2023-11-29

Stability, Longevity, and Regulatory Bionetworks

Christian N. K. Anderson
Brigham Young University

Stability, Longevity, and Regulatory Bionetworks

Christian N. K. Anderson

A dissertation submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Mark K. Transtrum, Chair
J. C. Price
Samuel Payne
Sean Warnick
Dennis Della Corte

Department of Physics and Astronomy

Brigham Young University

ABSTRACT

Stability, Longevity, and Regulatory Bionetworks

Christian N. K. Anderson
Department of Physics and Astronomy, BYU
Doctor of Philosophy

Genome-wide studies of diseases and chronic conditions frequently fail to uncover marked or consistent differences in RNA or protein concentrations. However, the developing field of kinetic proteomics has made promising discoveries in differences in the turnover rate of these same proteins, even when concentrations were not necessarily different. The situation can theoretically be modeled mathematically using bifurcation equations, but uncovering the proper form of these is difficult. To this end, we developed TWIG, a method for characterizing bifurcations that leverages information geometry to identify drivers of complex systems. Using this, we characterized the bifurcation and stability properties of all 132 possible 3- and 22,662 possible 4-node subgraphs (motifs) of protein-protein interaction networks. Analyzing millions of real world protein networks indicates that natural selection has little preference for motifs that are stable *perse*, but a great preference for motifs who have parameter regions that are exclusively stable, rather than poorly constrained mixtures of stability and instability. We apply this knowledge to mice on calorie restricted (CR) diets, demonstrating that changes in their protein turnover rates do indeed make their protein networks more stable, explaining why CR is the most robust way known to extend lifespan.

Keywords: bifurcations, protein turnover, stability, regulatory network, topological boundaries, longevity, calorie restriction

ACKNOWLEDGMENTS

I would like to acknowledge Dr. J.C. Price for his introduction to proteomics, MS/MS kinematics, and the wider BYU research community. His lab was a first landing place on my return to Utah after twenty years away, and it couldn't have been a more exciting or welcoming place. I also wish to thank his students, particularly Dr. Richard Hajime Carson, Dr. Lavender Lin, and Dr. Bradley Naylor whose prior years of research, software development, lab protocols, and humor as I fumblingly did my first MS/MS analyses made this project possible.

My thanks also to Dr. Sam Payne, fellow traveler across Middle Academia, for his many conversations, critiques, classes, and emergency debugging of biological database APIs. For much the same reasons, my thanks to Drs. Sean Warnick, and Dennis Della Corte, teachers and committee members extraordinaires.

I thank also the many people who helped me on my winding academic journey to a bachelor's degree in biology at Stanford, and a masters and a dissertation (but not a defense of same) in Marine Biology at Scripps Institution of Oceanography, UCSD; they are too numerous to mention, but include especially Drs. Elizabeth Hadly, Paul K. Dayton, George S. Sugihara, Nic Rawlence, Jessica Metcalf, and Scott V. Edwards.

My profoundest debt of gratitude goes to my advisor, Dr. Mark K. Transtrum, whose mathematical brilliance in the field of model analysis is matched only by his astonishing patience, compassion, and availability to explain to those of us behind him how to follow where he led. Every grad student should be so lucky in their mentors and friends.

Contents

List of Figures

List of Tables

DEDICATION

I dedicate this dissertation to my parents,
Paul Lawrence (1946-2018) and Lavina Fielding Anderson (1944-2023),
whose deaths bracketed my BYU grad school career
as tidily as their scholarship, love, and unending support
blessed and bracketed my life.

~ And ~

to Marina Capella, M.D., M.Ed.; wife and partner;
co-rescuer of 30+ rabbits, six chickens, and one beleaguered cat;
and for again and again being my island of stability
when everything around me felt like chaos.

Chapter 1

Introductory material

1.1 The origin story of this thesis

The idea behind this dissertation came to me shortly after Dr. Price had introduced me to Dr. Transtrum, and both had generously spent hours explaining to me the implications of the powerful tools they had developed. Dr Price was very nearly the world's sole practitioner of kinetic proteomics, having pioneered and perfected a technique to measure the turnover of thousands of proteins in a single experiment, which was revealing profound differences in the effort required by sick or senescent cells to maintain the normal protein levels that so frustrated generations of GWAS scientists searching for "the cancer gene". Dr. Transtrum had realized that information geometry could be used to identify which parts of a complex system were driving changes in state, and demonstrated that universally all real-world systems had a great many parts that were redundant ("sloppy") at least at the scale and condition we were studying them.

It suddenly occurred to me that together, these two tools might go far in explaining the Holy Grail of Systems Biology: why does a healthy cell suddenly change and become cancerous or senescent? The typical reason for some of our more interesting bifurcating mathematical systems

to change state is an increase in rate constants, as in the logistic map where a period-doubling bifurcation cascade leads to chaos in a figure so iconic it frequently appears in TV shows and movies without explanation (Fig. 1.1).[1] Could a change in a cell's state be driven by increasing rate constants? Better yet, could Dr Price's new kinetic proteomic methods prove that rates were increasing in sick cells, and could Dr Transtrum's sloppy modelling methods prove that this drove cells through a bifurcation into a new basin of attraction?

I pitched this idea to Dr Transtrum, and in about 20 minutes he sketched out the entirety of my thesis on a whiteboard. His outline convinced me that while I might have the biological knowledge to make headway, I certainly needed to learn a great deal more math, which is how I ended up in his lab and why this biologically flavored dissertation says "Department of Physics and Astronomy" on the cover. But what makes these tools different than others, how did they come together to prove this theory, and what does it imply about the future of human health and longevity?

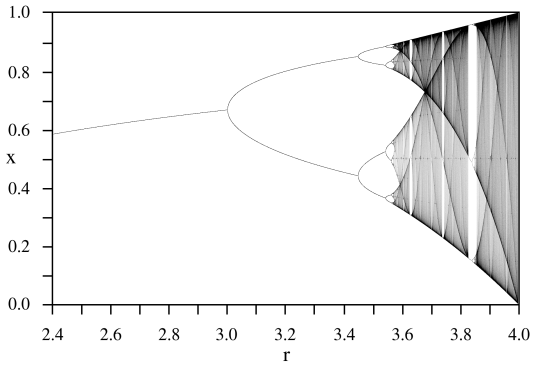

Figure 1.1 The iconic logistic map. [1]

1.2 Analytical tools: a gentle introduction

First, I provide a relatively high-level overview of each of the three tools that needed to be brought together for this study. I anticipate that people reading this will already have expertise in one or two of the necessary fields (bifurcation theory, information geometry, and kinetic proteomics), but would benefit from a general, and slightly less technical introduction to the others. A more technical,

[1] For example, in the Emmy-nominated *The Lost Room* (2006).

but still high-level, overview of the relevant fields can be found in the following section of the Introduction (1.3), and far more technical and specialized introductions can be found in the first section of each of the corresponding chapters.

1.2.1 Bifurcations and Network Theory (for non-mathematicians)

Most mathematical formulae exhibit only one behavior when their parameters are changed. The standard formula for a line, $y = mx + b$ can move up and down with changes to b, and will become steeper or shallower with changes to m, but will never be anything but a straight line. If we assume that y is some measurable quantity and x is time, then this can be recast as an ordinary differential equation (ODE), where we would say that $\dot{y} = m;\ y_0 = b$. No matter how we change the parameter values m and b, y will never display any behavior other than a smooth change at rate m from starting point b. We say the system's behavior is globally topologically homogeneous, because there is always a way to transform any straight line into any other straight line by smoothly adjusting b or m or both. That is, topological homeomorphism exists to turn any one line into any another.

However, nonlinear equations often exhibit more than one kind of behavior. The system $\dot{y} = ry - y^3$ is an interesting example (Fig. 1.2). While $r < 0$, no matter the initial value of y, it will approach 0 as time goes on; the topology is that of a global attractor. However, when $r > 0$, this attractor suddenly splits into two at $\pm\sqrt{r}$, and which value y is drawn to will depend on which side of the separatrix at $y_0 = 0$ the system starts at. While the system is homogeneous for any value above $r > 0$, because we can smoothly move the attractors nearer or farther apart by adjusting r, there is no way to continuously transform two attractors into one, nor vice versa. This bifurcation of attractors is a topological *inhomogeneity*, and because it is the most famous such example, any boundary for which no homeomorphic transformation exists is referred to as a bifurcation even if it doesn't result in a doubling. For example, a "saddle-node" or "blue-sky" bifurcation occurs when an unstable point suddenly becomes stable. In fact, while there are infinite numbers of equations

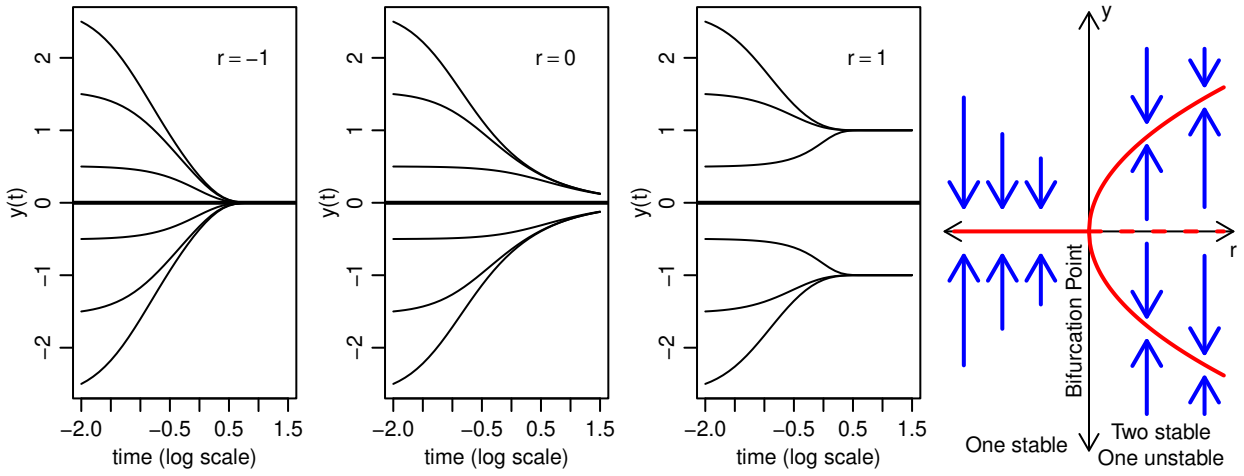

Figure 1.2 The pitchfork bifurcation $\dot{y} = ry - y^3$. shows a switch from one equilibrium to two stable and one unstable equilibrium at $r = 0$. When r is negative, all trajectories are "attracted" to 0, albeit less strongly as $r \to 0^-$. However, after $r > 0$ trajectories are repelled from 0, and attracted to either $\pm\sqrt{r}$ depending on the starting value of $y(0)$.

that show multiple behaviors, mathematicians have been able to reduce nearly all of them down to just five "normal form" classes (or combinations of them), much the same way there are an infinite number of equations that produce straight lines (e.g., $5y - 2x = 4$ or $\frac{y}{x} = 8 - \frac{2}{x}$ or $9^y = 27^x + 3$) but all of them can be reduced to the slope-intercept form $y = mx + b$.

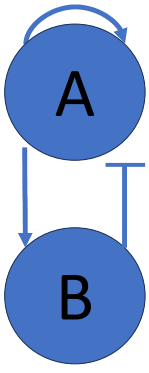

Figure 1.3 A simple network

ODEs are often used to describe protein interaction networks as well. For example, we might say that protein A is being constituitively expressed at rate α (or $\dot{A} = \alpha$ in ODE-ese), but also broken down when it comes in contact with a proteosome, an event that is more likely the more A there is, so $\dot{A} = \alpha - \delta A$. Let us add that A autocatalyzes itself, so $\dot{A} = \alpha + k_{AA}A - \delta A$. So far so good; mathematically it's fairly easy to say that this system has an equilibrium ($\dot{A} = 0$) when $A = \frac{\alpha}{\delta - k_{AA}}$ However, we are interested in protein networks, not isolated behavior. To this end, now let us say that there is a second protein, protein B, which is activated by A and also represses it. This system (Fig. 1.3) is almost as

simple a network as possible,[2] and can be represented by the coupled ODEs:

$$\begin{aligned} \dot{A} &= \alpha - \delta A - k_{BA}AB + k_{AA}A \\ \dot{B} &= \alpha - \delta B + k_{AB}AB \end{aligned} \tag{1.1}$$

Despite the cartoonish simplicity of the system, it is very difficult to say simply by inspection if a bifurcation exists in this system or not. In fact, there are several: by adjusting the parameters, the system can be unstable with A and B increasing without limit, or only B can increase while A is reduced to nothing, both can be reduced to nothing, or the two can reach an equilibrium, or the two can oscillate indefinitely. [4]

Because bifurcations are fascinating and among our best models at generating qualitatively distinct phenomena, but difficult to recognize or generate from formulae, many sophisticated methods have been developed to detect and analyze them (see Sec 1.3 and Sec 2.3.1 below). However, all rely on substantial mathematical skill applied to each unique case. Because a method for understanding bifurcations was needed for thousands of network shapes and millions of instances of protein-protein interactions, a method that could be applied automatically without human intervention was needed. This was the first great challenge of the thesis, and is the subject of Chapter 2.

1.2.2 Information Geometry (for non-physicsts)

IG is a powerful and astonishingly under-utilized tool for understanding what the “important parts” of a system are. Like many tools, it is more readily understandable when applied to a problem: it is easier to explain what a hammer is for to someone who already has a nail.

[2] In formal logic, this network is very similar to the Liar’s Paradox, which can be phrased as "A: Statement B is true; B: Statement A is false". In this case, we would modify the first statement to read "A: This statement and B are both true". One resolution to this paradox is to give different levels of credibility to the statements, which is analogous to assigning different values to the interaction constants.

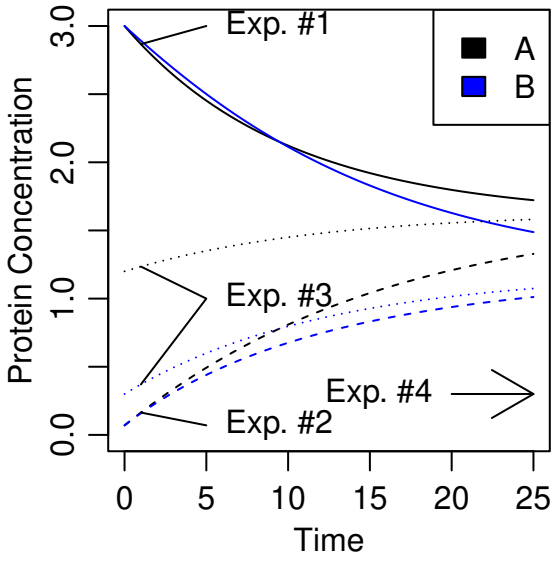

Figure 1.4 The example network above will run to equilibrium values where A (black) is higher than B (blue), though the initial conditions affect the trajectory. Experiment #1 has initial concentrations of 3.0, #2 starts at 0.07, while #3 begins at 1.2 and 0.3 for A and B respectively. Experiment #4 is observed when the system has reached equilibrium, so the initial concentrations do not matter.

Returning to the simple network in Fig. 1.3 and Eq. 1.1, we see the system has five parameters: the constitutive expression rate (α), the degradation rate (δ), autocatalysis rate of A (k_{AA}), and the effect sizes of A and B on each other ($-k_{BA}, k_{AB}$). Which of these parameters is "most important" clearly depends on the context. For example, if very large amounts of A and B are present far in excess of equilibrium values, then the baseline expression α is going to be (temporarily) masked by their degradation. Conversely, if almost zero A and B are present, then none of the density-dependent parameters matter very much, and the system will be (temporarily) driven exclusively by α as all the other parameters are being multiplied by a very small number. This constitutes the "information" part of IG: is α important or not at the moment?

The "geometry" part comes from picturing the effect of adjusting parameter values on how well the model matches your data. Imagine that the true parameter values for the system are

$\alpha = \delta = .1$, $k_{AA} = .05$, $k_{AB} = k_{BA} = .01$, so that the proteins' effects on each other are relatively weak. Starting the system at different initial concentrations will affect the trajectory, but the system will always end up at the same equilibrium values with $A > B$ as $t \to \infty$ (Fig. 1.4). An experiment where A and B both start at high concentrations will be dominated by degradation early, so we expect errors in the δ parameter to be more costly than errors in α. Simulating this network with $A_0 = B_0 = 3.0$ shows that small changes from the correct parameter values indeed result in deviations from the correct concentrations at t=0.5 of [2.932, 2.945], but these error costs increase much faster in the δ direction than the α direction (Fig. 1.5a). The relative rate of increase in the costs is quantified using the Fisher Information Matrix (FIM). The details can be found in Sec. 2.3.2, but for now suffice it to say that the FIM quantifies the curvature of the cost surface[3] in every possible direction. We calculate curvature (second derivatives) not slope (first derivatives) because it is assumed calculations are centered where the cost has a local minimum (the best fit), and at this point the slope is zero in all directions. Directions with high curvature increase costs rapidly, indicating that small changes in those parameter values will result in poor model fits, while those with small changes can be changed a great deal without affecting fit. This sounds like a good thing, but actually means these parameters are poorly constrained by the data, since even large errors might not be detectable. These directions are called "sloppy", and are ubiquitous across biological, physical, and information systems. [5–7]

An important side note here is that the FIM itself quantifies curvature along parameter axes (e.g., $\frac{\partial^2 A}{\partial \alpha^2}$) or combinations of two parameter axes (e.g., $\frac{\partial^2 A}{\partial \alpha \partial \delta}$). However, the greatest and/or least curvature might not align with these directions. For example, in Fig. 1.5 the canyon floor lies at a slight angle to the α direction. Fortunately, a standard tool of linear algebra called singular value decomposition provides a rotation of the FIM so it aligns with the curvatures in order from steepest

[3] Actually, the FIM measures the curvature of the model manifold in dataspace, but the two are coupled. That is, a manifold with a high curvature in one direction will have a highly curved cost surface in the same direction.

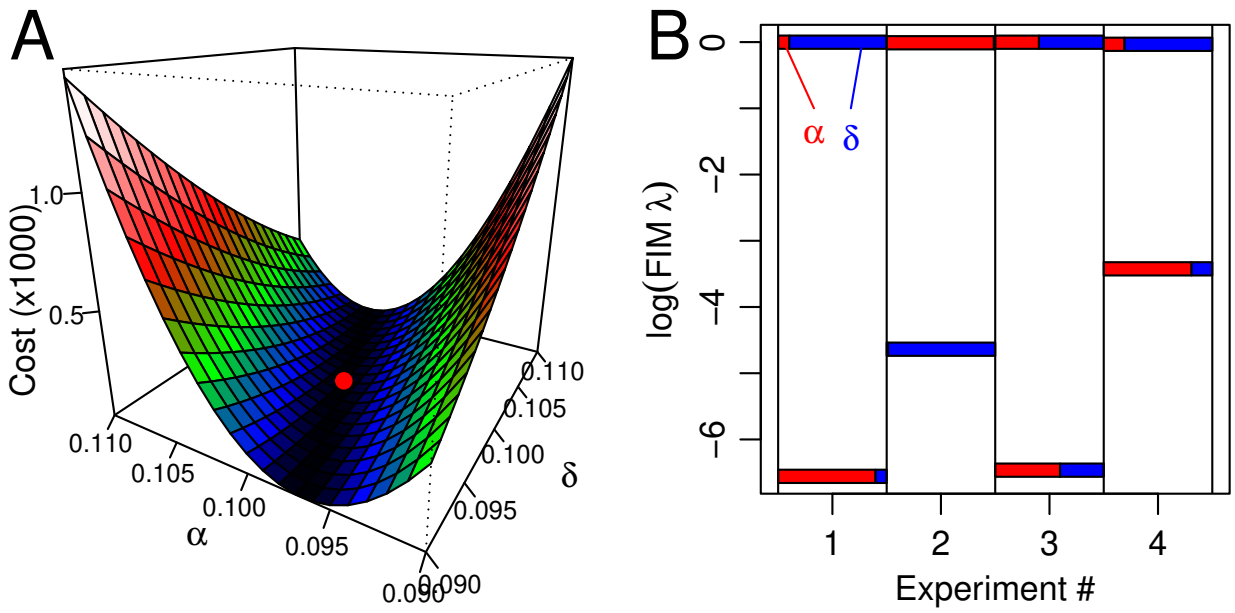

Figure 1.5 (A) Attempting to fit the observed data to Experiment #1 with slightly wrong parameter values is more costly in the δ than the α direction. (B) A graphical interpretation of the Fisher Information Matrix, using eigenanalysis to represent the entries. For Experiment #1, note that one direction is far more sensitive than the other (bars $\sim e^{6.5}$ units apart, a $\sim$700x difference), and this direction aligns mostly, but not entirely, along the δ (blue) axis, corresponding to the slightly-rotated steep canyon sides in (A), and the importance of degradation (δ) over synthesis (α) at high concentrations.

to shallowest (the eigenvectors), and simultaneously provides a numeric measure of how curved each of these directions are (the eigenvalues).

This is reflected in the FIM, which indicates for Experiment #1 that the axis of change is 90% δ and 10% α, and costs accrue orders of magnitude more rapidly in this direction than others (Fig 1.5B). That is, the difference in eigenvalues (the λs of FIM) tells us the sides of the canyon are $\sim$ 700 times steeper than the floor. Similarly, when samples are taken early in Experiment #2, the observations are driven strongly by α. The mixed conditions of Experiment #3 mean both α and

δ are relevant. Finally, the long timescale of Experiment #4 means we are measuring equilibrium concentrations, so both α and δ are relevant because both shift the equilibrium. These last two experiments illustrate a crucial and remarkable power of IG: they have the same separation of scales between the leading and lagging eigenvalues as the first two experiments. This means they are every bit as stiff (the canyon walls are just as steep) in one direction, it just happens to not be the direction of α or δ but a mix of the two. This tells you that the system can be rewritten with different variables that will better match the local geometry of the model than the one we started with, even though we have no idea what this parameter might be (though we know it must include both α or δ) much less having actually transformed the equations into this unknown coordinate system.

Note that, for the sake of simplicity, we have fixed the three k parameters and only considered changes to α and δ. This lets us draw 3D cost surfaces (the full system would require a 6D image) and helps think clearly about the effects of each parameter on the experimental results. Ordinarily an IG analysis of this system would produce a 5x5 FIM, whose five eigenvalues and eigenvectors would quantify the contributions of each of the five parameters of the ODE system in Eq. 1.1.

One of the main uses for Information Geometry is model reduction. Across disciplines, FIM analysis often informs researchers that some parameter directions are 10 or more orders of magnitude less important than the leading parameter direction, and so can be removed from the model (if mathematically possible) without appreciably affecting accuracy. [8] This tendency of almost all models to be locally reducible suggests a universality class among natural processes with built-in redundancy. [9, 10] Overall, IG seems not only to be an interesting tool, but also a case of Wigner's "unreasonably effectiveness" in understanding drivers of real world phenomena. [11]

1.2.3 Kinetic Proteomics (for non-biochemists)

The understanding that proteins break down and need to be replaced goes back to antiquity, with hints of the idea of "dynamic permanence" attributed to Alcmaeon by Aristotle (who expanded on them) and also in the writings of "father of medicine" himself, Hippocrates. [12, 13] The knowledge that proteins turn over at different rates, and sometimes the same protein turns over at different rates in different tissues, was also inferred centuries ago. [14] Yet as recently as 20 years ago, almost no study had attempted to measure these rates for individual proteins, with the exception of well-understood and easily acquired proteins like albumin, insulin, myosin, and collagen; instead, researchers focused on whole-body turnover rates, and speculated about movement between "bound" and "free" amino acid pools in thousands of papers across almost 50 years. [15]

This provided a highly pixelated and incomplete view of protein dynamics, and everyone knew it, but proceeding to the next level of complexity seemed intractably difficult. For one thing, there are in the neighborhood of 25,000 different proteins in every human cell (at least potentially), and separating them is tedious when it is possible at all. Worse, the overwhelming majority of them are extremely rare in almost all cell types, (six proteins make up $> 90\%$ of all blood serum proteins) making the quest to isolate most proteins like finding a few particular straws of hay in a very large hay stack. For another, the interactions between proteins are notoriously difficult to identify. While modern techniques like co-immunoprecipitation assays make it possible to find which proteins routinely form complexes, it is generally not possible to determine from this alone who is regulating who, and if this interaction is up- or down-regulating. As a result, the Protein-Protein Interaction (PPI) database StringDB has 12 million interactions among human proteins, with similar results for the high-quality Harvard BioPlex 3.0 database, [16] while RegNetwork's extraction from the Kyoto Encyclopedia of Genes and Genomes (KEGG)–the largest index of directional and signed regulatory interactions–has just 4,000 listed interactions, and a similar number for mice. [17] Thus, even if it was possible to extract proteins, their regulatory context would remain murky.

However, this started to change around 2000 with the advent of tandem mass spectrometry systems capable of identifying of the individual proteins making up a heterogeneous bulk sample. The full procedure is lengthy, but at its core it works by exposing the sample to a medium-energy laser, which fragments and ionizes the proteins. The resulting charged gas is then passed through a magnetic field, where the mass:charge ratio (m/z) causes different fragments to travel to a detector at different speeds. The fragment size is determined precisely using the time to reach the detector for quantitative time of flight (QToF) systems, or a combination of release time and angular velocity for OrbiTrap systems. Because each amino acid has its own unique mass and number of sites where deuteriums can replace hydrogens (thereby increasing the mass), isotope ratios and fragment identity can be derived rapidly and unambiguously.

Though considered an exotic application by almost all MS/MS labs, who were more comfortable identifying far more homogeneous inorganic mixtures, [18] the first proof-of-concept experiments proved unexpectedly and undeniably robust. The first attempt was isotope-coded affinity tags (ICATs), which involved coupling labeled cysteine with biotin, but had the disadvantage of having an intermediate reaction that could bias the results, and a bit of metabolic scrambling as cysteine was converted into other amino acids rather than being directly incorporated into proteins. [19] Both of these problems had already been solved by Stable Isotope Labeling by Amino Acids in Cell culture (SILAC), [20] though it took four years to publish the proof. [18] Further developments sped the process up, so that it was possible to use SILAC to quantify turnover across the entire yeast proteome in a single experiment by 2008, [21] while deuterium-labelled water had already been integrated into a software pipeline to simultaneously calculate turnovers in thousands of proteins by 2011. [22, 23] This method introduces 7% $^{2}H_2O$ into a model organism via injection (or growth medium for microorganisms), which is distributed through the body in minutes, then provisions it with similarly deuterated water. The use of hydrogen isotopes is useful since it allows proteins to be marked multiple times (providing statistical robustness), is as ubiquitous in the sample as water, and

can be distinguished from isotopes of other elements due to deuterium's unusually large quantum mass defect (a matter of a few thousands of an a.m.u., yet still detectable on modern Orbitrap MS/MS machines), all without interfering with ordinary cellular metabolism. Today, sections of a single tissue–*e.g.*, a brain–can be analyzed serially or in parallel to compare turnover rates of thousands of proteins within the same individual's organ, [24] a specificity that must have seemed impossible 20 years ago.

Full organism proteomics is still challenging due to the hyperabundance of a few proteins "oversaturating" the MS/MS detectors and making the detection of rare species challenging, yet it is already on the horizon with methods like boxcar sampling. Similarly, progress in mathematical analysis and software have made studies in humans feasible, where provision of isotopes is necessarily more erratic than for laboratory animals.

1.3 Synthesis and Overview

Largely because these techniques are so new, interdisciplinary attempts to use two of them in tandem have not yet been attempted formally, at least widely. However, hints already existed that such a fusion of methods would be possible.

1.3.1 Detecting bifurcations

Sethna's group had recently drawn a tight connection between information geometry and the renormalization group. [25] Because bifurcations are associated with the renormalization group, [26–28] this suggested that there was a link between the two. Removing the renormalization "middle-man" was the task of Chapter 2, where I created a new analytical tool called TWIG (for time-widening information geometry) as a nod to the tree-like structure of bifurcation cascades.

