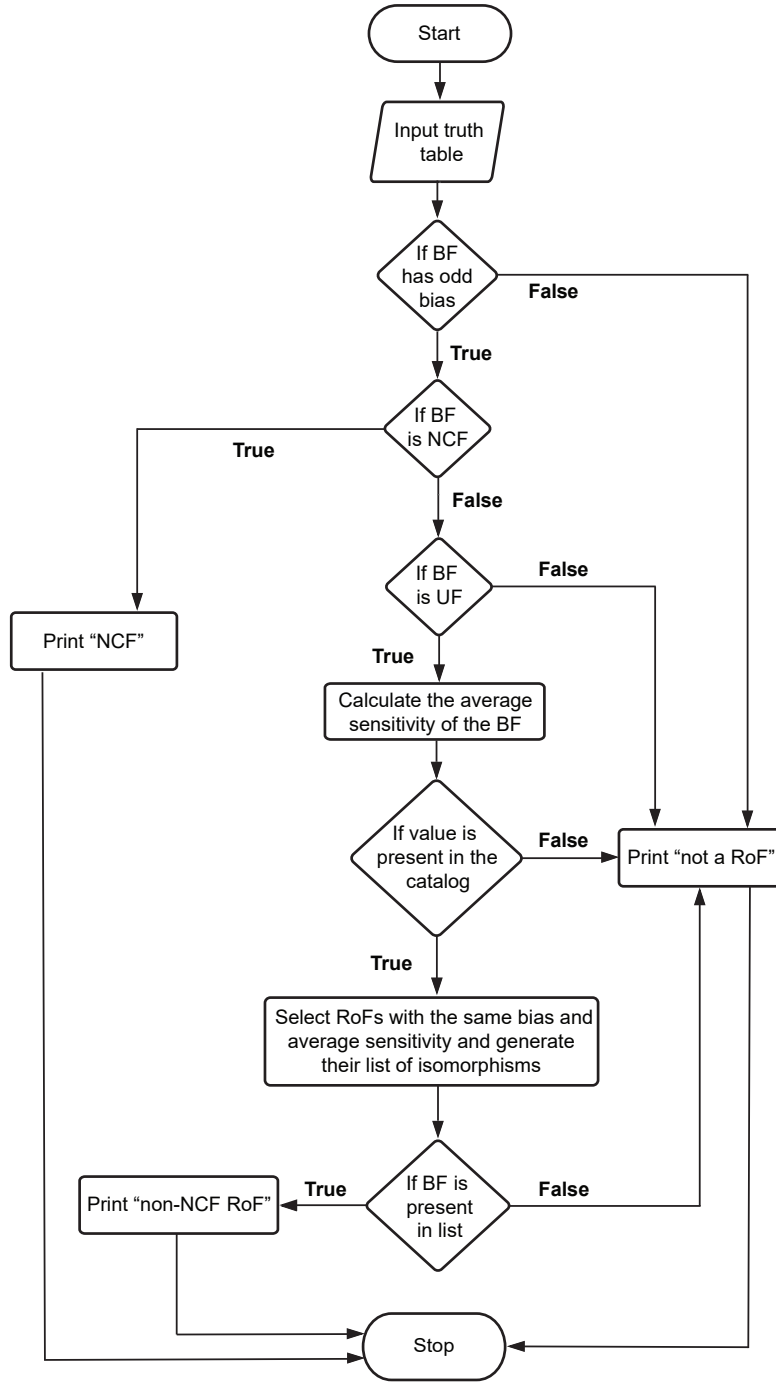


Figure 2.7: Flowchart describing our program to check whether a BF with $k \leq 10$ inputs entered by a user, is a read-once function (RoF). The program can also distinguish between a NCF and a non-NCF RoF.

is at least one representative RoF extracted, the program then loops over that list. For each such RoF we generate all RoFs belonging to that same equivalence class (just loop

over all isomorphisms), and for each such function the program directly checks whether its truth table is the same as that of the input BF. If an equality is found, the input BF is a RoF and we are done. If all the representative RoFs are tested without any success, then the input BF is not a RoF. The catalog of RoFs along with the python code to check for RoFs is available via the GitHub repository: <https://github.com/asamallab/MCBF>.

2.5 Discussion

In this chapter, we characterized the space of different biologically meaningful BFs and how they overlap with each other. We first explore several representations of the BF such as truth table, Boolean expression and the Boolean hypercube. Next, we present Feldman’s [70] classification scheme based on the number of inputs and the bias of BFs. In that classification, several BFs that are isomorphic can be clubbed into a compact representative BF. We term such a set of BFs as a $k[P]$ set, where k is the number of inputs to a BF and P is its *unnormalized* bias. Shifting the focus to the biologically meaningful types, we find that except for the EFs, the fraction of all of the biologically meaningful types (in the space of all BFs) decreases with an increasing number of inputs implying that effectiveness is a feature of a random BF and other properties such as unateness and canalyzation are not. Interestingly, the BF that have at least one ineffective input always have an even bias and a BF that has an odd bias is always an EF (see Property 2.4.4). Furthermore, all UFs are not EFs (see Property 2.4.7). This in fact naturally leads to another type of biologically meaningful BF, namely, the EUFs. Computationally, we find that the fraction of even and odd bias BFs is unequal for low input BFs for different types (except for RoFs and NCFs), and tends to half as the number of inputs increases. Following this, we show that both NCFs and RoFs are subsets of EUFs, NCFs are a subset of RoFs and both NCFs and RoFs always have odd bias (see Properties 2.4.8 and 2.4.12). Lastly, we provide a method to generate RoFs for an arbitrary number of inputs and an algorithm to check if a given BF is a RoF by leveraging the various properties of RoFs. These observations and simple properties can serve as powerful checks and balances in designing and testing algorithms to generate different types of biologically meaningful

BFs.

Data and code availability statement

All the codes necessary to generate the different types of biologically meaningful BFs and check whether a given BF belongs to any of the biologically meaningful types is provided in the GitHub repository: <https://github.com/asamallab/MCBF>.

Chapter 3

Minimum complexity drives regulatory logic in Boolean models of living systems

As alluded to in the previous chapter, hardly any previous efforts have been devoted towards a systematic understanding of the extent to which Boolean functions (BFs) encoding the associated regulatory logic rule in gene regulatory networks (GRNs) are far from *random*. Over the past two decades, Boolean dynamical models that capture a wide range of biological processes in several species have been reconstructed using literature based evidence. Some of these models include the development of roots in *Arabidopsis thaliana* [87], gene expression patterns of segment polarity genes in *Drosophila melanogaster* [26], the cell cycle transcriptional network in *fission yeast* [27], the cell cycle transcriptional network in mammalian systems [88] among several others [29, 45, 89, 90]. This has led to efforts toward building repositories of reconstructed Boolean GRNs such as Cell Collective [45] and GINSIM [91], which can now be exploited to answer the questions pertaining to the nature of regulatory logic rules employed in reconstructed Boolean models of living systems.

In this chapter, we systematically examine the preponderance of the different types of

biologically meaningful BFs (as defined in Chapter 2), namely, effective functions (EFs), unate functions (UFs), canalizing functions (CFs) and nested canalizing functions (NCFs) in a reference biological dataset of 2687 BFs obtained from 88 manually reconstructed discrete models. We begin by testing each type of BF for their enrichment in this dataset. Next, we designed relative enrichment tests to quantify the enrichment of a sub-type of BF within a given type of BF. Using these tests we make inferences on specific properties of regulatory logic rules that are preponderant in biological systems. Kauffman [22] had proposed that the occurrence of logical rules could be shaped by the constraint of being *chemically simple*. We borrow concepts from the computer science literature to quantify the notion of simplicity (or complexity) of a BF and then perform a thorough evaluation of preponderant biologically meaningful types of BFs from the perspective of complexity. The two measures of complexity which we exploit are Boolean complexity [69] and average sensitivity [71, 92]. We show that read-once Functions (RoFs) [84] that constitute all logical rules with minimal Boolean complexity are highly over-represented in the biological data. Further, we provide an analytical proof that NCFs [36], which are a subset of RoFs, minimize not only the Boolean complexity but also the average sensitivity across all BFs in Feldman’s associated $k[P]$ set. Our result that NCFs are minimally complex in terms of both complexity measures is a likely explanation for their prevalence in biological data. In a nutshell, our exploration of two complexity measures using 2687 BFs compiled from published models puts Kauffman’s conjecture of *preference for simplicity* on a sound footing while refining it, using a quantitative framework for rule complexity in GRNs. **The work reported in this chapter is contained in the published manuscript [49].**

3.1 Enrichments, relative enrichments and p -value tests

3.1.1 Enrichments and relative enrichments

Consider a given type of BF (say unate with k inputs) which we refer to as T . Denote by f_0 the fraction of functions that are of type T in the random ensemble and by f_1

the corresponding fraction in our reference biological dataset. The enrichment ratio E is simply f_1/f_0 . If $E > 1$, then T is enriched while if $E < 1$, T is depleted. If $E = 1$, there is neither enrichment nor depletion.

In our study we are also interested in relative enrichments to probe for possible causes of enrichments. For that we consider a type T and one of its sub-types, say T_s . For instance in our comparison of the two measures of complexity we examined the case where $T=\text{RoF}$ and $T_s=\text{NCF}$. In direct analogy with what was done for enrichments, we define the relative enrichment $E_R = (f_{s,1}/f_1)/(f_{s,0}/f_0)$ where the subscript s refers to type T_s . If biological enrichment is driven solely by the property of being in T , then the relative enrichment is expected to be close to 1. As a consequence, if E_R is large, then there must be other factors than *belonging to* T driving this relative enrichment.

3.1.2 Statistical significance tests

We developed a first statistical test to determine whether an observed enrichment E was statistically significant. The underlying statistical distribution of the random variable E is obtained by formalizing an underlying hypothesis referred to as H_0 . Here H_0 corresponds to hypothesizing that the functions in the reference biological dataset are drawn from the random ensemble where all 2^{2^k} BFs with k inputs are equiprobable. The (right-sided) p -value is then just the probability that such a drawing leads to a value of E as large as the one actually observed. This probability is computed as follows. The fraction f_0 is first determined. Then we consider drawing M BFs from the random ensemble and count the number m of these functions that belong to type T , wherein M is the number of BFs in the reference biological dataset. The probability of having a given value m is given by the binomial distribution: $\binom{M}{m} f_0^m (1 - f_0)^{M-m}$. The desired p -value is then just the sum of all such probabilities under the condition that m is larger or equal to Mf_1 .

The second type of test we perform concerns the statistical significance of a relative enrichment E_R deviating from 1. Again we formalize this by introducing an H_0 hypothesis. Using the notation of the previous sub-section, H_0 corresponds to assuming that although there is a selection for T (as evident from a large value of E), the elements that are drawn

within T have a uniform probability, that is members of T_s are not more probable than the other elements of T . Consider then drawing a sample of size M under H_0 . If it leads to M_T elements in T as in the reference biological dataset, the distribution of the number of elements in T_s is known. Specifically, the probability to have m elements in T_s is $\binom{M_T}{m} f_0^m (1 - f_0)^{M_T - m}$ where now f_0 is the ratio of sizes of T_s and T . The desired p -value is then just the sum of all such probabilities under the condition that m is larger or equal to the number of T_s elements in the reference biological dataset. The code used to compute these p -values is implemented in R and is available from the Github repository: <https://github.com/asamallab/MCBF>.

The number of functions belonging to a particular type of BF was obtained from both computation and theory. The number of CFs for $k = 6, 7, 8$ and NCFs for $k = 7, 8$ were obtained from [78] and [93]. In certain cases (i.e., EFs and UFs having 6, 7 or 8 inputs), it was computationally unfeasible to obtain the exact number of BFs in these types and there was no data in the literature as well, and hence, we used sampling to estimate the probability of a BF to belong to these types, for the specified number of inputs.

3.2 Enrichment of different types of BFs in reconstructed Boolean models of gene regulatory networks

In this section, we report on the relative abundance and associated statistical significance of the different types of BFs in a compiled dataset of 2687 BFs from 88 reconstructed models. For details on the compiled reference biological dataset see Section A.1, Appendix A. The delineation of the space of different biologically meaningful BFs is shown in the schematic Figure 3.1(a) and was explored in detail in Chapter 2. The in-degree distribution of these 2687 BFs, represented in Figure 3.1(b), shows that the number of these BFs decreases rapidly with increasing k . The key methodology hereafter consists in focusing on the relative abundances of the different types of BFs when comparing the ensemble of all BFs to the ensemble composed of our reference biological dataset. A statistically significant

enrichment is suggestive of some selection pressure on the BF_s in the biological networks.

Table 3.1: Number of different types of biologically meaningful BF_s in the reference biological dataset. Here, k is the number of inputs, “All” is the total number of BF_s for a given number of inputs, EF corresponds to effective functions, UF to unate functions (all sign combinations), CF to canalizing functions, EUF to effective and unate functions, ECF to effective and canalizing functions, UCF to unate and canalizing functions, EUCF to effective, unate and canalizing functions, NCF to nested canalizing functions and RoF to read-once functions.

k	Types of BF _s									
	All	EF	UF	CF	EUF	ECF	UCF	EUCF	NCF	RoF
1	934	934	934	934	934	934	934	934	934	934
2	687	671	687	687	671	671	687	671	671	671
3	412	392	411	398	391	378	398	378	378	378
4	258	251	257	239	250	232	239	232	230	244
5	156	149	153	136	146	129	135	128	120	133
6	107	98	107	93	98	85	93	85	67	83
7	51	48	50	49	47	46	48	45	34	41
8	45	45	43	40	43	40	38	38	27	34
9	19	19	18	17	18	17	17	17	7	16
10	13	12	13	11	12	10	11	10	3	6
11	1	1	1	1	1	1	1	1	0	0
12	3	3	3	3	3	3	3	3	2	2
13	0	0	0	0	0	0	0	0	0	0
14	1	1	1	1	1	1	1	1	1	1

Table 3.2: Fraction of different types of biologically meaningful BFs in the reference biological dataset. Here, k is the number of inputs, EF corresponds to effective functions, UF to unate functions (all sign combinations), CF to canalizing functions, EUF to effective and unate functions, ECF to effective and canalizing functions, UCF to unate and canalizing functions, EUCF to effective, unate and canalizing functions, NCF to nested canalizing functions and RoF to read-once functions.

k	Types of BFs								
	EF	UF	CF	EUF	ECF	UCF	EUCF	NCF	RoF
1	1	1	1	1	1	1	1	1	1
2	0.977	1	1	0.977	0.977	1	0.977	0.977	0.977
3	0.951	0.998	0.966	0.949	0.917	0.966	0.917	0.917	0.917
4	0.973	0.996	0.926	0.969	0.899	0.926	0.899	0.891	0.946
5	0.955	0.981	0.872	0.936	0.827	0.865	0.821	0.769	0.853
6	0.916	1	0.869	0.916	0.794	0.869	0.794	0.626	0.776
7	0.941	0.980	0.961	0.922	0.902	0.941	0.882	0.667	0.804
8	1	0.956	0.889	0.956	0.889	0.844	0.844	0.6	0.756
9	1	0.947	0.895	0.947	0.895	0.895	0.895	0.368	0.842
10	0.923	1	0.846	0.923	0.769	0.846	0.769	0.231	0.462
11	1	1	1	1	1	1	1	0	0
12	1	1	1	1	1	1	1	0.667	0.667
13	-	-	-	-	-	-	-	-	-
14	1	1	1	1	1	1	1	1	1

Table 3.3: p -value tests for enrichments of the different types of BFs in the reference biological dataset. A low p -value indicates that the corresponding type of BF is enriched in the reference biological dataset when compared to the ensemble of all BFs. For $k > 2$ when the p -value shown is 0, it was smaller than what we could measure, and when the p -value shown is 1, its deviation from 1 was smaller than we could measure. Here, EF corresponds to effective functions, UF to unate functions (all sign combinations), CF to canalizing functions, NCF to nested canalizing functions and RoF to read-once functions.

k	Odd bias	EF	UF	CF	NCF	RoF
2	3.742×10^{-177}	6.75×10^{-114}	0	0	3.74×10^{-177}	3.74×10^{-177}
3	6.253×10^{-76}	2.44×10^{-11}	6.65×10^{-162}	1.83×10^{-111}	3.09×10^{-184}	3.09×10^{-184}
4	2.714×10^{-62}	0.919	0	8.86×10^{-279}	0	0
5	2.718×10^{-25}	1	0	0	0	0
6	1.753×10^{-13}	1	0	0	0	0
7	6.058×10^{-08}	1	0	0	0	0
8	1.561×10^{-06}	1	0	0	0	0

Table 3.4: Fractions of functions that are RoFs, non-NCF RoFs or NCFs, in the space of all 2^{2^k} BFs (f_0) and in the reference biological dataset (f_1). E ($= f_1/f_0$) is the enrichment ratio; it indicates the extent of the over-representation of such functions in the reference biological dataset. Over-representation is highest for NCFs but clearly non-NCF RoFs are also highly over-represented. Computations are reported for functions with $k \leq 8$ inputs.

k	RoF			non-NCF RoF			NCF		
	f_0	f_1	E	f_0	f_1	E	f_0	f_1	E
1	0.5	1.000	2.000	0	0	-	0.5	1.000	2.00
2	0.5	0.977	1.953	0	0	-	0.5	0.977	1.95
3	0.250	0.917	3.670	0	0	-	0.25	0.917	3.67
4	1.27×10^{-2}	0.946	74.495	1.46×10^{-3}	5.43×10^{-2}	37.04	1.12×10^{-2}	8.91×10^{-1}	79.38
5	3.52×10^{-6}	0.853	2.42×10^5	1.04×10^{-6}	0.083	7.99×10^4	2.47×10^{-6}	0.769	3.11×10^5
6	1.91×10^{-14}	0.776	4.06×10^{13}	9.12×10^{-15}	0.150	1.64×10^{13}	9.97×10^{-15}	0.626	6.28×10^{13}
7	2.95×10^{-32}	0.804	2.73×10^{31}	1.86×10^{-32}	0.137	7.39×10^{30}	1.09×10^{-32}	0.667	6.11×10^{31}
8	2.92×10^{-69}	0.756	2.59×10^{68}	2.18×10^{-69}	0.156	7.14×10^{67}	7.41×10^{-70}	0.600	8.10×10^{68}

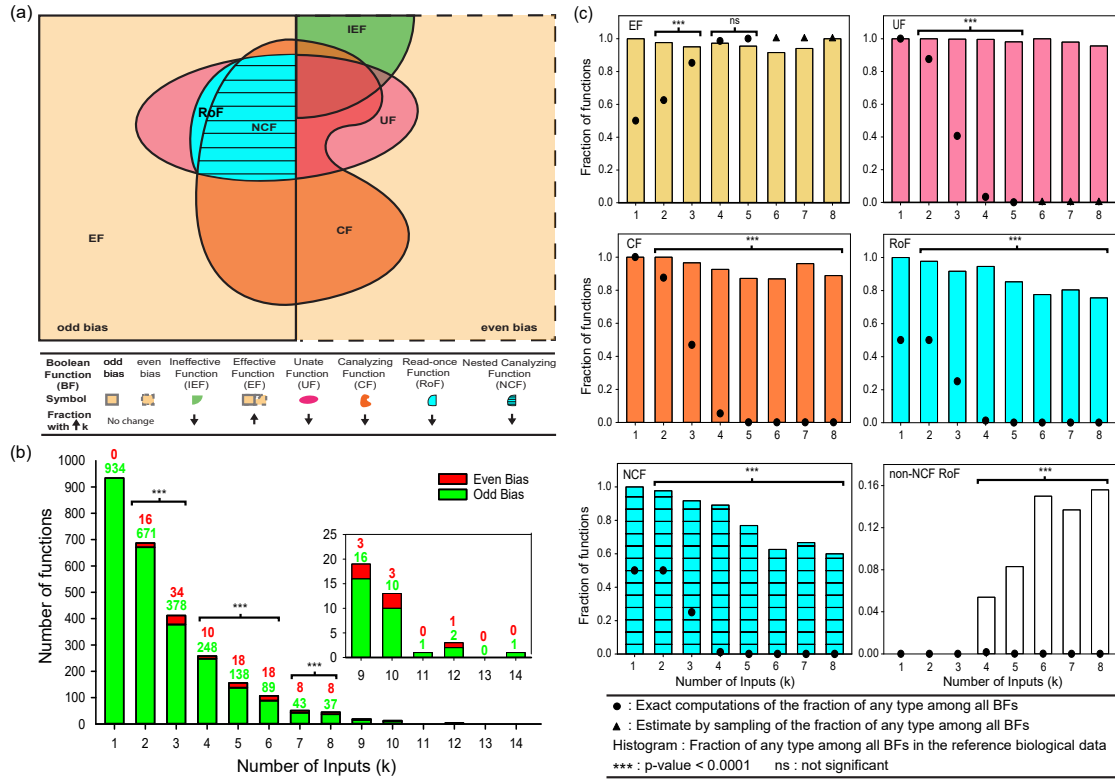


Figure 3.1: Overlap of different types of BFs and their distribution in the reference biological dataset. (a) In the space of all BFs, a schematic of the overlaps between different types of biologically meaningful BFs with 4 inputs. This figure is not drawn to scale but the sizes of the sets corresponding to different types of BFs and their intersections respect the order of the actual values. The legend gives the correspondence between shapes with specific color and the different types of BFs. Ordering the different types of BFs with 4 inputs (which are *not* mutually exclusive) based on their sizes in a descending order gives: EF > Odd bias = Even bias > CF > UF > RoF > NCF. The up (or down) arrows in the legend depict the increase (or decrease) in the fraction of BFs that belong to a specific type as k increases. (b) The in-degree distribution for nodes in the reference biological dataset. (c) The plots show the abundance and statistical significance of the biologically meaningful BFs for $k \leq 8$ in the reference biological dataset. The dot symbols which appear to coincide with the x -axis are very small non-zero numbers (except for non-NCF RoFs with $k = 1, 2, 3$).

3.2.1 Enrichment in types when comparing to the ensemble of random BFs

Figure 3.1(b) indicates that for in-degrees $1 \leq k \leq 8$, the odd bias BFs are dominant and statistically enriched in the reference biological dataset. It is not immediately apparent

why BFs with odd bias should be preferred over BFs with even bias as biologically meaningful BFs with even bias do exist, e.g., a subset of functions which are both unate and canalyzing can have even bias (see Figure 3.1(a)). Furthermore, among 2-input BFs, the XOR and XNOR functions have even bias but are completely absent from our reference biological dataset.

Figure 3.1(c) shows the relative abundances in the reference biological dataset as bars (see Tables 3.1 and 3.2 for exact values) and as dots in the ensemble of random BFs, of the various types of BFs. Statistical tests reveal that the relative abundances in the reference biological dataset are larger (one-sided p -values) than those expected under the null hypothesis whereby the reference BFs are drawn from the ensemble of random BFs (see stars above the bars in Figure 3.1(c) and Table 3.3 for p -values), with the exception of the EFs. This exception is justified by the fact that functions drawn randomly from the space of all BFs are typically EFs, particularly for BFs with at least 3 inputs (see Table 2.2). The ratios provided in Table 3.4 show that the RoF, NCF and the non-NCF RoF types are all strongly enriched in the reference biological dataset.

Table 3.5: The relative enrichment ratios E_R for the RoFs and NCFs in the ensemble of odd bias BFs, EFs and UFs. These ratios indicate the extent of the over-representation of such functions in the reference biological dataset. $E_R > 1$ suggests that there is indeed an enrichment of RoFs and NCFs within the EFs, UFs and CFs in the reference biological dataset when compared to that expected in the ensemble of all EFs, UFs and CFs.

k	E_R for RoF in:			E_R for NCF in:		
	Odd bias	EF	UF	Odd bias	EF	UF
1	1	1	2	1	1	2
2	1.0	1.25	1.75	1.0	1.25	1.75
3	2.0	3.284	1.567	2.0	3.284	1.567
4	38.770	75.483	2.536	41.279	80.367	2.700
5	1.37×10^5	2.54×10^5	13.633	1.76×10^5	3.25×10^5	17.47

Table 3.6: The relative enrichment ratio E_R of the NCFs in the CFs and RoFs. $f_{s,0}/f_0$ denotes the fractions of functions that are NCFs in the space of all CFs or RoFs and $f_{s,1}/f_1$, the equivalent fraction in the reference biological dataset. $E_R = (f_{s,1}/f_1)/(f_{s,0}/f_0)$ denotes the enrichment ratio and it indicates the extent of the over-representation of such functions in the reference biological dataset. Computations are reported for BF with $k \leq 8$ inputs. The low p -values indicate that there is an enrichment of NCFs within the CFs and RoFs in the reference biological dataset when compared to that expected in the ensemble of all CFs and RoFs.

k	NCF in CF				NCF in RoF			
	$f_{s,0}/f_0$	$f_{s,1}/f_1$	E_R	p -value	$f_{s,0}/f_0$	$f_{s,1}/f_1$	E_R	p -value
1	0.5	1	2	-	1	1	1	-
2	0.571	0.977	1.709	3.49×10^{-139}	1	1	1	-
3	0.533	0.950	1.781	2.47×10^{-78}	1	1	1	-
4	0.209	0.962	4.595	5.32×10^{-144}	0.885	0.943	1.066	6.86×10^{-04}
5	8.22×10^{-3}	0.882	1.07×10^2	1.56×10^{-233}	0.703	0.902	1.283	7.46×10^{-09}
6	1.78×10^{-06}	0.720	4.04×10^5	0	0.522	0.807	1.546	1.58×10^{-08}
7	7.19×10^{-15}	0.694	9.65×10^{13}	0	0.370	0.829	2.240	2.42×10^{-10}
8	7.88×10^{-33}	0.675	8.57×10^{31}	0	0.254	0.794	3.129	5.26×10^{-12}

3.2.2 Relative enrichment in sub-types when comparing to the ensemble of random BFs

Comparing the enrichments of the different types of biologically meaningful BFs can provide signatures of causes of enrichment. For instance, if selection operated only in favor of unateness, each sub-type therein (NCF, RoF or non-NCF RoF) would be expected to have its relative abundance (proportion within UF) be the same whether one considers the reference biological dataset or the ensemble of random BFs. In effect, the proportions of different sub-types of BFs in the two ensembles point to which factors drive the different enrichments. We thus developed a way to test the null hypothesis that a sub-type enrichment is solely due to the enrichment in one of its englobing types.

Let us first consider the enrichment ratios of NCFs and RoFs within the three englobing types of BFs: odd bias, EFs and UFs. From Table 3.5, it is clear that, for $k > 2$, the relative enrichment ratios E_R (when comparing the observed to the expected under the null hypothesis) of both the NCFs and RoFs are much greater than 1, implying that the enrichment of these sub-types does not follow from the enrichment of their super-sets. Thus biological selection *solely* in favor of being odd biased, effective or unate is not consistent with the enrichments found for the NCFs or RoFs in the reference biological dataset, some other factors must be at work.

Second, since NCFs are a subset of CFs, we can ask whether canalization is the factor driving the enrichment of NCFs. Since the relative enrichment ratios are high and the p -values low (see Table 3.6), we conclude that selection for canalization alone does not explain the enrichment observed for NCFs. Similarly, we can ask whether the fact that a function is a RoF, that drives the enrichment of NCFs (a sub-type of RoF). As shown in Table 3.6, the relative enrichment of NCF within RoF is quite modest, almost all k having E_R values in the range 1 to 2. Nevertheless our statistical method shows that these values are not consistent with 1 (absence of any enrichment) as indicated by the p -values in Table 3.6, so there must be some further cause of the enrichment of NCFs other than that of belonging to the RoF type. In order to understand the potential reasons for the enrichment of RoFs and NCFs, we explored different complexity measures derived from computer science literature as explained in the next section.

3.3 Complexity Measures

Various measures of *complexity* of BFs have been studied in the computer science literature [80, 92, 94]. We adopt two of them in this work, namely, Boolean complexity and average sensitivity.

3.3.1 Minimal expressions and Boolean complexity

The first measure of complexity we use, formulated in particular by Feldman [69], is the Boolean complexity. In principle there are an infinite number of logical expressions corre-

sponding to a given BF [69, 94]. Feldman [69] focused on the shortest possible expression when considering the number of literals it is composed of, the so called *minimal formula* for a BF. Feldman defined the *Boolean complexity* of a BF to be the number of literals in its minimal formula [69, 94]. Though Boolean expression types such as the minimal canonical disjunctive normal form (DNF) or the minimal canonical conjunctive normal form (CNF) are widely-used to represent BFs, they are typically distinct from the minimal formula as defined by Feldman [69].

For instance the 3-input BF in the minimal canonical DNF, $f(x_1, x_2, x_3) = (\bar{x}_1 \wedge \bar{x}_2 \wedge \bar{x}_3) \vee (\bar{x}_1 \wedge \bar{x}_2 \wedge x_3) \vee (\bar{x}_1 \wedge x_2 \wedge \bar{x}_3)$ containing 9 literals can be shown to be equivalent to a minimum formula containing 3 literals by applying the laws of Boolean algebra as follows:

$$\begin{aligned}
f(x_1, x_2, x_3) &= (\bar{x}_1 \wedge \bar{x}_2 \wedge \bar{x}_3) \vee (\bar{x}_1 \wedge \bar{x}_2 \wedge x_3) \vee (\bar{x}_1 \wedge x_2 \wedge \bar{x}_3) \\
&= \bar{x}_1 \wedge ((\bar{x}_2 \wedge \bar{x}_3) \vee (\bar{x}_2 \wedge x_3) \vee (x_2 \wedge \bar{x}_3)) \\
&= \bar{x}_1 \wedge ((\bar{x}_2 \wedge (\bar{x}_3 \vee x_3)) \vee (x_2 \wedge \bar{x}_3)) \\
&= \bar{x}_1 \wedge (\bar{x}_2 \vee (x_2 \wedge \bar{x}_3)) \\
&= \bar{x}_1 \wedge ((\bar{x}_2 \vee x_2) \wedge (\bar{x}_2 \vee \bar{x}_3)) \\
&= \bar{x}_1 \wedge (\bar{x}_2 \vee \bar{x}_3)
\end{aligned}$$

Here, x_i and \bar{x}_i represent a positive and negative literal respectively. In the above simplification, we employ the law $\bar{x}_i \vee x_i = 1$, and the distribution property over the OR (\vee) operator. Thus, the minimal irreducible expression $f(x_1, x_2, x_3) = \bar{x}_1 \wedge (\bar{x}_2 \vee \bar{x}_3)$ has 3 literals and the function has Boolean complexity equal to 3. However, note that the minimal DNF for this BF is $(\bar{x}_1 \wedge \bar{x}_2) \vee (\bar{x}_1 \wedge \bar{x}_3)$, which has 4 literals, and factorization of this expression is necessary to obtain the minimal expression with 3 literals for the above BF.

Computing the Boolean complexity

Obtaining a minimal formula for a given BF or expression is a computationally hard problem [95]. In practice, one has to resort to heuristic algorithms such as the QMV proposed by Vigo [96] for reducing expressions. Thus, barring exceptions, one can only obtain an upper bound on the Boolean complexity for BFs with several inputs. In our work, to obtain the factorized minimal expression of a BF, we employ the logic synthesis software “ABC” [97,98]. To improve the estimated Boolean complexity of a BF, we give as input to the ABC software four types of Boolean expressions, namely the full DNF, the full CNF, the Quine-McCluskey minimized DNF expression [99,100] and the Quine-McCluskey minimized CNF expression, corresponding to the same BF. As a result, 4 output Boolean expressions are obtained of which the one with the least number of literals is chosen as the minimal equivalent expression of the BF. The number of literals in this expression is then our estimate of the Boolean complexity of that BF. From the definition of Boolean complexity, the following 3 properties immediately follow:

Property 3.3.1. EFs with k inputs have Boolean complexity $\geq k$.

Proof: For a BF f to be effective, the k different input variables must appear at least once in the minimal expression or formula for the BF. This implies that the number of literals in the minimal expression or Boolean complexity is $\geq k$.

Property 3.3.2. NCFs with k inputs have Boolean complexity equal to k .

Proof: In Eq. (2.7), each variable or literal in the expression for an NCF appears exactly once, thus the Boolean complexity of a NCF is equal to k . Thus, NCFs have the *minimum* Boolean complexity among EFs with given number of inputs.

Property 3.3.3. RoFs with k inputs have the minimum Boolean complexity k among all the EFs.

Proof: Since RoFs are constructed such that each variable or literal in the expression for a RoF (Eq. (2.6)) appears exactly once, the Boolean complexity of a RoF is equal to k . Further, using Property 3.3.1, k is the minimum value in EF, hence RoFs have the minimum Boolean complexity among all EFs. In sum, RoFs correspond exactly to the set

of EFs with *minimum* Boolean complexity.

3.3.2 Average sensitivity of BFs

The second measure of complexity we use, the average sensitivity, is based on how sensitive a BF is to changes of its inputs [71]. For a BF f with k inputs, the *sensitivity* for a given assignment of the input variables $\mathbf{x} = (x_1 = a_1, x_2 = a_2, \dots, x_k = a_k)$ is the number of *neighbors* \mathbf{y} of \mathbf{x} for which the output $f(\mathbf{y})$ is different from $f(\mathbf{x})$ [71,92]. The assignments \mathbf{y} and \mathbf{x} are *neighbors* if they differ in the value of exactly one of their k variables. The average of the sensitivity over all input combinations gives the average sensitivity of a BF, and is given by the expression:

$$S_f = \left\langle \sum_{i=1}^k f(\mathbf{x} \oplus \mathbf{e}_i) \oplus f(\mathbf{x}) \right\rangle_{\mathbf{x}} \quad (3.1)$$

where \oplus is the XOR operator and $\mathbf{e}_i \in \{0, 1\}^k$ denotes the unit vector corresponding to having input variable $x_i = 1$ and all other input variables set to 0. \mathbf{x} can be mapped to a vertex V of a k -dimensional Boolean hypercube (or k -cube). The sensitivity at \mathbf{x} then has a geometric interpretation: it is the number (between 0 and k) of neighbors of V whose output value differs from that of V . The total sensitivity of f which is the sum of the sensitivities over all the vertices of the k -cube is equal to twice the number of k -cube edges whose two ends are vertices with complementary output values. It follows from the above definition that the lower the average sensitivity of a BF, the more *robust* it is to changes of its input variables [71].

Note that isomorphic BFs have identical average sensitivities. Indeed, the operations of rotations or reflections about any of the axes of the hypercube do not change the number of red and blue neighbors with output values 1 or 0, respectively, for any vertex (see Figure 2.1(d)). Moreover, a BF and its complement belonging to sets $k[P]$ and $k[2^k - P]$, respectively, also have the same average sensitivity. This is because under complementation of the BF, the red and blue vertices of the k -cube are exchanged, thereby leaving the number of edges E_{01} in the k -cube unchanged.

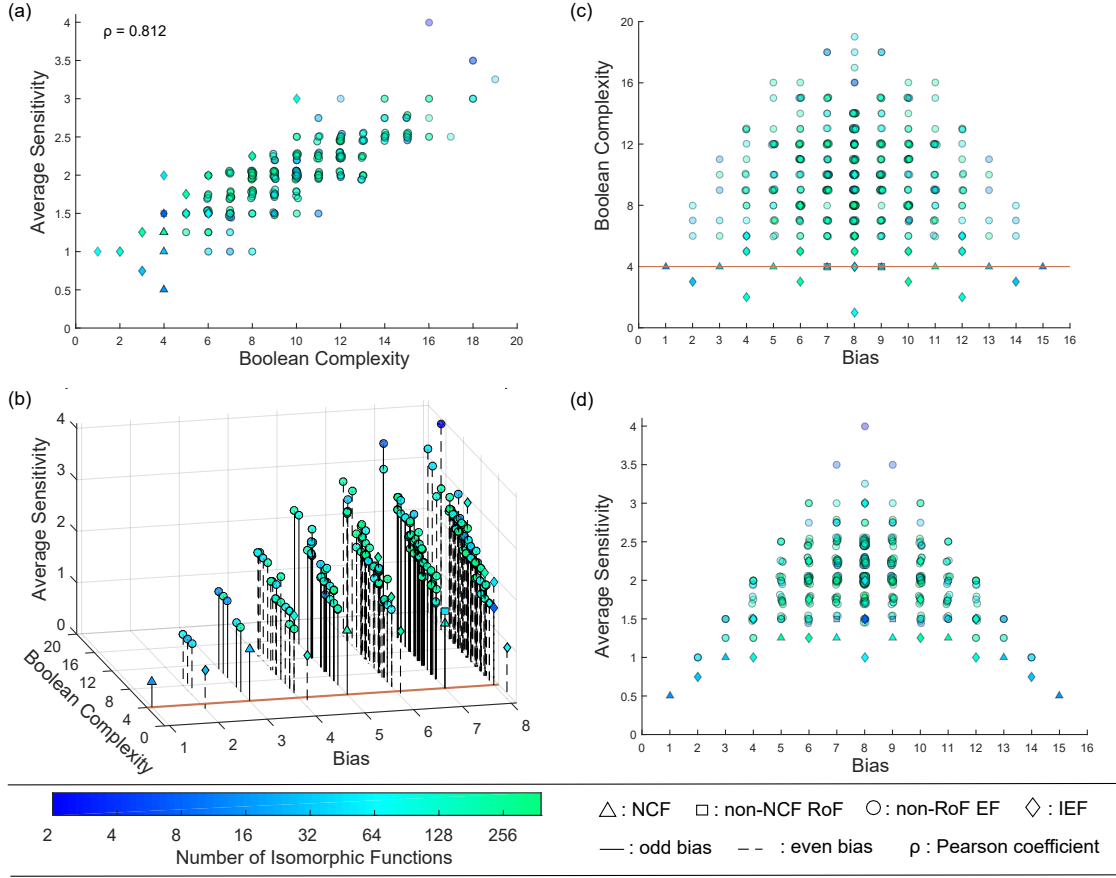


Figure 3.2: Dependence of the two complexity measures on the bias and associated 2D projections for all BFs with $k = 4$ inputs. In each sub-figure, a point corresponds to a class of (isomorphic) BF and is assigned a shape and a color. The shape of a point (triangle, square, circle or diamond) denotes the type of BF (NCF, non-NCF RoF, non-RoF EF or ineffective functions (IEFs)) whereas its color indicates the number of BFs contained in its corresponding class. The same shape and color scheme is applicable to all the plots. A slight jiggle is added at some points to resolve overlapping representative BFs. The type *non-RoF EF* refers to the subset of EFs which are not RoFs. (a) The Pearson correlation coefficient (ρ) between the two measures is large and positive and was calculated for all BFs with $k = 4$ and $P \leq 8$. (b) The 3D plot adds the third dimension of bias P to the preceding 2D plot. The solid and dashed vertical lines or needles, as we will refer to them henceforth, show the projections of the points onto the plane of bias and Boolean complexity. These needles have been included to enhance clarity while distinguishing between the odd bias BFs and even bias BFs. The brown line drawn at the Boolean complexity 4 highlights the functions that possess the minimum Boolean complexity and are effective as well. The RoFs are the only functions which lie along this line. Since the two complexity measures are invariant under complementation of the BF, the bias values have been shown only up to $P = 8$.

Figure 3.2 (previous page): (c) Variation of the Boolean complexity with the bias. With increasing bias upto $P = 8$, the number of representative BFs increases, but so does the range of Boolean complexity of these functions. The RoFs and IEFs have the minimum Boolean complexity in any $4[P]$ set. The brown line drawn at the Boolean complexity 4 highlights the functions that possess the minimum Boolean complexity and are effective as well. (d) Variation of average sensitivity with increasing bias. Clearly, the NCFs and IEFs have the minimum average sensitivity in any $4[P]$ set. Note that both sub-figures (c) and (d) are symmetric about $P = 8$ due to the complementarity property.

3.4 Enriched functions in biological data have minimum complexity

A plausible explanation for the enrichment of the RoFs and NCFs in the dataset is their *low complexity*. In terms of the first notion (Boolean complexity), the RoFs, of which NCFs are a subset, have the minimum Boolean complexity among all EFs (from Properties 3.3.2 and 3.3.3). RoFs and NCFs have the same Boolean complexity but differ in the second measure of complexity, namely average sensitivity (see Figure 3.2). This following section examines more closely the properties of these two complexity measures. Computationally, the NCFs appear to have the minimum average sensitivity for a given $k[P]$ set (see Figure 3.2). We also harness the fact that for any bias P , the minimum average sensitivity is obtained for a particular geometry of the “on” vertices of the k -dimensional hypercube. We will show that when the bias is odd, this geometry corresponds to a NCF, while if it is even the BF is ineffective.

3.4.1 Boolean complexity and average sensitivity are strongly correlated

We first explore how the two measures of complexity compare. The average sensitivity of a BF can be computed easily using Eq. (3.1) while computing the Boolean complexity of a BF is more challenging but was done as described in the section on complexity measures. A bivariate analysis of these two measures of complexity allows us to obtain the Pearson correlation coefficient ($\rho = 0.812$) for all BFs at $k = 4$ inputs. We find that there is a

strong positive linear relationship between the two measures (see Figure 3.2(a)). Looking closely at functions in the neighborhood of the brown line (which highlights the minimum Boolean complexity of 4 for EFs) in the 3D plot Figure 3.2(b), we observe that: (i) All EFs along this brown line have odd bias and are NCFs or non-NCF RoFs (see Figures 3.2(b) and 3.2(c)). (ii) At bias $P = 7$, NCFs have a lower average sensitivity than the non-NCF RoFs (see Figures 3.2(b) and 3.2(d)). (iii) At any even bias, the BF's having the minimum average sensitivity are ineffective functions (IEFs) of Boolean complexity strictly less than k (see Figures 3.2(b) and 3.2(c)). These computational observations led us to the two conjectures listed below which we prove in the subsequent sub-sections:

- When P is odd, NCFs have the minimum average sensitivity within their $k[P]$ set.
- When P is even, the functions with minimum average sensitivity are ineffective with Boolean complexity $< k$.

We were also curious as to whether two representative non-NCF RoFs within a $k[P]$ set could have the same average sensitivity and we find this is indeed true via exhaustive computational enumeration of RoFs with $k \leq 10$. We observe that such a case of two representative non-NCF RoFs first occur when $k = 7$ at the bias $P = 25$.

3.4.2 NCFs have the minimum average sensitivity within their $k[P]$ set when P is odd

Mapping average sensitivity to the number of edges between P vertices of a k -cube

In the k -cube representation of a BF, each vertex corresponds to a binary string \mathbf{x} that defines the BF's input. We thus assign 0s and 1s to each of the associated vertices to specify the BF's output for each input string \mathbf{x} . If P is the bias of the BF, there are P vertices carrying the label 1. The total number of edges stemming from these P vertices is kP . Of these, some edges may end at one of the other $P - 1$ vertices having the value 1; we refer to the associated set of edges as E_{11} . Similarly, we denote by E_{01} the remaining edges, ending at any of the $2^k - P$ other vertices having the value 0. These two quantities

satisfy $E_{01} + 2E_{11} = kP$ [72]. The average sensitivity of the BF is given by $2E_{01}/2^k$; clearly the problem of minimizing this quantity in the set $k[P]$ is equivalent to maximizing E_{11} since k and P are fixed.

Edge-maximizing arrangement between P vertices of the k -cube: Defining *good sets*

Hart [72] solved the problem of finding an arrangement of P vertices on a k -cube that maximizes the number of edges connecting them. This problem has also been solved by other authors [101, 102], though in other contexts. We choose to use Hart's approach due to its mathematical clarity and easy visualization. Hart introduces the notion of a *good set* of P vertices on a k -cube where $P < 2^k$ using the following recursive definition:

- (i) If $P = 1$, we always have a good set.
- (ii) Otherwise, find r such that $2^r < P \leq 2^{r+1}$. Select any $(r+1)$ -cube embedded in the k -cube. Then, select two r -cubes which are vertex disjoint subsets of the $(r+1)$ -cube. To select the P vertices, include first 2^r vertices by taking one of the r -cubes and include the remaining $P - 2^r$ vertices by imposing that they form a good set containing $P - 2^r$ vertices on the other r -cube.

By expressing P as a sum of powers of 2, i.e., $P = \sum_{i=1}^l 2^{r_i}$, the resulting set of strictly increasing exponents $\{r_1, r_2, \dots, r_l\}$ gives the dimensions of the successive cubes to be used to define a good set. Hart [72] was able to prove that good sets maximize the number of edges connecting P vertices at fixed P .

Good sets having an odd number of vertices correspond to NCFs

Claim: Given the k -cube representation of BFs in $k[P]$, our claim is that the P vertices (P odd) with output value 1 form a good set *iff* the BF is a NCF.

Proof: Consider the logical expression of a NCF (Eq. (2.7)) in a $k[P]$ set. The i^{th} canalyzing variable $x_{\sigma(i)}$ determines which partition (of the possible $k - (i - 1)$ partitions, $i - 1$ variables having already been fixed) of a $(k - (i - 1))$ -cube into 2 vertex disjoint $(k - i)$ -cubes is to be canalyzed. Furthermore, the canalyzing input value a_i ($x_{\sigma(i)} = a_i$) fixes the outputs of the vertices of one of the two vertex disjoint $(k - i)$ -cubes to the value

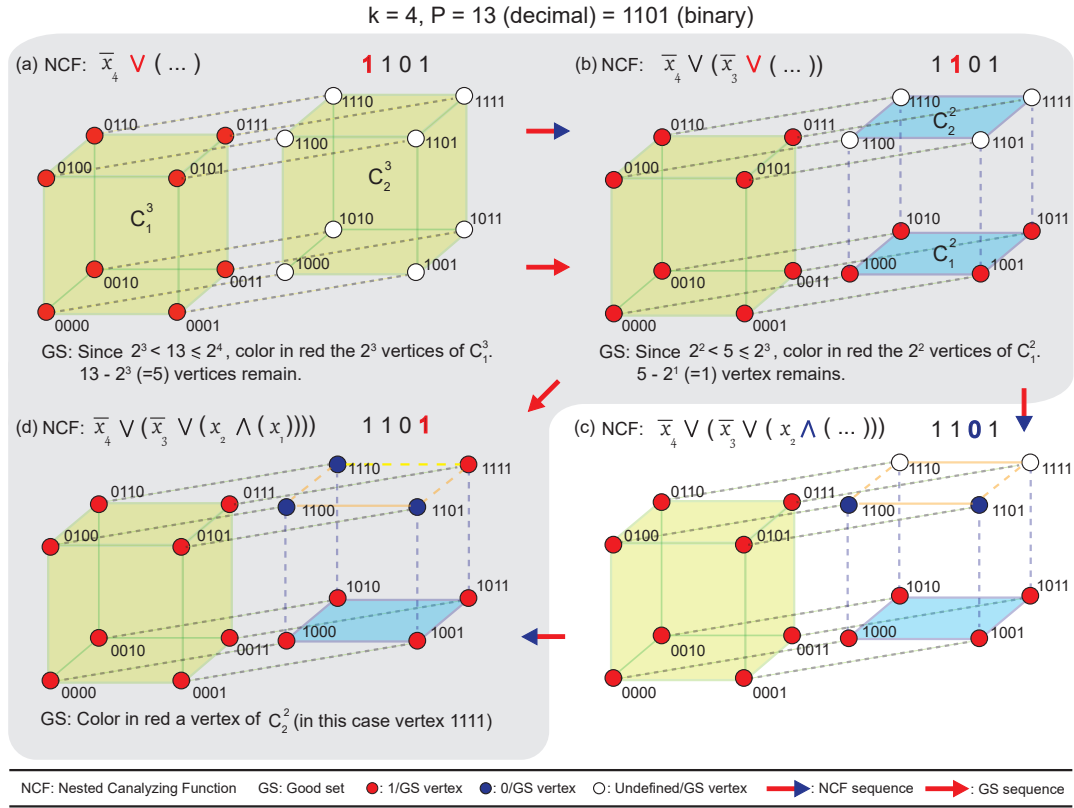


Figure 3.3: A Good set (GS) with P vertices where P is odd on a k -dimensional hypercube is equivalent to a NCF in $k[P]$ set ($k = 4, P = 13$). In parts (a), (b) and (d) shaded in grey, we show the recursive construction of a GS for $P = 13$ vertices in a 4-dimensional hypercube by coloring its vertices red, and in parts (a), (b), (c) and (d), we show the equivalence of that GS with 13 vertices to a NCF with bias 13. The vertices of the hypercube are labeled in the order x_4, x_3, x_2, x_1 wherein x_i is 0 or 1. Here, C_1^j and C_2^j denote the two vertex disjoint j -dimensional hypercubes of the $(j + 1)$ -dimensional hypercube. The active bit in each part (a), (b), (c) and (d) is the colored bit in the binary representation of 13 in that part. **(a)** Since $P = 13$ lies between 2^3 and 2^4 , 2^3 vertices of either C_1^3 or C_2^3 (here, C_1^3) form part of the GS. This leaves $13 - 8 = 5$ vertices to be colored red to complete the GS. This choice of 8 vertices in C_1^3 for the GS leads to the canalyzation of vertices labelled $x_4 = 0$ to the output value 1. In this step, the active bit is 1 and as a result the \vee operator follows the literal \bar{x}_4 . **(b)** Following the same procedure as in (a) for coloring the remaining 5 vertices of the GS leads to the choice of 4 vertices in C_1^2 . This leaves one vertex to be colored (which is the base case of the recursion to construct the GS). The choice of 4 vertices for the GS leads to the canalyzation of vertices with $x_4 = 1$ and $x_3 = 0$ to the output value 1. The active bit in this step is 1 and as a result the \vee operator follows the literal \bar{x}_3 . **(c)** For the corresponding NCF, the vertices with $x_4 = 1$, $x_3 = 1$ and $x_2 = 0$ are canalyzed to the output value 0. The active bit in this step is 0 and as a result the \wedge operator follows the literal x_2 . **(d)** For the last step, any vertex in C_2^2 can be colored to complete the 13 vertices in GS, and we color here the vertex 1111. The vertex with $x_4 = 1$, $x_3 = 1$, $x_2 = 1$ and $x_1 = 1$ is canalyzed to the output value 1, and the remaining vertex is set to output value 0.

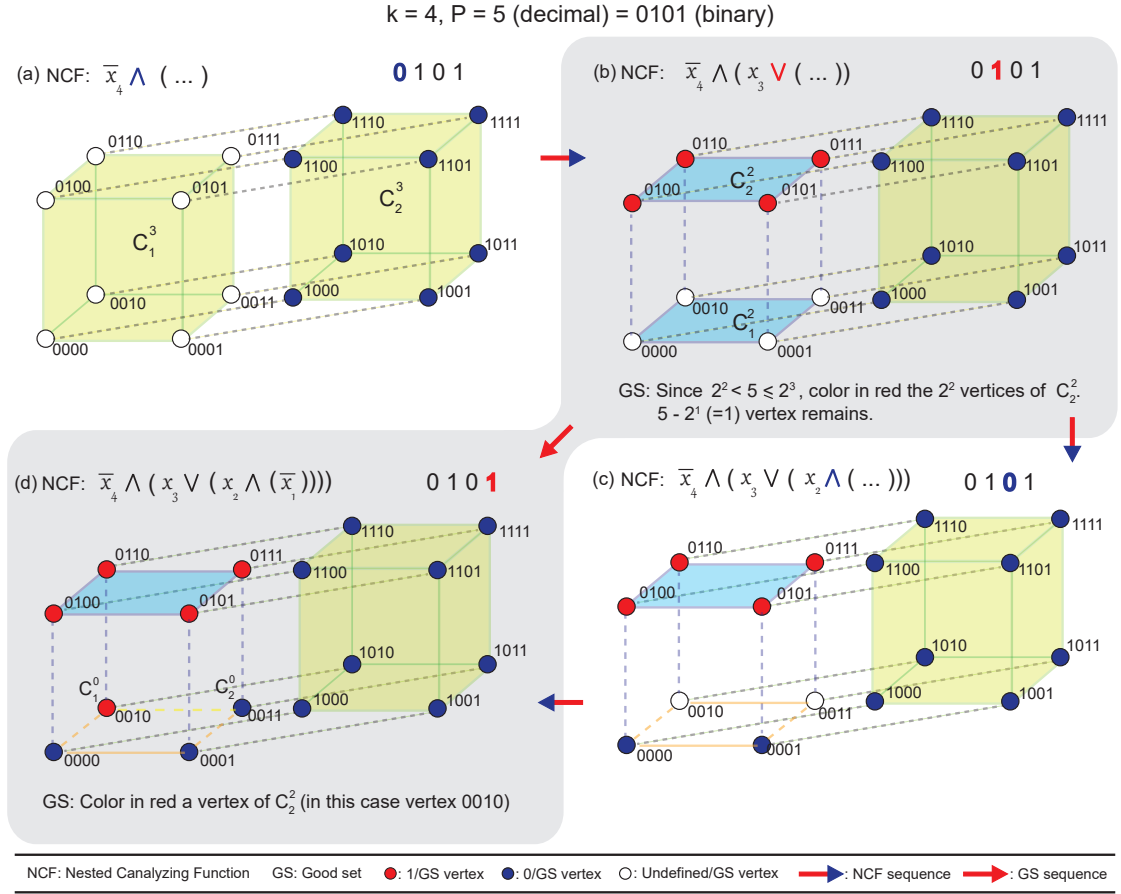


Figure 3.4: A Good set (GS) with P vertices where P is odd on a k -dimensional hypercube is equivalent to a NCF in $k[P]$ set ($k = 4, P = 5$). In parts (b) and (d) shaded in grey, we show the recursive construction of a GS for $P = 5$ vertices in a 4-dimensional hypercube by coloring its vertices red, and in parts (a), (b), (c) and (d), we show the equivalence of that GS of 5 vertices to a NCF with bias 5. The vertices of the hypercube are labeled in the order x_4, x_3, x_2, x_1 wherein x_i is 0 or 1. Here, C_1^j and C_2^j denote the two vertex disjoint j -dimensional hypercubes of the $(j + 1)$ -dimensional hypercube. The active bit in each part (a), (b), (c) and (d) is the colored bit in the binary representation of 5 in that part. **(a)** The vertices with $x_4 = 1$ are canalyzed to the output value 0. The active bit in this step is 0 and as a result the \wedge operator follows the literal \bar{x}_4 . **(b)** Since $P = 5$ lies between 2^2 and 2^3 , 2^2 vertices of either C_1^2 or C_2^2 (here, C_2^2) form part of the GS. This leaves $5 - 4 = 1$ vertex to be colored red to complete the GS. This choice of 4 vertices in C_2^2 for the GS leads to the canalyzation of vertices labeled $x_4 = 0$ and $x_3 = 1$ to the output value 1. The active bit in this step is 1 and as a result the \vee operator follows the literal x_3 . **(c)** The vertices with $x_4 = 0$, $x_3 = 0$ and $x_2 = 0$ are canalyzed to the output value 0. The active bit in this step is 0 and as a result the \wedge operator follows the literal x_2 . **(d)** For the last step, any vertex in C_1^0 can be colored to complete the 5 vertices in GS, and we color the vertex 0010. The vertex with $x_4 = 0$, $x_3 = 0$, $x_2 = 1$ and $x_1 = 0$ is canalyzed to the output value 1, and the one remaining vertex is set to output value 0.

b_i . Repeating the procedure recursively over $i \in \{1, 2, \dots, k\}$ gives the arrangement of 1s and 0s for a NCF on a k -cube. To obtain a NCF with a certain bias P , the i 's for which $b_i = 1$ have to be chosen appropriately so that $P = \sum_{i=1}^k b_i 2^{k-i}$.

The above procedure of setting the output values of P vertices to 1s and $2^k - P$ vertices to 0s on the k -cube is equivalent to obtaining a good set of P vertices, setting their output values to 1 and then setting the output of the remaining $2^k - P$ vertices to 0. This is true because:

1. The dimensions of the cubes whose vertices are to have the output value 1 are the same in either case (i.e., the set of exponents obtained by expressing P as a sum of powers of 2 is unique for a given P).
2. When some i -cube is chosen to place the 1s, there is only one other i -cube, which (along with the chosen i -cube) constitutes 2 vertex disjoint subsets of a $(i+1)$ -cube.

In both cases, this is an i -cube where the next set of 1s are placed.

Thus the P vertices with output value 1 in a NCF constitute a good set and inversely any good set with P odd corresponds to a NCF. Given Hart's proof, NCFs must then have the minimum average sensitivity among all BFs in $k[P]$. We provide a visual illustration of the equivalence of Hart's construction of the good set to the construction of the NCF for two 4-input BFs with biases $P = 13$ (see Figure 3.3) and $P = 5$ (see Figure 3.4).

3.4.3 Good sets having an even number of vertices has Boolean complexity strictly less than k

The logic of the above derivation can be extended to the case where the good set has an even number of vertices: one then sees that the resulting BFs have a hierarchical structure similar to the NCFs, but with some variables ineffective (see Figure 3.5). If all ineffective variables are ignored, one sees that a good set of even number of vertices leads to a NCF with fewer variables.

Claim: If an even number of vertices P having the output value 1 in the hypercube representation of a BF (in a $k[P]$ set) forms a good set, then its Boolean complexity is

$k = 4, P = 6$ (decimal) = 0110 (binary)

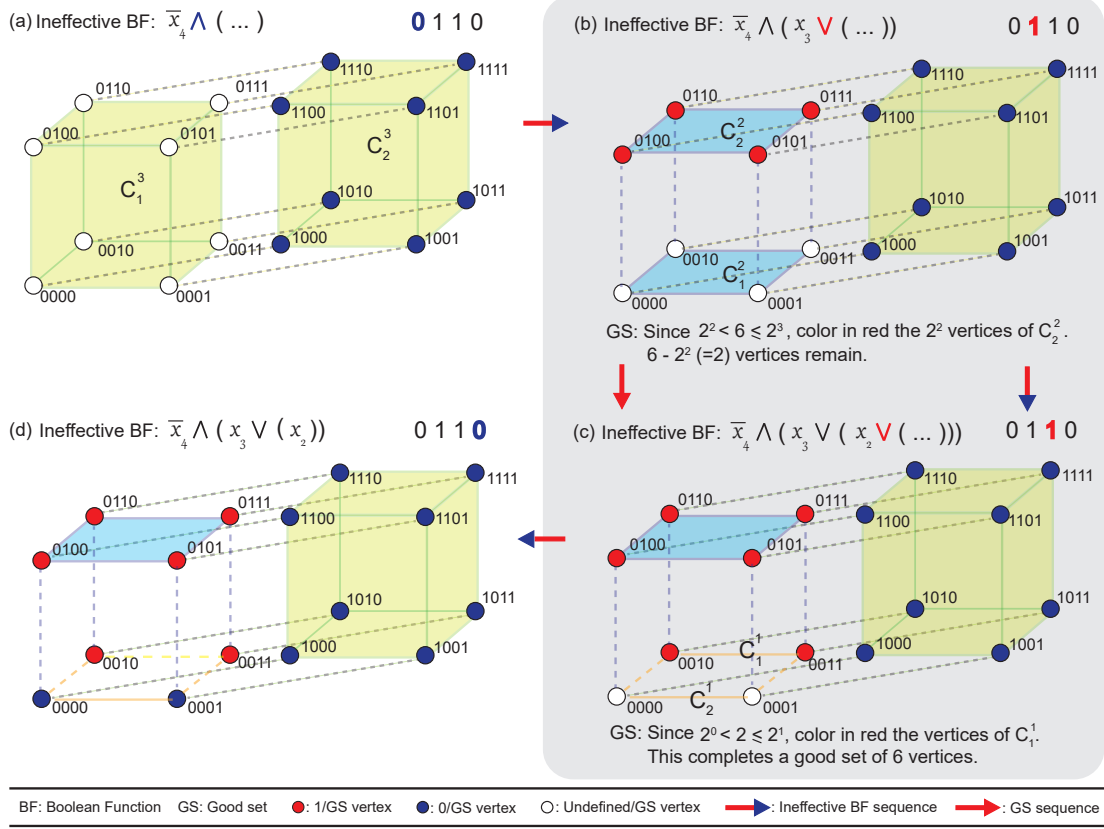


Figure 3.5: Good set (GS) for P vertices where P is even on a k -dimensional hypercube is equivalent to an IEF in that $k[P]$ set ($k = 4, P = 6$). In parts (b) and (c) shaded in grey, we show the recursive construction of a GS for $P = 6$ vertices in a 4-dimensional hypercube by coloring its vertices red, and in parts (a), (b), (c) and (d), we show the equivalence of that GS with 6 vertices to an IEF with bias 6. The vertices of the hypercube are labeled in the order x_4, x_3, x_2, x_1 wherein x_i is 0 or 1. Here, C_1^j and C_2^j denote the two vertex disjoint j -dimensional hypercubes of the $(j + 1)$ -dimensional hypercube. The active bit in each part (a), (b), (c) and (d) is the colored bit in the binary representation of 6 in that part. **(a)** The vertices with $x_4 = 1$ are set to the output value 0. The active bit in this step is 0 and as a result the \wedge operator follows the literal \bar{x}_4 . **(b)** Since $P = 6$ lies between 2^2 and 2^3 , 2^2 vertices of either C_1^2 or C_2^2 (here, C_2^2) form part of the GS. This leaves $6 - 4 = 2$ vertices to be colored to complete the GS. This choice of 4 vertices in C_2^2 for the GS leads to setting the output value of vertices labeled $x_4 = 0$ and $x_3 = 1$ to 1. The active bit in this step is 1 and as a result the \vee operator follows the literal x_3 . **(c)** Since the remaining 2 vertices lies between 2^1 and 2^2 , 2^1 vertices of either C_1^1 or C_2^1 (here, C_1^1) form part of the GS. This completes the GS of 6 vertices. This choice of 2 vertices in C_1^1 for the GS leads to setting the output value of vertices labeled $x_4 = 0$, $x_3 = 0$, and $x_2 = 1$ to 1. The active bit in this step is 1 and as a result the \vee operator follows the literal x_2 . **(d)** Since the two remaining undefined vertices take the same value 0, the variable x_1 does not appear in the expression of the BF. Thus, the BF corresponding to GS with 6 vertices is ineffective.

strictly less than k .