As noted briefly above, this represented a substantial improvement over existing analytical methods. The two major alternatives to RG analysis both involve substantial pen-and-paper work by mathematicians. Central Manifold Reduction involves making linear approximations to the system at the bifurcation point, whose relative slopes can reveal information about the system's stability. [29, 30] Similarly, methods involving Lyupanov exponents are able to characterize the variability of different parameters around the bifurcation, which is often (but not always) correlated with instability. [31, 32] Unlike these methods, TWIG is able to not just characterize the instability around a bifurcation point, with hints about which parameters are responsible, but actually determine how many parameters are involved in an optimal reparameterization of the model (the codimension), which bare parameters are involved in this optimal reparameterization and to what degree, the direction of the bifurcation hypersurface (the separatrix), and provides a self-check to make sure the analysis has run long enough to be valid.

1.3.2 Stable regions in real networks

Strides had also been made in the field of protein network analysis when it was realized that in order for a network to be unstable, pieces of it needed to be unstable as well. This gave rise to motif analysis, or detailed studies of subgraphs involving all the connections between a small number of nodes of the parent graph. A leader in the field of motif analysis was Uri Alon, who wrote a series of increasingly important papers through the early 2000s, culminating in a textbook summarizing and synthesizing the entire field. [33–38] The theme of this analysis was that while any interaction topology was technically possible, "evolution... converges again and again onto a defined set of circuit elements that obey general design principles...a rather small set of basic building-block circuits." He called these overrepresented topologies motifs, and posited that "whereas many circuit designs can perform a given function on paper, we will see that very few can work robustly in a cell." [38]

Robustness proved a slippery concept to define mathematically, for Alon and colleagues in systems biology as it had in ecology a few decades earlier. [39] Fortunately, network theorists (some of them at BYU) had worked out an unassailable definition provided the equations for the network could be fully specified and the equilibrium calculated. Their method was called spectral analysis, and involved calculating the eigenvalues of the system's Jacobian at the equilibrium; provided the largest was below a minimum threshold (what that value was depends on the method by which the system is iterated) the system was stable, but unstable if larger. [40] We combined motif-centered thinking with spectral analysis to break the intractable problem of a cell's globally stability into analyzable chunks of subgraphs of 3- or 4-nodes each. We sampled 1000 parameter values for each of the 132 possible 3-motifs and 22,662 4-motifs to determine stability metrics for each one across the range of biologically plausible variables, an undertaking that took over 3 years of CPU time on the BYU supercomputer cluster (Chap. 3). Unlike the relatively small number of behaviors shown by normal-form bifurcations–but like the simple 2-motif example above (Fig. 1.3)–these small networks had many bifurcations and a surprisingly large volume of biologically possible parameter space where they were unstable. This suggested that protein networks were potentially under constant threat of breaking down, and must have some control to prevent this.

Armed with TWIG and the stability analysis of all possible small networks, we were now able to investigate bifurcations in the real world. As noted above (Sec. 1.2.3), there are huge databases of protein-protein interactions because it is relatively easy to experimentally determine that two proteins clump together; however, there are few databases that tease apart who is regulating who in the clump, and if it is positive or negative regulation. For this reason, we were forced to use the relatively small RegNetwork database of approximately 4,000 interactions of 1,000 proteins in humans and a parallel network of the same size for mice. [17] Building on a substantial cottage industry of software to count motifs, [41, 42] we discovered millions of motifs from these links (Chap. 4; Table 4.1). As Alon had promised, many of these motifs were vastly over-represented

relative to literature-derived null expectations. [43] However, this over-representation did not seem to be correlated with any of our measures of stability (Table 4.2).

We considered several reasons for this unexpected finding, eventually deciding that a need for adaptability was apparently trumping the need for stability. However, we realized our analysis assumed network bifurcations were similar to those of the normal-form bifurcations in Fig. 1.2: a well-defined and smooth boundary between two behaviors tidily partitioning parameter space. However, our investigation of higher dimension and chaotic systems in Chap. 2 (especially Sec. 2.6) showed that some parameter spaces were not like this at all; their boundaries were so convoluted as to be fractal and breakdown the very concept of a "behavior region". Careful scrutiny of the stability data from Chap. 3 made us realize that some motifs were far more well-partitioned than others (Fig. 5.6), and these well-partitioned motifs had a very strong tendency to be over-represented in real world networks (Fig. 5.7). It wasn't the amount of stability in the parameter space that mattered, it was how well-bounded that stability was that mattered.

1.3.3 Considering longevity

There were several reasons to believe that this could have implications for longevity. One of the primary hallmarks of aging is the loss of proteostasis, [44] so processes that stabilize protein networks should also postpone aging. One of the most robust methods to induce longevity is a calorie restricted (CR) diet, [45–47] which was also known to decrease the rate of protein turnover in mice. [48] We were now finally in a position to link these two concepts: was it true that CR's tendency to slow turnover led to proteostasis which then led to longevity? This was our task in Chap. 5, where we found that the stability was indeed increased in mice on calorie restricted diets. The increase in stability, though modest, was highly significant despite numerous challenges in linking the turnover and network databases together, then analyzing the gappy results.

Finally, this thesis considers the implications of the demonstrated link between longevity and protein networks operating on the stable side of bifurcations. What does this mean for human anti-aging treatments? Does it have implications for proximate causes of death, such as cancer or Parkinson's disease? What does the future for protein network stability look like?

Chapter 2

Sloppy model analysis provides bifurcation characterization

2.1 Abstract

Bifurcation phenomena are common in multi-dimensional multi-parameter dynamical systems. Normal form theory suggests that bifurcations are driven by relatively few combinations of parameters. Models of complex systems, however, rarely appear in normal form, and bifurcations are controlled by nonlinear combinations of the bare parameters of differential equations. Discovering reparameterizations to transform complex equations into a normal form is often very difficult, and the reparameterization may not even exist in a closed-form. Here, we show that information geometry and sloppy model analysis using the Fisher information matrix can be used to identify the combination of parameters that control bifurcations. By considering observations on increasingly long time scales, we find those parameters that rapidly characterize the system's topological inhomogeneities, whether the system is in normal form or not. We anticipate that this novel analytical

method, which we call time-widening information geometry (TWIG), will be useful in applied network analysis.

2.2 Introduction

This paper provides a method for extracting bifurcation parameters from a set of dynamic equations by combining information geometry and bifurcation theory. Both are useful for modeling multi-parameter systems and systems with multiple regimes of behavior respectively, but together they provide methods for data-driven analysis of a wide array of natural phenomena. By creating an explicit connection between the information in the signal (model output) and the model parameters, we identify the combinations of parameters responsible for topological change in the dynamics, the codimension of the bifurcation, and the time scale necessary to resolve this information. The information further provides the directions normal to the separatrix, which divides behavioral regimes of the system.

Traditionally, when confronted with a high-dimensional, multi-parameter system of dynamic equations, bifurcation analysis proceeds by attempting to simplify the system to a manageable size. Center Manifold Reduction exploits the Hartman-Grobman theorem [29] to create a lower-dimensional linear map in the region of a critical point that is locally accurate and is a rapid way to determine the system stability. Shoshitaivishili extended this method to non-hyperbolic equilibria, creating a container for critical modes to straighten out non-linear terms and, ideally, drop some of them [30]. Such methods have been used to describe phenomena as diverse as neural network optimization and foraging decisions in monkeys [49, 50].

A related approach is the method of Poincaré-Birkhoff normal forms. It uses appropriately centered manifolds to analyze which nonlinear terms are essential and must remain even under optimal coordinate transformations. Such transformations are useful, because the reduced normal-

form equations typically have greater symmetry than the initial problem, a property that can be exploited by many analytical tools. Though powerful, "in practice lengthy calculations may be necessary to extract the relevant normal-form coefficients from the initial equations." [30] Even if such coefficients can be found, neither their interrelationship nor their relative sensitivities are always apparent. It is often the case that some parameters differ by many orders of magnitude in their effect on long-term dynamics, and a method that doesn't distinguish among them is sub-optimal for most applications.

The method of Lyupanov exponents is an admirably general tool for analyzing the global stability of a system. Unfortunately, it provides little information about which specific parameter combinations lead to system (in)stability. For the purposes of bifurcation analysis, it is therefore sometimes paired with sensitivity analyses based on the global sensitivity metrics of Sobol' [31]. These measures, along with useful extensions such as FAST (Fourier amplitude sensitivity test) and Importance Measures [32, 51, 52], are able to determine exactly how much of a model's variability is due to each of its parameters. While this often works in practice, there are two potential pitfalls in this approach. First, it assumes that the parameters responsible for variability are also responsible for instability, which is not always the case. Second, if the bifurcation is caused by combinations of many parameters (as frequently happens), then variability will often be high across all these parameters even though the bifurcation itself has a low codimension. In other words, a low-dimensional bifurcation surface generally cuts diagonally across parameter space unless appropriately reparameterized. Once such a transformation is applied and the system is reduced to a normal form (see Sec. 2.4), then the codimension should be apparent, but finding that reparameterization is still likely to be cumbersome, if not impossible, in closed-form. Just one such transformation can require several papers, as in the case of high-dimensional diffusion-activated processes from Kramers, through Langer, and finally to one dimension, derived using iterations of singular value decomposition by Berezhkovskii [53].

A third, independent line of analysis comes from Renormalization Group (RG) methods, which are usually applied to study universal power-laws near critical points. Feigenbaum [54] was the first to note such universalities in bifurcations of the discrete period-doubling type, a result extended by himself and others until it included all major bifurcation types [27, 55–58]. Working from the other direction, scientists investigating critical phenomena with RG (e.g., many behaviors of quantum chromodynamics) have discovered bifurcations, and used the tools of one to analyze the other [59, 60]. A remarkable study found deep equivalence between RG transformations and normal form theory, showing that the difficult transformation of an ODE system into a normal form could often be accomplished to at least second order by applying three RG transforms [28].

More broadly, universal scaling laws and RG analysis of critical points is often associated with emergence and the systematic irrelevance of many degrees of freedom. Recent work has extended these ideas to a broader class of systems known as "sloppy models" [5, 8, 9, 61, 62]. The moniker "sloppy" is meant to convey that these systems have a few combinations of parameters that are many orders-of-magnitude more influential than other parameter combinations. More precisely, one unit change in a "stiff" parameter direction has as much influence as a million or more unit change in a different "sloppy" direction. Sloppy model analysis relies heavily on the techniques of information geometry [6, 61, 63] and in this paper we use the terms interchangeably. These techniques have motivated novel reduction algorithms by removing unimportant, sloppy parameters [63–65]. Recent work [25] demonstrates that as coarse-graining of RG models proceeds, the flow causes information of "relevant" parameter combinations to be maintained while "irrelevant" parameters are compressed and become sloppy. These ideas share a common goal with bifurcations analysis in which many diverse systems are collected into a few universal, normal forms. This paper closes the loop, showing how information geometry applies directly to bifurcation analysis without passing through the "middleman" of renormalization group theory. The usefulness of such

an analysis, which we call Time Widening Information Geometry (TWIG), also circumvents the need for the other types of analyses described above.

In this work, we demonstrate similar notions of "relevant" and "irrelevant" parameters near a bifurcation using the formalism of information geometry and sloppy models. The intuition behind this approach is as follows. Topological inhomogeneities in the flow field produce trajectories containing different information on either side of a bifurcation. For example, on one side of a Hopf bifurcation, all trajectories collect into a central fixed point, while they flow into an orbit (limit cycle) on the other side. TWIG works by measuring the information content in these trajectories at increasingly long time scales and identifying those combinations of parameters to which the trajectory is most sensitive. At long time scales, these are the parameters responsible for the bifurcation, while parameters that cause only local variability have less impact.

Information geometry can be applied to complex systems from many disciplines–but especially systems biology–to iteratively "reverse engineer" optimal statistical models by removing parameters whose value has little influence on the macroscopic behavior of the system [8, 62, 64, 66]. However, it was recently appreciated that such reverse engineering can be done even if the underlying system bifurcates into qualitatively different behaviors, because the information geometry of parameters participating in the bifurcation show a characteristic "sand dune" shape when crossing from one behavioral state to another [67]. These results imply that if the functional form of the system is known, it should be even easier to determine bifurcation parameters than if the system's equations need to be inferred.

This paper is organized as follows: In Section 2.3, we provide background information on bifurcations and information geometry generally, and, specifically, how we conceptualize them for the purposes of applying the latter to the analysis of the former. In Section 2.4, we show how an IG analysis of the normal form bifurcations rapidly provides insight into the structure of bifurcations

simple enough to be understood by other methods. Section 2.5 shows how this analysis extends to more difficult bifurcations, the implications of which are summarized in Section 2.7.

2.3 Background and Problem Formulation

2.3.1 Bifurcations

Bifurcations frequently arise in the analysis of dynamical systems, where one typically characterizes the flow field with special attention to any fixed points or stable oscillations [68]. Consider a generalized system of n coupled dynamic equations, where each equation is of the form $\dot{\mathbf{y}} = f(\mathbf{y};\theta)$, where θ is a vector of m parameters. Small changes to any of the θ_i values typically result in correspondingly small changes to the n-dimensional vector field, such as small changes to the position of a fixed point or radius of a limit cycle. Such deformations are topologically equivalent (meaning the number and properties of the attractors / repellers in the field do not change) and homeomorphic (continuous with a continuous inverse). However, there may be critical parameter values where a small change causes new fixed points to emerge from old ones, or two fixed points to approach and be mutually annihilated, or limit cycles to be broken. Since one common form of nonhomeomorphic transformation is the emergence of two fixed points from one, the phenomenon is generically called bifurcation, though we discuss other possibilities below.

Several types of simple bifurcations have been identified and reduced to their simplest possible mathematical expression. These are the so-called “normal forms" and are enumerated in the section below. These forms are convenient starting points for analysis, since they have clearly defined rate parameters that are unambigiously responsible for causing topological inhomogeneities. However, even elegant mathematical descriptions of real-world dynamical systems rarely conform exactly to one of the normal forms.

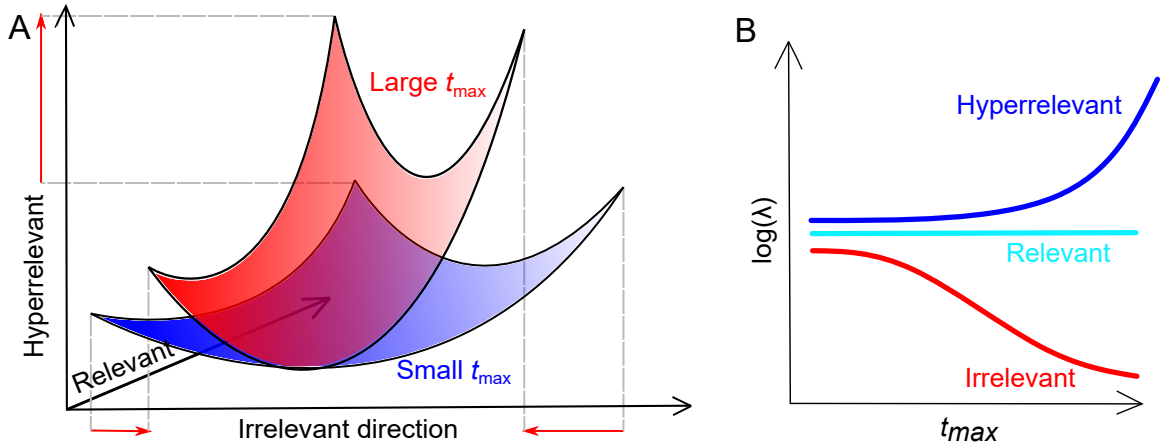

Figure 2.1 (A) The model manifold in data space represents all values that can be reached by changing parameters. The axes represent directions that are distorted in characteristic ways as t_{max} increases. They can be contracted (irrelevant), expanded (hyperrelevant) or unchanged (relevant). (B) Relevance can be quantified by observing the eigenvalues of the Fisher information matrix as t_{max} is increased. Eigenvalues that do not change at longer time scales retain their relevance, while those that increase or decrease become either more or less relevant.

Bifurcation parameters for physical models often do not align with the bare parameters. In the classic example of boiling liquid, the bifurcation parameter is some combination of temperature, pressure, salinity, and others. In general, a reparameterization to a single, unambiguous bifurcation parameter may be possible in principle, but often requires either substantial additional physical insight, or mathematical sophistication, or both. Some researchers have even recommended building an analogous physical circuit as the fastest method to detect the bifurcation [69]. Complex models can have hundreds of coupled dynamic equations with thousands of parameters (e.g., models of sophisticated mobile phone circuit boards [70], or complex protein networks [71]). How can we determine which parameter (or more likely, combination of parameters) is responsible for the bifurcation in such cases?

2.3.2 Information Geometry

The fundamental object of information geometry is the Fisher information matrix (FIM or $\mathscr{I}$), which quantifies the information that the observations $\mathbf{y}$ contain about the parameters θ of a dynamical system. Here we introduce the FIM for dynamical systems.

Consider a system of ordinary differential equations where the parameters are tuned to be exactly at their critical values, i.e., the system is at (one of) its bifurcation point(s). The system is allowed to evolve, and the trajectory of one of its equations y_j is sampled at several time points $y_j(t_i)$, where $t_i = t_0 + \frac{i}{n} t_{max}$. To help visualize this process, let us imagine a one-dimensional system

$$y(t) = \theta_1 + e^{-\theta_2 t} + e^{\theta_3 t} \tag{2.1}$$

sampled at $t = \{1,2,3\}$ to create a vector of three observations $\mathbf{y} = \{y(t_1), y(t_2), y(t_3)\}$ which we plot in $\mathbb{R}^3$, i.e., data space. If $\theta_3 > 0$, then there is no equilibrium; if $\theta_3 = 0$ and $\theta_2 > 0$ then the equilibrium is at $\theta_1 + 1$ or $\theta_1 + 2$ if $\theta_2 = 0$. As the parameters of θ change, the position of $\mathbf{y}$ will also change, but except for extreme values of θ_i, it cannot reach all possible values in $\mathbb{R}^3$. The space filled by values of $\mathbf{y}$ that can be reached for a given range of parameter values defines the model manifold. A schematic of such a manifold is drawn in Fig. 2.1A.

The Fisher information is most-commonly defined in probabilistic terms as the expected Hessian matrix of the log-likelihood:

$$\mathscr{I} = -E\left[\frac{\partial^2}{\partial \theta^2} \log \mathscr{L}(\theta | d)\right] \tag{2.2}$$

where θ is a vector of parameters, and d is the data. For deterministic systems such as we consider here, it is standard practice to assume that measurements are obscured by additive Gaussian noise,

$$d_i = y(t_i) + \zeta \tag{2.3}$$

where $y(t_i)$ is the (deterministic) output of the model at time t_i and ζ is standard normal random variable $\zeta \sim \mathscr{N}(0,1)$. This assumption defines a probability distribution to which Eq. 2.2 can be

applied [6] Because this construction is so common in information theory, it is often referred to as the sensitivity Fisher information matrix or sFIM [72] for reasons that will soon be apparent. In general $\mathscr{I}$ can be expressed in terms of the first derivatives only

$$\mathscr{I} = -E\left[\frac{\partial^2}{\partial\theta^2}\log\mathscr{L}(\theta|d)\right] \tag{2.4}$$

$$= E\left[\frac{\partial}{\partial\theta}\log\mathscr{L}(\theta|d)\frac{\partial}{\partial\theta}\log\mathscr{L}(\theta|d)\right] \tag{2.5}$$

Using the second form, one can show that sFIM becomes

$$\mathscr{I}_{i,j} = \sum_{k=1}^{M} J_{k,i}J_{k,j} = (J^T J)_{i,j} \tag{2.6}$$

where we have introduced the Jacobian or sensitivity matrix $J_{k,j} = \frac{\partial y_k}{\partial\theta_j}$ whose entries denote the sensitivity of prediction k to changes in parameter j. In Eq. (2.6), M denotes the number of observations.

The entries of the FIM indicate the sensitivity of the model's trajectory to changes in each pair of parameters. A high score indicates that a parameter pair has a strong influence on model dynamics, while a small score indicates a "sloppy" direction (parameter values can change a great deal without much changing $\mathbf{y}$). The curvature of the likelihood function converts distances in parameter space to distances on the manifold in data space, making the FIM a Riemannian metric tensor on the model manifold in data space. It is important to note that the physical units of parameters can strongly affect the values of the FIM. For this reason, it is common to perform dimensional analysis before sloppy model analysis as we do throughout this study.

In general the curvature of the likelihood surface does not align with the bare parameters. Rather, the characterization of the model's sloppiness aligns with the eigenvectors of $\mathscr{I}$. Eigenvalues of the FIM are related to the singular value decomposition of $J = U\Sigma V^T$:

$$\mathscr{I} = V\Sigma^2 V^T. \tag{2.7}$$

where U and V are matrices of the left and right singular vectors of J, and Σ is the diagonal matrix of its singular values. This implies that the right singular vectors of the Jacobian V are also the eigenvectors of the FIM. The eigenvectors of $\mathscr{I}$ "orient" the parameter-space into the parameter combinations most relevant for changing the model's behavior.

Imagine now that we coarsen the sampling rate by changing t_{max}. In our simple example, increase t_{max} from 3 to 6 means the model $\mathbf{y}$ is sampled at $t = \{2,4,6\}$. This procedure stretches the manifold in some directions and compresses it in others. This distortion is measured by an increase or decrease in the eigenvalues of $\mathscr{I}$, respectively. Compression of the manifold (i.e., decreasing eigenvalue) with increasing t_{max} indicates that the combination of parameters is less important to the long-term dynamics. We call the corresponding eigendirection "irrelevant". Similarly, if the manifold stretches (i.e., increasing eigenvalue), we call the corresponding direction "hyperrelevant". Directions that are neither compressed nor stretched are called "relevant" direction (Fig. 2.1B). Returning to the example in Eq. 2.1, θ_1 is relevant since its effect on the model's output is unchanged with observation time. In contrast, θ_2 is irrelevant since the exact rate of the decay matters less as time scales become very large, and θ_3 is hyperrelevant since small changes have large effects at large t. Note that θ_2 and θ_3 are functionally interchangeable if either is negative.

This procedure is similar to coarse-graining under RG flow described in reference [25] and is used to generate their Fig. 1. In our case, however, because we are coarsening the sampling rate, the total observation time increases and includes new information, i.e., observations at later times. As such, it is not a true coarse-graining and introduces the possibility of hyperrelevant directions, i.e., directions that become increasingly important such as θ_3. We will see that hyperrelevant directions are associated with the stability or instability of the equilibrium.

This method is also somewhat analogous to studies that use Sobol' sensitivity analysis to track importance at different time scales, either bare parameters or eigenvalue combinations. Such methods are excellent at providing estimates of model variability at a given point in parameter space,

and have noted both increasing and decreasing importance for model parameters of biophysical systems [73, 74]. Critics note that these methods are computationally expensive, even when implementing Morris acceleration [75], and the implications for bifurcation analysis are not immediately obvious.

In addition to characterizing bifurcations, TWIG analysis reveals two other features of bifurcating systems. First, there can be parameters (or combination of parameters) that move the location of a fixed point without causing a bifurcation. Such parameter combinations appear as "relevant" eigendirections, as the new equilibrium appears in long-time observations. These parameters need to be removed in order to correctly identify the codimension of the bifurcation. We do this by solving for the location of the fixed point with a numeric RootFind algorithm and subtracting it from the trajectory at every point. This effectively translates the fixed point to the origin and is analogous to the recentering step of Center Manifold Analysis. For limit cycle trajectories, we recenter by subtracting off the (unstable) fixed point that must exist within the cycle (according to the Poincaré-Bendixson theorem [76]).

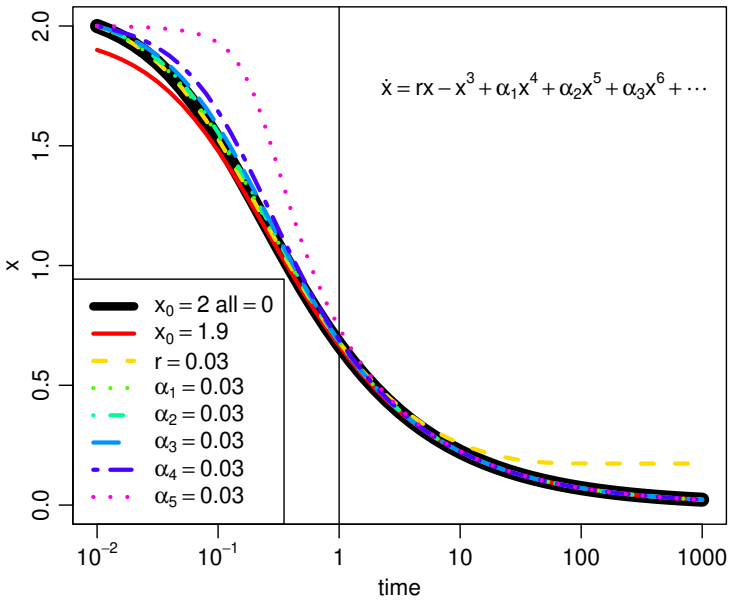

Figure 2.2 The trajectory of a supercritical pitchfork at the bifurcation point (heavy black line), and slightly perturbed from it (thin colored lines). At short time scales, y_0 (thin red) and high-order parameters (long-dashed blue) appear relevant. But, as the dynamics progress, r (short-dashed yellow) emerges as the only parameter that changes the long-term equilibrium point. This change from relevant to not (and vice versa) occurs at $t \approx 10$, and is reflected in the arch shape and changing colors of Fig. 2.3.

The second feature arises in such oscillating systems. Parameters that change the phase or frequency of oscillation without destroying the equilibrium itself appear as hyperrelevant as the accumulating phase differences becomes increasingly important at late times. Previous research has shown that such systems frequently cause problems in an information geometry framework by introducing “ripples" into the likelihood surface of Eq. 2.2. The solution is to perform a coordinate transformation so the period itself becomes a parameter. In one formulation of the FIM, this causes the manifold to “unwind”, creating a smooth likelihood surface [77] and thereby eliminating a misleading eigendirection.

Four important pieces of information come from this Time Widening Information Geometry (TWIG) analysis. First, the number of hyperrelevant and relevant directions corresponds to the codimension of the bifurcation system. Second, the square of each element of the eigenvector matrix V_{ij} indicates the participation factor of each bare parameter θ_i in eigenvector j. This last fact follows because the participation factor $p_{ij} \equiv U_i^j V_i^j = V_{ij}^2$ as can be seen by combining the definition of a participation factor [78,79] with Eq. 2.7 above. Third, the eigendirections themselves will change as t_{max} increases and parameters that influence the short-term dynamics lose their salience at long time scales. If initial conditions are included as parameters, their loss of relevance is a strong indicator that the system has been simulated “long enough" to capture equilibrium behavior. This is not a trivial concern in practice, where long numeric simulations are always fighting the accumulation of computer round-off error. Finally, at equilibrium the relevant eigendirections point along the (potentially) high-dimensional separatrix surface, and so the bifurcation can be mapped through all parameter space.

Note that this procedure works no matter the number of dynamical variables involved in the differential equation system. However, it presupposes that the model can be simulated on at least one side of the bifurcation to arbitrarily long times, *i.e.* it analyzes stable dynamics on the threshold of instability. A bifurcation that switches between two different forms of instability will not be

easily detectable with this method, since trajectories will diverge on both sides of the bifurcation. In the next section, we demonstrate how this procedure works for all common normal forms of bifurcations.

2.4 Normal-form Bifurcations

Local bifurcations can be described mathematically in a potentially infinite number of ways, but nearly all of them can be reparameterized, at least locally, to one of five kinds of normal forms. These are:

- Saddle-node: $\dot{x} = r + x^2$, where one stable and one unstable fixed point emerge from an previously uninterrupted flow at a critical value r_{crit}
- Transcritical: $\dot{x} = rx - x^2$, where a stable and an unstable fixed point exist everywhere but the bifurcation, and swap stability at the critical value
- Supercritical Pitchfork: $\dot{x} = rx - x^3$, where symmetric stable fixed points emerge from a single fixed point, which itself becomes unstable
- Subcritical Pitchfork: $\dot{x} = rx + x^3$, symmetric unstable fixed points emerge from an unstable fixed point, which swaps stability
- Hopf: a stable limit cycle emerges from what had previously been a stable point attractor. Depending on the coordinate system, the normal form is $\dot{z} = z(a + b|z|^2)$ (complex), $\dot{x} = -y + x(\mu - r^2)$; $\dot{y} = x + y(\mu - r^2)$ (Cartesian), or $\dot{r} = r(\mu - r^2)$; $\dot{\theta} = -1$ (Polar).

A method able to detect bifurcation parameters for these types of bifurcations will detect the overwhelming majority of bifurcations we are likely to encounter. The Fisher information as a function of t_{max} for each bifurcation type has a closed-form solution, which complements and

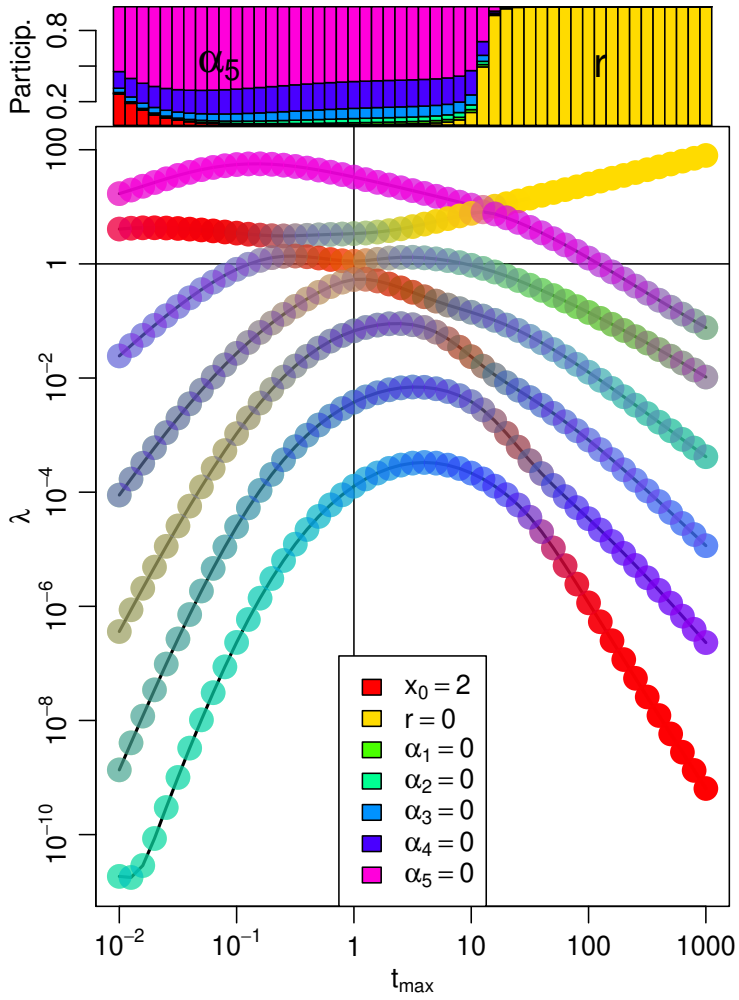

Figure 2.3 The "rainbow diagram" of the system from Fig. 2.2, showing the FIM's eigenanalysis at each t_{max}. The top panel represents the participation of each parameter in the first eigenvector ($V_{i,1}^2$ in Eq. 2.7). The leading eigenvector changes from α_5-dominant (purple) to r-dominant (yellow) at $t_{max} \approx 10$, *i.e.*, just where the α_5 trajectory is replaced by the r trajectory as most divergent in Fig. 2.2. The large panel below shows all seven eigenvalues ($\lambda_i = \sqrt{\Sigma_{ii}^2}$ in Eq. 2.7) at each t_{max}, colored as the weighted average RGB of each parameter's participation factor. Thus, the top line, corresponding to the largest eigenvalue in the top panel, starts mostly purple (α_5), but turns yellow as r dominates the first eigenvalue at larger t_{max} values. For all parameters, a small change to parameter values influences trajectory at short time scales (the rising limb) but, with the exception of r, *not* at long time scales (the descending limb). The red color in the bottom-right indicates that the initial value x_0 eventually becomes the least relevant parameter in the model. Pure colors indicate an eigenvector pointing along a parameter axis, while mixed colors like browns and greys indicate many parameters participate in the eigenvector.

validates the numerical results that we present here (see Appendix A.1). In each case, the sensitivity with respect to the bifurcation parameter, r, dominates the long-term dynamics of the system in the neighborhood of the bifurcation, no matter how many other higher order parameters are added to the normal form.

For example, a supercritical pitchfork of the form $\dot{x} = rx - x^3 + \alpha_1 x^4 + \alpha_2 x^5 \ldots$ experiences a bifurcation when $r = \alpha_i = 0$. At short time scales (e.g., where $t_{max} < 1$), the system's trajectory is strongly influenced by changes to its initial value x_0 and the higher order α terms (for $x_0 > 1$). However, later dynamics show that changes to the α_i's (and x_0) barely affect the trajectory of approach to equilibrium at 0, while small modifications to r move the equilibrium itself (Fig. 2.2). An eigenanalysis of the FIM (Fig. 2.3) quantifies these insights and clearly demonstrate the effect of coarse-graining on the system (i.e., increasing t_{max} while keeping the number of samples constant). At very short time scales ($t_{max} < .05$), x_0 and the highest order α term are the main participants of the leading eigenvector, and x_0 soon falls off as t_{max} increases; recall from Fig. 2.2 that this high-order term was equivalently able to bend the trajectory significantly until $t \approx 1$. Around $t_{max} = 10$, r begins to have a noticeable influence on the observed trajectory, and correspondingly this is the point where r becomes the dominant participant in the leading eigenvector. For large t_{max}, the leading eigenvalue increases while all other eigenvalues decrease, indicating that the system's bifurcation is codimension one. Note that in this range, small changes to the initial value x_0 have fallen all the way to the last eigenvector, indicating that the system has been allowed to run long enough that transient dynamics are removed, or at least have orders of magnitude less influence than any of the nuisance parameter α_i's. There is no significance to the fact that in this and subsequent "rainbow diagrams", the leading eigenvalue eventually begins to increase; this is simple case of an increasing line overtaking non-increasing ones and nothing inherent about the highest eigenvalue at small time scales. This can be confirmed by the change in color, indicating that the parameter responsible for the leading eigenvector has changed.

Similar figures can be produced for the saddle-node, transcritical, and subcritical pitchfork bifurcation classes. In each case, the eigenanalysis of the FIM indicates

- how long the system should be simulated, by the time it takes for the effect of the initial conditions to reach the least relevant eigenvector
- the codimension of the bifurcation, by the number of non-decreasing eigenvalues ($= 1$ for each normal form),
- the participation factor of each parameter in the hyper/relevant directions by the square of the corresponding eigenvectors (asymptotically approaching 100% r in each normal form)
- the null-space of the bifurcation surface, making it possible to track the bifurcation hypersurface through parameter space.

These are relatively simple bifurcations, where the separatrix is the hyper-plane $r = 0$. In more complicated situations where the separatrix is a nonlinear combination of bare parameters, this analysis identifies the vector normal to the separatrix. In principle, this local characterization could be extended to map that separatrix (along the hyper/relevant directions) through the high-dimensional parameter space.

Hopf bifurcations present more of a challenge, as they have a fundamentally more complex normal form without an easy analytic solution, and a trajectory which can be manipulated in more than one way. Where the first four bifurcation classes are characterized by the presence and stability of fixed points, Hopf bifurcations are characterized by a limit cycle that emerges from a fixed point, whose radius *and* velocity can be manipulated by model parameters.

Consider the following Hopf bifurcation in polar coordinates, where, as above, additional high order terms have been added:

$$\begin{aligned} \dot{r} &= \mu r - r^3 + \alpha_1 r^4 + \alpha_2 r^5 \\ \dot{\theta} &= \omega + \beta r^2 + \alpha_3 r^3 + \alpha_4 r^4 + \alpha_5 r^5 \end{aligned} \tag{2.8}$$

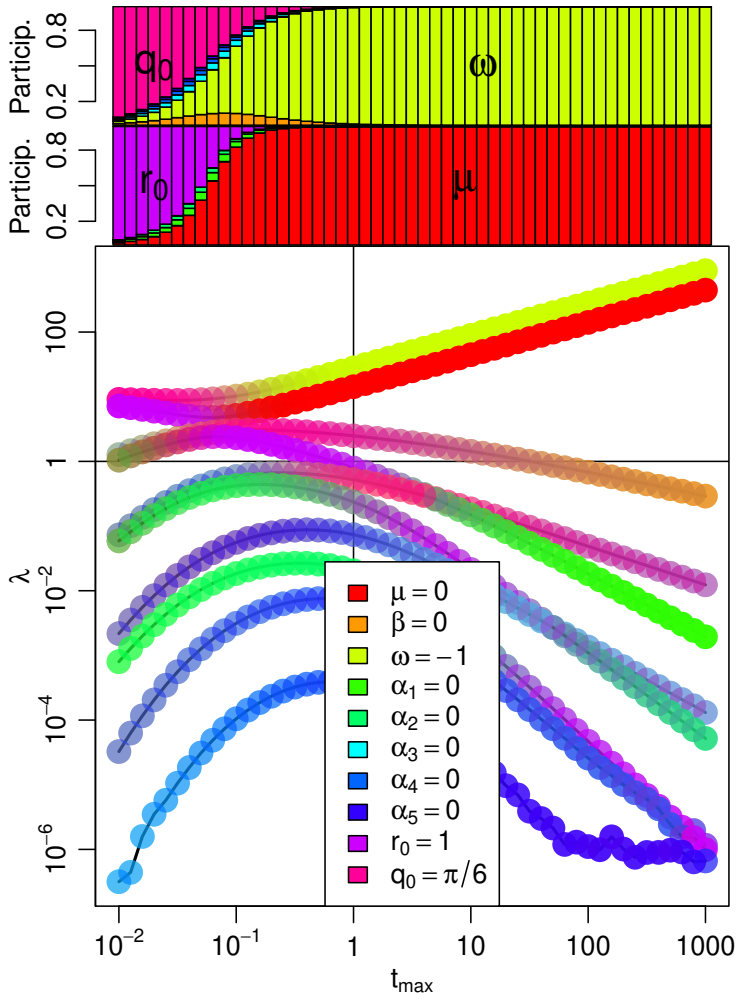

Figure 2.4 TWIG analysis of the Hopf bifurcation. The first of the hyperrelevant (rising) eigenvalues comes from the periodicity of the trajectory, whose velocity is set by ω. The second hyperrelevant eigenvalue comes from the bifurcation itself, indicating that the Hopf bifurcation is codimension-1, and the bifurcation depends simply on μ, and not some complicated combination of parameters. Note that the Hopf bifurcation is far easier to simulate at long time scales in polar form than in cartesian coordinates.

At the bifurcation point $\mu = 0$, a fixed point at the origin expands into a limit cycle. The velocity of trajectories around this cycle are primarily driven by ω, provided y values are small. Note that the periodicity of the Hopf bifurcation introduces a second hyperrelevance to long-term dynamics. Infinitesimal changes to velocity make little difference to the final position of the trajectory $F(t_{max}; y, \theta)$ *if* t_{max} *is small*, but will have an increasing effect as t_{max} grows. By contrast, μ is hyperrelevant because it is the bifurcation parameter. The increasing importance of these two parameters, in contrast to all others, is clearly illustrated in Fig. 2.4.

As noted above, this ability to characterize all normal-form bifurcations depends on the ability to isolate changes in information due to the bifurcation itself. This depends on the only source of variation in long-term behavior coming from the bifurcation, and so the preceding analyses were conducted for systems exactly at the bifurcation point. We now consider how the picture changes for dynamics near, but not exactly at, the bifurcation point. Applying TWIG just to the left and right of the bifurcation point of a pitchfork ($r = \pm.01$) shows characteristic patterns (Fig. 2.5). In these cases, we find that the bifurcation parameter is hyper-relevant on intermediate time scales ($t_{max} < 100$ in Fig. 2.5). However, on longer time scales ($t_{max} > 100$), the leading eigenvalue either asymptotes or decreases once the trajectories have converged to the fixed point, depending on whether the location of the fixed point can or cannot be controlled, respectively. In other words, when approached from the $r < 0$ side, small changes to r don't move the equilibrium ($y(t) \to 0$), meaning the exact value of r is irrelevant. But approaching from the $r > 0$ side causes trajectories to run to $y(t) \to \pm\sqrt{r}$, meaning r is relevant. Moving the system closer to bifurcation, this intermediate regime extends further and further, until at $r = 0$ it occupies the entire trajectory and r is hyperrelevant at all times.

In general, being slightly off the bifurcation obscures the effect of the bifurcation parameter to an extent proportional to the distance from the bifurcation. This is particularly useful in the case of hemi-stable bifurcations, which need to be approached from the stable side or else test trajectories will diverge to infinity (and cause computer overflow). In the case of the subcritical pitchfork, at the

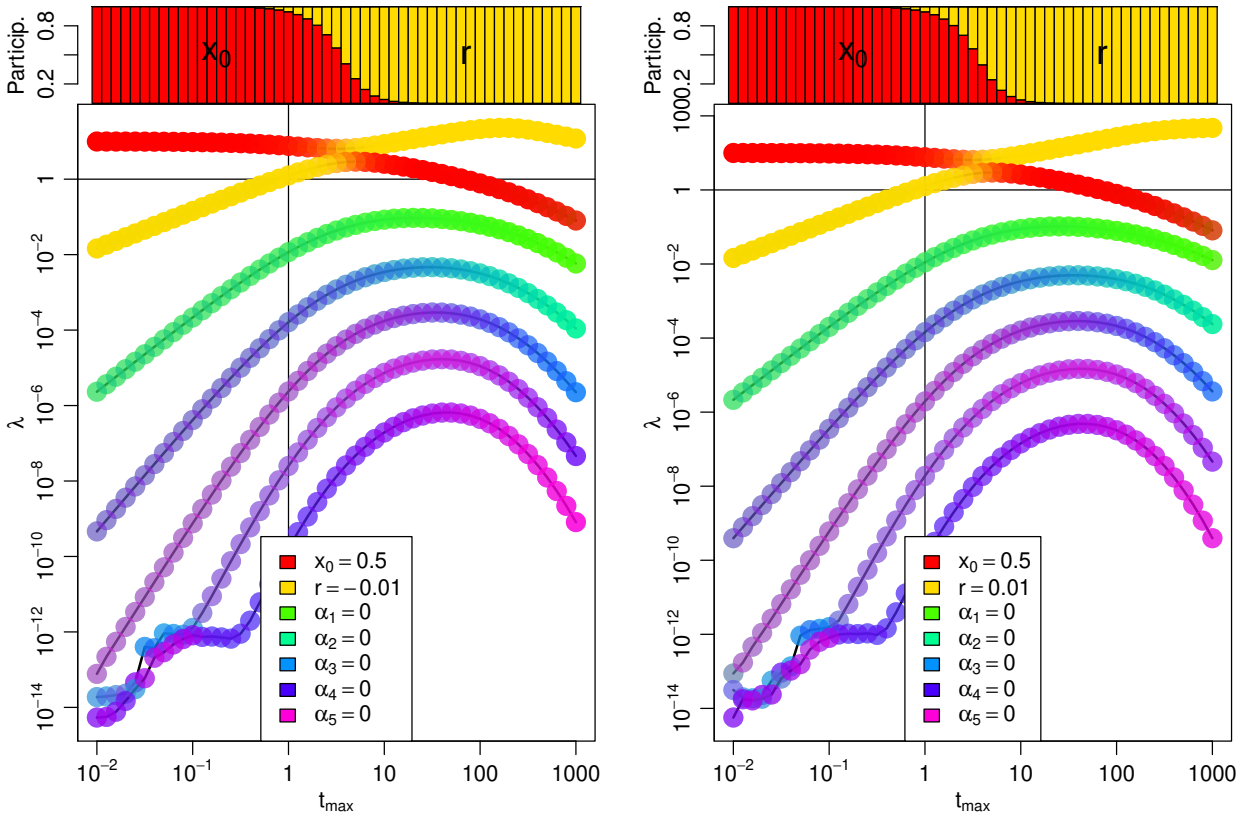

Figure 2.5 TWIG analysis near-but-not-at the bifurcation values show the diagnostic pattern of an increasing eigenvalue at intermediate time scales, rather than at all time scales above a certain limit. It is still possible to identify parameters participating in the bifurcation, and the bifurcation's codimension, though the signal becomes obscured the further one moves away from the bifurcation in either direction.

bifurcation itself ($r = 0$) the system is unstable. However, at values of r just less than bifurcation value, TWIG can be performed and the bifurcation characterized as above (Fig. 2.6).

2.5 Bifurcations in Non-normal Forms

Equations describing real systems are not typically written in one of these normal forms. So even when a researcher knows a system contains a bifurcation, it might not be apparent which one of

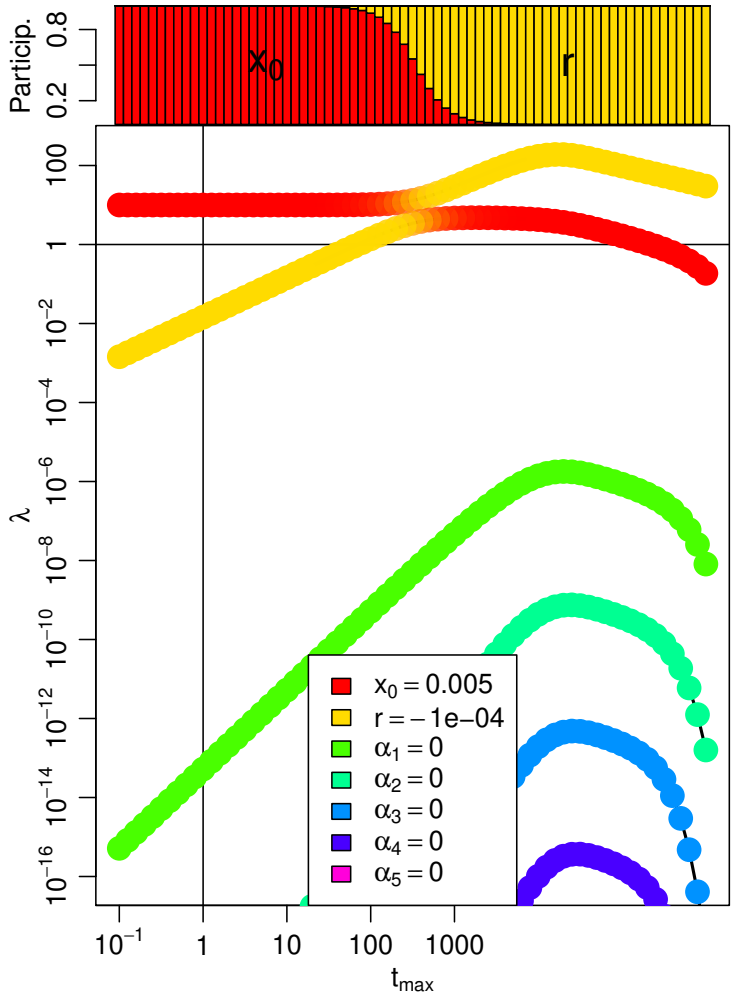

Figure 2.6 The subcritical pitchfork cannot be analyzed using TWIG at the bifurcation point ($r = 0$) because the system is unstable. However, simulations slightly to the stable side of the bifurcation ($r \to 0^-$) reveals the bifurcation parameter, though because analysis happens off the bifurcation, the peak occurs at intermediate values instead of reaching an asymptote.

these it is. For example, a model of a bead on a rotating hoop

$$mr\frac{\partial^2\phi}{\partial t^2} = -b\,\frac{\partial\phi}{\partial t} - mg\sin\phi + mr\omega^2\sin\phi\cos\phi$$

has a supercritical pitchfork bifurcation, though it might require simulating many values of r and ω to appreciate this [80]. Similarly, the equation:

$$\dot{x} = r\ln x + x - 1 + \alpha_1(x-1)^2 + \alpha_2(x-1)^3 + \cdots \tag{2.9}$$

contains a transcritical bifurcation at $x = 1$ when $r = -1$. However, this only becomes clear after reparameterizing the equation by $R = r + 1$, and $X = \frac{r}{2}(x-1)$, when the equation assumes the normal form $\dot{X} = RX - X^2 + \mathcal{O}(X^3)$. Such a substitution might not be immediately apparent to a researcher; however, time-widening information geometry clarifies the situation.

If the dynamics in Eq. (2.9) are run long enough, we observe that one eigenvalue is relevant while all others are irrelevant. Furthermore, the corresponding participation factor becomes dominated exclusively by r (Fig. 2.7). This tells us that (1) the process has codimension 1, and (2) the reparameterization involves only r. We confirm that our analysis has converged since the initial condition y_0 is the dominant participation factor in the smallest eigenvalues. However, we note that convergence occurs at a somewhat larger value of t_{max} than in the normal form examples above. Also note that transcritical bifurcations have a leading eigenvalue that is *relevant* rather than hyperrelevant, due to a quirk of the normal-form algebra. See Appendix A.2 for a thorough explanation.

But what happens when the situation is not so straightforward? Modifying the above example to the equation

$$\dot{x} = r\ln(x) + a(x-\alpha) + b(x-\alpha)^2 + \cdots \tag{2.10}$$

should still have a transcritical bifurcation for certain parameter values, but no simple reparameterization to create a normal form exists. From above, we can recognize that when a transcritical bifurcation occurs at $x = 1$ for $r = -1, \alpha = 1$. However, when $\alpha \neq 1$, in the neighborhood of

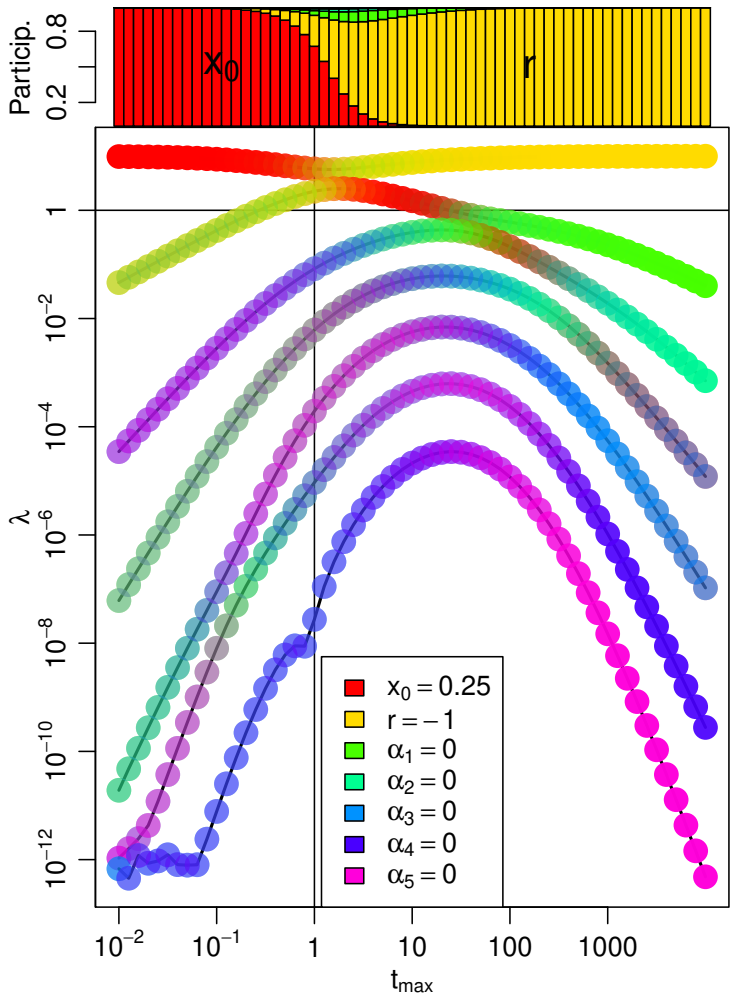

Figure 2.7 Equations such as Eq. (2.9) that are not in normal form can be interpreted using the same procedure as for normal form bifurcations. As above, the presence of just one non-decreasing eigenvalue, whose corresponding eigenvector is dominated by the single parameter r, indicates that the system has codimension 1 and the bifurcation parameter involves only r. The relevant (not hyperrelevant) leading eigenvalue is characteristic of a transcritical bifurcation.

$x = \alpha$ all the power terms are zero, but the term $r\ln(x) > 0$ if $\alpha < 1$, suggesting that no fixed point exists in that region. The appearance or disappearance of a fixed point is the hallmark of a saddle-node bifurcation, and indicates that allowing a bit of variability in the fixed point's location has introduced a second codimension to the dynamical system. This is borne out by TWIG analysis, which shows that the equation indeed produces a hyperrelevant eigenvector corresponding to the saddle-node parameter α, which controls the existence–not just the location–of an equilibrium. The transcritical bifurcation still exists and is controlled by r, as implied by the previous analysis. This example shows that even in situations with two different bifurcation classes, neither of which can be reparameterized into normal form, TWIG still allows us to efficiently and unambiguously identify co-dimension and bifurcation parameters.

2.5.1 A biophysical example

Glycolysis is a multi-step process which uses the bond energy of glucose to catabolize energy-carrying biomolecules easily usable by cells, which represents one of the dominant processes of all heterotrophic life on earth. A bottleneck in this crucial process is the phosphorylation of fructose-6-phosphate into fructose-1,6-bisphosphate catalyzed by the enzyme phosphofructokinase. The complicated five-species mass-action equation describing this reaction's kinetics can be simplified using Tikhonov's theorem and assuming low concentrations of ATP to the simple dimensionless system: [2, 81]

$$\begin{aligned} \dot{x} &= -x + ay + c_1 x^2 y + c_2 x^3 \\ \dot{y} &= b - ay + c_3 x^2 y + c_4 y^2 \end{aligned} \tag{2.11}$$

where x and y are the concentrations of ADP and F6P respectively, and the four c_i constants are nuisance parameters added to mask the system dynamics. There is a curved bifurcation surface that separates the range of kinetic parameters a, b which lead to either a fixed point at $(b, b/(a+b^2))$ when $c_1 = 1,\ c_3 = -1$ as in the canonical model, or a stable limit cycle. The separation between

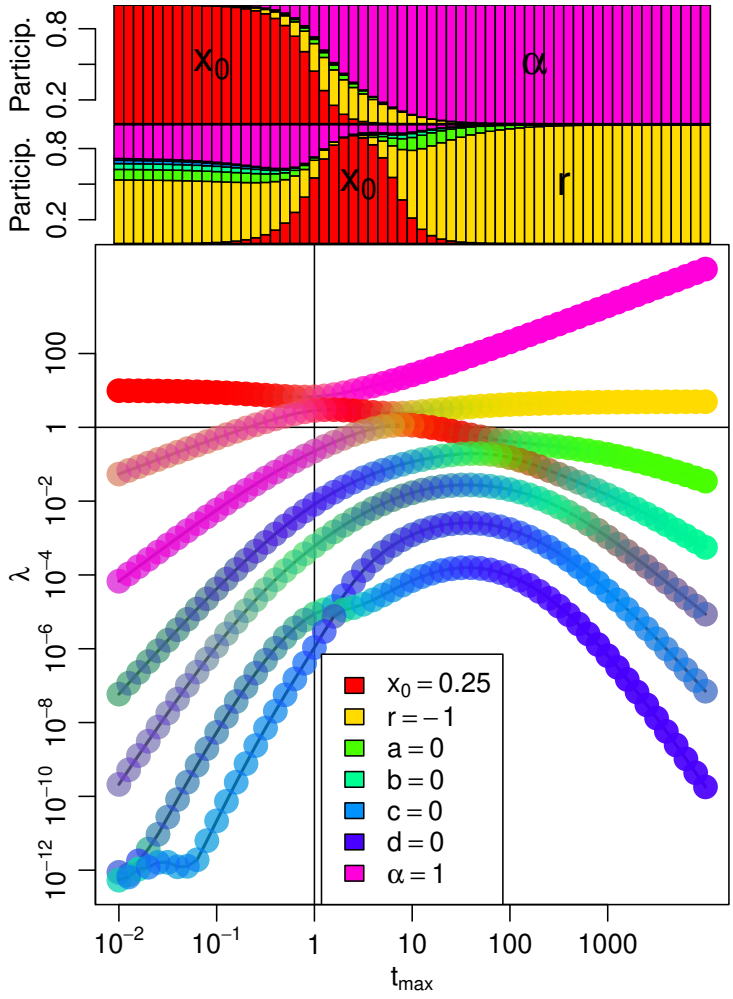

Figure 2.8 A difficult non-normal-form transcritical bifurcation such as Eq. (2.10) can be extremely challenging to analyze analytically, but sloppy analysis indicates one hyperrelevant parameter (corresponding in this case to a saddle-node) and one relevant parameter (as usual, indicating transcritical bifurcation). This means that this system has a bifurcation of codimension two. Note that the participation factor of the two leading eigendirections runs to 1.0 in the direction of α and r respectively, indicating that the system can be placed into normal form without a complicated recombination of parameters.

the fixed point and limit cycle regimes has the form $b^2 = \frac{1}{2}\left(1 - 2a \pm \sqrt{1-8a}\right)$ [82]. The resulting oscillations in glycolytic activity predicted by this analysis have been observed *in vivo* since the early 1970s [83].