Proof: The arrangement of P 1s and $2^k - P$ 0s on the k -cube in the case where P is even is almost the same as in a NCF, with the exception that the vertices of the last 1-cube (composed of 2 vertex disjoint sets of 0-cubes) will have the same output values b_k . By direct computation, we have $P = \sum_{i=1}^k b_i 2^{k-i}$ which is always even.

Now consider the construction of the DNF of a BF (with bias P) defined by such a good set. Suppose that in the recursive construction of the good set one begins by assigning 1s to the vertices of a j -cube ($j < k$). The first clause of the DNF is then just the AND (product) of all the $k - j$ literals involved to fill the vertices of that j -cube. If the next step of the recursive construction of the good set consists in assigning 1s to a i -cube ($i < j$), the second clause of the DNF will be the product of all $k - i$ previous literals. We can thus iteratively construct the DNF for the BF represented by the given good set.

Since the vertices get filled by 0s or 1s hierarchically from a j -cube to $(j - 1)$ -cube, after filling the 1-cube, we are left with another 1-cube to be filled. When output values of the vertices of this last 1-cube are to be fixed, both vertices have to be set to the same output value since P is even. Thus they will either contribute a clause with $k - 1$ variables to the DNF expression (if the output values are set to 1) or they will not contribute any clause (if the output values are set to 0). Importantly, the variable which is missing in this clause is not present in any of the other clauses, therefore making that BF ineffective in that input. In constructing such a function, there will be at most $k - 1$ variables in the Boolean expression. This implies that the resulting function has a Boolean complexity strictly less than k . See Figure 3.5 for a visual proof of the above argument.

3.5 Implications of using biologically meaningful BFs for Boolean network dynamics

3.5.1 Computing the distributions of network average sensitivities

A natural question that emerges from our results is: what are the implications of selecting these various types of BFs for the network dynamics? To answer this, we exploit the indicator defined in [46, 71] referred to as *network average sensitivity*. This quantity is the mean, over all nodes of the network, of each node’s average sensitivity. Daniels *et al.* [46] found that by fixing the biological network structure and selecting CFs over random BFs for all nodes, the network average sensitivity s of the resulting Boolean network is brought close to the critical value $s \sim 1$. We extend this approach to consider the effects of selecting for the different biologically meaningful BFs, determining the distribution of network average sensitivities over the 88 models (see Figure 3.6). To determine the consequences of using different types of BFs in a network, we keep its structure (list of inputs to each node) but assign to each node a random function belonging to a particular type of BF (for example EF or CF), and compute the network average sensitivity of the resulting Boolean model. For each biological network and a particular type of BF, we repeat the above procedure 1000 times and store the sampled data points. We performed this for all 88 models in our reference biological dataset using a broad range of BFs such as: EF, EUF, CF, ECF, NCF, RoF, non-NCF RoF. Finally, we plot the distribution for the obtained data points as a violin plot (see Figure 3.6). Note that for the biological case there are only 88 data points corresponding to 88 networks or models, whereas in all other cases there are 88000 data points, as we sample 1000 data points for each type of BF per network. The computer programs used to generate biologically meaningful types of BFs and check if a BF belongs to one of them is available at: <https://github.com/asamallab/MCBF>. The procedure used to generate random k -input BFs for each of the types mentioned above is provided in Section B.1, Appendix B. Having generated the distributions, we then compare

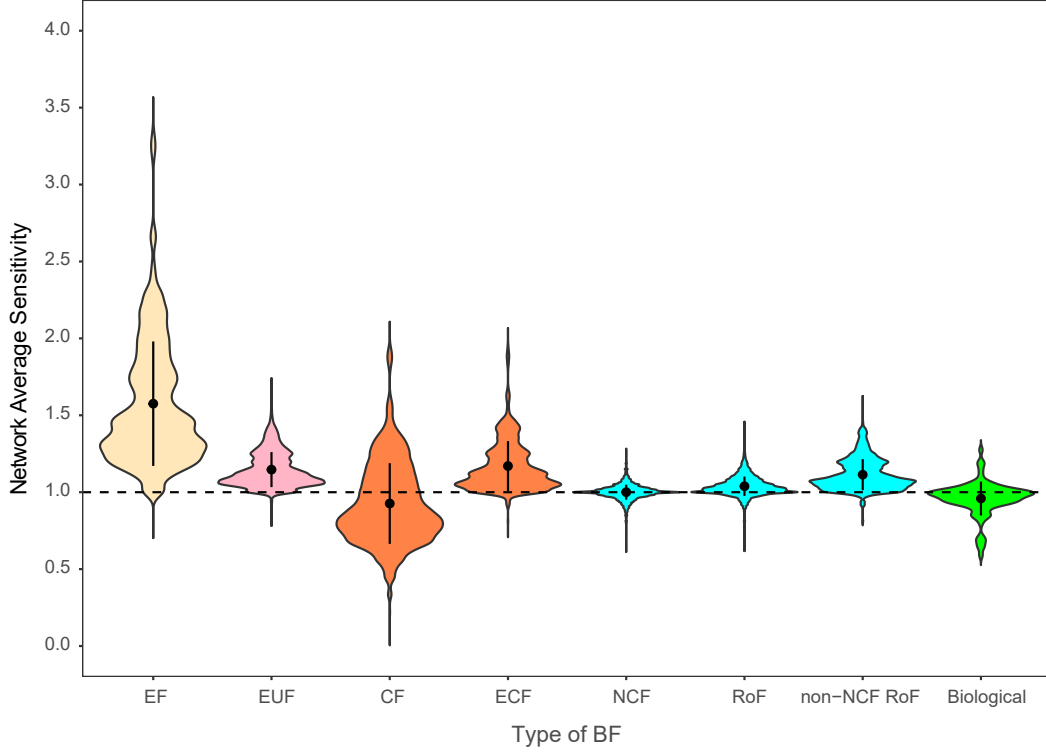


Figure 3.6: Distribution of the network average sensitivity when using the list of inputs from 88 biological models but enforcing different types of BFs to the nodes, namely effective functions (EF), effective and unate functions (EUF), canalizing functions (CF), effective and canalizing functions (ECF), nested canalizing functions (NCF), read-once functions (RoF) and non-NCF RoFs. The right-most case is the distribution when using the actual BFs in the biological models. This plot has been generated by keeping the maximum width of each of the violins fixed.

these distributions to that of the biological case.

3.5.2 Estimating the overlap between the distributions of network average sensitivities for various types of BFs and the biological case

Next, we quantify the overlaps of these different distributions and find that all types of BFs except for the NCFs and RoFs have a substantial fraction of their distributions lying outside the 95% confidence interval of the distribution of the biological case. To estimate the extent of overlap between the distribution of network average sensitivities corresponding to a particular type of BF and the biological case, we compute the fraction of data points (of the distribution of the BF we are interested in) which are outliers when

considering the biological distribution. The outlying regions are defined via the 5% of data points which fall on any one side of the biological distribution (one-sided test) or 2.5% of data points on either side (two-sided test) of the distribution. Note that if at the 2.5% (or 5%) threshold we get an network average sensitivity value for which there are multiple data points, then it may be that only some of these data points may fall in that 2.5%. If so, then that value of network average sensitivity is assigned a probability equal to the number of occurrences in the outlier divided by the total number of data points having that network average sensitivity (in the biological distribution). Thus when counting the number of data points (in the distribution of some type of BF) falling in the outliers of the biological distribution, only a fraction (equal to the probability) of those data points having the threshold value of network average sensitivity are counted as outliers.

Table 3.7: Quantifying the fraction of models in different ensembles with network average sensitivities (s) lying outside the distribution of s for biological networks. The percentage of data points that fall outside the 95% confidence interval of the biological case in the distribution of network average sensitivities when using the list of inputs from biological models but enforcing different types of BFs to the nodes, namely effective functions (EF), effective and unate functions (EUF), canalizing functions (CF), effective and canalizing functions (ECF), nested canalizing functions (NCF), read-once functions (RoF) and non-NCF RoFs. The distribution of network average sensitivities is shown in Figure 3.6 and data for both one-sided tests and two-sided tests are provided here.

Type of BF	One-sided (upper 5%)	One-sided (lower 5%)	Two-sided (2.5% on either side)
EF	92.61	0.0	87.27
EUF	39.35	0.0	30.69
CF	20.38	16.61	26.75
ECF	43.13	0.0	35.05
NCF	0.75	0.0	0.04
RoF	5.92	0.0	2.29
non-NCF RoF	32.4	0.0	22.4

3.5.3 Ensembles generated with NCFs and RoFs have the maximum overlaps with biological case

By quantifying the overlaps of these different distributions, we find that all types of BFs except for the NCFs and RoFs have a substantial fraction of their distributions lying

outside the 95% confidence interval of the distribution of the biological case (see Table 3.7). It is clear that the larger the fraction of data points that are outliers to the biological distribution, the more distant that distribution is from the biological case. From the data in Table 3.7 for the two-sided test, we can arrange various BFs based on their increasing proximity to the biological distribution in the following manner: $EF < ECF < EUF < CF < \text{non-NCF RoF} < \text{RoF} < \text{NCF}$. Furthermore, we see that RoFs and NCFs have rather narrow distributions that are peaked near $s = 1$ (see Figure 3.6).

3.6 Results from the repeat analyses after discarding the ineffective inputs to BFs in the reference biological dataset

In our reference biological dataset of 2687 BFs from 88 models, there are 63 IEFs. Such IEFs in the reference biological dataset are likely reconstruction errors in the model, and a possible way to mitigate any influence of these IEFs on the results from our analyses is by considering the truncated BF without the ineffective inputs. That is, for all of the 63 IEFs in the reference biological dataset, we discard the ineffective inputs and consider the corresponding truncated EF. For instance, if a k -input BF has j ineffective inputs (where $k > j$), then the effective number of inputs in the BF is $k_{eff} = k - j$, which is also equal to the number of inputs in the truncated EF.

To confirm that the conclusions of this study are not affected by these IEFs, we repeated our analyses (including relative abundance of biologically meaningful BFs and associated statistical tests, and distributions of network average sensitivities for the 88 models) by considering a modified reference biological dataset of 2687 BFs wherein each of the 63 IEFs are replaced by their corresponding truncated EFs. The associated results are reported in Figures C.1 and C.2 in Appendix C, and Tables C.1 - C.7 in Appendix C. From these additional figures and tables, it is evident that all the conclusions we reach using the 2687 BFs (including the IEFs) in the reference biological dataset, remain unchanged when IEFs are replaced by their corresponding truncated EFs with k_{eff} effective inputs

in the modified dataset.

3.7 Discussion

One of our main conclusions is that these biologically meaningful types of BFs represent a tiny fraction of the space of all BFs, and yet we find that they cover nearly all BFs found in our reference biological dataset. Of course this dataset may reflect some biases introduced by the researchers who built the associated models but the diversity of groups involved in building these models points to the solidity of our conclusions.

Another major conclusion we reach is that RoFs and their subset NCFs are specifically and strongly enriched in the reference biological dataset. We remark that while the relative abundance of CFs and NCFs in biological networks has been previously reported in several publications [22, 36, 37, 46, 86, 103, 104], our work provides a systematic study of 7 different types of BFs in a large curated reference biological dataset. In fact, previous studies neither carried out statistical tests nor assessed the relative enrichments in sub-types, e.g. NCFs within CFs or RoFs, and in this respect, our study is able to shed light on possible factors driving enrichment. The specific enrichment of RoFs and NCFs can be tied to their minimizing two measures of complexity namely, Boolean complexity [69, 94] and average sensitivity [71, 92]. RoFs turn out to be the set of BFs minimizing Boolean complexity. Furthermore, extending previous studies realizing that NCFs have low average sensitivity [39, 105, 106], we show that in fact NCFs achieve the *theoretical minimum* of this complexity measure in their $k[P]$ set, a result that was also reported in [107, 108].

The framework we use both supports and formalizes Kauffman’s [22] qualitative view in which *simplicity* should be a driver of the regulatory logic in biological systems. Kauffman argued that CFs were simpler than random functions, and therefore should be expected to arise quite frequently in biological systems [22, 37]. Our use of an extensive curated dataset generated from published Boolean models of biological networks enabled us to compare different notions of simplicity, and thereby confront Kauffman’s view to real data in a well defined quantitative framework. By identifying simplicity with minimum complexity defined in terms of either Boolean complexity or average sensitivity, NCFs are

the simplest of all BFs. We can thus justify the much stronger preponderance of the NCF type in comparison to the CF type conjectured by Kauffman.

In the reference biological dataset, we found occurrences of IEFs even though the corresponding models had been curated by their authors. Most likely such cases are modeling errors. A possible way to handle an IEF in such a biological context is by considering the truncated BF without its ineffective inputs. We have confirmed that all our conclusions remain unchanged by repeating the analysis starting with a modified reference biological dataset wherein every IEF is replaced by its corresponding truncated effective BF.

Lastly, our methods and results have implications for the problem of model selection within the Boolean framework [54, 109] as we will see in the next chapter. By model selection we mean the process of selecting Boolean models from the ensemble of Boolean models which satisfy given constraints such as having specified steady states. During model selection, the preferential use of NCFs or RoFs could serve as a relevant criterion to constrain network reconstruction [54, 110].

Data and code availability statement

The data on the 2687 BFs and the type of biologically meaningful BF they belong to, and codes related to statistical tests and complexity measures is provided in the GitHub repository: <https://github.com/asamallab/MCBF>.

Chapter 4

Leveraging developmental landscapes for model selection in Boolean models of gene regulatory networks

Current efforts to reconstruct Boolean developmental gene regulatory networks (DGRNs) are generally underdetermined, *i.e.*, there exists many combinations of regulatory logic rules that can recover the desired gene expression patterns [109], even for a given network structure [54]. Without additional information, modelers typically have to fix somewhat arbitrarily certain logic rules, a process that introduces hidden biases and preferences that are never made explicit. This chapter aims to address this unmet need of providing a systematic framework for *model selection* of Boolean DGRNs from an ensemble of models that are equally plausible at the level of their logic rules. To do so, we work with the *hierarchy* of cell types emerging from the *relative stability* (RS) associated with the system's developmental landscape and demonstrate how that information can be used to select between otherwise equivalent models. Though the genesis of using the developmental landscape in Boolean model selection goes back to the work by Zhou *et al.* [54], to

date the idea has been explored only at a small scale. In what follows we concretize and extend the ideas in [54,67] into a systematic framework that leverages the developmental landscape to perform model selection for larger networks.

First, we explore the literature for various measures of RS in Boolean models of DGRNs that have been introduced so far. We then quantify the concordance between these measures of RS using different ensembles derived from a Root Stem Cell Niche (RSCN) network [52] and a Pancreas differentiation network [54]. Using one of those RS measures, namely the MFPT, we show how to construct an associated potential cellular lineage tree and determine the frequency of occurrence of different lineage trees in the above mentioned ensembles. In addition, because the matrix formalism to calculate MFPT as proposed in [54] does not scale up computationally, we take a stochastic approach to compute the MFPT. With this method, we identify the relative orderings and cellular lineage trees for the successive root development models of Alvarez-Buylla’s group [73–75] that have increasing complexity. We find that the latest model proposed by that group does not satisfy the expected hierarchy, which indicates that RS between cell types was not one of their (conscious or not) criteria for selecting the logic rules. Lastly, we propose an iterative greedy search algorithm that leverages the expected developmental landscape (or hierarchies) to perform model selection from an ensemble of models that reproduce the biologically desired gene expression patterns. Thanks to these conceptual and computational developments, we provide a systematic framework to perform model selection within a biologically plausible ensemble of Boolean models using the associated developmental landscapes. **The work reported in this chapter is contained in the published manuscript [68].**

4.1 Relative stability and ordering of fixed points

In developmental dynamics, the propensity of a less differentiated cell type to transform into a more differentiated one is higher than the converse. This inherent asymmetry in cell state transitions forms the conceptual basis of RS [54,57] and is illustrated via a developmental landscape [111] as shown in Figure 4.1(a). To make this notion quantitative,

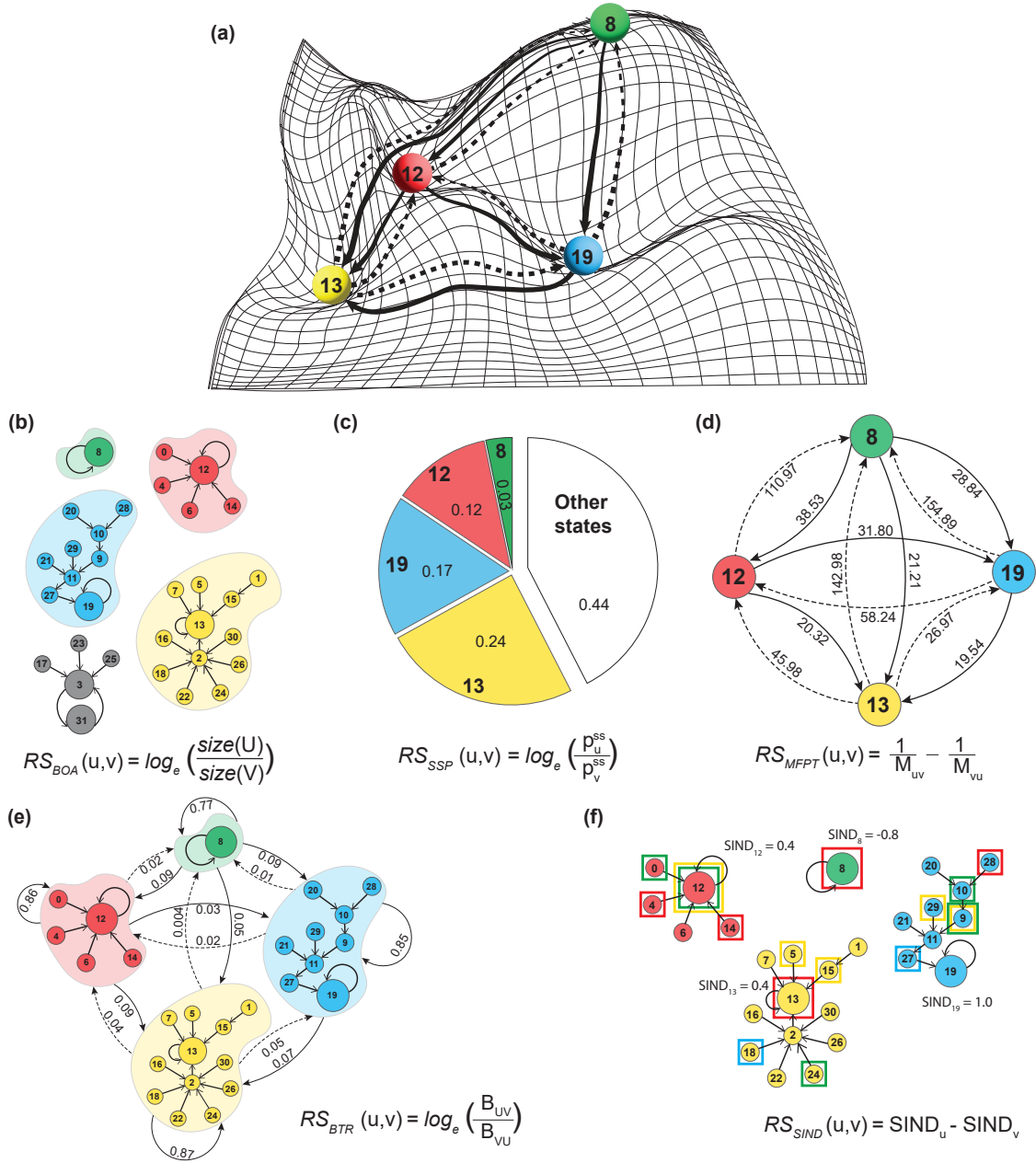


Figure 4.1: Developmental landscape inspired by Waddington and relative stability measures. The epigenetic landscape and its measures of RS are depicted here for the toy model displayed in Figure 1.1. Biological attractors are denoted by the symbols u and v and their basin of attraction by U and V respectively. (a) The colored balls and their numeric labels represent different cell types and their corresponding fixed point states (integer equivalent of the binary expression pattern) in the toy model. The balls are trapped in local minima but at different altitudes that indicate their RS . Solid and dashed lines indicate greater and lesser propensity respectively, for a cell state transition. (b) **Basin of attraction (BOA).** The graph's connected components correspond to the basins of attraction (same colors as the corresponding cell states on the landscape).

Figure 4.1 (previous page): (c) **Steady state probability (SSP)**. The pie chart gives the steady state probabilities of each fixed point (color coded) and of all the other states (non fixed points, in white). (d) **Mean First Passage Time (MFPT)**. The complete oriented digraph provides all the MFPT between fixed points. Solid and dashed lines indicate smaller and larger MFPTs respectively. (e) **Basin Transition Rate (BTR)**. A complete oriented digraph where the nodes are the basins of attraction of fixed points and the edges are the BTR from one basin to the other. Solid and dashed lines indicate larger and smaller BTRs respectively. (f) **Stability Index (SIND)**. A colored square indicates the 1-Hamming neighbors of the fixed point associated with that color. This information is used to compute the SIND.

we work in the mathematical framework proposed by Zhou *et al.* [54] that is outlined below. Let l and m be integer representations of any states of the network with N nodes so that $l, m \in \{0, 1, \dots, 2^N - 1\}$. Then, the deterministic dynamics can be represented via a matrix \mathbf{T} whose elements $T_{lm} = 1$ if updating the state m via Boolean functions (BFs) $\mathbf{F} = \{f_1, f_2, \dots, f_N\}$ gives the state l . Stochastic dynamics are introduced via a noise parameter η that flips the state of each gene independently with a probability η . Thus, let \mathbf{P} be the *perturbation* matrix whose entries P_{lm} give the probability that η alone drives the transition from m to l . If $d(l, m)$ is the Hamming distance between l and m , then P_{lm} is defined via:

$$P_{lm} = \begin{cases} \eta^{d(l,m)}(1-\eta)^{N-d(l,m)} & \text{if } l \neq m \\ 0 & \text{if } l = m \end{cases}$$

The (stochastic) dynamics of the DGRN is then defined via the transition matrix:

$$\mathbf{T}^* = (1 - \eta)^N \mathbf{T} + \mathbf{P} \tag{4.1}$$

It is easy to see that \mathbf{T}^* is an ergodic Markov chain, so quantities defined for such chains can be carried over to these Boolean networks (BNs) [54, 67]. In particular, each column

of \mathbf{T}^* sums to 1. Specifically, $\sum_{l=0}^{2^N-1} T_{lm}^* = 1, \forall m \in \{0, 1, \dots, 2^N - 1\}$ since:

$$\sum_{l=0}^{2^N-1} T_{lm}^* = \sum_{l=0}^{2^N-1} (1-\eta)^N T_{lm} + \eta^{d(l,m)} (1-\eta)^{N-d(l,m)} \quad (4.2)$$

$$= (1-\eta)^N + \sum_{r=1}^N \binom{N}{r} \eta^r (1-\eta)^{N-r} \quad (4.3)$$

$$= (\eta + (1-\eta))^N \quad (4.4)$$

$$= 1$$

In brief, their framework allows for transitions between all pairs of states in the state space (which is not possible in the deterministic Boolean modeling) by introducing stochasticity into the dynamics. This stochasticity has been interpreted as intrinsic noise arising from the stochastic gene expression dynamics [67].

In the ensuing text, we describe the 5 measures of RS proposed based on [54, 67], see also Figure 4.1(b-f) for the formulas of the RS between any pair of biological attractors u and v . Let the *basin of attraction* (BOA) of the attractors u and v be denoted by U and V . We denote by $RS_{measure}(u, v)$ the RS between u and v , where $measure \in \{BOA, SSP, MFPT, BTR, SIND\}$. $RS_{measure}(u, v) > 0$ implies that u is more stable than v . The 5 different measures are explained below:

4.1.1 Basin of Attraction

The basin of attraction (BOA) U of an attractor u comprises the set of all states that can reach the attractor u by applying deterministic update rules. The size of the basin of attraction U is the number of states in U , which we denote by $size(U)$. The size of a basin of attraction of an attractor plays an important role in determining that attractor's robustness to perturbations. Note that features of the structure of the basin of attractions of attractors may also be important determinants of the dynamics of Boolean networks, but are not pursued in this thesis. Perturbations to attractors with a small basin of attraction are generally more likely to take the system to a larger basin of attraction of a different attractor. Therefore, cell states (which are biological fixed point attractors) with

larger basin sizes are expected to exhibit greater stability in terms of its robustness to perturbations, compared to cell states with smaller basin sizes. We quantify this notion of relative stability between the pair of attractors u and v as follows:

$$RS_{BOA}(u, v) = \log_e \left(\frac{\text{size}(U)}{\text{size}(V)} \right) \quad (4.5)$$

4.1.2 Steady State Probability

Under the dynamics \mathbf{T}^* , the network is found in different states with certain probabilities which do not change over time at the steady state, corresponding to the steady state probability distribution. The *steady state probability* (SSP) of u (p_u^{ss}) is the probability of finding the network in state u under the steady state distribution. For any specified initial condition, let $\mathbf{p}(t)$ be the vector whose entry $p_l(t)$ is the probability of being in state l at time t . Note that $\sum_{l=0}^{2^N-1} p_l = 1$. Then the probability vector at the next time step is given by:

$$\mathbf{p}(t+1) = \mathbf{T}^* \mathbf{p}(t) \quad (4.6)$$

If $\mathbf{p}(t+1) = \mathbf{p}(t)$, then $\mathbf{p}(t)$ is necessarily (unique because of ergodicity) the steady state probability distribution \mathbf{p}^{ss} of the network states.

$$RS_{SSP}(u, v) = \log_e \left(\frac{p_u^{ss}}{p_v^{ss}} \right) \quad (4.7)$$

where p_u^{ss} is the SSP of the fixed point u and similarly for v .

4.1.3 Mean First Passage Time

The number of steps along a state space trajectory starting at state m and terminating at the first occurrence of l in a stochastic process is called the first passage time from state m to l . Its average over a large number of trajectories is then the *mean first passage time* (MFPT) from m to l and is denoted by M_{lm} . The MFPT of an ergodic Markov chain

(\mathbf{T}^*) can be calculated analytically using the fundamental matrix \mathbf{Z} [112] as follows:

$$\mathbf{Z} = (\mathbf{I} - \mathbf{T}^* + \mathbf{W})^{-1} \quad (4.8)$$

where \mathbf{I} is the identity matrix and \mathbf{W} is a matrix whose columns are the vector \mathbf{p}^{ss} . Note the order of both these matrices is $2^N \times 2^N$. M_{lm} is given by:

$$M_{lm} = \frac{Z_{ll} - Z_{lm}}{p_l^{ss}} \quad (4.9)$$

The RS associated with the MFPT is then defined as [54]:

$$RS_{MFPT}(u, v) = \frac{1}{M_{uv}} - \frac{1}{M_{vu}} \quad (4.10)$$

4.1.4 Basin Transition Rate

The *basin transition rate* (BTR) B_{UV} is the probability to transition from any state in basin V to any state in basin U when applying the stochastic dynamics for one time step. Mathematically, it is defined [67] via the formula:

$$B_{UV} = \sum_{l \in U} \sum_{m \in V} T_{lm}^* / \text{size}(V) \quad (4.11)$$

where l and m denote states in U and V respectively, from which we define:

$$RS_{BTR}(u, v) = \log_e \left(\frac{B_{UV}}{B_{VU}} \right) \quad (4.12)$$

4.1.5 Stability Index

The *stability index* (SIND) of a fixed point u , $SIND_u$, is defined following [67] as:

$$SIND_u = \sum_l O_l^u - \sum_{m \neq u} O_u^m \quad (4.13)$$

where the first sum is over all fixed points and the second over all fixed points other than u . O_l^m is a ratio, whose numerator is the number of 1-Hamming neighbors of fixed point l belonging to the basin of fixed point m , and whose denominator is the total number of 1-Hamming neighbors of fixed point l . Using this we define:

$$RS_{SIND}(u, v) = SIND_u - SIND_v \quad (4.14)$$

The 5 measures of RS for the toy model in Figure 1.1 are illustrated in Figure 4.1(b-f). The equations that define these measures via the matrix formalism are referred to as *exact* values in what follows. Given n fixed points, a few of the $n(n-1)/2$ associated RS inequalities may be known from biology, thus providing a partial view of the landscape. In case all such inequalities are available, it is desirable to combine them to obtain a complete picture of the landscape. For example, for the 4 fixed points in Figure 4.1(a), R (red), B (blue), G (green) and Y (yellow), one gets 6 pairwise relations: $G < R$, $G < B$, $G < Y$, $R < B$, $R < Y$, $B < Y$ using the RS_{MFPT} . These can be combined into the linear hierarchy, $G < R < B < Y$, corresponding to a *total* order. But finding such a total order may not always be possible as there may be *inconsistencies* amongst the inequalities. If instead, we had the inequalities $G < R$, $G < B$, $G < Y$, $B < R$, $R < Y$ and $Y < B$, the last three make it impossible to find a total order. Such a situation leads us to go beyond linear hierarchies by using tree-based (partial) hierarchies as we explain in later sections.

4.2 Constructing biologically plausible ensembles for model selection

We now describe a part of our model selection framework where by successively imposing different biologically motivated constraints on a Boolean model with a fixed network structure, it is possible to converge to a smaller subset of models that are biologically relevant [54]. Let k_i be the number of inputs to a gene $i \in \{1, 2, \dots, N\}$ in the BN. Keeping the network structure fixed (i.e., list of input genes to each gene), without imposing any constraints on the logical update rules or truth tables, there are

$2^{2^{k_1}} \times 2^{2^{k_2}} \times 2^{2^{k_3}} \times \dots \times 2^{2^{k_i}} \times \dots \times 2^{2^{k_N}}$ possible combinations of BFs (and thus Boolean models), which is often times astronomical. The first constraint is to restrict the truth tables to respect the fixed point condition for each of the desired biological fixed points. Every fixed point will constrain the output of 1 row in all of the truth tables, so N fixed points will constrain at most N rows of every truth table. Note that multiple fixed points may lead to redundant constraints. This constraint guarantees the recovery of all the biological fixed points but nevertheless does not exclude the presence of other (possibly, irrelevant) attractors. When feasible, one could also demand that there be no irrelevant attractors (i.e., attractors that do not correspond to biological fixed points). The Boolean GRNs then recover *only* the desired biological attractors. Secondly, we impose that the BF at each node conforms to the activatory or inhibitory signs of its regulators. In other words, we ensure that the BFs are *sign conforming* with respect to the network structure. Third, the choice of BFs is restricted to nested canalizing functions (NCFs) (or some other choice such as effective function (EFs), or unate functions (UFs), or effective and unate functions (EUFs)). As a motivation for this constraint, it has recently been shown that NCFs possess the minimum average sensitivity [49] among BFs and confer critical dynamics to the model, a hallmark of the dynamics of GRNs [22, 46, 49, 113]. Finally, known *RS* constraints on the biological fixed points can be used to select for models that conform to the expected developmental landscape as we will detail in the subsequent sections.

We now quantitatively illustrate the above methodology using a Boolean model of *Arabidopsis thaliana* RSCN [52] whose network structure is provided in Figure D.1(a), Appendix D. This RSCN model (*model A* in [52]) has 9 nodes, 19 edges and 4 fixed points that correspond to the cell types: Quiescent center (QC), Vascular initials (VI), Cortex-Endodermis initials (CEI) and Columella epidermis initials (CEpI) (see Figure D.1(a), Appendix D). In what follows, we have ordered the nodes of this RSCN network as: PLT, AUXIN, ARF, AUXIAA, SHR, SCR, JKD, MGP, WOX5. The total number of models possible for this network without any constraints on the truth tables is $4 \times 4 \times 4 \times 4 \times 4 \times 65536 \times 16 \times 256 \times 4294967296 \approx 1.18 \times 10^{21}$. By imposing the fixed point constraints on the truth tables the number becomes $2 \times 2 \times 2 \times 2 \times 1 \times 4096 \times 2 \times 16 \times 268435456 = 562949953421312 \approx 5.63 \times 10^{14}$. Next, on imposing the sign conforming constraint, we get

$2 \times 2 \times 2 \times 2 \times 1 \times 70 \times 2 \times 2 \times 848 = 3799040 \approx 3.8 \times 10^6$. Further imposing the NCF constraint, the total number of models becomes $1 \times 1 \times 1 \times 1 \times 1 \times 17 \times 1 \times 1 \times 75 = 1275$. Demanding that *only* the desired fixed points be recovered, the number is further reduced to 170. Finally, imposing *RS* constraints (via MFPT) from the expected developmental landscape, namely, that QC (Quiescent center) be the least stable of all the fixed points [114], we are left with 80 models. Thus from approximately 10^{21} models, the space of viable models can be shrunk to just 80 models.

Since the last 2 constraints are imposed at the level of the model (not truth table), it may be computationally cumbersome to apply them to networks where the number of models are typically large even after imposing constraints on the truth tables. This necessitates the development of stochastic methods to enable model selection on larger ensembles of biologically plausible models.

4.2.1 Two biological models and their ensembles of DGRNs

Statistical analyses presented in this work are performed on ensembles of GRNs derived from two *benchmark* biological models. The first is a RSCN model of *Arabidopsis thaliana* [52] (see Figure D.1(a), Appendix D) and the other is a pancreatic cell differentiation model [54] (see Figure D.1(b), Appendix D). Keeping the network structure fixed for the RSCN [52] or Pancreas development [54] models constrain the truth table at each node using the desired biological fixed points and NCFs that are *sign conforming* with respect to the network structure. This gives us the first type of ensemble: DGRNs that recover *at least* the desired biological attractors using *sign conforming* NCFs (sc-NCFs). The ensembles for the RSCN and Pancreas differentiation models denoted by $Root_{sc-NCF}$ and $Panc_{sc-NCF}$ consist of 1275 and 3600 models respectively. The other type of ensemble is obtained from the previous one by discarding models that include non-biological attractors. These ensembles for the RSCN and Pancreas differentiation models are denoted by $Root_{sc-NCF}^*$ and $Panc_{sc-NCF}^*$, and consist of 170 and 109 models respectively. We remark here that the BF at the AUX node alone was fixed to the choice made in the original RSCN model. We further constructed analogous ensembles by imposing that the BFs at each node be

sc-EUFs rather than sc-NCFs, leading to 4 other ensembles: $Root_{sc-EUF}$ (36600 models), $Root_{sc-EUF}^*$ (1400 models), $Panc_{sc-EUF}$ (7056 models) and $Panc_{sc-EUF}^*$ (159 models).

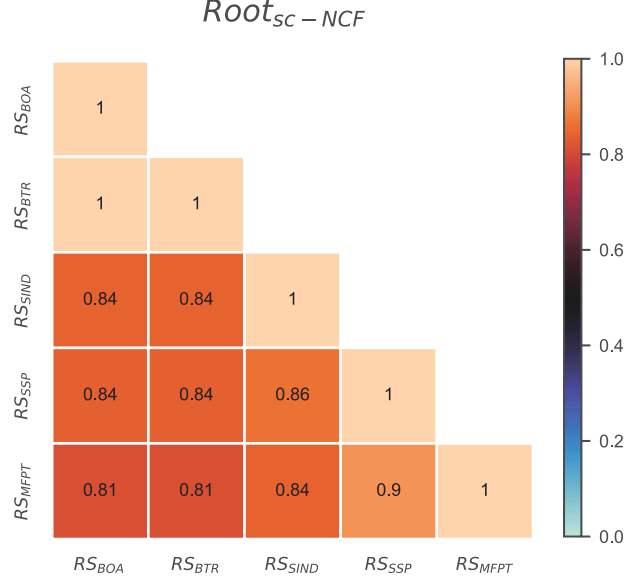


Figure 4.2: Pearson correlation between pairs of relative stability measures in the ensemble $Root_{sc-NCF}$. The rows and columns correspond to choices for the 5 RS measures. These 5 measures are based on size of basin of attraction (RS_{BOA}), basin transition rates (RS_{BTR}), a stability index (RS_{SIND}), steady state probabilities (RS_{SSP}) and mean first passage times (RS_{MFPT}). The heatmap indicates the value of the Pearson correlation coefficient between pairs of these measures. The measures were computed by exact means across all pairs of biological fixed points, for all 1275 models of $Root_{sc-NCF}$ ensemble, using a noise intensity parameter value of 1%.

4.3 The five measures of relative stability are strongly correlated with each other

The Pearson correlations between all 5 RS measures were computed for the ensemble $Root_{sc-NCF}$ (see Figure 4.2), showing that all measures are strongly correlated. Similar correlation heatmaps were constructed for the other sc-NCF ensembles, namely, $Root_{sc-NCF}^*$, $Panc_{sc-NCF}$ and $Panc_{sc-NCF}^*$ (see Figure D.2, Appendix D), and in all cases the RS measures are strongly correlated. Also from Figure 4.2 and Figure D.2, Appendix D, RS_{BOA} and RS_{BTR} are *perfectly* correlated, a result we prove in the following sub-section. The scatter plots for all pairs of measures (excluding RS_{BTR} since it is equivalent to RS_{BOA}) for the $Root_{sc-NCF}$ is shown in Figure 4.3. Similar scatter

plots for the other sc-NCF ensembles are provided in Figures D.3 - D.5, Appendix D. The Pearson correlation coefficients typically stay high even if one considers pairs of fixed points *separately* (see Figure 4.4 for all pairs of biological attractors of the $Root_{sc-NCF}$ ensemble). Similar correlation heatmaps were generated for the other sc-NCF ensembles and are provided in Figures D.6 - D.8, Appendix D. Individual scatter plots for all pairs of RS measures, for each pair of attractors were also plotted for the $Root_{sc-NCF}$ ensemble, one of which is shown in Figure 4.5 and the remaining are shown in Figures D.9 - D.13, Appendix D.

We also tested whether the correlations are dependent on the *type* of BFs assigned to the nodes, specifically using sc-EUFs instead of sc-NCFs. The corresponding Pearson correlation heatmaps reveal that despite using a different type of BF, namely the sc-EUFs, the correlations between the measures remains very high (see Figure 4.6 for the ensemble $Root_{sc-EUF}$) and Figure D.14, Appendix D for the remaining ensembles.

Since all 5 measures are strongly correlated, they will usually provide quite similar hierarchies in the landscapes. So we proceed with the MFPT for the remainder of this work as it captures cell state transitions more naturally. Indeed, MFPT is measured via trajectories traced out in the gene expression state space while transitioning from one cell type to another under the stochastic dynamics, whereas other measures do not refer to such dynamics. Furthermore, MFPT offers a richer representation of the landscapes in the form of trees (arborescences) as we later illustrate.

4.3.1 The relative stability measure for the BTR is identical to that of the BOA

Here, we prove that RS_{BOA} and RS_{BTR} are identical. We denote by l and m states that belong to the basins of attraction U and V respectively. Note that transitions between states belonging to different basins are caused by the presence of noise, specified via the matrix \mathbf{P} . Hence $T_{lm}^* = P_{lm}$ (if $l \in U$ and $m \in V$), where P_{lm} are the entries of the matrix \mathbf{P} . Since \mathbf{P} is a symmetric matrix, $T_{lm}^* = P_{lm} = P_{ml} = T_{ml}^*$. Then, starting with the

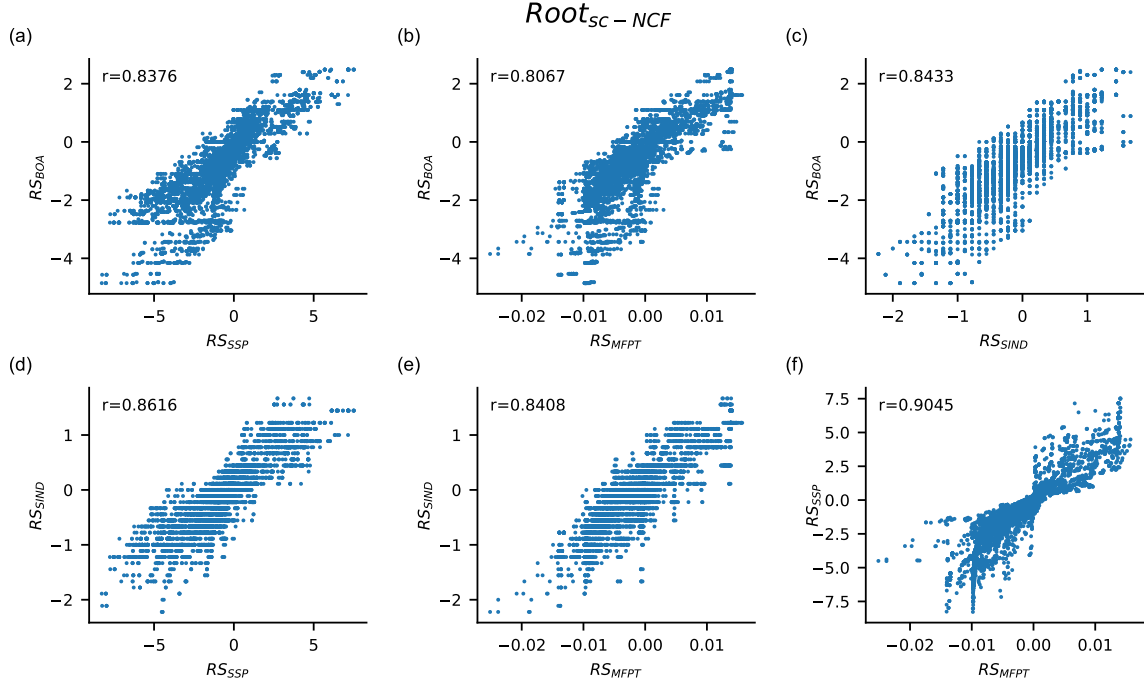


Figure 4.3: Scatter plots displaying values of relative stability in the ensemble $Root_{sc-NCF}$. Each sub-figure from (a) to (f) is a scatter plot where the x and y axes are for different measures of RS . These 5 measures are based on size of basin of attraction (RS_{BOA}), basin transition rates (RS_{BTR}), a stability index (RS_{SIND}), steady state probabilities (RS_{SSP}) and mean first passage times (RS_{MFPT}). These measures have been computed by the exact method for all pair of biological fixed points, for all 1275 models belonging to the ensemble $Root_{sc-NCF}$, at 1% noise. Of the 10 possible scatter plots for distinct pairs of the 5 RS measures, only 6 are shown here as RS_{BOA} and RS_{BTR} are equivalent. The Pearson correlation coefficient (r) for each scatter plot is computed and reported in the plot. These plots indicate that the correlation between the different RS measures is quite strong.

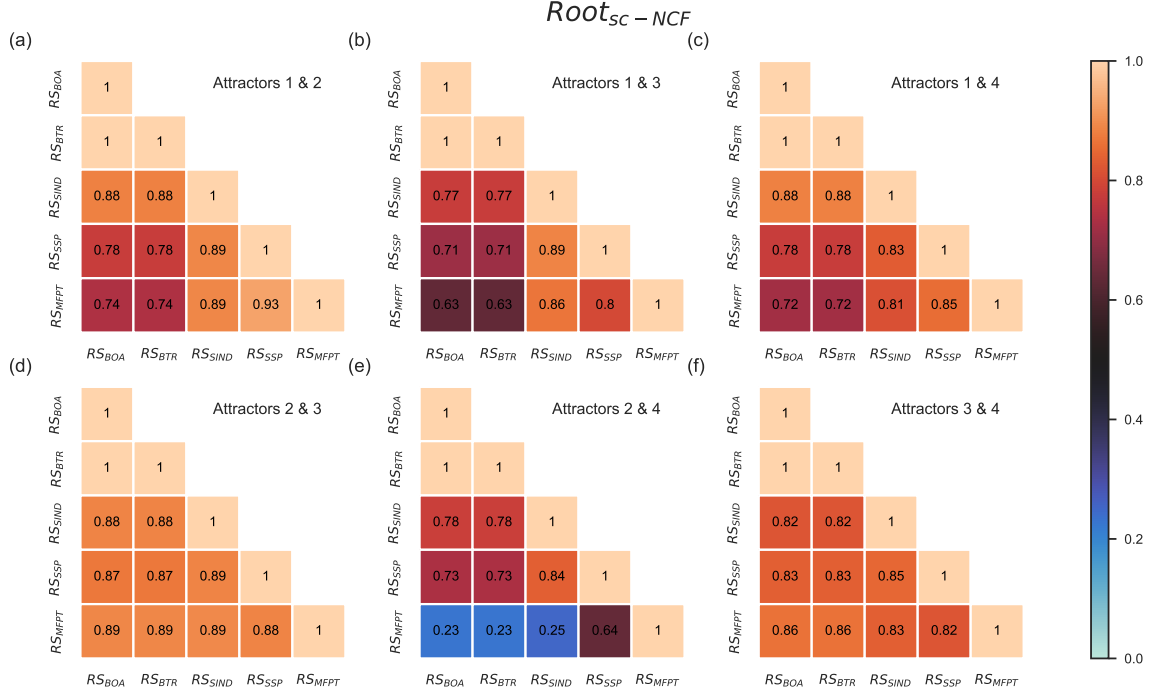


Figure 4.4: Pearson correlation between different pairs of relative stability measures for a given pair of fixed points for the ensemble $Root_{sc-NCF}$. The rows and columns of all heatmaps correspond to choices for the 5 measures of RS . These 5 measures are based on size of basin of attraction (RS_{BOA}), basin transition rates (RS_{BTR}), a stability index (RS_{SIND}), steady state probabilities (RS_{SSP}) and mean first passage times (RS_{MFPT}). The heatmaps indicate the Pearson correlation coefficient between pairs of these measures. For a particular sub-figure, these measures are computed by exact means for the pair of biological fixed points specified in that sub-figure, for all 1275 models in the ensemble $Root_{sc-NCF}$ using a noise intensity parameter value of 1%. Each biological attractor (fixed point) is numbered as follows. 1: Quiescent center (QC), 2: Vascular initials (VI), 3: Cortex-Endodermis initials (CEI), 4: Columella epidermis initials (CEpI). The upper triangular portion of the heatmap is not displayed as the heatmap entries constitute a symmetric matrix. Furthermore, RS_{BOA} and RS_{BTR} are perfectly correlated, an observation which we prove theoretically by showing that RS_{BOA} and RS_{BTR} are in fact equivalent.

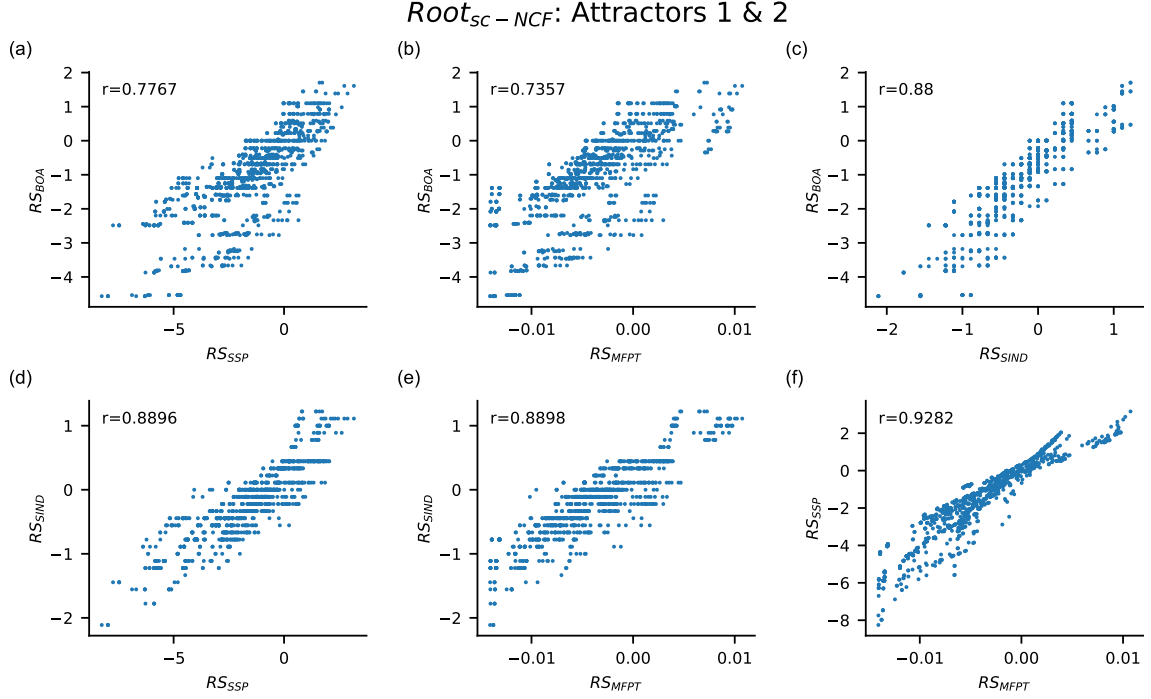


Figure 4.5: Scatter plots between the different pairs of relative stability measures for the pair of attractors 1 and 2 for the ensemble $Root_{sc-NCF}$. Each sub-figure from (a) to (f) is a scatter plot where the x and y axes are for different measures of RS . These 5 measures are based on size of basin of attraction (RS_{BOA}), basin transition rates (RS_{BTR}), a stability index (RS_{IND}), steady state probabilities (RS_{SSP}) and mean first passage times (RS_{MFPT}). All these measures have been computed by the exact method for the pair of biological fixed points 1 (Quiescent center (QC)) and 2 (Vascular initials (VI)), for all 1275 models belonging to the ensemble $Root_{sc-NCF}$, at 1% noise. Of the 10 possible scatter plots for distinct pairs of the 5 RS measures, only 6 are shown here as RS_{BOA} and RS_{BTR} are equivalent. The Pearson correlation coefficient (r) for each scatter plot is computed and reported in the plot.

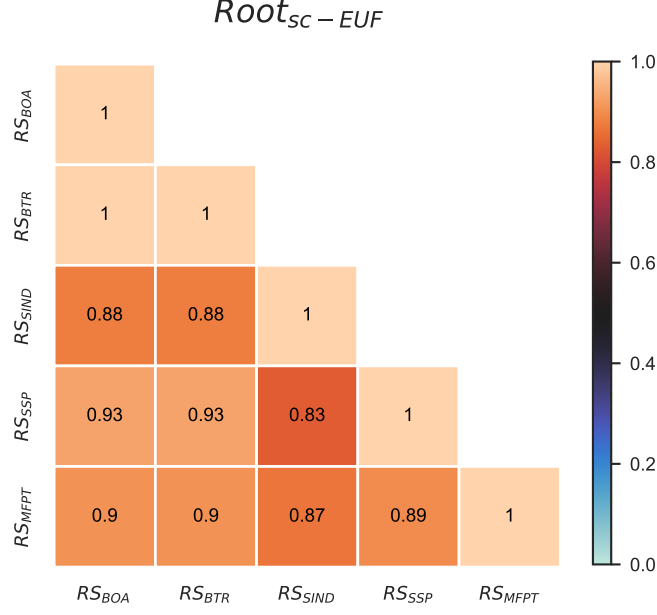


Figure 4.6: Pearson correlation between different pairs of relative stability measures for the ensemble $Root_{sc-EUF}$. The rows and columns correspond to choices for the RS measures. These 5 measures are based on size of basin of attraction (RS_{BOA}), basin transition rates (RS_{BTR}), a stability index (RS_{SIND}), steady state probabilities (RS_{SSP}) and mean first passage times (RS_{MFPT}). The heatmap indicates the value of the Pearson correlation coefficient between pairs of these measures. Note that these measures are computed by exact means across all pairs of biological fixed points, for all 36600 models in this ensemble $Root_{sc-EUF}$ using a noise intensity parameter value of 1%. The upper triangular portion of the heatmap is not displayed as the heatmap entries constitute a symmetric matrix. Furthermore, RS_{BOA} and RS_{BTR} are perfectly correlated, an observation which we prove theoretically by showing that RS_{BOA} and RS_{BTR} are in fact equivalent.

expression for $size(V)B_{UV}$, we have:

$$size(V)B_{UV} = \sum_{l \in U} \sum_{m \in V} T_{lm}^* = \sum_{l \in U} \sum_{m \in V} T_{ml}^* = \sum_{l \in V} \sum_{m \in U} T_{lm}^* = size(U)B_{VU} \quad (4.15)$$

As a consequence, $B_{UV}/B_{VU} = size(U)/size(V)$. Hence $RS_{BTR}(u, v) = RS_{BOA}(u, v)$.

4.4 Inferring cellular lineage trees using MFPT

4.4.1 Minimum Spanning Arborescence

A spanning tree of a connected undirected graph is a subgraph that is a tree and contains all the vertices of the graph. A minimum spanning tree of an undirected graph with

weighted edges is a spanning tree that has the minimum sum over its edge weights. An *arborescence* is a rooted, directed tree in which all edges are oriented away from the root. A minimum spanning arborescence (MSA) is the directed analog of the minimum spanning tree constructed from a directed graph with weighted edges. The total number of spanning arborescences for a complete digraph with n vertices (having distinctly labelled nodes) is n^{n-1} . This can be reasoned as follows. From Cayley’s theorem [115], the number of distinct (undirected) trees of n labeled vertices is n^{n-2} . To get an arborescence from an undirected tree, one simply has to specify the root, which gives a directed tree with edges that point away from the root. Since there are n ways to choose the root, there are $n \times n^{n-2} = n^{n-1}$ arborescences for n nodes.

4.4.2 Constructing a potential cellular lineage tree using the MFPT and MSA

Developmental trajectories are expected to follow paths of least resistance on the epigenetic landscape that thus can be summarized via a lineage tree taking one from undifferentiated to differentiated cells. A transition from an undifferentiated state to a more differentiated state should be more probable and take less time than a transition in the opposite direction, these times being provided in the associated MFPTs. We thus infer the lineage tree from the matrix \mathbf{M} whose entries M_{uv} give the MFPT for going from fixed point v to fixed point u (in presence of noise). \mathbf{M} thus corresponds to a complete weighted directed graph $G(\mathbf{M})$ whose nodes are biological attractors and edges carry the weights M_{uv} . The cell lineage tree should then correspond to the directed rooted tree that minimizes the sum of the MFPTs over its edges. Such a tree is precisely the MSA of $G(\mathbf{M})$. To construct a MSA from $G(\mathbf{M})$, we use the implementation of Edmond’s algorithm from the NetworkX package [116], namely, the *minimum_spanning_arborescence* module. This method may also be applied using other types of transition rates such as the BTR.

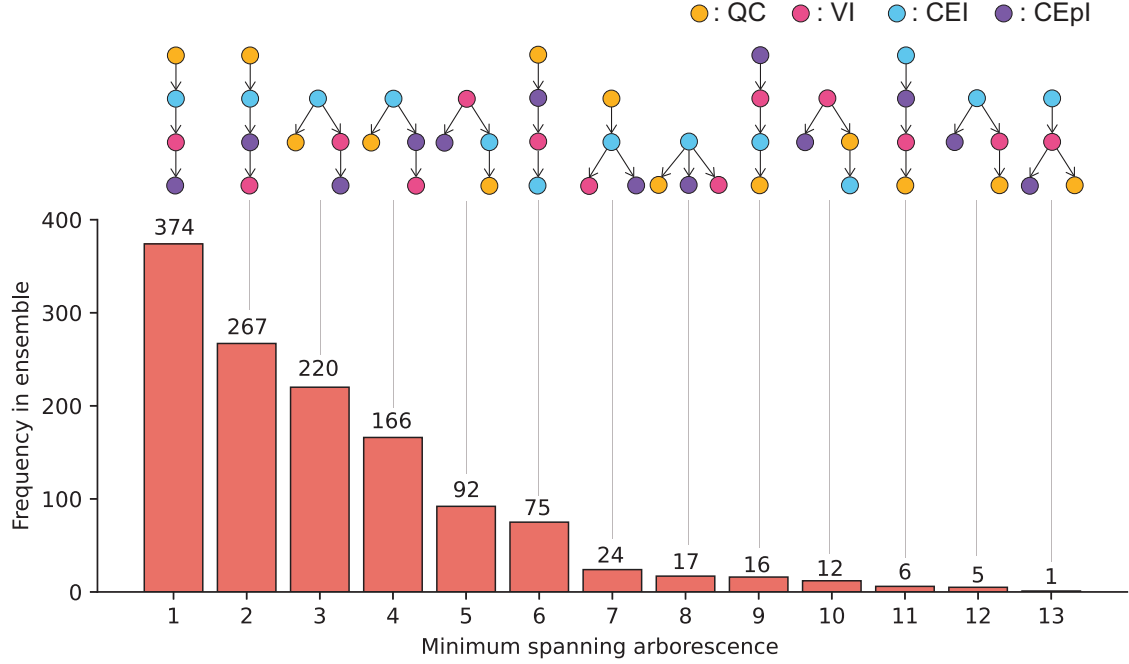


Figure 4.7: Frequency distribution of the minimum spanning arborescences (MSAs) in the ensemble $Root_{sc-NCF}$. The x -axis (top) labels the different MSAs that occur in this ensemble containing 1275 models with 1% noise. Of the 64 possible (labeled and oriented) trees for 4 fixed points, only 13 are realized. The y -axis is the frequency of each of these trees. The biological fixed points of the $Root_{sc-NCF}$ ensemble are as follows. QC: Quiescent center, VI: Vascular initials, CEI: Cortex-Endodermis initials and CEPI: Columella epidermis initials.

4.4.3 Distribution of lineage trees computed using MFPT for various ensembles

Above, we provided a prescription to generate a MSA from a MFPT matrix. Figure 4.7 shows the distribution of such MSAs for the $Root_{sc-NCF}$ ensemble (where MFPT is computed using the exact scheme with 1% noise). An immediate observation is that, of 64 possible trees and 5 possible tree topologies, only 13 trees and 4 tree topologies actually occur in the ensemble. Furthermore, not all 13 trees found respect the RS conditions suggested by the underlying biology. Specifically, the QC cell type is expected to be the least stable compared to the other cell types and therefore is expected to be the *root* of the tree. Thus only 4 trees out of the 13 appear to be biologically realistic. For the distribution of the MSA for other sc-NCF ensembles, see Figures D.15 - D.17, Appendix D.

4.5 Scaling up the computation of MFPT to reliably infer the relative stability of attractors in larger Boolean networks

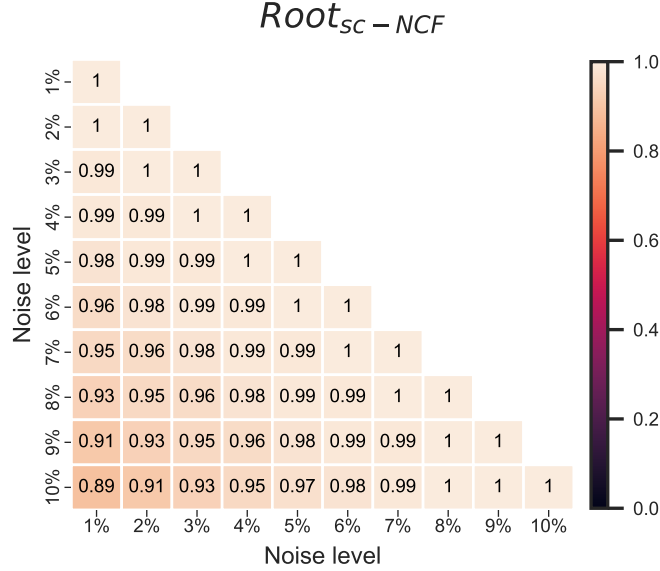


Figure 4.8: Pearson correlation between RS_{MFPT} values computed by exact methods for different pairs of noise values for the ensemble $Root_{sc-NCF}$. Rows and columns correspond to the noise intensities ranging from 1% to 10%. The heatmap gives the value of the Pearson correlation coefficient of RS_{MFPT} values when considering all pairs of biological fixed points and all 1275 models within the ensemble $Root_{sc-NCF}$, for different pairs of noise intensities. The upper triangular portion of the heatmap is not displayed because it constitutes a symmetric matrix. The correlation between the RS_{MFPT} for different values of noise is found to be very strong even for pairs of noise values which have a large difference.

4.5.1 Inferences drawn from MFPT are insensitive to changes in noise intensities

Before applying the MFPT to obtain a hierarchy of states for larger models, we test it on smaller ones. First, we compute the RS_{MFPT} values (for all pairs of biological fixed points) using the exact method, for different noise intensities ranging from 1% to 10%, for all models in the ensemble. Then for different pairs of noise values, we calculate the Pearson correlation coefficient between these RS_{MFPT} values. We find that for all pairs

of noise values, RS_{MFPT} values are strongly correlated for all 4 ensembles. The heatmap for the $Root_{sc-NCF}$ ensemble is shown in Figure 4.8. See Figure D.18, Appendix D for the correlation heatmaps associated with the ensembles $Root_{sc-NCF}^*$, $Panc_{sc-NCF}$ and $Panc_{sc-NCF}^*$ respectively.

Next, we obtain the set of partial orders (inequalities for all pairs of fixed points) using RS_{MFPT} (computed using the exact method) for different noise intensities ranging from 1% to 10%. For each pair of noise values, we compute the number and fraction of models with at least one disagreement in their (partial) orders using the ensemble $Root_{sc-NCF}$, and plot them as a heatmap in Figure 4.9. Clearly, the fractions of disagreements are quite low even for large differences in noise values. These observations are recapitulated in other ensembles as well (see Figures D.19 - D.21, Appendix D).

These results reveal that the outcome of using a noise intensity of 5% will not differ much from that using a noise intensity of 1%. Thus, it is possible to use a larger noise intensity without affecting prediction power, the benefit being that it can greatly speed up the stochastic simulations.