A TWIG analysis of this system provides several insights, summarized in Fig. 2.9. First, even though the separatrix between fixed point and limit cycle in $a,b-$space is a nonlinear curve, because b can be reparameterized as a function of a, it is codimension one. Second, the "nuisance" parameter c_4 introduces a change in the period of the oscillations, which means that infinitesimal changes in its value cause larger deviations in final trajectory the longer the simulation runs. This shows up as a hyperrelevant direction in TWIG; however, as discussed above, it is *not* a second codimension.

2.6 Chaotic systems

Systems showing chaotic behavior have long represented a challenge to traditional categories of thinking, and the difficulty distinguishing deterministic chaos from randomness is practically its own subdiscipline [68, 84–87]. In the context of TWIG analysis, there are two characteristics of the system that need to be considered carefully.

First, unlike other systems considered here, one hallmark of chaos is long-term sensitive dependence on initial conditions, or the "butterfly effect". Because of this, a TWIG analysis carried out in the chaotic regime, in contrast to Fig. 2.3 where the parameter x_0 becomes the least relevant, will classify initial condition parameters as relevant. Note that if the chaotic system produces a strange attractor, then the initial conditions will change the location of the system on the attractor at long time scales, but not the shape of the attractor itself, which prevents these parameters from becoming hyperrelevant. That is, the maximum distance between two trajectories begun at slightly different initial conditions will eventually saturate on opposite sides of the attractor, and not increase without bound.

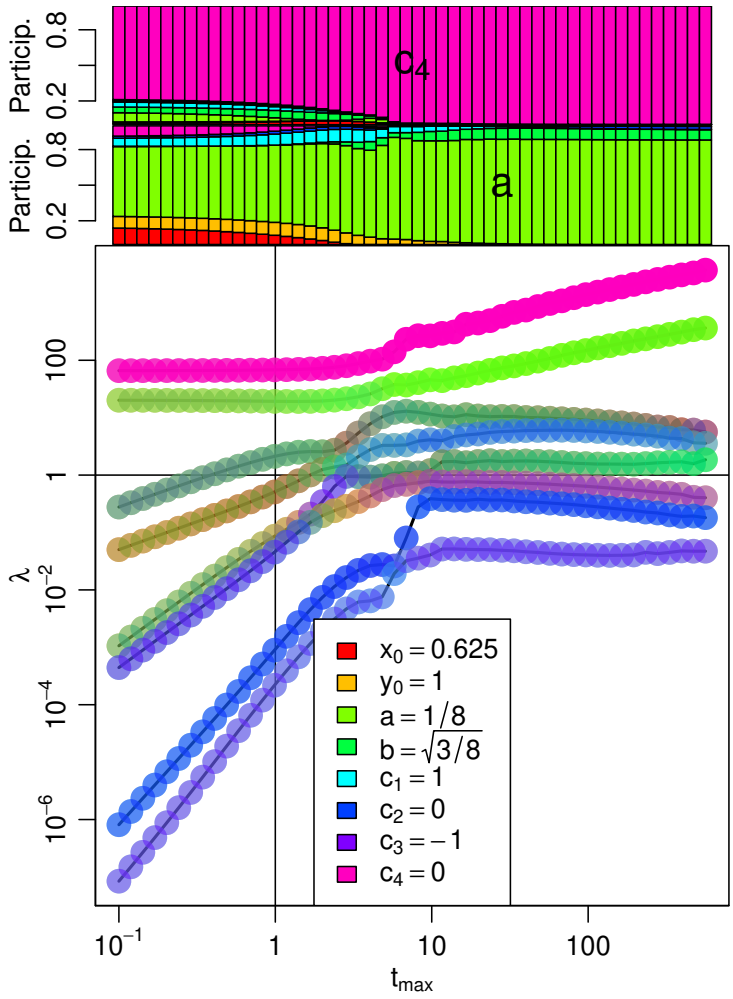

Figure 2.9 Analysis of the “glycoscillator" bifurcation (Eq. 2.11). The frequency of the oscillations are driven by c_4, while the radius of oscillations can be controlled with just one of the a, b parameters discovered by Sel’kov [2].

Second, the four classic examples of chaotic systems approach chaos through a complicated series of bifurcations, rather than a singular event as in the normal-form bifurcations above. The logistic map famously contains a period-doubling "bifurcation cascade", with the distance between these bifurcation events decreasing geometrically by the Feigenbaum constant α universally [54, 55, 88]. There is thus a "fuzzy boundary" between the periodic behavior of, say, an 8-cycle and the chaotic region as we pass through the increasingly narrow 16-cycle region, 32-cycle region, and so forth. The Hénon map experiences a similar bifurcation cascade along the line $b = 0.3$ as a increases from 1 to 1.5 [89], while the Rössler attractor has a bifurcation cascade in the opposite direction on the plane $a = 0.2,\ c = 5.7$ as b decreases from 1.5 towards 0 [90]. As we show below for a Rössler system, these boundaries are not just fuzzy, but also fractal-like. Most complex of all, the single fixed point of a Lorenz system experiences a pitchfork bifurcation at $r = 1$, whose two stable points then experience Hopf bifurcations at $r \approx 24.74$, while the unstable point undergoes a "homoclinic explosion" at $r \approx 13.926$ that produces an "a thicket of infinitely many saddle-cycles and aperiodic orbits [68]." If even these pedagogical "toy models" of chaos have such indeterminate boundaries, it is likely that examples of chaotic systems encountered "in the wild" will as well.

While the FIM may be evaluated at any point in this fuzzy region, its interpretation is less clear. The eigenvectors, which indicate the direction normal to the separatrix in other systems, lose this meaning since there is no direction normal to a fractal surface. Note that this also holds true for the intermittency route to chaos as well. Abrupt changes to chaos, with or without smooth changes in fractal dimension, also exist and would be expected to give cleaner results in the TWIG analysis [91, 92], but unfortunately are expected to be less common and less familiar to readers.

That being said, TWIG still can provide powerful qualitative insights into the nature of a period-doubling chaotic system. The Rössler attractor defined by

$$\begin{aligned} \dot{x} &= -y - z \\ \dot{y} &= x + ay \\ \dot{z} &= b + z(x - c) \end{aligned} \tag{2.12}$$

has a well-known period doubling map revealed by decreasing b along the parameter-space line $a = .2, c = 5.7$, with the fuzzy transition from an 8-cycle to chaos occurring in the region near $b \approx .70$. TWIG analysis carried out near this point reveals that while changes to b or c in this region can lead to long term divergent behavior, changes to a have a much stronger effect (Fig. 2.10). In other words, even though we "walked up" to the bifurcation region in the b direction, TWIG was able to tell us that the fuzzy bifurcation boundary was strongly angled normal to the a direction. This insight is not found in the usual treatments of the Rössler attractor (e.g., in the citations above), but can be easily verified by simulating the system in Eq. 2.12 at many sample parameter values in the region around the bifurcation. This reveals flat "sheets" of periodic behavior sandwiched between strata of chaos in the a direction (Fig. 2.11); these sheets can eventually be encountered for a fixed value of a by moving far enough in b or c, which is essentially the process diagrammed in the period-doubling map with which we started this exercise.

Above, we made the claim that TWIG analysis could be used to determine four characteristics of the system, the first being the length of time to run an analysis by the decay of sensitivity to initial conditions. For chaotic systems, this is no longer the case due to the butterfly effect. However, by removing the initial values as parameters, we see that TWIG can still be used *qualitatively* to determine the other three characteristics: bifurcation co-dimension, the null space, and the (fuzzy) normal to the bifurcation region.

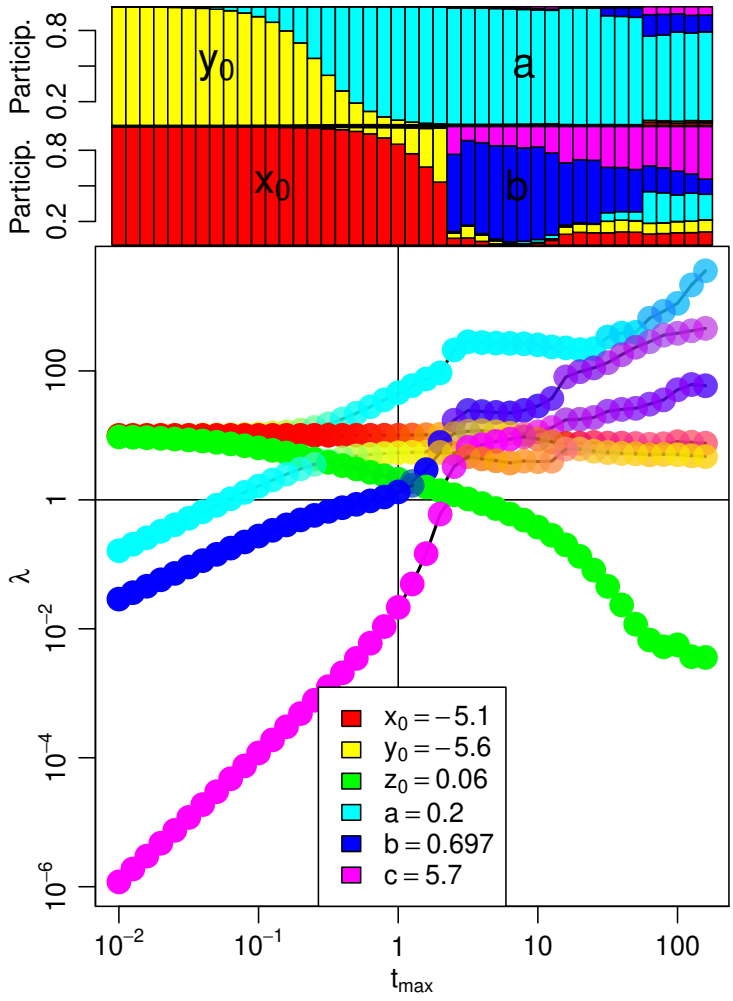

Figure 2.10 TWIG analysis of the Rössler attractor, a chaotic system, evaluated in the region of rapid period doubling just before the onset of chaos. Due to the butterfly effect, the initial conditions remain relevant at long time scales, and cannot be used to determine appropriate simulation length. However, excluding these from analysis, we are still able to qualitatively see that there is one hyperrelevant direction, dominated by a. This came as a surprise to the authors, because the bifurcation region was approached by changing values of b until a period-doubling cascade was observed, yet TWIG uncovered a greater sensitivity to a than b even in this region. This was confirmed by sampling the parameter space in Fig. 2.11.

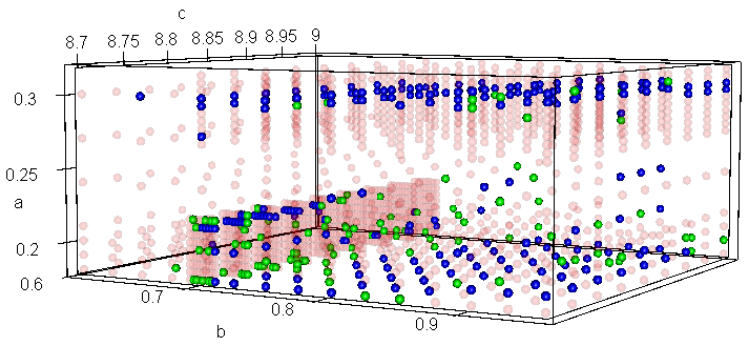

Figure 2.11 The parameter space in the period-doubling region of the Rössler attractor shows flat sheets of 8-cycle behavior (solid blue spheres) sandwiched between chaos (transparent red) in the a direction. Green spheres are simulations difficult to classify as either 8-cycle or chaotic.

2.7 Conclusion

Progressive time-dilation of the Fisher Information Matrix as realized by our Time-Widening Information Geometry (TWIG) analysis is an efficient way of characterizing bifurcations in a dynamical system. Researchers have long used eigenanalysis of $\mathscr{I}$ to characterize the "sloppiness" of a system, *i.e.* its exponential range of sensitivities to parameter changes, and recently leveraged this accumulated expertise with coarse-graining to understand phenomena occurring at distinct time scales [25, 93]. Building on these insights, we here demonstrate that as t_{max} increases, the changing eigenvalues of $\mathscr{I}$ (and the composition of the corrresponding eigenvectors) allow us to (1) characterize the codimension of the bifurcation, (2) quantify the participation of each bare parameter in the bifurcation, (3) characterize the bifurcation's hyper-surface, and (4) have an internal check on the length of time necessary to simulate the system to reach equilibrium. These are substantial insights to be gained relatively cheaply. Sloppy bifurcation analysis constitutes a powerful tool to supplement traditional analytical analysis [68, 94], and other specialized analytical tools for high-dimensional problems [32, 51, 54, 71, 95–97].

Insights derived from TWIG are useful not just for theoreticians interested in characterizing a bifurcation or reparameterizing a system to emphasize the bifurcation; it is also critical for

the process of fitting parameter values. The rainbow plots in this paper demonstrate that at the bifurcation point, simulations frequently show a separation of over 10 orders of magnitude in their parameter sensitivity, a gap that gets larger the longer the simulations run (or the more data is collected, in an experimental context). If researchers care about fitting all parameters, it is crucial to recognize that the effect of hyperrelevant parameters will overwhelm the others, so only if these parameters are fixed in the experiment can the less relevant ones be inferred [67, 97]. Future work may naturally extend the method to large systems including those derived from partial differential equations.

Our TWIG analysis has some inherent limitations. It presupposes that the model can be simulated on at least one side of the bifurcation to arbitrarily long times, *i.e.* it analyzes stable dynamics on the threshold of instability. A bifurcation that switches between two different forms of instability will not be easily detectable with this method, as trajectories will diverge on both sides of the bifurcation. However, such doubly-unstable bifurcations may be of limited practical interest anyway, as loss of stability is generally a far more common real-world problem than a change in the type of instability of a system that never was stable to begin with. Hemi-stable points (as in saddle-node or subcritical pitchfork bifurcations) are easily analyzed when approached from the stable side (see Fig. 2.6); otherwise test trajectories can diverge beyond computer tolerance at moderate time scales. A notable limitation of the method as presented here is the inability to analyze hyperbolic fixed points. Future work may additionaly leverage center manifold techniques to investigate bifurcations in such systems. We note here that absolutely unstable fixed points (*i.e.*, where every eigenvalue is positive) can be conveniently analyzed in TWIG simply by running time backwards, and analyzing trajectories at ever-closer instants to the initial divergence from the instability.

Because it is a particularly efficient method of determining important information about high-dimensional bifurcations, we anticipate that TWIG will be useful in situations with many compo-

nents where one or a few bifurcations are expected in each component. These include power grids, circuit boards, interatomic models, complex protein regulatory networks, and ecosystem-based management systems of multiple interacting populations. Such complexity presents substantial difficulties for closed-form analysis but can be tamed with insights gleaned from this method.

Acknowledgements: We thank Archishman Raju for helpful discussions and comments on the manuscript. We also thank two anonymous reviewers for their careful reading and suggestions for extensions of TWIG. This work was supported by the US National Science Foundation under Award NSF-1753357.

Chapter 3

Realistic Small Regulatory Networks Have a Rich Behavior Space

3.1 Abstract

Protein-protein interaction networks (PPIs) are large and complex, yet self-organize to sustain all known life. To help understand the building blocks of these systems, we analyze small subgraphs (called "motifs" in this study) where all proteins can interact positively, negatively or not at all. There are 132 such 3-protein networks that are topologically unique (A→B→C is a permutation of B→C→A and C→B→A and B→A→C etc, so only one such network is simulated), and 22,662 4-protein networks. Using a flexible model, we choose 1,000 biologically plausible values for the parameters, find all fixed points of the system, and determine the stability of those points. Previous research had suggested that a switch from a stable fixed-point equilibrium, to a limit cycle, to divergence to infinity was to be expected as the three possible behaviors. Instead, we determined that the behavior space of the typical motif included many different numbers of un/stable fixed points in complex combinations (a median of 6 and 12 distinct behaviors). We also discover that

divergence to infinity is not at all uncommon at known biological values, despite such a result likely being fatal to the cell, necessitating a substantial degree of selection and regulation. The effects of adding an additional node to 3-motifs is discussed, with implications for scaling up to full biochemical pathways.

- **Data dashboard:** https://oceanchaos.shinyapps.io/motif/
- **Full ouptut:** DOI:10.17632/2vsj7wr7wz.1 (hosted by Mendeley Data)

3.2 Networks and Motifs

Networks are a powerful mathematical abstraction for summarizing a system of interacting elements. Scientists have used the lens of network analysis to interrogate many different kinds of complex systems, including food webs, electric circuits, social networks, macro- and micro-scale economies, and protein-protein interactions (PPI). [98, 99]

Studies of real-world networks tend to be challenging due to their size and complexity. The internal dynamics of each node of the system is not always known, due to the large number of nodes and the difficulty in separating them from a potentially large number of interactions with other nodes. For example, in a food web there are reasons to believe the limiting factors on primary producers and top predators are different, but asking what the internal dynamics of a predator population are in the absence of other species (including its prey) has no clear meaning. For this reason, many early papers about such systems focused on collecting summary statistics of the system, rather than in-depth simulation of the (often murky) behaviors of each part of the system. Such studies of PPIs determined that networks from very different species and genome sizes had similar summary statistics, such as degree distribution, radius, and betweenness. [100]

Another option in dealing with this overwhelming complexity is to analyze small pieces of the network in isolation in the hopes of assembling overall behavior after these building blocks

are understood. This approach has the advantages of being analytically tractable, and (at least for some of the real-world subgraphs) scientists have more confidence in the internal behavior and interaction (or lack thereof) among the actors/components/species/proteins [100]. Particular subgraph topologies that occur frequently are called motifs, [41] and have been studied in detail since the 1970s, beginning with Simberloff and Diamond's rival mechanisms for food web assembly. [39, 101–103] Unfortunately, even this simple concept is fraught, since whether a motif occurs more (or less) "frequently" than expected depends on the null model one uses. A randomly constructed Erdös-Remy produces different subgraphs than a rich-get-richer scale-free network; and despite numerous detailed studies, little consensus has been reached about which assembly rules apply to different real-world phenomena. [99, 100, 104]

This paper contributes to the effort to understand the building blocks of large real-world networks by analyzing the properties and stability of all possible graphs of 3 or 4 nodes across scales of interaction rates known to be biologically attainable. The results give us insight into which topological components are difficult to push out of equilibrium, and which would need to be carefully regulated to avoid runaway feedback that could prove fatal. We use the term "motif" to refer to 3- or 4-node networks to emphasize we are considering these as small pieces of a larger system, without implying anything about their frequency relative to expectations in that larger system. However, this terminology reinforces our thinking that each such topology is under selective pressure, and so its frequency in the real world likely is shaped by the utility of its dynamics in biological systems. Inasmuch as these pressures are universal, findings that apply to PPIs are expected to also have implications for neuronal connections, economies, and the World Wide Web. [39, 41, 104] Of course, other constraints impinge on real-life networks, as will be discussed in more detail in later work.

3.2.1 Simplifications of our model

Box famously declared that "all models are wrong; the practical question is how wrong do they have to be to not be useful." [105, 106] The guiding principle that "useful" models simplify complexity as far as possible without losing contact with observed phenomenological behavior has motivated a great deal of science in recent decades. [11, 62, 64, 107, 108] Philosophers of science have even observed that model utility takes precedence over "truth" in practice, if not in theory. [109] Protein-protein interactions (PPIs) undergird the most information-rich process in the known universe–life itself–and so manifest impenetrable complexity if approached from a purely mechanistic standpoint. Our models elide a great diversity of biochemical mechanisms into a single mathematical process; the omissions we are aware of, and justifications for leaving them out, are discussed in the following paragraphs.

This paper follows the convention of representing interactions as either positive (with an arrow) or negative (with a plunger) in a network. This shorthand, of course, fails to distinguish between the very wide range of biochemical processes that can lead to these effects. PPIs are sometimes conceptualized as existing in a three-dimensional continuum from homodimers to heterologous oligomeric complexes, from obligate through facultative to true monomers, and from the permanent through the transient to truly instantaneous in a quantum sense. [110–112] Attempts to predict where in this space a protein will fall based on its three-dimensional structure have failed to find simple deterministic variables, though sophisticated classifiers are consistently improving, *e.g.*, by including quantum molecular dynamics or meta-learning strategies. [113, 114] Thinking of proteins as existing at just one point in this space may be a flawed analogy, as meta-analysis indicates as many as one protein in seven has more than one quaternary structure, [111] and the recent discovery of numerous intrinsically disordered proteins (IDPs) suggest that tertiary structure is not necessarily fixed either. [115]

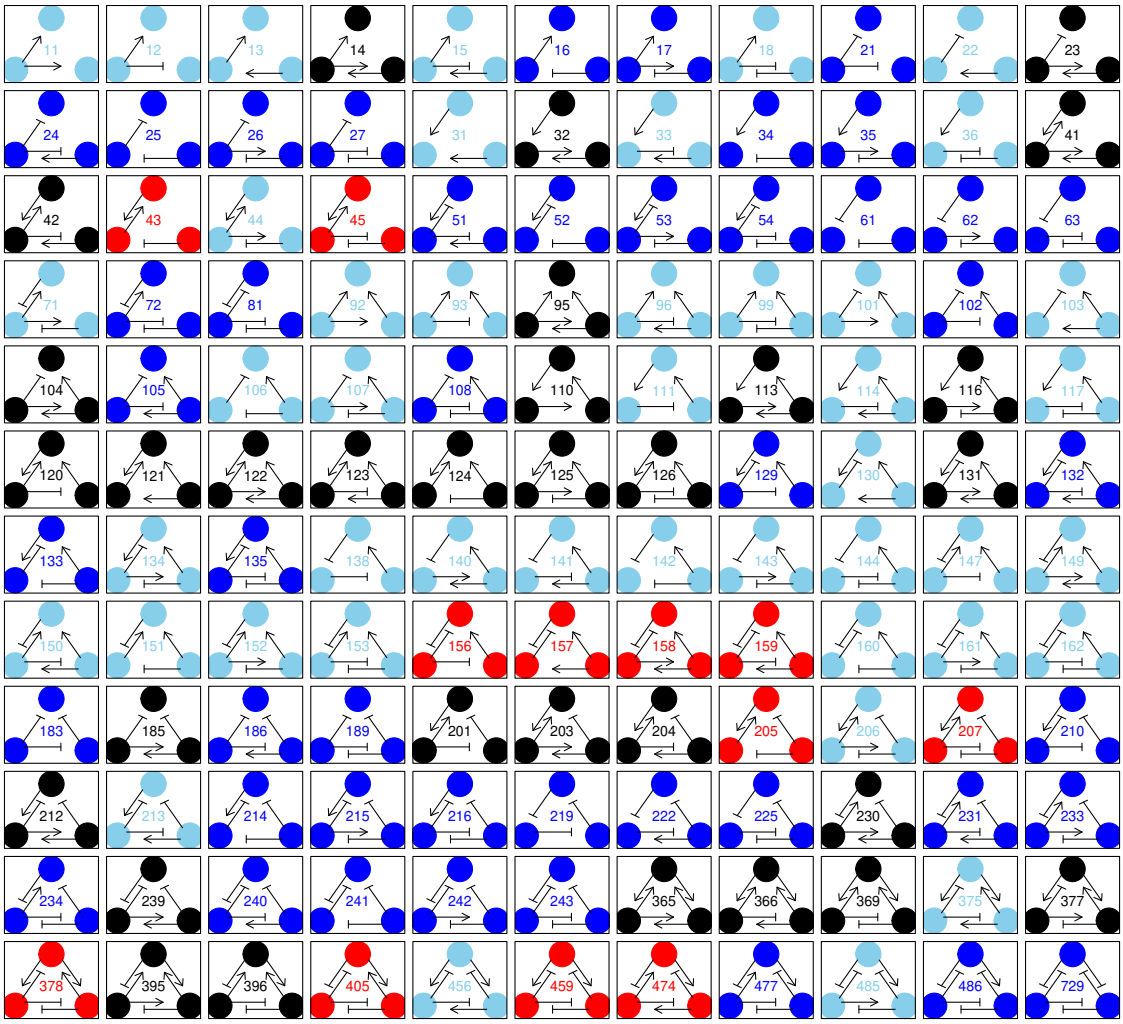

Figure 3.1 All connected, directed, 3-node motifs. Black graphs diverge to infinity, *i.e.*, have no fixed points, in at least 50% of simulations; red graphs average at least 15% unstable fixed points across simulations, and the light/dark blue graphs were the least/most stable of the remainder. N.b., each motif pictured is topologically isomorphic to up to five others motif IDs simply by switching the node order.

Complexation can be particularly difficult to model or even study. An extreme example is the 110MDa nuclear pore complex, which consists of about 1000 proteins of 34 types; the structure of the outer face was recently solved; it consist of an octomer of hetero-hexadecimer complexes (128 total proteins). [116] Even identifying such complexes in the output of a genome-wide survey experiment can be difficult, as discovering higher order correlations is geometrically more difficult than pairwise comparisons, though techniques like iterative random forests have recently sped such discoveries considerably. [117] For the purposes of this paper, we assume that complexes form as a stepwise progression of pairwise associations (A+B $\rightarrow$ AB; AB+C $\rightarrow$ ABC, etc), and that the effect of a complex is adequately captured by a topology where two or three protein species have the same direction of effect on another.

The rates at which PPIs occur span many orders of magnitude, as one would expect given the vast diversity of tasks that proteins carry out in living systems. It is perhaps not as well appreciated that the rate at which any one PPI occurs can also span orders of magnitude depending on the physio-chemical environment. As an extreme example, the collagen fibers that had remained mostly intact in a *Tyrannosaurus rex* specimen for 68My fell apart in minutes when exposed to proteases in a warm liquid bath. [118, 119] Another relatively slow process, though orders of magnitude faster than the stasis above, is the activation of a transcription factor, which will regulate the speed at which mRNAs for a particular protein are transcribed, which sets a maximum but not necessarily a minimum on the rate of the protein's actual translation. Generally, components of complexes have similar promoters so increasing the transcription of one increases the transcription of all, though evidence of strong negative feedback is lacking at least in eukaryotes. [120] Transcription factors, modifications of promoter regions, alternative splicing, and alternative coding of amino acids [121] are justifiably well-studied and certainly play a role in long-term dynamics; however, global analyses suggest that the concentration of proteins is primarily regulated post-transcription. [122] It is thought that much of this global rate is controlled by regulation of active ribosomes, [123, 124] by regulating

initiation factors or inhibiting core proteins rather than actually changing the number of proteins via degredation or stabilization. [125–127] The dynamics of this process have been mathematically modeled, and agree well with observational studies. [128, 129] However, active destruction or production through splicing of proteins certainly also plays a role, and can be activated nearly instantaneously via chemical tagging (such as methylation, acetylation, and phosphorylation), which can act as an on/off toggle for many enzymes.

As a result of this extreme diversity, a recent meta-analysis of 120 studies of protein turnover rates concluded that little consensus had been reached about average rates or even distribution shape. [130] However, it can be confidently asserted that the different methods of regulation have overlapping scales of effect, resulting in a continuous distribution of effect sizes and rates spanning many orders of magnitude. Two very large-scale surveys of mammalian proteome dynamics indicate that protein lifespan varies from one hour to one month; [48, 122, 131] we pin the extremes of degredation and synthesis rates to these numbers ($k_{deg}, k_{synth} = e^{[-6.91,0]}$ in hours^{-1}) and mirror this range for the maximum protein-protein effect sizes ($k_{AB}, k_{BA}, k_{AC}, ...$). Because we are interested in dynamics across parameter space rather than the observed frequency of these dynamics, we sample log-uniformly across the range, rather than using the more biologically realistic weighting of a log-normal distribution centered around $\tilde{3}$ days observed in some studies. For this study, we also normalize all rates to the synthesis rate (which is always 1). We do this because the rate values in isolation are arbitrary; that is, a "low" rate can be made arbitrarily large by measuring in units of month^{-1} instead of second^{-1}, and vice versa. It is the separation of scales that drives behavior, not the rate value itself.

The shape of the effect vs concentration on a PPI also depends on the underlying molecular mechanism. Simple molecular scissors, direct competition, or complete promoter-region exclusion by a transcription factor would each have a negative effect that scales linearly with concentration, albeit at different timescales; while non-competitive inhibition, localized concentration of reactants,

or reactions that depend on heteromeric complexes can be highly nonlinear. To capture this diversity, we follow the suggestion found in [132] that standard model methods such as Nonlinear Normal Mode analysis (NNM) and total quasi-steady state assumptions (tQSSA) coincide, and can be roughly captured by providing every effect with a Hill coefficient which varies from linear ($h = 0$) to highly nonlinear in either the asymptotic ($h < 0$) or exponential ($h > 0$) direction.

This paper is also motivated by making two sweeping generalizations, both of which are valid for the majority of cases but have many known exceptions. The first is that higher rates decrease stability, an assumption motivated by the unambigious effect rate parameters have on codimension-1 bifurcations (See Chap. 2) While valid in simple systems, there are many caveats depending on the complexity of the system and the definition one uses for stability. [133] Nonetheless, instabilities are known to arise in PPIs as rates increase, *e.g.*, driving a Hopf-like system past the bifurcation point from dynamic equilibrium into endogenous oscillations. [71, 97, 134, 135] Stability, in the context of dynamical systems, is an invariant set of points or subspace S such that for a set of equations $\lim_{t\to\infty} f_n(x,t) \in S$. This includes strange attractors and limit cycles; however, for the purposes of this paper, we consider "stability" to mean the presence of a fixed point with only negative eigenvalues in its Jacobian matrix (see below). [40] The presence of multiple unstable fixed points, or the total absence of fixed points, constitutes instability as understood in this paper. As the scales of interaction separate in such a system, it is common for bifurcations of fixed points to occur, leading to topological inhomogeneities and distinct behaviors. For this reason, we simulate a wide variety of rates, assuming that this is the dominant driver of instability within any one topology.

The second generalization is that unstable fixed points are, generally, detrimental to cellular health. This is, of course, not universally true; oscillatory and periodic processes such as sleeping, breathing, and the beating of the heart are necessary for life, [136] though chaotic and disordered dynamics appear to be actively suppressed even in networks where they would be expected through stabilizing PPI links. [137] Nevertheless, homeostasis implies that most things are mostly stable

most of the time, as codified in the mathematically rigorous "steady state assumption" ubiquitous in biochemical research. [138, 139] Perhaps "at equilibrium" is a preferable term to "stable", as it implies a dynamic process underlying the relatively constant level. While it is not known how much variability is "healthy", it is certainly the case that a cell unable to maintain protein behavior within very tight limits is not going to survive very long.

These two generalizations in some sense may cancel each other out, as some philosophers of biology have argued that only dynamic mechanistic explanations (cyclically organized mechanisms with complex dynamics) are capable of explaining how living organisms can be. [140] With these caveats supporting a hopeful perspective, we push forward to analyze the role topology can play in the stability of living networks.

3.3 Determining stability

We are interested in protein networks, which are, by their nature, large and complex. Exhaustive searching of the entire network space is, at present, not possible, as the number of connected topologically non-equivalent networks increases far faster than exponentially with the number of nodes.[1] We therefore turn our attention to small motifs of 3 and 4 nodes, with the understanding that these smaller building blocks will provide insight to the network as a whole. Additionally, the methods outlined here should be adaptable for a targeted investigation of some possible large topologies, a sampling of the complete topology space, or possibly even complete surveys in the future.

The first step is to generate all possible networks. In our attempt to make this method scalable to larger motifs (at least in principle), we used a vector of N edges representing the protein-protein interaction parameters k_{AB}, k_{BA}, k_{CA}, etc. There are 6 such interactions in 3-motifs, 12 for 4-motifs,

[1] The first 40 terms of this sequence are available at https://oeis.org/A053517/b053517.txt from OEIS

and $n(n-1)$ for an n-motif. In this paper, +/- 1 represented an enhancing or repressing interaction respectively, and 0 represented no interaction. For example, the repressilator might be represented by $[0,-1,-1,0,0,-1]$ assuming an edge order of $[AB,BA,AC,CA,BC,CB]$. It is relatively easy to extend such a scheme to large values of N and k if, say, one wished to model allosteric and non/competitive inhibitors with distinct dynamical equations. A (k^N,N) matrix $\mathscr{T}$ is then initialized to null interactions, representing the full topology space, then populated using the following pseudocode:

Algorithm 1 Produce all motifs of n nodes with k interaction types

```
E ← [1,1,1,...]                                   ▷ size N, edge status
while True do
    next empty row in 𝒯 ← E
    E[N] ← E[N] + 1
    if ∑E = kN then
        break
    end if
    i ←N
    while E[i] > k do
        E[i] ← 1
        i ← i − 1
        E[i] ← E[i] + 1
    end while
end while
```

The second step is to filter these possibilities to remove (1) disconnected graphs and (2) topologically equivalent graphs (Fig. 3.2) Fortunately, efficient routines for determining connectivity exist in most mathematical languages and can be implemented directly. The second step is more difficult. It first requires a list of the n! valid reorderings of edges (n.b.: not N!). This can be generated efficiently by creating an adjacency matrix of the index of each interaction. One can then swap rows and columns using the same index reordering for both (*e.g.*, if rows 1 and 2 are swapped, columns 1 and 2 must also be swapped) and store the result. In the example above this would be:

$$[1\,2\,3\,4\,5\,6] \rightarrow \begin{pmatrix} 0 & 1 & 6 \\ 2 & 0 & 3 \\ 5 & 4 & 0 \end{pmatrix} \xrightarrow{\text{permute}} \begin{bmatrix} 2 & 1 & 5 & 6 & 3 & 4 \\ 3 & 4 & 1 & 2 & 6 & 5 \\ 5 & 6 & 2 & 1 & 4 & 3 \\ 4 & 3 & 6 & 5 & 1 & 2 \\ 6 & 5 & 4 & 3 & 2 & 1 \end{bmatrix}$$

Doing this for all distinct orderings of rows and columns creates an $N \times n!$ matrix of all possible reorderings of any row in $\mathscr{T}$ that will result in a topologically equivalent motif. Starting with the first connected motif, it can then be reordered by each of these $n!$ reorderings to create a set of $n! - 1$ new E' vectors that match another topologically equivalent entry in $\mathscr{T}$.

Simply searching through all k^N topologies to find these matches is computationally inefficient. However, the row number of each reordered topology within $\mathscr{T}$ can be computed rapidly as:

$$R_i = \sum_i (E_i k^{N-i}) \tag{3.1}$$

This method reduced the hunt time for all isomorphisms by a factor of 70,000 relative to a naïve search on a typical laptop for the 3-node network, and improvements would be even greater for larger networks.

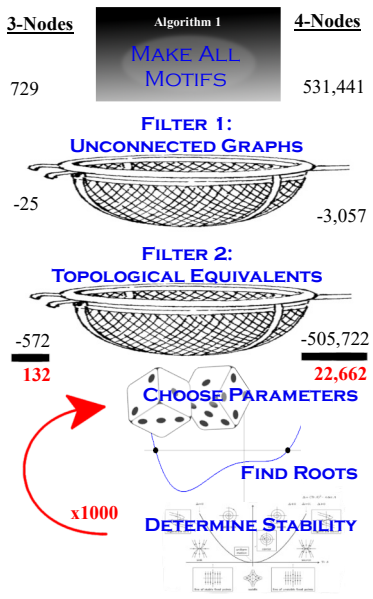

Figure 3.2 A schematic demonstrating how all motifs were generated and their stability determined. See text for detailed description of each step.

These two filter steps greatly reduce the number of topologies to investigate (132 instead of 729 for 3 nodes, and 22,662 instead of 531,441 for 4 nodes). These 132 topologically inhomogenous 3-motifs can be seen in Fig. 3.1. To explore the behavior space of each of the remaining topologies, we used the

following steps. Each protein is modeled as

$$\dot{x_i} = \alpha - \delta x_i + \sum_{i \neq j} K_{i,j} x_j x_i^h \tag{3.2}$$

where α represents the autocatalysis rate, δ the degradation rate, $K_{i,j}$ the interaction effect of j on i, and h the hill coefficient determining the degree of nonlinearity in the interaction. (For justification, see above) First, parameter values are selected for autocatalysis, degradation rates, and interaction strength. (The sign and existence of the interaction $E_{i,j}$ is determined by the network topology under investigation.) The range of reasonable parameters was derived from global surveys of mammalian kinetic rates and transcriptomics [122] [141]. Note that the parameter values are pulled from a uniform distribution, even though the distribution of rates observed in nature as reported in these studies is much closer to log-normal. This is because, for the purposes of this paper, we are interested only in surveying the space of possible conditions rather than reproducing those conditions. The range of parameter values is similar in human cancer cells, though the average turnover rates are higher in the quickly dividing cell lines. [142] Thus far, our approach is similar to that of [143], who also produced the 729 3-motifs noted above (filtering for connectedness but not for isometry) and sampled them at 1000 different parameter values to determine which motifs were able to achieve adaptation behaviors.

We then used a sophisticated root-finding algorithm to solve where the system of equations in Eq. 3.2 was equal to zero, thus determining the fixed points of the system. We implemented the roots() command in the `IntervalRootFinding` package of Julia, using the Krawczyk operator to contract the range and tolerances set to a relatively loose 10^{-6}.[2] This algorithm uses a branching tree to search the parameter space, and provides guarantees of finding all roots in a given range. Note that this finds both stable and unstable fixed points. We soon realized that because the range of possible rates spans many orders of magnitude, many such fixed points are very close to zero

[2]Written by Luis Bennet and David Sanders (UNAM). Full API documentation available at gitub. For full details on Julia itself, see [144]

for at least some of the proteins, causing round-off boundary errors in the numeric root finder. We addressed this problem by log-transforming the system (see Appendix for derivation):

$$\dot{X}_i = -\delta + e^{-X_i}\left(\alpha + e^{hX_i}\sum_{i\neq j} K_{i,j}e^{X_j}\right) \tag{3.3}$$

where $X_i = \log x_i$ and the other parameters are as above. We used the bounding box $[e^{-100}, 1000]$ in $\mathscr{R}^n$ for the $n-$motifs. In practice, root-finding was the rate-limiting step in the whole process.

Once this set of fixed points had been determined, the values of each fixed point were plugged into the corresponding Jacobian matrix for the system

$$J_{ij} = \begin{cases} e^{-x_i}[\alpha + e^{hx_i}(h-1)(e^{x_j}K_{j,i} + \ldots)] & i = j \\ K_{j,i}(x_i e^{h-1} + x_j) & i \neq j \end{cases} \tag{3.4}$$

The stability of each fixed point was then determined by calculating the spectral radius of each Jacobian matrix:

$$\sigma = \sup[\Re(\lambda)] \tag{3.5}$$

that is, the largest real part of any of the Jacobian's eigenvalues. For continuous time PDEs, spectral radii < 0 are diagnostic of intrinsic stability (*sensu* [40]). It was recently demonstrated that this criterion not only demonstrates that a network is intrinsically stable when the interactions happen instantaneously, but also stable under any time-lag condition. [145]

Note that there are at least nine other acknowledged definitions of stability, not all of which yield the same conclusions on a macroscale. [133] While it is likely that most definitions will correlate with spectral radius, there would be some quantitative differences in the shape–if not the direction–of the relationship between topology, rate, and stability. For example, Holing's resistance (the barriers to switching between steady states) is related to the number of fixed points, and the absence of chaos similarly depends on the absence of unstable fixed points. Similarly, a limit cycle can be stable; but because it necessarily forms around at least one unstable fixed point by the

Poincaré-Bendixson Theorem, even in high dimensions, [76, 146] it misleadingly appears in our analysis as unstable (though not divergent).

Having thus arrived at a stability estimate for one parameter set for one motif, we then bootstrap, choosing $B = 1,000$ sets of parameters for each motif. These runs were distributed across 100 nodes of the supercomputer cluster at BYU's Office of Research Computing, requiring approximately 25,000 hours of total CPU time on Intel Broadwell (2.4 GHz) processors with access to 4G RAM equivalents.

We note here a computational difficulty in determining the number of fixed points. Because the rate parameters can span four orders of magnitude, it is entirely possible to find a *nearly* steady state where one protein level is strongly constrained, and secondary dynamics among the other proteins are very slow. These slow dynamics, if occurring below the tolerance limit of the root finding algorithm, may show up as a "fixed plane", or rather multiple coplanar "nearly fixed points". A secondary step was included in our algorithm to ensure fixed points were genuine, rather than simply below a root-finder's threshold for ≈ 0: the set of fixed points was scanned, and if any two were found to have a value of the same coordinate within 10^{-5}, this value was set and a lower-dimensional root-find with more stringent limits was performed.

3.4 Behavior space

The number of distinct behaviors revealed by this analysis is astonishing. Far from the simple transition between equilibrium to oscillation under certain rigorous conditions we expected from previous studies, every 3-motif has at least two topologically inhomogenous regions of parameter space, and 99.97% of 4-motifs do as well (all but 7 of 22,662). The average across all topologies was 5.0 and 9.1 distinct numbers of fixed points for a sample of 1,000 parameter combinations for 3- and 4-motifs respectively.

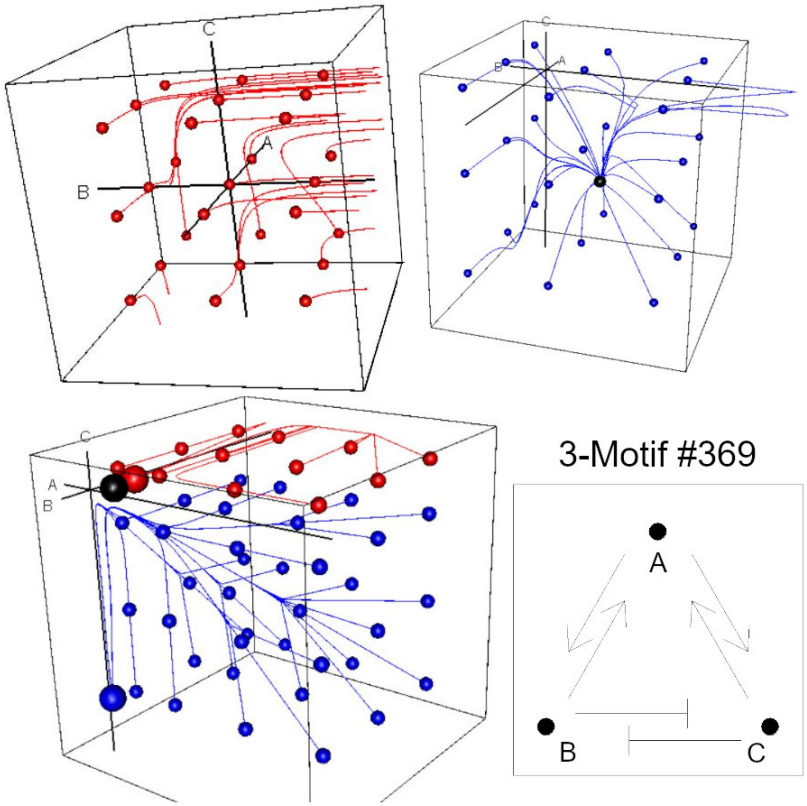

Figure 3.3 The behavior space of simple motifs was unexpectedly rich. The most complicated of the 3-motifs was #369, which showed ten different numbers of fixed points corresponding to 10 inhomogenous flow-field topologies Top left: all starting points (small red balls) diverge to infinity. Top right: all converge to a global equilibrium (large blue ball). Bottom left: two stable (red and blue) / one unstable (black) fixed point. Seven more complex behaviors exist, but are difficult to visualize.

This finding came as a surprise, since a good deal of previous theoretical and experimental effort had focused on the represselator 3-motif (#219), demonstrating that it could switch from a point equilibrium to a stably oscillating limit cycle. [147–150] It turns out that 3-motif #219 is by several measures the least behavior-rich of all motifs, and also the least likely to run away to infinity (have no fixed points, stable or not). By contrast, 3-motif #369 had 14 different behaviors (combinations of un/stable fixed points), including 0 fixed points (runaway), 0+1 un/stable points, (1 or 2)+0 un/stable points, (0 or 1)+2 un/stable points, and (0 up to 9)+1 un/stable fixed points (Fig. 3.3) Another notable topology is #314, which has the 1/0 behavior across 21% of its parameter space, nearly double any other and much more than the median 0.3% across all 3-motifs. Of course, 4-motifs vary even more, with a median of nine distinct counts of fixed points and as many as 30 distinct fixed points, corresponding to a median of 12 and maximum of 31 distinct behaviors.

We use 12 different measures of behavioral diversity to characterize this wide range of behaviors (Tab. 3.1). Different measures of the diversity of behavior space tended to correlate with each other and the total regulatory weight of the motifs, defined as $E_t = E_+ - E_-$ where E_+ is the number of up-regulating edges, and E_- is the number of down-regulating edges. Notably, the likelihood of runaway dynamics increases with E_t ($\rho = 0.76$ for 4-motifs), while the regions of parameter space including at least 1 unstable fixed point decreased ($\rho = -0.75$ for 4-motifs, both values similar for 3-motifs). Note that the relationships between statistics and regulatory weight is not necessarily linear. The number of behaviors and fixed points is particularly sinusoidal across the possible range (Fig. 3.4).

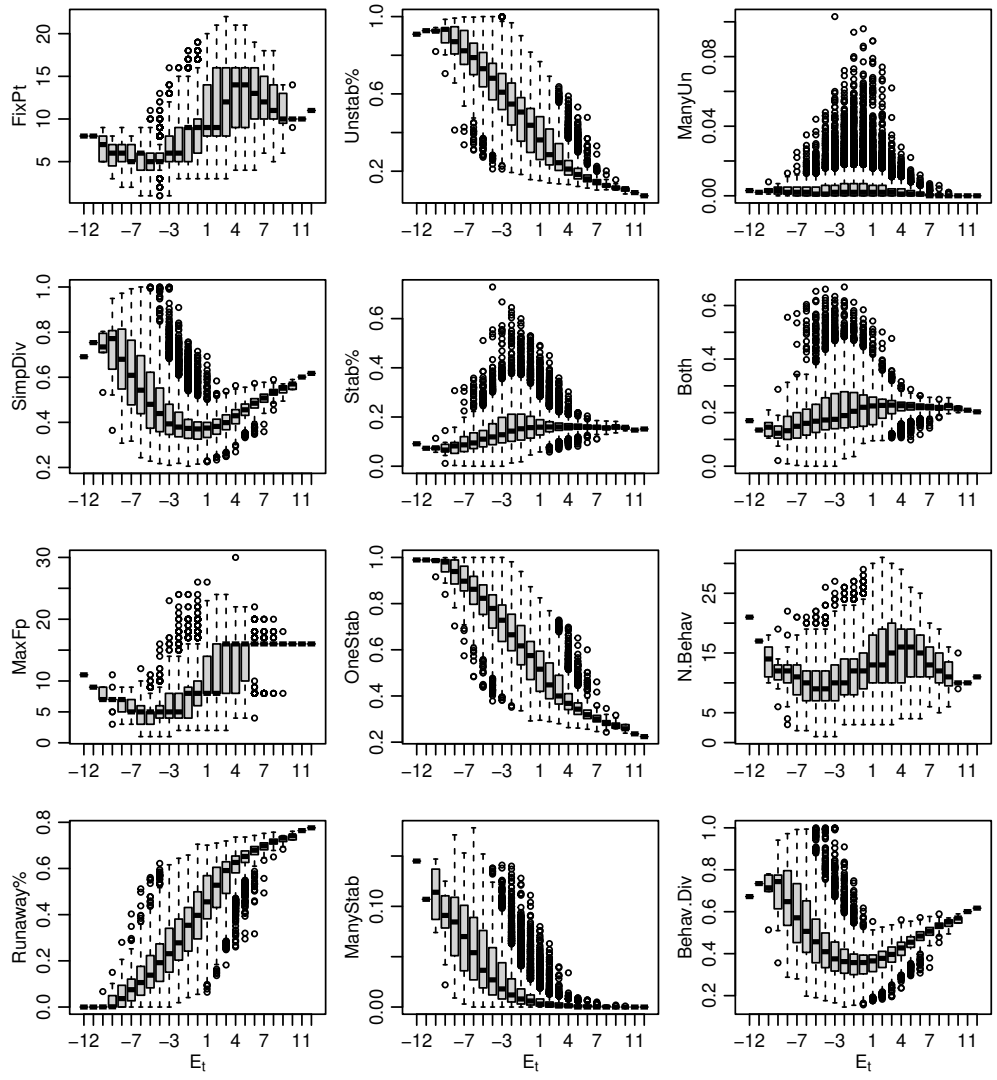

Figure 3.4 Various statistical measures of stability correlate with the total regulatory direction of the 4-motif E_t (defined in the text). Similar patterns appear for the 3-motifs but are less apparent due to the smaller number of total distinct topologies.

Table 3.1 Variables measuring motifs' behavioral diversity

VARIABLE	DEFINITION
FixPt	The number of unique counts of fixed points.
SimpDiv	Simpson's diversity of fixed point counts $\Lambda = \sum p_i^2$, or the probability of two simulations having the same number of fixed points
MaxFp	The largest number of fixed points observed
Runaway%	The fraction of simulations with no fixed points (all trajectories diverge to infinity)
Unstab% Stab%	% un/stable fixed points averaged across all sims (no fixed points = 0)
OneStab	Sims with 0 un/1 stable fixed points
ManyStab	Sims with 0 un/2+ stable fixed points
MinCycle	Sims with 1 un/0 stable fixed points (related to the number of limit cycles)
ManyUn	Sims with 2+ un/0 stable fixed points
Both	Simulations with 1+ un/1+ stable fixed points
N.Behav	Number of unique combinations of un/stable fixed points. 2/1 and 0/3 count as two behaviors, but just one FixPt
Behav.Div	Simpson's diversity of behavior counts

3.5 Predicting 4-Node behavior from 3-Node subgraphs

One of the most critical, yet least understood, aspects of this area of research is the degree to which the stability of subgraphs determines the stability of the overall system. An exhaustive search of behaviors across all topology space is not feasible. A back-of-the-envelope calculation suggests that if it takes 25,000 hours to sample the 22,662 motifs with 4-nodes, then it would take 3,000 years of CPU time to sample the 29,174,514 motifs with 5-nodes, though in reality it would take much longer due to the increased time required to find individual roots in the more complicated system of dynamical equations. Therefore, a thorough understanding of the in/stability generated by adding a fourth node to a 3-motif system may provide at least qualitative shortcuts to understanding larger interaction networks.

The addition of a fourth node creates the possibility for six new edges, positive or negative or zero, and thus in principle $3^6 - 1 = 728$ possible 4-motifs can be created from any 3-motif (the -1 representing the disconnected 000 option). In practice, over 75% of these possible extensions

are isomorphic to each other, or extensions of different 3-motifs. For example, for 3-motifs with three-way symmetry like the repressilator (#219), all additions of a node via one positive edge are isomorphic when directed inward (3) or outward (3); the same holds for adding the node via one negative edge, and becomes even more involved for the $3\ C(6,2) = 45$ ways of adding two edges.

The small gains theorem states that any two systems $S_1,\ S_2$ that are stable themselves ($|S_1| < 1,\ |S_2| < 1$) can be coupled in such a way to make a new system that will be stable if $|S_1| \cdot |S_2| < 1$, where the norm operator is a test for system stability. In our case, we use the spectral radius, defined as the L_∞ norm of the system's eigenvalues; however, the result holds for any induced norm. [151] Defining S_1 as the 3-motif system and S_2 as the fourth node implies that the addition of the fourth node should ONLY increase total system stability as $|S_2| < 1$ due to the $-\delta x_i$ term in Eq. 3.2 and $-\delta$ term in Eq. 3.3.

This turns out not to be the case when the nature of the coupling between the systems includes positive feedback loops of sufficient magnitude to overcome the degradation term; that is, the gains are not small. For example, 3-motif #219 (the repressilator) is the most stable of all motifs with only 2.2% of biological parameter space diverging (lacking any fixed points), 8.7% with one unstable fixed point (surrounded by a stable limit cycle), and 89.1% with one stable fixed point. These three behaviors of the repressilator also represented the minimum number of behaviors observed for any 3-motif. Adding one edge with a direct positive feedback loop (*i.e.*, k_{CD} and k_{DC} are both positive) creates a network isomorphic to 4-motif ID#1839.[3] This motif diverged in 50.4% of its sample space and had at least one stable and unstable fixed point in 31.4% of cases. Including positive feedback loops to all three nodes of the repressilator (4-motif #147767[4]) pushes the fraction of

[3]The 23 other 4-motifs isomorphic to this one are: 4723, 13647, 16527, 27759, 30643, 39571, 42451, 118607, 121519, 144847, 158927, 158959, 159247, 237711, 240595, 249807, 275731, 354515, 357427, 371795, 371827, 372115, and 380755

[4]Due to symmetries, only 8 distinct topologies belong to this group. The other seven are 147767, 159287, 252759, 264279, 267163, 278683, 372155, and 383675

runaway sims up to 68.5%, and drops our stability index from 28.9% to 16.0%. However, another way of looking at this topology is a negative feedback loop imposed on the edges of a three-star of positive feedback. Without this negative feedback loop (4-motif #365[5]), stability drops slightly to 14.8%, while divergence remains all but constant at 68.3%.

These results illustrate a common theme: it seems to be easier to disrupt a stable motif than to stabilize an unstable one. The 10 least divergent 3-motifs had an average of 5.1% of simulations run away (no fixed points); adding a node with a single positive feedback loop increased the runaway risk by nearly an order of magnitude to 48.7% (n=24 unique 4-motifs, mean pairwise difference = 43.6%). By contrast, the 10 *most* divergent 3-motifs had no fixed points in 64.7% of parameter space, but adding an activated repressor (A activates D represses A) was only able to decrease this percentage to 56.9% (n=24, mean pairwise difference = 7.8%).[6] This generally unimpressive trend obscures an interesting finding: a sub-population of 4-motifs was far more sensitive to stabiliza-

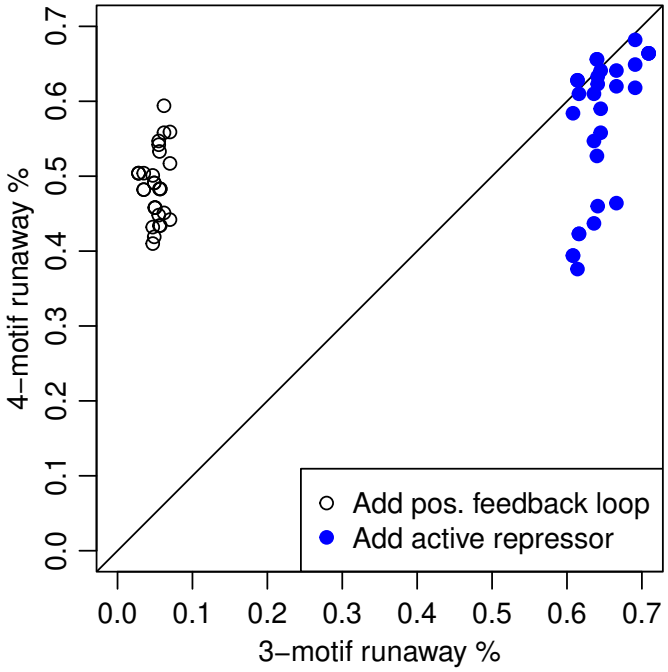

Figure 3.5 The 10 most stable 3-motifs are disrupted significantly when a 4th node is added to the network with a positive feedback loop to any node. By contrast, the 10 least stable 3-motifs become only slightly less unstable when an activated repressor is added to the network, though there are a minority of nodes where the effect is larger than the others; n=24 unique topologies generated in both cases, +48.7% vs -7.8% simulations runaway (*i.e.,* have no fixed points).

[5]Also 29165, 239149, and 262765

[6]The ten most stable 3-motifs are (in order) 219, 27, 243, 81, 729, 222, 225, 189, 25, and 231. The ten least stable are 185, 369, 395, 131, 41, 203, 113, 366, 122, and 365

tion than others, including many that included the the same 3-motif as a graph minor. This finding implies that stabilization can be effective if applied to key nodes in the topology, while destabilization can enter the network anywhere to a similar strong effect (Fig. 3.5).