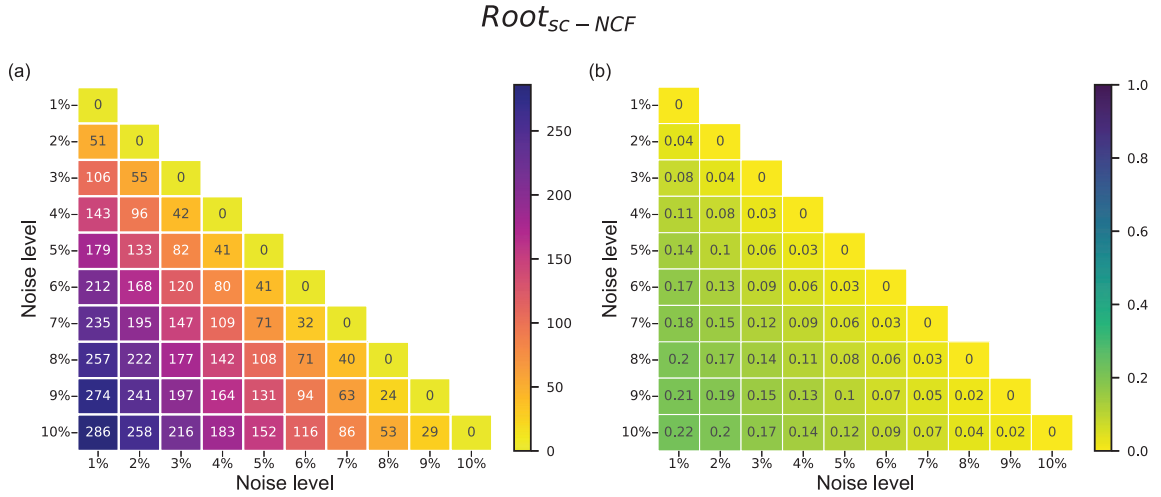


Figure 4.9: Number and fraction of models which differ in at least one comparison of partial ordering of the different biological fixed points when considering two different noise values, in the ensemble $Root_{sc-NCF}$. The (partial) order of two fixed points is specified via the MFPT values for going from one to the other, computed here using an exact method. Rows and columns correspond to the noise intensity. The heatmap (a) gives the number of models (out of a total of 1275) that differ in at least one (partial) order across pairs of biological fixed points. The heatmap (b) provides the same information but using the fraction of such models.

4.5.2 Stochastic approach to estimate the MFPT

The matrix (or exact) method [54, 67] for calculating the MFPT is not scalable to large network sizes: it requires storing a $2^N \times 2^N$ matrix for a network of N nodes, which is computationally unfeasible even for networks of size 16. We thus devised an approach based on stochastic simulations of trajectories with the following dynamics: If the noise does not alter the state of the network, then one applies the deterministic dynamics. The MFPT M_{uv} is the average of the number of time steps taken over a large number of trajectories starting at state v and evolved iteratively under the above-mentioned dynamics till state u is reached.

4.5.3 Comparison of the stochastic approach to the exact method of computing MFPT

Here we provide a comparison of our stochastic approach to compute MFPT described above, to its exact counterpart. First, for a given sc-NCF ensemble, we compute M_{uv} (MFPT from a biological fixed point v to another biological fixed point u) using both methods for various noise intensities (3%, 4%, 5%), taking for the stochastic method 500, 1500 and 2500 trajectories. For a given noise and number of trajectories, the results are presented as a scatter plot (see Figure 4.10) and via a table with Spearman and Kendall rank correlation coefficients (see Table D.1, Appendix D) for the ensemble $Root_{sc-NCF}$. Similar scatter plots and tables are provided for the other ensembles: $Root_{sc-NCF}^*$ (see Figure D.22, Appendix D and Table D.2, Appendix D), $Panc_{sc-NCF}$ (see Figure D.23, Appendix D and Table D.3, Appendix D) and $Panc_{sc-NCF}^*$ (see Figure D.24, Appendix D and Table D.4, Appendix D). As expected, these plots reveal that the stochastic method is in excellent agreement with the exact one, all the more so that one adds more and more trajectories.

To make this last claim more quantitative, we have used one model (chosen at random) from each of the 4 sc-NCF ensembles to test whether the M_{uv} values obtained from the stochastic method are statistically reliable. In effect, we ensure that by choos-

$Root_{sc-NCF}$

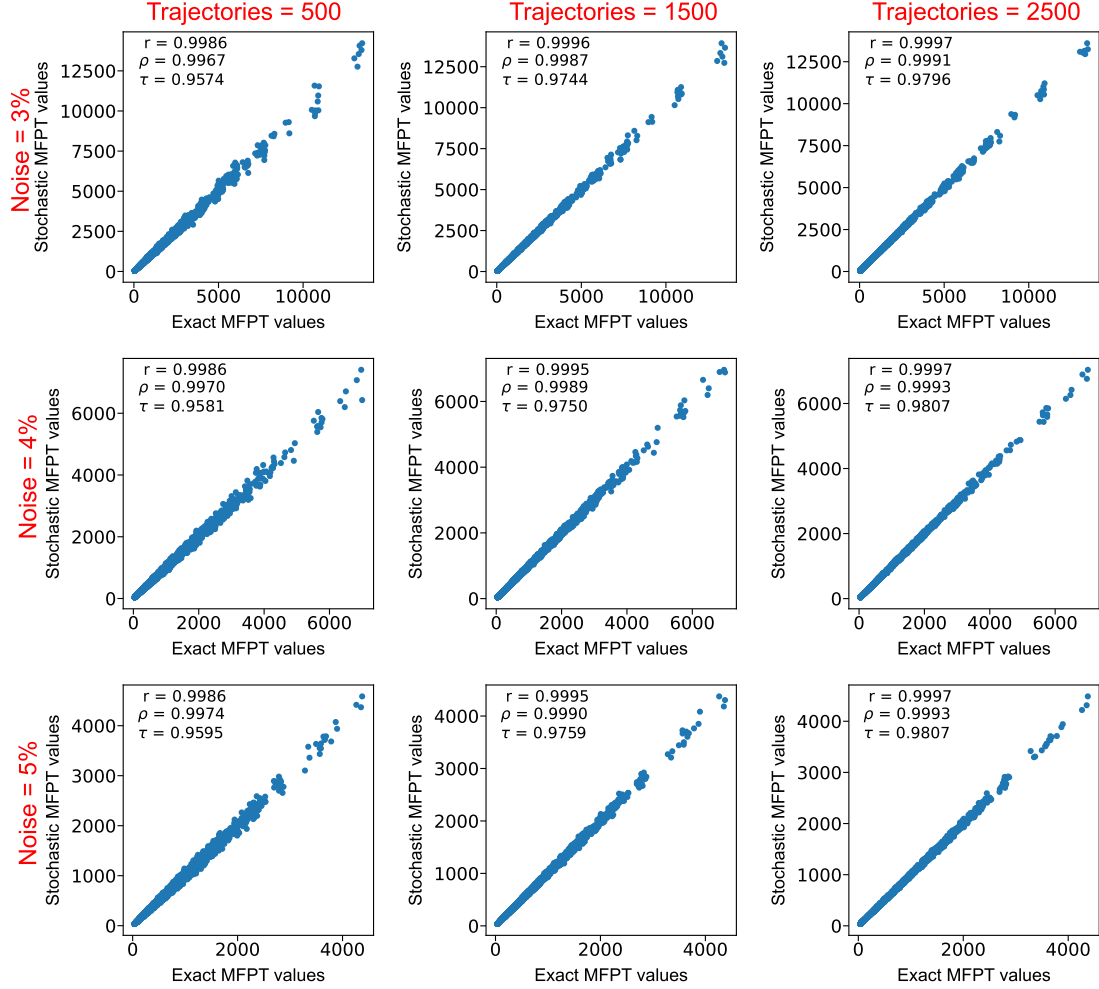


Figure 4.10: Correlation between the MFPT obtained via the exact method versus the proposed stochastic method using the ensemble $Root_{sc-NCF}$. The x and y axis of all scatter plots represent the MFPT from the biological fixed point v to the biological fixed point u (denoted by M_{uv}) computed via exact and stochastic means respectively, for all pairs of fixed points and for all 1275 models belonging to the ensemble $Root_{sc-NCF}$. Each scatter plot is generated for a particular noise (3%, 4% or 5%) and number of trajectories (500, 1500 or 2500) going from fixed point v to fixed point u . The exact and stochastic MFPT values are strongly correlated as can be seen from the 3 measures of correlation, namely, Pearson correlation coefficient (r), Spearman rank correlation coefficient (ρ) and Kendall rank correlation coefficient (τ). It can be seen that at a fixed noise as the number of trajectories are increased from 500 to 2500, the correlation becomes stronger across all 3 correlation measures.

ing a sufficiently large number of trajectories, the statistical (standard) error associated with the stochastic method is within acceptable limits. This is clear from the bar plots shown for each model from the 4 ensembles: $Root_{sc-NCf}$ (see Figure 4.11), $Root_{sc-NCf}^*$, $Panc_{sc-NCf}$ and $Panc_{sc-NCf}^*$ (see Figure D.25, Appendix D). Given that the stochastic approach to compute MFPT provides excellent agreement with the exact computation approach when using a sufficient number of simulated trajectories, this computational tool should provide a useful way to derive developmental landscapes for larger BN models.

4.6 *Arabidopsis thaliana* root development: A case study

As case studies, we considered 3 Boolean models of the *Arabidopsis thaliana* root development that have been reconstructed and published between the years 2013 and 2020 [73–75]. We show on Boolean DGRN models of *Arabidopsis thaliana* root development [73–75] how our methods can be used to obtain landscapes and enable model selection. These include: a 2013 RSCN model [73], a 2017 Root Apical Meristem (RAM) model [74] and a 2020 RSCN model [75]. The 2013 [73] and 2017 [74] studies presented multiple Boolean models from which we chose one per published article. Our choice was based on 2 simple criteria. One, that the model should recover most of the expected biological fixed points with the levels of the phytohormone auxin being high (see [75]). The other, that the fraction of state space occupied by the basins of biological fixed points having high auxin levels be the largest among all models proposed in that article. Of the 10 models in the 2013 [73] publication, those criteria led to choosing *model 4*. Of the 2 models in the 2017 [74] publication, it was the *GHRN1 model* that satisfied these criteria. In the 2020 article [75], only 1 model was provided. The models are given in the BoolNet format [117] - 2013 model [73] (see Table D.5, Appendix D), 2017 model [74] (see Table D.6, Appendix D) and 2020 model [75] (see Table D.7, Appendix D). The network structures and the biological fixed points (auxin being ON) for the 2013, 2017 and 2020 models are shown in Figures D.26(a), D.26(b) and D.27, Appendix D respectively. Starting from 2010, the Alvarez-Buylla group has refined their root development models over the years by the

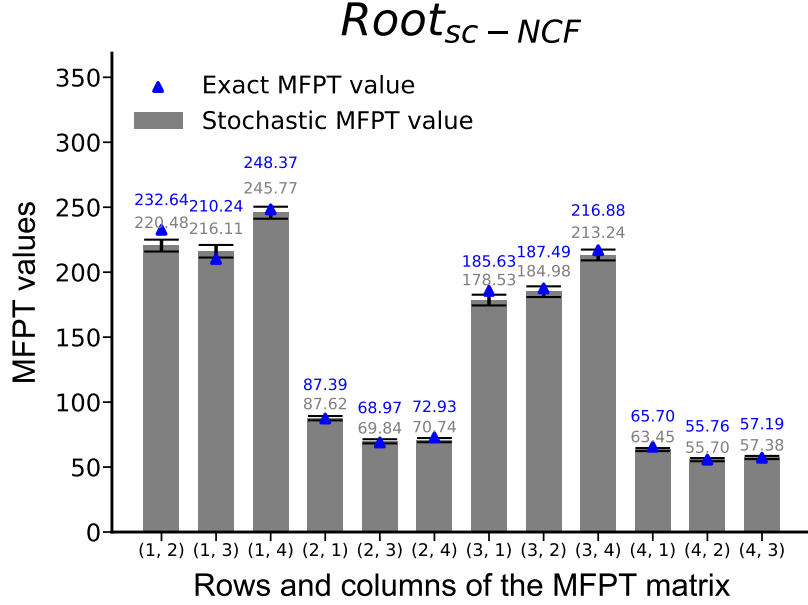


Figure 4.11: Barplot of the mean first passage time (MFPT) from one biological fixed point to another, computed via the stochastic method for a model taken from the ensemble $Root_{sc-NCF}$. The x -axis labels the rows and columns of the MFPT matrix entries, so for instance (1, 3) denotes the case of matrix element M_{13} when going from fixed point 3 to fixed point 1. The numbering of the fixed points are as follows. 1: Quiescent center (QC), 2: Vascular initials (VI), 3: Cortex-Endodermis initials (CEI) and 4: Columella epidermis initials (CEpI). The y -axis represents the associated mean first passage time (MFPT). It is computed following the stochastic approach, averaging over 2500 different trajectories of the dynamics starting from one fixed point and stopping as soon as the other fixed point is reached when using the rules for a particular Boolean model in the ensemble $Root_{sc-NCF}$, at 5% noise level. The tiny error bars indicate that the statistical error in the estimation of the MFPT value is very low. For comparison, the MFPT values obtained via the exact method are displayed via blue triangles (numerical values are provided above the bars in blue). The MFPT obtained via the proposed stochastic method is very close to that obtained via exact means. The grouping of bars according to the target fixed point visually illustrates the ease or difficulty of reaching a particular biological fixed point from the other three. For instance, in this particular Boolean model, it is difficult to go to the QC starting from any other fixed point and it is relatively easy to go to the CEpI starting from any other fixed point.

addition of new genes and interactions, resulting in more diverse cell types as can be seen in Figures D.26 and D.27, Appendix D.

4.6.1 Relative orderings of the biological fixed points

Though the published Boolean models [73–75] of *Arabidopsis thaliana* DGRNs all recover the expression patterns for the cell types observed in the root, they need not all conform to the expected developmental landscape. We thus computed the landscape hierarchies in the 2013 RSCN model [73], 2017 RAM model [74] and 2020 RSCN model [75] of *Arabidopsis thaliana*. Figure 4.12 shows the hierarchies obtained via 2 methods, using the basin of attraction and MSA. The MSA was constructed from the MFPTs computed using the stochastic method at 5% noise with 10000 trajectories. Experiments have shown that QCs undergo asymmetric cell division wherein one of the daughter cells differentiates to another cell type and the other stays of the QC type [114]. It is also known that the de-differentiation of other cell types to QC is rare. With this information we can impose an expected partial landscape of root development: QC is relatively less stable compared to all other cell types. We find that the 2013 and 2017 models recover this landscape hierarchy when using as *RS* measures BOA and MFPT (to obtain MSAs) (see Figure 4.12). However the 2020 model, though it is the most recent, does not recover the partial landscape, be-it via basin size or via MFPT as can be seen from Figure 4.12 (note that all these *Arabidopsis* models have been reconstructed by the same group of scientists). Specifically, the QC is more stable than the CEI/Endodermis, a relative ordering that is incompatible with experimental findings. It is also known that the cell type *Transition domain* (C. PTD2) is expected to more stable than the cell type *Central Pro-vascular initials* (C. PPD) [75] since the latter differentiates to the former. For this case it is worth noting that the *RS* associated with the basin of attraction of these cell types (C. PTD2 < C. PPD) violates the expected hierarchy (C. PPD < C. PTD2) whereas the *RS* associated with MFPT is in concordance with the expected hierarchy. These observations suggest that the developmental landscape was not considered during the reconstruction of this last Boolean model published in 2020. It also leads one to ask: how do we search for models that successfully recapitulate the expected landscape?

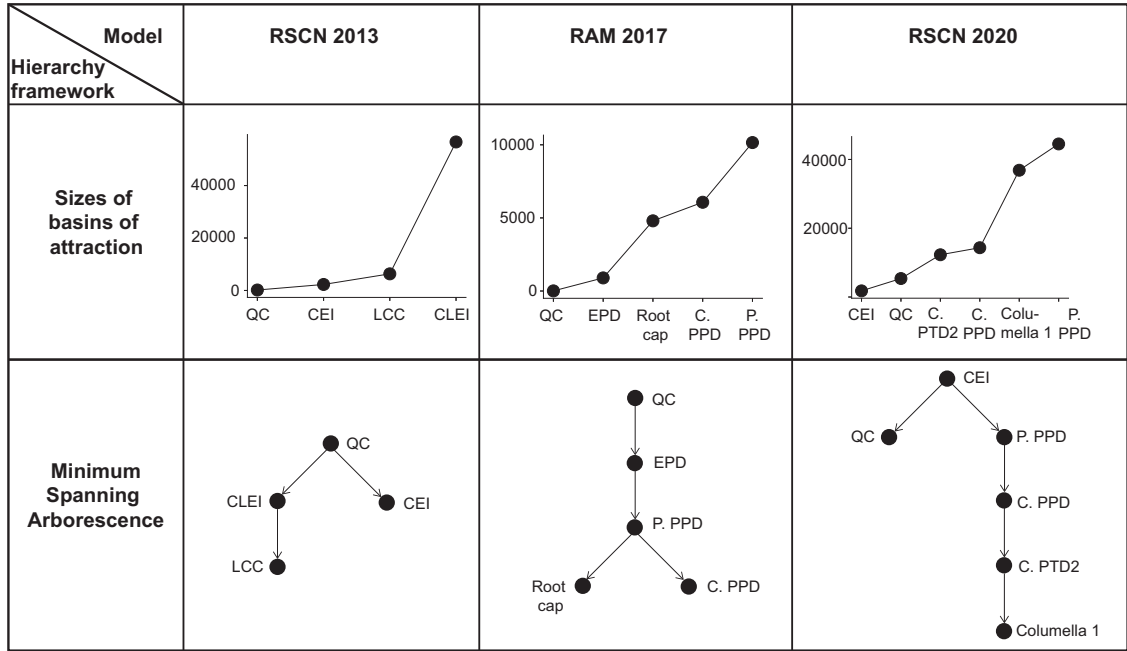


Figure 4.12: Hierarchy of fixed points based on basin sizes and minimum spanning arborescence (MSA) for the 2013, 2017 and 2020 root development models of *Arabidopsis thaliana*. The first row of this table shows the hierarchy of fixed points according to the basin size. For the 2013 and 2017 models, the QC has the smallest basin size whereas for the 2020 model, the CEI has the smallest basin size. The second row gives the hierarchy of states according to the MSA. The MSA was constructed using MFPTs obtained using the stochastic method under a noise intensity parameter value of 5% and averaging over 10000 trajectories. For the 2013 and 2017 models, the QC is at the root of the tree whereas for the 2020 model it is the CEI that is at the root of the tree. The expansions of the abbreviations for the 2013 model are as follows: QC: Quiescent center, CEI: Cortex-endodermis initials, LCC: Lateral root-cap and CLEI: Columella and lateral root-cap-epidermis initials. The expansions of the abbreviations for the 2017 model are as follows: EPD: Endodermis Proliferation Domain, P. PPD: Peripheral Pro-vascular Proliferation Domain, C. PPD: Central Pro-vascular Proliferation Domain. The expansions of the abbreviations for the 2020 model are as follows: CEI: Cortex/endodermis initial cell, P. PPD: Peripheral Pro-vascular initials, C. PPD: Central Pro-vascular initials, C. PTD2: Transition domain, Columella 1: Columella initials.

4.6.2 A greedy algorithm for Boolean model selection

Exhaustive exploration of ensembles are rarely feasible computationally so one needs to develop alternative approaches. Here, we present an iterative greedy algorithm (see Algorithm 2) to search for Boolean models that respect predefined constraints on the developmental landscape. Our algorithm is inspired in part by the Potts model (a generalization of the Ising model) [118] because for each gene we assign a given set of *Potts-like states* corresponding to its possible BFs. The algorithm requires: (1) an initial Boolean model that generally will not respect the expected pairwise orderings, (2) a dictionary of the allowed BFs for each gene, the genes being sorted in ascending order according to their number of possible BFs. The BFs allowed for a particular gene are obtained by imposing multiple constraints on the truth tables (see Section 4.2) and (3) the expected pairwise orderings of the cell types comprising the predefined constraints. Genes associated with a single BF are removed from this dictionary since there is no decision to make for their Boolean rule. Note: This does not mean that such nodes are removed from the GRN. It only means that those genes do not have to be visited during the selection of a model as there is only one allowed BF at that gene that recovers the biological attractors and satisfies the other constraint on the BFs. Therefore, such genes may be removed from the dictionary without altering the ensemble of models to be searched. Moreover, this results in a computational benefit - to visit only the set of genes with at least two allowed BFs - thereby improving the efficiency of our search algorithm.

Our algorithm repeatedly sweeps through the whole list of genes replacing the gene's current BF by a new choice at each iteration, and determining for the modified model all the pairwise orderings of fixed points. If the resulting trial model is worse with respect to the developmental landscape than the current model (not respecting as many predefined orderings) or with respect to the threshold set on the fraction of state space occupied by all basins of the biological attractors, the change is rejected and the algorithm continues without implementing the modification, otherwise the change is accepted.

Algorithm 2 Greedy algorithm to find Boolean models that obey a specified developmental landscape

```

1: GeneBFDictionary  $\leftarrow$  dictionary providing allowed BFs for each gene
2: CurrentBN  $\leftarrow$  initial BN whose hierarchy may not match the expected relative
   orderings
3: ThresholdSize  $\leftarrow$  minimum total basin size over all biological fixed points
4: while CurrentBN does not match the expected relative orderings do
5:   for Gene in GeneBFDictionary do
6:     TrialBF  $\leftarrow$  random BF from the BFs allowed at Gene
7:     TrialBN  $\leftarrow$  CurrentBN after replacing BF of Gene with TrialBF
8:     if basin sizes of fixed points of TrialBN  $<$  ThresholdSize then
9:       break
10:    else
11:      CurrentBNHierarchy  $\leftarrow$  List of pairwise orderings for biological
        fixed points of CurrentBN
12:      TrialBNHierarchy  $\leftarrow$  List of pairwise orderings for biological fixed
        points of TrialBN
13:      if TrialBNHierarchy matches the expected landscape constraints
        then
14:        TrialBNStatus  $\leftarrow$  “Best”
15:        CurrentBN  $\leftarrow$  TrialBN
16:      else if TrialBNHierarchy is equivalent or better than
        CurrentBNHierarchy then
17:        TrialBNStatus  $\leftarrow$  “Equivalent or better”
18:        CurrentBN  $\leftarrow$  TrialBN
19:      else
20:        TrialBNStatus  $\leftarrow$  “Worse”
21:      end if
22:    end if
23:    if TrialBNStatus is “Best” then
24:      break
25:    end if
26:  end for
27:  if TrialBNStatus is “Best” then
28:    break
29:  end if
30: end while

```

\triangleright *CurrentBN* is the best model

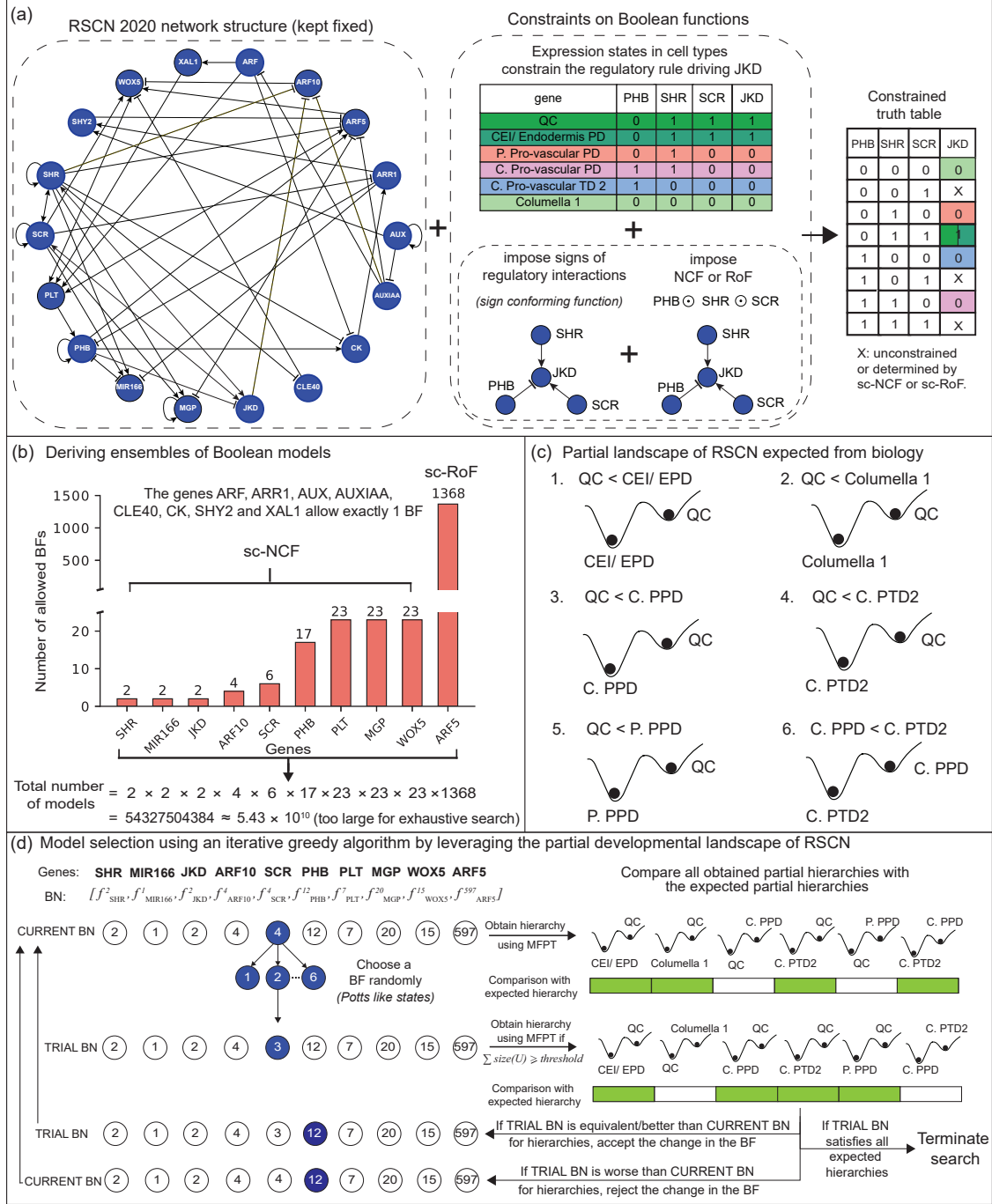


Figure 4.13: Workflow of our methodology for model selection, illustrated on the 2020 model of the Boolean GRN of *Arabidopsis thaliana* RSCN. (a) Procedure to generate ensembles of Boolean models that recover the desired attractors while keeping the network structure of the reconstructed GRN fixed. (b) The bar plot shows for each gene the number of BFs that have at least two allowed BFs after imposing all the constraints in (a).

Figure 4.13 (previous page): (c) The partial expected landscape represented as inequalities between different pairs of fixed points. Here QC: Quiescent center, CEI/EPD: Cortex/endodermis initial cell, P. PPD: Peripheral Pro-vascular initials, C. PPD: Central Pro-vascular initials, C. PTD2: Transition domain, Columella 1: Columella initials. (d) Schema of the iterative greedy algorithm to perform model selection. A BN is represented as a vector of BFs with each entry given by f_g^i where $i \in \{1, 2, \dots, m\}$, g is the gene and m is the number of allowed BFs for that gene. From CURRENT BN we generate a TRIAL BN by assigning to a node, a BF randomly chosen from the allowed BFs at that node. TRIAL BN may be accepted or rejected based on criteria shown in the right portion of this sub-figure. $\sum size(U)$ is the sum of sizes of the BOA of all biological attractors and *threshold* is that same sum for the 2020 model. In either case, the iterative scheme is continued by generating a random BF for the following gene and is terminated when TRIAL BN satisfies all the expected hierarchies.

4.6.3 Greedy algorithm generates many models satisfying the expected developmental landscape

To use the developmental landscape constraint in model selection, we designed an iterative greedy algorithm and applied it to the 2020 model of the *Arabidopsis thaliana* RSCN. We first generate an ensemble of Boolean models using the procedure depicted in Figure 4.13(a) (for details see Section 4.2). In particular, we impose that the type of BF assigned to a gene be the same as the one assigned in the 2020 Boolean model [75]. With that prescription, the BFs of all genes were sc-NCFs except for ARF5 which was of the *sign conforming* read-once function (sc-RoF) type. The number of allowed BFs for each gene is displayed in Figure 4.13(b), leading to an ensemble of models of size of the order 10^{10} , too large for an exhaustive search.

To efficiently explore this ensemble, we use our iterative greedy search. The algorithm takes as inputs the 2020 Boolean model [75] (the initial model as a vector), the BFs allowed at each gene (see Figure 4.13(b)), and the set of expected hierarchies (see Figure 4.13(c)). Starting from the 2020 model we iterate through each gene as shown in Figure 4.13(d), computing at each iteration the hierarchies of the resulting model via our stochastic method (at 5% noise intensity with 2500 trajectories) and terminate the algorithm as soon as a model satisfies all the expected partial hierarchies. Out of 1000 runs, we obtain 990 distinct models which conform to the expected partial landscape and

are listed in Supplementary Table S5.1. A flowchart of this model selection framework with further generalizations is shown in Figure 4.14.

4.7 Discussion

The principal findings of our work reported in this chapter can be stated as follows. First, the 5 different measures of RS [54, 67], based on size of BOA, SSP, MFPT, BTR and SIND are strongly correlated with each other. Second, MFPT can be used to generate cellular lineage trees in addition to providing a hierarchy of cell states, offering a richer picture of developmental dependencies than just a linear ordering of the fixed points. Third, noise intensities need not be particularly small in order to reliably identify the hierarchies between cell types (fixed points of the DGRN). Fourth, the exact method to calculate MFPT based on matrix algebra is not feasible for large networks, whereas our alternative stochastic approach has no such limitation and is particularly simple to implement. Fifth, we developed an iterative greedy search algorithm that can identify models conforming to the expected developmental landscape from a large *ensemble* of biologically plausible models. Lastly, multiple Boolean models were produced that satisfied the expected developmental landscape of the *Arabidopsis thaliana* RSCN.

The challenge of modeling complex biological systems is an old one [22] that has led to many efforts to integrate information from multiple datasets. Several efforts have gone into inference of BNs using a wide range of methods [55, 58, 119–123]. Also, the epigenetic landscape of Boolean models has also been quantified [124], particularly in some model systems such as the flower specification GRN of *Arabidopsis thaliana* [125, 126] and RSCN [127]. However little has been done to leverage that landscape to infer viable Boolean models. Furthermore, during the reconstruction of Boolean models of DGRNs from biological data, modelers are often forced to make arbitrary choices at the level of logic rules. The present work advocates a more streamlined process to assign logic rules to genes by leveraging RS constraints derived from biological developmental landscapes. This idea had its genesis in [54] but it was only applied to a minimal 5-gene Boolean DGRN of Pancreas cell differentiation. Since then, no effort has gone into scaling such a

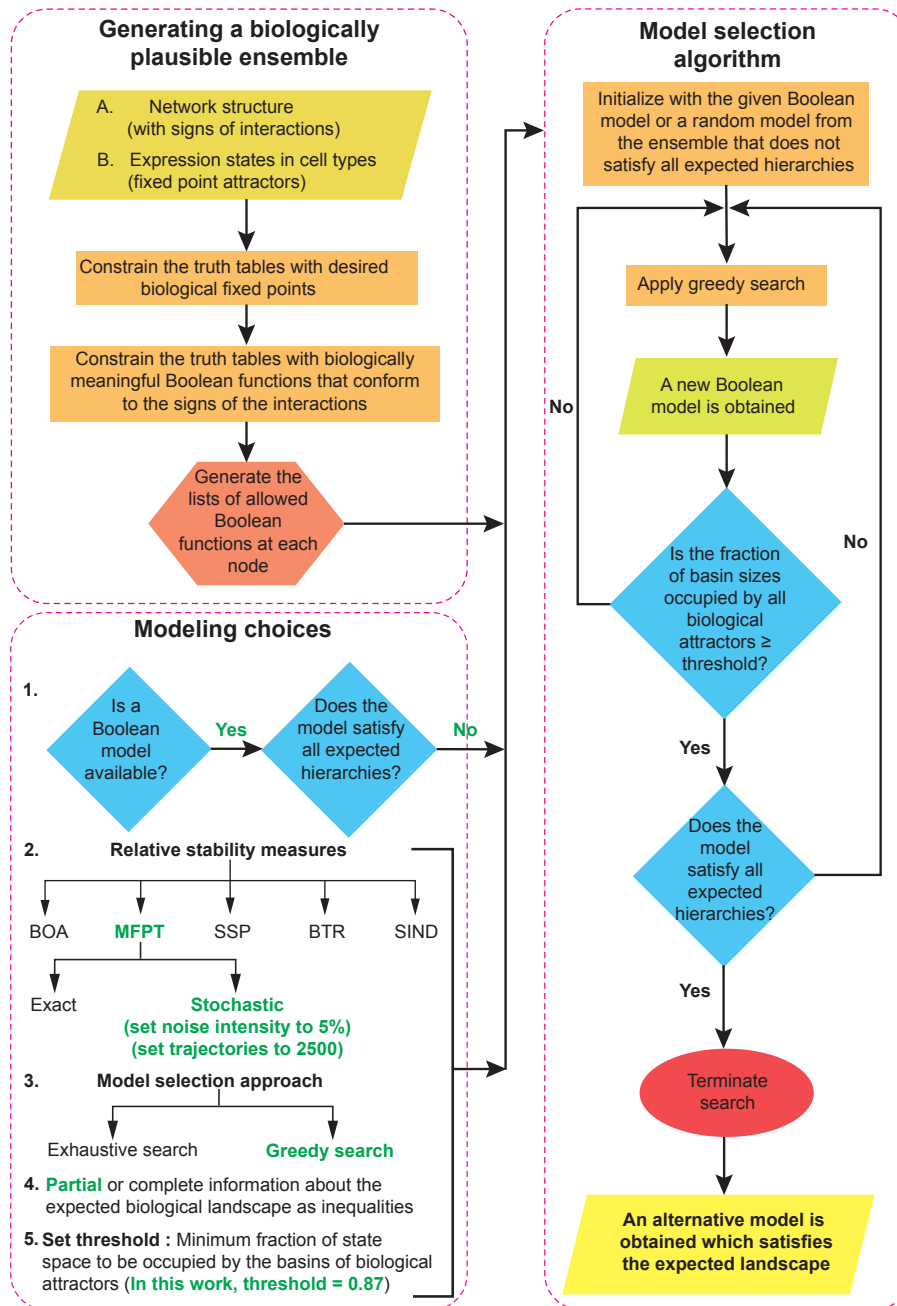


Figure 4.14: A flowchart of the workflow of the model selection procedure. The flowchart consists of 3 portions. The first is “Generating a biologically plausible ensemble”. Two inputs are necessary, namely, the network structure (including the signs of the interactions) and the expected biological fixed points (cell states). The truth table at each node of the network (obtained from the network structure) is constrained using the fixed points and biologically meaningful BFs that conform to the signs of the given interactions. This gives a list of allowed BFs at each gene, from which the ensemble of biologically plausible models can be generated. (Note: In case a Boolean model is provided a priori and the BFs assigned are biologically meaningful, then we constrain the truth tables as per the type of BFs originally assigned in that model).

Figure 4.14 (previous page): The second portion consists of various “Modeling choices”. These include choices of *RS* measures, whether an exhaustive or greedy search should be carried out, information about the expected developmental landscape as inequalities and the value for the threshold of the fraction of state space to be occupied by the states of the basins of biological attractors. The text colored green are the particular choices that we implement in this work. But, other possible choices (colored black) are shown to illustrate the generalization of this framework. The last portion, namely, “Model selection algorithm” takes as input a Boolean model that does not satisfy all the expected hierarchies (this could be a model known a priori or a random model from the generated ensemble), and based on the “modeling choices”, searches for a model that satisfies all the hierarchies via the iterative greedy search. The search terminates once a Boolean model that satisfies the expected hierarchies is found.

methodology to larger models or refining it. Here we build upon that work, taking as a case study an 18 gene Boolean model of *Arabidopsis thaliana* RSCN [75]. Although that model is quite recent, it does not satisfy the expected hierarchies between cell types. With the help of a simple but very effective greedy search algorithm, we were able to improve that model to obtain multiple DGRNs having satisfactory developmental landscapes.

Our methodology provides multiple benefits. For instance it can easily handle the addition of further constraints on the DGRNs, such as robustness to noise, that will restrict even more the space of relevant models. Furthermore, it should impact experimental work in at least two ways. First, since it strongly reduces the space of possible models, it allows for validation via a limited number of experimental measurements. For instance, when considering the ensemble based on the RSCN 2010 GRN [52], the imposition of the different biologically motivated constraints led to just 80 models. Performing additional experiments will lead to a few – if not just one – models compatible with all measurements. Second, our methodology provides testable predictions. Even if multiple models remain possible, there can be features shared by all models that were not previously recognized by biologists. For instance all models may predict that, when knocking out a particular gene, one or more of its targets will change expression in a given direction. Similarly, all models are likely to share some properties in their cellular lineage trees such as preferential de-differentiation paths. Such predictions will inevitably stimulate experimental work.

In conclusion, the results reported in this chapter hold promise for developing standardized workflows for Boolean model reconstruction of DGRNs by leveraging biological

constraints and computational methods.

Supplementary Information

Supplementary Table S5.1 associated with this chapter is available for download from the GitHub repository: https://github.com/asamallab/PhDThesis-Subbaroyan_Ajay/blob/main/SI/ST_Chapter5.xlsx

Data and code availability statement

All the data and codes necessary to reproduce the results in this chapter are available for download from the GitHub repository: <https://github.com/asamallab/LDLM>

Chapter 5

A preference for link operator functions can drive Boolean biological networks towards critical dynamics

The notion that complex biological systems are situated in the neighbourhood of a critical dynamical regime has been studied quite extensively both outside [128] and within the Boolean framework [46, 113, 129–131]. The study of damage spreading in Boolean network (BN) models of gene regulatory networks (GRNs) provides an insight into their dynamical *regime*. More explicitly, the damage to a network state in a Boolean model mathematically translates to performing bit flip(s) to the state of node(s) of the network. In other words, perturbations to a network state via bit flips lead to damage in the network state. The temporal evolution of such a damage under the dynamics of the system is known as damage spreading. Damage spreading in Boolean networks is generally illustrated using the Derrida plot [36, 71, 132]. The Derrida plot is partitioned into three regions: ordered, critical and chaotic regimes. For models in the ordered regime, perturbations (small random changes in the state of the system) tend to remain small or disappear. In the

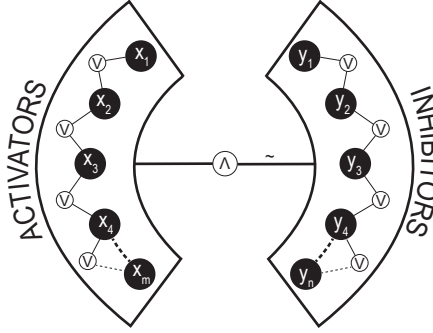
case of models falling in the chaotic regime, perturbations spread out over many nodes in the network. In the critical regime, the dynamics is neither ordered nor chaotic. More recently, Daniels *et al.* [46] considered a static measure as a proxy for damage spreading, specifically the authors used the network average sensitivity of a BN and showed that most biological models largely fall in the critical regime for which the network average sensitivity of the BN is equal to 1.

In this chapter, we focus our attention on a certain type of Boolean functions (BFs) called link operator functions (LOFs) [35, 90]. Firstly, we show the relationship between the different LOFs, and subsequently enumerate the LOFs for different numbers of inputs. Thereafter, we quantify the fraction occupied by LOFs in the space of all BFs and also within effective andunate functions (EUFs). Next, we ask what fraction of regulatory logic rules in a reference biological dataset of regulatory logic rules extracted from reconstructed Boolean models are LOFs. Following this, we present two case studies wherein we impose a given network structure but allow different BFs to examine the consequences of having to satisfy steady-state constraints corresponding to biological phenotypes [54, 109]. In particular, we show that limiting the choice of BFs to LOFs during such model selection can dramatically shrink the size of the search space. Lastly, by computing the *static* network average sensitivity for a wide range of fixed biological network structures, but imposing different types of functions (EFs, EUFs and LOFs), we find AND-NOT and OR-NOT logic in LOFs are closest to reproducing the network average sensitivity distribution of biological regulatory logic. **The work reported in this chapter is contained in the published manuscript [50].**

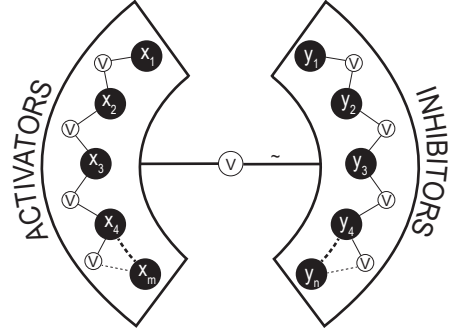
5.1 Link operator functions

Mendoza and Xenarios [90] defined a type of veto regulatory logic in BNs which they used to model the differentiation of T-helper cells. In the above-mentioned work, the veto logic operates as follows. If any inhibitor is present (ON), the regulated gene is turned OFF. If all inhibitors are absent and at least one activator is present, then the regulated gene is turned ON, otherwise the gene is turned OFF. In a subsequent contribution, Zobolas *et*

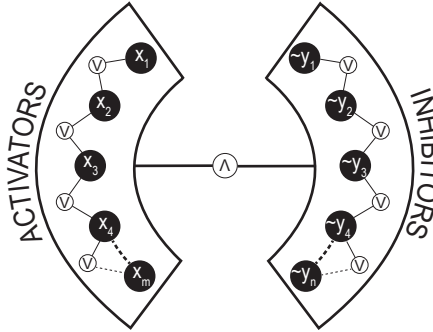
$$(a) f_{\text{AND-NOT}} = (x_1 \vee x_2 \vee \dots \vee x_m) \wedge \sim (y_1 \vee y_2 \vee \dots \vee y_n)$$



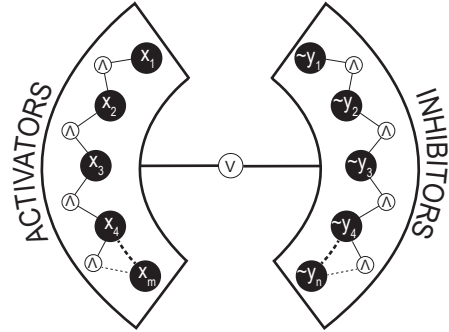
$$(b) f_{\text{OR-NOT}} = (x_1 \vee x_2 \vee \dots \vee x_m) \vee \sim (y_1 \vee y_2 \vee \dots \vee y_n)$$



$$(c) f_{\text{AND-pairs}} = (x_1 \vee x_2 \vee \dots \vee x_m) \wedge (\sim y_1 \vee \sim y_2 \vee \dots \vee \sim y_n)$$



$$(d) f_{\text{OR-pairs}} = (x_1 \wedge x_2 \wedge \dots \wedge x_m) \vee (\sim y_1 \wedge \sim y_2 \wedge \dots \wedge \sim y_n)$$



V - OR operator Λ - AND operator ~ - NOT operator

Figure 5.1: Schematic figure illustrating the various types of consistent LOFs. The inputs to LOF BF are divided into two sets namely, activators and inhibitors, denoted by the variables x_i and y_j respectively. There are m activators and n inhibitors. The logical operators which connect the variables are the AND (\wedge), OR (\vee) and NOT (\sim) operators. The four types of LOFs shown are: (a) AND-NOT logic (b) OR-NOT logic (c) AND-pairs logic and (d) OR-pairs logic.

al. [35] use the structure of the logical expression of these veto BFs to explore a number of other BFs possessing similar logical structure and define these as *link operator functions* (LOFs). Their Boolean expression is constructed by linking a set of m activators (labelled as x_i) to a set of n inhibitors (labelled as y_j) by a logical operator shown as \otimes in Eq. (5.1). We use the symbol k to denote the total number of regulators of the considered node, i.e., $k = m + n$. The general expression for these functions is given by:

$$(x_1, x_2, \dots, x_m) \otimes (y_1, y_2, \dots, y_n) \quad (5.1)$$

where the link operator \otimes can be: NOR, NAND, AND-NOT, OR-NOT, NOR-NOT, NAND-NOT, XOR, pairs, XNOR among others. The activators (or inhibitors) x_1, x_2, \dots, x_m (or y_1, y_2, \dots, y_n) are typically connected by only AND or OR operators. The LOFs are defined for functions which have at least one activator ($m \geq 1$) and one inhibitor ($n \geq 1$). Notably, Zobolas *et al.* [35] showed that only some link operators in Eq. (5.1) satisfy biologically relevant *consistency* properties namely, monotonicity and essentiality (or effectiveness). Biological regulatory logic rules are typically expected to possess both of these consistency properties [33, 34, 49]. BFs which possess both of the properties of unateness (or monotonicity) and essentiality (or effectiveness) in all inputs are known as EUFs (see Section 2.2, Chapter 2). In their recent work, Zobolas *et al.* [35] focussed on three types of LOFs namely, AND-NOT, OR-NOT and their pairs function (which in this paper we call AND-pairs), that satisfy the above-mentioned two consistency properties (see Table 5.1 for the exact definition). The AND-NOT, OR-NOT and AND-pairs are given by:

$$f_{AND-NOT} = (x_1 \vee x_2 \vee \dots \vee x_m) \wedge \sim (y_1 \vee y_2 \vee \dots \vee y_n) \quad (5.2)$$

$$f_{OR-NOT} = (x_1 \vee x_2 \vee \dots \vee x_m) \vee \sim (y_1 \vee y_2 \vee \dots \vee y_n) \quad (5.3)$$

$$f_{AND-pairs} = (x_1 \vee x_2 \vee \dots \vee x_m) \wedge (\bar{y}_1 \vee \bar{y}_2 \vee \dots \vee \bar{y}_n) \quad (5.4)$$

where \vee is the OR operator, \wedge is the AND operator and \sim is the NOT operator. For an illustration of the LOFs, see Figure 5.1. We find that in addition to these three types of LOFs, another type of LOF can be constructed which satisfies the two consistency

properties, and is complementary to the AND-pairs in a manner that the OR-NOT is complementary to the AND-NOT. Note that if one complements an AND-NOT function, we get an OR-NOT function but with the activators and inhibitors exchanged. Similarly, if we complement the AND-pairs, we get the OR-pairs but with the activators and inhibitors exchanged. We call this the OR-pairs and it is given by the expression:

$$f_{OR-pairs} = (x_1 \wedge x_2 \wedge \dots \wedge x_m) \vee (\bar{y}_1 \wedge \bar{y}_2 \wedge \dots \wedge \bar{y}_n) \quad (5.5)$$

The biological interpretation for each of the LOFs is as follows:

- **AND-NOT**: The presence of a single inhibitor represses transcription independent of the presence of multiple activators. Thus, transcription takes place only in the absence of inhibitors and in the presence of at least one activator.
- **OR-NOT**: The presence of any activator guarantees transcription independent of the presence of inhibitors. In the absence of both inhibitors and activators, gene transcription takes place.
- **AND-pairs**: The presence of at least one activator and the absence of at least one inhibitor is sufficient to ensure transcription.
- **OR-pairs**: All activators must be present, or all inhibitors must be absent in order for transcription to take place.

Table 5.1 lists the four consistent types of LOFs, their expression and the additional types of BFs to which they belong and Figure 5.1 depicts the various LOFs. Henceforth, we reserve the word LOF to mean only the 4 consistent types, namely, AND-NOT, OR-NOT, AND-pairs and OR-pairs. Note from Table 5.1 that AND-NOT and OR-NOT LOFs are NCFs whereas AND-pairs and OR-pairs LOFs are not NCFs. It follows that AND-NOT and OR-NOT LOFs will have all properties that NCFs possess that will differentiate them from BFs that are not NCFs. However, to the best of our knowledge, subsets of NCFs such as AND-NOT and OR-NOT are not easily distinguished based on existing metrics on BFs and are a subject for future investigation.

Table 5.1: The different types of consistent LOFs. The four different types of LOFs are AND-NOT, OR-NOT, AND-pairs and OR-pairs. From this table, it can be ascertained that these four types of LOFs satisfy all the consistency properties considered in Zobolas *et al.* [35]. Note that Zobolas *et al.* [35] have only considered the first three types in their work. Here EF is effective function, UF is unate function, CF is canalyzing function, NCF is nested canalyzing function and CCF is collectively canalyzing function

LOF type	Boolean Expression	EF	UF	CF	NCF	CCF
AND-NOT	$(x_1 \vee x_2 \vee \dots \vee x_m) \wedge \sim (y_1 \vee y_2 \vee \dots \vee y_n)$	Yes	Yes	Yes	Yes	No
OR-NOT	$(x_1 \vee x_2 \vee \dots \vee x_m) \vee \sim (y_1 \vee y_2 \vee \dots \vee y_n)$	Yes	Yes	Yes	Yes	No
AND-pairs ($n > 1$)	$(x_1 \vee x_2 \vee \dots \vee x_m) \wedge (\bar{y}_1 \vee \bar{y}_2 \vee \dots \vee \bar{y}_n)$	Yes	Yes	No	No	Yes
OR-pairs ($m > 1$)	$(x_1 \wedge x_2 \wedge \dots \wedge x_m) \vee (\bar{y}_1 \wedge \bar{y}_2 \wedge \dots \wedge \bar{y}_n)$	Yes	Yes	No	No	Yes

5.2 Characterizing the space of LOFs

5.2.1 Relationship between the different types of LOFs

We note that there may be overlaps between two different types of LOFs, and between LOFs and other types of biologically meaningful BFes [49]. Within the space of LOFs we observe that:

- (a) AND-NOT and OR-NOT do not overlap.
- (b) AND-pairs and OR-pairs do not overlap.
- (c) The AND-NOT LOF is equivalent to the AND-pairs LOF if there is only one inhibitory input ($n = 1$), for any value of k .
- (d) The OR-NOT LOF is equivalent to the OR-pairs LOF if there is only one activatory input ($m = 1$), for any value of k .

The above observations (c) and (d) serve as a motivation to construct a set of four non-overlapping types of LOFs (see Table 5.1). We first define two non-overlapping types of LOFs:

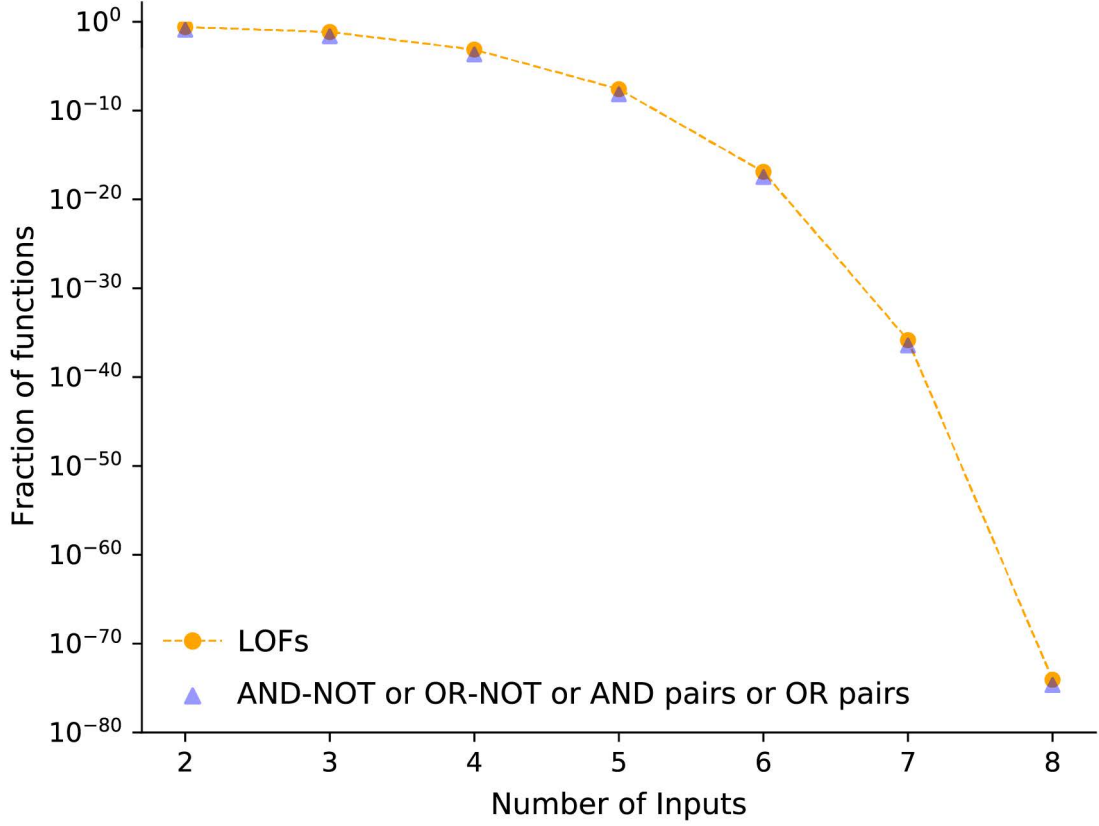


Figure 5.2: The reduction in the size of the space of consistent LOFs in comparison to the space of all BFs with increasing number of inputs. The decrease of the fraction of consistent LOFs with increasing number of inputs is extremely rapid. Here LOFs (orange circles) refer to the sum of the fractions of all four consistent LOFs, namely the AND-NOT, OR-NOT, AND-pairs and OR-pairs (with any redundancies removed). The blue triangles represents any one of the aforementioned four types of LOFs, since each of them has the same number of functions.

- (i) AND-pairs ($n > 1$) as the AND-pairs with more than one inhibitory input.
- (ii) OR-pairs ($m > 1$) as the OR-pairs with more than one activatory input.

AND-pairs ($n > 1$) and OR-pairs ($m > 1$) do not overlap with the AND-NOT and OR-NOT LOFs, respectively. Moreover, we observe that both AND-NOT and OR-NOT LOFs are *nested canalizing functions* (NCFs). The AND-pairs ($n > 1$) and OR-pairs ($m > 1$) on the other hand are *collectively canalizing functions* (CCFs). A k -input BF is said to be *collectively canalizing* if by fixing a certain subset of i inputs (such that $1 < i < k$), the output of the function is determined [81] while it is not when fixing fewer than i inputs.

Table 5.2: Number of LOFs as a function of the number of activators (m), the number of inhibitors (n) and the total number of inputs (k). Note that $k = m + n$. Importantly, a LOF should have at least one activating input ($m \geq 1$) and at least one inhibiting input ($n \geq 1$), and thus, LOFs can exist only for nodes with 2 or more inputs ($k \geq 2$). Here, we give the number of LOFs for different possible combinations of m activators and n inhibitors for a given number of inputs k . Moreover, we report separately the number of functions in the four different types of consistent LOFs namely, AND-NOT, OR-NOT, AND-pairs ($n > 1$) and OR-pairs ($m > 1$). In addition, the table also gives the number of effective and unate functions (EUFs) for different possible combinations of m and n . As k increases, it can be seen that the LOFs become a tiny fraction of the EUFs.

k	m	n	EUFs	LOFs				Total	Fraction of EUFs that are LOFs
				AND-NOT	OR-NOT	AND-pairs ($n > 1$)	OR-pairs ($m > 1$)		
2	1	1	4	2	2	0	0	4	1
3	1	2	27	3	3	0	3	9	0.333
3	2	1	27	3	3	3	0	9	0.333
4	1	3	456	4	4	0	4	12	0.0263
4	2	2	684	6	6	6	6	24	0.0351
4	3	1	456	4	4	4	0	12	0.0263
5	1	4	34470	5	5	0	5	15	4.35×10^{-4}
5	2	3	68940	10	10	10	10	40	5.80×10^{-4}
5	3	2	68940	10	10	10	10	40	5.80×10^{-4}
5	4	1	34470	5	5	5	0	15	4.35×10^{-4}

5.2.2 Cardinality of the different types of LOFs

It is straightforward to count the number of LOFs. Consider the AND-NOT LOFs for instance. For a given number of inputs (k), and for a given number of activators (m) and inhibitors (n), there are $C(k, m)$ (the binomial coefficient) ways to assign m activators and n inhibitors. Since all the activators are connected by an AND or an OR operator, the permutations between them do not alter the BF. Hence there are exactly $C(k, m)$ BFs in the AND-NOT category. A similar argument holds for the number of functions in the OR-NOT category. For the AND-pairs ($n > 1$) and OR-pairs ($m > 1$), the number of functions for m activators and n inhibitors is $C(k, m) - C(k, 1)$ and $C(k, n) - C(k, 1)$ respectively. To calculate the total number of LOFs of a given type for k inputs, we sum

over all the values of m . Hence for both AND-NOT and OR-NOT, there are a total of $2^k - 2$ BFs each (we subtract 2 because LOFs do not include the cases where there are no activators or inhibitors, i.e., $C(k, m = 0)$ and $C(k, n = 0)$ are not counted).

Based on this exact counting of LOFs, it can be easily seen that LOFs form an extremely small subset of the space of all BFs and that their corresponding fraction decreases fast with increasing number of inputs (see Table 5.2 and Table E.1, Appendix E). Furthermore, even within the space of EUFs, LOFs form a tiny subset. Figure 5.2 is a semi-log plot that shows this decrease in the fraction of LOFs with the increase in the number of inputs. Note that even if one pools the four classes of LOFs under consideration, the number of functions (for a given number of inputs) increases approximately by a factor of 4, which nevertheless does not affect our conclusion. Figure 5.2 and Table E.1, Appendix E illustrate this point.

5.3 Preponderance of LOFs in reconstructed Boolean models of gene regulatory networks

Even though the LOFs are *consistent* in terms of the effectiveness and monotonicity properties, it remains to be shown how frequently they arise in biological systems. To investigate this, we take as our reference biological dataset BFs extracted from a collection of 57 Boolean models of biological systems from the Cell Collective database that are a result of the works of many authors, covering a wide variety of biological processes in a number of species spanning the multiple kingdoms of life. Only those models in the Cell Collective database where both the biological network and BFs were curated manually were considered for this analysis (see Appendix A.2, Table A.1). It is clear from Figure 5.3, Table 5.3 and Table E.2, Appendix E that the AND-NOT are particularly abundant in reconstructed Boolean models whereas the other types of LOFs such as OR-NOT, AND-pairs and OR-pairs are almost absent. Recall that BFs with at least one activator and one inhibitor can be LOFs. Hence it is meaningful to calculate the fraction of LOFs in the reference biological dataset among those BFs with at least one activator and one inhibitor

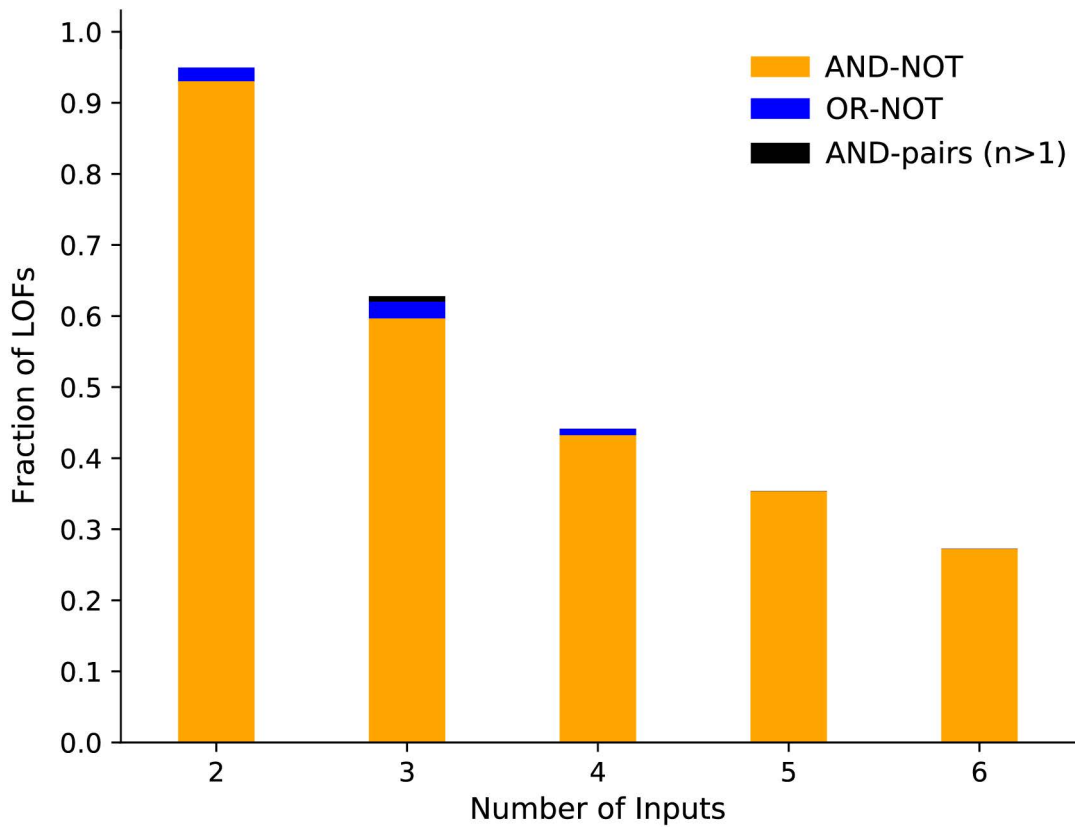


Figure 5.3: The fractions of the various types of consistent LOFs in the reference biological dataset. The AND-NOT LOFs are clearly abundant among the biological functions with at least one activator and one inhibitor, whereas the other types though present, are not as abundant as the AND-NOT functions. Note that the fractions for each of the LOFs are calculated with respect to the number of BFs with at least one activator and one inhibitor as input.

Table 5.3: The abundance of LOFs in the collection of BFs from reconstructed models of biological systems. The reference biological dataset consists of BFs from 57 Boolean models compiled in the Cell Collective database (<https://cellcollective.org/>). Notably, a LOF should have at least one activating input ($m \geq 1$) and at least one inhibiting input ($n \geq 1$), and thus, LOFs can exist only for nodes with 2 or more inputs ($k \geq 2$). Focussing on the subset of BFs in the 57 reconstructed models that have at least one activating input ($m \geq 1$) and at least one inhibiting input ($n \geq 1$), the table classifies the BFs in the reference biological dataset into effective andunate functions (EUFs) and different types of consistent LOFs. It is evident that EUFs, and moreover, the AND-NOT LOFs within EUFs, are abundant in the reference biological dataset regardless of k . In this table, we display the statistics for BFs in the reference biological dataset up to 5 inputs ($k \leq 5$). In Table E.2, Appendix E we display the statistics for all BFs in the reference biological dataset up to $k = 12$ inputs. ‘NA’ means ‘not applicable’, corresponding to values of m and n for which the LOF under consideration does not exist.

k	m	n	BFs in reference biological dataset	EUFs	LOFs				Total
					AND- NOT	OR- NOT	AND- pairs ($n > 1$)	OR- pairs ($m > 1$)	
2	1	1	158	150	147	3	NA	NA	150
3	1	2	35	32	30	1	1	NA	32
3	2	1	94	87	47	2	NA	0	49
4	1	3	16	16	13	1	0	NA	14
4	2	2	38	35	17	0	0	0	17
4	3	1	57	48	18	0	NA	0	18
5	1	4	4	4	1	0	0	NA	1
5	2	3	16	15	10	0	0	0	10
5	3	2	25	24	8	0	0	0	8
5	4	1	20	17	4	0	NA	0	4

(see Table E.3, Appendix E).

The dominance of AND-NOT LOFs in the reference biological dataset implies that regulatory logic is primarily governed by a special type of veto mechanism wherein the presence of a single inhibitor determines the output of the gene, independent of the presence of activators. In other words,

- (i) the activators can function only in the absence of the inhibitors, and
- (ii) the *vetoing power* of all inhibitors is the same.

Thus, even though activators are far more numerous than inhibitors in the reference biological dataset, the inhibitors generally control the logic output. Results inferred from reference biological dataset are typically and rightly subject to scrutiny in that they could

be artefacts of a biased dataset. In the present case we believe that this is highly unlikely given the diversity of biological processes being modeled. Note that 54 out of 57 models in our dataset belong to the eukaryota domain; the biological literature therein is abundant with cases where the repressor (or inhibitor) is able to suppress transcription even in the presence of many activators [133].

Table 5.4: Model selection by using different types of BFs with and without the steady state constraints. The two Boolean models, pancreas development and epithelial-mesenchymal transition (EMT), with 5 nodes each are used to illustrate the reduction in allowed models achieved by using various biologically meaningful BFs, both with and without the steady state constraints.

BF constraint	Pancreas development		EMT	
	No constraint	Steady state constraints	No constraint	Steady state constraints
None	17179869184	1048576	268435456	262144
EUF	104976	7056	1458	140
NCF	65536	3600	1024	96
LOF	1296	100	54	8

5.4 LOFs as facilitators of Boolean model reconstruction and selection: Two case studies

Model selection is the problem of searching for models that exhibit high fidelity to observations from biological data. In general there are many ways to satisfy constraints derived from such observations [54, 134–136]. In this work, we follow the model selection procedure by Zhou *et al.* [54]: One begins by determining the network structure of the system via experimental data providing information on the regulatory interactions between the biological components. Next, dynamical models must reproduce the biological steady states, and in the Boolean framework, this corresponds to imposing constraints on the truth table or BFs assigned to every node of the network. Finally, among the various types of BFs, biologically meaningful functions can be chosen. Thus, by applying such successive constraints, we can zero-in on a much smaller subset of models within the space of all possible models. We illustrate such a model selection procedure on two reconstructed GRNs

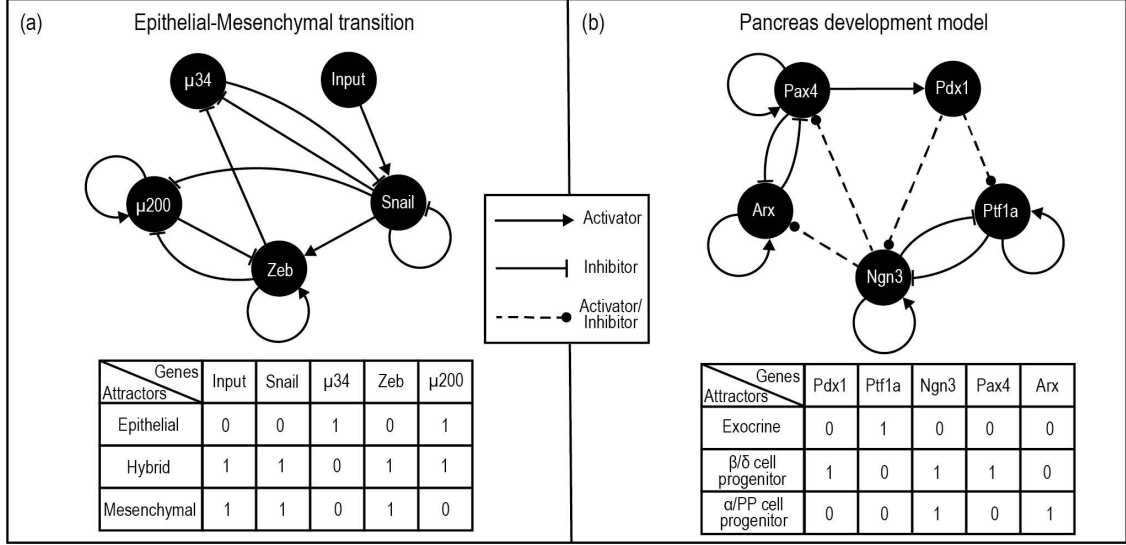


Figure 5.4: Schematic figure showing the two models of pancreas development and Epithelial-mesenchymal transition (EMT) GRNs along with their attractors. Nodes are associated with genes and edges correspond to directed interactions. The biologically relevant attractors in both models are steady states. In the pancreas development network, the edges labeled “Activator/Inhibitor” correspond to interactions whose signs were denoted as unknown in Zhou *et al.* [54].

(see Figure 5.4): a pancreas differentiation model [54] and an Epithelial–mesenchymal transition (EMT) model [67].

Table 5.4 illustrates the reduction in the number of possible models when imposing our successive constraints. Following Zhou *et al.* [54], the network connectivity is imposed as well as the sign of each interaction when it is known. The problem is then to search the space of BFs at each node. The constraint of reproducing the steady states factorizes and thus the number of models satisfying the constraints is given by the product of the number of BFs satisfying the constraints on each node. For instance in the EMT model, there are a total of 268435456 ($= 1 \times 256 \times 16 \times 256 \times 256$) models if one imposes neither steady state constraints nor constraints on the type of BFs, whereas there are 262144 ($= 1 \times 64 \times 4 \times 32 \times 32$) models satisfying the steady state constraints but ignoring further constraints on the type of BFs. These numbers also reflect the fact that even with a fixed network structure along with steady state constraints on BFs, the number of models is astronomical.

By taking advantage of the tiny fraction of LOFs in the space of all BFs, we can

tremendously shrink the number of models which are biologically relevant. More explicitly, in the case of the EMT model, the 262144 models obtained by applying only steady state constraints are reduced to just 8 by demanding that the BFs are LOFs whereas in the pancreatic development model, of 1048576 models which are obtained by imposing steady state constraints, only 100 models satisfy the conditions of both reproducing the steady states and using LOFs for their regulatory logic. We emphasize that both these models primarily serve as toy models to illustrate the procedure of model selection and consequently the shrinkage of the space of models that satisfy both the attractors constraints and the LOF constraints. In essence, using LOFs can tremendously shrink the space of Boolean models to be explored.

5.5 Implications of using LOFs for Boolean network dynamics

Damage spreading [132] in discrete dynamical systems measures how two trajectories diverge and thus provides a measure of sensitivity to initial conditions, much like Lyapunov exponents do in continuous systems. A question that is relevant for investigating the long-term behaviour of the system is whether damage spreads across the network states with time. This is in essence equivalent to asking how sensitive the dynamics of the system is to slightly different initial conditions of a network state (in this case, a chosen network state, and the other a perturbed network state). Damage spreading and sensitivity to initial conditions in Boolean networks may be thought of as different sides of the same coin and are primarily investigated via the same method, namely, Derrida plot [36, 71, 132].

The Derrida plot is typically used to quantify damage spreading and is constructed as follows. Choose two initial network states (usually with a small Hamming distance). Evolve each of the network states by one time step and compute the Hamming distance between them. Now, repeat this procedure for a large number of pairs of initial network states. A plot of the Hamming distance between the initially chosen network states and the Hamming distance between the states evolved after a single time step gives the Derrida

plot [36, 132]. Note that static measures of damage spreading such as network average sensitivity, as is explained below, do not require the simulation of state space trajectories and can be calculated directly from the BFs. This in fact leads to a major advantage of using such static measures over the Derrida plot, namely, its scalability to very large networks. Studies in Boolean models of biological GRNs suggest that they exhibit neither ordered nor chaotic behaviour, but rather an intermediate kind of behaviour, known as *critical*. Here we employ a static measure of damage spreading, as opposed to the one used to construct Derrida plots, namely the network average sensitivity of a BN [71]. Firstly, the average sensitivity of a BF is given by the proportion of cases where changing one of the inputs at random changes the output value, averaged over all possible input combinations. The network average sensitivity of the BN is then the mean of the average sensitivity of all its BFs (Eq. (3.1)).

Shmulevich and Kauffman [71] showed that under synchronous updation, when using randomly drawn representatives of classes of functions, it is possible to infer the damage spreading regime of a BN without resorting to dynamical simulations by simply determining the network average sensitivity. Typically, networks with sensitivity $s \approx 1$ indicate that they fall in the critical regime, $s < 1$ in the ordered regime and $s > 1$ in the chaotic regime. Furthermore, by computing the sensitivity s of a wide range of biological Boolean models, Daniels *et al.* [46] showed that most biological models fall in the *critical* regime $s \approx 1$.

In this work, we compare sensitivities of biological networks with fixed connectivity structure but varying functions, namely: effective function (EF), EUF, AND-NOT, OR-NOT, AND-pairs, OR-pairs and biological functions (i.e., the functions as assigned by model builders). We perform this analysis on 57 models collected from the Cell Collective database [45] (<https://cellcollective.org>). In the case of EFs and EUFs, each node can be assigned BFs ranging over numerous values of average sensitivities whereas for the LOFs of a given kind, there exists multiple functions, but all with the same value of average sensitivity. In Figure 5.5, we see that networks driven by LOF regulatory logic push the biological network dynamics towards criticality ($s = 1$) (see Table E.4, Appendix E). Based on the fraction of networks lying in the outliers of the distribution of network

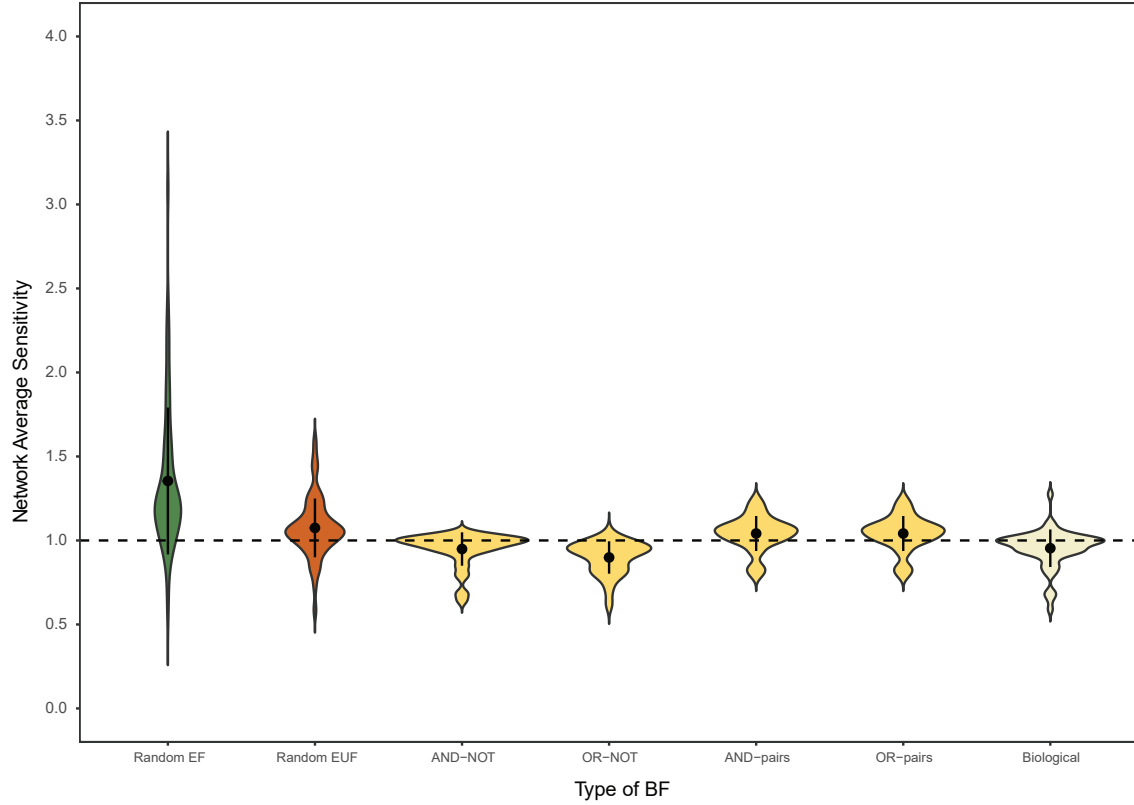


Figure 5.5: Network average sensitivity distribution of the various models in the reference biological dataset using various types of BFs. The sensitivity of models where the structure of the reconstructed biological network is preserved but with the BFs replaced by one of the following types: random effective functions (EFs), random effective and unate functions (EUFs), AND-NOT, OR-NOT, AND-pairs, OR-pairs LOFs. For comparison, we also include the case where the functions are as assigned originally in the reconstructed biological model. Since nodes with only activators or only inhibitors as inputs cannot be assigned LOFs, we assigned the biological functions to them and calculated the average sensitivity of the resulting network. This is done even in the case of EFs and EUFs so as to ensure a fair comparison between the distributions of the average sensitivities of the various BFs being considered.

average sensitivity of biological networks (see Table E.5, Appendix E), AND-NOT and OR-NOT logic lead to more realistic behavior than other types of logic functions. Details about the procedure to generate EFs and EUFs and other assumptions used in these computations can be found in Section B.2, Appendix B.

5.6 Discussion

A large-scale analysis to assess the abundance of LOFs in Boolean models of biological networks had not been carried out so far. In this chapter, we perform such an analysis which reveals the large preference for AND-NOT logic in the regulatory rules of genetic networks. This preference coupled with the fact that LOFs occupy a minute region in not only the space of all BFs, but also within the EUFs, raises the question: why are LOFs, specifically the AND-NOT logic, preferred over other choices of BFs? We tackle this question by determining how the imposition of various types of regulatory rules affects damage spreading in such networks. Daniels *et al.* [46], by using network average sensitivity that is a static measure of damage spreading, showed that having canalyzing rules pushes Boolean models towards criticality. We go one step further to show that within both canalyzing functions and consistent logic functions (i.e. EUFs), though LOF logic drives network dynamics towards criticality, eukaryotic mechanisms are predominantly driven by the AND-NOT logic. Biological networks governed by OR-NOT logic fall slightly in the ordered regime, in comparison to other types of BFs.

Given that there are multiple advantages to choosing LOFs as regulatory logic, it is worth noting that LOFs are also limited in their scope as they require at least one activator and one inhibitor. We observe that such nodes, all of whose inputs are either only activators or only inhibitors, are abundant in biological networks. Thus, it is the combined effect of those logic functions and the AND-NOT logic that shape the models we have studied here.

Data and code availability statement

All the data and codes necessary to reproduce the results in this chapter are available for download from the GitHub repository: <https://github.com/asamallab/LOF>

Chapter 6

Relative importance of composition structures and biologically meaningful logics in bipartite Boolean models of gene regulation

Mounting evidence over the past two decades, obtained via biological network reconstruction using large-scale data from high-throughput experiments [7, 10, 137], has shown that the architecture of real gene networks is far from random, both for their network structure [7, 10, 17, 46, 103, 129, 138, 139] and for their logical update rules [26, 27, 29, 45, 53, 87, 91, 103, 140, 141], i.e., the Boolean functions (BFs) assigned to each associated gene. To date, the bipartite Boolean models proposed [43, 142, 143] for transcriptional gene regulation are theoretical propositions without a solid grounding in empirical evidence. This chapter seeks to approach the question of prevalence of composition structures in real gene regulatory networks (GRNs) from a data-centric perspective. The central theme of this chapter is in effect to examine how plausible it is for both composi-

tion structures and composed BFs to occur in real transcriptional regulatory networks by analyzing published experimental data.

We begin by estimating the prevalence of composition structures arising in two different scenarios of gene regulation. The first scenario is gene regulation by heteromeric protein complexes which act as transcription regulators [142, 143]. The other scenario, which is a novel aspect of this work, accounts for transcriptional regulation via *cis*-regulatory elements, in particular promoters and enhancers [144, 145], that can be bound by transcription factors (TFs). Next, we build upon the work of Fink and Hannam [43] on Boolean compositions and augment their approach for counting the number of possible BFs under Boolean compositions by accounting for the fact that the different input variables are distinguishable and so are non-equivalent under permutation. We then compare the restriction in the logic rules in GRNs due to Boolean compositions with the restriction due to different types of biologically meaningful BFs, and thereafter analyze how often Boolean compositions display biologically meaningful properties. Finally, we evaluate the enrichment (depletion) and *relative* enrichment (depletion) of composed BFs in a compiled empirical dataset of 2687 BFs from published reconstructed Boolean models of biological systems. **The work reported in this chapter is contained in the published manuscript [51].**

6.1 Bipartite Boolean networks, composition structures and composed BFs

6.1.1 Bipartite Boolean networks

Evidently, the Boolean network (BN) model of gene regulation (see Figure 6.1(a)) is a coarse-grained picture of biological reality. There have been proposals to incorporate more realistic features within the Boolean framework [142, 146–150]. Graudenzi *et al.* [146] were the first to propose a bipartite BN model of gene regulation with an aim to incorporate more realistic assumptions about the timescales of genetic processes. Their model, called as gene protein BN or gene product BN (GPBN), could explicitly capture the interactions

between genes and proteins, or genes and gene products (e.g., microRNAs), respectively. Hannam *et al.* [142] generalized the notion of GPBNs further and proposed a bipartite BN model that could also account for the formation of heteromeric protein complexes in regulatory processes. More precisely, the biological basis behind such a bipartite model of transcriptional regulation is as follows [142, 143]. Firstly, a factor affecting a gene’s transcription rate can be either a single TF or a complex of TFs (e.g. heterodimer of TFs [151]). We refer to either type as a *transcriptional regulator* (TR). Thus the presence of a TR may depend on the expression of one or more genes. Secondly, multiple TRs can control the expression of a given gene. Note that genes are regulated not only by TFs but by other types of molecules such as miRNAs and hormones, and accounting for these in the bipartite formalism proposed by Fink and Hannam [43] requires a further exploration of the framework. We do believe however that it may be possible to explicitly account for complexes containing different molecules such as RNA-binding proteins or hormone-receptor complexes in this framework but do not pursue this further in this chapter. Fink and Hannam [43] capture the gene-TF-gene interactions in bipartite BNs via subgraphs called *composition structures* and elucidate how composition structures allow for a *composition* of BFs to be defined on genes. Further, they show that the presence of composition structures can severely restrict the space of allowed BFs. Note that the restrictive nature of the composition of BFs, albeit in unipartite Boolean models, has also been considered by Shmulevich *et al.* [113].

6.1.2 Composition structures

Fink and Hannam [43] introduced the term *composition structure* to denote specific subgraphs of gene-TF-gene interactions in a bipartite BN. More precisely, a composition structure $\{t_1, t_2, \dots, t_r\}$ is assigned to a given gene if its transcriptional regulation depends on the states of r TRs according to a BF of r inputs. Further, the state of each TR i , where $i \in \{1, 2, \dots, r\}$, in turn depends on the states of t_i genes according to a BF of t_i inputs. We now provide here a formal definition of composition structures. Consider a subgraph in the bipartite network model of transcriptional regulation wherein a given gene has r incoming links from r TRs, that is, the expression of the given gene is directly

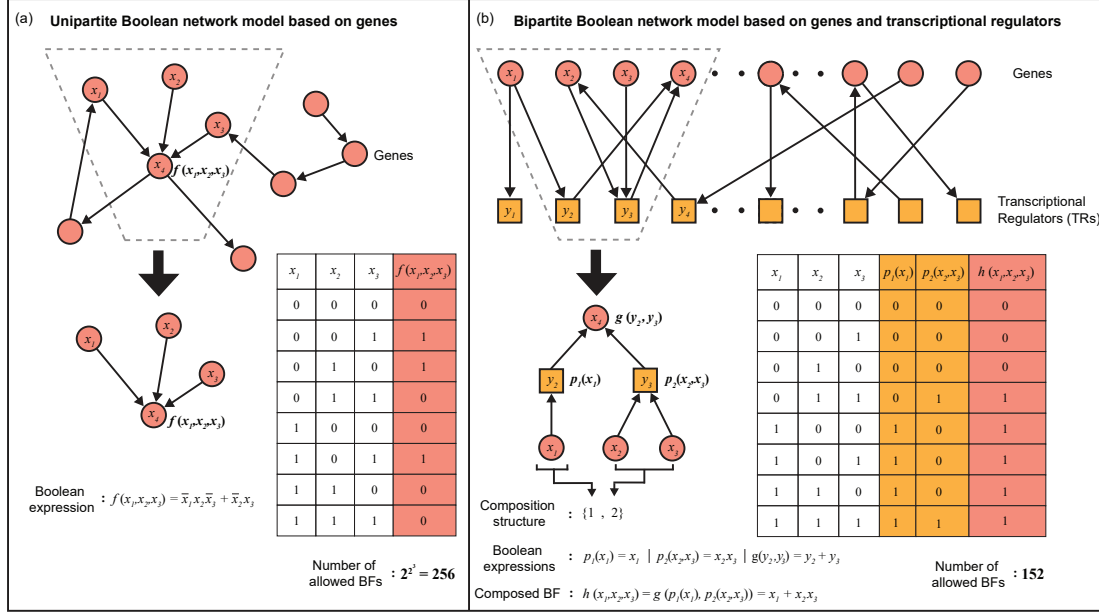


Figure 6.1: Boolean functions in unipartite versus bipartite network models of transcriptional gene regulation. (a) A unipartite BN model consisting of only genes. The dashed trapezium highlights a subgraph wherein 3 genes with expression states x_1, x_2 and x_3 , directly regulate the gene with expression state x_4 . Thus, the BF f determining the state x_4 of the output gene depends on the states of the 3 input genes x_1, x_2 and x_3 , and the truth table for this 3-input BF f is shown in the figure; its *bias*, defined as the number of 1's in the output column, is 3 for this case. Note that any one of the $2^{2^3} = 256$ possible 3-input BFs can be assigned to BF f . (b) A bipartite BN model accounting for the two types of molecular species involved in transcriptional regulation namely, the genes and TRs. In this bipartite BN model, the states of genes are denoted by variables x_1, x_2, \dots, x_i and the states of TRs are denoted by variables y_1, y_2, \dots, y_j . The dashed trapezium highlights the subgraph wherein the gene with state x_1 determines the TR with state y_2 according to a 1-input BF $p_1(x_1) = x_1$, and the genes with states x_2 and x_3 determine the state of the TR y_3 according to a 2-input BF $p_2(x_2, x_3) = x_2 x_3$. Moreover, the TRs with states y_2 and y_3 in this subgraph directly regulate the gene with state x_4 according to a 2-input BF $g(y_2, y_3) = y_2 + y_3$. Ultimately, the regulation of the output gene with state x_4 depends on the states of the input genes x_1, x_2 and x_3 according to a 3-input BF $h(x_1, x_2, x_3) = g(p_1(x_1), p_2(x_2, x_3)) = x_1 + x_2 x_3$. Fink and Hannam [43] called such a subgraph a *composition structure* and the BF h corresponding to the subgraph a *composed BF*. The truth table of a composed BF h allowed by this particular composition structure $\{1, 2\}$ is shown in the figure. Moreover, for the composition structure $\{1, 2\}$, there are $2^{2^1} 2^{2^2} 2^{2^2} = 256$ ways to combine the BFs g, p_1 and p_2 and these combinations result in only 152 unique BFs h after accounting for the permutations of the inputs x_1, x_2 and x_3 .

controlled by r TRs and each of these r TRs in turn have t_i incoming links from t_i genes where $i \in \{1, 2, \dots, r\}$, that is, each TR i is directly dependent on t_i genes. In this work, there are two biological interpretations of the TRs that we investigate. The first interpretation is that TRs are either TFs or a complex of TFs [43, 142, 143]. In this interpretation, one set of nodes of the bipartite network are the TFs or their complexes, and the other set of nodes are the genes that code for TFs. An edge from a TR to a gene denotes the transcriptional regulation of the gene's expression, whereas an edge from a gene to the TR indicates the expression of TFs or formation of TF complexes. This picture attempts to capture the regulation of a gene's expression by another gene via TFs or their complexes as the intermediary between the genes. The second interpretation, which is novel to this work, is that TRs are the enhancers and promoters of a gene that are bound by TFs. In this interpretation, one set of nodes are genomic regulatory elements such as enhancers and promoters bound to TFs and the other set of nodes are the genes that code for TFs. Here, edges from enhancers and promoters bound to TFs, to genes, indicate the regulation of a gene's expression by its regulatory elements and edges from genes to enhancers and promoters bound to TFs indicate the TFs that bind to the enhancers and promoters. This picture attempts to capture the regulation of a gene's expression by TFs binding to the promoter and enhancer regions of a gene. In recent work, Fink and Hannam [43] termed such a subgraph in the bipartite model as a *composition structure*, and denoted it as $\{t_1, t_2, \dots, t_r\}$ (see Figure 6.1(b)); since the composition graph is a tree of depth 2, the ordering of the degrees (i.e., t_i s) is arbitrary and so one can force the sequence $\{t_1, t_2, \dots, t_r\}$ to be increasing. In their work, Fink and Hannam [43] assumed that the t_1, t_2, \dots, t_r genes directly controlling the r TRs in the subgraph are distinct. Evidently, the sum $k = t_1 + t_2 + \dots + t_r$ gives the number of genes whose products directly regulate the targeted gene in the bipartite model. In other words, this sum k in the bipartite model gives the number of inputs k to a gene in the corresponding unipartite model.

Clearly, for a given value of k , there are multiple composition structures possible. For instance, the possible composition structures for $k = 4$ are: $\{1, 1, 1, 1\}$, $\{1, 1, 2\}$, $\{1, 3\}$, $\{2, 2\}$ and $\{4\}$. Fink and Hannam [43] refer to the subset of functions within all 2^{2^k} BFs resulting from the restrictions imposed by the composition

structure as *composed Boolean functions* or simply *composed BFs*.

6.1.3 Composed BFs

Consider a composition structure $\{t_1, t_2, \dots, t_r\}$ in the bipartite BN framework. Let there be a gene whose transcriptional regulation depends on the states of r TRs according to a BF g with r inputs. We denote the BF g as $g = g(y_1, y_2, \dots, y_r)$, where y_1, y_2, \dots, y_r are the states of the r TRs. The state of each TR i , where $i \in \{1, 2, \dots, r\}$, in turn depends on the states of t_i genes according to a BF p_i with t_i inputs.

Let us denote the states of the $k = t_1 + t_2 + \dots + t_r$ genes directly controlling the r TRs as $x_1, \dots, x_{t_1}, \dots, x_{t_1+t_2}, \dots, x_k$. It follows that:

$$\begin{aligned} y_1 &= p_1(x_1, \dots, x_{t_1}), \\ y_2 &= p_2(x_{t_1+1}, \dots, x_{t_1+t_2}), \\ &\vdots \\ y_r &= p_r(x_{t_1+t_2+\dots+t_{r-1}+1}, \dots, x_k). \end{aligned}$$

The regulation of a gene in the composition structure $\{t_1, t_2, \dots, t_r\}$ ultimately depends on the states of k genes according to some BF h of k inputs. This BF h is in fact the composition of the BFs p_1, p_2, \dots, p_r fed into g , that is:

$$\begin{aligned} &g(y_1, y_2, \dots, y_r) \\ &= g(p_1(x_1, \dots, x_{t_1}), \dots, p_r(x_{t_1+t_2+t_{r-1}+1}, \dots, x_k)) \\ &= h(x_1, x_2, \dots, x_k). \end{aligned}$$

In the above equation, the BF h is said to be a composed BF. There are no restrictions on the BFs that can be assigned to p_1, p_2, \dots, p_r or g . Therefore, the upper limit on the possible number of composed BFs h is:

$$2^{2^{t_1}} 2^{2^{t_2}} \dots 2^{2^{t_r}} 2^{2^r}.$$

However, the $2^{2^{t_1}} 2^{2^{t_2}} \dots 2^{2^{t_r}} 2^{2^r}$ BFs thereby composed are generally not all distinct, and it is necessary to remove the redundancies to obtain the set of (non-redundant) composed BFs. Such a non-redundant set of composed BFs is referred to as *biological logics* by Fink and Hannam [43], and in the present work we will refer to this non-redundant set of BFs as simply the *composed BFs*.

From the definition of composed BFs, it follows that if a BF h is associated with a composition structure $\{t_1, t_2, \dots, t_r\}$, then its complement \bar{h} is also associated with the same composition structure as we will show in Property 6.3.1. Figure 6.1(b) provides a schematic illustration of a composed BF belonging to the composition structure $\{1, 2\}$. Here, the state of a given gene x_4 depends on its input TRs y_2 and y_3 according to a 2-input BF $g(y_2, y_3)$. Further, y_2 depends on the input gene x_1 according to a 1-input BF $p_1(x_1)$, and y_3 depends on input genes x_2 and x_3 according to a 2-input BF $p_2(x_2, x_3)$. Thus, in this composition structure $\{1, 2\}$, the state of x_4 ultimately depends on x_1, x_2 and x_3 according to a composed BF of 3-inputs $h(x_1, x_2, x_3) = g(p_1(x_1), p_2(x_2, x_3))$. Such a composition of BFs reduces the number of allowed 3-input BFs.

Fink and Hannam [43] have provided exact analytical expressions for the number of composed BFs in a composition structure. Following these analytical expressions, it can be easily shown that the composed BFs belonging to the two composition structures $\{1, 1, 1, \dots, 1\}$ and $\{k\}$ do not restrict the space of k -input BFs, and they each include all 2^{2^k} possible BFs as we will show in Property 6.3.2. Thus, for k -input BFs, these two composition structures can be considered as trivial whereas the remaining composition structures are in fact non-trivial. Further, it is easy to see that there are no non-trivial composition structures for 1-input and 2-input BFs. Importantly, we excluded all the trivial composition structures from the analyses reported in this work, and in particular, we focus on non-trivial composition structures corresponding to 3, 4 and 5 input BFs. Finally, composition structures also allow for autoregulation, an important feature in determining the attractor landscape [152], wherein a TF associated with a gene can regulate the expression of that same gene.

6.2 Empirical evidence for the presence of composition structures

6.2.1 Quantifying the presence of protein complexes that can act as transcriptional regulators in Humans

Genes often come in families following either segmental or whole genome duplications, and that is the case in particular for those coding for TFs. There are several organisms where it has been shown that TFs within a given family form complexes in the form of heterodimers or even multimers contributing to gene regulation [153–156]. For instance the family of TFs called *auxin response factors* (ARFs) includes over 20 members in numerous plants and it has been shown that they form heterodimers that activate gene transcription [155, 157]. However, a quantitative assessment of the frequency at which heteromeric complexes contribute to gene regulation has not been carried out. The prevalence of such complexes in real-world GRNs can provide empirical support for the (frequent or not) occurrence of non-trivial composition structures.

We obtained a list of 1325 macromolecular complexes in *H. sapiens* from the EBI Complex Portal database [158], and the list of 1639 human TFs from <http://humantfs.ccb.utoronto.ca/> provided by Lambert *et al.* [159]. Among the 1639 human TFs, we selected only those TFs that were reviewed in the SWISS-PROT [160] protein database, resulting in a list of 1617 human TFs that was used for further analysis. We found that among the 1325 complexes in *H. sapiens*, 169 satisfy the constraint of being heteromeric with all subunits corresponding to TFs (see Table F.1, Appendix F). Of those, 165 are heterodimers and the remaining 4 are heterotrimers. Furthermore, there are 84 unique TFs composing these 169 complexes. Second, we manually searched for DNA binding evidence for each of these 169 heteromeric complexes and found that DNA binding has been verified for 86 of them. This then leaves us with 86 validated complexes of TFs that act as TRs, and thus, are likely candidates for forming composition structures.

Another approach we take to estimate the number of protein complexes acting as TRs

is to determine from the literature if there are TF families known to form heterodimers. Two genes coding for a protein can derive from a common ancestor (by duplication) leading to paralogs, and in particular to proteins with similar sequences, structure and function. Thus the complex forming propensity of a TF is expected to be conserved across the elements in that particular family. In evolution, this phenomenon is so common that one often has dozens or more genes belonging to the same family. We thus explored the specific importance of heteromers of TFs belonging to particular families. Indeed, it is known that certain classes of TFs, for instance basic leucine zipper (bZIP) [154, 161] and basic helix-loop-helix (bHLH) [162] classes, bind to DNA as homo- or hetero-dimers [156, 163, 164]. Knowing the prevalence of such TFs could shed light on the abundance of dimeric complexes which act as TRs. Thus for the 1617 TFs in *H. sapiens* we used the JASPAR database [165] to obtain the associated TF families; JASPAR provides a manually curated list of DNA TF binding motifs, the corresponding TFs, family information etc. Focusing on the TFs of the bZIP and bHLH families, we found 36 TFs in the first family and 38 in the second family (see Table F.2, Appendix F). Although our current data suggests that TF complexes are not so prevalent, we cannot rule out that this conclusion is an artifact of insufficient experimental evidence on complexes regulating genes.

6.2.2 Quantifying the presence of protein complexes that can act as transcriptional regulators in Yeast

A similar count of complexes involved in transcriptional regulation in *Saccharomyces cerevisiae* is presented in Tables F.3 and F.4, Appendix F. To perform the empirical analysis in *S. cerevisiae*, we first obtained a list of 617 macromolecular complexes in *S. cerevisiae* from the EBI Complex portal database [158]. Then we obtained the list of TFs in *S. cerevisiae* from the Yeastract database [166]. To do this, we obtained a list of 5195 verified genes from the SGD YeastMine database [167], and provided these genes as input to the Yeastract database. Additionally, we used the query *DNA binding evidence or expression evidence* in the Yeastract database. *DNA binding evidence* includes TF regulation verified by experiments such as EMSA, ChIP, ChIP-chip and ChIP-seq, DNA footprinting,

whereas *expression evidence* includes those interactions established by comparing gene expression in wild-type strains with mutant strains in which the gene encoding the TF is mutated. Expression evidence is obtained via northern blotting, RT-PCR, DNA microarrays and RNA-seq experiments [166]. The above query in the YeastRACT database resulted in a list of 217 TFs that were used for further analysis. Note that 153 out of these 217 TFs show evidence for both DNA binding and effects on expression. Finally, among the 617 macromolecular complexes, we selected only those complexes in which all the protein subunits correspond to TFs.

Using the EBI complex portal and the TFs from YeastRACT database, we found that there are 17 heteromeric complexes among the 617 complexes in *S. cerevisiae* such that each of their protein subunits correspond to TFs (see Table F.3, Appendix F). Thereafter, we ascertained via manual curation of the literature associated with these 17 complexes that 15 of them act as TRs, 1 of them binds to DNA but whether it regulates gene expression is uncertain, and the remaining 1 acts as a transcriptional co-repressor. Among the 15 complexes that act as TRs, 9 complexes are formed by 2 proteins, 5 complexes are formed by 3 proteins and 1 complex is formed by 4 proteins. There are 30 unique TFs whose combination results in these 15 complexes. Additionally, we found that 11 out of these 15 complexes are such that their protein subunits show both DNA binding and expression evidence.

It is known from experiments that TFs belonging to the bZIP [161, 163] and bHLH [162, 164] classes typically bind DNA as dimers. To determine the classes of the 217 TFs in *S. cerevisiae*, we utilized the JASPAR database [165]. In *S. cerevisiae*, we found 10 TFs that belong to the bZIP class and 7 TFs that belong to the bHLH class (see Table F.4, Appendix F). Further, we find that all the 17 TFs in the bZIP or bHLH classes display evidence for both DNA binding and effects on expression.

6.2.3 TF binding regions and active enhancers

We relied on two types of published datasets for estimating the prevalence of composition structures arising from cis-regulatory modules: (i) TF binding regions and (ii) active

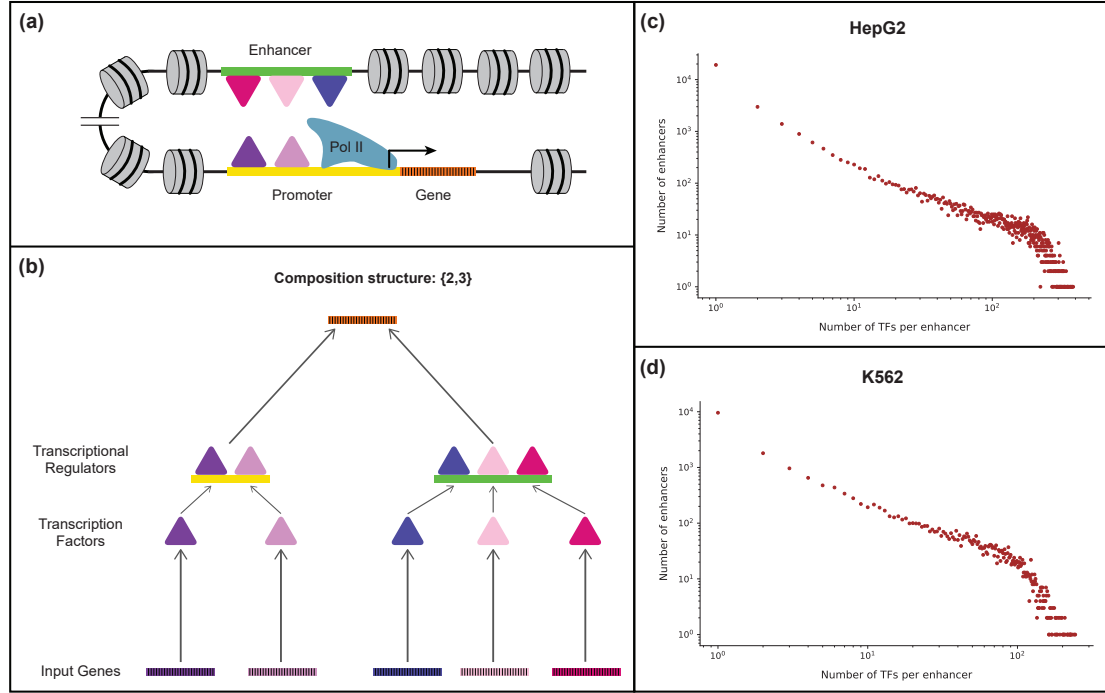


Figure 6.2: Non-trivial composition structures arising due to enhancers bound by multiple TFs. (a) A biologically plausible mechanism revealing the occurrence of non-trivial composition structures in transcriptional gene regulation. Multiple TFs can bind to the promoter as well as the enhancer region(s) of a target gene. The enhancers and promoters bound by TFs then act as TRs of their target genes, resulting in non-trivial composition structures. (b) A schematic representation of the composition structure $\{2,3\}$ arising in sub-figure (a). The target gene is regulated by an active promoter that is bound by 2 TFs, and an active enhancer that is bound by 3 TFs. (c) Scatter plot showing the number of active enhancers bound by a given number of TFs in the HepG2 cell line in humans. We found that 32.68% of the active enhancers in HepG2 are bound by at least 2 TFs. (d) Scatter plot showing the number of active enhancers bound by a given number of TFs in the K562 cell line in humans. We found that 44.31% of the active enhancers in K562 are bound by at least 2 TFs. The x and y axes in part (c) and (d) are in log scale. These results suggest that non-trivial composition structures are prevalent in GRNs.

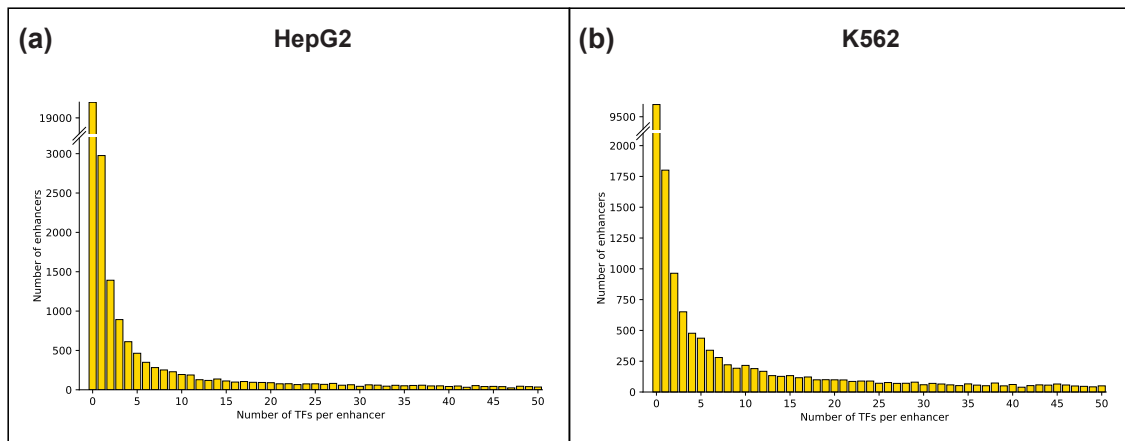


Figure 6.3: Prevalence of active enhancers bound by multiple TFs in human cell lines. (a) Histogram showing the number of active enhancers bound by a given number of TFs in the HepG2 cell line in humans. We found that 32.68% of the active enhancers in HepG2 are bound by at least 2 TFs. (b) Histogram showing the number of active enhancers bound by a given number of TFs in the K562 cell line in humans. We found that 44.31% of the active enhancers in K562 are bound by at least 2 TFs. These results suggest that non-trivial composition structures are prevalent in GRNs.

enhancers. We focused on the two well-studied human cell lines HepG2 and K562 because there is ample published data for them. We obtained the DNA binding regions of the TFs as ChIP-seq narrowPeak bed files for the two cell lines from the human ENCODE project [168]. The active enhancers are obtained from data processed using the STARRPeaker peak-calling software [169]. Employing these two datasets, we consider that a TF binds to an active enhancer if and only if both the midpoint and the summit of the ChIP-seq peaks for that TF fall within the active enhancer region. Notably, there were no cases wherein the summit of the peaks were not provided in the ChIP-seq files obtained from the human ENCODE project. The ChIP-seq narrowPeak bed files for the HepG2 and K562 cell lines were last downloaded on April 28th 2022 and April 29th 2022, respectively, from the human ENCODE project website: <https://www.encodeproject.org>. The processed datasets from human ENCODE used for this analysis and the associated codes are available at: <https://github.com/asamallab/CoSt>. This study was carried out in accordance with relevant guidelines and regulations.

6.2.4 Composition structures arising through enhancers

Bipartite BN models provide a quite general framework and so for instance composition structures can accommodate other mechanisms of eukaryotic gene regulation than the one involving complexes as covered in the previous sub-section. Here, we propose one such alternative picture where the intermediate TRs are no longer protein complexes but are associated with *cis*-regulatory modules such as promoters, enhancers, or insulators. In eukaryotes, transcription is typically regulated via the binding of TFs upstream of the gene [144, 145, 170]. Promoters are located close to the transcription start site where RNA polymerases and TFs assemble to initiate transcription [171]. Enhancers on the other hand may be located at rather large distances (in fact both upstream or downstream) of the target gene they regulate [172]. Enhancers are *active* or *inactive* based on whether their chromatin state is accessible or not; in the former case, TF binding sites within these enhancers can attract specific TFs and thus modulate transcription of nearby genes [145]. Interestingly, a given enhancer typically contains multiple such binding sites and is thus considered to be a *cis*-regulatory module [145, 173].

Figures 6.2(a) and 6.2(b) illustrate how enhancers and promoters may act as TRs in the composition structure $\{2, 3\}$ where we have chosen to have 2 TFs binding to the promoter and 3 TFs binding to the enhancer. One can suppose that an abundance of enhancers containing multiple TF binding sites is suggestive of the prevalence of non-trivial composition structures in real-world GRNs. In view of this possibility, we perform an analysis to provide a quantitative estimate of the number of TFs that bind to active enhancers in two widely-studied human cell lines namely, HepG2 and K562.

For the cell line HepG2, we used ChIP-seq peaks provided for 458 unique TFs and a total of 32929 enhancers detected as active (see Section 6.2.3). 2976 enhancers had exactly one TF binding within their region while 10754 enhancers had two or more TFs binding within their region, representing 32.68% of the total number of enhancers in HepG2 (see Figure 6.2(c)). Additionally, of the 458 TFs for which data is available in HepG2, we found that 456 TFs bind to at least one of the enhancers detected as active. For the cell line K562, we used ChIP-seq peaks provided for 323 unique TFs and a total of 20471 enhancers

detected as active. 1801 enhancers had exactly one TF binding within their region while 9071 enhancers had two or more TFs binding within their region, representing 44.31% of the total number of enhancers in K562 (see Figure 6.2(d)). Additionally, of the 323 TFs for which data is available in K562, we found that 322 TFs bind to at least one of the enhancers detected as active. The fact that 32.68% and 44.31% of the active enhancers in HepG2 and K562, respectively can be bound by at least two TFs suggest that non-trivial composition structures indeed do arise frequently in gene regulatory logics. Figures 6.3(a) and 6.3(b) are the same as Figures 6.2(c) and 6.2(d) respectively, except that Figures 6.2(c) and 6.2(d) are plotted on a log-log scale.

6.3 Characterizing the space of composed BFs

6.3.1 Properties of composed BFs

Property 6.3.1. Given a composition structure $\{t_1, t_2, \dots, t_r\}$, if h is a possible composed BF, then its complement \bar{h} is also a possible composed BF.

Proof: h is a composed BF of the form $g(p_1, p_2, \dots, p_r)$, where each p_i is a BF of t_i inputs. There are $2^{2^{t_1}} 2^{2^{t_2}} \dots 2^{2^{t_r}} 2^{2^r}$ combinations of p_1, p_2, \dots, p_r and g which comprise the composed BFs h . Let us consider one such combination $p_1^*, p_2^*, \dots, p_r^*$ and g^* , that corresponds to a BF h^* from the space of all composed BFs. Now, there also exists another combination $p_1^*, p_2^*, \dots, p_r^*$ and \bar{g}^* among the $2^{2^{t_1}} 2^{2^{t_2}} \dots 2^{2^{t_r}} 2^{2^r}$ combinations, which corresponds to the BF \bar{h}^* . Hence, a function and its complement are both present in the composed BFs of any composition structure.

Property 6.3.2. The composition structure $\{k\}$ does not restrict the space of BFs.

Proof: The composition structure $\{k\}$ corresponds to a k -input BF of the form $h = g(p_1(x_1, x_2, \dots, x_k))$. There are 2^{2^k} BFs that can be assigned to p_1 , and $2^{2^1} = 4$ BFs that can be assigned to g . Among the 4 BFs that can be assigned to g , if we consider the $g(p_1) = p_1$, then $h = p_1(x_1, x_2, \dots, x_k)$. It follows that any BF among the 2^{2^k} possible k -input BFs can be assigned to h . Since h spans all the possible k -input BFs, the composition structure $\{k\}$ cannot restrict the space of BFs.

Table 6.1: Comparison of the number and fraction of BFs allowed by different composition structures, with and without including all possible permutations of input variables. The composition structures in the bipartite BN framework of transcriptional gene regulation are categorized based on the number of inputs k to a gene in the corresponding unipartite BN framework. The column “Number of composed BFs” gives the number of distinct BFs in a composition structure, and the subcolumns provide a comparison of the number of such BFs both without and with the accounting for all possible permutations of the input variables. The column “Fraction of composed BFs” gives the fraction of distinct BFs in a composition structure among all possible BFs for a given number k of inputs, and the subcolumns provide a comparison of the fraction of such BFs both without and with the accounting for all possible permutations of the input variables.

k	Composition structure	Number of composed BFs		Fraction of composed BFs	
		without permutation	with permutation	without permutation	with permutation
1	{1}	4	4	1	1
2	{2}	16	16	1	1
	{1,1}	16	16	1	1
3	{3}	256	256	1	1
	{1,2}	88	152	0.344	0.594
	{1,1,1}	256	256	1	1
4	{4}	65536	65536	1	1
	{1,3}	1528	4864	0.023	0.074
	{2,2}	520	1208	0.008	0.018
	{1,1,2}	1696	6216	0.026	0.095
	{1,1,1,1}	65536	65536	1	1
5	{5}	4294967296	4294967296	1	1
	{1,4}	393208	1921928	9.16×10^{-05}	4.47×10^{-04}
	{2,3}	9160	71608	2.13×10^{-06}	1.67×10^{-05}
	{1,1,3}	30496	263488	7.10×10^{-06}	6.13×10^{-05}
	{1,2,2}	11344	100768	2.64×10^{-06}	2.35×10^{-05}
	{1,1,1,2}	457216	3446488	1.06×10^{-04}	8.02×10^{-04}
	{1,1,1,1,1}	4294967296	4294967296	1	1

6.3.2 Accounting for all the permutations of inputs of composed BFs

In their procedure to count BFs arising from a composition structure, Fink and Hanam [43] do not account for permutations of the input variables, that is they ignore

the indices of the inputs. In the present work, we have extended Fink and Hannam’s counting approach by accounting for all the permutations of input variables in a given composition structure. Consider a composed BF of the type $g(p_1(x_1), p_2(x_2, x_3))$ that belongs to the composition structure $\{1, 2\}$ and corresponds to a 3-input BF $h(x_1, x_2, x_3)$. Taking $p_1(x_1) = x_1$, $p_2(x_2, x_3) = x_2x_3$, and $g(x, y) = x + y$ leads to the composed BF $h(x_1, x_2, x_3) = x_1 + x_2x_3$. However the BFs obtained by permuting the indices of these variables, namely $x_2 + x_1x_3$ and $x_3 + x_1x_2$, are just as relevant biologically; indeed, the indices point to genes and these are hardly ever equivalent. Thus, we count all three of the cases above as valid composed BFs. In contrast, Fink and Hannam [43] count them as one composed BF. A code to generate all the composed BFs for any given composition structure after accounting for all the permutations of the input variables is available from the associated GitHub repository: <https://github.com/asamallab/CoSt>. Note that this example shows that the two ways of counting are not generally related by the number of permutations ($k!$) of k indices because of possible symmetries within these expressions. Including all possible permutations of inputs is sufficient to ensure that all isomorphisms (i.e., permutations and negations of inputs) of a BF in a composition structure are also present therein. Note that the procedure used to count the corrected values of the number of composed BFs in the present work is purely computational and is based on enumeration. Such a computational approach limits our ability to count the number of BFs belonging to most composition structures beyond $k = 5$ inputs.

Table 6.1 provides a comparison of the number of distinct BFs allowed by different composition structures for $k \leq 5$ inputs, both with and without including all possible permutations of the input variables. Table 6.1 also provides these results as fractions among all possible BFs for $k \leq 5$ inputs. Naturally, we find that accounting for all possible permutations of inputs increases the number of BFs in a composition structure in comparison to those reported by Fink and Hannam [43]. However, this does not alter the central result of Fink and Hannam [43], that is, composition structures significantly restrict the space of possible BFs. This is evident from the trends for the fractions of composed BFs among all possible BFs as a function of the number of inputs (see Table 6.1).

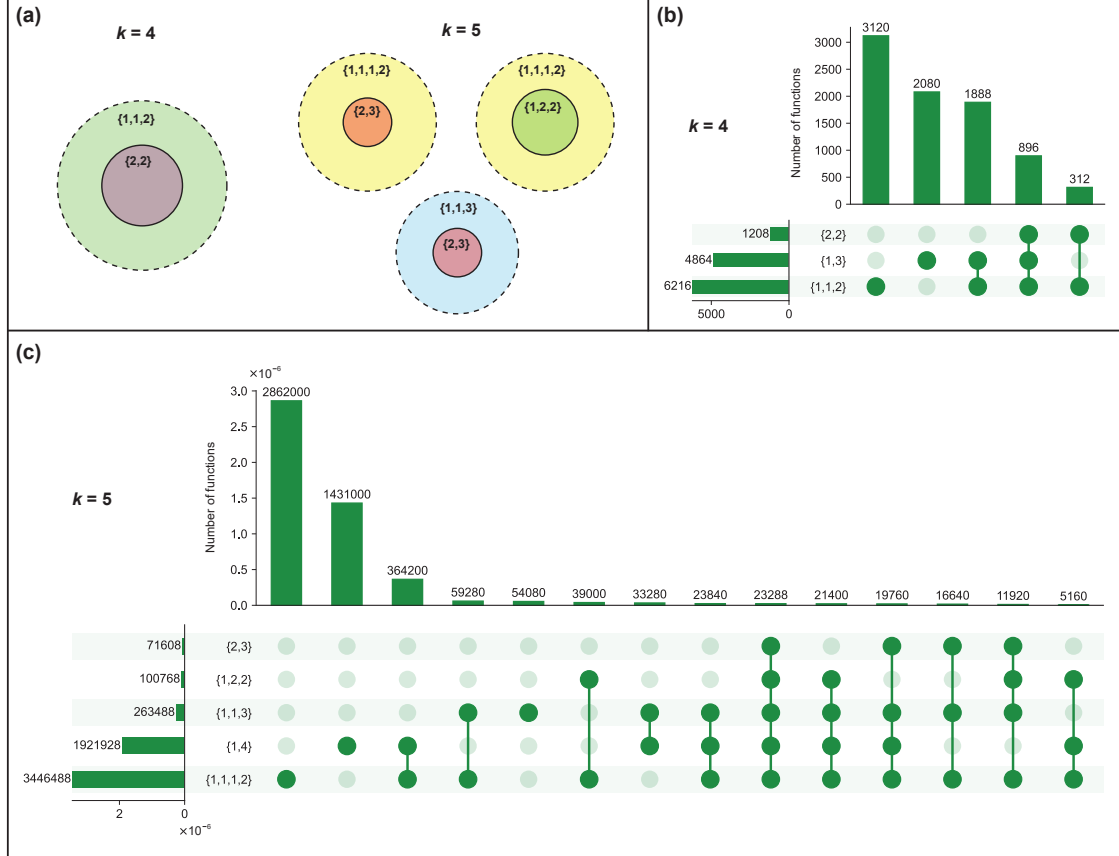


Figure 6.4: Overlaps between the sets of BFs compatible with different composition structures at $k = 4$ and $k = 5$ inputs. (a) Venn diagrams illustrating *proper* subsets among the sets of non-trivial composition structures at $k = 4$ and $k = 5$ inputs. (b) UpSet plot illustrating the number of BFs that are present in all possible intersections of non-trivial composition structures at $k = 4$ inputs. (c) UpSet plot illustrating the number of BFs that are present in all possible intersections of non-trivial composition structures at $k = 5$ inputs. The horizontal bars in the UpSet plots indicate the number of BFs that are present in different composition structures. The vertical bars indicate the number of BFs that are simultaneously present in some and absent from other composition structures, as specified by the underlying dark and light green circles.

6.3.3 Overlap of composed BFs across various k -input composition structures

There are multiple composition structures $\{t_1, \dots, t_r\}$ possible for a given number of inputs k such that $t_1 + t_2 + \dots + t_r = k$, and each composition structure allows a certain set of BFs. However, composed BFs can belong to more than one composition structure. Therefore, it is worthwhile to examine the overlaps of composed BFs across all non-trivial composition structures with a given number of inputs k . Here, we analysed the intersections of composed BFs across non-trivial composition structures for $k = 4$ and $k = 5$ inputs. We reiterate that there are no non-trivial composition structures for $k = 1$ and $k = 2$ inputs, and note that $\{1,2\}$ is the only non-trivial composition structure for $k = 3$.

For $k = 4$ inputs, we find that the set of BFs in the composition structure $\{2,2\}$ is a proper subset of the set of BFs in $\{1,1,2\}$ (see Figure 6.4(a)). For $k = 5$ inputs, we find that the set of BFs in the composition structure $\{2,3\}$ is a proper subset of the set of BFs in $\{1,1,3\}$ as well as $\{1,1,1,2\}$, and the set of BFs in the composition structure $\{1,2,2\}$ is a proper subset of the set of BFs in $\{1,1,1,2\}$ (see Figure 6.4(a)). Further, we give the number of BFs in all possible intersections of non-trivial composition structures for $k = 4$ and $k = 5$ inputs through UpSet plots [174] in Figures 6.4(b) and 6.4(c), respectively.

6.3.4 Comparing restriction levels: composition structures versus biologically meaningful types

Clearly, imposing a non-trivial composition structure significantly restricts the space of allowed BFs within the complete space of BFs with k inputs. As shown by some of us recently [49], the same holds when imposing certain biologically meaningful properties. Here, we compare the level of restriction achieved by four established biologically meaningful types of BFs, namely unate functions (UFs), canalizing functions (CFs), nested canalizing functions (NCFs) and read-once functions (RoFs), to that achieved by composed BFs of a given composition structure, in the space of all BFs with k inputs. See Section 2.2, Chapter 2 for the formal definitions of biologically meaningful BFs. Among

Table 6.2: Number of BFs in different composition structures that display biologically meaningful properties. The number of BFs within a non-trivial composition structure that also belong to each of the four types of biologically meaningful functions, namely unate functions (UFs), canalyzing functions (CFs), nested canalyzing functions (NCFs) and read-once functions (RoFs). The column “Number of composed BFs” gives the number of BFs that are allowed in a given composition structure.

Composition structure	Number of composed BFs	Number of biologically meaningful BFs in composition structure			
		UF	CF	NCF	RoF
$\{1,2\}$	152	96	120	64	64
$\{1,3\}$	4864	1210	3514	736	736
$\{2,2\}$	1208	634	730	224	320
$\{1,1,2\}$	6216	1370	1850	736	832
$\{1,4\}$	1921928	41676	1292276	10624	12544
$\{2,3\}$	71608	13676	33596	3264	6784
$\{1,1,3\}$	263488	26156	80996	10624	14144
$\{1,2,2\}$	100768	17836	25236	5504	9984
$\{1,1,1,2\}$	3446488	61516	122516	10624	15104

the four different types of biologically meaningful BFs, it is known that the NCFs represent the smallest fraction in the space of all BFs [49] (see Chapter 2). For $k = 4$ and $k = 5$ inputs, we find that certain composition structures restrict very strongly, though less than NCFs (see Table F.5, Appendix F). Specifically, at $k = 4$ inputs, $\{2,2\}$ is the most restrictive one. The composed BFs in $\{2,2\}$ occupy a fraction of 0.018 among all BFs, which is 1.63 times greater than the fraction occupied by NCFs at $k = 4$ (whose value is 0.011). For $k = 5$ inputs, $\{2,3\}$ is the most restrictive composition structure. The BFs in $\{2,3\}$ occupy a fraction of 1.67×10^{-5} , which is about 6.76 times greater than the fraction occupied by NCFs at $k = 5$ (whose value is 2.47×10^{-6}). In Table F.5, Appendix F, we compare the fraction of BFs in the most restrictive composition structure to the fractions for each of the four types of biologically meaningful BFs for $k \leq 5$ inputs.

We next evaluated how often a BF in a composition structure also displays biologically meaningful properties. Table 6.2 shows the number of composed BFs that belong to each of the four types of biologically meaningful BFs for non-trivial composition structures with

$k \leq 5$ inputs. Clearly imposing BFs to be biologically meaningful and to be compatible with a given composition structure severely restricts the possible BFs. We also find that certain types of biologically meaningful BFs, in particular NCFs, are proper subsets of BFs in certain composition structures. Specifically, all the 64 NCFs with $k = 3$ inputs are contained in the composition structure $\{1, 2\}$, all the 736 NCFs with $k = 4$ inputs are contained in the composition structures $\{1, 3\}$ and $\{1, 1, 2\}$, and all the 10624 NCFs with $k = 5$ inputs are contained in the composition structures $\{1, 4\}$, $\{1, 1, 3\}$ and $\{1, 1, 1, 2\}$. Moreover, all CFs with $k = 3, 4$, and 5 inputs are a subset of the composition structures $\{1, 2\}$, $\{1, 3\}$ and $\{1, 4\}$, respectively, whereas all RoFs with $k = 4$ and 5 inputs are a subset of the composition structures $\{1, 1, 2\}$ and $\{1, 1, 1, 2\}$, respectively. In Table F.6, Appendix F, we provide the fraction of composed BFs that belong to each of the four types of biologically meaningful BFs for non-trivial composition structures with $k \leq 5$ inputs.

We also computed the number and fraction of composed BFs for different composition structures which have odd bias. Recently, some of us showed that BFs with odd bias are preponderant among BFs in reconstructed BN models of biological systems [49]. Furthermore, it was shown that NCFs [86] and RoFs [49] have odd bias. Here, we find that the fraction of BFs with odd bias in any composition structure with $k \leq 5$ inputs is less than 0.5 (see Table F.7, Appendix F). Additionally, we find that BFs – with any given even bias – occur in all composition structures with $k \leq 5$ inputs. In Table F.7, Appendix F, we list the odd biases of BFs that are present in composition structures with $k \leq 5$ inputs.