Thus far, this study has emphasized the role topology has in determining behavior, but it is important to recognize that within a topology decoupling of rate scales can drastically alter the stability of the system. So, while each motif has its own set-point for where the degradation rate is able to maintain equilibrium, within the motif is a strong inverse correlation between the risk of runaway and the degradation rate. A sufficiently high degradation rate can make even the least stable 3-motif have as low a probability of diverging as the most stable 3-motif at a low degradation rate (Fig. 3.6). (Recall that all rates are rescaled relative to the synthesis rate, which therefore is always represented by 1 or log(1)=0. Thus, while the raw degradation rate spans four orders of magnitude form 10^{-3} to 10^{0}, the scaled degradation rate can span more.) Similarly, the nonlinearity of the protein interactions–from asymptotic when the Hill coefficient is negative, to linear at 0, to exponential when positive–has a strong, though complex, effect on the stability index of the motif. Generally, slightly negative values show a peak in stability; lower values tend to allow unstable fixed points to occur more frequently, while higher values lead to runaway dynamics. Unlike the degradation rate, Hill coefficients are unable to reverse the effects of network structure; *i.e.*, at very low Hill coefficient values, extremely stable motifs become only moderately stable, while extremely unstable motifs become only moderately unstable. (Fig. 3.6)

3.6 Conclusion

Previous work on the complex dynamics of protein networks emphasized either the different behaviors of one network (*e.g.*, the transition from stable equilibrium to a limit cycle about an unstable equilibrium in the represselator), [149, 150] or scan all networks to find one behavior. [143]

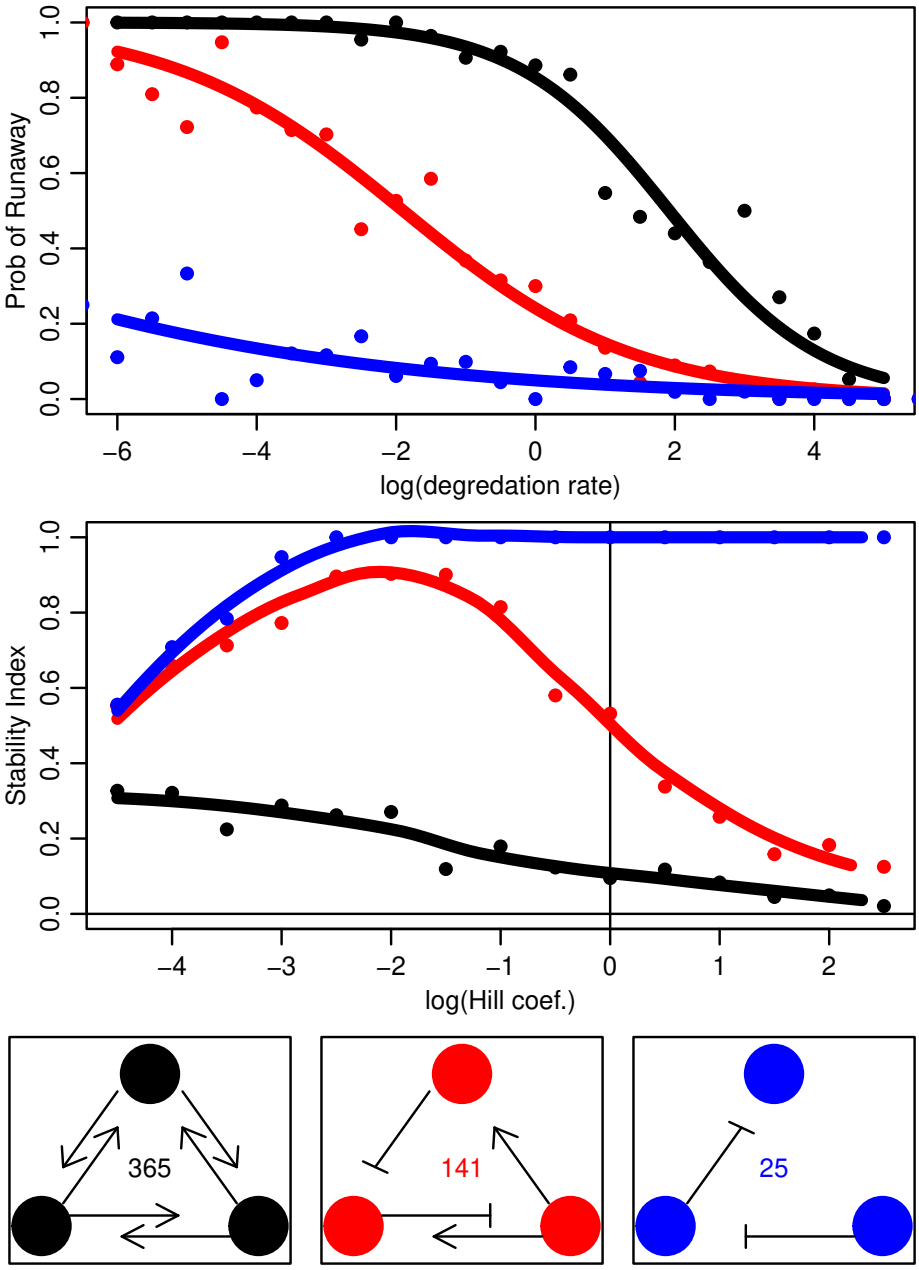

Figure 3.6 The effects of model parameters within the least (black), median (red), and most (blue) stable motifs. High degradation rates can make even the most unstable motif as stable as the most stable motif at low degradation rates. The Hill coefficient typically has a motif-dependent optimal value for creating stability, indicating that some degree of nonlinearity helps the system maintain equilibrium. In some cases, this optimum is shifted so far to one direction that the optimum lies outside the simulated range.

Our general survey discovered that almost all motifs display far richer behaviors than the previous work would suggest. Furthermore, the jump from a median of 6 distinct configurations of stable and unstable fixed points for 3-motifs to 12 for 4-motifs suggests that this richness will only increase as protein networks become larger and larger. While some of these networks were dominated by just one behavior, 91.7% of 4-motifs had a Simpson's diversity of under 50% (meaning, any two sets of random parameters were more likely than not to generate different behaviors).

This work also clearly demonstrates that, while topology exerts a strong central tendency on the diversity of behaviors, rates can be adjusted upwards to disrupt even the most stable structures and down to tame even the most divergent, albeit to a lesser extent. The dual role of rate and structure suggests that no topology, no matter how unstable, will be completely prohibited by selection, as it can be controlled by careful regulation. However, these unstable topologies are likely "weak points" in the overall regulatory superstructure of life; a global increase in protein turnover rates is more likely to cause network collapse in these motifs than others. This opens up a potentially fruitful line of investigation for investigating what specific biomolecules are responsible for shortening life-spans in organisms with high protein turnover rates from across the animal kingdom. [152–154] Conversely, it suggests that moderating these interactions may decrease protein turnover-related processes like the aging rate and/or cancer risk. [155] Future studies will investigate these intriguing possibilities.

We hope these results will encourage investigators to think more expansively about the possibilities of protein behaviors, beyond the standard "steady state vs oscillation" paradigm. Navigating this diversity is a conceptual challenge and computationally intensive. To help researchers broaden hypotheses, we make the results of nearly 3 years of CPU time freely available on Mendeley Data here, and can be browsed via a dashboard here. A wider appreciation for behavioral diversity in the building blocks of networks should generate fascinating research into the multiple roles molecules can play in the diversity of cellular functions.

Acknowledgements: This work was supported by the U.S. National Science Foundation under Awards No. NSF-1753357, and No. EPCN-1710727. We thank the BYU Office of Research Computing and the Fulton Supercomputing Lab for use of their facilities.

Chapter 4

The instability of known protein regulatory networks

4.1 Abstract

The study of motifs, small subgraphs of the overall protein regulatory network, has been limited by data to graphs that are undirected, unsigned, small, prokaryotic, or some combination thereof. Using a recent combination of databases, we analyze over 100,000 3-motifs and 3,000,000 4-motifs in humans and a parallel dataset in mice, representing much of the total biological knowledge of signed directed protein-protein interactions (PPIs) in the two species. While we confirm a similar small set of enriched topologies as previous studies, analyzing the stability properties of these motifs suggests that positive selection for stability has counter-intuitively *not* played a major role in determining motif abundance. This suggests that function and adaptability may play such a major role in network evolution that the role of stability is, for the most part, masked by these other concerns.

4.2 Introduction

4.2.1 Stability and systems biology

Stability is widely seen as a requirement of biological processes and indeed underlies biological reasoning in general. [156–158] Although there has been some philosophical resistance to including evolutionary thinking in systems biology as a rigidly defined subdiscipline, [159] the assumption that biological networks at all scales (biomolecules through ecosystems) adapt to conditions via natural selection was so widespread as to never need a serious defense among practitioners. [160–162] Evolution toward greater stability, or toward the closely related concept of robustness to perturbations, has therefore been a *leit motif* in the systems biology literature. [163] An iconic example was study of the anterior-posterior patterning of *Drosophila* embryos by morphogens; an elegant series of articles (1) argued that a simple diffusion-from-source model was not sufficiently stable to explain observed successful embryological development in the face of natural variability, (2) deduced strong limitations on the parameter space of a model that could maintain the pattern, and (3) demonstrated experimentally that biological mechanisms did indeed exist (though they were previously unknown) to maintain the system in that small region of parameter space. [38, 164–166] Similar reasoning has motivated experiments in evolution for optimal resource use, [167, 168] been used to explain inefficiencies in processes like chemotaxis and transcription that require high fidelity as a sort of "kinetic proofreading", [169–171] and–most relevant for us–the observed frequencies of biological network structures. [41]

Unstable networks are expected to be rare for the same reason as unstable airplanes: they soon vanish and are replaced by more stable alternatives. This occurs not just because of regulation catastrophes, where necessary metabolites are driven to zero or protein concentrations are driven towards infinity, [172, 173] but also due to information failure in biochemical circuits necessary for life. [174] (Note that in this paper, we define a network as unstable for a given set of rate constants

if at least one protein concentration will go to infinity for all starting conditions.) Furthermore, it is known that metabolic and regulatory networks "rewire" on-the-fly in response to temporary conditions, such as DNA damage [175], nutrient pulses, and many others. [176, 177] Such rapid responses suggest that a great deal of potential regulatory network plasticity is being canalized (*sensu* Waddington [178]), and experimental evidence is mounting to support this supposition. [95,179,180] Therefore, regulatory networks should be readily malleable to evolutionary pressure in the face of better alternatives. The general proposition that the most stable way of performing a defined function would be favored and repeated at greater frequency in future generations is thought to explain the observation that some motifs were greatly enriched in real-life bionetworks while others are conspicuously absent; many networks can perform a function, but relatively few can perform them robustly and become common. [37, 38, 41]

However, stability is not the only force at work on bionetworks; after all, a universally suppressed system with no activity represents an ultimate in stability, but in a biological context also represents death. Networks must have activity, that is have a dynamic function, to exist at all, and this function must persist even when conditions change; function, stability, and adaptability interact in complex ways, and realized bionetworks represent the end-product of a complicated evolutionary past of trade-offs among these goals. [181] As the example of segmentation shows, unstable networks can persist if the function is important enough to justify the cost of maintaining a regulatory structure in the small region of parameter space that is stable. Similarly, simulations in a constant environment evolve accurate and efficient networks that are far less modular than those observed in nature, but simulations in unpredictable environments justifies naturally occurring inefficiencies because they result in necessary adaptability. [182–185]

This paper tests the consensus view that stability is one of the dominant evolutionary forces on regulatory PPIs, and seeks to detect its effects across known eukaryotic pathways in humans and mice. If stability is the dominant, or one of the dominant, factors shaping regulatory networks, then

such real-world networks should show clear signs of evolving away from unstable structures. In general terms, this would mean (1) few subgraphs that drive protein levels to infinity at biological interaction strengths, (2) relatively few up-regulating links, (3) networks that are difficult to fragment if an interaction or protein vanishes due to mutation or a changing environment, and (4) many feedback loops to prevent runaway processes. Surprisingly, the evidence for these reasonable suppositions is either weak or contradictory.

4.2.2 Biological data

Some background on how the technical details behind this surprising conclusion was reached is called for. Methods for studying protein network regulation at the genome-wide scale are dominated by microarray analysis of coimmunoprecipitation (coIP) assays. This powerful method has allowed scientists to efficiently survey cell cultures at huge scales, then create large databases of proteins found in complexes, including the large repositories BioGRID, IntAct, STRING, HPRD, TRED, and RegNetwork. This method sadly does not preserve information about regulator/regulatee relationships, much less the sign (up- or down-regulating) of such regulation. Even databases built on co-expression datasets tend to elide this crucial information; *e.g.*, the Gene Regulatory Network database (GRNdb) [186] scores transcription factor $\rightarrow$ target relationships using GENIE3 [187] but convolved with predicted binding domains in the SCENIC pipeline, preserving direction but not sign. Indeed, it is entirely possible that many of the relationships in these databases should not be characterized as "regulatory" at all, but could instead be post-transcriptionally collaborative or antagonistic. As a result, the overwhelming majority of the millions of interactions between tens of thousands of genes, proteins, and miRNAs reported in the twenty databases examined for this study would be appropriate for undirected and unsigned graphs only.

However, a subsection of the RegNetwork database [17] contains both regulatory direction and sign (up/down), consisting of two databases of approximately 4,000 interactions for 1,000

Table 4.1 Statistics of the KEGG-RegNetwork datasets.
p: probability of two random proteins interacting
δ: probability of that interaction being down-regulation.

	mouse	human
proteins	1,033	983
interactions	4,034	3,954
3-motifs	109,195	102,809
- groups (of 132)	26	29
- top group	71.1%	71.1%
- 3-cycles	3	2
4-motifs	3,397,715	3,162,070
- groups (of 22.6k)	190	209
- top group	45.4%	45.2%
- 4-cycles	0	0
- 3-cycles	98	115
p	0.38%	0.41%
δ	2.4%	2.3%

distinct regulatory elements each in *Homo sapiens* and *Mus musculus* (Tab. 4.1). Even this database required a complicated coordination of known interactions with pathways described in the Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathway Map system. [188] Plans to expand this dataset to more proteins and other species, as alluded to in [17] have not yet materialized.[1] Though far from complete, the two kegg.reg.direction tables of the RegNetwork database represent a good-faith attempt to gather everything that has been published about the up/down-regulation of specific genes in known pathways.

Because we know the database is far from complete, it is even more crucial than usual to remember that the results of this paper represent the state of scientific knowledge about protein regulation, not the properties of all regulatory networks that exist in real life. The completeness of even undirected interaction networks has always been a fraught question. A 2009 review of the six largest Protein-Protein Interaction (PPI) databases found that disagreements occurred at all levels. [189] The largest database (HPRD) in the most completely studied species (*H. sapiens*)

[1] Pers. comm. CNKA with both ZP Liu and H Wu

did not have entries for roughly one third of interactions found in the other five databases. These differences occurred not just due to disagreements among what constituted an interaction ("matrix" belonging to the same complex, as opposed to "spoke" confirmed interaction with a single marked protein), but also different annotation methods, and even disagreement among databases about how many interactions were reported by the same paper. Attempts at resolving these differences by meta-databases like AGIL and STRING have been hampered by disagreement about annotation method. Even today, the 11.9 million links reported in the STRING database for humans (v11.5, July 2022) [190] imply an average of 560 interactions per protein-encoding-gene if we accept the maximum of 21,306 reported in CHESS, [191] only 23.6 of which have a high confidence score of $\geq$0.7 support.[2] The Harvard BioPlex3 database, which is built on affinity purification experiments that rigorously remove false positives, finds 118,162 interactions among 14,586 proteins for an average of 8.1 interactions per protein. [16] By contrast, long-term evolutionary experiments followed by genome-wide coexpression assays suggest $\sim$1,000 interactions per protein. [192, 193][3] Indeed, studies of complex heritable traits like schizophrenia and Crohn's disease suggest the "omnigenic" involvement of the entire genome either directly or indirectly. [194]

4.2.3 Graph theory

Much as it is more difficult to make databases of the regulatory properties of protein-protein interactions than databases cataloging their existence, the mathematical/theoretical analysis of directed graphs presents its own challenges above those of undirected graphs. While computational biologists have long exploited graph theory to optimize phylogenetic inference [195] and evaluate ecosystem structure, [196] protein interactions have perhaps the clearest parallels and most rigorous applications to its abstractions. [68, 132, 197–199] The information in a typical PPI database can be

[2]The STRING-db.org help manual acknowledges that the cutoff for a significant level of support is arbitrary, but suggests 0.4 as "medium" and 0.7 as "high" levels. See https://string-db.org/cgi/help

[3]No comparison is available, as the AGIL data server was taken offline in May 2022.

thought of as an undirected graph, with vertices representing proteins and edges representing the existence of their interaction as determined by coIP or a similar experimental procedure. Several different kinds of mutations also have direct analogies to graph theory: a mutation that eliminates the interaction is analogous to an edge deletion ($G\backslash e$ in graph theory symbols), one that causes interactions to be piped through a complex of two proteins (or other gene products) rather than interacting with each protein separately can be thought of as an edge contraction (G/e), and a loss-of-function mutation is like a vertex deletion ($G-v$). All such transformations create a minor H of the original graph G, a concept exhaustively studied in a classic series of twenty papers spanning nearly two decades by Robertson and Seymour that establish analytical tools on solid theoretical ground. [200]

By contrast, we are interested in who is regulating whom, a process better characterized by digraphs, *i.e.*, a directed graph where the edge A→B is distinct from A←B. These abstractions are analytically difficult, and the study of their properties has lagged behind that of undirected graphs. However, in recent years, a great deal of progress has been made in showing that several of the most famous properties of undirected graphs apply to digraphs as well, including ones relevant to mutable bionetworks such as Wagner's conjecture, [201] Hadwiger's conjecture and chromatic number, [202, 203] Mader's problem, [204] and the directed grid theorem. [205, 206] These results suggest that the tools of undirected graph theory can provisionally be used to analyze directed graphs as well, even when not already explicitly extended to them.

4.3 Counting observed motifs

One of the first tasks of this project is to count the representation of each motif in the KEGG-RegNetwork dataset. Note that only two undirected 3-motifs exist (a triangle and a V-shaped 2-star) so these building blocks contain little information, though they have still been studied by

PPI researchers. [207–209] However, with the addition of direction (13) and sign (132) many more topologically distinct 3-motifs exist(*i.e.* 132 isomorphic graph groups). Similarly, only five undirected 4-motifs exist, but there are 199 unsigned directed ones, and 22,662 signed directed ones. [38] Counting motifs on large networks is a difficult and time-intensive problem that has received a great deal of attention, with one recent survey finding 58 different published approaches. [210] In general, these methods either sample the network (via a random-walk, color coding, enumeration-generalization, etc.) to get a representative distribution of motifs, or fully enumerate all motifs either by direct counts that avoid duplication via hash tables (as in the classic MFinder [41] and its descendants like FANMOD/ESU [211]) or in linear-time using counting-motifs. [42] Many recent methods have vastly improved on the efficiency of these early techniques by counting specific motifs via matrix-based (such as ORCA [212] and JESSE [213]) and/or decomposition methods (RAGE [214, 215] and ESCAPE [216]). Other methods focused on applying these theoretical advances to parallel computing, some with fairly minor changes to existing methods (DM-ESU [217] or GPU-ORCA [218]) and others more radically by parallelizing over edges instead of vertices (SubEnum [219] and MR-GTries [220]). As we were interested in all 3- and 4-motifs, we chose not to operationalize the motif-specific counting advances, instead combining aspects of the full enumeration methods MFinder and Itzhack, designed to "mostly" avoid counting the same motifs many times but to run in parallel without the need for work-sharing crosstalk.

After the parallel runs finished, results were gathered from across processes and duplicates eliminated, creating lists of all combinations of proteins that interacted to create 3- and 4-protein subgraphs (belonging to Ω_3 and Ω_4). The direction and sign of all links among these subgraphs was then loaded from the data tables, and each subgraph was assigned to its corresponding motif group (the smaller sets Ω'_3 and Ω'_4, consisting of just one representative from each isomorphic group in Ω_3 and Ω_4). To clarify the difference between these sets, the six blue networks in Fig. 4.3 all belong to Ω_3; but, because they are isomorphs of each other, only one belongs to Ω'_3 as a representative of the

group. The process of assigning interacting proteins to isogroups was somewhat complicated by the context-dependent nature of the up/down regulation in the KEGG dataset; some links were known to be sometimes positive *and* sometimes negative, and so a single set of proteins was sometimes assigned to multiple motif groups. Assignment was performed efficiently using algorithm 1 and equation 3 in Chap. 3.

With this technical background now in hand, we can restate the hypotheses from Section 4.2.1 in quantitative terms. While the abundance of motifs in Ω'_3 and Ω'_4 will be driven by connectivity p, up-regulation bias δ, and the number of isoforms of each motif, (1) departures from the random expectation of Erdös-Renyi graphs with these constraints will correlate with their stability indices.3 We can also predict many features of stable graphs that should be enhanced in a stability-driven system: (2) down-regulation would be favored over potentially destabilizing up-regulation, (3) fragile motifs (with bridges and articulation points) should be rare, and (4) there is no *a priori* reason to expect pass-through elements (which potentially lead to signal amplification) or co-regulation elements (which can make sure elements of complexes are produced concurrently) to be particularly favored one over the other. All of these hypotheses were reasonable, in line with theory and findings from other data sets, and completely wrong.

4.3.1 Observed trends

In both mice and humans, the overwhelming number of links are positive, and the connecting networks are sparse; as a result, the counts across motif space Ω' are very strongly biased towards just a few of the positive and simple motifs (Tab. 4.1). In humans, 96.3% of observed 3-motifs belonged to just three groups, all of them with the minimum two links required for connectivity, and both of those links positive (Fig. 4.1). Turning our attention to the 4-motifs, we find that fewer than 1% of the possible isogroups occurred even once in the data (209 of 22,660), and 45% of the 3.2M observed motifs belonged to just one isogroup, the "A co-promotes $\{B,C,D\}$" group. Only

eight 4-motifs in Ω'_4 have three positive edges and no negative edges (0.035%), but they comprise 88.5% of all observed 4-motifs in humans, including seven of the eight most abundant. Results are qualitatively similar for mice, with a strong bias towards simple, positive networks, but with even fewer motif groups represented in the marginally larger data set: 26 rather than 29 of the 132 possible 3-motifs, and 190 rather than 209 of the 22,660 possible 4-motifs. (See Appendix: Figs C.2, C.3)

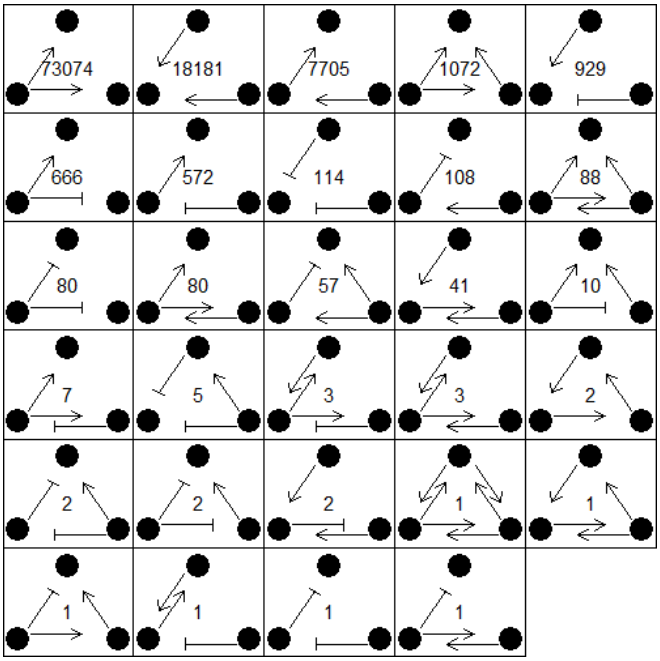

Figure 4.1 Counts of all observed 3-motifs in humans

4.4 Expected Frequencies

The total for each motif was then compared to expectations, calculated several different ways. The expected frequency of directed and signed motifs is a poorly characterized problem. Even in the relatively simpler case of undirected motifs, there is little agreement about which assembly rule best characterizes biological networks, and different rules produce different motif distributions.

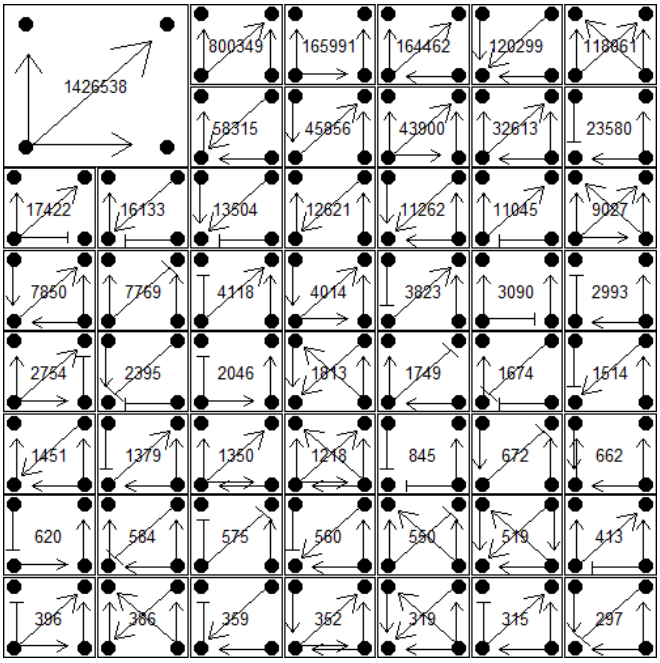

Figure 4.2 Counts of the most common 4-motifs in humans. Only the 53 shown here made up at least ∼0.01% of the total count of 3.16M observed motifs.

[39, 221, 222] We extended the method recently proposed by Fushimi *et al.* [43] as one of the most efficient (considerably faster than either Possible Graph Sampling or LINking and Counting, both described in [207]) and one of the few designed for digraphs. As in their case, we let the graph structure be $G = (\mathcal{V}, \varepsilon)$, noting that self-loops are assumed for all proteins $v \in \mathcal{V}$. Note also that $|\mathcal{V}| = |\varepsilon|(|\varepsilon| - 1)$. The connection probability of any edge $e \in \varepsilon$ is derived from the dataset as $p = \frac{L-d}{N(N-1)}$ where L is the observed number of interactions, d is the number of duplicate entries that occurred earlier in the data table ($d = 4$ for both data sets, when a given interaction could be both positive and negative), and $N = |\mathcal{V}|$ is the number of unique regulatory elements represented (see Tab. 4.1). We then define two indicator functions: $x_+(e) = 1$ if e is present and up-regulating, $x_-(e) = 1$ if e is present and down-regulating, and both are 0 otherwise. We further define δ as the proportion of down-regulating interactions in the data set and assume that up- and down-regulation are the only types of interactions, so the proportion of up-regulation is $1 - \delta$. Note that our edge system is more complicated than Fushimi's since each edge has three possible states, and so the

space of all graphs $\mathbf{x} = \{\ldots, x(e), \ldots\} \in \Omega$ has cardinality $|\Omega| = 3^{|\varepsilon|}$, though many of these will be forbidden because they generate a disconnected graph; however, it is also simplified by assuming the probability of each edge is equal, $p(e) = p \ \ \forall e \in \varepsilon$. With these definitions, the probability of any motif can be calculated from its edge-list $\mathbf{x}$ as:

$$m(\mathbf{x}) = \prod_{e \in \varepsilon} p^{x_+(e)+x_-(e)} \delta^{x_-(e)} (1-p)^{1-x_-(e)-x_+(e)} \tag{4.1}$$

$$= \prod_{e \in \varepsilon_x} p\delta^{x_+(e)} (1-\delta)^{x_-(e)} \prod_{e \in \varepsilon \setminus \varepsilon_x} (1-p) \tag{4.2}$$

Thus far we have followed Fushimi's notation, but here note this system can be simplified even further if we let

$$\pi(e) = \begin{cases} p\delta, & e \text{ is down-regulating} \\ 1-p, & e \text{ is absent} \\ p(1-\delta), & e \text{ is up-regulating} \end{cases}$$

We can then substitute these values to find

$$m(\mathbf{x}) = \prod_{e \in \varepsilon} \pi(e) \tag{4.3}$$

This equation represents the probability of any possible motif in Ω. We then optimize by introducing a mapping function to take the index of any possible motif G_i to the index of its isomorphic group G'_i, via $f(i) \to i'$, let $T_i = |G'_i|$ be the number of members of isomorphic group G'_i, and denote the space of distinct isomorphic groups as Ω'. Because $|\Omega'| << |\Omega|$, far fewer expectations, stability metrics, and motif statistics need be derived than otherwise. The probability of all motifs in the group is equal, since each $G_i \in G'_i$ has the same number of positive, negative, and absent links), so the final probability of each group can be calculated as

$$m'_{i'}(\mathbf{x}) = C \sum_{i=1}^{|\Omega'|} T_i m_i(\mathbf{x}) \tag{4.4}$$

where C is the correction factor for those $G' \notin \Omega'$ because they are disconnected graphs. It can be found by temporarily letting $C = 1$, finding all the incorrect m'_i values as μ_i, and then solving

$C^{-1} = \sum \mu_i'$. On a standard laptop, the probability of the 22,660 4-motifs in Ω_4' could be calculated in 0.1 seconds using this method.