6.4 Enrichments of composed BFs in reconstructed Boolean models of gene regulatory networks

In this section, we present the results of our analyses of the abundances of composed BFs in a compiled reference biological dataset of 2687 BFs from 88 published BN models of biological systems [49]. More explicitly, we do not reconstruct the composition structures associated with each of the 2687 BFs in our database (see Section A.1, Appendix A),

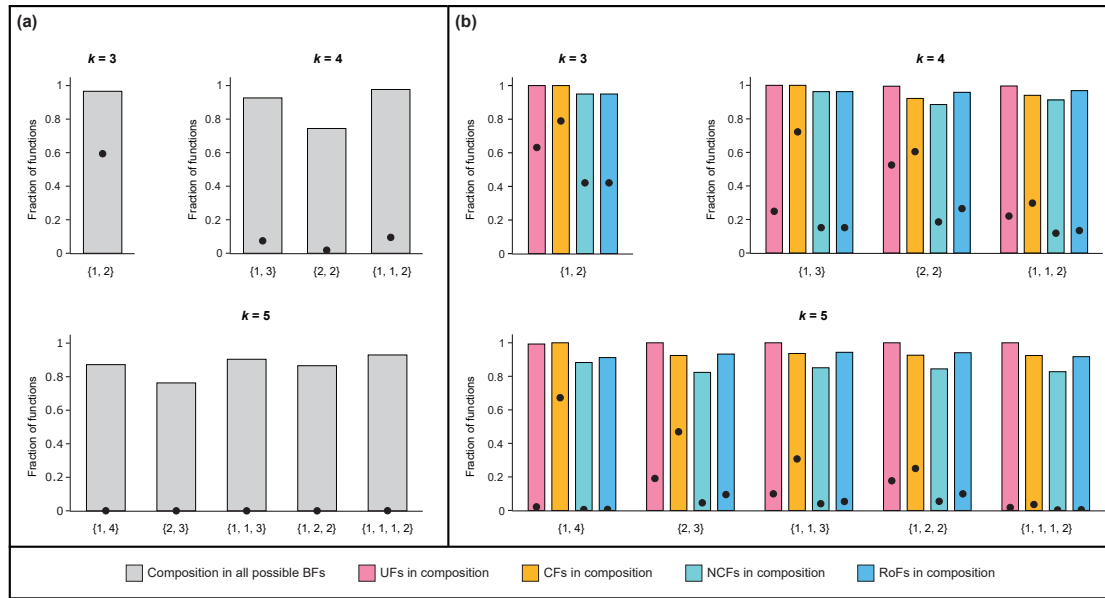


Figure 6.5: Abundance of composed BF in reconstructed biological networks. (a) Bar plots give the fraction of BF in the reference biological dataset that are compatible with each of the composition structures. The black dots indicate the fraction when considering all possible BF instead of only the ones in the reference biological dataset. Note that since the sets of BF allowed by different composition structures overlap with each other, the sum of the bar plot values may be larger than 1. (b) For all BF of the reference biological dataset compatible with a given composition structure, the bars give the fraction of these BF that belong to each of the four biologically meaningful sub-types: unate functions (UFs), canalizing functions (CFs), nested canalizing functions (NCFs), and read-once functions (RoFs). Again, the black dots give these fractions when considering instead all possible BF.

Table 6.3: Relative enrichment of biologically meaningful BF’s among composed BF’s of different composition structures in the reference biological dataset. This table gives the relative enrichment values E_R in the reference biological dataset for the four biologically meaningful sub-types within composed BF’s for different non-trivial composition structures with number of inputs $k \leq 5$. These four biologically meaningful sub-types within composed BF’s include those BF’s in a composition structure that also happen to be unate functions (UFs), canalizing functions (CFs), nested canalizing functions (NCFs), or read-once functions (RoFs).

Composition structure	E_R of biologically meaningful sub-types in a given composition structure			
	UF	CF	NCF	RoF
{1,2}	1.58	1.27	2.26	2.26
{1,3}	4.02	1.38	6.36	6.36
{2,2}	1.90	1.53	4.77	3.62
{1,1,2}	4.52	3.16	7.71	7.23
{1,4}	45.78	1.49	159.62	139.70
{2,3}	5.24	1.97	18.07	9.85
{1,1,3}	10.07	3.05	21.11	17.57
{1,2,2}	5.65	3.70	15.46	9.49
{1,1,1,2}	56.03	26.00	268.47	209.30

but rather determine which composition structure each of the 2687 BF’s belong to, by comparing the real BF’s with the composed BF’s of various composition structures.

To begin, we computed two proportions for each possible composition structure. The first is the proportion of BF’s with k inputs in the reconstructed biological networks that belong to the given composition structure (bar plots in Figure 6.5(a)). The second is the corresponding proportion in the *random ensemble* with k inputs; that proportion is thus given by the number of BF’s with k inputs that are compatible with the given composition structure, divided by the total number of BF’s (black dots in Figure 6.5(a)). The results show that the proportions of composed BF’s in the *reference biological dataset* are larger than in the ensemble of random BF’s, for each composition structure, indicating that non-trivial composed BF’s are enriched in real biological networks. Note that the sets of BF’s allowed by different composition structures overlap with each other (see Figure 6.4), allowing for the sum of the height of the bars in Figure 6.5(a) to be larger than 1. To

Table 6.4: Comparison between the enrichments of composed BFs and biologically meaningful BFs of minimum complexity in the reference biological dataset. The table provides the enrichment factors when composed BFs in non-trivial composition structures (denoted as CS in the first column) with $k \leq 5$ inputs are compared with two classes of biologically meaningful BFs of minimum complexity namely, nested canalizing functions (NCFs) and read-once functions (RoFs). T_C denotes the set of composed BFs allowed by a composition structure at a given number of inputs k , T_{NCF} denotes the set of all k -input NCFs, and T_{RoF} denotes the set of all k -input RoFs. \cap represents the intersection of two sets and \setminus represents the set-theoretic difference. “–” in the columns $T_{NCF} \setminus T_C$ or $T_{RoF} \setminus T_C$ indicates that the NCFs or RoFs are a subset of the set of BFs allowed by the composition structure.

CS	$T_C \cap T_{NCF}$	$T_C \setminus T_{NCF}$	$T_{NCF} \setminus T_C$	$T_C \cap T_{RoF}$	$T_C \setminus T_{RoF}$	$T_{RoF} \setminus T_C$
$\{1,2\}$	3.67	0.14	–	3.67	0.14	–
$\{1,3\}$	79.38	0.55	–	79.38	0.55	37.04
$\{2,2\}$	192.78	5.68	29.77	146.06	2.29	29.77
$\{1,1,2\}$	79.38	1.02	–	74.49	0.38	–
$\{1,4\}$	310977.13	230.48	–	272157.87	173.03	96791.63
$\{2,3\}$	826630.05	8459.68	82296.27	450476.77	3397.73	72800.54
$\{1,1,3\}$	310977.13	2286.48	–	258889.63	883.34	0.00
$\{1,2,2\}$	570245.27	6069.12	32263.88	350214.73	2426.14	32263.88
$\{1,1,1,2\}$	310977.13	200.33	–	242434.78	96.28	–

consider this question in greater depth, we define the *enrichment factor* as the ratio of the first and the second proportions. For instance, for the composition structures $\{2,2\}$ and $\{2,3\}$ that are the most restrictive composition structures for $k = 4$ and $k = 5$ inputs, the corresponding enrichment factors are 40.37 and 45760.08. To check the level of significance of this effect, we applied our statistical tests (see Section 3.1, Chapter 3). In Table F.8, Appendix F, we list the enrichment factors for all non-trivial composition structures having $k \leq 5$ inputs and we give the corresponding one-sided p -values. These p -values show that the enrichment effects are indeed statistically significant, providing evidence in biological systems of a selection pressure in favor of each of the non-trivial composition structures.

Figure 6.5(b) is a bar plot of the fractions in the reference biological dataset of the four biologically meaningful sub-types when focusing on the BF's satisfying a given composition structure. In addition, the black dots give the corresponding fractions when using the random ensemble instead of the reference biological dataset. We call *relative enrichment*

E_R the ratio of these fractions that focuses on both a given composition structure and a given biologically meaningful sub-type of BF. The E_R s are larger than 1 for all non-trivial composition structures with number of inputs $k \leq 5$, suggesting that the four biologically meaningful sub-types of composed BFs are enriched within any composition structure in the reference biological dataset. Table 6.3 gives the E_R values for the four biologically meaningful sub-types in all non-trivial composition structures with number of inputs $k \leq 5$. Furthermore, the computed relative enrichment values are statistically significant as determined by one-sided p -values (see Table F.9, Appendix F).

A previous analysis [49] showed that biologically meaningful BFs are enriched in our reference biological dataset. Notably, those enrichments are likely driven by complexity minimization, with NCFs and RoFs respectively minimizing two complexity measures namely, average sensitivity and Boolean complexity [49]. An immediate question that then arises is whether the enrichments of composed BFs as found in Figure 6.5 might just be driven by enrichments of NCFs and RoFs. To examine that possibility, let T_C denote the set of BFs allowed by a composition structure C at a given number of inputs k , and let T_{NCF} denote the set of NCFs with k inputs. We have determined the enrichment factors of three disjoint sets of BFs: composed BFs that are also NCFs (i.e., $T_C \cap T_{NCF}$), composed BFs that are not NCFs (i.e., $T_C \setminus T_{NCF}$), and NCFs that are not composed BFs (i.e., $T_{NCF} \setminus T_C$). Table 6.4 shows the enrichment factors for these three disjoint sets of BFs, for all non-trivial composition structures with $k \leq 5$ inputs. We find that the BFs belonging to the set $T_C \cap T_{NCF}$ display a very high enrichment factor. Moreover, for composition structures $\{2, 2\}$, $\{2, 3\}$ and $\{1, 2, 2\}$, we find that both the sets $T_C \setminus T_{NCF}$ and $T_{NCF} \setminus T_C$ are enriched in the biological datasets. However, the enrichment factor is much larger for the set $T_{NCF} \setminus T_C$. Finally, for the composition structures $\{1, 2\}$, $\{1, 3\}$, $\{1, 1, 2\}$, $\{1, 4\}$, $\{1, 1, 3\}$ and $\{1, 1, 1, 2\}$ that are a superset of the corresponding NCFs, we find that the set $T_C \setminus T_{NCF}$ is either depleted or shows a lower enrichment factor compared to the set $T_C \cap T_{NCF}$. After repeating the above analysis for RoFs to estimate the enrichment factors for $T_C \cap T_{RoF}$, $T_C \setminus T_{RoF}$ and $T_{RoF} \setminus T_C$, we find that the results are similar to those for NCFs (see Table 6.4). Furthermore, all these enrichment factors are statistically significant as determined by one-sided p -values (see Table F.10, Appendix F).

These results suggest that although composed BFs are subject to positive selection in real biological networks, the primary driving force for enrichment is the property of being an NCF or an RoF.

We have also examined these questions for the other sub-types of biologically meaningful BFs. Table F.11, Appendix F lists the corresponding enrichment factors while Table F.12, Appendix F lists the associated p -values. First, we find that the set of BFs that are UFs but not composed BFs (i.e., $T_{UF} \setminus T_C$) are enriched whereas those BFs that are composed BFs but not UFs (i.e., $T_C \setminus T_{UF}$) are highly depleted. This suggests that UFs could also be a possible driving factor for the enrichment of composed BFs in biological networks. Second, BFs that are composed BFs but not CFs (i.e., $T_C \setminus T_{CF}$) are highly enriched compared to BFs that are CFs but not composed BFs (i.e., $T_{CF} \setminus T_C$). Though this result provides evidence for composition structures as a driving factor for the enrichment of CFs in real biological networks, we reiterate our earlier result that this enrichment is primarily driven by the property of being NCFs or RoFs.

6.5 Discussion

We began our empirical study into the potential biological relevance of non-trivial composition structures arising in bipartite GRNs by investigating two different scenarios. In the first, we estimated the degree of occurrence of heteromeric complexes formed by DNA-binding proteins while in the second we characterized co-occurrences of TF binding sites in enhancers.

In the scenario of transcriptional regulation by heteromeric complexes as proposed by Hannam *et al.* [142], a non-trivial composition structure arises when a gene is regulated by at least two TRs, of which at least one is a heteromeric protein complex made up of at least two monomers. From the data on macromolecular complexes in humans obtained from the EBI Complex Portal [158], we find that for approximately 6.5% of the complexes (86 out of 1325), all of their monomeric subunits are identified as TFs. (For such an identification, we imposed that they be present in the database of 1617 human TFs from Lambert *et al.* [159] and come with strong evidence for DNA binding as ascertained by manual curation of the

literature.) Furthermore, we find that 4.57% of the human TFs belong to the bZIP and bHLH classes that are known to bind to DNA as homodimers or heterodimers. It is likely that the collection of complexes in the EBI Complex Portal are biased towards complexes which do not act as TRs given that the detection and characterization of heteromeric protein complexes which act as TRs is experimentally challenging. Though our empirical analysis provides some support for Hannam’s picture of heteromeric protein complexes acting as TRs, the existing data on such complexes is insufficient to quantitatively estimate the prevalence of composition structures in real-world GRNs. Another point that requires critical assessment in this picture of gene regulation is the number of logic rules that govern the formation of the heteromeric complexes. Since a heteromeric complex is a conjunction of all its monomeric subunits, the only Boolean logic rule which captures the formation of a complex is the one linking all the components by the AND operator. In a general bipartite BN, the upper limit for the number of logics possible for the composition structure $\{t_1, t_2, \dots, t_r\}$ is $2^{2^{t_1}} 2^{2^{t_2}} \dots 2^{2^{t_r}} 2^{2^r}$, whereas if one imposes the AND logic for the formation of protein complexes only 2^{2^r} logics are possible.

The flexibility of the bipartite formalism allows us to capture a more nuanced scenario in gene regulation that involves *cis*-regulatory elements (such as enhancers and promoters) and the TFs which bind to them. In our picture, a target gene is regulated by *cis*-regulatory elements which act as TRs and each *cis*-regulatory element acts in a way that depends on the TFs that bind therein (see Figure 6.2(a)). Thus, a non-trivial composition structure is realized when a gene is regulated by at least two *cis*-regulatory elements, one of which is regulated by at least two TFs. We thus inferred whether non-trivial composition structures of this kind arise in GRNs by determining how often the enhancers of a gene are bound by at least two TFs. By analyzing ChIP-seq and enhancer datasets in the two human cell lines HepG2 and K562, we find that 32.68% and 44.31% of their respective active enhancers bind to at least two TFs. Our result suggests that composition structures with *cis*-regulatory elements acting as TRs are likely to be prevalent in bipartite GRNs. We remark that experimental limitations do not allow for the detection of all the enhancers for a given target gene in a given cell type, preventing the identification of exact composition structures from empirical data.

We also address many questions from the perspective of providing a comprehensive comparison between the BFs of a composition structure and the known types of biologically meaningful BFs. We first provide the corrected values for the number of BFs belonging to a given composition structure by accounting for all the isomorphisms for each of the composed BFs. Fink and Hannam [43] leave these out of their work. We find that NCFs and RoFs are more restrictive than the most restrictive composition structures. Next, by quantifying the overlaps between different composition structures, we find that BFs belonging to different composition structures may partially overlap but some composition structures may in fact be subsets of other composition structures, e.g. $\{2,2\}$ is a subset of $\{1,1,2\}$. Following this, we compute the intersections between composition structures and biologically meaningful BFs and find that of 9 composition structures (up to 5-input BFs), the NCFs are a subset of 6 composition structures.

Moving to results derived from the reference biological dataset of 2687 BFs, we find that the composed BFs are indeed enriched in our dataset in comparison to the space of all BFs. Then by computing the relative enrichment of a biologically meaningful sub-type in non-trivial composition structures (for instance, the relative enrichment of NCFs when considering BFs compatible with the composition structure $\{2,2\}$), we find that these sub-types are enriched, though the cause of its enrichment could be attributed either to the property of being biologically meaningful or to the property of belonging to the composition structure. To decide between these two possibilities, we compare the relative enrichments of biologically meaningful BFs which do not belong to the composed BFs to the relative enrichment of the composed BFs which do not belong to the biologically meaningful BFs. In a nutshell, these tests confirm that the property of being minimally complex in terms of the Boolean complexity or the average sensitivity, i.e., being either an RoF or a NCF, is most likely what drives the enrichment of composition structures.

Data and code availability statement

All the data and codes necessary to reproduce the results in this chapter are available for download from the GitHub repository: <https://github.com/asamallab/CoSt>

Chapter 7

Summary and future outlook

7.1 Summary

This thesis addresses two main objectives. First, to investigate whether regulatory logic rules in reconstructed Boolean models of biological networks are random or not. Second, to leverage relative stability as a constraint for model selection of Boolean models of developmental gene regulatory networks (DGRNs).

To address the first objective, we considered different types of Boolean functions (BFs) in the Boolean modeling literature that are known to be biologically meaningful based on the properties that real regulatory logic are expected to possess. Some of these properties include effectiveness [33], unateness [34] and canalization [22] leading to effective functions (EFs),unate functions (UFs), canalizing functions (CFs) and nested canalizing functions (NCFs) [36, 79]. We also proposed as a potential biologically meaningful type of BF, the read-once functions (RoFs) [84] that have hitherto been unexplored in the context of biological systems. By computationally enumerating the different types of biologically meaningful BFs, we conjectured several interesting properties pertaining to the intersections and overlaps between these types for different number of inputs, and proved them theoretically. We showed that these results could be leveraged in algorithms for both checking and generating different types of BFs. These results are reported in Chapter 2 of the thesis.

Next, in Chapter 3, we shifted our focus to the real regulatory logic rules. From a large corpus of reconstructed Boolean models for living systems spanning several species, we first extracted the regulatory logic rules [45, 140]. We then computed the enrichments and their statistical significance for the various biologically meaningful types in this dataset. Furthermore, we showed that the enrichment of certain types of BFs may be due to the enrichment of their sub-types. In particular, we found that the NCFs and RoFs are the most enriched among all the types of BFs, and NCFs are enriched even within the RoFs [49]. To tackle the question regarding which properties may be responsible for the enrichment of the RoFs and NCFs, we revisited the idea of complexity proposed by Stuart Kauffman from a computer science perspective. More explicitly, we looked at two measures of complexity, namely, the Boolean complexity [69] and average sensitivity [71]. We showed that for a given number of inputs and a given bias, the RoFs attain the theoretical minimum for Boolean complexity, and NCFs attain the theoretical minimum for average sensitivity [49]. This revealed *minimum complexity* as a likely design principle of regulatory logic in biological systems. Lastly we demonstrated the implication of choosing RoFs and NCFs for network dynamics, and show that such a choice renders the dynamics closest to the critical regime when compared to using other types of BFs.

In Chapter 4, we address the second objective by first determining the different measures of relative stability that have been introduced to date - size of the basin of attraction (BOA), mean first passage time (MFPT), steady state probability (SSP), basin transition rate (BTR) and stability index (SIND) [54, 67]. We then used a pancreas differentiation GRN [54] and a root stem cell niche (RSCN) GRN [52] to construct benchmark ensembles that are equally plausible at the level of the biological attractors they recovered and the type of regulatory logic rules they employed. We showed that the relative stability for a given pair of biological attractors in a given benchmark ensemble is strongly correlated for different pairs of measures. Furthermore, we tested that these correlations held across a diverse range of ensembles, thereby supporting our conclusion that the measures of relative stability are strongly correlated. This enabled us to choose any of the 5 measures for further computations of relative stabilities, and we chose the MFPT as it captures best the transition between different cell states. Next, we proposed that the potential cellular

lineage tree associated with a Boolean GRN is the minimum spanning arborescence computed from a complete directed network in which the nodes are cell states and the edges are the MFPTs. Using this method, we computed the distribution of potential lineage trees for the different benchmark ensembles. Interestingly, we found that several Boolean models in the RSCN GRN ensemble could be eliminated based on the fact that the quiescent center (QC) was not the root of the lineage tree associated with those models. Note that the QC is the stem cell in the RSCN and is expected to be the least stable of all the cell types since it differentiates to the surrounding initial cell types [114].

As the exact computation of MFPT [54, 67] for larger Boolean networks (BNs) was infeasible due to the requirement of performing operations on extremely large matrices, we devised a stochastic method to compute the MFPT. Since larger noise values allowed for greater computational efficiency to obtain MFPT using the stochastic method, we tested and confirmed that hierarchies between cell states obtained with MFPT are relatively insensitive to small changes in noise. Furthermore, to find an optimum number of iterations and noise levels that allow for reliable computation of the MFPT using the stochastic method, we computed the relative stabilities between all pairs of attractors for each benchmark dataset using both the stochastic and the exact methods, and showed that they are highly correlated. Using our stochastic method, we computed the hierarchies and lineage trees for three different reconstructed Boolean models for the root development of *Arabidopsis thaliana* published in 2013 [73], 2017 [74] and 2020 [75]. We found that the latest model (2020 model) does not satisfy the expected relative stability criteria, namely, that the QC cell type is not the least stable of the cell states and is not at the root of the lineage tree, indicating that the relative stability was not a criteria considered during model reconstruction. Lastly, we propose an iterative greedy algorithm that takes as input a Boolean model which does not satisfy the expected constraints, the list of allowed BFs at each node (obtained using the known biological fixed points, signs of interactions and type of logic rule) and the biologically expected hierarchies, and outputs a Boolean model that satisfies all the expected hierarchies. Using the 2020 Boolean model as our initial condition to our iterative greedy algorithm, to our surprise, we found 990 improved models in 1000 different simulations. In sum, we developed a systematic model selection

workflow that leverages known constraints on the developmental landscape quantified via the relative stability (in particular, the MFPT) of various pairs of cell states to select from an ensemble of models that are otherwise equally plausible both at the level of the logic rule employed and the biological attractors they recover [68].

Though we have broadly addressed the objectives that were stated at the beginning of this summary, we further explored in Chapters 5 and 6, the biological significance of link operator functions (LOFs) and composition structures respectively, and strengthen some of the conclusions we reached in Chapter 3. In Chapter 5 we showed that of the different consistent types of LOFs [35], namely, AND-NOT, OR-NOT, AND-pairs and OR-pairs, the AND-NOT was most abundant in the reconstructed Boolean models, indicating that the presence of a single inhibitor determines the gene expression, independent of the presence of other activators. Furthermore, we showed that LOFs can act as a very strong constraint on BFs in model selection owing to its very small size in the space of all BFs. Finally, using the static measure of network dynamics, namely, the network average sensitivity, we showed that if the network structure of the reconstructed Boolean models are kept fixed and the BFs are replaced by random LOFs that are consistent with the signs of the regulatory interactions, then, on average, the dynamics is near the critical regime [50]. In Chapter 6, we explore composition structures [43, 51]. Here, our goal was two fold. One, to quantify using empirical datasets, the biological plausibility of composition structures arising from different types of transcriptional regulators (TRs), and two, to quantify the enrichments of composed BFs arising from composition structures in reconstructed Boolean models of GRNs. We first consider the case in which TRs are protein complexes, and estimated the fraction of complexes in which the protein subunits are themselves transcription factors in both humans and yeast. We found that in both organisms the fractions were quite low suggesting that this biological scenario may not be so plausible. However, due to the lack of sufficient data we could not arrive at a firm conclusion. In light of this, we proposed an alternate biological scenario where TRs in composition structures are *cis*-regulatory elements in the DNA such as promoters and enhancers. So, we quantified the fraction of active enhancers that are bound by at least 2 transcription factors (as the restriction on BFs requires at least one element

of the composition structure to be at least 2) and showed that this fraction is greater than 30% in each of the human cell lines, namely, K562 and HepG2. With the above analysis, we addressed our first goal. We also quantified the intersections between the various composition structures and the biologically meaningful types of BFs presented in Chapter 2. Lastly, we computed the enrichment ratios of composed BFs in our dataset of regulatory logic rules obtained from reconstructed BNs and found that though composed BFs are enriched, yet their enrichment can be ascribed to their minimal complexity rather than their property of being *composed* [51].

7.2 Future outlook

Regulatory logic is fundamental to the smooth functioning of cellular processes such as cell growth, cell division and cell differentiation. In Chapters 2, 3, 5 and 6, we explored various types of BFs that are based on biological properties. In Chapter 3, we reached a conclusion that minimum complexity is a likely design principle in regulatory logic rules. However, as a subtlety to that conclusion, it is appropriate to stress that NCFs minimize the average sensitivity for a given number of inputs and a given bias. Therefore an immediate question is to explore whether BFs having certain biases are more enriched than others, and if so, why? Another question along similar lines is whether the enrichment of the NCFs is itself due to an enrichment of some of its sub-types. Answers to both these questions may shed light on other aspects that can distinguish logic rules with minimum complexity, which could further lead to uncovering other design principles. Regulatory logic is shaped by evolution. It may be worthwhile to understand how and why a particular logic rule at a gene arises in the context of evolution. Such a study may be infeasible with the available reconstructed Boolean models and would require the inference of the regulatory rules at homologous genes across species. Stretching this line of thought further, an evolutionary perspective on regulatory logic can enable the use of existing evolution-based approaches such as directed evolution for synthetic design of regulatory logic. Note that the regulatory logic rules that are assigned in reconstructed BNs are manually curated from the literature. Therefore, it is necessary to develop methods that infer regulatory logic directly from gene

expression datasets. This approach can aid in expelling doubts regarding why regulatory logic rules are minimally complex - is it due to the underlying biology or due to subjective influence.

Chapter 4 illustrated a framework that leveraged the hierarchies between cell states on the developmental landscape via the relative stability, as a constraint for model selection. Several natural questions arise regarding the improvement of the proposed framework. The first is to find more computationally scalable measures of relative stability. Such measures will immediately allow model selection in very large BNs. Our framework uses fixed point constraints and the type of regulatory logic to impose restrictions at the level of the logic rule. However, there is a wealth of biological datasets such as transcriptomics, perturbation and mutant datasets that may be used to devise novel constraints at the level of the truth table (or BF). In Chapter 4 we also proposed a method to generate potential cellular lineage trees using the minimum spanning arborescence. Due to only the availability of partial information on the lineage tree in *Arabidopsis thaliana* root development we could not completely explore the implications of our method. So it is imperative that we provide stronger validation of our method for generating potential lineage trees for GRNs that have a well characterized cellular lineage tree. Furthermore, biological situations can arise where the developmental lineage is not a tree because it has loops. Accounting for loops is sometimes necessary in evolutionary phylogenies or in certain developmental processes (one then speaks of convergence of cell fates). It is not clear what should then replace the minimum spanning arborescence. Also, we implemented an iterative greedy algorithm to search and select models that conform to the expected hierarchies on the developmental landscape from a large ensemble of models that are equally biologically plausible. This was achieved by altering at most a single rule in each iteration. Such a method may not be ideal always as it does not efficiently sample the space of biologically plausible Boolean models as the space of Boolean models satisfying all the relative stability constraints amounts to an ever smaller fraction of the whole space (and may sometimes even be empty). Therefore, the development of improved sampling algorithms to explore a large space of Boolean models is warranted. Lastly and most importantly, the relative stability quantifies the propensity to transition from one cell state to another. Therefore, the tools

developed here have (in)direct implications in cellular reprogramming, dedifferentiation and transdifferentiation as well. For instance, we may be able to hypothesize the set of changes to regulatory logic rules that can induce an increased propensity of transition from a differentiated cell state to a pluripotent cell state.

In sum, the central future direction this thesis entails is a combination of data-centric approaches along with computer science based measures to better understand the design principles of regulatory logic in GRNs.

Appendix A

Reference biological datasets of Boolean functions

A.1 Reference biological dataset compiling reconstructed discrete models of living systems used to quantify the preponderance of different biologically meaningful Boolean functions

To assess the abundance of different types of biologically meaningful Boolean functions (BFs) in reconstructed discrete models of living systems, we first compiled a large dataset of 88 models that have been published to date. These 88 models were either downloaded from databases such as Cell Collective [45] (<https://cellcollective.org/>), GINSIM [91] (<http://ginsim.org/>) or BioModels [140] (<http://www.ebi.ac.uk/biomodels/>), or directly obtained from the corresponding published article. Notably, most of these 88 models were downloaded from the Cell collective database [45]. The majority of these models pertain to mammalian systems and a much smaller fraction pertain to plant systems.

The mammalian models include networks for signaling pathways [53,175,176], differentiation [141,177] and various cancers [29,30,178]. Among the plant models, this compilation includes cases from flower organ specification [141], root stem cells [52], and guard cell signaling [179]. Overall, the 88 discrete models in this compilation capture a very diverse collection of biological processes throughout multiple kingdoms of life.

This study [49] is focused only on properties of BFs assigned to different nodes in reconstructed models of biological networks. Some of those networks included nodes taking more than two discrete states; in our compilation, we included only BFs assigned to nodes with binary states which further also had inputs only from other nodes with binary states. While compiling the BFs from these 88 models, we have also gathered the information on the *signs* of interactions between regulators (input nodes) and target gene (output node). Such information is typically obtained from associated experimental literature.

Across the 88 models in this compilation, the number of nodes in a model varies between 4 and 128. From these 88 models, we have compiled 2687 BFs pertaining to 2687 nodes that have number of inputs $k \geq 1$. The BFs assigned to each node in these 88 models are the result of many authors manually identifying appropriate input-output relations during network reconstruction. In other words, the 2687 BFs in the reference biological dataset were chosen during model reconstruction process such as to capture the known regulatory information. This reference dataset of 2687 BFs is available via the GitHub repository: <https://github.com/asamallab/MCBF>.

A.2 Models in the Cell Collective database used for quantifying the preponderance of link operator functions

Table A.1 gives the list of models and their associated PMIDs in the Cell Collective database that were considered for quantifying the preponderance of link operator functions (LOFs) in real reconstructed Boolean networks [50]. This reference dataset of 1741 BFs is available via the GitHub repository: <https://github.com/asamallab/LOF>.

Table A.1: Various models from the Cell Collective database which have been considered for analysis in the study on LOFs. This table reflects the diversity of models which have been considered for analysis in the study on LOFs [50].

PMID	Biological process being modeled
21563979	Lac Operon in <i>E. coli</i>
20862356	Mammalian Cortical Area Development
24970389	Breast Cancer Erbb Skbr3 Cell Line over Short Term
23868318	Drosophila Hh Signalling Pathway
24970389	Breast Cancer Erbb Hcc1954 Cell Line over Long Term
23134720	Oxidative Stress Response
22102804	T Cell Survival Network (Small)
28584084	Lymphoid and Myeloid Cells Transdifferentiation
27148350	pc 12 Cell Differentiation
21968890	IL-1 Signalling
26751566	B Cell Differentiation
22102804	T Cell Survival Network (Large)
24970389	Breast Cancer Erbb Bt474 Cell Line over Long Term
24970389	Breast Cancer Erbb Bt474 Cell Line over Short Term
23868318	Drosophila Spz Signalling Pathway
23868318	Drosophila Toll Signalling Pathway
22267503	Fanconi Anemia/ Breast Cancer Pathway
16873462	Mammalian Cell Cycle
17010384	Neurotransmitter Signalling Pathway
16542429	T Helper Cell Differentiation
22253585	Immune Response against <i>B. bronchiseptica</i> and <i>T. retortaeformis</i>
26408858	Lymphopoiesis Regulatory Network
26385365	Fanconi Anemia Pathway
26090929	Cd4 T Cell Differentiation
28639170	L-Arabinose Operon in <i>E. coli</i>
18463633	Cell Cycle Transcription Network
23868318	Drosophila Wg Signalling Pathway
26616283	Aurka Network of Neuroblastoma
22253585	Immune Response against <i>B. bronchiseptica</i>
22253585	Immune Response against <i>T. Retortaeformis</i>
23868318	Drosophila Fgf Signalling Pathway
19144179	Glucose Repression Signalling Pathways in <i>S. cerevisiae</i>
23868318	Drosophila Vegf Signalling Pathway
29206223	Senescence Associated Secretory Phenotype
24970389	Breast Cancer Erbb Skbr3 Cell Line over Long Term
23056457	Cardiac Gene Regulatory Network II
19118495	Drug Targets in Mammalian Cell Cycle
26340681	<i>A. thaliana</i> Cell Cycle

PMID	Biological process being modeled
19185585	Budding Yeast Cell Cycle I
27594840	Regulatory Hspc-Msc Network.
21968890	IL-6 Signalling
25908096	Iron Acquisition in <i>A. Fumigatus</i>
19025648	Cholesterol Regulatory Pathway
24250280	MAPK Network Influence on Cancer Cell Fate
24970389	Breast Cancer Erbb Hcc1954 Cell Line over Short Term
26573569	Human Gonadal Sex Determination
27542373	Guard Cell Signalling
19422837	Apoptosis Pathway
26446703	Colitis Associated Colon Cancer Network
19662154	Egfr/Erbb Signalling
28361666	Prostate Cancer Signal Transduction Network
20221256	Cell-Fate Decision in Response to Cytokines
16968132	A Dynamic Model of Guard Cell Absciscic Acid Signaling
17722974	T Cell Receptor Signalling
23171249	TOL Network of <i>P. putida</i>
23233838	Yeast Apoptosis Network
22962472	HGF-induced Keratinocyte Migration

Appendix B

Procedure to generate random BFs belonging to various types of BFs

B.1 Randomized generation of biologically mean- ingful BFs used in Chapter 3

Effective functions

Choose a random integer between 0 and 2^{2^k} and convert the integer to its binary vector representation (see Section 2.1.1, Chapter 2) and check if the resulting Boolean function (BF) is effective. If not, repeat the procedure till an effective function (EF) is obtained.

Effective and unate functions

Up to $k = 6$ inputs, all the unate functions (UFs) which are effective can be generated, hence a random choice from this list returns an effective and unate function (EUF). If $k > 6$, a random partition of k is generated such that each element of the partition is a number less than or equal to 6. In other words, $k = k_1 + k_2 + k_3 + \dots$ such that $k_i \leq 6$.

A random EUF with k_1 variables is generated and combined with a random EUF with k_2 variables by either an AND or OR logic function. This is repeated till all elements of the partition are covered. For example, if $k = 10$, then an acceptable partition is $(2, 5, 3)$ and the EUF which is generated is $(EUF(2) \odot EUF(5)) \odot EUF(3)$ where \odot is AND or OR (which is also chosen randomly for each occurrence). Since generating the UFs with greater than 6 inputs is computationally expensive, we resort to the heuristic algorithm provided above. The functions obtained using this heuristic may not give a uniform distribution over all EUFs.

Canalyzing functions

We implement the algorithm provided in the software BoolNet [117] to generate random CFs. Generate a random integer between 0 and 2^{2^k} and convert it to a binary vector. This is a random BF. If the BF is not canalyzing, choose one of the k inputs randomly and also choose a random canalyzing input value (0 or 1). Set the outputs corresponding those 2^{k-2} entries of the binary vector to 0 or 1 (also chosen randomly). Thus the generated function is guaranteed to be canalyzing in at least one input.

Effective and canalyzing functions

Generate a CF based on the procedure given above and check if the resulting BF is effective. If not, repeat the procedure till an EF is obtained. We abbreviate such effective and canalyzing functions as ECFs.

Nested canalyzing functions

We leverage the fact that for each bias, there is exactly one nested canalyzing function (NCF) upto isomorphisms. Hence, given k , randomly choose an odd bias between 1 and 2^{k-1} , say P . For bias P , the NCF is generated by setting the first P bits of the output binary vector to 1 and the remaining $2^k - P$ bits to 0. Note that since the average sensitivity is invariant under change of signs in the inputs, one can simply calculate the average sensitivity of any of the isomorphic forms of a NCF with bias P .

Read-once functions

Since all the representative read-once functions (RoFs) can be generated for $k \leq 10$, a RoF can be chosen randomly from such a list of representative RoFs. In case $k > 10$, we partition k into two parts such that $k = k_1 + k_2$, where $k_1 \leq 10$ and $k_2 \leq 10$. A randomly chosen k_1 input RoF is then combined with a randomly chosen k_2 input RoF by either an AND or OR operator which is also chosen randomly.

non-NCF read-once functions

For $k \geq 4$, generate a random RoF and check if it is a NCF. If so, repeat the procedure till a non-NCF RoF is generated. In case a node has less than 4 inputs, random NCFs are assigned to them as there are no non-NCF RoF for $k < 4$.

B.2 Randomized generation of biologically meaningful BFs used in Chapter 5

We obtain the distributions of sensitivities for a fixed network structure when assigning candidate nodes of the network, BFs belonging to the category of interest, namely: EFs, EUFs and link operator functions (LOFs) such as AND-NOT, OR-NOT, AND-pairs and OR-pairs. Here, candidate nodes are the nodes that qualify as having at least one activator and one inhibitor as inputs, with the other ones being assigned the biological function. In case of assigning a certain type of LOF to a node, all possible functions would have the same average sensitivity, hence there is only one possibility in terms of the average sensitivity of the network. This is not the case when choosing EFs or EUFs. Hence, it was necessary to devise a randomized procedure to generate such functions. When generating EUFs for $k = 7, 8$ and 9 inputs, our procedure is based on a greedy heuristic and so may not be truly uniform.

Effective functions

For generating a k -input EF, we choose a random integer between 0 and 2^{2^k} , and then accept the obtained BF if it is an EF, otherwise we reject it and repeat the operation until a function is accepted. Note that the output column of the truth table of a BF can be considered as a string of bits of length 2^k and thus be encoded as an integer ranging from 0 to 2^{2^k} .

Effective and unate functions

Firstly, we observe that in the context of obtaining the average sensitivity, the signs of the inputs to the node under consideration are immaterial. Hence generating an EUF with all activators, as opposed to some particular sign combination, suffices for our purpose. The reasons are two-fold:

- (i) For a EUF (at a given k) where all inputs are, say activators, one can exchange 0s and 1s in columns of the truth table to obtain a UF with activators and inhibitors. Thus if we know all the EUFs with all inputs being activators, we can generate EUFs with all other combinations of input signs by such operations.
- (ii) The average sensitivity of an EUF is unchanged if an activator input is inverted (by a column operation on the truth table) to an inhibitor and vice versa.

For nodes with up to 6 inputs, we can generate all the UFs, hence that set can be sampled uniformly without any bias. If the UF function obtained is not an EF, we reject it and keep sampling until we get an EF. The resulting EUF is then assigned to the node under consideration. For nodes with greater than 6 inputs, say 7 inputs, two uniformly sampled 6-input UFs are chosen with one of them assigned to the input value “0” of the newly added 7th variable, and the other, assigned to the input value “1” of the 7th variable. We then check if the resulting function is a UF with all signs as activators. If so, we store it in a list. We repeat this to obtain a large number of BFs with 7 inputs. We then uniformly sample from this list of obtained functions and assign it to a node if the function is an EF. Similarly, to obtain UFs with 8 inputs, we uniformly sample from the list of 7 input

functions which we generated previously and repeat the previous procedure: assign one of the UFs to the input value “0” of the 8th input variable, the other to the input value “1” of the same variable and check if the resulting function is a UF with all inputs as activators. If so, we store the UF in a list. Finally, we sample uniformly from this list and assign a function to a node if the function is an EF.

Appendix C

Additional Figures and Tables for Chapter 3

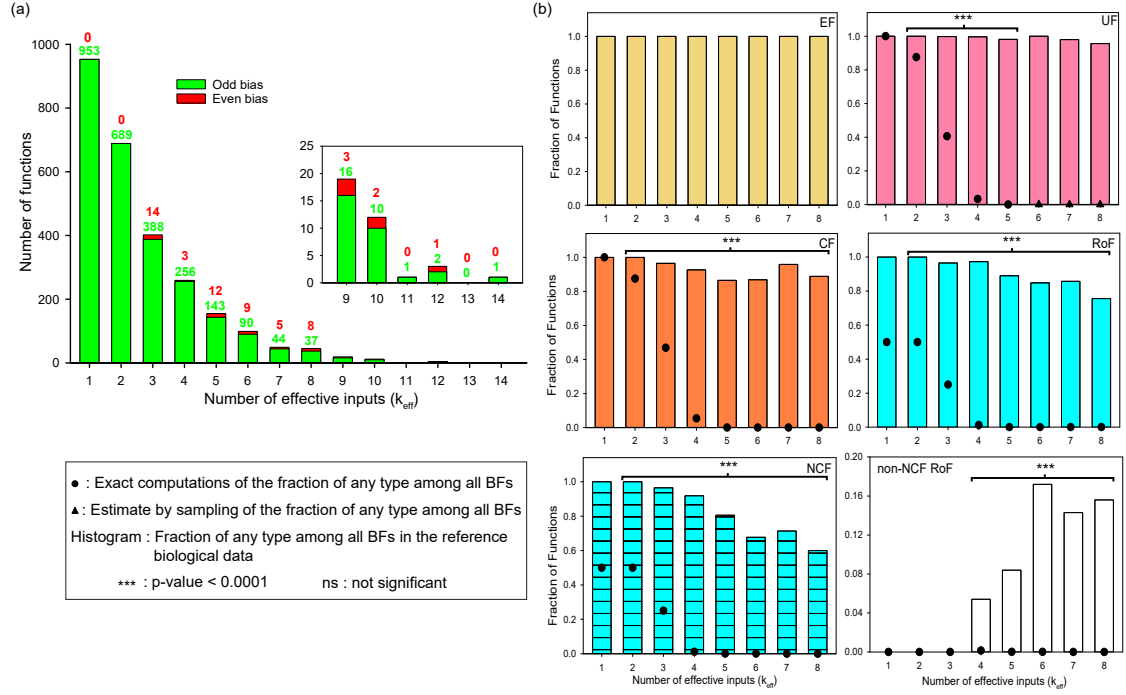


Figure C.1: Overlap of different types of BFs and their distribution in the modified reference biological dataset. (a) The in-degree distribution for nodes in the modified reference biological dataset after discarding the ineffective inputs in ineffective functions. Here, k_{eff} is the number of effective inputs after stripping a BF of its ineffective inputs. (b) The plots show the abundance and statistical significance of the biologically meaningful BFs for $k_{eff} \leq 8$ in the modified dataset. The dot symbols which appear to coincide with the x -axis are very small non-zero numbers (except for non-NCF RoFs with $k = 1, 2, 3$). We do not show the p -values of the EF case since checking its statistical significance is not meaningful as all BFs are forced to be effective in the modified dataset. The raw data associated with these plots along with results from the statistical test for over-representation are included as Tables C.1 - C.4 in Appendix C.

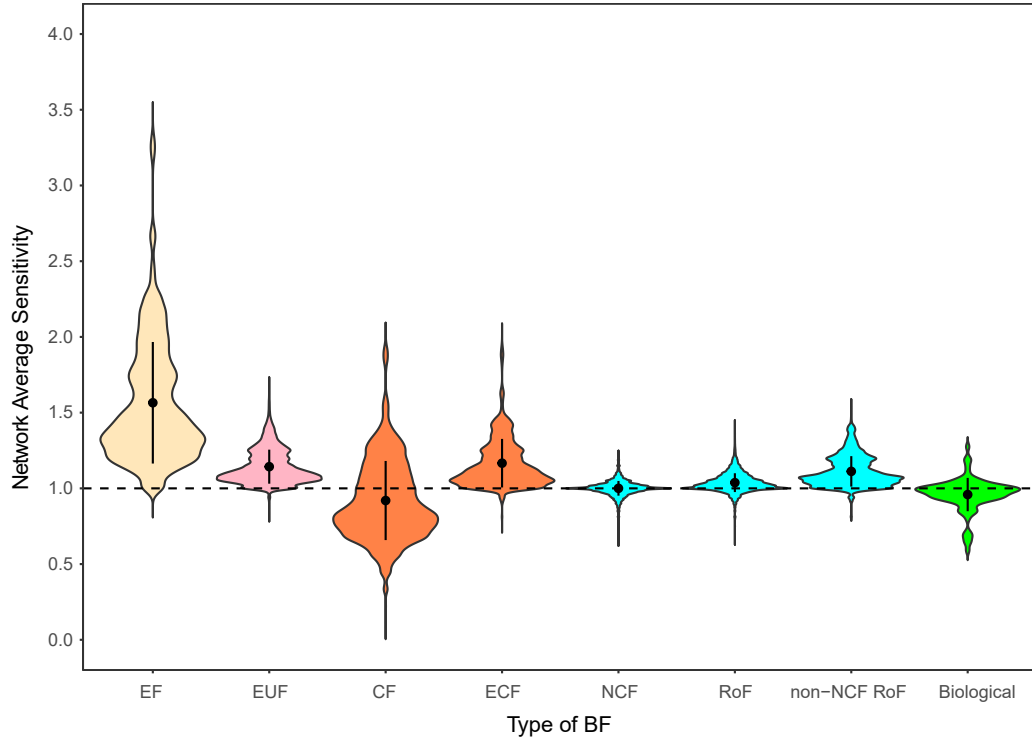


Figure C.2: Distribution of the network average sensitivity of 88 models after discarding the ineffective inputs. Distribution of the network average sensitivity when using the list of (effective) inputs from biological models but enforcing different types of BFs to the nodes, namely effective functions (EF), effective and unate functions (EUF), canalyzing functions (CF), effective and canalyzing functions (ECF), nested canalyzing functions (NCF), read-once functions (RoF) and non-NCF RoFs. For this computation, we start with the modified reference biological dataset wherein all ineffective inputs to nodes are discarded in each of the 88 networks or models. The right-most case is the distribution when using the actual BFs in the biological models. This plot has been generated by keeping the maximum width of each of the violins fixed.

Table C.1: Number of different types of biologically meaningful BFs in the modified reference biological dataset. Here, k_{eff} is the number of effective inputs after stripping a BF of its ineffective inputs. “All” is the total number of BFs for a given number of effective inputs, EF corresponds to effective functions, UF to unate functions (all sign combinations), CF to canalizing functions, EUF to effective and unate functions, ECF to effective and canalizing functions, UCF to unate and canalizing functions, EUCF to effective, unate and canalizing functions, NCF to nested canalizing functions and RoF to read-once functions.

k_{eff}	Types of BFs									
	All	EF	UF	CF	EUF	ECF	UCF	EUCF	NCF	RoF
1	953	953	953	953	953	953	953	953	953	953
2	689	689	689	689	689	689	689	689	689	689
3	402	402	401	388	401	388	388	388	388	388
4	259	259	258	240	258	240	240	240	238	252
5	155	155	152	134	152	134	133	133	125	138
6	99	99	99	86	99	86	86	86	67	84
7	49	49	48	47	48	47	46	46	35	42
8	45	45	43	40	43	40	38	38	27	34
9	19	19	18	17	18	17	17	17	7	16
10	12	12	12	10	12	10	10	10	3	6
11	1	1	1	1	1	1	1	1	0	0
12	3	3	3	3	3	3	3	3	2	2
14	1	1	1	1	1	1	1	1	1	1

Table C.2: Fraction of different types of biologically meaningful BF_s in the modified reference biological dataset. Here, k_{eff} is the number of effective inputs after stripping a BF of its ineffective inputs. EF corresponds to effective functions, UF to unate functions (all sign combinations), CF to canalizing functions, EUF to effective and unate functions, ECF to effective and canalizing functions, UCF to unate and canalizing functions, EUCF to effective, unate and canalizing functions, NCF to nested canalizing functions and RoF to read-once functions.

k_{eff}	Types of BF _s								
	EF	UF	CF	EUF	ECF	UCF	EUCF	NCF	RoF
1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	1	1	1	1	1
3	1	0.998	0.965	0.998	0.965	0.965	0.965	0.965	0.965
4	1	0.996	0.927	0.996	0.927	0.927	0.927	0.919	0.973
5	1	0.981	0.865	0.981	0.865	0.858	0.858	0.806	0.890
6	1	1	0.869	1	0.869	0.869	0.869	0.677	0.848
7	1	0.980	0.959	0.980	0.959	0.939	0.939	0.714	0.857
8	1	0.956	0.889	0.956	0.889	0.844	0.844	0.600	0.756
9	1	0.947	0.895	0.947	0.895	0.895	0.895	0.368	0.842
10	1	1	0.833	1	0.833	0.833	0.833	0.25	0.5
11	1	1	1	1	1	1	1	0	0
12	1	1	1	1	1	1	1	0.667	0.667
14	1	1	1	1	1	1	1	1	1

Table C.3: p -value tests for statistical enrichments of the different types of BFs in the modified reference biological dataset. Here, k_{eff} is the number of effective inputs after stripping a BF of its ineffective inputs. A low p -value indicates that the corresponding type of BF is enriched in the modified reference biological dataset when compared to the ensemble of all BFs. For $k_{eff} > 2$ when the p -value shown is 0, it was smaller than what we could measure. Here, UF corresponds to unate functions (all sign combinations), CF to canalyzing functions, NCF to nested canalyzing functions and RoF to read-once functions.

k_{eff}	Odd bias	UF	CF	NCF	RoF
2	0	0	0	0	0
3	9.46×10^{-98}	5.43×10^{-158}	2.59×10^{-108}	1.43×10^{-212}	1.43×10^{-212}
4	3.63×10^{-74}	0	5.10×10^{-280}	0	0
5	5.12×10^{-31}	0	0	0	0
6	2.95×10^{-19}	0	0	0	0
7	4.11×10^{-10}	0	0	0	0
8	1.56×10^{-6}	0	0	0	0

Table C.4: Enrichment of the different types of BFs in the modified reference biological dataset. Fractions of functions that are RoFs, non-NCF RoFs or NCFs, in the space of all $2^{k_{eff}}$ BFs (f_0) or in the modified reference biological dataset (f_1). E ($= f_1/f_0$) is the enrichment ratio; it indicates the extent of the over-representation of such functions in the modified reference dataset. Over-representation is highest for NCFs but clearly non-NCF RoFs are also highly over-represented. Computations are reported for functions with $k_{eff} \leq 8$ inputs.

k_{eff}	RoF			non-NCF RoF			NCF		
	f_0	f_1	E	f_0	f_1	E	f_0	f_1	E
1	0.5	1.0	2.0	0	0	-	0.5	1.0	2.0
2	0.5	1.0	2.0	0.0	0.0	-	0.5	1.0	2.0
3	0.25	1.0	4.0	0.0	0.0	-	0.25	1.0	4.0
4	0.0127	0.965	76.012	0.001	0.0	0.0	0.011	0.965	85.927
5	3.517×10^{-06}	0.973	2.77×10^5	1.04×10^{-06}	0.054	5.18×10^4	2.47×10^{-06}	0.919	3.71×10^5
6	1.909×10^{-14}	0.89	4.66×10^{13}	9.12×10^{-15}	0.084	9.21×10^{12}	9.97×10^{-15}	0.806	8.08×10^{13}
7	2.950×10^{-32}	0.848	2.87×10^{31}	1.86×10^{-32}	0.172	9.26×10^{30}	1.092×10^{-32}	0.677	6.20×10^{31}
8	2.918×10^{-69}	0.857	2.94×10^{68}	2.18×10^{-69}	0.143	6.57×10^{67}	7.404×10^{-70}	0.714	9.64×10^{68}

Table C.5: Relative enrichment of the different types of BF in the modified reference biological dataset. The relative enrichment ratios E_R for the RoFs and NCFs in the ensemble of odd bias BFs, EFs and UFs in the modified dataset. Here, k_{eff} is the number of inputs after stripping a BF of its ineffective inputs. These enrichment ratios indicate the extent of the over-representation of such functions in the modified reference biological dataset. $E_R > 1$ suggests that there is indeed an enrichment of RoFs and NCFs within the odd bias BFs, EFs and UFs in the modified reference biological dataset when compared to that expected in the ensemble of all odd bias BFs, EFs and UFs.

k_{eff}	E_R for RoF in:			E_R for NCF in:		
	Odd bias	EF	UF	Odd bias	EF	UF
1	1.0	1.0	2	1.0	1.0	2
2	1.0	1.25	1.75	1.0	1.25	1.75
3	2.0	3.40625	1.625	2.0	3.40625	1.625
4	39.385	74.920	2.517	44.522	84.692	2.845
5	1.40×10^5	2.77×10^5	14.851	1.88×10^5	3.71×10^5	19.94

Table C.6: Statistical test for relative enrichment of NCFs in CFs and RoFs in the modified reference biological dataset. The relative enrichment ratio E_R of the NCFs in the CFs and RoFs in the modified dataset. $f_{s,0}/f_0$ denotes the fractions of functions that are NCFs in the space of all CFs or RoFs and $f_{s,1}/f_1$, the equivalent fraction in the modified reference biological dataset. Here, k_{eff} is the number of inputs after stripping a BF of its ineffective inputs. $E_R = (f_{s,1}/f_1)/(f_{s,0}/f_0)$ denotes the enrichment ratio and it indicates the extent of the over-representation of such functions in the modified reference dataset. Computations are reported for BFs with $k_{eff} \leq 8$ inputs. The low p -values indicate that there is an enrichment of NCFs within the CFs and RoFs in the modified reference biological dataset when compared to that expected in the ensemble of all CFs and RoFs.

k_{eff}	NCF in CF				NCF in RoF			
	$f_{s,0}/f_0$	$f_{s,1}/f_1$	E_R	p - value	$f_{s,0}/f_0$	$f_{s,1}/f_1$	E_R	p - value
2	0.571	1.0	1.75	0	1.0	1.0	1.0	0
3	0.533	1.0	1.875	0	1.0	1.0	1.0	0
4	0.209	1.0	4.774	1.04×10^{-160}	0.885	1.0	1.130	3.87×10^{-4}
5	0.008	0.991	120.588	3.73×10^{-251}	0.703	0.944	1.343	2.02×10^{-9}
6	1.78×10^{-6}	0.932	5.22×10^5	0	0.522	0.906	1.734	4.06×10^{-8}
7	7.19×10^{-15}	0.779	1.08×10^{14}	0	0.370	0.798	2.157	1.04×10^{-10}
8	7.87×10^{-33}	0.744	9.45×10^{31}	0	0.254	0.833	3.283	5.26×10^{-12}

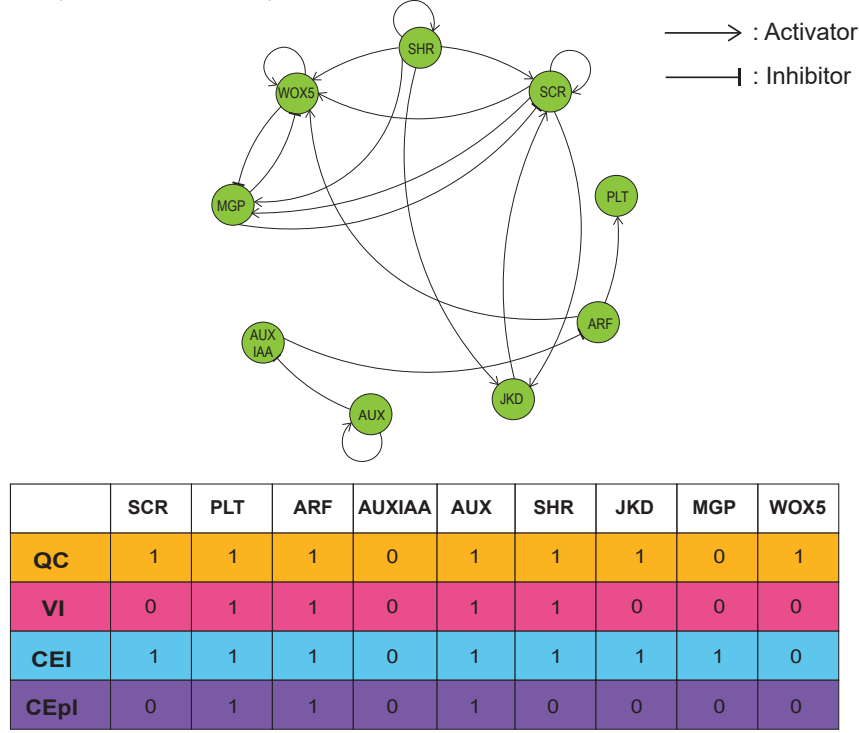
Table C.7: Quantifying the fraction of models in different ensembles with network average sensitivities (s) lying outside the distribution of s for biological networks (without ineffective inputs). The percentage of data points that fall outside the 95% confidence interval of the modified biological dataset in the distribution of network average sensitivities when using the list of inputs from biological models but enforcing different types of BFs to the nodes, namely effective functions (EF), effective and unate functions (EUF), canalyzing functions (CF), effective and canalyzing functions (ECF), nested canalyzing functions (NCF), read-once functions (RoF) and non-NCF RoFs, The distribution of network average sensitivities is shown in Figure C.2, Appendix C and data for both one-sided tests and two-sided tests are provided here. From the data for the two-sided test, we can arrange various BFs based on their increasing proximity to the biological distribution in the following manner: $EF < ECF < EUF < CF < \text{non-NCF RoF} < \text{RoF} < \text{NCF}$.

Type of BF	One-sided (upper 5%)	One-sided (lower 5%)	Two-sided (2.5% on either side)
EF	91.85	0.0	86.26
EUF	38.68	0.0	29.84
CF	19.7	17.31	26.72
ECF	42.5	0.0	34.51
NCF	0.75	0.01	0.05
RoF	5.53	0.0	2.06
non-NCF RoF	31.88	0.0	22.05

Appendix D

Additional Figures and Tables for Chapter 4

(a) Gene regulatory network and attractors of *Arabidopsis thaliana* Root Stem Cell Niche (RSCN) (Azpeitia *et al.* 2010)



(b) Gene regulatory network and attractors of Pancreas cell differentiation (Zhou *et al.* 2016)

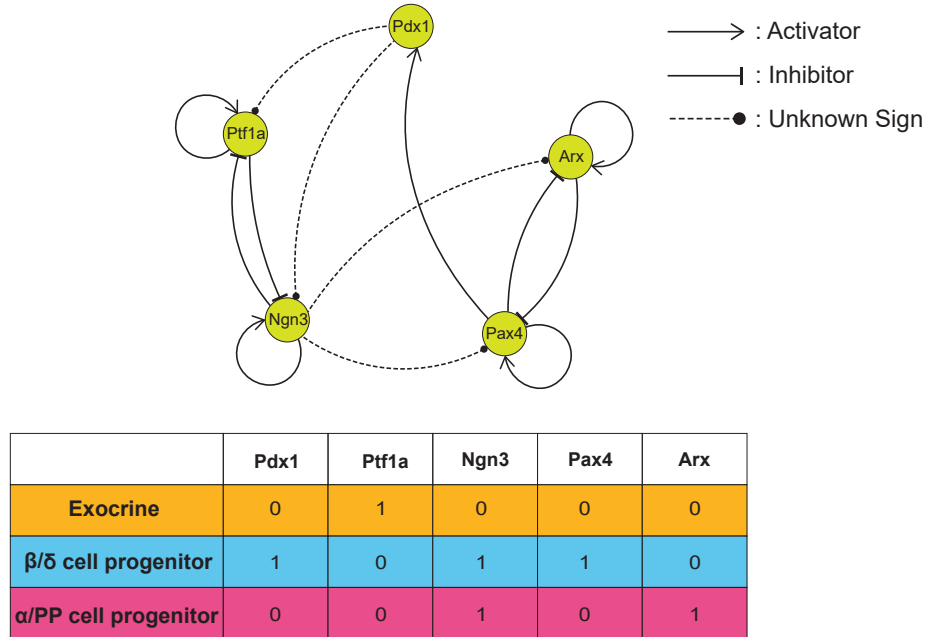


Figure D.1: Biological networks used to generate ensembles of Boolean models. (a) *Arabidopsis thaliana* RSCN Boolean gene regulatory network (GRN) and its attractors. The network is constructed using regulatory interactions obtained from the Boolean functions (BFs) of *model A* in [52]. Here, QC: Quiescent center, VI: Vascular initials, CEI: Cortex-Endodermis initials, CEpI: Columella epidermis initials (CEpI) (b) Pancreas cell differentiation model Boolean GRN and its attractors.

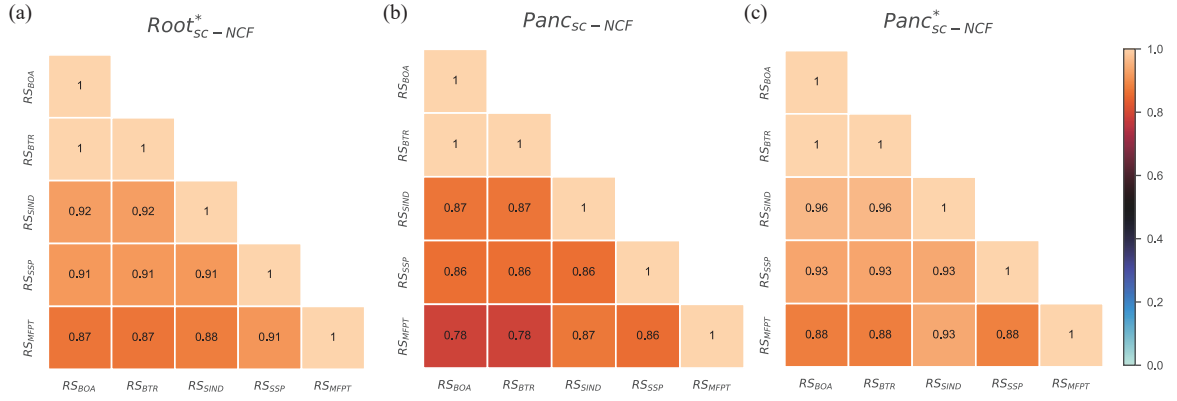


Figure D.2: Pearson correlation between different pairs of relative stability measures for the ensembles $Root_{sc-NCF}^*$, $Panc_{sc-NCF}$ and $Panc_{sc-NCF}^*$. The rows and columns correspond to choices for the relative stability measures. The heatmap indicates the value of the Pearson correlation coefficient between pairs of these measures. These 5 measures are based on size of basin of attraction (RS_{BOA}), basin transition rates (RS_{BTR}), a stability index (RS_{IND}), steady state probabilities (RS_{SSP}) and mean first passage times (RS_{MFPT}). Note that these measures are computed by exact means across all pairs of biological fixed points, for all 170, 3600, 109 models in the ensembles $Root_{sc-NCF}^*$, $Panc_{sc-NCF}$ and $Panc_{sc-NCF}^*$ respectively, using a noise intensity parameter value of 1%. The upper triangular portion of the heatmap is not displayed as the heatmap entries constitute a symmetric matrix. Furthermore, RS_{BOA} and RS_{BTR} are perfectly correlated, an observation which we prove theoretically in Section 4.3.1 by showing that RS_{BOA} and RS_{BTR} are in fact equivalent.

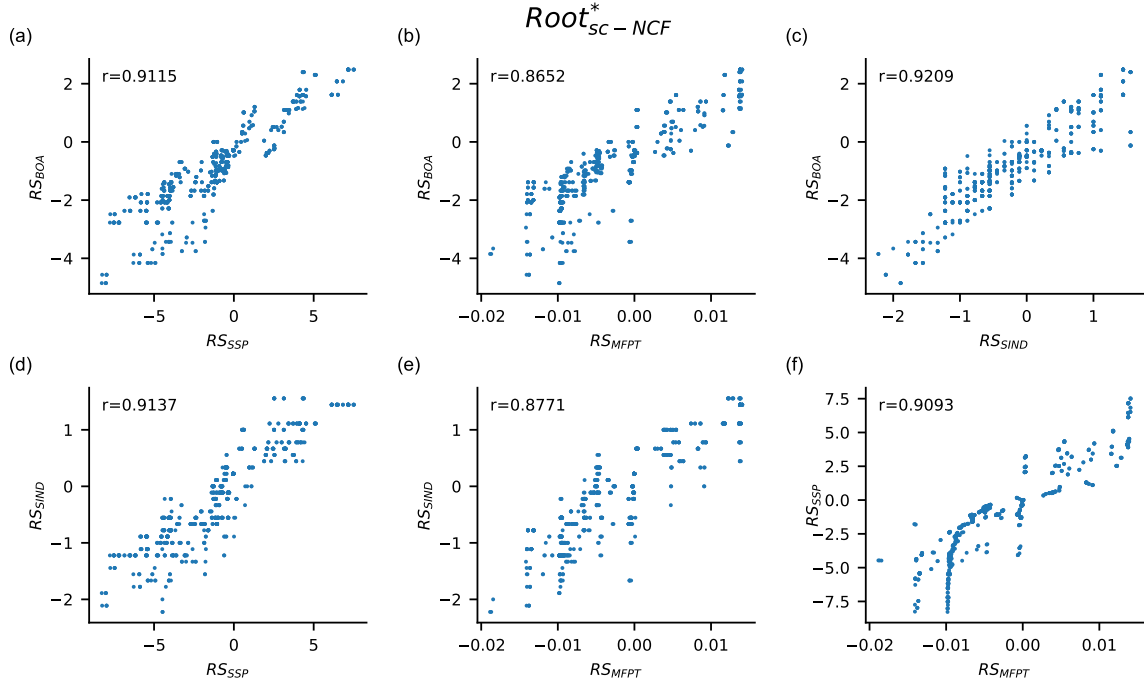


Figure D.3: Scatter plots displaying values of relative stability in the ensemble $Root_{sc-NCF}^*$. Each sub-figure from (a) to (f) is a scatter plot where the x and y axes are for different measures of relative stability. These 5 measures are based on size of basin of attraction (RS_{BOA}), basin transition rates (RS_{BTR}), a stability index (RS_{SIND}), steady state probabilities (RS_{SSP}) and mean first passage times (RS_{MFPT}). These measures have been computed by the exact method for all pair of biological fixed points, for all 170 models belonging to the ensemble $Root_{sc-NCF}^*$, at 1% noise. Of the 10 possible scatter plots for distinct pairs of the 5 relative stability measures, only 6 are shown here as RS_{BOA} and RS_{BTR} are equivalent. The Pearson correlation coefficient (r) for each scatter plot is computed and reported in the plot. These plots indicate that the correlation between the different relative stability measures is quite strong.

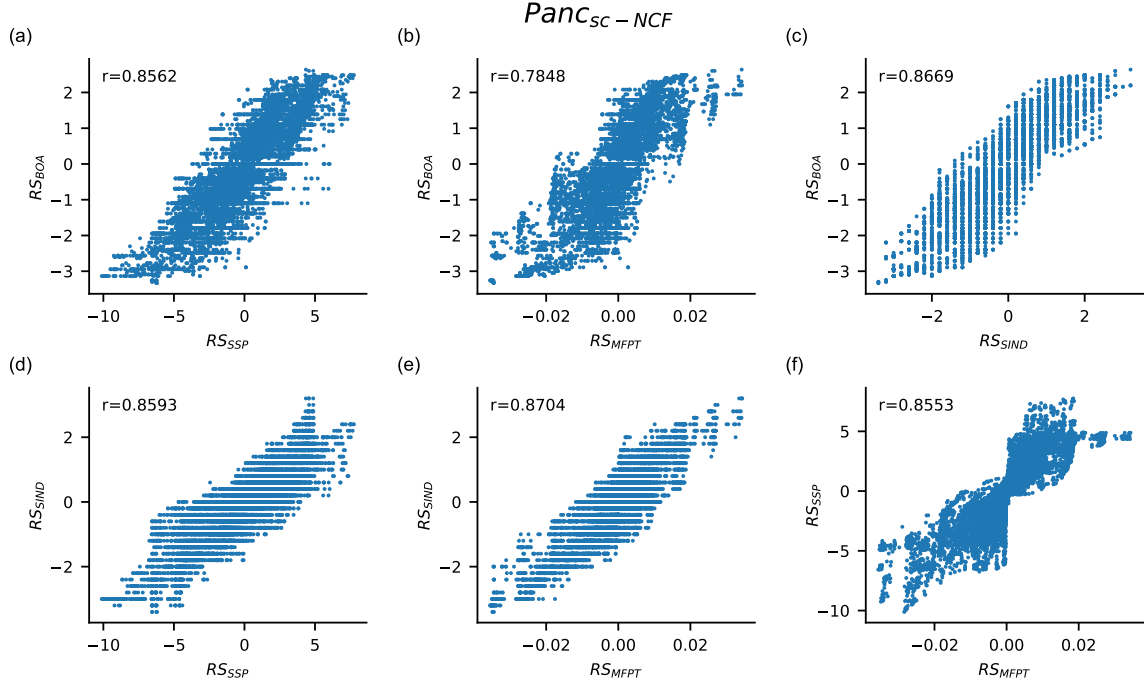


Figure D.4: Scatter plots displaying values of relative stability in the ensemble *Panc_{sc}-NCF*. Each sub-figure from (a) to (f) is a scatter plot where the x and y axes are for different measures of relative stability. These 5 measures are based on size of basin of attraction (RS_{BOA}), basin transition rates (RS_{BTR}), a stability index (RS_{SIND}), steady state probabilities (RS_{SSP}) and mean first passage times (RS_{MFPT}). These measures have been computed by the exact method for all pair of biological fixed points, for all 3600 models belonging to the ensemble *Panc_{sc}-NCF*, at 1% noise. Of the 10 possible scatter plots for distinct pairs of the 5 relative stability measures, only 6 are shown here as RS_{BOA} and RS_{BTR} are equivalent. The Pearson correlation coefficient (r) for each scatter plot is computed and reported in the plot. These plots indicate that the correlation between the different relative stability measures is quite strong.

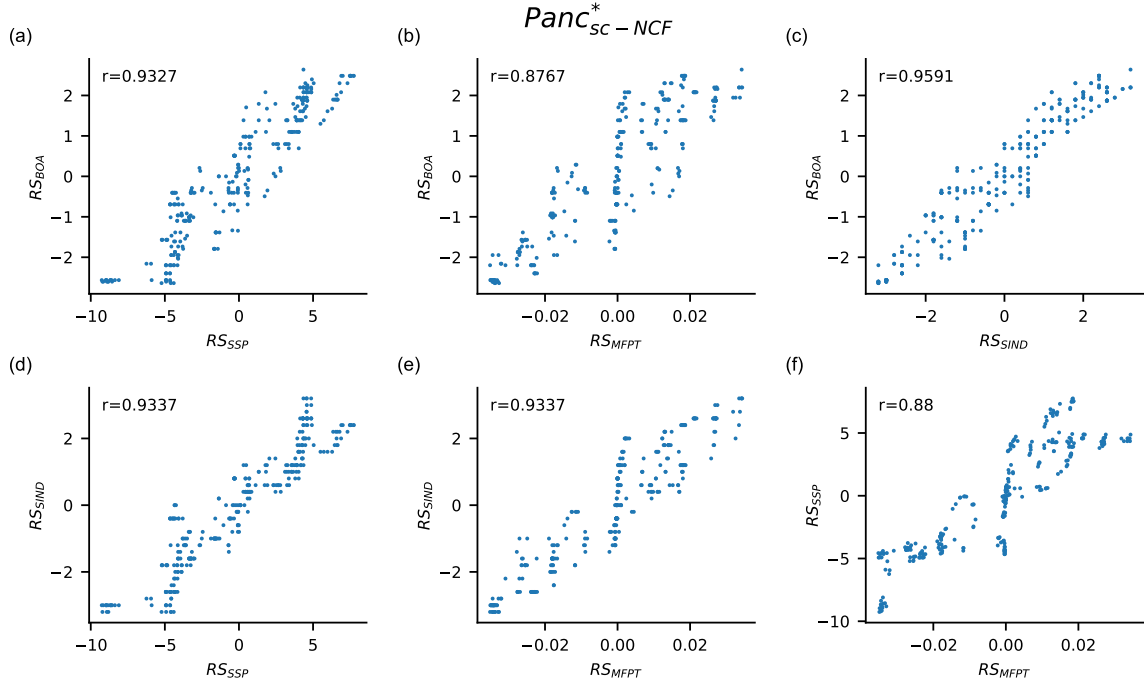


Figure D.5: Scatter plots displaying values of relative stability in the ensemble $Panc_{sc-NCF}^*$. Each sub-figure from (a) to (f) is a scatter plot where the x and y axes are for different measures of relative stability. These 5 measures are based on size of basin of attraction (RS_{BOA}), basin transition rates (RS_{BTR}), a stability index (RS_{SIND}), steady state probabilities (RS_{SSP}) and mean first passage times (RS_{MFPT}). These measures have been computed by the exact method for all pair of biological fixed points, for all 109 models belonging to the ensemble $Panc_{sc-NCF}^*$, at 1% noise. Of the 10 possible scatter plots for distinct pairs of the 5 relative stability measures, only 6 are shown here as RS_{BOA} and RS_{BTR} are equivalent. The Pearson correlation coefficient (r) for each scatter plot is computed and reported in the plot. These plots indicate that the correlation between the different relative stability measures is quite strong.

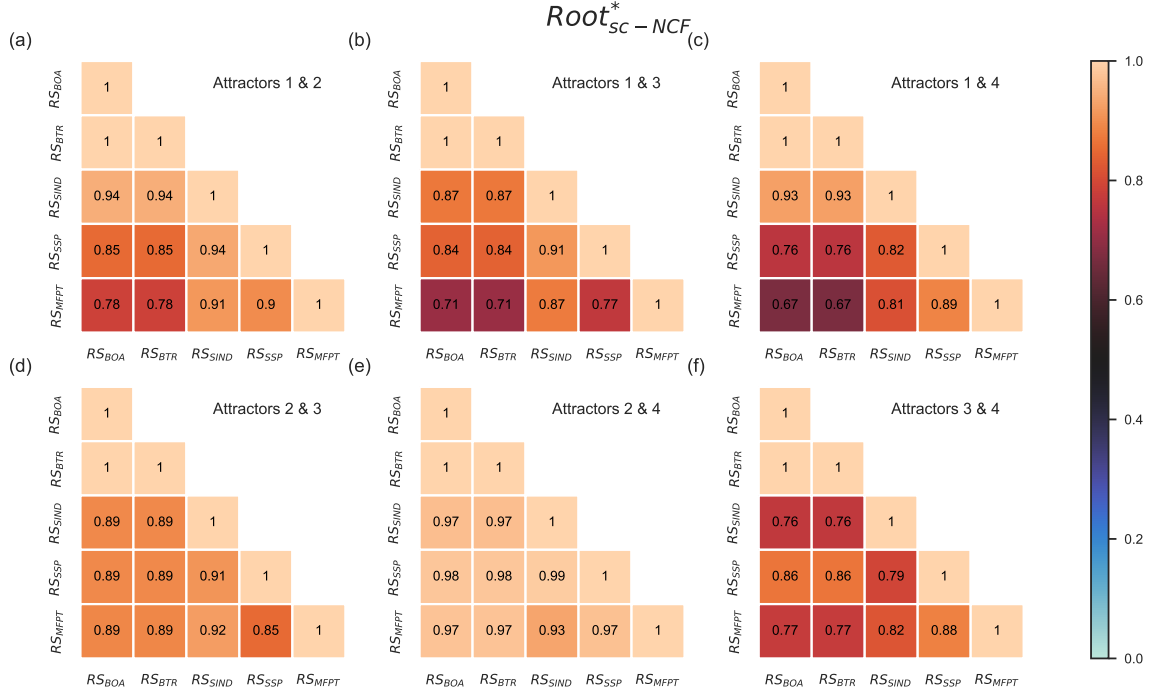


Figure D.6: Pearson correlation between different pairs of relative stability measures for a given pair of fixed points for the ensemble $Root_{sc-NCF}^*$. The rows and columns of all heatmaps correspond to choices for the relative stability measures. The heatmaps indicate the Pearson correlation coefficient between pairs of these measures. These 5 measures are based on size of basin of attraction (RS_{BOA}), basin transition rates (RS_{BTR}), a stability index (RS_{SIND}), steady state probabilities (RS_{SSP}) and mean first passage times (RS_{MFPT}). For a particular sub-figure, these measures are computed by exact means for the pair of biological fixed points specified in that sub-figure, for all 170 models in the ensemble $Root_{sc-NCF}^*$ using a noise intensity parameter value of 1%. Each biological attractor (fixed point) is numbered as follows. 1: Quiescent center (QC), 2: Vascular initials (VI), 3: Cortex-Endodermis initials (CEI), 4: Columella epidermis initials (CEpI). The upper triangular portion of the heatmap is not displayed as the heatmap entries constitute a symmetric matrix. Furthermore, RS_{BOA} and RS_{BTR} are perfectly correlated, an observation which we prove theoretically in Section 4.3.1 by showing that RS_{BOA} and RS_{BTR} are in fact equivalent.

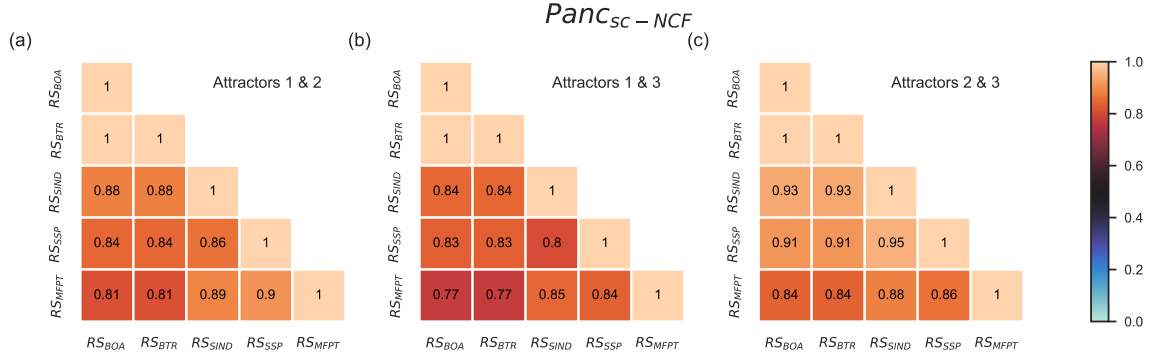


Figure D.7: Pearson correlation between different pairs of relative stability measures for a given pair of fixed points for the ensemble $Panc_{sc-NCF}$. The rows and columns of all heatmaps correspond to choices for the relative stability measures. The heatmaps indicate the Pearson correlation coefficient between pairs of these measures. These 5 measures are based on size of basin of attraction (RS_{BOA}), basin transition rates (RS_{BTR}), a stability index (RS_{SIND}), steady state probabilities (RS_{SSP}) and mean first passage times (RS_{MFPT}). For a particular sub-figure, these measures are computed by exact means for the pair of biological fixed points specified in that sub-figure, for all 3600 models in the ensemble $Panc_{sc-NCF}$ using a noise intensity parameter value of 1%. Each biological attractor (fixed point) is numbered as follows. 1: Exocrine, 2: β/δ progenitor, 3: α/PP progenitor. The upper triangular portion of the heatmap is not displayed as the heatmap entries constitute a symmetric matrix. Furthermore, RS_{BOA} and RS_{BTR} are perfectly correlated, an observation which we prove theoretically in Section 4.3.1 by showing that RS_{BOA} and RS_{BTR} are in fact equivalent.

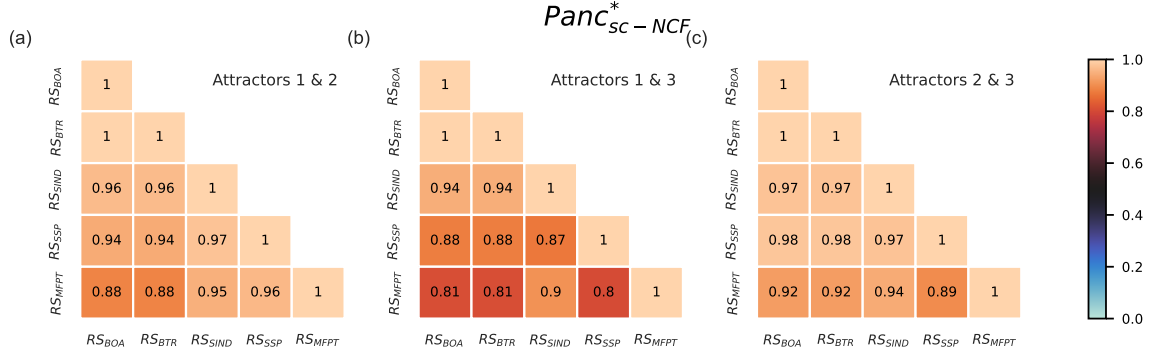


Figure D.8: Pearson correlation between different pairs of relative stability measures for a given pair of fixed points for the ensemble $Panc_{sc-NCF}^*$. The rows and columns of all heatmaps correspond to choices for the relative stability measures. The heatmaps indicate the Pearson correlation coefficient between pairs of these measures. These 5 measures are based on size of basin of attraction (RS_{BOA}), basin transition rates (RS_{BTR}), a stability index (RS_{SIND}), steady state probabilities (RS_{SSP}) and mean first passage times (RS_{MFPT}). For a particular sub-figure, these measures are computed by exact means for the pair of biological fixed points specified in that sub-figure, for all 109 models in the ensemble $Panc_{sc-NCF}^*$ using a noise intensity parameter value of 1%. Each biological attractor (fixed point) is numbered as follows. 1: Exocrine, 2: β/δ progenitor, 3: α/PP progenitor. The upper triangular portion of the heatmap is not displayed as the heatmap entries constitute a symmetric matrix. Furthermore, RS_{BOA} and RS_{BTR} are perfectly correlated, an observation which we prove theoretically in Section 4.3.1 by showing that RS_{BOA} and RS_{BTR} are in fact equivalent.

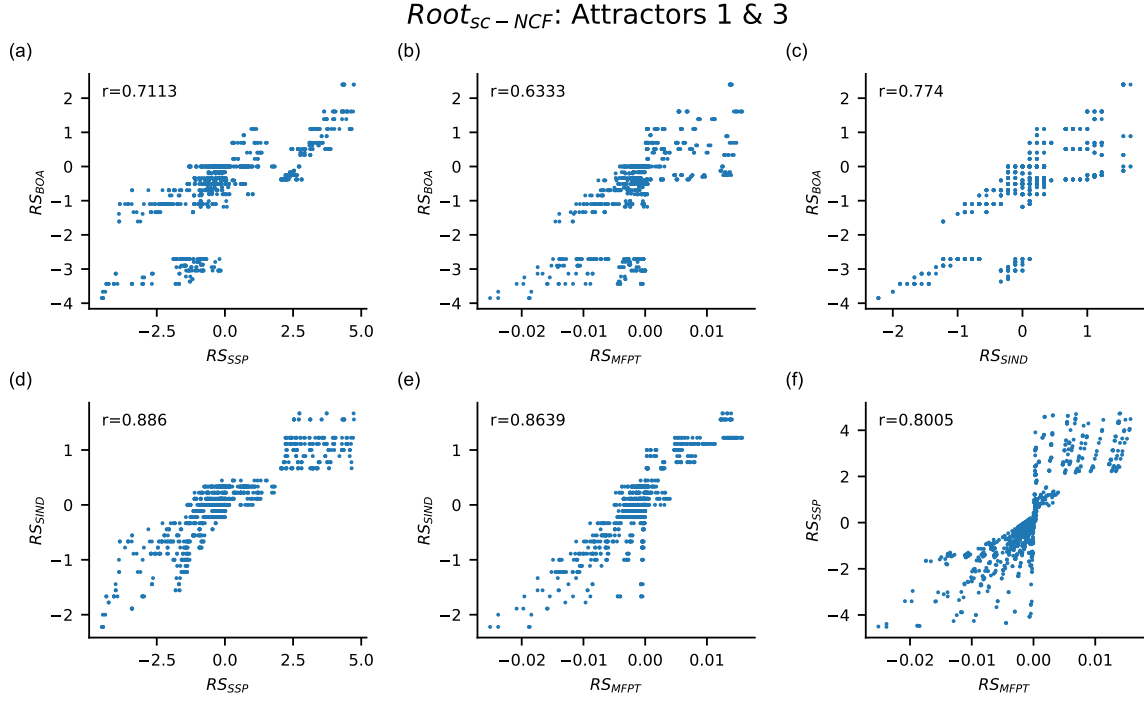


Figure D.9: Scatter plots between the different pairs of relative stability measures for the pair of attractors 1 and 3 for the ensemble $Root_{sc-NCF}$. Each sub-figure from (a) to (f) is a scatter plot where the x and y axes are for different measures of relative stability. These 5 measures are based on size of basin of attraction (RS_{BOA}), basin transition rates (RS_{BTR}), a stability index (RS_{SIND}), steady state probabilities (RS_{SSP}) and mean first passage times (RS_{MEFT}). All these measures have been computed by the exact method for the pair of biological fixed points 1 (Quiescent center (QC)) and 3 (Cortex-Endodermis initials (CEI)), for all 1275 models belonging to the ensemble $Root_{sc-NCF}$, at 1% noise. Of the 10 possible scatter plots for distinct pairs of the 5 relative stability measures, only 6 are shown here as RS_{BOA} and RS_{BTR} are equivalent. The Pearson correlation coefficient (r) for each scatter plot is computed and reported in the plot.

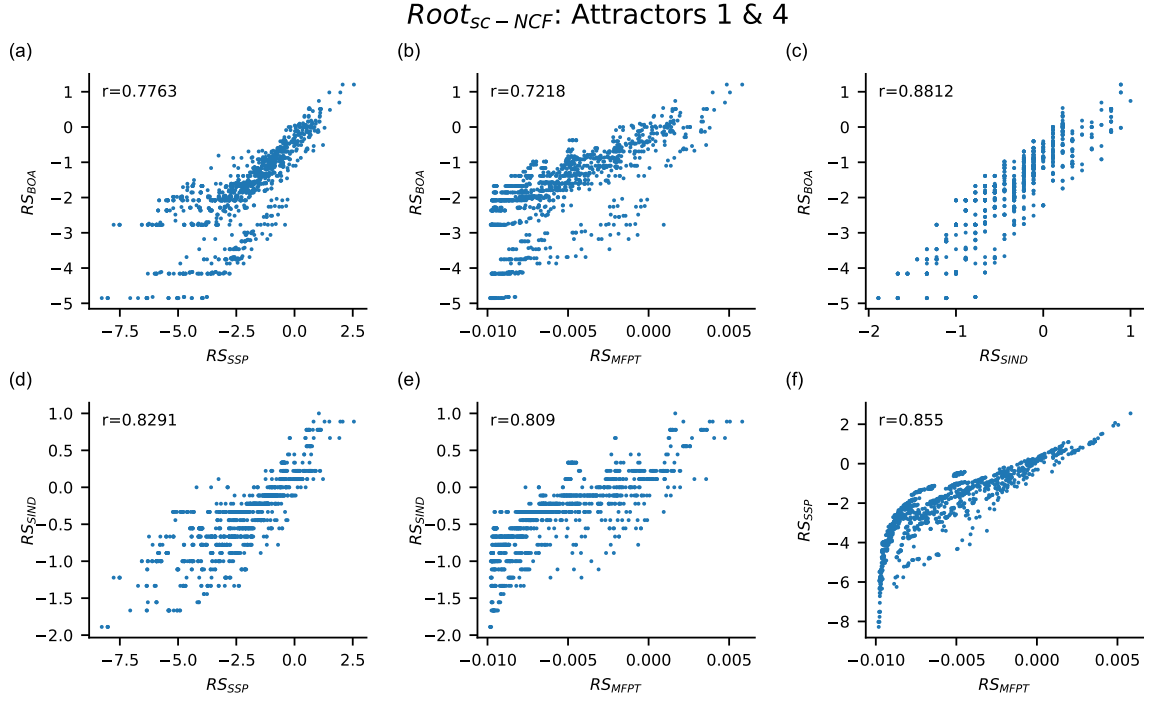


Figure D.10: Scatter plots between the different pairs of relative stability measures for the pair of attractors 1 and 4 for the ensemble $Root_{sc-NCF}$. Each sub-figure from (a) to (f) is a scatter plot where the x and y axes are for different measures of relative stability. These 5 measures are based on size of basin of attraction (RS_{BOA}), basin transition rates (RS_{BTR}), a stability index (RS_{SIND}), steady state probabilities (RS_{SSP}) and mean first passage times (RS_{MEFT}). All these measures have been computed by the exact method for the pair of biological fixed points 1 (Quiescent center (QC)) and 4 (Columella epidermis initials (CEpI)), for all 1275 models belonging to the ensemble $Root_{sc-NCF}$, at 1% noise. Of the 10 possible scatter plots for distinct pairs of the 5 relative stability measures, only 6 are shown here as RS_{BOA} and RS_{BTR} are equivalent. The Pearson correlation coefficient (r) for each scatter plot is computed and reported in the plot.

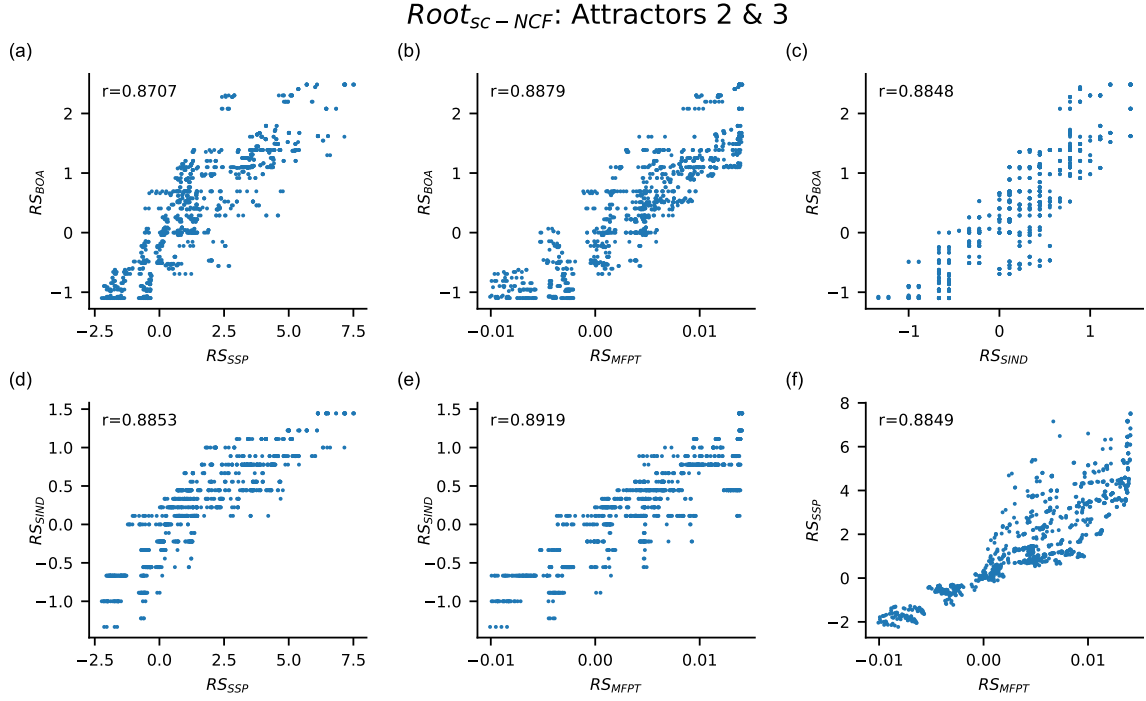


Figure D.11: Scatter plots between the different pairs of relative stability measures for the pair of attractors 2 and 3 for the ensemble $Root_{sc-NCF}$. Each sub-figure from (a) to (f) is a scatter plot where the x and y axes are for different measures of relative stability. These 5 measures are based on size of basin of attraction (RS_{BOA}), basin transition rates (RS_{BTR}), a stability index (RS_{SIND}), steady state probabilities (RS_{SSP}) and mean first passage times (RS_{MEFT}). All these measures have been computed by the exact method for the pair of biological fixed points 2 (Vascular initials (VI)) and 3 (Cortex-Endodermis initials (CEI)), for all 1275 models belonging to the ensemble $Root_{sc-NCF}$, at 1% noise. Of the 10 possible scatter plots for distinct pairs of the 5 relative stability measures, only 6 are shown here as RS_{BOA} and RS_{BTR} are equivalent. The Pearson correlation coefficient (r) for each scatter plot is computed and reported in the plot.

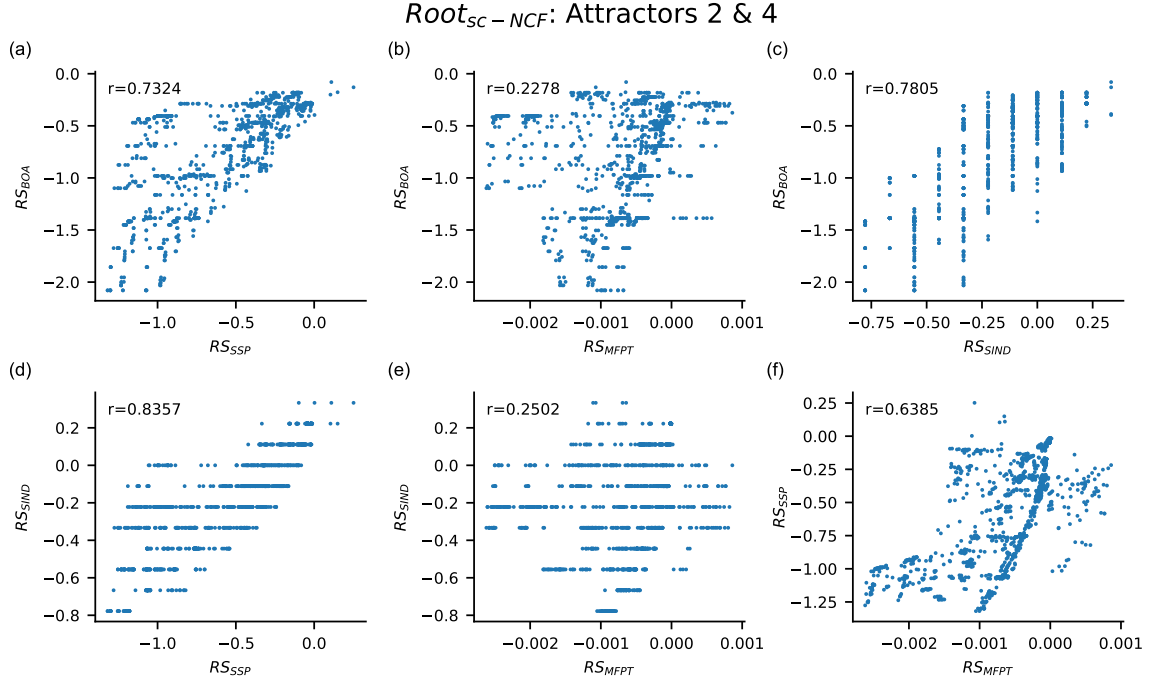


Figure D.12: Scatter plots between the different pairs of relative stability measures for the pair of attractors 2 and 4 for the ensemble $Root_{sc-NCF}$. Each sub-figure from (a) to (f) is a scatter plot where the x and y axes are for different measures of relative stability. These 5 measures are based on size of basin of attraction (RS_{BOA}), basin transition rates (RS_{BTR}), a stability index (RS_{SIND}), steady state probabilities (RS_{SSP}) and mean first passage times (RS_{MFPT}). All these measures have been computed by the exact method for the pair of biological fixed points 2 (Vascular initials (VI)) and 4 (Columella epidermis initials (CEpI)), for all 1275 models belonging to the ensemble $Root_{sc-NCF}$, at 1% noise. Of the 10 possible scatter plots for distinct pairs of the 5 relative stability measures, only 6 are shown here as RS_{BOA} and RS_{BTR} are equivalent. The Pearson correlation coefficient (r) for each scatter plot is computed and reported in the plot.

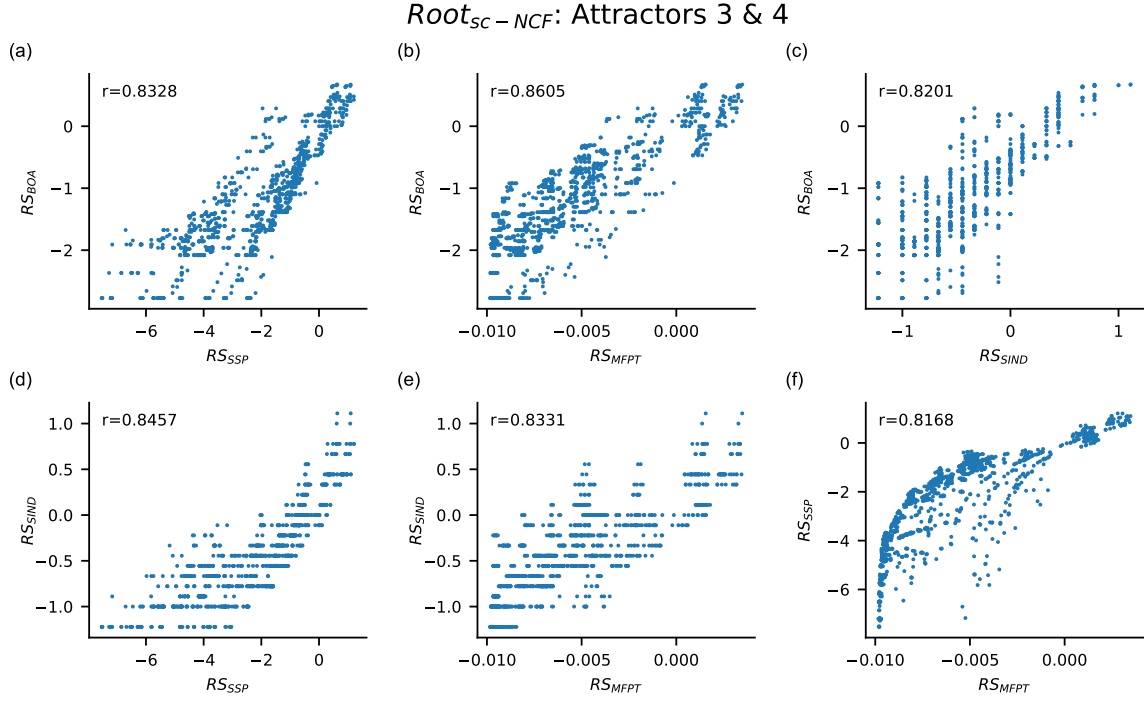


Figure D.13: Scatter plots between the different pairs of relative stability measures for the pair of attractors 3 and 4 for the ensemble $Root_{sc-NCF}$. Each sub-figure from (a) to (f) is a scatter plot where the x and y axes are for different measures of relative stability. These 5 measures are based on size of basin of attraction (RS_{BOA}), basin transition rates (RS_{BTR}), a stability index (RS_{SIND}), steady state probabilities (RS_{SSP}) and mean first passage times (RS_{MEFT}). All these measures have been computed by the exact method for the pair of biological fixed points 3 (Cortex-Endodermis initials (CEI)) and 4 (Columella epidermis initials (CEpI)), for all 1275 models belonging to the ensemble $Root_{sc-NCF}$, at 1% noise. Of the 10 possible scatter plots for distinct pairs of the 5 relative stability measures, only 6 are shown here as RS_{BOA} and RS_{BTR} are equivalent. The Pearson correlation coefficient (r) for each scatter plot is computed and reported in the plot.

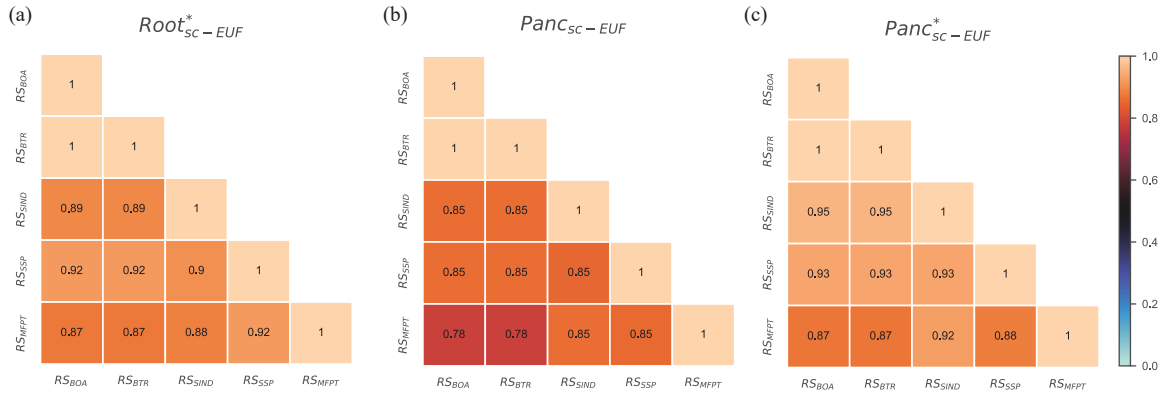


Figure D.14: Pearson correlation between different pairs of relative stability measures for the ensembles $Root_{sc-EUF}^*$, $Panc_{sc-EUF}$ and $Panc_{sc-EUF}^*$. The rows and columns correspond to choices for the relative stability measures. The heatmap indicates the value of the Pearson correlation coefficient between pairs of these measures. These 5 measures are based on size of basin of attraction (RS_{BOA}), basin transition rates (RS_{BTR}), a stability index (RS_{SIND}), steady state probabilities (RS_{SSP}) and mean first passage times (RS_{MFPT}). Note that these measures are computed by exact means across all pairs of biological fixed points, for all 1400, 7056 and 159 models in the ensembles $Root_{sc-EUF}^*$, $Panc_{sc-EUF}$ and $Panc_{sc-EUF}^*$ respectively, using a noise intensity parameter value of 1%. The upper triangular portion of the heatmap is not displayed as the heatmap entries constitute a symmetric matrix. Furthermore, RS_{BOA} and RS_{BTR} are perfectly correlated, an observation which we prove theoretically in Section 4.3.1 by showing that RS_{BOA} and RS_{BTR} are in fact equivalent.

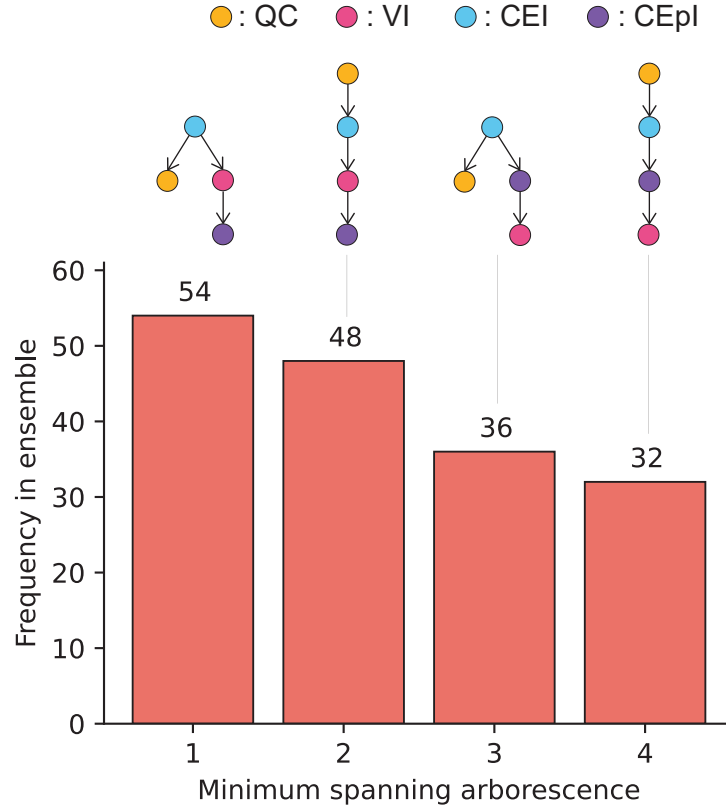


Figure D.15: Frequency distribution of the minimum spanning arborescences (MSAs) for the ensemble $Root_{sc-NCF}^*$. The MSA for a Boolean model is constructed from a complete digraph whose vertices are biological fixed points and directed edges are the MFPTs. The x axis labels the different MSAs that occur in the ensemble $Root_{sc-NCF}^*$. Of the 64 possible (labeled and oriented) trees for 4 fixed points, only 4 occur in the $Root_{sc-NCF}^*$. The y axis is the frequency of each of these trees among the 170 models of the $Root_{sc-NCF}^*$ ensemble. The biological fixed points of the $Root_{sc-NCF}^*$ ensemble are as follows. QC: Quiescent center, VI: Vascular initials, CEI: Cortex-Endodermis initials and CEpI: Columella epidermis initials.

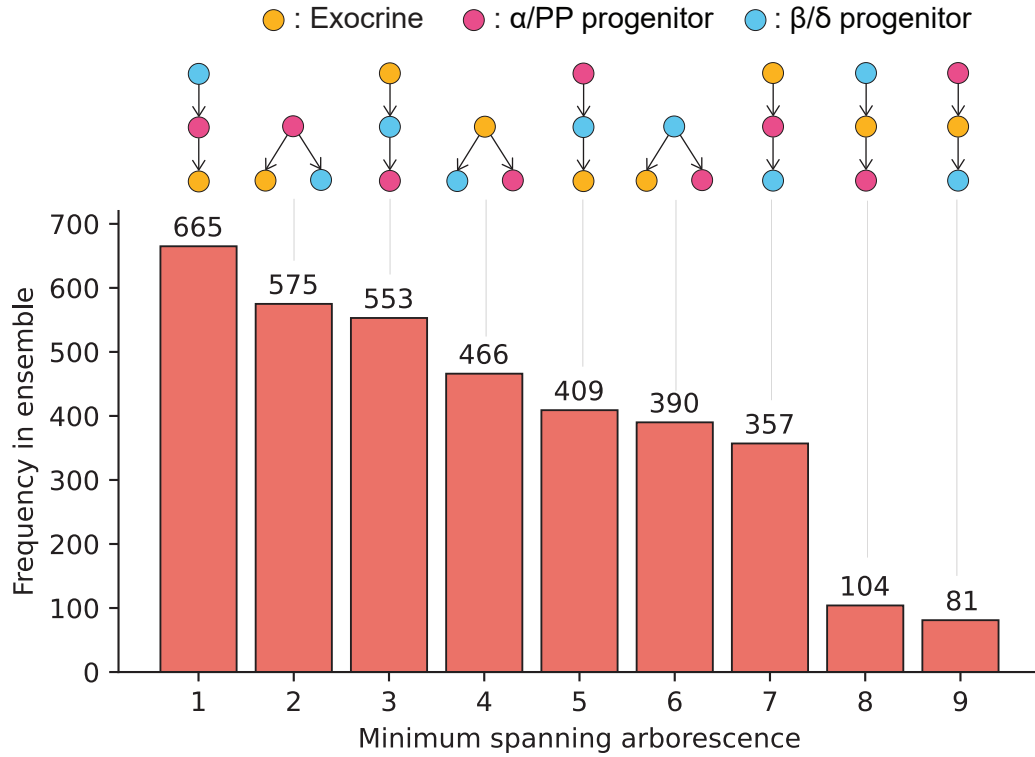


Figure D.16: Frequency distribution of the minimum spanning arborescences (MSAs) for the ensemble $Panc_{sc-NCF}$. The MSA for a Boolean model is constructed from a complete digraph whose vertices are biological fixed points and directed edges are the MFPTs. The x axis labels the different MSAs that occur in the ensemble $Panc_{sc-NCF}$. Of the 9 possible (labeled and oriented) trees for 3 fixed points, all 9 occur in the $Panc_{sc-NCF}$. The y axis is the frequency of each of these trees among the 3600 models of the $Panc_{sc-NCF}$ ensemble.

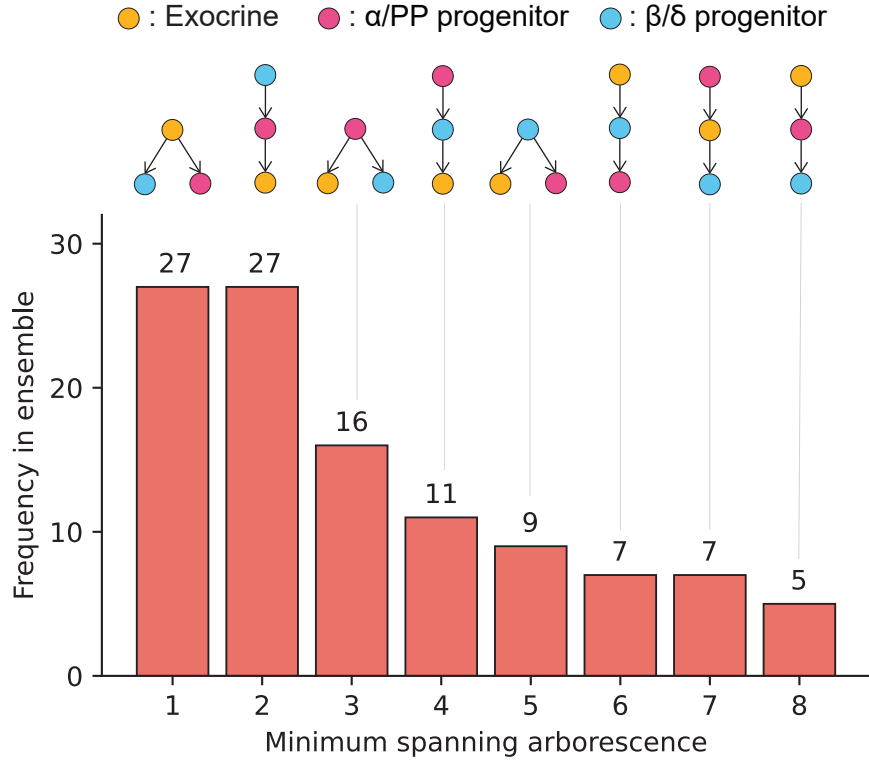


Figure D.17: Frequency distribution of the minimum spanning arborescences (MSAs) for the ensemble $Panc_{sc-NCF}^*$. The MSA for a Boolean model is constructed from a complete digraph whose vertices are biological fixed points and directed edges are the MFPTs. The x axis labels the different MSAs that occur in the ensemble $Panc_{sc-NCF}^*$. Of the 9 possible (labeled and oriented) trees for 3 fixed points, only 8 occur in the $Panc_{sc-NCF}^*$. The y axis is the frequency of each of these trees among the 109 models of the $Panc_{sc-NCF}^*$ ensemble.

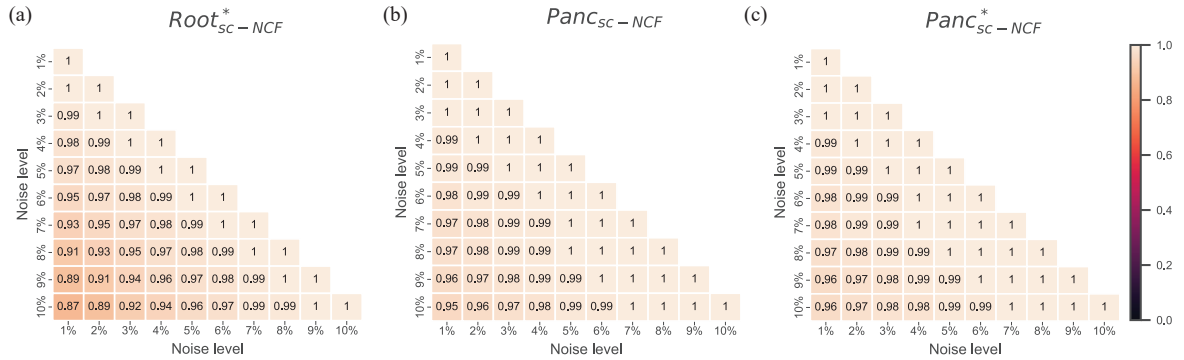


Figure D.18: Pearson correlation between RS_{MFPT} values computed by exact methods for different pairs of noise values for the ensembles $Root_{sc-NCF}^*$, $Panc_{sc-NCF}$ and $Panc_{sc-NCF}^*$. Rows and columns correspond to the noise intensities ranging from 1% to 10%. The heatmap gives the value of the Pearson correlation coefficient of RS_{MFPT} values for all pairs of attractors in different ensembles. The upper triangular portion of the heatmap is not displayed because it constitutes a symmetric matrix. The correlation between the RS_{MFPT} for different values of noise is found to be very strong even for pairs of noise values which have a large difference.

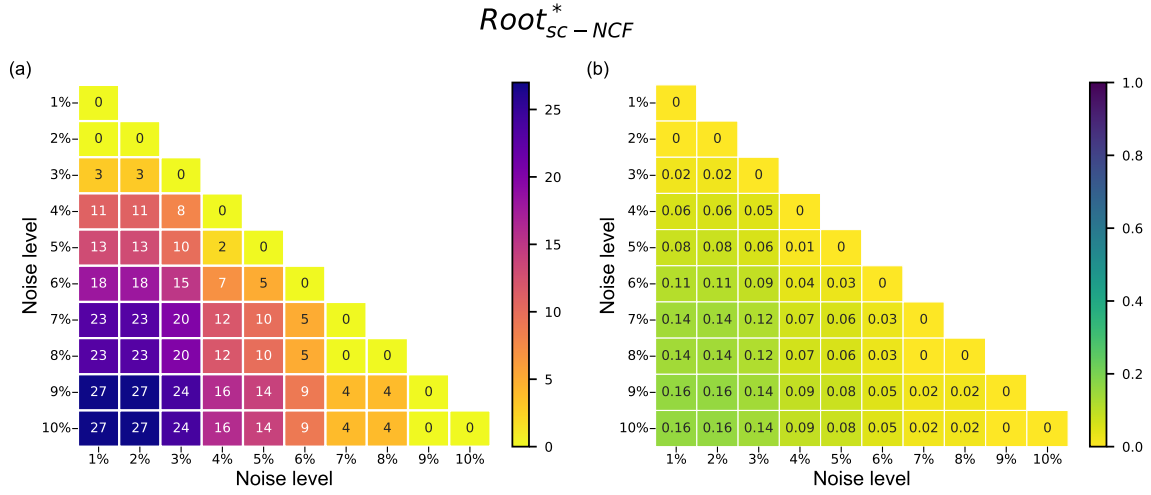


Figure D.19: Number and fraction of models which differ in at least one comparison of partial ordering of the different biological fixed points when considering two different noise values, in the ensemble $Root_{sc-NCF}^*$. The (partial) order of two fixed points is specified via the MFPT values for going from one to the other, computed here using an exact method. Rows and columns correspond to the noise intensity. The heatmap (a) gives the number of models (out of a total of 170 models in the ensemble $Root_{sc-NCF}^*$) that differ in at least one (partial) order across pairs of biological fixed points. The heatmap (b) provides the same information but using the fraction of such models. The upper triangular portions of the heatmaps are not displayed because they constitute a symmetric matrix.

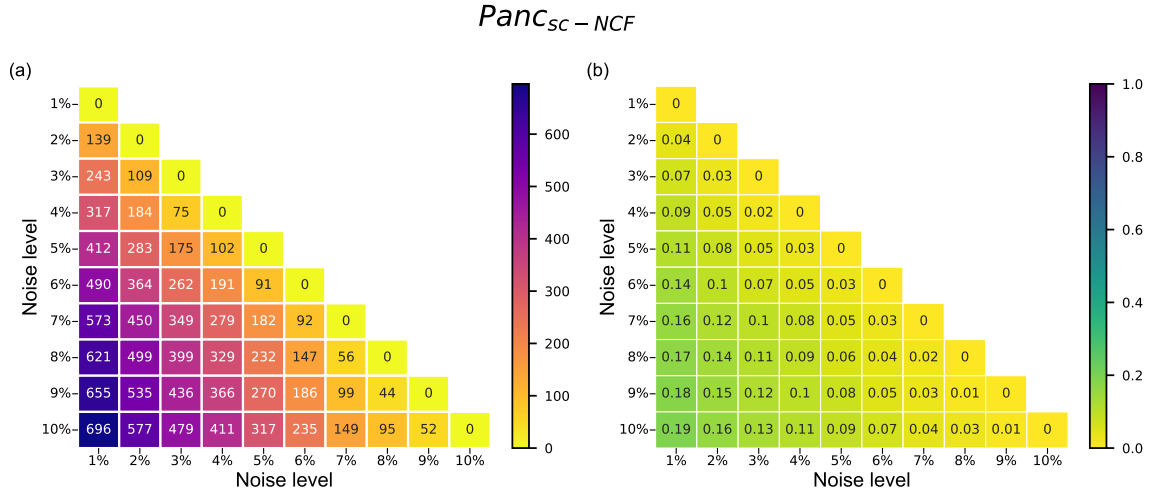


Figure D.20: Number and fraction of models which differ in at least one comparison of partial ordering of the different biological fixed points when considering two different noise values, in the ensemble $Panc_{sc-NCF}$. The (partial) order of two fixed points is specified via the MFPT values for going from one to the other, computed here using an exact method. Rows and columns correspond to the noise intensity. The heatmap (a) gives the number of models (out of a total of 3600 models in the ensemble $Panc_{sc-NCF}$) that differ in at least one (partial) order across pairs of biological fixed points. The heatmap (b) provides the same information but using the fraction of such models. The upper triangular portions of the heatmaps are not displayed because they constitute a symmetric matrix.

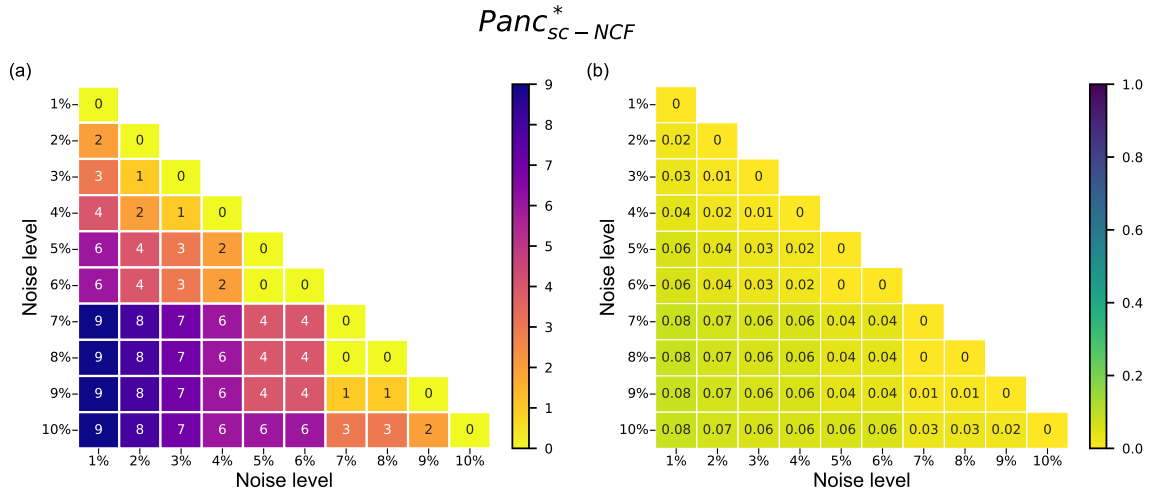


Figure D.21: Number and fraction of models which differ in at least one comparison of partial ordering of the different biological fixed points when considering two different noise values, in the ensemble $Panc_{sc-NCF}^*$. The (partial) order of two fixed points is specified via the MFPT values for going from one to the other, computed here using an exact method. Rows and columns correspond to the noise intensity. The heatmap (a) gives the number of models (out of a total of 109 models in the ensemble $Panc_{sc-NCF}^*$) that differ in at least one (partial) order across pairs of biological fixed points. The heatmap (b) provides the same information but using the fraction of such models. The upper triangular portions of the heatmaps are not displayed because they constitute a symmetric matrix.

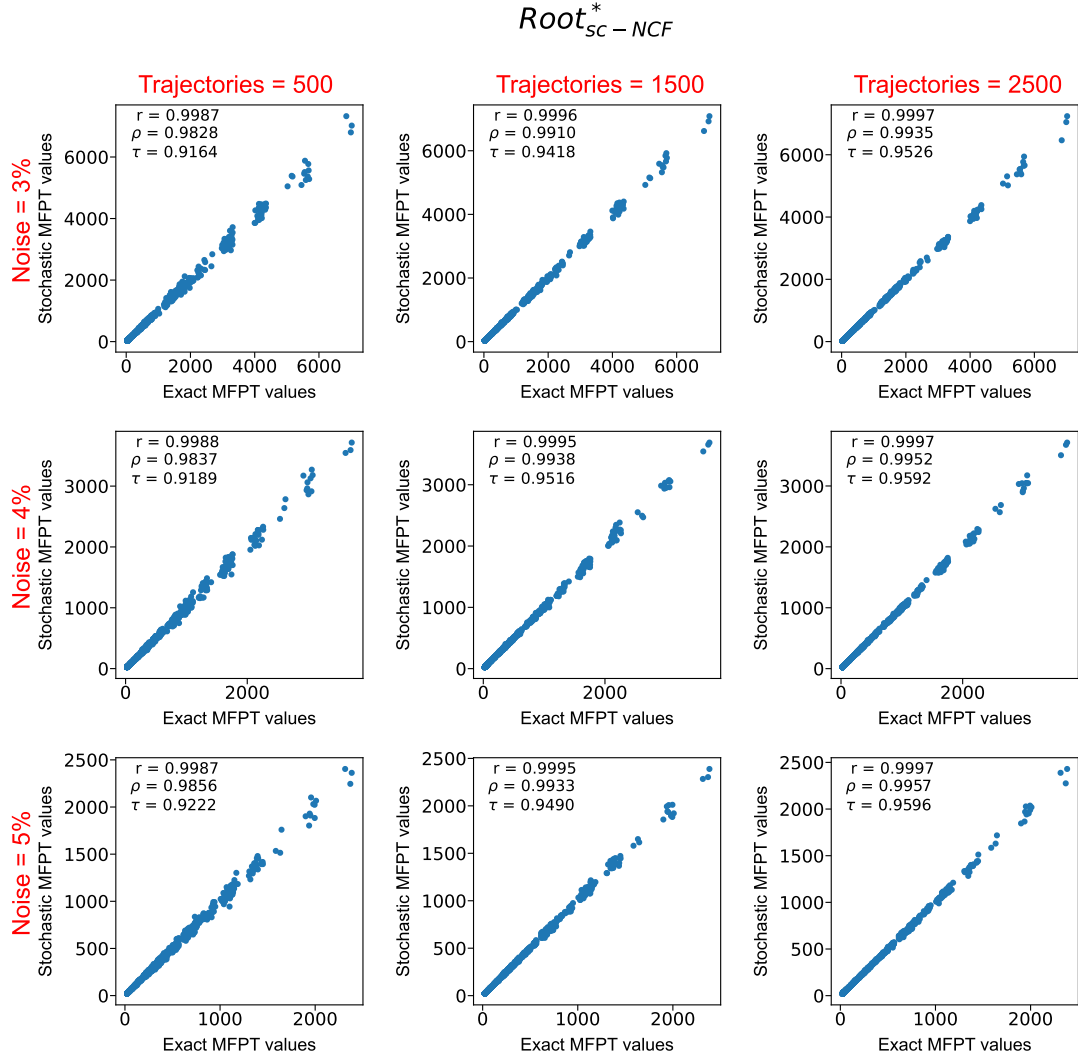


Figure D.22: Correlation between the MFPT obtained via the exact method versus the proposed stochastic method using the ensemble $Root_{sc-NCF}^*$. The x and y axes of all scatter plots represent the MFPT from the biological fixed point v to the biological fixed point u (denoted by M_{uv}) computed via exact and stochastic means respectively, for all pairs of fixed points and for all 170 models belonging to the ensemble $Root_{sc-NCF}^*$. Each scatter plot is generated for a particular noise (3%, 4% or 5%) and number of trajectories (500, 1500 or 2500) going from fixed point v to fixed point u . The exact and stochastic MFPT values are strongly correlated as can be seen from the 3 measures of correlation, namely, Pearson correlation coefficient (r), Spearman rank correlation coefficient (ρ) and Kendall rank correlation coefficient (τ). It can be seen that at a fixed noise as the number of trajectories are increased from 500 to 2500, the correlation becomes stronger across all 3 correlation measures.