4.4.1 Comparison to observations

Despite the dense notation, this null model is remarkably simple in the sense that it makes no assumptions about network assembly rules, selective pressures, or stability. Instead, it assumes that any PPI is as likely to exist, and as likely to be up or down, as any other. However, these assumptions do not imply that any *motif* is as likely as any other. Because only $\sim$0.4% of the possible interactions are realized in the two datasets, and $< 3\%$ of the interactions are down-regulating, the expectations are strongly skewed towards motifs with few and positive links. Thus far, the null model is consonant with the results.

However, the introduction of T_i biases the results in the wrong direction. Most 3-motifs have six isoforms, but some have three, two, or even one (Fig. 4.3). Note that in Fig. 4.1 and 4.2, motifs that co-regulate (A regulates B, C, and D if present) are the most common, while pass-through motifs (A regulates B regulates C ...) are nearly an order of magnitude less frequent. Ignoring T_i, the two positive links and four absent links in the 3-motif should make these motifs equally likely. However, our null model predicts the opposite: there can never be more than N ways to create a simple co-regulatory model, but there are potentially $N!$ pass-through models. (Compare the green and blue isoforms in Fig. 4.3.) This means we expect twice as many pass-throughs as co-regulators in the 3-motif (6 to 3), and a six-fold bias among the 4-motifs (24 pass-throughs to 4 coregulators), yet we observe a 4:1 bias in the other direction.

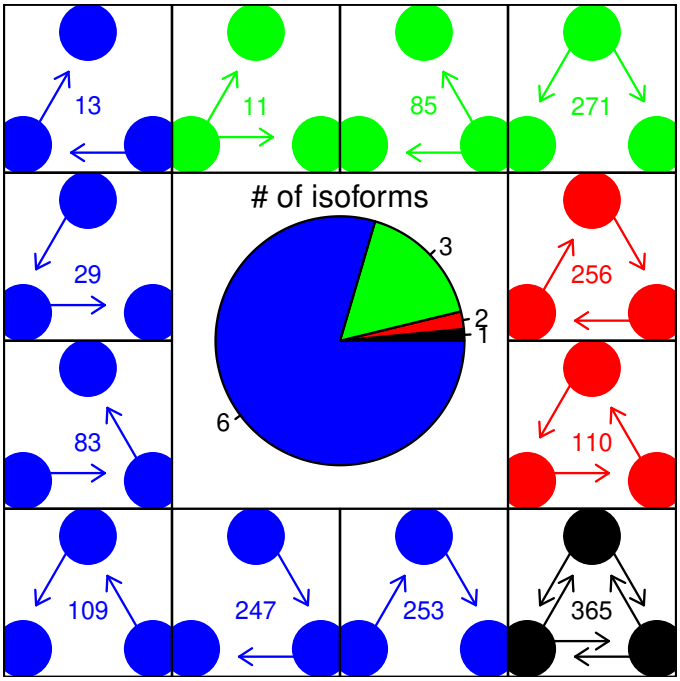

Figure 4.3 Most 3-motifs have six isoforms (105, blue), but 22 have three (green), 3 have two (red), and the remaining two have just one (black, the positive and negative 3-cliques). Representatives of each of these four classes, with all of their isoforms and corresponding motif IDs, are shown around the pie chart. Note that all edges are up-regulating in this figure, but the results hold for any combination of up- and down-regulation.

4.5 Topological correlates

The number of co-regulating elements can be efficiently calculated using an adjacency matrix. For example, if we define the edges of the motif as 1 if present and 0 as absent and enter them in an adjacency matrix A, then the number of co-regulating elements is

$$\sum_i \binom{\sum_j a_{i,j}}{2}$$

(The number of co-regulat*ed* elements can be calculated by reversing the order of the sums.) The number of pass-through elements can also be calculated as $\text{grandsum}(A^2) - tr(A^2)$, in other words, the sum of all the elements of A^2 (because the elements of A^2, α_{ij}, represent the number of 2-step paths from i to j) except the diagonal (which represents forbidden out-and-back loops). This is a nicely compact formulation, but not particularly efficient. A faster method is to fill the adjacency matrix with the *index* of each edge, then take the Cartesian product $B = a_{ij} \times a_{jk}$ where $i \neq j \neq k$. This produces a (2xN!) matrix of indices, and the number of pass-through elements is the collection of rows where both columns are non-zero: $\sum_i y_{b_{i,1}} \wedge y_{b_{i,2}}$. These formulae reduce computation time of these properties across all 22k members of Ω'_4 to $\sim$1.5 seconds on a standard laptop.

For both the 3- and 4-motif set, there was a strong nonlinear relationship between residuals from the null model and the number of co-regulatory and pass-through elements (Fig. 4.4). In all cases, it was their absence that had substantial explanatory power: motifs with no coregulating elements were approximately 3.5x and 3.2x less common than predicted by the null model in Ω'_3 and Ω'_4 respectively, while motifs lacking pass-through elements were $\sim$7.5 and $\sim$80 times more common. These correlates decreased the overall error, but failed to explain many of the largest outliers, including the reinforcing co-regulating motif shown in the inset, many of which consisted of exclusively positive edges. “All-positive” motifs were moderately more common than predicted by the null model, but gains in predictability from including “all-positive” as a covariable were modest (Tab. 4.2).

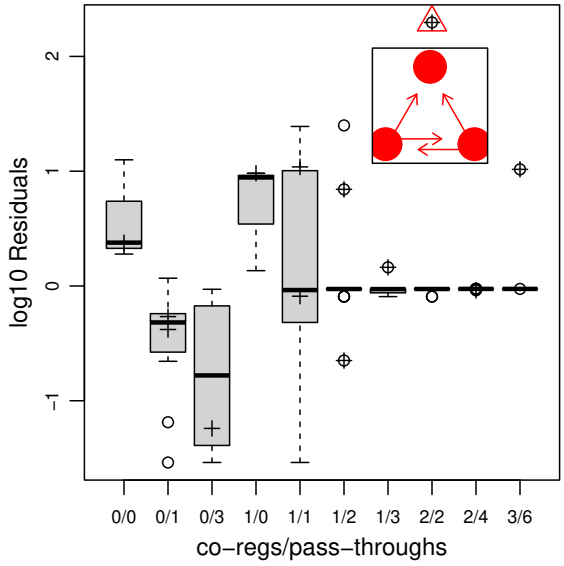

Figure 4.4 A null model captures much of the variability in known bio-networks. Motifs with no co-regulatory element (A regulates both B and C) occur ∼3.5x less frequently than the null model predicts, while motifs with no pass-through elements (A regulates B regulates C) occur ∼7.5x more frequently. Many of the largest deviations occur on motifs consisting solely of up-regulating edges (marked with +), which occur on average 3x more frequently than expected, after taking co-regulation and pass-through elements into account.

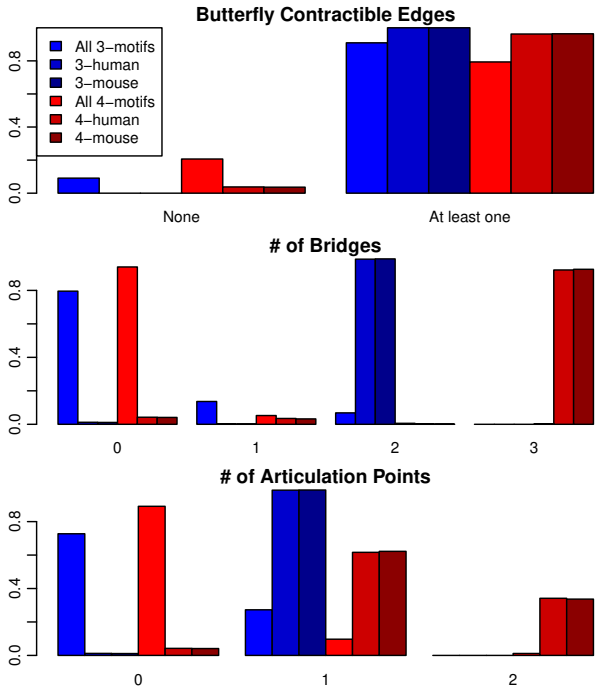

Figure 4.5 The total number of butterfly contractible edges, bridges, and articulation points in all connected 3- and 4-node motifs was an exceptionally poor predictor of how many were observed in the KEGG-RegNetwork datasets. The deviations from expectation were not only large, but all in directions that increased energetic costs or decreased stability, contrary to our hypothesis.

There are several other notable properties of the observed motif distribution. First, real-world protein networks are expected to be difficult to cut, as this would lead to disregulation. Bridges are edges that cause a graph to become disconnected when removed, that is $G \in \Omega$ but $G \backslash e \notin \Omega$. Articulation points are the vertex equivalent, $G - v \notin \Omega$. Across the theoretical Ω' spaces, bridges and articulation points are relatively rare; 80% of 3-motifs and 94% of 4-motifs have no bridges, while 73% and 89% lack articulation points. Contrary to expectations, these "uncuttable" motifs are heavily depleted in the datasets with only about 1.2% and 4.2% of observed 3- and 4-motifs lacking bridges or articulation points in both humans and mice (Fig. 4.5bc).

Butterfly minors are a special class of reduction that exists only on digraphs. They are created through one of three moves:

- The deletion of an edge,
- The contraction of edge e from (a,b) into a if e is a's only outgoing edge, or
- Into b if e is the only incoming edge of b.

The latter two procedures are analogous to the elimination of a redundancy in the network; if many proteins regulate A but A regulates only B, then they may as well regulate B directly. Similarly, if B regulates only one protein, then all its regulators could regulate A just as simply. The energetic costs of producing such an unnecessary "middle step" into a protein regulation pathway would likely be evolutionarily unfavorable in the absence of the need for an integrator or other secondary process. An edge is said to be butterfly contractible if it meets either of the latter two conditions above, a concept that has proven useful to theorists who have used it to extend theories of tree width and the Erdös-Posá property to directed graphs, and to prove Norine's Conjecture. [223–226]

Unlike bridges and articulation points, butterfly contractible edges are abundant; 90.9% and 79.4% of 3- and 4-motifs in Ω' have at least one such edge, respectively. Far from eliminating such edges, however, the data set shows that they are all but required: exactly one observed 3-motif (out

of >100,000) in each species had no butterfly contractible edges, while over 96% of 4-motifs in both species had at least one (Fig. 4.5a).

The networks also showed far more clustering than expected in the random null model. This effect is somewhat masked by the very low connectivity of networks in the data, which meant motifs with large numbers of edges were expected to be even more rare than observed. Motifs with the minimum number of edges (2 for 3-motifs, 3 for 4-motifs) occurred very slightly less often than expected, but the more edges, the more enrichment was observed (Fig C.6). 3-Motifs with four, five, and six edges were only expected to occur twice, but instead appeared 96 and 74 times in human and mouse respectively; 4-motifs of 6+ edges were only expected $\sim$1.5 times in the data, but instead were counted 1,502 and 1,075 times respectively. This over-representation of high edge count motifs demonstrates a good deal of cross-talk within functionally linked genes.

Clustering was also observed in the co-occurrence of regulatory sign. We have already discussed how motifs made up of only positive (i.e. up-regulating) edges were enriched across data sets (Fig. 4.4, Tab. 4.2), but negative edges clustered as well. For example, 4-motifs with four negative edges were expected only $\sim$0.013 times in either species, but were instead observed 36 times in humans and 39 times in mice. Overall, motifs with two or more negative edges were seen $\sim$3.6 times more often than random in 3-motifs in both species, but a more modest 52% and 23% enrichment in 4-motifs in humans and mice; however 4-motifs that were all negative were 9.8- and 8.6-fold enriched.

The number of k-cycles in a directed graph is an NP-hard problem, but can be calculated fairly quickly on small motifs using a modified color-coding method, [33] and sped even faster for k-cycles on k-motifs by recognizing that any cycle includes every vertex, so only one starting point need be tried. Similarly, on 4-motifs, every 3-cycle must include either vertex 1 or 2 (or both), so only these two starting points need be tried, the second further sped by removing all incoming edges to vertex 1. Across Ω_3, 43% of motifs have one cycle, and 9.1% have two; the null model predicts

81 one-cycle and 3 two-cycle motifs should be observed in both species, but instead only 3 and 1 appear in humans, and 2 and 1 in mice. Similarly, 33% of Ω_4 motifs have zero and one 4-cycles, 20% have two, and on down to 0.8% with six; while these are predicted to be relatively rare, the null model still expects ∼3,597 motifs with at least one 4-cycle, yet none are observed out of the over 3,000,000 motifs counted in each data-set. Three cycles are a bit more widely distributed, rising from 11.8% of Ω_4 motifs with zero to a peak of 25.0% with two, and tailing off to 0.08% with eight. The null model predicts ∼5,820 motifs with one 3-cycle, 1,146 with two, and 1,101 with three or more. However, the observed distributions are 79, 36, and 0 for humans, and 68, 30, and 0 for mice. We discuss these astonishingly low numbers in the conclusion.

The out-degree distribution of both species has ∼700 nodes with zero outgoing edges, but a very long tail ending in two "hubs" or "global regulators" with over 100 targets each (NF-κB-1 and RELA=NF-κB-3 in both). As found in [227], the distribution follows a power law until limited by sample size. The in-degree distribution also follows prior research, having a shorter tail and a log-linear rank-count curve that drops below one protein per in-degree at ∼30 regulators, though three proteins in each species have an in-degree > 45. [34, 228] The truncated distribution found in *E. coli* and other prokaryotes was thought to result from their relatively short promoter region, and therefore could be longer in eukaryotic organisms with their more baroque regulatory methods; this appears to not be the case. Indeed, approximately 20% of all targets in both datasets were regulated exclusively by one transcription factor (160/776 in humans, 156/806 in mice), and over 45% of these mapped back to just four transcription factors (42/43 to TP53/Trp53, 15/15 to NR3C2, 13/7 to IRF3, and 6/8 to NR1H4 in humans/mice). This confirms the prevalence of Single Input Modules (SIMs) as seen in other organisms. [34, 177]

4.6 Stability

Stability was determined by creating a nonlinear ODE for each motif in Ω'_3 and Ω'_4, drawing 1,000 sets of random parameter values for each, solving the system for fixed points, and finally calculating the eigenvalues at each fixed point to determine its stability properties (see details in 3). The ODE is of the form:

$$[\dot{P}_i] = \alpha - \delta[P_i] + \sum_{j \neq i} k_{ij}[P_j][P_i]^h \tag{4.5}$$

where $[P_i]$ is the concentration of protein i, α is the (universal) autocatalysis rate at which all proteins are made independent of regulation, δ is the degradation rate (n.b.: density dependent), k_{ij} is the effect size of protein j on protein i, and h is the hill coefficient which determines the shape of the response of protein i to protein j. A network was considered "unstable" for a given parameter vector if it had no stable fixed points, implying the suicidal consumption of all cellular resources as a protein's concentration rushes toward infinity. This calculation process took over three years of CPU time, and resulted in a thoroughly sampled parameter space, revealing numerous distinct behaviors and markedly different stability profiles for all possible motifs. Two intuitive measures of stability were

1. runaway: the fraction of parameter space with no fixed points; and hence any starting concentration of proteins would be driven to zero or infinity.

2. SI: A Stability Index, modified from the SSS of [221], where each draw of random parameters received 0 points if there were no fixed points, and between 0 and 1 points based on the fraction of fixed points that were stable, averaged over all 1,000 draws. For example, a motif where 50% of simulations had no fixed points, 25% had 1 stable fixed point, and 25% had 1 stable and 1 unstable fixed point would have a stability index of $.5 \times 0 + .25 \times 1 + .25 \times .5 = 0.375$.

Stability was also calculated using 12 other statistics, but they all correlated with these two and were not included in subsequent tests.

Table 4.2 Models predicting observed motif abundance.
pt0: 0 pass-through elements; crg0: 0 co-regulating elements; ap: all edges positive; (models containing these three parameters are considered "full"); SI: stability index; runaway: fraction of parameter space with no fixed points

Model	r^2		MAE		AIC	
$n-$motif	3	4	3	4	3	4
Null						
human	0.836	0.6468	512	108.8	193	-16310
mouse	0.828	0.608	546.5	117.6	198.5	-13560
+pt0+crg0						
human	0.863	0.6495	413.8	108.8	173.3	-16480
mouse	0.8626	0.6115	451.3	117.6	172.8	-13760
+pt0+crg0+ap						
human	0.8731	0.6509	370	108.6	165.2	-16560
mouse	0.8661	0.612	420	117.5	171.4	-13790
Null+SI						
human	0.8361	0.6469	512.5	108.8	194.9	-16310
mouse	0.8284	0.6082	545.8	117.6	200.2	-13570
full+SI						
human	0.8633	0.6496	414.5	108.8	175	-16480
mouse	0.8631	0.6117	448.7	117.6	174.3	-13770
Null+runaway						
human	0.8364	0.6469	512.7	108.8	194.7	-16310
mouse	0.8281	0.6081	546.2	117.6	200.4	-13560
full+runaway						
human	0.8634	0.6496	415.8	108.8	174.8	-16480
mouse	0.8628	0.6116	448.8	117.6	174.6	-13760

These two measurements were used to correct both the null predictions above, and the "full" model including co-regulation, pass-through, and all-positive information. These four models were used to predict the four distributions (human and mouse x 3- and 4-motif) and evaluated using three metrics for 48 total tests; in no case did the addition of stability information improve model performance (Tab. 4.2). Of course, other metrics of stability are possible, and there are reasons to think our data are non-representative of networks as a whole. Nevertheless, the total absence of any correlation at all cannot be ignored.

4.7 Similarities to other networks

The surprising results for stability are all the more striking since the network is conventional in many other ways. Like previously studied networks, it is sparse (linkage $< 1\%$), has a long-tailed out-degree distribution and a more compact in-degree (even though regulation in the eukaryotes studied here is less sterically constrained by short promoter regions than the prokaryotic networks where these trends were discovered), is dominated by a relatively small number of motifs, and these motifs show the classic Single Input Module (SIM) pattern. [228] While these results must be approached with caution given the incomplete nature of the data, there are reasons to believe SIMs would be common in an evolving network: a master regulator is useful if multiple proteins need to respond to a particular environmental condition, such as metabolizing enzymes in reponse to a nutrient pulse or damage repair proteins in response to heat-shock or DNA damage. [229] They would also be useful in creating a multi-protein complex, as it would ensure the transcription of all components of the complex is started and stopped at once. [230] Less obviously, SIMs generate temporal expression programs by regulating different targets at different concentrations and/or binding affinities, so one protein's transcription is activated after another as regulator concentrations rise in LIFO order. [177, 231, 232] Fairly minor adjustments in such a network can result in a 100x change in regulatory speed, [233] and synthesized networks using this motif are able to create a "binary ripple" to count cellular events in base 2. [234] Given all these benefits, it is perhaps unsurprising that by far the most abundant 3- and 4-motifs were small-scale SIMs (Figs 4.1,4.2). By contrast, pass-through or "cascade" motifs, though often discussed as having important roles (*e.g.*, amplifying extracellular signals) also tend to amplify noise in the system, [38] and have proven vulnerable to collapse in ecosystems if not stabilized by numerous weak interactions. [235] Perhaps this increased variability (as opposed to instability *per se*) goes partway to explaining why coregulatory elements were more common than cascade elements (top left of Figs 4.1, 4.2, C.2,

C.4), despite the opposite being in line with null expectations due to the larger number of cascade isoforms.

Another similarity to other regulatory networks was the strikingly low number of cycles. Feedback loops are often conspicuously absent in protein networks, a phenomenon researchers have imputed to the instability or multistability of loops. [221, 229] Some argue that the core "plant" of a gene network is always stable, while the regulatory "controller" feedback structure creates any observed instability. [236] However, the instability of feedback motifs has been greatly overstated. Positive feedback loops indeed lead to instability across most of parameter space; 3-motif #365 where all three nodes positively regulate the other two runs to infinity 71% of the time. However, consider the prototypical negative feedback loop: the A-suppresses-B-suppresses-C-suppresses-A "represselator" or "rock-paper-scissors game" (3-motif #219), widely used in both theory and experiment for its elegant Hopf bifurcation. While the shift from a fixed equilibrium to a limit cycle is indeed a form of instability, it is not one that runs to infinity; indeed, our analysis suggests that the represselator runs to infinity in only 2.2% of parameter space, the lowest of all 3-motifs in Ω'_3.3 The inclusion of negative feedback loops is intuitively one of the best ways to add stability to a system through self-regulation, which is why they are so common in human-designed electronic circuits. [41] We must seek an explanation besides their supposed instability to account for their near-total absence in protein networks.

Two other notable motifs are diamonds, and feed-forward loops (FFLs). Diamonds (Fig 4.6), where $A \rightarrow \{B,C\} \rightarrow D$, are highly enriched in foodwebs and neural wiring networks, but not generally in protein networks, [41] and not in this data set either. While the all-positive version of this pattern occurs a modest 66% more than expected (a typical bonus for positive links clustering together), the 10 possible diamond patterns are on average about 33% less abundant than predicted by the null model. Because negative links are so rare, 7/10 such motifs are expected < 2 times in the data and indeed six of those did not occur. The exception is Diamond H, which is enriched

almost 50-fold, perhaps because it performs the unique function of turning off an OR logic gate. That is, a protein co-regulated by two transcription factors, either one of which is sufficient to start transcription, must have both transcription factors turned off together to be turned off itself, and Diamond H is the only 4-motif capable of performing this task. Similarly, Diamonds B and C are both incoherent subgraphs, where one limb activates a gene while another limb shuts it off, and both are expected to occur $\sim$50 times. But B requires the energetically costly production of a repressor protein to accomplish this while C does not, instead down-regulating a promoter. This may explain why B does not appear in the human data while C is $\sim$2-fold enriched in both species. According to Savageau's "demand rules", [237] this implies that the ultimate target of the diamond is generally needed in the cell, so it is energetically favorable to switch off an activator rather than switch on a repressor. Similar logic governs FFL loops, which–unlike diamonds–are typically overexpressed in protein networks. [3, 34, 238] In our data, FFLs were enriched $\sim$3-fold on average, the enrichment coming from both coherent and incoherent structures (Fig C.5). Enrichment of both classes is expected as both have distinct roles in cells: coherent FFLs introduce delay elements into transcription, [35], while incoherent FFLs accelerate reaction speeds. [36] This ability to respond in time-appropriate ways to the environment implies that FFLs are always superior to cascades in fluctuating environments, [168] though Pareto evolution is able to find other motifs that are even better. [239]

We note in passing that bifans ($\{A,B\} \rightarrow \{C,D\}$, an abundant motif in neuronal and artificial sensory networks, [34] and whose mathematical properties have been studied elegantly in [67, 240]) exist in seven distinct isoforms; the all-positive bifan (4-motif #7381) was observed over 100,000 times in both species (top-right corner of Fig 4.2), far in excess of the $\sim$545 expected occurrences; the all-negative bifan (#14761) was $\sim$125-fold enriched, and the total bifan count was 200x higher than expected (242,523 vs 1,199.4 across both species). This overexpression suggests that generalized bifans, or dense overlapping regulons (DORs), are at least as common in eukaryotes as

they are in prokaryotes. [241] In none of the three case-studies above (diamonds, FFLs, nor bifans) do the departures from expectations correlate with stability metrics.

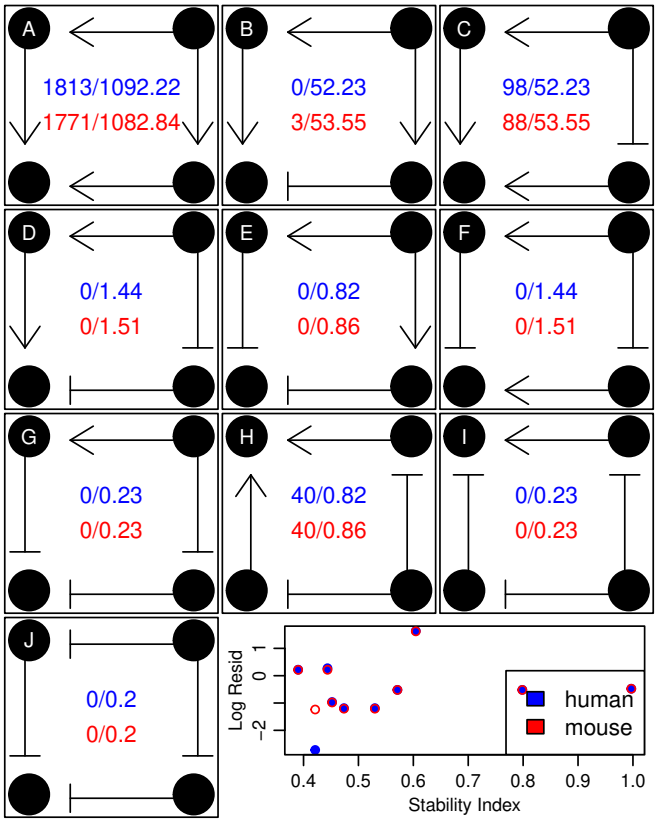

Figure 4.6 On average, diamonds motifs occur slightly less than expected (numbers represent observed/expected, blue=human, red=mouse). The exceptions to this trend are interesting. Diamonds B and C are expected to occur equally and both perform the function of a delayed AND logic gate via incoherent feed-forward. However C is about twice as common as expected while B barely occurs at all, because B requires the production of inhibitor while C does not, and so C would be energetically favorable. Similarly, H has the highest Z-score, perhaps because it is the only motif that can shut down a logical OR gate. As the graph at the bottom shows, stability provides a poor explanation for residual enrichment relative to null expectations.

It is unclear how much of the clustering observed in the data represents actual necessary cross-talk among elements of complex regulatory circuits, and how much is a bias of a few pathways being studied intensely by scientists, and thus being over-represented in the database. Perhaps even more plausibly, the co-occurrence of negative and positive links may be a bias of methods that

are better at detecting positive and negative regulation; only pathways investigated using both a positive-results- and negative-results-biased method would have an accurate mix of positive and negative links represented.

4.8 Conclusion

However, bias in scientific interest or methods cannot explain the instability in motifs discovered across the database. The average amount of biological parameter space where observed motifs would runaway to infinity is 36.5% for 3-motifs and 42.9% for 4-motifs, which is actually higher than the 33.6% and 40.2% expected from sampling evenly across $\Omega_{3,4}$. A bias toward instability clearly can't be the whole story, or life would not continue. Note that this "inescapable tendency towards infinity" definition differs from other uses of the term "instability" in biological literature, where it often implies a transiently untennable network topology that collapses to a stable state by the elimination of a node, by gene silencing or local extinction of a species in a food web. Note that using the equations

One implication of this instability is that there must be many down-regulating processes not included in the current body of knowledge to keep the system from regulatory catastrophe. Indeed, many such processes exist which we did not model; [242] even within the RegNetwork database there are large numbers of miRNAs and siRNAs which are assumed to have a nearly-exclusively down-regulating effect on cellular processes, [17] and affect as much as 30% of the genome. [243] They were not included in this study because it is not well understood who regulates these regulators, though it is known that somebody does: studies of vertebrate embryos indicate that approximately half of such iRNA regulators are significantly regulated themselves during development. [244] Including them in this study would therefore have introduced a strong bias into our motif counts by inserting huge numbers of nodes with a (probably false) in-degree of 0. Furthermore, many proteins

interact with each other not through transcription factor regulation, which is a relatively slow process, but directly in the cytoplasm by binding to activate, deactivate, or flagging for destruction via methylation, [245] acetylation, [246] phosphorylation, [247, 248] or ubquitination, [249] a process whose very name implies how wide spread it is; and these interactions can happen far faster than a protein can be manufactured, enter the nucleus, then find and bind to a promoter. Perhaps the implication is that activation happens slowly via transcription factors, while degradation happens quickly via the mechanisms listed above. Finally, scientific methods are better at measuring large increases from baseline than decreases, since most proteins are at low concentrations most of the time. For example, of the 289 proteins known in blood plasma, hemoglobin, albumin, complement factor H, and the immunoglobulins make up 99% of the mass. [250] With baselines so low for so many proteins, even strong negative regulation may not be detectable.