Panc_{sc} – NCF

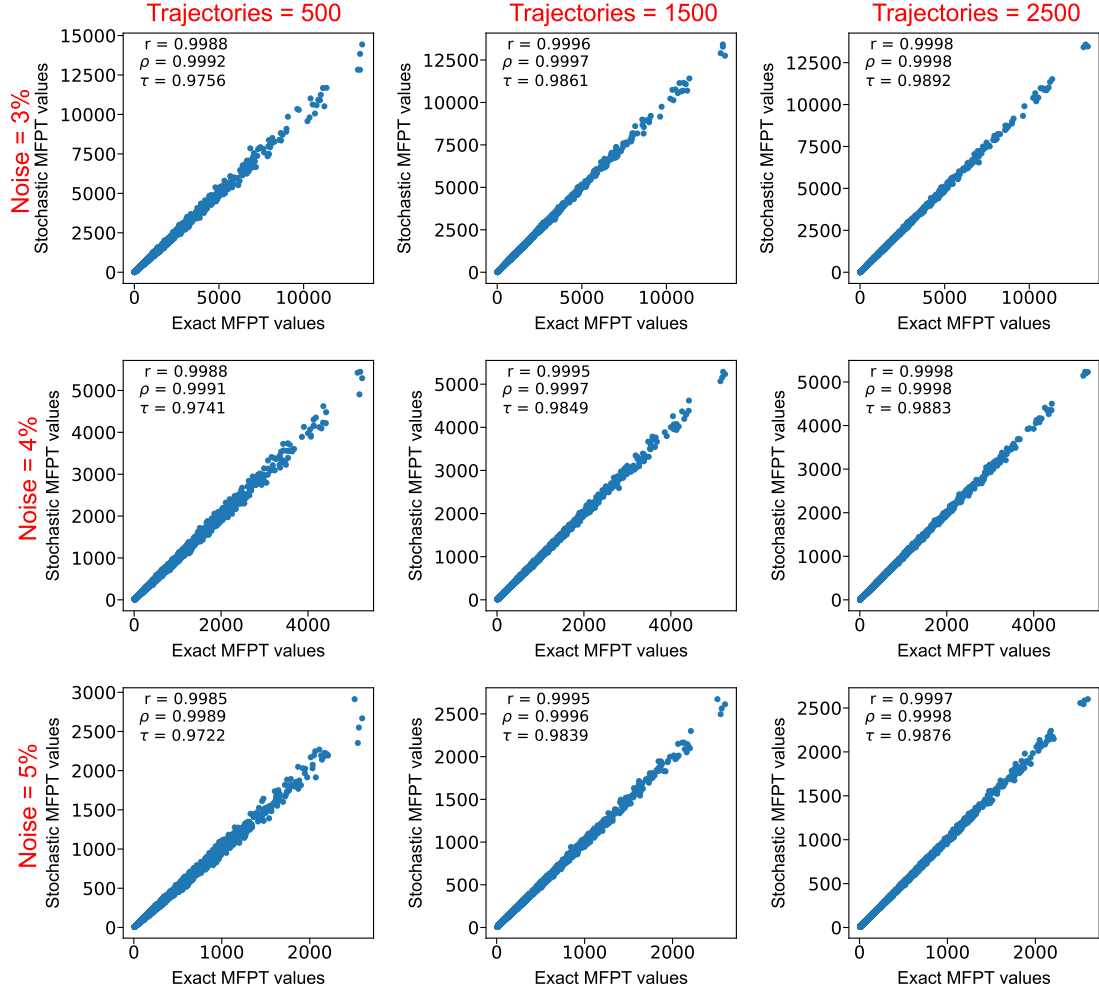


Figure D.23: Correlation between the MFPT obtained via the exact method versus the proposed stochastic method using the ensemble *Panc_{sc}–NCF*. The x and y axes of all scatter plots represent the MFPT from the biological fixed point v to the biological fixed point u (denoted by M_{uv}) computed via exact and stochastic means respectively, for all pairs of fixed points and for all 3600 models belonging to the ensemble *Panc_{sc}–NCF*. Each scatter plot is generated for a particular noise (3%, 4% or 5%) and number of trajectories (500, 1500 or 2500) going from fixed point v to fixed point u . The exact and stochastic MFPT values are strongly correlated as can be seen from the 3 measures of correlation, namely, Pearson correlation coefficient (r), Spearman rank correlation coefficient (ρ) and Kendall rank correlation coefficient (τ). It can be seen that at a fixed noise as the number of trajectories are increased from 500 to 2500, the correlation becomes stronger across all 3 correlation measures.

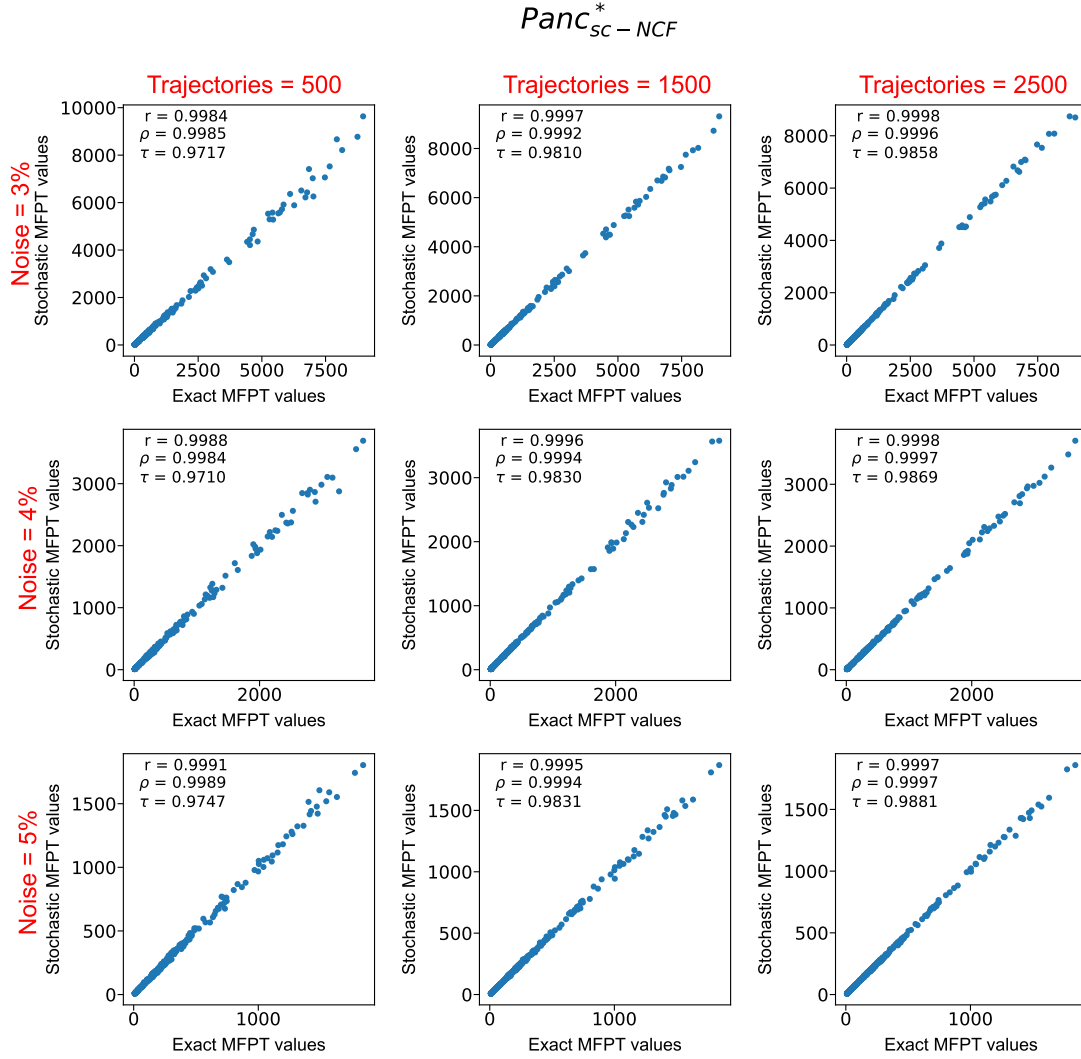


Figure D.24: Correlation between the MFPT obtained via the exact method versus the proposed stochastic method using the ensemble $Panc_{sc-NCF}^*$. The x and y axes of all scatter plots represent the MFPT from the biological fixed point v to the biological fixed point u (denoted by M_{uv}) computed via exact and stochastic means respectively, for all pairs of fixed points and for all 109 models belonging to the ensemble $Panc_{sc-NCF}^*$. Each scatter plot is generated for a particular noise (3%, 4% or 5%) and number of trajectories (500, 1500 or 2500) going from fixed point v to fixed point u . The exact and stochastic MFPT values are strongly correlated as can be seen from the 3 measures of correlation, namely, Pearson correlation coefficient (r), Spearman rank correlation coefficient (ρ) and Kendall rank correlation coefficient (τ). It can be seen that at a fixed noise as the number of trajectories are increased from 500 to 2500, the correlation becomes stronger across all 3 correlation measures.

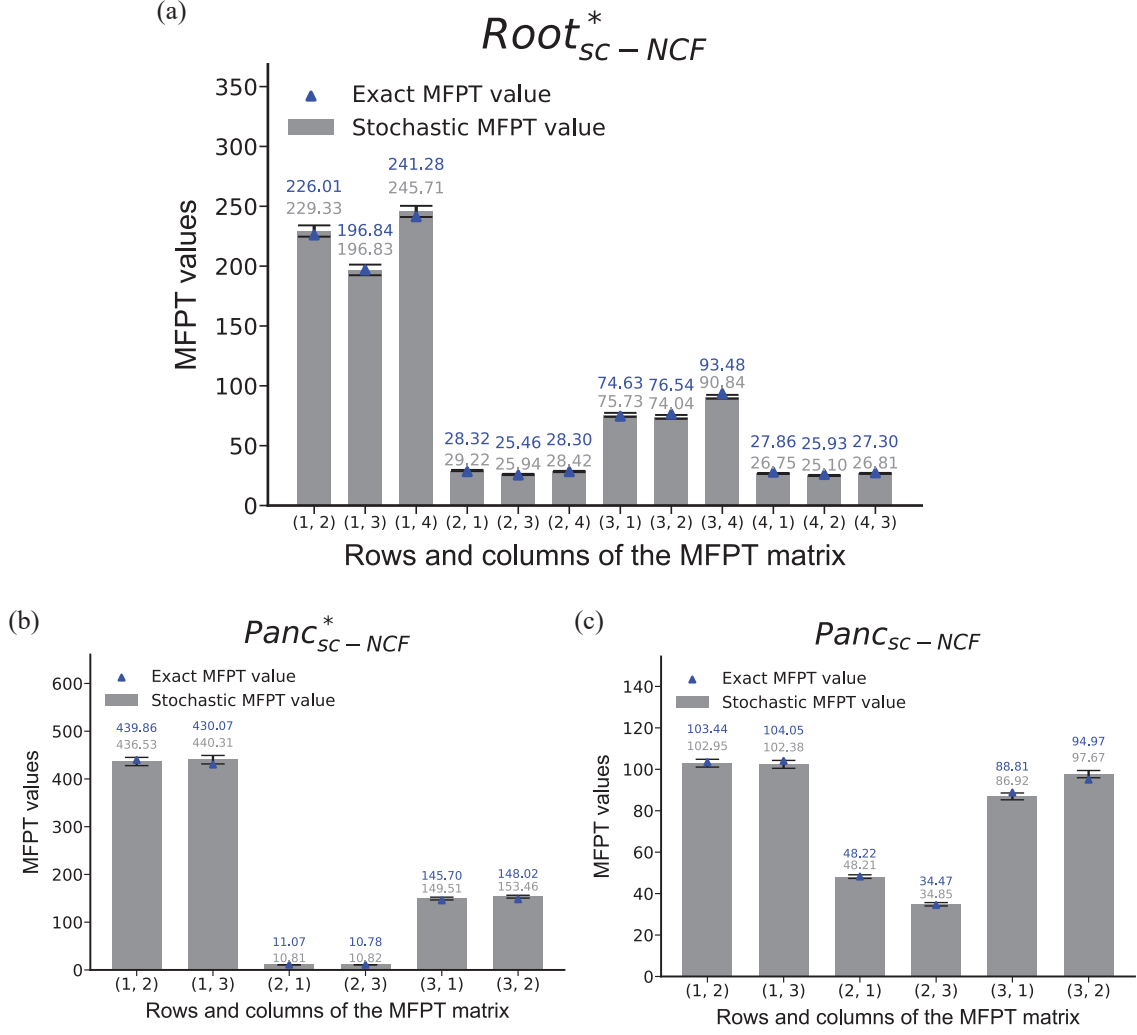
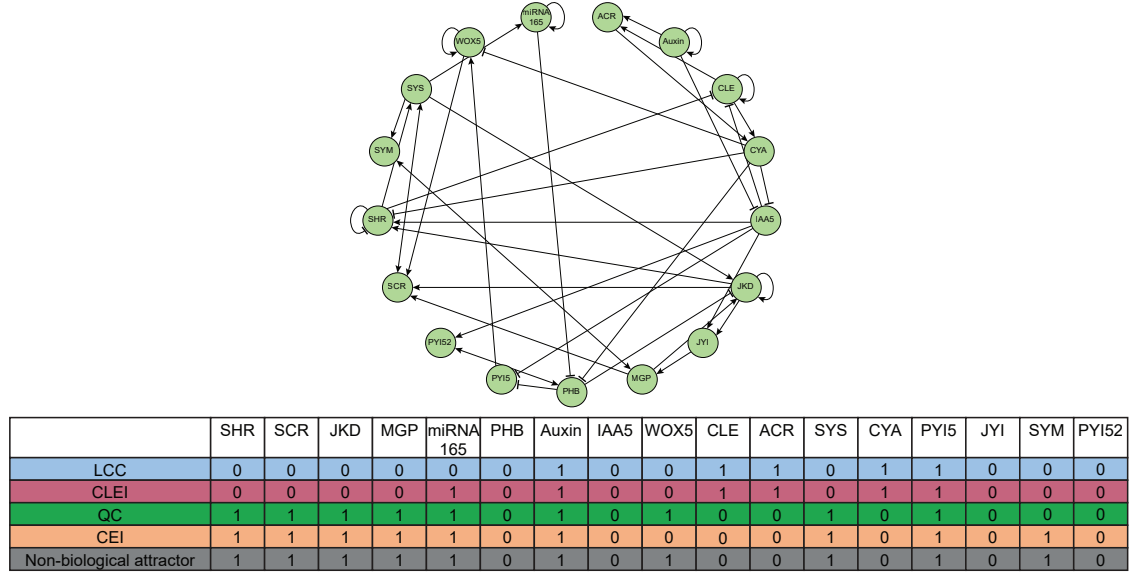


Figure D.25: Barplot of the mean first passage time (MFPT) from one biological fixed point to another, computed via stochastic methods for 3 models, each specific to one of 3 ensembles $Root_{sc-NCF}^*$, $Panc_{sc-NCF}$ and $Panc_{sc-NCF}^*$. The x axis labels the rows and columns of the MFPT matrix entries, so for instance (1, 3) denotes the case of matrix element M_{13} when going from fixed point 3 to fixed point 1. The numbering of the fixed points for the $Root_{sc-NCF}^*$ models are as follows. 1: Quiescent center (QC), 2: Vascular initials (VI), 3: Cortex-Endodermis initials (CEI) and 4: Columnella epidermis initials (CEpI). The numbering of the fixed points for the $Panc_{sc-NCF}$ and $Panc_{sc-NCF}^*$ ensembles are as follows. 1: Exocrine, 2: β/δ progenitor, 3: α /PP progenitor. The y axis represents the associated MFPTs. It is computed following our stochastic approach, averaging over 2500 different trajectories of the dynamics starting from one fixed point and stopping as soon as the other fixed point is reached when using the rules for a particular Boolean model in an ensemble at 5% noise level. The tiny error bars indicate a very low statistical error in the estimation of the MFPT value. For comparison, the MFPT values obtained via the exact method are displayed via blue triangles (numerical values are provided above the bars in blue). The MFPT obtained via the proposed stochastic method is very close to that obtained via exact means.

(a) Gene regulatory network and attractors of *Arabidopsis thaliana* root stem cell niche (Azpeitia *et al.* 2013)



(b) Gene regulatory network and attractors of *Arabidopsis thaliana* root stem cell niche (García-Gómez *et al.* 2017)

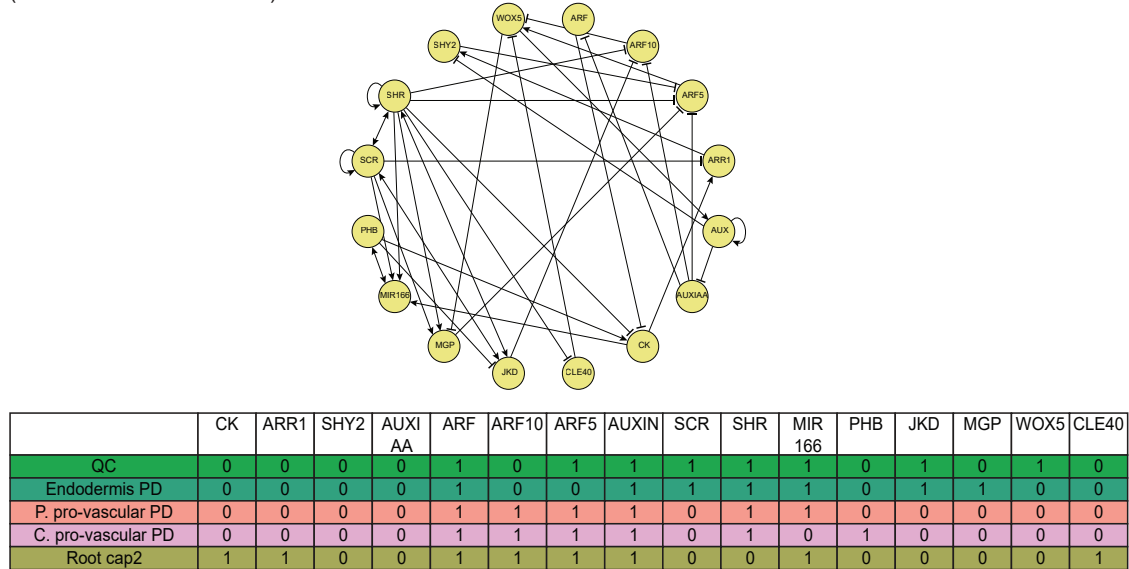
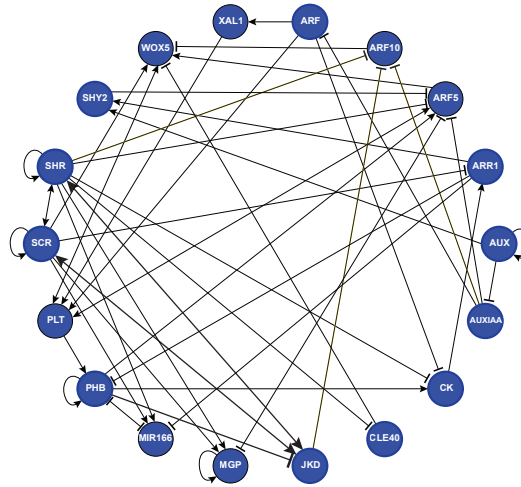


Figure D.26: 2013 and 2017 GRNs of the *Arabidopsis thaliana* root development. (a) 2013 model of the *Arabidopsis thaliana* RSCN Boolean GRN and its fixed points with $AUX = 1$. This model has 17 nodes and 42 edges of which 5 nodes are intermediate nodes. The network is constructed using regulatory interactions obtained from the BFs of *model 4* in [73]. Here, QC: Quiescent center, CEI: Cortex-endodermis initials, LCC: Lateral root-cap and CLEI: Columella and lateral root-cap-epidermis initials. (b) 2017 model of the *Arabidopsis thaliana* RAM Boolean GRN and its fixed points with $AUX = 1$. This model has 16 nodes and 39 edges. The network is constructed using regulatory interactions obtained from the BFs of the *GHRN1 model* in [74].

Gene regulatory network and attractors of *Arabidopsis thaliana* root stem cell niche
(García-Gómez *et al.* 2020)



gene	CK	ARR1	SHY2	AUXIAA	ARF	ARF10	ARF5	XAL1	PLT	AUX	SCR	SHR	MIR166	PHB	JKD	MGP	WOX5	CLE40
QC	0	0	0	0	1	0	1	1	1	1	1	1	1	0	1	0	1	0
CEI/ Endodermis PD	0	0	0	0	1	0	0	1	1	1	1	1	1	0	1	1	0	0
P. Pro-vascular PD	0	0	0	0	1	1	1	1	1	1	0	1	1	0	0	0	0	0
C. Pro-vascular PD	0	0	0	0	1	1	1	1	1	1	0	1	0	1	0	0	0	0
C. Pro-vascular TD2	1	1	0	0	1	1	1	1	1	1	0	0	0	1	0	0	0	1
Columella 1	1	1	0	0	1	1	1	1	1	1	0	0	1	0	0	0	0	1

Figure D.27: 2020 model of the *Arabidopsis thaliana* RSCN Boolean GRN and its fixed points with $AUX = 1$. This network has 18 nodes and 51 edges. The network is constructed using regulatory interactions obtained from the BFs of the model in [75]. Here, QC: Quiescent center, CEI: Cortex-endodermis initials, P. Pro-vascular PD: Peripheral Pro-vascular initials, C. Pro-vascular PD: Central Pro-vascular initials, C. Pro-vascular TD2: Transition domain, Columella 1: Columella initials.

Table D.1: Correlation between the exact and stochastic methods to compute MFPT using the $Root_{sc-NCF}$ ensemble. The Spearman (ρ) and Kendall (τ) rank correlation coefficients for MFPT values computed by the exact method and stochastic method for the ensemble $Root_{sc-NCF}$ for different combinations of number of trajectories and noise intensities is given. The associated p-values of the correlation coefficients are also provided. The column “ N_A ” of the table gives the number of MFPT values obtained from exact computations that lie outside the error bar obtained via the stochastic method.

η (%)	Number of Trajectories	Spearman		Kendall		N_A
		ρ	p-val	τ	p-val	
3	500	0.9969	$< 10^{-323}$	0.9592	$< 10^{-310}$	4950
3	1000	0.9982	$< 10^{-323}$	0.9703	$< 10^{-310}$	4825
3	1500	0.9987	$< 10^{-323}$	0.9749	$< 10^{-310}$	4935
3	2000	0.9990	$< 10^{-323}$	0.9780	$< 10^{-310}$	4938
3	2500	0.9991	$< 10^{-323}$	0.9801	$< 10^{-310}$	4884
4	500	0.9975	$< 10^{-323}$	0.9615	$< 10^{-310}$	4872
4	1000	0.9986	$< 10^{-323}$	0.9718	$< 10^{-310}$	4928
4	1500	0.9990	$< 10^{-323}$	0.9768	$< 10^{-310}$	4919
4	2000	0.9992	$< 10^{-323}$	0.9796	$< 10^{-310}$	4851
4	2500	0.9994	$< 10^{-323}$	0.9816	$< 10^{-310}$	4853
5	500	0.9976	$< 10^{-323}$	0.9616	$< 10^{-310}$	4891
5	1000	0.9988	$< 10^{-323}$	0.9725	$< 10^{-310}$	4885
5	1500	0.9991	$< 10^{-323}$	0.9774	$< 10^{-310}$	4861
5	2000	0.9993	$< 10^{-323}$	0.9803	$< 10^{-310}$	4851
5	2500	0.9994	$< 10^{-323}$	0.9822	$< 10^{-310}$	4895

Table D.2: Correlation between the exact and stochastic methods to compute MFPT using the $Root_{sc-NCF}^*$ ensemble. The Spearman (ρ) and Kendall (τ) rank correlation coefficients for MFPT values computed by the exact method and stochastic method for the ensemble $Root_{sc-NCF}^*$ for different combinations of number of trajectories and noise intensities is given. The associated p-values of the correlation coefficients are also provided. The column “ N_A ” of the table gives the number of MFPT values obtained from exact computations that lie outside the error bar obtained via the stochastic method.

η (%)	Number of Trajectories	Spearman		Kendall		N_A
		ρ	p-val	τ	p-val	
3	500	0.9826	$< 10^{-323}$	0.9154	$< 10^{-310}$	623
3	1000	0.9878	$< 10^{-323}$	0.9322	$< 10^{-310}$	632
3	1500	0.9903	$< 10^{-323}$	0.9406	$< 10^{-310}$	655
3	2000	0.9926	$< 10^{-323}$	0.9483	$< 10^{-310}$	665
3	2500	0.9943	$< 10^{-323}$	0.9548	$< 10^{-310}$	613
4	500	0.9869	$< 10^{-323}$	0.9260	$< 10^{-310}$	628
4	1000	0.9913	$< 10^{-323}$	0.9417	$< 10^{-310}$	651
4	1500	0.9936	$< 10^{-323}$	0.9510	$< 10^{-310}$	631
4	2000	0.9946	$< 10^{-323}$	0.9557	$< 10^{-310}$	652
4	2500	0.9958	$< 10^{-323}$	0.9610	$< 10^{-310}$	640
5	500	0.9849	$< 10^{-323}$	0.9204	$< 10^{-310}$	669
5	1000	0.9902	$< 10^{-323}$	0.9382	$< 10^{-310}$	679
5	1500	0.9937	$< 10^{-323}$	0.9504	$< 10^{-310}$	632
5	2000	0.9948	$< 10^{-323}$	0.9553	$< 10^{-310}$	644
5	2500	0.9956	$< 10^{-323}$	0.9597	$< 10^{-310}$	629

Table D.3: Correlation between the exact and stochastic methods to compute MFPT using the $Panc_{sc-NCF}$ ensemble. The Spearman (ρ) and Kendall (τ) rank correlation coefficients for MFPT values computed by the exact method and stochastic method for the ensemble $Panc_{sc-NCF}$ for different combinations of number of trajectories and noise intensities is given. The associated p-values of the correlation coefficients are also provided. The column “ N_A ” of the table gives the number of MFPT values obtained from exact computations that lie outside the error bar obtained via the stochastic method.

η (%)	Number of Trajectories	Spearman		Kendall		N_A
		ρ	p-val	τ	p-val	
3	500	0.9992	$< 10^{-323}$	0.9757	$< 10^{-310}$	6874
3	1000	0.9995	$< 10^{-323}$	0.9828	$< 10^{-310}$	6851
3	1500	0.9997	$< 10^{-323}$	0.9859	$< 10^{-310}$	6842
3	2000	0.9998	$< 10^{-323}$	0.9878	$< 10^{-310}$	6871
3	2500	0.9998	$< 10^{-323}$	0.9892	$< 10^{-310}$	6764
4	500	0.9990	$< 10^{-323}$	0.9739	$< 10^{-310}$	6961
4	1000	0.9995	$< 10^{-323}$	0.9817	$< 10^{-310}$	6854
4	1500	0.9997	$< 10^{-323}$	0.9852	$< 10^{-310}$	6738
4	2000	0.9998	$< 10^{-323}$	0.9869	$< 10^{-310}$	6938
4	2500	0.9998	$< 10^{-323}$	0.9883	$< 10^{-310}$	6958
5	500	0.9989	$< 10^{-323}$	0.9723	$< 10^{-310}$	6938
5	1000	0.9995	$< 10^{-323}$	0.9806	$< 10^{-310}$	6888
5	1500	0.9996	$< 10^{-323}$	0.9841	$< 10^{-310}$	6777
5	2000	0.9997	$< 10^{-323}$	0.9860	$< 10^{-310}$	6872
5	2500	0.9998	$< 10^{-323}$	0.9875	$< 10^{-310}$	6889

Table D.4: Correlation between the exact and stochastic methods to compute MFPT using the $Panc_{sc-NCF}^*$ ensemble. The Spearman (ρ) and Kendall (τ) rank correlation coefficients for MFPT values computed by the exact method and stochastic method for the ensemble $Panc_{sc-NCF}^*$ for different combinations of number of trajectories and noise intensities is given. The associated p-values of the correlation coefficients are also provided. The column “ N_A ” of the table gives the number of MFPT values obtained from exact computations that lie outside the error bar obtained via the stochastic method.

η (%)	Number of Trajectories	Spearman		Kendall		N_A
		ρ	p-val	τ	p-val	
3	500	0.9984	$< 10^{-323}$	0.9715	$< 10^{-310}$	213
3	1000	0.9990	$< 10^{-323}$	0.9788	$< 10^{-310}$	191
3	1500	0.9993	$< 10^{-323}$	0.9829	$< 10^{-310}$	210
3	2000	0.9995	$< 10^{-323}$	0.9845	$< 10^{-310}$	216
3	2500	0.9995	$< 10^{-323}$	0.9859	$< 10^{-310}$	227
4	500	0.9987	$< 10^{-323}$	0.9730	$< 10^{-310}$	187
4	1000	0.9992	$< 10^{-323}$	0.9795	$< 10^{-310}$	213
4	1500	0.9994	$< 10^{-323}$	0.9828	$< 10^{-310}$	214
4	2000	0.9996	$< 10^{-323}$	0.9860	$< 10^{-310}$	199
4	2500	0.9996	$< 10^{-323}$	0.9867	$< 10^{-310}$	206
5	500	0.9987	$< 10^{-323}$	0.9732	$< 10^{-310}$	217
5	1000	0.9993	$< 10^{-323}$	0.9804	$< 10^{-310}$	226
5	1500	0.9995	$< 10^{-323}$	0.9838	$< 10^{-310}$	212
5	2000	0.9996	$< 10^{-323}$	0.9868	$< 10^{-310}$	188
5	2500	0.9997	$< 10^{-323}$	0.9872	$< 10^{-310}$	203

Table D.5: Boolean functions for the 2013 RSCN model in the BoolNet format. The column with header “Target gene name” contains the list of genes whose regulation is captured by the corresponding row entry in the column “Regulatory logic rule”. The symbols &, | and ! correspond to the logic operators AND, OR and NOT respectively.

Serial Number	Target gene name	Regulatory logic rule
1	SHR	JKD (IAA5 & (!CYA !SHR))
2	SCR	(JKD & (MGP WOX5)) (SYS & (MGP & WOX5))
3	JKD	(SYS & JKD & (!PHB MGP)) (SYS & MGP) (!PHB & MGP)
4	MGP	JYI SYM
5	miRNA165	SYS miRNA165
6	PHB	!miRNA165 & (!CYA PYI52)
7	Auxin	Auxin
8	IAA5	!Auxin & !CYA
9	WOX5	PYI5 & !CYA & WOX5
10	CLE	(CLE ! IAA5) & !SHR
11	ACR	CLE & Auxin
12	SYS	SHR & SCR
13	CYA	CLE & ACR
14	PYI5	!PHB & !IAA5
15	JYI	JKD & IAA5
16	SYM	SYS & MGP
17	PYI52	PHB & IAA5

Table D.6: Boolean functions for the 2017 RAM model in the BoolNet format. The column with header “Target gene name” contains the list of genes whose regulation is captured by the corresponding row entry in the column “Regulatory logic rule”. The symbols &, | and ! correspond to the logic operators AND, OR and NOT respectively.

Serial Number	Target gene name	Regulatory logic rule
1	CK	(PHB & !ARF) !SHR
2	ARR1	!SCR & CK
3	SHY2	ARR1 & !AUX
4	AUXIAA	!AUX
5	ARF	!AUXIAA
6	ARF10	!(JKD & SHR) & !AUXIAA
7	ARF5	!(SHR & MGP) & !SHY2 & !AUXIAA
8	AUX	WOX5 AUX
9	SCR	SHR & JKD & SCR
10	SHR	SHR (SCR & JKD)
11	MIR166	(SHR & SCR & ! CK) !PHB
12	PHB	!MIR166
13	JKD	!PHB & SHR & SCR
14	MGP	!WOX5 & SHR & SCR
15	WOX5	!ARF10 & ARF5 & !CLE40
16	CLE40	!SHR

Table D.7: Boolean functions for the 2020 RSCN model in the BoolNet format.
The column with header “Target gene name” contains the list of genes whose regulation is captured by the corresponding row entry in the column “Regulatory logic rule”. The symbols &, | and ! correspond to the logic operators AND, OR and NOT respectively.

Serial Number	Target gene name	Regulatory logic rule
1	CK	(PHB & !ARF) !SHR
2	ARR1	!SCR & CK
3	SHY2	ARR1 & !AUX
4	AUXIAA	!AUX
5	ARF	!AUXIAA
6	ARF10	!(JKD & SHR) & !AUXIAA
7	ARF5	((PHB PLT) & !(SHR & MGP)) & !SHY2 & !AUXIAA
8	XAL1	ARF
9	PLT	ARF5 ARF WOX5 XAL1
10	AUX	AUX
11	SCR	SHR & JKD & SCR
12	SHR	SHR (SCR & JKD)
13	MIR166	(SCR & SHR & !ARR1) !PHB
14	PHB	((!ARR1 & PLT) PHB) & !MIR166
15	JKD	!PHB & SHR & SCR
16	MGP	!ARF5 & SHR & SCR & MGP
17	WOX5	!ARF10 & ARF5 & !CLE40 & SCR & PLT
18	CLE40	!SHR

Appendix E

Additional Figures and Tables for Chapter 5

Table E.1: Fraction of link operator functions (LOFs) among the complete space of Boolean functions (BFs) for a given number of inputs k . Evidently, $k = m + n$ where m and n are the number of activators and inhibitors, respectively. A LOF should have at least one activating input ($m \geq 1$) and at least one inhibiting input ($n \geq 1$), and thus, LOFs can exist only for nodes with 2 or more inputs ($k \geq 2$). Here, we give the total number of LOFs for a specific k which is the cumulative number across the different possible combinations of m activators and n inhibitors. Moreover, we report separately the number of functions in the four different consistent types of LOFs namely, AND-NOT, OR-NOT, AND-pairs ($n > 1$) and OR-pairs ($m > 1$). Finally, the table also gives the total number of BFs for a specific k . As k increases, it can be seen that the LOFs become an infinitesimal fraction of the complete space of BFs.

k	Total number of BFs	LOFs				Total	Fraction of LOFs in BFs
		AND- NOT	OR- NOT	AND- pairs ($n > 1$)	OR- pairs ($m > 1$)		
2	16	2	2	0	0	4	2.50×10^{-1}
3	256	6	6	0	0	12	4.69×10^{-2}
4	65536	14	14	10	10	48	7.32×10^{-4}
5	4294967296	30	30	25	25	110	2.56×10^{-8}
6	1.84467×10^{19}	62	62	56	56	236	1.28×10^{-17}
7	3.40282×10^{38}	126	126	119	119	490	1.44×10^{-36}
8	1.15792×10^{77}	254	254	246	246	1000	8.64×10^{-75}
9	1.34078×10^{154}	510	510	501	501	2022	1.51×10^{-151}
10	1.80×10^{308}	1022	1022	1012	1012	4068	2.26×10^{-305}

Table E.2: The abundance of link operator functions (LOFs) in the collection of Boolean functions (BFs) from reconstructed models of biological systems. The reference biological dataset consists of BFs from 57 Boolean models compiled in the Cell Collective database (<https://cellcollective.org/>). Notably, a LOF should have at least one activating input ($m \geq 1$) and at least one inhibiting input ($n \geq 1$), and thus, LOFs can exist only for nodes with 2 or more inputs ($k \geq 2$). Focussing on the subset of Boolean functions in the 57 reconstructed models that have at least one activating input ($m \geq 1$) and at least one inhibiting input ($n \geq 1$), the table classifies the BFs in the reference biological dataset into effective and unate functions (EUFs) and different types of consistent LOFs. It is evident that EUFs, and moreover, the AND-NOT LOFs within EUFs, are abundant in the reference biological dataset regardless of k .

k	m	n	BFs in reference biological dataset	EUFs	LOFs				Total
					AND- NOT	OR- NOT	AND- pairs ($n > 1$)	OR- pairs ($m > 1$)	
2	1	1	158	150	147	3	0	0	150
3	1	2	35	32	30	1	1	0	32
3	2	1	94	87	47	2	0	0	49
4	1	3	16	16	13	1	0	0	14
4	2	2	38	35	17	0	0	0	17
4	3	1	57	48	18	0	0	0	18
5	1	4	4	4	1	0	0	0	1
5	2	3	16	15	10	0	0	0	10
5	3	2	25	24	8	0	0	0	8
5	4	1	20	17	4	0	0	0	4
6	2	4	3	3	1	0	0	0	1
6	3	3	14	11	5	0	0	0	5
6	4	2	14	13	4	0	0	0	4
6	5	1	13	8	2	0	0	0	2
7	2	5	1	1	1	0	0	0	1
7	3	4	1	1	0	0	0	0	0
7	4	3	5	5	1	0	0	0	1
7	5	2	3	1	1	0	0	0	1
7	6	1	8	4	0	0	0	0	0
8	3	5	1	1	0	0	0	0	0
8	4	4	3	3	0	0	0	0	0
8	6	2	1	1	0	0	0	0	0
8	7	1	5	1	0	0	0	0	0

k	m	n	BFs in reference biological dataset	EUFs	LOFs				Total
					AND- NOT	OR- NOT	AND- pairs ($n > 1$)	OR- pairs ($m > 1$)	
9	4	5	1	1	1	0	0	0	1
9	6	3	1	0	0	0	0	0	0
9	7	2	2	1	0	0	0	0	0
9	8	1	2	2	0	0	0	0	0
10	3	7	1	1	0	0	0	0	0
10	7	3	1	0	0	0	0	0	0
10	8	2	1	1	0	0	0	0	0
10	9	1	1	0	0	0	0	0	0
12	10	2	3	3	0	0	0	0	0

Table E.3: The number of Boolean functions (BFs) in the reference biological dataset for each input with only activators or only inhibitors, and those with at least one activator and one inhibitor.

k	Number of BFs		
	Total	With only either acti- vators or inhibitors	With at least one ac- tivator and one in- hibitor
1	658	658	0
2	465	307	158
3	254	125	129
4	162	51	111
5	92	27	65
6	55	11	44
7	20	2	18
8	16	6	10
9	9	3	6
10	5	1	4
11	1	1	0
12	3	0	3
14	1	1	0

Table E.4: The network average sensitivity of models where the numbers of activating and inhibiting inputs are as in the reconstructed biological networks but have particular types of Boolean functions namely, random effective functions (EFs), random effective and unate functions (EUFs), AND-NOT, OR-NOT, AND-pairs, OR-pairs. The network average sensitivity for each network was simulated 1000 times and the average of those values are presented here. Here N , k , m and n are the number of nodes in the network, number of inputs, number of activators and the number of inhibitors respectively. The network average sensitivity is calculated with respect to the number of nodes with at least one input N ($k \geq 1$). $N(m \geq 1$ and $n \geq 1)$ is the number of nodes in the network which have at least one activator and one inhibitor. The column “If Model is EUF” is TRUE whenever all the functions assigned in the original Boolean model are effective and unate. EUFs could be assigned to nodes with at most 9 inputs. Our randomization procedure for choosing EF and EUFs is explained in Appendix B. If the number of inputs to any node exceeds 9, we assign the Boolean function provided in the original model. When computing the sensitivity of the given network structure with different types of LOFs, we perform the following modifications: (a) If a function had an ineffective input and was unate, the ineffective input was removed and an LOF with the remaining signs was assigned. (b) If a function was not unate (regardless of its effectiveness), then the signs which were assigned in the original model were used to choose an LOF. (c) If a function possessed only activators or only inhibitors, an LOF could not be assigned, and hence the functions used in the original Boolean model were kept.

Model PMID	N	N ($k \geq 1$)	N ($m \geq 1$ and $n \geq 1$)	If ($k \leq 9$) is EUF	If model is EUF	Network average sensitivity of the models							
						EF	EUF	AND- NOT	OR- NOT	AND- pairs	OR- pairs	Biological	
20862356	5	5	4	TRUE	TRUE	1.463	1.132	0.9	1.05	1.15	1.15	0.8	
23868318	11	9	0	TRUE	TRUE	1	1	1	1	1	1	1	
18463633	9	9	2	TRUE	TRUE	1.221	1.076	1	1	1.111	1.111	0.944	
21563979	13	10	4	TRUE	TRUE	1.27	1.084	1	0.975	1.1	1.1	1	
16873462	10	10	7	TRUE	TRUE	1.869	1.298	0.775	1	1.163	1.163	1.019	
23868318	18	10	2	TRUE	TRUE	0.965	0.925	0.925	0.925	0.925	0.925	0.925	
24970389	16	11	3	TRUE	TRUE	1.049	0.838	0.688	0.722	0.835	0.835	0.688	
23868318	24	11	2	TRUE	TRUE	0.89	0.834	0.795	0.841	0.841	0.841	0.841	
24970389	16	11	3	TRUE	TRUE	1.081	0.815	0.653	0.619	0.784	0.784	0.653	

Model PMID	N	N ($k \geq 1$)	N ($m \geq 1$ and $n \geq 1$)	If ($k \leq 9$)	If model is EUF	Network average sensitivity of the models						
						EF	EUF	AND- NOT	OR- NOT	AND- pairs	OR- pairs	Biological
24970389	16	11	3	TRUE	TRUE	1.152	0.861	0.688	0.619	0.807	0.807	0.688
26090929	18	12	11	FALSE	TRUE	3.096	1.589	0.634	0.807	1.163	1.163	0.59
23868318	26	12	6	TRUE	TRUE	1.472	1.151	0.97	0.915	1.048	1.048	0.957
23056457	15	14	5	TRUE	TRUE	1.203	1.027	0.968	0.896	1.008	1.008	0.99
26340681	14	14	13	TRUE	TRUE	2.412	1.502	0.881	0.798	1.256	1.256	1.275
23171249	24	14	8	FALSE	TRUE	1.974	1.231	0.954	0.805	1.041	1.041	0.884
17010384	16	16	4	TRUE	TRUE	1.072	1.01	1.016	0.984	1.016	1.016	0.984
26751566	22	17	7	TRUE	TRUE	1.468	1.158	0.972	0.911	1.071	1.071	0.999
23134720	19	18	8	TRUE	TRUE	1.195	1.046	1.014	0.972	1.056	1.056	0.986
22102804	18	18	9	TRUE	TRUE	1.217	1.042	0.986	0.986	1.069	1.069	1.014
23868318	24	18	1	TRUE	TRUE	0.945	0.934	0.934	0.934	0.934	0.934	0.934
16542429	23	19	8	TRUE	TRUE	1.192	1.046	1.036	0.938	1.036	1.036	1.036
26616283	23	19	10	TRUE	TRUE	1.364	1.097	0.999	0.912	1.06	1.06	1.06
25908096	22	20	6	TRUE	TRUE	1.192	1.067	1.022	0.997	1.05	1.05	1.044
24970389	25	21	13	TRUE	TRUE	1.542	1.132	0.961	0.798	1.086	1.086	0.961
24970389	25	21	14	TRUE	TRUE	1.586	1.158	0.938	0.83	1.095	1.095	0.938
24970389	25	21	14	TRUE	TRUE	1.776	1.229	0.934	0.786	1.129	1.129	0.934
22253585	26	25	7	TRUE	TRUE	1.246	1.085	1.03	0.96	1.085	1.085	1.01
20221256	28	25	10	TRUE	TRUE	1.17	1.039	1.05	0.94	1.05	1.05	1.03
27594840	26	26	12	TRUE	TRUE	1.591	1.192	0.997	0.847	1.104	1.104	0.967
28361666	42	28	10	TRUE	TRUE	1.162	1.033	1.007	0.949	1.025	1.025	0.998
22253585	33	33	9	TRUE	TRUE	1.24	1.084	1.045	0.947	1.083	1.083	1.011
19025648	34	34	3	TRUE	TRUE	1.011	0.993	0.993	0.993	0.993	0.993	0.993
19422837	41	39	16	TRUE	TRUE	1.242	1.069	1.032	0.942	1.077	1.077	1.006
16968132	44	40	12	TRUE	TRUE	1.246	1.083	1.006	0.975	1.084	1.084	1
29206223	51	49	10	TRUE	TRUE	1.046	0.982	0.974	0.949	0.985	0.985	0.985
24250280	53	49	15	TRUE	TRUE	1.222	1.071	1.018	0.972	1.079	1.079	1.005

Model PMID	N	N ($k \geq 1$)	N ($m \geq 1$ and $n \geq 1$)	If ($k \leq 9$)	If model is EUF	Network average sensitivity of the models						
						EF	EUF	AND- NOT	OR- NOT	AND- pairs	OR- pairs	Biological
22253585	53	52	12	TRUE	TRUE	1.258	1.084	1.016	0.927	1.054	1.054	1.003
22102804	60	54	53	TRUE	TRUE	1.877	1.282	1.006	0.847	1.216	1.216	0.926
19144179	73	55	10	TRUE	TRUE	1.105	1.035	0.984	1.016	1.025	1.025	1.011
23233838	73	60	9	FALSE	TRUE	1.249	0.988	0.965	0.934	0.994	0.994	0.929
27148350	62	61	0	TRUE	TRUE	0.934	0.934	0.934	0.934	0.934	0.934	0.934
22962472	68	62	5	TRUE	TRUE	1.036	1.002	0.996	0.976	1.004	1.004	0.988
26446703	70	69	30	TRUE	TRUE	1.267	1.076	1.014	0.957	1.09	1.09	0.992
21968890	86	71	18	TRUE	TRUE	1.195	1.047	0.967	0.966	1.047	1.047	0.897
19662154	104	76	17	FALSE	TRUE	1.308	1.005	0.975	0.873	0.982	0.982	0.937
17722974	101	94	14	TRUE	TRUE	1.11	1.034	1.006	0.974	1.023	1.023	0.962
21968890	118	104	18	TRUE	TRUE	1.095	1.008	0.972	0.953	1.004	1.004	0.927
28584084	33	33	15	FALSE	FALSE	1.462	1.071	1.038	0.934	1.11	1.11	1.028
22267503	28	28	25	TRUE	FALSE	2.219	1.435	1.028	0.693	1.198	1.198	1.019
26408858	81	81	25	TRUE	FALSE	1.171	0.938	0.938	0.925	1.027	1.027	0.929
26385365	15	15	14	TRUE	FALSE	2.231	1.426	1.015	0.744	1.231	1.231	0.978
28639170	13	13	1	TRUE	FALSE	0.746	0.731	0.808	0.808	0.808	0.808	0.808
23868318	23	23	1	TRUE	FALSE	0.596	0.587	0.848	0.848	0.848	0.848	0.848
19118495	20	20	5	TRUE	FALSE	1.078	0.937	0.847	0.953	0.978	0.978	0.847
19185585	18	18	11	TRUE	FALSE	1.717	1.241	1	0.806	1.135	1.135	1.108
26573569	19	19	14	FALSE	FALSE	2.072	1.269	0.947	0.808	1.191	1.191	1.053
27542373	49	49	15	TRUE	FALSE	1.612	1.217	0.998	0.856	1.056	1.056	1.092

Table E.5: Overlap of the network average sensitivity (s) distribution of various BFs with the outliers of the distribution of s of the biological models. The fraction of data points in the first distribution which fall outside the 95% confidence interval of the biological distribution is calculated. Three ways to define outliers are provided: “upper” corresponding to being greater than the 95 percentile value of the biological distribution, “lower” corresponding to being less than the 5 percentile value, and “two sided” as being less than the 2.5 percentile or greater than the 97.5 percentile.

Type of BF	upper	lower	two sided
EF	0.72	0.017	0.35
EUf	0.3	0.017	0.12
AND-NOT	0	0.017	0
OR-NOT	0	0.035	0
AND-pairs	0.21	0	0
OR-pairs	0.21	0	0

Appendix F

Additional Figures and Tables for Chapter 6

Table F.1: A list of 169 complexes from the set of 1325 complexes in *H. sapiens* such that all the protein subunits of these complexes are transcription factors. The 1325 complexes in *H. sapiens* were obtained from the EBI Complex Portal database. The list of human TFs was obtained from <http://humantfs.ccb.utoronto.ca/>. “Complex ID” is the identifier of the complex as given in the EBI Complex Portal database. “Complex name” is the name of the complex as given in the EBI Complex Portal database. “Size of the complex” is the number of protein subunits the complex is constituted of. “Uniprot ID of protein subunits” gives the Uniprot IDs of the subunits in a complex and their stoichiometric coefficients as integers within brackets. Stoichiometric coefficients which are unknown are represented by (–). “Is a TR” column is “yes” if the complex acts as a transcriptional regulator (TR) based on manual literature curation.

Complex ID	Complex name	Size of the complex	UniProt ID of protein subunits	Is a TR
CPX-6405	bZIP transcription factor complex, ATF1-CREB1	2	P16220(1) P18846(1)	yes
CPX-6414	bZIP transcription factor complex, ATF2-BATF3	2	P15336(1) Q9NR55(1)	yes
CPX-6416	bZIP transcription factor complex, ATF2-FOS	2	P01100(1) P15336(1)	yes
CPX-6420	bZIP transcription factor complex, ATF2-JUN	2	P05412(1) P15336(1)	yes
CPX-6421	bZIP transcription factor complex, ATF2-JUNB	2	P15336(1) P17275(1)	yes
CPX-6467	bZIP transcription factor complex, ATF3-BATF	2	P18847(1) Q16520(1)	yes

CPX-6468	bZIP transcription factor complex, ATF3-BATF3	2	P18847(1) Q9NR55(1)	yes
CPX-6469	bZIP transcription factor complex, ATF3-CEBPA	2	P18847(1) P49715(1)	yes
CPX-6471	bZIP transcription factor complex, ATF3-CEBPG	2	P17676(1) P18847(1)	yes
CPX-6477	bZIP transcription factor complex, ATF3-FOS	2	P01100(1) P18847(1)	yes
CPX-6478	bZIP transcription factor complex, ATF3-FOSL1	2	P15407(1) P18847(1)	yes
CPX-6474	bZIP transcription factor complex, ATF3-JUN	2	P05412(1) P18847(1)	yes
CPX-6476	bZIP transcription factor complex, ATF3-JUNB	2	P17275(1) P18847(1)	yes
CPX-6523	bZIP transcription factor complex, ATF4-BATF2	2	P18848(1) Q8N1L9(1)	yes
CPX-6524	bZIP transcription factor complex, ATF4-BATF3	2	P18848(1) Q9NR55(1)	yes
CPX-6525	bZIP transcription factor complex, ATF4-CEBPA	2	P18848(1) P49715(1)	yes
CPX-6527	bZIP transcription factor complex, ATF4-CEBPG	2	P18848(1) P53567(1)	yes
CPX-6563	bZIP transcription factor complex, ATF4-JUNB	2	P17275(1) P18848(1)	yes
CPX-6567	bZIP transcription factor complex, ATF4-MAFB	2	P18848(1) Q9Y5Q3(1)	yes
CPX-6585	bZIP transcription factor complex, ATF5-BATF	2	Q16520(1) Q9Y2D1(1)	yes
CPX-6586	bZIP transcription factor complex, ATF5-CEBPA	2	P49715(1) Q9Y2D1(1)	yes
CPX-6589	bZIP transcription factor complex, ATF5-CEBPE	2	Q15744(1) Q9Y2D1(1)	yes
CPX-6588	bZIP transcription factor complex, ATF5-CEBPG	2	P53567(1) Q9Y2D1(1)	yes
CPX-7006	bZIP transcription factor complex, BATF-CEBPA	2	P49715(1) Q16520(1)	yes
CPX-7010	bZIP transcription factor complex, BATF-CEBPE	2	Q15744(1) Q16520(1)	yes
CPX-7008	bZIP transcription factor complex, BATF-CEBPG	2	P53567(1) Q16520(1)	yes

CPX-7011	bZIP transcription factor complex, BATF-HLF	2	Q16520(1) Q16534(1)	yes
CPX-7003	bZIP transcription factor complex, BATF-JUNB	2	P17275(1) Q16520(1)	yes
CPX-7017	bZIP transcription factor complex, BATF-NFIL3	2	Q16520(1) Q16649(1)	yes
CPX-7063	bZIP transcription factor complex, BATF2-JUN	2	P05412(1) Q8N1L9(1)	yes
CPX-7061	bZIP transcription factor complex, BATF2-JUNB	2	P17275(1) Q8N1L9(1)	yes
CPX-7095	bZIP transcription factor complex, BATF3-CEBPA	2	P49715(1) Q9NR55(1)	yes
CPX-7097	bZIP transcription factor complex, BATF3-CEBPG	2	P53567(1) Q9NR55(1)	yes
CPX-7100	bZIP transcription factor complex, BATF3-JUN	2	P05412(1) Q9NR55(1)	yes
CPX-7101	bZIP transcription factor complex, BATF3-JUNB	2	P17275(1) Q9NR55(1)	yes
CPX-486	bZIP transcription factor complex, FOS-JUN	2	P01100(1) P05412(1)	yes
CPX-504	c-Myb-C/EBPbeta complex	2	P10242(1) P17676(2)	yes
CPX-1956	CCAAT-binding factor complex	3	P23511(1) P25208(1) Q13952(1)	yes
CPX-3229	CLOCK-BMAL1 transcription complex	2	O00327(1) O15516(1)	yes
CPX-3230	CLOCK-BMAL2 transcription complex	2	O15516(1) Q8WYA1(1)	yes
CPX-5156	ERalpha-NCOA2 activated estrogen receptor complex	2	P03372(2) Q15596(2)	yes
CPX-1123	FOXO3-MYC complex	2	O43524(-) P01106(-)	yes
CPX-6016	ISGF3 complex	3	P42224(-) P52630(-) Q00978(-)	yes
CPX-5834	NF-kappaB DNA-binding transcription factor complex, p65/c-Rel	2	Q04206(1) Q04864(1)	yes
CPX-517	PXR-NCOA1 activated nuclear receptor complex	2	O75469(2) Q15788(2)	yes
CPX-496	RXRalpha-PXR nuclear receptor complex	2	O75469(2) P19793(2)	yes

CPX-508	RXRalpha-RARalpha retinoic acid receptor complex	2	P10276(1) P19793(1)	yes
CPX-654	RXRalpha-TRbeta nuclear hormone receptor complex	2	P10828(1) P19793(1)	yes
CPX-54	SMAD1-SMAD4 complex	2	Q13485(1) Q15797(2)	yes
CPX-3252	SMAD3-SMAD4 complex	2	P84022(2) Q13485(1)	yes
CPX-6041	STAT1/STAT3 complex	2	P40763(1) P42224(1)	yes
CPX-6042	STAT1/STAT4 complex	2	P42224(1) Q14765(1)	yes
CPX-6043	STAT3/STAT5A complex	2	P40763(1) P42229(1)	yes
CPX-6044	STAT3/STAT5B complex	2	P40763(1) P51692(1)	yes
CPX-91	Transcriptional activator Myc-Max complex	2	P01106(1) P61244(1)	yes
CPX-104	Transcriptional repressor Mad-Max complex	2	P61244(1) Q05195(1)	yes
CPX-3079	USF1-USF2 upstream stimulatory factor complex	2	P22415(1) Q15853(1)	yes
CPX-6419	bZIP transcription factor complex, ATF2-JDP2	2	P15336(1) Q8WYK2(1)	yes
CPX-6422	bZIP transcription factor complex, ATF2-JUND	2	P15336(1) P17535(1)	yes
CPX-6595	bZIP transcription factor complex, ATF6-ATF6B	2	P18850(1) Q99941(1)	yes
CPX-6600	bZIP transcription factor complex, ATF6B-XBP1	2	P17861(1) Q99941(1)	yes
CPX-2500	bZIP transcription factor complex, BACH1-MAF	2	O14867(1) O75444(1)	yes
CPX-7165	bZIP transcription factor complex, BACH1-MAFF	2	O14867(1) Q9ULX9(1)	yes
CPX-2872	bZIP transcription factor complex, BACH1-MAFG	2	O14867(1) O15525(1)	yes
CPX-2493	bZIP transcription factor complex, BACH1-MAFK	2	O14867(1) O60675(1)	yes
CPX-2483	bZIP transcription factor complex, BACH2-MAF	2	O75444(1) Q9BYV9(1)	yes
CPX-2484	bZIP transcription factor complex, BACH2-MAFF	2	Q9BYV9(1) Q9ULX9(1)	yes
CPX-2482	bZIP transcription factor complex, BACH2-MAFK	2	O60675(1) Q9BYV9(1)	yes

CPX-7005	bZIP transcription factor complex, BATF-JUN	2	P05412(1) Q16520(1)	yes
CPX-7102	bZIP transcription factor complex, BATF3-JUND	2	P17535(1) Q9NR55(1)	yes
CPX-509	bZIP transcription factor complex, CEBPA-CEBPB	2	P17676(1) P49715(1)	yes
CPX-1971	E2F1-DP1 transcription factor complex	2	Q01094(1) Q14186(1)	yes
CPX-1972	E2F2-DP1 transcription factor complex	2	Q14186(1) Q14209(1)	yes
CPX-711	PPARgamma-NCOA1 activated nuclear receptor complex	2	P37231(1) Q15788(1)	yes
CPX-702	PPARgamma-NCOA2 activated nuclear receptor complex	2	P37231(1) Q15596(1)	yes
CPX-525	RARalpha-NCOA1 activated retinoic acid receptor complex	2	P10276(2) Q15788(2)	yes
CPX-666	RARalpha-NCOA2 activated retinoic acid receptor complex	2	P10276(2) Q15596(2)	yes
CPX-632	RXRalpha-LXRalpha nuclear hormone receptor complex	2	P19793(1) Q13133(1)	yes
CPX-678	RXRalpha-LXRbeta nuclear hormone receptor complex	2	P19793(1) P55055(1)	yes
CPX-513	RXRalpha-NCOA2 activated retinoic acid receptor complex	2	P19793(2) Q15596(2)	yes
CPX-816	RXRalpha-RARalpha-NCOA2 retinoic acid receptor complex	3	P10276(1) P19793(1) Q15596(-)	yes
CPX-631	RXRalpha-VDR nuclear hormone receptor complex	2	P11473(1) P19793(1)	yes
CPX-716	RXRbeta-LXRalpha nuclear hormone receptor complex	2	P28702(1) Q13133(1)	yes
CPX-652	RXRbeta-LXRbeta nuclear hormone receptor complex	2	P28702(1) P55055(1)	yes
CPX-871	RXRbeta-VDR nuclear hormone receptor complex	2	P11473(1) P28702(1)	yes
CPX-6062	SMAD3-TTF-1 complex	2	P43699(-) P84022(-)	yes
CPX-2497	bZIP transcription factor complex, BACH1-MAFB	2	O14867(1) Q9Y5Q3(1)	uncertain

CPX-6407	bZIP transcription factor complex, ATF2-ATF3	2	P15336(1) P18847(1)	uncertain
CPX-6408	bZIP transcription factor complex, ATF2-ATF4	2	P15336(1) P18848(1)	uncertain
CPX-6415	bZIP transcription factor complex, ATF2-DDIT3	2	P15336(1) P35638(1)	uncertain
CPX-6417	bZIP transcription factor complex, ATF2-FOSL1	2	P15336(1) P15407(1)	uncertain
CPX-6385	bZIP transcription factor complex, ATF3-ATF4	2	P18847(1) P18848(1)	uncertain
CPX-6472	bZIP transcription factor complex, ATF3-CEBPE	2	P18847(1) Q15744(1)	uncertain
CPX-6473	bZIP transcription factor complex, ATF3-DDIT3	2	P18847(1) P35638(1)	uncertain
CPX-6542	bZIP transcription factor complex, ATF4-CREBZF	2	P18848(1) Q9NS37(1)	uncertain
CPX-6543	bZIP transcription factor complex, ATF4-DDIT3	2	P18848(1) P35638(1)	uncertain
CPX-6564	bZIP transcription factor complex, ATF4-FOS	2	P01100(1) P18848(1)	uncertain
CPX-6565	bZIP transcription factor complex, ATF4-FOSL1	2	P15407(1) P18848(1)	uncertain
CPX-6562	bZIP transcription factor complex, ATF4-JUN	2	P05412(1) P18848(1)	uncertain
CPX-6597	bZIP transcription factor complex, ATF6-XBP1	2	P17861(1) P18850(1)	uncertain
CPX-2485	bZIP transcription factor complex, BACH2-MAFG	2	O15525(1) Q9BYV9(1)	uncertain
CPX-7014	bZIP transcription factor complex, BATF-DBP	2	Q10586(1) Q16520(1)	uncertain
CPX-7004	bZIP transcription factor complex, BATF-DDIT3	2	P35638(1) Q16520(1)	uncertain
CPX-7108	bZIP transcription factor complex, BATF3-DBP	2	Q10586(1) Q9NR55(1)	uncertain
CPX-7106	bZIP transcription factor complex, BATF3-DDIT3	2	P35638(1) Q9NR55(1)	uncertain
CPX-69	bZIP transcription factor complex, CEBPA-DDIT3	2	P35638(1) P49715(1)	uncertain
CPX-70	bZIP transcription factor complex, CEBPB-DDIT3	2	P17676(1) P35638(1)	uncertain

CPX-6047	STAT2/STAT6 complex	2	P42226(1) P52630(1)	uncertain
CPX-6046	STAT3/STAT4 complex	2	P40763(1) Q14765(1)	uncertain
CPX-6045	STAT5A/STAT5B complex	2	P42229(1) P51692(1)	uncertain
CPX-480	AP-1 transcription factor complex FOS-JUN-NFATC2	3	P01100(1) P05412(1) Q13469(1)	uncertain
CPX-9	bZIP transcription factor complex, ATF1-ATF4	2	P18846(1) P18848(1)	uncertain
CPX-6402	bZIP transcription factor complex, ATF1-BACH1	2	O14867(1) P18846(1)	uncertain
CPX-6404	bZIP transcription factor complex, ATF1-NFIL3	2	P18846(1) Q16649(1)	uncertain
CPX-6409	bZIP transcription factor complex, ATF2-ATF7	2	P15336(1) P17544(1)	uncertain
CPX-6412	bZIP transcription factor complex, ATF2-BACH1	2	O14867(1) P15336(1)	uncertain
CPX-6413	bZIP transcription factor complex, ATF2-BATF	2	P15336(1) Q16520(1)	uncertain
CPX-6418	bZIP transcription factor complex, ATF2-FOSL2	2	P15336(1) P15408(1)	uncertain
CPX-6466	bZIP transcription factor complex, ATF3-ATF7	2	P17544(1) P18847(1)	uncertain
CPX-6470	bZIP transcription factor complex, ATF3-CEBPB	2	P17676(1) P18847(1)	uncertain
CPX-6479	bZIP transcription factor complex, ATF3-FOSL2	2	P15408(1) P18847(1)	uncertain
CPX-6480	bZIP transcription factor complex, ATF3-MAFF	2	P18847(1) Q9ULX9(1)	uncertain
CPX-6481	bZIP transcription factor complex, ATF3-MAFG	2	O15525(1) P18847(1)	uncertain
CPX-6522	bZIP transcription factor complex, ATF4-BATF	2	P18848(1) Q16520(1)	uncertain
CPX-6526	bZIP transcription factor complex, ATF4-CEBPB	2	P17676(1) P18848(1)	uncertain
CPX-6528	bZIP transcription factor complex, ATF4-CEBPD	2	P18848(1) P49716(1)	uncertain
CPX-6529	bZIP transcription factor complex, ATF4-CEBPE	2	P18848(1) Q15744(1)	uncertain
CPX-8	bZIP transcription factor complex, ATF4-CREB1	2	P16220(1) P18848(1)	uncertain

CPX-6541	bZIP transcription factor complex, ATF4-CREB3	2	O43889(1) P18848(1)	uncertain
CPX-6566	bZIP transcription factor complex, ATF4-MAF	2	O75444(1) P18848(1)	uncertain
CPX-6568	bZIP transcription factor complex, ATF4-NFE2	2	P18848(1) Q16621(1)	uncertain
CPX-6570	bZIP transcription factor complex, ATF4-NFE2L2	2	P18848(1) Q16236(1)	uncertain
CPX-6572	bZIP transcription factor complex, ATF4-NFE2L3	2	P18848(1) Q9Y4A8(1)	uncertain
CPX-6601	bZIP transcription factor complex, ATF6B-CREBZF	2	Q99941(1) Q9NS37(1)	uncertain
CPX-6781	bZIP transcription factor complex, ATF7-BACH1	2	O14867(1) P17544(1)	uncertain
CPX-6782	bZIP transcription factor complex, ATF7-CEBPG	2	P17544(1) P53567(1)	uncertain
CPX-6784	bZIP transcription factor complex, ATF7-DDIT3	2	P17544(1) P35638(1)	uncertain
CPX-6783	bZIP transcription factor complex, ATF7-FOS	2	P01100(1) P17544(1)	uncertain
CPX-6785	bZIP transcription factor complex, ATF7-FOSL2	2	P15408(1) P17544(1)	uncertain
CPX-6786	bZIP transcription factor complex, ATF7-JUN	2	P05412(1) P17544(1)	uncertain
CPX-6787	bZIP transcription factor complex, ATF7-JUNB	2	P17275(1) P17544(1)	uncertain
CPX-6788	bZIP transcription factor complex, ATF7-JUND	2	P17535(1) P17544(1)	uncertain
CPX-6789	bZIP transcription factor complex, ATF7-NFE2	2	P17544(1) Q16621(1)	uncertain
CPX-7012	bZIP transcription factor complex, BACH1-BATF	2	O14867(1) Q16520(1)	uncertain
CPX-2494	bZIP transcription factor complex, BACH1-CREB1	2	O14867(1) P16220(1)	uncertain
CPX-2496	bZIP transcription factor complex, BACH1-DDIT3	2	O14867(1) P35638(1)	uncertain
CPX-2491	bZIP transcription factor complex, BACH1-FOS	2	O14867(1) P01100(1)	uncertain
CPX-7093	bZIP transcription factor complex, BACH2-BATF3	2	Q9BYV9(1) Q9NR55(1)	uncertain

CPX-2479	bZIP transcription factor complex, BACH2-MAFB	2	Q9BYV9(1) Q9Y5Q3(1)	uncertain
CPX-2471	bZIP transcription factor complex, BACH2-NFE2L3	2	Q9BYV9(1) Q9Y4A8(1)	uncertain
CPX-7018	bZIP transcription factor complex, BATF-BATF3	2	Q16520(1) Q9NR55(1)	uncertain
CPX-7007	bZIP transcription factor complex, BATF-CEBPB	2	P17676(1) Q16520(1)	uncertain
CPX-7009	bZIP transcription factor complex, BATF-CEBPD	2	P49716(1) Q16520(1)	uncertain
CPX-7013	bZIP transcription factor complex, BATF-JUND	2	P17535(1) Q16520(1)	uncertain
CPX-7065	bZIP transcription factor complex, BATF2-CEBPA	2	P49715(1) Q8N1L9(1)	uncertain
CPX-7067	bZIP transcription factor complex, BATF2-CEBPE	2	Q15744(1) Q8N1L9(1)	uncertain
CPX-7066	bZIP transcription factor complex, BATF2-CEBPG	2	P53567(1) Q8N1L9(1)	uncertain
CPX-7068	bZIP transcription factor complex, BATF2-DBP	2	Q10586(1) Q8N1L9(1)	uncertain
CPX-7064	bZIP transcription factor complex, BATF2-DDIT3	2	P35638(1) Q8N1L9(1)	uncertain
CPX-7081	bZIP transcription factor complex, BATF2-HLF	2	Q16534(1) Q8N1L9(1)	uncertain
CPX-7085	bZIP transcription factor complex, BATF2-MAFF	2	Q8N1L9(1) Q9ULX9(1)	uncertain
CPX-7096	bZIP transcription factor complex, BATF3-CEBPB	2	P17676(1) Q9NR55(1)	uncertain
CPX-7098	bZIP transcription factor complex, BATF3-CEBPD	2	P49716(1) Q9NR55(1)	uncertain
CPX-7099	bZIP transcription factor complex, BATF3-CEBPE	2	Q15744(1) Q9NR55(1)	uncertain
CPX-7109	bZIP transcription factor complex, BATF3-CREB3	2	O43889(1) Q9NR55(1)	uncertain
CPX-7107	bZIP transcription factor complex, BATF3-HLF	2	Q16534(1) Q9NR55(1)	uncertain
CPX-7103	bZIP transcription factor complex, BATF3-MAFF	2	Q9NR55(1) Q9ULX9(1)	uncertain
CPX-7105	bZIP transcription factor complex, BATF3-MAFG	2	O15525(1) Q9NR55(1)	uncertain

CPX-5342	RXRalpha-NCOA1 activated retinoic acid receptor complex	2	P19793(2) Q15788(2)	uncertain
----------	--	---	------------------------	-----------

Table F.2: Classification of transcription factors in *H. sapiens*. The table provides a list of 74 TFs in *H. sapiens* that belong to the basic helix-loop-helix (bHLH) or basic leucine zipper (bZIP) classes. The classes of the TFs are determined using the JASPAR database. The column “UniProt ID” provides the UniProt IDs for each of the 74 TFs, whereas “TF name” gives the name of the TF.

UniProt ID	TF name	Class
Q9HBZ2	ARNT2	Basic helix-loop-helix factors (bHLH)
P50553	ASCL1	Basic helix-loop-helix factors (bHLH)
Q92858	ATOH1	Basic helix-loop-helix factors (bHLH)
Q8N100	ATOH7	Basic helix-loop-helix factors (bHLH)
O15516	CLOCK	Basic helix-loop-helix factors (bHLH)
Q6QHK4	FIGLA	Basic helix-loop-helix factors (bHLH)
P61296	HAND2	Basic helix-loop-helix factors (bHLH)
Q14469	HES1	Basic helix-loop-helix factors (bHLH)
Q9Y543	HES2	Basic helix-loop-helix factors (bHLH)
Q5TA89	HES5	Basic helix-loop-helix factors (bHLH)
Q96HZ4	HES6	Basic helix-loop-helix factors (bHLH)
Q9BYE0	HES7	Basic helix-loop-helix factors (bHLH)
Q9Y5J3	HEY1	Basic helix-loop-helix factors (bHLH)
Q9UBP5	HEY2	Basic helix-loop-helix factors (bHLH)
Q16665	HIF1A	Basic helix-loop-helix factors (bHLH)
P61244	MAX	Basic helix-loop-helix factors (bHLH)
O75030	MITF	Basic helix-loop-helix factors (bHLH)
Q9UH92	MLX	Basic helix-loop-helix factors (bHLH)
Q99583	MNT	Basic helix-loop-helix factors (bHLH)
A6NI15	MSGN1	Basic helix-loop-helix factors (bHLH)
P50539	MXI1	Basic helix-loop-helix factors (bHLH)
P01106	MYC	Basic helix-loop-helix factors (bHLH)
P04198	MYCN	Basic helix-loop-helix factors (bHLH)
P13349	MYF5	Basic helix-loop-helix factors (bHLH)
P23409	MYF6	Basic helix-loop-helix factors (bHLH)
P15172	MYOD1	Basic helix-loop-helix factors (bHLH)

P15173	MYOG	Basic helix-loop-helix factors (bHLH)
Q8TAK6	OLIG1	Basic helix-loop-helix factors (bHLH)
Q13516	OLIG2	Basic helix-loop-helix factors (bHLH)
Q7RTU3	OLIG3	Basic helix-loop-helix factors (bHLH)
O43680	TCF21	Basic helix-loop-helix factors (bHLH)
Q9UL49	TCFL5	Basic helix-loop-helix factors (bHLH)
Q01664	TFAP4	Basic helix-loop-helix factors (bHLH)
P19532	TFE3	Basic helix-loop-helix factors (bHLH)
P19484	TFEB	Basic helix-loop-helix factors (bHLH)
O14948	TFEC	Basic helix-loop-helix factors (bHLH)
P22415	USF1	Basic helix-loop-helix factors (bHLH)
Q15853	USF2	Basic helix-loop-helix factors (bHLH)
P15336	ATF2	Basic leucine zipper factors (bZIP)
P18847	ATF3	Basic leucine zipper factors (bZIP)
P18848	ATF4	Basic leucine zipper factors (bZIP)
P17544	ATF7	Basic leucine zipper factors (bZIP)
O14867	BACH1	Basic leucine zipper factors (bZIP)
Q9BYV9	BACH2	Basic leucine zipper factors (bZIP)
Q16520	BATF	Basic leucine zipper factors (bZIP)
Q9NR55	BATF3	Basic leucine zipper factors (bZIP)
P49715	CEBPA	Basic leucine zipper factors (bZIP)
P17676	CEBPB	Basic leucine zipper factors (bZIP)
P49716	CEBPD	Basic leucine zipper factors (bZIP)
Q15744	CEBPE	Basic leucine zipper factors (bZIP)
P53567	CEBPG	Basic leucine zipper factors (bZIP)
P16220	CREB1	Basic leucine zipper factors (bZIP)
O43889	CREB3	Basic leucine zipper factors (bZIP)
Q03060	CREM	Basic leucine zipper factors (bZIP)
Q10586	DBP	Basic leucine zipper factors (bZIP)
P01100	FOS	Basic leucine zipper factors (bZIP)
P15407	FOSL1	Basic leucine zipper factors (bZIP)
P15408	FOSL2	Basic leucine zipper factors (bZIP)
Q16534	HLF	Basic leucine zipper factors (bZIP)
Q8WYK2	JDP2	Basic leucine zipper factors (bZIP)

P05412	JUN	Basic leucine zipper factors (bZIP)
P17275	JUNB	Basic leucine zipper factors (bZIP)
P17535	JUND	Basic leucine zipper factors (bZIP)
O75444	MAF	Basic leucine zipper factors (bZIP)
Q8NHW3	MAFA	Basic leucine zipper factors (bZIP)
Q9ULX9	MAFF	Basic leucine zipper factors (bZIP)
O15525	MAFG	Basic leucine zipper factors (bZIP)
O60675	MAFK	Basic leucine zipper factors (bZIP)
Q16621	NFE2	Basic leucine zipper factors (bZIP)
Q16649	NFIL3	Basic leucine zipper factors (bZIP)
P54845	NRL	Basic leucine zipper factors (bZIP)
Q10587	TEF	Basic leucine zipper factors (bZIP)
P17861	XBP1	Basic leucine zipper factors (bZIP)
Q16656	NRF1	Basic leucine zipper factors (bZIP)

Table F.3: A list of 17 complexes from the set of 617 complexes in *S. cerevisiae* such that all the protein subunits of these complexes are transcription factors. “Complex ID” and “Complex name” are the identifier and names of complexes of *S. cerevisiae* as given in the EBI Complex Portal database. “Size of the complex” is the number of protein subunits the complex is constituted of. “Uniprot ID of protein subunits” gives the Uniprot IDs of the subunits (TFs obtained from the Yeastract database) in a complex along with their stoichiometric coefficients as integers within brackets. Unknown stoichiometric coefficients are marked as “(–)”. “Is a TR” column is “yes” if the complex acts as a transcriptional regulator (TR) based on manual literature curation. “TR binding and expression evidence” is “yes” if for a given complex, all subunits show evidence for both TR binding and effects on expression, otherwise, the entry is “no”.

Complex ID	Complex name	Size of the complex	Uniprot ID of protein subunits	Is a TR	TR binding and expression evidence
CPX-575	Ste12/Dig1/Dig2 transcription regulation complex	3	P13574(–) Q03063(–) Q03373(–)	uncertain	no
CPX-576	Tec1/Ste12/Dig1 transcription regulation complex	3	P13574(–) P18412(–) Q03063(–)	yes	no
CPX-828	RTG transcription factor complex	2	P32607(1) P38165(1)	yes	yes
CPX-946	SBF transcription complex	2	P09959(1) P25302(1)	yes	no
CPX-950	MBP transcription complex	2	P09959(–) P39678(–)	yes	no
CPX-999	MET4-MET28-MET31 sulfur metabolism transcription factor complex	3	P32389(–) P40573(–) Q03081(–)	yes	yes
CPX-1015	MET4-MET28-MET32 sulfur metabolism transcription factor complex	3	P32389(–) P40573(–) Q12041(–)	yes	yes
CPX-1016	CBF1-MET4-MET28 sulfur metabolism transcription factor complex	3	P17106(2) P32389(–) P40573(–)	yes	yes
CPX-1038	PIP2-OAF1 transcription factor complex	2	P39720(1) P52960(1)	yes	yes
CPX-1042	GAL3-GAL80 transcription regulation complex	2	P04387(2) P13045(2)	yes	no
CPX-1044	GAL4-GAL80 transcription repressor complex	2	P04386(2) P04387(2)	yes	yes
CPX-1200	RAP1-GCR1 transcription activation complex	2	P07261(2) P11938(–)	yes	yes
CPX-1229	RAP1-GCR1-GCR2 transcription activation complex	3	P07261(2) P11938(–) Q01722(2)	yes	yes
CPX-1277	INO2-INO4 transcription activation complex	2	P13902(–) P26798(–)	yes	yes
CPX-1415	IME1-UME6 transcription activation complex	2	P21190(–) P39001(–)	yes	yes
CPX-1663	CYP8-TUP1 corepressor complex	2	P14922(1) P16649(4)	corepressor	no
CPX-1830	CCAAT-binding factor complex	4	P06774(1) P13434(1) P14064(1) Q02516(1)	yes	yes

Table F.4: Classification of transcription factors in *S. cerevisiae*. The table provides a list of 17 TFs from the Yeastract database, that belong to the basic helix-loop-helix (bHLH) or basic leucine zipper (bZIP) classes. The classes of the TFs are determined using the JASPAR database. The column “UniProt ID” provides the UniProt IDs for each of the 17 TFs, whereas “TF name” corresponds to the name of the TF. All the 17 TFs shown in the table display evidence for DNA binding as well as effects on expression.

UniProt ID	TF name	DNA binding and expression evidence	Class
P33122	TYE7	Yes	Basic helix-loop-helix factors (bHLH)
P38165	RTG3	Yes	Basic helix-loop-helix factors (bHLH)
P13902	INO4	Yes	Basic helix-loop-helix factors (bHLH)
P26798	INO2	Yes	Basic helix-loop-helix factors (bHLH)
P07270	PHO4	Yes	Basic helix-loop-helix factors (bHLH)
P17106	CBF1	Yes	Basic helix-loop-helix factors (bHLH)
P14164	ABF1	Yes	Basic helix-loop-helix factors (bHLH)
P32389	MET4	Yes	Basic leucine zipper factors (bZIP)
P40573	MET28	Yes	Basic leucine zipper factors (bZIP)
P41546	HAC1	Yes	Basic leucine zipper factors (bZIP)
P03069	GCN4	Yes	Basic leucine zipper factors (bZIP)
Q02100	SKO1	Yes	Basic leucine zipper factors (bZIP)
Q06596	ARR1	Yes	Basic leucine zipper factors (bZIP)
Q08182	YAP7	Yes	Basic leucine zipper factors (bZIP)
Q03935	YAP6	Yes	Basic leucine zipper factors (bZIP)
P40574	YAP5	Yes	Basic leucine zipper factors (bZIP)
P38749	YAP3	Yes	Basic leucine zipper factors (bZIP)

Table F.5: Comparison of the fractions of four types of biologically meaningful BFs and the fraction of the BFs allowed by the most restrictive composition structures for $k \leq 5$ inputs. The fractions are computed with respect to all possible BFs for $k \leq 5$ inputs. The four types of biologically meaningful BFs include unate functions (UFs), canalyzing functions (CFs), nested canalyzing functions (NCFs) and read-once functions (RoFs). The column “Composed BF” represents BFs contained in the most restrictive composition structure for a given k . The most restrictive composition structure is the composition structure that has the least number of BFs when compared to other composition structures with the same number of inputs k . For $k = 1$ and 2, there are no restrictions in possible BFs due to composition structure since only trivial composition structures such as $\{1\}$, $\{1, 1\}$ and $\{2\}$ exist. For $k = 3, 4$ and 5, the most restrictive composition structures are $\{1, 2\}$, $\{2, 2\}$ and $\{2, 3\}$, respectively.

k	Fraction of				
	UF	CF	NCF	RoF	Composed BF
1	1	1	0.5	0.5	1
2	0.875	0.875	0.5	0.5	1
3	0.406	0.469	0.25	0.25	0.594
4	0.033	0.054	0.011	0.013	0.018
5	5.37×10^{-5}	3.01×10^{-4}	2.47×10^{-6}	3.52×10^{-6}	1.67×10^{-5}

Table F.6: Fraction of BFs in different composition structures that display biologically meaningful properties. The fraction of BFs in different non-trivial composition structures that also belong to each of the four types of biologically meaningful BFs, namely unate functions (UFs), canalyzing functions (CFs), nested canalyzing functions (NCFs) and read-once functions (RoFs). The fraction is computed with respect to all BFs allowed by a composition structure.

Composition structure	Fraction of biologically meaningful BFs in composition structure			
	UF	CF	NCF	RoF
$\{1, 2\}$	0.632	0.789	0.421	0.421
$\{1, 3\}$	0.249	0.722	0.151	0.151
$\{2, 2\}$	0.525	0.604	0.185	0.265
$\{1, 1, 2\}$	0.220	0.298	0.118	0.134
$\{1, 4\}$	0.022	0.672	0.006	0.007
$\{2, 3\}$	0.191	0.469	0.046	0.095
$\{1, 1, 3\}$	0.099	0.307	0.040	0.054
$\{1, 2, 2\}$	0.177	0.250	0.055	0.099
$\{1, 1, 1, 2\}$	0.018	0.036	0.003	0.004

Table F.7: Number and fraction of BFs with odd bias in different composition structures. The fractions of BFs with odd bias in a composition structure are computed with respect to all allowed BFs in the composition structure. The column “Number of composed BFs” gives the number of allowed BFs in a composition structure. The column “Odd biases present” gives the list of odd biases of BFs that are present in a composition structure. Note that the table only gives data for non-trivial composition structures with $k \leq 5$ inputs.

Composition structure	Number of composed BFs	BFs with odd bias in composition structure		
		Number	Fraction	Odd biases present
{1,2}	152	64	0.421	1,3
{1,3}	4864	1760	0.361	1,3,5,7
{2,2}	1208	320	0.264	1,3,7
{1,1,2}	6216	2368	0.381	1,3,5,7
{1,4}	1921928	646144	0.336	1,3,5,7,9,11,13,15
{2,3}	71608	17024	0.238	1,3,5,7,9,11,15
{1,1,3}	263488	75584	0.287	1,3,5,7,9,11,13,15
{1,2,2}	100768	25344	0.252	1,3,5,7,9,11,13,15
{1,1,1,2}	3446488	1266944	0.368	1,3,5,7,9,11,13,15

Table F.8: Enrichment of composed BFs in the reference biological dataset. The enrichment factors for composed BFs in different non-trivial composition structures with number of inputs $k \leq 5$ and the associated one-sided p -values.

Composition structure	Enrichment factor	p -value
{1,2}	1.63	6.15×10^{-72}
{1,3}	12.48	4.33×10^{-245}
{2,2}	40.37	3.90×10^{-274}
{1,1,2}	10.30	7.57×10^{-250}
{1,4}	1948.23	0
{2,3}	45760.08	0
{1,1,3}	14732.62	0
{1,2,2}	36887.66	0
{1,1,1,2}	1158.31	0

Table F.9: p -values corresponding to the relative enrichment values of biologically meaningful BF's within different composition structures. The p -values corresponding to the relative enrichment values E_R for the four biologically meaningful sub-types of BF's within different composition structures with number of inputs $k \leq 5$. The four biologically meaningful sub-types within composed BF's include those BF's in a composition structure that also happen to be unate functions (UF's), canalyzing functions (CF's), nested canalyzing functions (NCF's) and read-once functions (RoF's).

Composition structure	p -values corresponding to E_R of sub-types in composition structure			
	UF	CF	NCF	RoF
$\{1,2\}$	0	0	1.79×10^{-115}	1.79×10^{-115}
$\{1,3\}$	0	0	2.27×10^{-176}	2.27×10^{-176}
$\{2,2\}$	1.75×10^{-54}	1.98×10^{-24}	5.61×10^{-100}	3.74×10^{-96}
$\{1,1,2\}$	3.08×10^{-166}	1.30×10^{-105}	1.42×10^{-185}	4.53×10^{-202}
$\{1,4\}$	5.19×10^{-227}	0	2.26×10^{-254}	3.13×10^{-258}
$\{2,3\}$	0	1.74×10^{-27}	1.62×10^{-111}	6.60×10^{-105}
$\{1,1,3\}$	0	1.26×10^{-57}	7.92×10^{-146}	8.03×10^{-160}
$\{1,2,2\}$	0	4.12×10^{-64}	7.77×10^{-123}	2.05×10^{-118}
$\{1,1,1,2\}$	0	1.32×10^{-181}	2.18×10^{-277}	9.30×10^{-301}

Table F.10: p -values for comparison between the enrichments of composed BFs and biologically meaningful BFs with minimum complexity in the reference biological dataset. T_C denotes the set of composed BFs allowed by a composition structure at a given number of inputs k , T_{NCF} denotes the set of all k -input nested canalizing functions (NCFs), and T_{RoF} denotes the set of all k -input read-once functions (RoFs). “ \cap ” represents the intersection of two sets and “ \setminus ” represents the set-theoretic difference. “ $-$ ” in the columns $T_{NCF} \setminus T_C$ or $T_{RoF} \setminus T_C$ indicates that the NCFs or RoFs are a subset of the set of BFs allowed by the composition structure.

Composition structure	$T_C \cap T_{NCF}$	$T_C \setminus T_{NCF}$	$T_{NCF} \setminus T_C$	$T_C \cap T_{RoF}$	$T_C \setminus T_{RoF}$	$T_{RoF} \setminus T_C$
$\{1,2\}$	3.09×10^{-184}	1	—	3.09×10^{-184}	1	—
$\{1,3\}$	0	0.966	—	0	0.966	1.66×10^{-19}
$\{2,2\}$	0	1.59×10^{-11}	7.07×10^{-70}	0	0.009	7.07×10^{-70}
$\{1,1,2\}$	0	0.406	—	0	0.999	—
$\{1,4\}$	0	2.17×10^{-35}	—	0	7.77×10^{-26}	1.06×10^{-47}
$\{2,3\}$	0	9.10×10^{-80}	4.44×10^{-106}	0	4.83×10^{-30}	8.09×10^{-105}
$\{1,1,3\}$	0	2.87×10^{-67}	—	0	8.85×10^{-25}	3.43×10^{-05}
$\{1,2,2\}$	0	1.36×10^{-76}	1.31×10^{-30}	0	1.00×10^{-28}	1.31×10^{-30}
$\{1,1,1,2\}$	0	7.82×10^{-52}	—	0	1.51×10^{-22}	—

Table F.11: Comparison between the enrichments of composed BFs and biologically meaningful BFs in the reference biological dataset. The table provides the enrichment factors of composed BFs allowed by non-trivial composition structures with $k \leq 5$ inputs and of two biologically meaningful BFs namely unate functions (UFs) and canalizing functions (CFs). T_C denotes the set of composed BFs allowed by a composition structure at a given number of inputs k , whereas T_{UF} and T_{CF} denote the set of all k -input UFs and CFs, respectively. “ \cap ” represents the intersection of two sets and “ \setminus ” represents the set-theoretic difference. “–” in the column $T_{CF} \setminus T_C$ indicates that the CFs are a subset of the set of BFs allowed by the corresponding composition structure.

Composition structure	$T_C \cap T_{UF}$	$T_C \setminus T_{UF}$	$T_{UF} \setminus T_C$	$T_C \cap T_{CF}$	$T_C \setminus T_{CF}$	$T_{CF} \setminus T_C$
$\{1,2\}$	2.58	0.00	1.01	2.06	0.00	–
$\{1,3\}$	50.17	0.00	4.76	17.28	0.00	–
$\{2,2\}$	76.53	0.44	10.91	61.59	7.97	5.66
$\{1,1,2\}$	46.54	0.05	1.91	32.54	0.87	0.31
$\{1,4\}$	89183.19	14.64	2623.97	2897.47	0.00	–
$\{2,3\}$	239564.87	0.00	4316.45	90144.74	6518.64	568.71
$\{1,1,3\}$	148416.79	0.00	1616.48	44868.92	1357.79	90.92
$\{1,2,2\}$	208387.45	0.00	2329.87	136371.86	3645.06	239.02
$\{1,1,1,2\}$	64895.59	0.00	1303.10	30112.53	91.11	47.07

Table F.12: p -values for comparison between the enrichments of composed BFs and biologically meaningful BFs in the reference biological dataset. The table provides the enrichment factors of composed BFs allowed by non-trivial composition structures with $k \leq 5$ inputs and of two biologically meaningful BFs namely, unate functions (UFs) and canalizing functions (CFs). T_C denotes the set of composed BFs allowed by a composition structure at a given number of inputs k , whereas T_{UF} and T_{CF} denote the set of all k -input UFs and CFs, respectively. “ \cap ” represents the intersection of two sets and “ \setminus ” represents the set-theoretic difference. “–” in the column $T_{CF} \setminus T_C$ indicates that the CFs are a subset of the set of BFs allowed by the corresponding composition structure.

Composition structure	$T_C \cap T_{UF}$	$T_C \setminus T_{UF}$	$T_{UF} \setminus T_C$	$T_C \cap T_{CF}$	$T_C \setminus T_{CF}$	$T_{CF} \setminus T_C$
$\{1,2\}$	3.23×10^{-149}	1	0.413	1.83×10^{-111}	1	–
$\{1,3\}$	0	0.999	1.39×10^{-08}	8.86×10^{-279}	0.995	–
$\{2,2\}$	0	0.661	5.93×10^{-49}	1.14×10^{-280}	1.39×10^{-10}	1.02×10^{-29}
$\{1,1,2\}$	0	0.999	0.041	0	0.652	0.960
$\{1,4\}$	0	0.002	2.02×10^{-59}	0	0.023	–
$\{2,3\}$	0	0.002	3.64×10^{-116}	0	5.16×10^{-36}	5.39×10^{-66}
$\{1,1,3\}$	0	0.009	1.99×10^{-38}	0	3.34×10^{-29}	1.24×10^{-09}
$\{1,2,2\}$	0	0.003	1.93×10^{-58}	0	1.16×10^{-36}	1.17×10^{-25}
$\{1,1,1,2\}$	0	0.116	2.68×10^{-26}	0	1.17×10^{-20}	1.22×10^{-05}

References

- [1] Balázsi, G., van Oudenaarden, A. & Collins, J. J. Cellular decision making and biological noise: from microbes to mammals. *Cell* **144**, 910–925 (2011).
- [2] Reményi, A., Schöler, H. R. & Wilmanns, M. Combinatorial control of gene expression. *Nature Structural and Molecular Biology* **11**, 812–815 (2004).
- [3] Spirin, V. & Mirny, L. A. Protein complexes and functional modules in molecular networks. *Proceedings of the National Academy of Sciences* **100**, 12123–12128 (2003).
- [4] Zhang, J., Maslov, S. & Shakhnovich, E. I. Constraints imposed by non-functional protein–protein interactions on gene expression and proteome size. *Molecular Systems Biology* **4**, 210 (2008).
- [5] O’Brien, J., Hayder, H., Zayed, Y. & Peng, C. Overview of MicroRNA Biogenesis, Mechanisms of Actions, and Circulation. *Frontiers in Endocrinology* **9** (2018).
- [6] Jordan, J. D., Landau, E. M. & Iyengar, R. Signaling networks: the origins of cellular multitasking. *Cell* **103**, 193–200 (2000).
- [7] Barabási, A. L. & Oltvai, Z. N. Network biology: understanding the cell’s functional organization. *Nature Reviews Genetics* **5**, 101–113 (2004).
- [8] Levine, M. & Davidson, E. H. Gene regulatory networks for development. *Proceedings of the National Academy of Sciences* **102**, 4936–4942 (2005).
- [9] Milo, R. *et al.* Network Motifs: Simple Building Blocks of Complex Networks. *Science* **298**, 824–827 (2002).

- [10] Palsson, B. Ø. *Systems Biology: Properties of Reconstructed Networks* (Cambridge University Press, 2006).
- [11] Davidson, E. H. *et al.* A Genomic Regulatory Network for Development. *Science* **295**, 1669–1678 (2002).
- [12] Förster, J., Famili, I., Fu, P., Palsson, B. Ø. & Nielsen, J. Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. *Genome research* **13**, 244–253 (2003).
- [13] Samal, A. *et al.* Low degree metabolites explain essential reactions and enhance modularity in biological networks. *BMC Bioinformatics* **7**, 1–10 (2006).
- [14] Hall, D. A. *et al.* Regulation of Gene Expression by a Metabolic Enzyme. *Science* **306**, 482–484 (2004).
- [15] Weidemüller, P., Kholmatov, M., Petsalaki, E. & Zaugg, J. B. Transcription factors: Bridge between cell signaling and gene regulation. *Proteomics* **21**, 2000034 (2021).
- [16] Alon, U. Biological Networks: The Tinkerer as an Engineer. *Science* **301**, 1866–1867 (2003).
- [17] Alon, U. *An Introduction to Systems Biology: Design Principles of Biological Circuits* (Chapman and Hall/CRC., 2006).
- [18] Tyson, J. J., Chen, K. & Novak, B. Network dynamics and cell physiology. *Nature Reviews Molecular Cell Biology* **2**, 908–916 (2001).
- [19] Bornholdt, S. Less is more in modeling large genetic networks. *Science* **310**, 449–451 (2005).
- [20] Kauffman, S. A. Metabolic Stability and Epigenesis in Randomly Constructed Genetic Nets. *Journal of Theoretical Biology* **22**, 437–467 (1969).
- [21] Kauffman, S. A. Homeostasis and Differentiation in Random Genetic Control Networks. *Nature* **224**, 177–178 (1969).

- [22] Kauffman, S. A. *The origins of order: self-organization and selection in evolution* (Oxford University Press, New York, 1993).
- [23] Huang, S. & Ingber, D. E. Shape-Dependent Control of Cell Growth, Differentiation, and Apoptosis: Switching between Attractors in Cell Regulatory Networks. *Experimental Cell Research* **261**, 91–103 (2000).
- [24] Huang, S., Eichler, G., Bar-Yam, Y. & Ingber, D. E. Cell fates as high-dimensional attractor states of a complex gene regulatory network. *Physical Review Letters* **94**, 128701 (2005).
- [25] Saadatpour, A., Albert, I. & Albert, R. Attractor analysis of asynchronous Boolean models of signal transduction networks. *Journal of Theoretical Biology* **266**, 641–656 (2010).
- [26] Albert, R. & Othmer, H. G. The topology of the regulatory interactions predicts the expression pattern of the segment polarity genes in *Drosophila melanogaster*. *Journal of Theoretical Biology* **223**, 1–18 (2003).
- [27] Davidich, M. & Bornholdt, S. The transition from differential equations to Boolean networks: a case study in simplifying a regulatory network model. *Journal of Theoretical Biology* **255**, 269–277 (2008).
- [28] Yachie-Kinoshita, A. *et al.* Modeling signaling-dependent pluripotency with Boolean logic to predict cell fate transitions. *Molecular Systems Biology* **14**, e7952 (2018).
- [29] Saadatpour, A. *et al.* Dynamical and Structural Analysis of a T Cell Survival Network Identifies Novel Candidate Therapeutic Targets for Large Granular Lymphocyte Leukemia. *PLoS Computational Biology* **7**, e1002267 (2011).
- [30] Von der Heyde, S. *et al.* Boolean ErbB network reconstructions and perturbation simulations reveal individual drug response in different breast cancer cell lines. *BMC Systems Biology* **8**, 1–22 (2014).
- [31] Buchler, N. E., Gerland, U. & Hwa, T. On schemes of combinatorial transcription logic. *Proceedings of the National Academy of Sciences* **100**, 5136–5141 (2003).

- [32] Samaga, R. & Klamt, S. Modeling approaches for qualitative and semi-quantitative analysis of cellular signaling networks. *Cell Communication and Signaling* **11**, 43 (2013).
- [33] Raeymaekers, L. Dynamics of Boolean Networks Controlled by Biologically Meaningful Functions. *Journal of Theoretical Biology* **218**, 331–341 (2002).
- [34] Aracena, J. Maximum Number of Fixed Points in Regulatory Boolean Networks. *Bulletin of Mathematical Biology* **70**, 1398 (2008).
- [35] Zobolas, J., Monteiro, P. T., Kuiper, M. & Flobak, A. Boolean function metrics can assist modelers to check and choose logical rules. *Journal of Theoretical Biology* **538**, 111025 (2022).
- [36] Kauffman, S. A., Peterson, C., Samuelsson, B. & Troein, C. Random Boolean network models and the yeast transcriptional network. *Proceedings of the National Academy of Sciences* **100**, 14796–14799 (2003).
- [37] Harris, S. E., Sawhill, B. K., Wuensche, A. & Kauffman, S. A. A Model of Transcriptional Regulatory Networks Based on Biases in the Observed Regulation Rules. *Complexity* **7**, 23–40 (2002).
- [38] Jarrah, A. S., Raposa, B. & Laubenbacher, R. Nested Canalizing, Unate Cascade, and Polynomial Functions. *Physica D: Nonlinear Phenomena* **233**, 167–174 (2007).
- [39] Li, Y., Adeyeye, J. O., Murrugarra, D., Aguilar, B. & Laubenbacher, R. Boolean nested canalizing functions: A comprehensive analysis. *Theoretical Computer Science* **481**, 24–36 (2013).
- [40] Layne, L. *Biologically Relevant Classes of Boolean Functions*. Ph.D. thesis, Clemson University (2011).
- [41] Gat-Viks, I. & Shamir, R. Chain functions and scoring functions in genetic networks. *Bioinformatics* **19**, i108–i117 (2003).

- [42] Shmulevich, I., Lähdesmäki, H., Dougherty, E. R., Astola, J. & Zhang, W. The role of certain Post classes in Boolean network models of genetic networks. *Proceedings of the National Academy of Sciences* **100**, 10734–10739 (2003).
- [43] Fink, T. & Hannam, R. Boolean composition restricts biological logics. *arXiv preprint arXiv:2109.12551* (2021).
- [44] Park, P. J. Chip-seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics* **10**, 669–680 (2009).
- [45] Helikar, T. *et al.* The Cell Collective: Toward an open and collaborative approach to systems biology. *BMC Systems Biology* **6**, 1–14 (2012).
- [46] Daniels, B. C. *et al.* Criticality Distinguishes the Ensemble of Biological Regulatory Networks. *Physical Review Letters* **121**, 138102 (2018).
- [47] Borriello, E. & Daniels, B. C. The basis of easy controllability in Boolean networks. *Nature Communications* **12**, 5227 (2021).
- [48] Skarja, M., Remic, B. & Jerman, I. Boolean networks with variable number of inputs (K). *Chaos: An Interdisciplinary Journal of Nonlinear Science* **14**, 205–216 (2004).
- [49] Subbaroyan, A., Martin, O. C. & Samal, A. Minimum complexity drives regulatory logic in Boolean models of living systems. *PNAS Nexus* **1**, pgac017 (2022).
- [50] Subbaroyan, A., Martin, O. C. & Samal, A. A preference for link operator functions can drive Boolean biological networks towards critical dynamics. *Journal of Biosciences* **47**, 17 (2022).
- [51] Yadav, Y., Subbaroyan, A., Martin, O. C. & Samal, A. Relative importance of composition structures and biologically meaningful logics in bipartite Boolean models of gene regulation. *Scientific Reports* **12**, 18156 (2022).
- [52] Azpeitia, E., Benítez, M., Vega, I., Villarreal, C. & Alvarez-Buylla, E. R. Single-cell and coupled GRN models of cell patterning in the *Arabidopsis thaliana* root stem cell niche. *BMC Systems Biology* **4**, 1–19 (2010).

- [53] Guberman, E., Sherief, H. & Regan, E. R. Boolean model of anchorage dependence and contact inhibition points to coordinated inhibition but semi-independent induction of proliferation and migration. *Computational and Structural Biotechnology Journal* **18**, 2145–2165 (2020).
- [54] Zhou, J. X., Samal, A., d’Hérouël, A. F., Price, N. D. & Huang, S. Relative stability of network states in Boolean network models of gene regulation in development. *Biosystems* **142-143**, 15–24 (2016).
- [55] Lähdesmäki, H., Shmulevich, I. & Yli-Harja, O. On Learning Gene Regulatory Networks Under the Boolean Network Model. *Machine Learning* **52**, 147–167 (2003).
- [56] Martin, S., Zhang, Z., Martino, A. & Faulon, J.-L. Boolean dynamics of genetic regulatory networks inferred from microarray time series data. *Bioinformatics* **23**, 866–874 (2007).
- [57] Zhou, J. X., Aliyu, M. D. S., Aurell, E. & Huang, S. Quasi-potential landscape in complex multi-stable systems. *Journal of the Royal Society Interface* **9**, 3539–3553 (2012).
- [58] 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)*, 34–41 (2019).
- [59] Beneš, N., Brim, L., Huvar, O., Pastva, S. & Šafránek, D. Boolean network sketches: a unifying framework for logical model inference. *Bioinformatics* **39**, btad158 (2023).
- [60] Liang, S., Fuhrman, S. & Somogyi, R. Reveal, a general reverse engineering algorithm for inference of genetic network architectures. In *Biocomputing*, 18–29 (1998).
- [61] Hopfensitz, M. *et al.* Multiscale Binarization of Gene Expression Data for Reconstructing Boolean networks. *IEEE/ACM transactions on computational biology and bioinformatics* **9**, 487–498 (2012).

- [62] Barman, S. & Kwon, Y.-K. A Boolean network inference from time-series gene expression data using a genetic algorithm. *Bioinformatics* **34**, i927–i933 (2018).
- [63] Dorier, J. *et al.* Boolean regulatory network reconstruction using literature based knowledge with a genetic algorithm optimization method. *BMC Bioinformatics* **17**, 410 (2016).
- [64] Maucher, M., Kracher, B., Köhl, M. & Kestler, H. A. Inferring Boolean network structure via correlation. *Bioinformatics* **27**, 1529–1536 (2011).
- [65] Ghaffarizadeh, A., Podgorski, G. J. & Flann, N. S. Applying attractor dynamics to infer gene regulatory interactions involved in cellular differentiation. *Biosystems* **155**, 29–41 (2017).
- [66] Yordanov, B. *et al.* A method to identify and analyze biological programs through automated reasoning. *NPJ Systems Biology and Applications* **2**, 16010 (2016).
- [67] Joo, J. I., Zhou, J. X., Huang, S. & Cho, K. H. Determining Relative Dynamic Stability of Cell States Using Boolean Network Model. *Scientific Reports* **8**, 12077 (2018).
- [68] Subbaroyan, A., Sil, P., Martin, O. C. & Samal, A. Leveraging developmental landscapes for model selection in Boolean gene regulatory networks. *Briefings in Bioinformatics* **24**, bbad160 (2023).
- [69] Feldman, J. Minimization of Boolean complexity in human concept learning. *Nature* **407**, 630–633 (2000).
- [70] Feldman, J. A catalog of Boolean concepts. *Journal of Mathematical Psychology* **47**, 75–89 (2003).
- [71] Shmulevich, I. & Kauffman, S. A. Activities and Sensitivities in Boolean Network Models. *Physical Review Letters* **93**, 48701 (2004).
- [72] Hart, S. A note on the edges of the n -cube. *Discrete Mathematics* **14**, 157–163 (1976).

- [73] Azpeitia, E., Weinstein, N., Benítez, M., Mendoza, L. & Alvarez-Buylla, E. R. Finding Missing Interactions of the *Arabidopsis thaliana* Root Stem Cell Niche Gene Regulatory Network. *Frontiers in Plant Science* **4**, 110 (2013).
- [74] García-Gómez, M. L., Azpeitia, E. & Álvarez Buylla, E. R. A dynamic genetic-hormonal regulatory network model explains multiple cellular behaviors of the root apical meristem of *Arabidopsis thaliana*. *PLoS Computational Biology* **13**, 1–36 (2017).
- [75] García-Gómez, M. L. *et al.* A system-level mechanistic explanation for asymmetric stem cell fates: *Arabidopsis thaliana* root niche as a study system. *Scientific Reports* **10**, 3525 (2020).
- [76] Wagner, A. *Robustness and evolvability in living systems* (Princeton University Press, 2005).
- [77] Grefenstette, J., Kim, S. & Kauffman, S. An analysis of the class of gene regulatory functions implied by a biochemical model. *Biosystems* **84**, 81–90 (2006).
- [78] Just, W., Shmulevich, I. & Konvalina, J. The number and probability of canalizing functions. *Physica D: Nonlinear Phenomena* **197**, 211–221 (2004).
- [79] Szallasi, Z. & Liang, S. Modeling the normal and neoplastic cell cycle with ‘realistic Boolean genetic networks’: Their application for understanding carcinogenesis and assessing therapeutic strategies. In *Pacific Symposium on Biocomputing*, vol. 3, 66–76 (1998).
- [80] Wegener, I. *The Complexity of Boolean Functions* (John Wiley & Sons, Inc., USA, 1987).
- [81] Reichhardt, C. J. O. & Bassler, K. E. Canalization and symmetry in Boolean models for genetic regulatory networks. *Journal of Physics A: Mathematical and Theoretical* **40**, 4339–4350 (2007).
- [82] Anthony, M. *Discrete Mathematics of Neural Networks* (Society for Industrial and Applied Mathematics, Philadelphia, 2001).

- [83] Kadelka, C., Kuipers, J. & Laubenbacher, R. The influence of canalization on the robustness of Boolean networks. *Physica D: Nonlinear Phenomena* **353**, 39–47 (2017).
- [84] Golumbic, M. C., Gurvich, V., Crama, Y. & Hammer, P. L. *Read-once functions*, 448–486. Encyclopedia of Mathematics and its Applications (Cambridge University Press, 2011).
- [85] Hayes, J. P. The Fanout Structure of Switching Functions. *Journal of the Association for Computing Machinery* **22**, 551–571 (1975).
- [86] Nikolajewa, S., Friedel, M. & Wilhelm, T. Boolean networks with biologically relevant rules show ordered behavior. *Biosystems* **90**, 40–47 (2007).
- [87] Mendoza, L., Thieffry, D. & Alvarez-Buylla, E. R. Genetic control of flower morphogenesis in *Arabidopsis thaliana*: a logical analysis. *Bioinformatics* **15**, 593–606 (1999).
- [88] Fauré, A., Naldi, A., Chaouiya, C. & Thieffry, D. Dynamical analysis of a generic Boolean model for the control of the mammalian cell cycle. *Bioinformatics* **22**, e124–e131 (2006).
- [89] de Jong, H. Modeling and Simulation of Genetic Regulatory Systems: A Literature Review. *Journal of Computational Biology* **9**, 67–103 (2002).
- [90] Mendoza, L. & Xenarios, I. A method for the generation of standardized qualitative dynamical systems of regulatory networks. *Theoretical Biology and Medical Modelling* **3**, 13 (2006).
- [91] Gonzalez, A. G., Naldi, A., Sanchez, L., Thieffry, D. & Chaouiya, C. GINsim: A software suite for the qualitative modelling, simulation and analysis of regulatory networks. *Biosystems* **84**, 91–100 (2006).
- [92] Cook, S., Dwork, C. & Reischuk, R. Upper and Lower Time Bounds for Parallel Random Access Machines without Simultaneous Writes. *SIAM Journal on Computing* **15**, 87–97 (1986).

- [93] Sasao & Kinoshita. On the number of fanout-free functions and unate cascade functions. *IEEE Transactions on Computers* **C-28**, 66–72 (1979).
- [94] Givone, D. D. *Introduction to Switching Circuit Theory*. McGraw-Hill computer science series (McGraw-Hill, New York, 1970).
- [95] Garey, M. R. & Johnson, D. S. *Computers and Intractability; A Guide to the Theory of NP-Completeness* (W. H. Freeman & Co., USA, 1990).
- [96] Vigo, R. A note on the complexity of Boolean concepts. *Journal of Mathematical Psychology* **50**, 501–510 (2006).
- [97] Berkeley Logic Synthesis & Verification Group. ABC: A System for Sequential Synthesis and Verification (2010).
- [98] Brayton, R. & Mishchenko, A. ABC: An Academic Industrial-Strength Verification Tool. In *Computer Aided Verification*, 24–40 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2010).
- [99] Quine, W. V. A Way to Simplify Truth Functions. *The American Mathematical Monthly* **62**, 627–631 (1955).
- [100] McCluskey, E. J. Minimization of Boolean Functions. *The Bell System Technical Journal* **35**, 1417–1444 (1956).
- [101] Harper, L. H. Optimal Assignments of Numbers to Vertices. *Journal of the Society for Industrial and Applied Mathematics* **12**, 131–135 (1964).
- [102] Bernstein, A. J. Maximally Connected Arrays on the n -Cube. *SIAM Journal on Applied Mathematics* **15**, 1485–1489 (1967).
- [103] Samal, A. & Jain, S. The regulatory network of *E. coli* metabolism as a Boolean dynamical system exhibits both homeostasis and flexibility of response. *BMC Systems Biology* **2**, 1–18 (2008).
- [104] Murrugarra, D. & Laubenbacher, R. Regulatory patterns in molecular interaction networks. *Journal of Theoretical Biology* **288**, 66–72 (2011).

- [105] Klotz, J. G., Heckel, R. & Schober, S. Bounds on the Average Sensitivity of Nested Canalizing Functions. *PLoS ONE* **8**, 1–8 (2013).
- [106] Gherardi, M. & Rotondo, P. Measuring Logic Complexity Can Guide Pattern Discovery in Empirical Systems. *Complexity* **21**, 397–408 (2016).
- [107] Çoban, H. & Kabakçioğlu, A. Nested canalizing functions minimize sensitivity and simultaneously promote criticality. *arXiv preprint arXiv:2109.01117* (2021).
- [108] Çoban, H. & Kabakçioğlu, A. Proof for Minimum Sensitivity of Nested Canalizing Functions, a Fractal bound, and Implications for Biology. *Physical Review Letters* **128**, 118101 (2022).
- [109] Henry, A., Monéger, F., Samal, A. & Martin, O. C. Network function shapes network structure: the case of the *Arabidopsis* flower organ specification genetic network. *Molecular Biosystems* **9**, 1726–1735 (2013).
- [110] Hinkelmann, F. & Jarrah, A. S. Inferring Biologically Relevant Models: Nested Canalizing Functions. *ISRN Biomathematics* **2012** (2010).
- [111] Waddington, C. H. & Kacser, H. *The Strategy of the Genes: A Discussion of Some Aspects of Theoretical Biology* (Allen & Unwin, 1957).
- [112] Grinstead, C. M. & Snell, J. L. *Grinstead and Snell's Introduction to Probability* (University Press of Florida, 2009).
- [113] Shmulevich, I., Gluhovsky, I., Hashimoto, R. F., Dougherty, E. R. & Zhang, W. Steady-state analysis of genetic regulatory networks modelled by probabilistic Boolean networks. *Comparative and Functional Genomics* **4**, 601–608 (2003).
- [114] Aichinger, E., Kornet, N., Friedrich, T. & Laux, T. Plant Stem Cell Niches. *Annual Review of Plant Biology* **63**, 615–636 (2012).
- [115] Cayley, A. *A theorem on trees*, vol. 23 (Cambridge University Press, 1889).
- [116] Hagberg, A., Swart, P. & Chult, D. Exploring network structure, dynamics, and function using NetworkX. Tech. Rep., Los Alamos National Laboratory (LANL), Los Alamos, NM (United States) (2008).

- [117] Müssel, C., Hopfensitz, M. & Kestler, H. A. BoolNet – an R package for generation, reconstruction and analysis of Boolean networks. *Bioinformatics* **26**, 1378–1380 (2010).
- [118] Potts, R. B. Some generalized order-disorder transformations. *Mathematical Proceedings of the Cambridge Philosophical Society* **48**, 106–109 (1952).
- [119] Akutsu, T., Miyano, S. & Kuhara, S. Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. In *Biocomputing'99*, 17–28 (World Scientific, 1999).
- [120] Pal, R., Ivanov, I., Datta, A., Bittner, M. L. & Dougherty, E. R. Generating Boolean networks with a prescribed attractor structure. *Bioinformatics* **21**, 4021–4025 (2005).
- [121] Hickman, G. J. & Hodgman, T. C. Inference of gene regulatory networks using Boolean-network inference methods. *Journal of Bioinformatics and Computational Biology* **07**, 1013–1029 (2009).
- [122] Trinh, H. & Kwon, Y.-K. A novel constrained genetic algorithm-based Boolean network inference method from steady-state gene expression data. *Bioinformatics* **37**, i383–i391 (2021).
- [123] Hérault, L., Poplineau, M., Duprez, E. & Remy, E. A novel Boolean network inference strategy to model early hematopoiesis aging. *Computational and Structural Biotechnology Journal* **21**, 21–33 (2023).
- [124] Villarreal, C., Padilla-Longoria, P. & Alvarez-Buylla, E. R. General Theory of Genotype to Phenotype Mapping: Derivation of Epigenetic Landscapes from N-Node Complex Gene Regulatory Networks. *Physical Review Letters* **109**, 118102 (2012).
- [125] Álvarez Buylla, E. R. *et al.* Floral Morphogenesis: Stochastic Explorations of a Gene Network Epigenetic Landscape. *PLoS ONE* **3**, 1–13 (2008).

- [126] Davila-Velderrain, J., Villarreal, C. & Alvarez-Buylla, E. R. Reshaping the epigenetic landscape during early flower development: induction of attractor transitions by relative differences in gene decay rates. *BMC Systems Biology* **9**, 20 (2015).
- [127] Davila-Velderrain, J., Caldu-Primo, J. L., Martinez-Garcia, J. C. & Alvarez-Buylla, E. R. Modeling the Epigenetic Landscape in Plant Development. In *Computational Cell Biology: Methods and Protocols*, 357–383 (Springer New York, New York, 2018).
- [128] Mora, T. & Bialek, W. Are Biological Systems Poised at Criticality? *Journal of Statistical Physics* **144**, 268–302 (2011).
- [129] Nykter, M. *et al.* Gene expression dynamics in the macrophage exhibit criticality. *Proceedings of the National Academy of Sciences* **105**, 1897–1900 (2008).
- [130] Villani, M. *et al.* Dynamical regimes in non-ergodic random Boolean networks. *Natural Computing* **16**, 353–363 (2017).
- [131] Villani, M., La Rocca, L., Kauffman, S. A. & Serra, R. Dynamical Criticality in Gene Regulatory Networks. *Complexity* **2018**, 5980636 (2018).
- [132] Derrida, B. & Pomeau, Y. Random networks of automata: a simple annealed approximation. *Europhysics Letters* **1**, 45 (1986).
- [133] Gaston, K. & Jayaraman, P. S. Transcriptional repression in eukaryotes: repressors and repression mechanisms. *Cellular and molecular life sciences : CMLS* **60**, 721–741 (2003).
- [134] Cho, K.-H. *et al.* Reverse engineering of gene regulatory networks. *IET Systems Biology* **1**, 149–163 (2007).
- [135] Dimitrova, E. *et al.* Parameter estimation for Boolean models of biological networks. *Foundations of Formal Reconstruction of Biochemical Networks* **412**, 2816–2826 (2011).
- [136] Laubenbacher, R. & Stigler, B. A computational algebra approach to the reverse engineering of gene regulatory networks. *Journal of Theoretical Biology* **229**, 523–537 (2004).

- [137] Pandey, S. *et al.* Boolean modeling of transcriptome data reveals novel modes of heterotrimeric G-protein action. *Molecular Systems Biology* **6**, 372 (2010).
- [138] Balleza, E. *et al.* Critical dynamics in genetic regulatory networks: examples from four kingdoms. *PLoS One* **3**, e2456 (2008).
- [139] Chowdhury, S. *et al.* Information propagation within the Genetic Network of *Saccharomyces cerevisiae*. *BMC Systems Biology* **4**, 1–10 (2010).
- [140] Li, C. *et al.* BioModels Database: An enhanced, curated and annotated resource for published quantitative kinetic models. *BMC Systems Biology* **4**, 1–14 (2010).
- [141] Méndez, A. & Mendoza, L. A Network Model to Describe the Terminal Differentiation of B Cells. *PLoS Computational Biology* **12**, e1004696 (2016).
- [142] Hannam, R., Kühn, R. & Annibale, A. Percolation in bipartite Boolean networks and its role in sustaining life. *Journal of Physics A: Mathematical and Theoretical* **52**, 334002 (2019).
- [143] Torrisi, G., Kühn, R. & Annibale, A. Percolation on the gene regulatory network. *Journal of Statistical Mechanics: Theory and Experiment* **2020**, 083501 (2020).
- [144] Shlyueva, D., Stampfel, G. & Stark, A. Transcriptional enhancers: from properties to genome-wide predictions. *Nature Reviews Genetics* **15**, 272–286 (2014).
- [145] Reiter, F., Wienerroither, S. & Stark, A. Combinatorial function of transcription factors and cofactors. *Current Opinion in Genetics & Development* **43**, 73–81 (2017).
- [146] Graudenzi, A. *et al.* Dynamical Properties of a Boolean model of Gene Regulatory Network with Memory. *Journal of Computational Biology* **18**, 1291–1303 (2011).
- [147] Graudenzi, A., Serra, R., Villani, M., Colacci, A. & Kauffman, S. A. Robustness Analysis of a Boolean Model of Gene Regulatory Network with Memory. *Journal of Computational Biology* **18**, 559–577 (2011).
- [148] Schwanhäusser, B. *et al.* Global quantification of mammalian gene expression control. *Nature* **473**, 337–342 (2011).

- [149] Flöttmann, M., Krause, F., Klipp, E. & Krantz, M. Reaction-contingency based bipartite Boolean modelling. *BMC Systems Biology* **7**, 1–12 (2013).
- [150] Mori, T., Flöttmann, M., Krantz, M., Akutsu, T. & Klipp, E. Stochastic simulation of Boolean rxncon models: towards quantitative analysis of large signaling networks. *BMC Systems Biology* **9**, 1–9 (2015).
- [151] Rottensteiner, H., Kal, A. J., Hamilton, B., Ruis, H. & Tabak, H. F. A heterodimer of the Zn2Cys6 transcription factors Pip2p and Oaf1p controls induction of genes encoding peroxisomal proteins in *Saccharomyces cerevisiae*. *European Journal of Biochemistry* **247**, 776–783 (1997).
- [152] Montagna, S., Braccini, M. & Roli, A. The impact of self-loops on Boolean networks attractor landscape and implications for cell differentiation modelling. *IEEE/ACM transactions on computational biology and bioinformatics* **18**, 2702–2713 (2020).
- [153] Fernandes, L., Rodrigues-Pousada, C. & Struhl, K. Yap, a novel family of eight bZIP proteins in *Saccharomyces cerevisiae* with distinct biological functions. *Molecular and Cellular Biology* **17**, 6982–6993 (1997).
- [154] Wolberger, C. Multiprotein-DNA Complexes in Transcriptional Regulation. *Annual Review of Biophysics and Biomolecular Structure* **28**, 29–56 (1999).
- [155] Vernoux, T. *et al.* The auxin signalling network translates dynamic input into robust patterning at the shoot apex. *Molecular Systems Biology* **7**, 508 (2011).
- [156] Funnell, A. P. W. & Crossley, M. Homo- and Heterodimerization in Transcriptional Regulation. In *Protein Dimerization and Oligomerization in Biology*, vol. 747, 105–121 (Springer New York, New York, 2012).
- [157] Guilfoyle, T. J. & Hagen, G. Auxin response factors. *Current Opinion in Plant Biology* **10**, 453–460 (2007).
- [158] Meldal, B. H. M. *et al.* Complex Portal 2022: new curation frontiers. *Nucleic Acids Research* **50**, D578–D586 (2022).

- [159] Lambert, S. A. *et al.* The human transcription factors. *Cell* **172**, 650–665 (2018).
- [160] Consortium, T. U. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research* **49**, D480–D489 (2020).
- [161] Dröge-Laser, W., Snoek, B. L., Snel, B. & Weiste, C. The Arabidopsis bZIP transcription factor family — an update. *Current Opinion in Plant Biology* **45**, 36–49 (2018).
- [162] Chen, L. & Lopes, J. M. Multiple bHLH proteins regulate CIT2 expression in *Saccharomyces cerevisiae*. *Yeast* **27**, 345–359 (2010).
- [163] Rodríguez-Martínez, J. A., Reinke, A. W., Bhimsaria, D., Keating, A. E. & Ansari, A. Z. Combinatorial bZIP dimers display complex DNA-binding specificity landscapes. *eLife* **6**, e19272 (2017).
- [164] Jones, S. An overview of the basic helix-loop-helix proteins. *Genome Biology* **5**, 6 (2004).
- [165] Castro-Mondragon, J. A. *et al.* JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Research* **50**, D165–D173 (2021).
- [166] Teixeira, M. C. *et al.* Yeasttract: an upgraded database for the analysis of transcription regulatory networks in *Saccharomyces cerevisiae*. *Nucleic Acids Research* **46**, D348–D353 (2018).
- [167] Balakrishnan, R. *et al.* Yeastmine—an integrated data warehouse for *Saccharomyces cerevisiae* data as a multipurpose tool-kit. *Database* **2012** (2012).
- [168] Project Consortium, E. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57 (2012).
- [169] Lee, D. *et al.* STARRPeaker: uniform processing and accurate identification of STARR-seq active regions. *Genome Biology* **21**, 1–24 (2020).

- [170] Spitz, F. & Furlong, E. E. Transcription factors: from enhancer binding to developmental control. *Nature Reviews Genetics* **13**, 613–626 (2012).
- [171] Sandelin, A. *et al.* Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nature Reviews Genetics* **8**, 424–436 (2007).
- [172] Blackwood, E. M. & Kadonaga, J. T. Going the distance: a current view of enhancer action. *Science* **281**, 60–63 (1998).
- [173] Rao, S., Ahmad, K. & Ramachandran, S. Cooperative binding between distant transcription factors is a hallmark of active enhancers. *Molecular Cell* **81**, 1651–1665.e4 (2021).
- [174] Lex, A., Gehlenborg, N., Strobel, H., Vuilleumot, R. & Pfister, H. UpSet: Visualization of Intersecting Sets. *IEEE transactions on visualization and computer graphics* **20**, 1983–1992 (2014).
- [175] Singh, A., Nascimento, J. M., Kowar, S., Busch, H. & Boerries, M. Boolean approach to signalling pathway modelling in HGF-induced keratinocyte migration. *Bioinformatics* **28**, i495–i501 (2012).
- [176] Song, K. S. *et al.* Induction of MUC8 gene expression by interleukin-1 β is mediated by a sequential ERK MAPK/RSK1/CREB cascade pathway in human airway epithelial cells. *Journal of Biological Chemistry* **278**, 34890–34896 (2003).
- [177] Herrmann, F., Groß, A., Zhou, D., Kestler, H. A. & Köhl, M. A Boolean Model of the Cardiac Gene Regulatory Network Determining First and Second Heart Field Identity. *PLoS ONE* **7**, e46798 (2012).
- [178] Remy, E. *et al.* A Modeling Approach to Explain Mutually Exclusive and Co-Occurring Genetic Alterations in Bladder Tumorigenesis. *Cancer Research* **75**, 4042–4052 (2015).
- [179] Gan, X. & Albert, R. Analysis of a dynamic model of guard cell signaling reveals the stability of signal propagation. *BMC Systems Biology* **10**, 1–14 (2016).