Some alternatives to the "missing down-regulating links" hypothesis can be discarded because of the way networks were simulated. Note that the effects of entropy and general housekeeping degradation processes is assumed to be rolled into the $\delta[P_i]$ term of eq. 4.5. Any 1-motif would reach equilibrium $[\dot{P}] = 0$ at α/δ, and can only escape into instability if the regulatory effects k_{ij} are large enough to overcome δ. Similarly, temporal dampening–a situation where genes simply turn themselves off eventually when their promoters wear away–was covered in simulations where $\alpha = 0$, yet instabilities existed there as well. Alternatively, perhaps could we indeed have discovered instability in the “transient topology collapsing to a more stable state” sense? The difficulty with this hypothesis is that transient states are, by definition, short-lived, and therefore very unlikely to be detected in large-scale PPI survey experiments. While perhaps a few such interactions could be in our database, they couldn’t possibly make up $\sim$40% of all interactions and still be considered “transient”.

In addition to missing down-regulating links, some degree of instability is desirable in biosystems. [193] As we pointed out in 3, oscillatory and periodic processes such as sleeping, breathing,

and the beating of the heart are necessary for life, [136] though chaotic and disordered dynamics appear to be actively suppressed even in networks where they would be expected through stabilizing PPI links. [137] Nevertheless, homeostasis implies that most things are mostly stable most of the time, as codified in the mathematically rigorous "steady state assumption" ubiquitous in biochemical research. [138, 139] It clashes with our view of biology to claim that $\tilde{4}0\%$ of regulation truly has no equilibrium at all, as implied here.

An additional explanation is that while some motifs are a good deal more stable than others, no motif is universally unstable. As in the example of *Drosophila* morphogenesis, a network where only 3% of its parameter space was stable could still be used for critical biological processes by maintaining system parameters in that small parameter space region. [165] Similar to our findings of stability's weak role in motif counts, studies of induction strength in metabolic regulatory networks find that, although a universally optimal structure and parameter combination exists, not all organisms adopt it. [251] This is difficult to explain on theoretical grounds, since mathematical studies demonstrate that perfect adaptation consists of a network with a large initial response to a change in input followed by a return to baseline afterwards, and that these two goals do not involve trade-offs so a path to adaptation always exists and should be quite rapid. [252]

However, assuming that the data represent at least a reasonable sampling of regulatory pathways, then there must be a better explanation for the lack of stable structures than a huge number of missing stabilizing negative links, globally fine-tuned parameters that keep unstable networks carefully contained in small stable regions, or instability of a special kind that is beneficial. As mentioned in the introduction, regulatory networks represent a balance of the demand for stability with the need for function and adaptability, [181] and there are reasons to think the latter two may overwhelm the former. First, Waddington canalization appears to be ubiquitous, despite being costly. Genes identified by Ahab software as part of the *Drosophila* segmentation pathway mutate about as often as any other locus, but their functional role is conserved between species. [253] This suggests

that function is prioritized over stability on long time scales. Similarly, organisms placed under artificially extreme selective pressures adapt quickly to the new environment: fish adjust their size and even the number of their vertebrae in both the laboratory and altered ecosystems, [254–256] *E. coli* rapidly evolves to optimal metabolic activity and chemotaxic behavior, [167] and fruit flies respond to early harvesting of their eggs by increasing life-span many-fold within a few years. [257] However, genome-wide screens of these processes suggest that one allele does not sweep to fixation in the new conditions, but diversity is maintained everywhere in the genome, [192] which strongly implies that flexibility and adaptability are being maintained at all costs in the face of artificially powerful selective pressure. Since it appears true that networks can rewire quickly in response to changing conditions, [176] perhaps life is able to escape instability by quickly switching to a different topology, or globally downregulating to prevent runaways, all in the name of preserving function and diversity. Warmflash *et al* [239] argue that dynamic properties of the network should be abstracted away from the network structure itself as the phenotype upon which selection acts, and only then can simulations mimic nature. Functional modules, not precise pathway structures, may be the locus of selective pressure. [258] The multifunctionality of bifans is already well documented, [240] and, according to our results in 3, all motifs can generate multiple distinct behaviors at biological parameter values. This implies that if the cell needs a given function, a large number of motifs can provide it. Perhaps natural selection chose from this set of motifs the most simple and favored positive regulation of proteins, so they are synthesized only when needed in order to conserve energy–with the exception of a few constitutively expressed proteins. In this way, the network motifs themselves may be of only secondary importance to their ability to maintain output in the face of an uncertain environment.

This favoring of adaptation and flexibility over stability gains some support by analogy with food webs. Mathematical simulations of ecosystems that employ adaptive foraging show an increase in stability and robustness to deletion, while preserving trophic levels and guild functionality

(analogous to hierarchical regulation and metabolic pathways) relative to those that do not. This effect overwhelms that of the initial topology, and the functional form of the links between nodes (Lotka-Volterra, Holling, or Ivlev). [259, 260] This can function even on networks, like ours, where complexity is and remains relatively low. [261]

One of the primary founders of motif research, Uri Alon, ended his book about the issues discussed here saying "one can only write an introduction, since we are only at the beginning of the adventure to find the design principles of biological systems." [38] The fact that fourteen years later so many fascinating possibilities remain open to explain evolution's apparent indifference to stability indicates that the adventure is still just beginning. As more data become available and the trade-offs between function, flexibility, and stability come into sharper focus, perhaps the role of structurally stable networks will also resolve.

Chapter 5

Calorie-restriction leads to longevity by stabilizing protein networks

5.1 Abstract

Calorie restriction has long been known to be a robust way of extending lifespan, and slowing protein turnover rates. Classical bifurcation theory also suggests that slowing turnover rates may be a way to stabilize complex systems. In this paper, we use large-scale proteome kinetics to quantify the effects of a CR diet on turnover in the house mouse *Mus musculus*, and place these (generally but not universally) slower rates in the context of all known regulatory pathways. The results indicate that networks are indeed stabilized by a CR regime. Furthermore, regulatory networks appear to be strongly shaped not by the total volume of stability a topology can support, but by the well-contained boundaries of the stable region; network features where stable and unstable parameter regions are intermixed are rare, suggesting they are targets of negative pressure by natural selection.

5.2 Turnover Rate and Calorie Restriction

Santa Fe Institute founder George Cowen has pointed out that life maintains itself at the boundary of order and disorder, where it “has enough stability to sustain itself and enough creativity to deserve the name of life," and understanding this boundary is a central goal of “the sciences of the twenty-first century.” [262] Unstable bionetworks have short lifespans for the same reasons as unstable economies and unstable airplanes; yet must balance the need for stability against the need for adaptability. It has been argued that aging itself exists as a method to introduce diversity and evolutionary options into a population, [263] and the theory of antagonistic pleiotropy argues that any innovation that increases reproductive success early in life must be paid for by decreased fitness later. [264] The balance of such forces can tip a population from favoring early reproduction to favoring long life under realistic conditions, [265] and is likely to maintain protein-protein regulatory networks (PPRNs) at the edge of disorder to adapt to these changing conditions for most species most of the time.

It is reasonable to think that interventions that lead to longer life work through the mechanism of slowing protein turnover, since decreasing the separation between rate scales is a general strategy for stabilizing complex systems. For example, any of the normal-form bifurcations can develop more fixed points and instabilities simply by holding the time-scale constant, but increasing the rate parameter(s). [68] Calorie restriction (CR) is arguably the most robust and reproducible way known to extend lifespan across species, significantly outperforming genetic and pharmaceutical interventions. [45–47] This phenomenon has been studied since the 1930s, [266] and numerous biochemical mechanisms have been proposed to explain it in the ensuing decades. [267] Many of these involve changes to either the action or the production of proteins, such as the “hallmarks of aging”: reduced damage from reactive oxidation species (ROS), deregulated nutrient sensing, telomere attrition, epigenetic alterations, shifts from investment in reproduction to delayed senescence, genomic instability, and loss of proteostasis. [44]

We focus on the last of these (loss of proteostasis), while noting that many of the others could contribute to protein disregulation as causative mechanisms. Specifically, we propose that an increase in protein turnover rate makes it more challenging to maintain concentrations proper for the healthy functioning of a cell. The reason is analogous to someone trying to keep several leaky buckets filled to a line: if water drips from the buckets a drop at a time, it is relatively simple to check every few minutes and refill as needed; however, if water leaks out in steady streams of different flux, it can be energetically demanding to keep up, and any brief distraction can cause some to run fatally dry. That is, the rate scale of outflow no longer matches the rate at which a person can check water levels and refill them. In much the same way, the mechanisms of proteostasis are both complex and energetically demanding; the proteostasis network consists of four basic areas–protein synthesis, protein folding, disaggregation, and degradation–each consisting of dozens of pathways and hundreds of molecular components (ribosomes, for example, contain over 200 proteins and rRNA components). [268] Breakdowns in any of these processes are responsible for over 20 known diseases in humans and animals, ranging from cataracts and cardiomyopathy to Nakajo and Angelman syndromes, to Parkinson's and amyotrophic lateral sclerosis (Lou Gehrig's disease). [48, 269] One of the common assumptions of systems biology generally is that disease consists of some form of protein disregulation, [4] wherein some necessary response to stress is happening at a timescale faster than permitted by the machinery of synthesis by ribosomes or degredation by proteosomes or inactivation by any of the many forms of protein modification and inhibition.

Though there is widespread agreement from both a mathematical and biochemical point-of-view that increasing flux decreases stability, [38, 68, 96] turnover *per se* has rarely been investigated as a disease causing-agent. The reason for this is that flux is difficult to measure, while concentration is fairly straightforward. RNA-Seq methods can reveal the entire transcriptome, while MSMS methods capture entire proteomes in matters of hours. [270, 271] However, the search for "the

aging gene" or "the cancer gene" by searching for proteins that massively change their expression under disease conditions has largely been fruitless, despite the explosion of advanced data science techniques. [272, 273] Certainly, part of the reason is that these conditions are multigenic (if not "omnigenic" [194]) and environmentally triggered; but part of the reason may be that turnover can change greatly without changing concentrations. Returning to our metaphor of the leaky buckets, water can flow out faster and faster while the water level remains at the same height as long as our harried refiller can keep up. This means the system can be at the edge of collapse without any warning from the water level, but plenty of warning from the flow rate. By analogy, a senescent or oncogenic cell may have the same protein levels as a healthy one, but is working much harder to maintain those levels. [274]

Multiple studies have created overwhelming evidence that calorie restriction has the net effect of slowing down protein turnover in mice, though this effect is both tissue- and pathway-specific. [152, 275–279] A multi-omic study that measured thousands of RNA concentrations, protein concentrations, and protein turnover simultaneously (addressing long-standing calls for the need for such integrated studies [280]) provided strong evidence that because RNA concentrations are unaffected by calorie restriction, the regulation of turnover was occurring post-transcription. [48] At least some of this may be due to an increasing backlog of ubiquitin-tagged proteins waiting to be degraded. [281] Others may be due to a sequestration of ribosomal subunits in "stress granules" or differential degradation of subunits to create nonstoichiometric ratios. [48, 127] This is in stark contrast to the general tendency of complexes to exhibit the same turnover rate. [120, 282] Similar trends have been found in basal organisms like *C. elegens*. [153, 283] A cross-species investigation found that whole-body protein turnover rate was a better predictor of longevity than even body mass; for example, the tiny blind mole-rat's slow turnover rate helped it live to age 20 while similarly-sized rats and hamsters have a maximum lifespan of 4 years. [154] This trend is so strong that one researcher, writing in an admirably concise 1974 Nature paper of 200 words, concludes he has

"a sufficient amount of experimental data to" show that protein turnover rates is "the biochemical mechanism [that determines] the life spans of various mammlaian species." [284]

All of the preceding leads to the following argument: (1) CR leads to longevity. (2) CR slows protein turnover. (3) Stable protein networks also lead to longevity. We therefore hypothesize that (4) lower rates of protein turnover are not merely correlated with longer lifespans, [285] but *cause* longer lifespans by stabilizing regulatory bionetworks. CR's life-extending properties are thus a result of CR lowering turnover, which leads to greater regulatory stability. If true, then mice on a CR diet should not only have overall lower rates of protein turnover, but the specific networks these proteins are involved in should become more stable as a result. We test this theory by linking databases of protein regulatory networks with databases of proteome-wide MS/MS turnover rate studies of mice on CR or ad libitum (AL) diets. The hypothesis predicts that lower turnover in CR mice should move regulatory protein-protein interaction networks (PPINs) away from mathematically unstable regions.

Although this particular study was enabled by a great deal of recent technology, the hypothesis underlying it is in fact very old. Historians of medicine write that the understanding that the body is in a constant state of "dynamic permanence", where structures are being constantly broken down and replaced with substances derived from food, goes back to Alcmaeon in the 6th century B.C.E. [12] Hints of a link between diet and longevity can also be found in The Aphorisms of the *Corpus Hippocraticum*, suggesting that fasting (calorie restriction) and a "cold temperment" (low turnover) were hallmarks of longevity, though writers sometimes switched the causal arrow. [13, 286] The Italian scientist Alvise Cornaro, writing at age 83 in the 1560s according to legend, recommended a high-quality calorie restricted diet to extend not only the duration but also the quality of life, a prescription that remained influential into the early 1800s. [13, 287] The French polymath François Magendie revived the idea of turnover on a molecular level, writing in his 1829 textbook "It is extremely probable that all parts of the body of man experience an intestine movement, which has

the double effect of expelling the molecules that can or ought no longer to compose the organs, and replacing them by new molecules. This internal, intimate motion, constitutes nutrition." He even understood that this turnover was tissue-specific: "Nutrition is more or less rapid according to the tissues. The glands, the muscles, skin, etc. change their volume, colour, consistence, with great quickness; the tendons, fibrous membranes, the bones, the cartilages, appear to have a much slower nutrition, for their physical properties change but slowly by the effect of age and disease." [14] Sadly, this understanding was largely eclipsed by more mechanical views of the body until the 1940s. [15, 288] Thanks to the advent of large biodatabases and -omics methods, we are now in a position to investigate if these old intuitions were correct.

In the remainder of this paper, proteins that decrease their turnover under CR will be referred to as "toughened", while those that decrease their turnover rate will be called "embrittled". This terminology reflects a change in their need for repair and replacement, emphasizing that this does not imply a change in their abundance. Section 5.3 discusses how turnover data were acquired, linked to (incomplete) regulatory network databases, and analyzed for stability given the gaps in the network. Section 5.4 presents evidence that real-life networks favor toughened proteins, and are indeed stabilized under a CR regime. Section 5.5 examines the assumption that "stable regions" have well-defined, smooth borders, and finds strong evidence that network topologies that violate this assumption are profoundly underrepresented. This suggests natural selection doesn't try to maximize the inherent stability of a network, but rather favors networks that can easily be maintained in their stable region (be it small or large). Finally, Section 5.6 discusses the prospects for applying the foregoing insights to extend human lifespan, given the difficulty maintaining a CR diet.

5.3 Methods

5.3.1 Turnover Data

Animal Handling and Labelling: All animal procedure protocols were approved by the Institutional Animal Care and Use Committee (IACUC) of Brigham Young University. Housing for mice was provided in a pathogen-free facility for the duration of the experiment. 10-week old C57/BL6 male mice were purchased from Charles River Laboratory, and fed ad-libitum for one week after arrival to acclimate to the facility. Mice were then randomly divided into ad-libitum (AL) and dietary-restricted (DR) dietary cohort groups; the mice were housed individually to ensure equal access to food. Animals were fed a low-protein NIH31 chow. AL animals had constant access to food, and DR animals were fed daily a pellet of 3g $\pm$ .1g in size (65% of expected AL consumption). Mice were weighed weekly during one of the daily feedings of DR animals to monitor health and weight loss. After 10 weeks of this regime, (except for the 0-day time points) the mice were given an intraperitoneal bolus injection of sterile D_2O saline at 35 μL/g body weight. This injection brought the mice up to 5% Molar Percent Excess (MPE) deuterium as previously described. Mice were then provided drinking water containing 8% MPE to maintain the 5% MPE in the animals' body water.

Euthenasia, tissue collection, turnover rate: Mice were anesthetized with CO_2 and then euthanized by cardiac puncture. Mice were then immediately dissected and all tissues except for blood were flash frozen on solid CO_2 and then stored at -80°C. Blood was stored on ice until it could be centrifuged at 800g for 10 minutes at 4°C. The centrifugation separated serum and red blood cells, which were stored in separate containers at -80°C. The AL and DR cohorts had 9 mice each. Two animals from each diet were sacrificed at each of these time points: 1 day, 3 days, 9 days and 27 days after bolus injection, with one animal from each group sacrificed without receiving a bolus injection or any other D_2O labeling.

Aliquots of serum were distilled in 2.0 mL screw cap tubes overnight in a 90°C sand bath, and the distillates collected. The distillate was diluted 1:300 in ddHO, and MPE of deuterium was directly measured against a D2O standard curve using a cavity ring-down water isotope analyzer (Los Gatos Research [LGR], Los Gatos, CA, USA) according to the published method. [289] Mouse liver tissue was placed in ammonium bicarbonate (ABC) (25 mM, pH 8.5) along with protease inhibitor cocktail (Sigma) and homogenized using a MP Biomedicals FastPrep-4 bead homogenizer at 6 m/s for 60 seconds. Volumes were calculated to give approximately 10 mg/mL protein. Protein concentration was measured using a bicinchoninic acid (BCA) protein assay (Thermo Fisher). 300-500 microg of protein were placed on 30 kDa centrifugal filters (VWR). 100 μL of a concentrated guanidine solution (6 M guanidine HCl, 100 mM Tris-HCl pH 8.5) was added to each sample and centrifuged at 14,000g for 15 minutes. This guanidine wash was repeated, and the flow-through discarded. Disulfide bonds were reduced using a 10 mM dithiothreitol/ M guanidine HCl/ mM Tris-HCl (pH 8.5) solution (100 μL total volume) added directly to the filters, with an incubation of 1 hour at 60°C in a sand bath. After 5 minutes of cooling, cysteine sulfhydryl groups were protected by reaction with iodoacetamide (IAM, 20 mM) for 60 minutes in the dark. Afterwards, the samples were centrifuged at 14,000g for 15 minutes, and the flow-through was discarded. The samples were washed twice with ABC (200 μL, centrifuged 15 minutes at 14,000g). Pierce MS-Grade trypsin was added (1:50 w:w) in 300 μL ABC to each sample, followed by incubation at 37°C overnight. Resulting peptides were eluted by centrifugation at 14,000g for 30 minutes, followed by a wash with 100 μL of ABC and an additional centrifugation for 30 minutes at 14,000g. Filters were discarded and the filtrate was dried using a speedvac (Sorval); the dried samples were stored at 4°C until use.

LC-MS Data Acquisition: Protein identification and kinetic acquisition were performed on two different mass spectrometers. First, the Agilent 6530 Q-ToF mass spectrometer coupled to capillary and nanoflow Agilent 1260 HPLC using the chipcube nano-spray source. [290] Peptides were eluted from the Agilent C18 Polaris chip at 300 nL/min using an H_2O-Acetonitrile gradient acidified to

pH 4 by use of Pierce LC-MS grade formic acid. Buffer A was 3% acetonitrile, 0.1% formic acid. Buffer B was 97% acetonitrile, 0.1% formic acid. The elution gradient was as follows: 0 minutes, 100% A; 0.1 minutes, 95% A; 27 minutes, 40% A; this was followed by high percentage B column washing and low percentage B equilibration. The Agilent 6530 Q-ToF mass spectrometer was run in 2 Ghz high dynamic range mode. Protein identification runs were performed in MS/MS mode using collision-induced dissociation (CID) with nitrogen gas. MS and MS/MS data were collected at a maximum rate of 4 spectra/second with CID fragmentation on the top 10 most abundant precursors. Dynamic exclusion was set to 0.2 minutes. Kinetic acquisitions were performed in MS only mode and collected at 1 spectra/second. MS only mode increases signal intensity, improves signal-to-noise, and gives more scan points per elution chromatogram, greatly enhancing isotopomer analysis accuracy.

Data were also collected on the Orbitrap Fusion-Lumos mass spectrometer. Samples were resuspended in 0.1% formic acid (Pierce LC-MS grade) in H_2O (Optima grade Thermo Fisher), and analyzed with a Thermo Lumos Tribrid (Orbitrap). Tryptic peptides were separated using a reverse phase C18 column (Acclaim PepMap trademark 100) and a Thermo Easy-Spray source. Mobile phase for the liquid chromatography was 0.1% formic acid in H2O (Buffer A) and 0.1% formic acid in 80% acetonitrile (Optima grade Thermo Fisher) with 20% H2O (Buffer B) on an Easy-nLC 1200 HPLC system. Samples were eluted using a gradient of 5% B to 22% B over 85min, 22% to 32% B over 15min, and a wash of 32% to 95% B over 10min, which was held at 95% B for 10min. Sample loading and equilibration were performed using the HPLC's built in methods. MS only runs were performed using 2400 V in the ion source, 60,000 resolution with a scan range of 375-1700 m/z, 30% RF Lens, quadrupole isolation, 80,000 AGC target, and a maximum injection time of 50 ms. MS/MS scans were performed using the same settings as MS only scans, with 3 seconds allowed per MS/MS after each MS scan, using the following filters: peptide monoisotopic peak determination, intensity threshold of 50,000, fragmentation of charge states +2 to +6, dynamic

exclusion that excluded a peak after being chosen once within 60 seconds, an error tolerance of 10 ppm high and low, and isotopes excluded. The fragmentation scan used an isolation window of 1.6 m/z, CID fragmentation with an energy of 30%, detection in the linear ion trap in Rapid Scan mode with an AGC target of 10,000, a maximum injection time of 35ms, and used the "Inject Ions for All Available Parallelizable Time" option.

Protein identification: Peak lists obtained from MS/MS spectra were identified using Mascot version 2.2.04, OMSSA version 2.1.9, X!Tandem version X!Tandem Sledgehammer (2013.09.01.1), MS-GF+ version Beta (v1), Comet version 2016.01 rev. 2 and MyriMatch version 2.2.140. The search was conducted using SearchGUI version 3.2.7. [291, 292] Protein identification was conducted against a concatenated target/decoy version of the Mus musculus complement of the UniProtKB (created September 2016, 16806 (target) sequences); decoy sequences were created by reversing the target sequences in SearchGUI. The identification settings were as follows: Trypsin, Specific, with a maximum of 2 missed cleavages 10.0 ppm as MS1 and 0.5 Da as MS2 tolerances; fixed modifications: Carbamidomethylation of C (+57.021464 Da); variable modifications: Oxidation of M (+15.994915 Da), Pyrolidone from Q (-17.026549 Da), Acetylation of protein N-term (+42.010565 Da), Pyrolidone from E (-18.010565 Da), Pyrolidone from carbamidomethylated C (-17.026549 Da); fixed modifications during refinement procedure: Carbamidomethylation of C (+57.021464 Da). Peptides and proteins were inferred from the spectrum identification results using PeptideShaker version 1.15.1. [293] Peptide Spectrum Matches (PSMs), peptides and proteins were validated at a 1.0% False Discovery Rate (FDR) estimated using the decoy hit distribution. Post-translational modification localizations were scored using the D-score and the phosphoRS score with a threshold of 95.0 as implemented in the compomics-utilities package.

Identification files and the MS-only mass spectrometry data were analyzed with the DeuteRater software package, which provided the protein turnover rates used for later analyses. [294] The resulting kinetic data was filtered to remove data with extreme outliers or other issues: First, rates

that were greater than 1 or less than 0.03 were eliminated from further analysis, as these rates represented extrapolations outside of the range of rates that could be calculated confidently with the time points used in the experiment. Since the kinetic proteomics data come from curve fits of relevant measurements, all curves with a Pearson's R^2 less than 0.5, or with a covariance (standard deviation/rate value) of greater than 0.2 were also removed from further analysis (the standard deviation was divided by the turnover rate for normalization).

The 10,601 oligopeptide fragments identified by MS/MS were sequence-matched to 3,426 distinct protein products across all diet conditions. Of these, exactly 1,600 met the data quality standards well enough to be assigned valid turnover rates for both the AL and CR conditions.

5.3.2 Stability

Previous research had simulated all possible network motifs using the equation

$$\begin{aligned} [\dot{P}_i] &= \alpha - \delta[P_i] + \sum_{i \neq j} k_{ji}[P_j][P_i]^h \\ \text{let } X_i &= \log[P_i] \rightarrow \\ \dot{X}_i &= -\delta + e^{-X_i}\left(\alpha + \sum_{i \neq j} k_{ji} e^{X_j + hX_i}\right) \end{aligned} \tag{5.1}$$

where $[P_i]$ is the concentration of protein i and X_i is the log transformation, α is the autocatalysis rate, δ is the degradation rate, k_{ji} is the rate constant for the effect of protein j on protein i (negative if downregulating, positive if upregulating, and 0 if j does not regulate i), and h is the Hill coefficient, which allows for nonlinearities in the regulatory system. In qualitative terms, Eq. 5.1 means that protein concentration increases by α due to natural background expression, decreases in a density-dependent way δ as proteins randomly become misfolded or fragment, and either up- or downregulate each other by k_{ji} in possibly non-linear ways (if $h \neq 1$).

The stability of the 132 topologically distinct 3-motifs and 22,650 topologically distinct 4-motifs was determined by selecting 1,000 vectors of parameters, and determining the number and location

of roots (protein concentrations such that $[\dot{P}_i] = 0 \forall i$) for each vector. The stability of each root was further determined by calculating its spectral radius at that point, that is, the largest real part of any of the system's Jacobian's eigenvalues. For continuous time PDEs, spectral radii < 0 are diagnostic of intrinsic stability (*sensu* [40]). Note that it was recently demonstrated that this criterion not only demonstrates that a network is intrinsically stable when the interactions happen instantaneously, but also stable under any time-lag condition. [145] This process took approximately 3 years of CPU time on the BYU supercomputer cluster, and is further detailed in Ch 2.

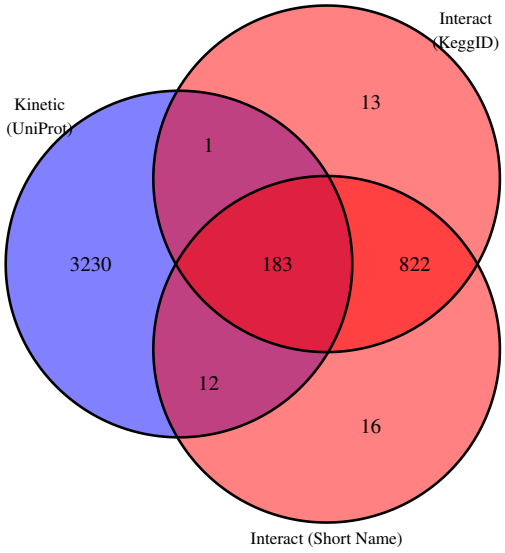

Figure 5.1 Venn diagram of proteins with turnover rate data (blue) and known network interactions (red, note that these were identified by both official short gene names and KeggID). Relatively few proteins were in both data sets, but of the 196 that were, 183 matched both short names and Kegg ID to the UniProtID of the turnover data set.

We then identify every 3- and 4-protein motif with known regulatory sign (up- or downregulating) in the house mouse *Mus musculus* (UniProt ID T01002) recorded in the RegNetwork database. [17] This database includes 1,033 proteins listed by official name and KEGG ID; [188] they were involved in 4,034 interactions (four in both a positive and negative direction, depending on cellular context), belonging to 109,000 distinct 3-motifs and 3.4 million 4-motifs. The stability metrics in Ch 2 were used to predict the abundance of these motifs relative to null expectations in Ch 3, and are used here to match known turnover rates.

Despite bi-directional querying of the API of four different databases (UNIPROT, KEGG, BiocManager, and DAVID) these UNIPROT IDs failed to cross-match for 837 of the KEGG ID and/or official abbreviation proteins in the RegNetwork database. The 196 successful cross-matches did not form a single complete 3- or 4-protein motif,

though they matched all-but-one protein in 2,322 and 12,580 motifs respectively. The proteins that did cross-match tended to be well-supported, with $> 93\%$ of matches occurring both directions in all four databases (183 of 196), representing 8-fold link confirmation (Fig. 5.1) Though not ideal, this was enough to significantly constrain the parameter space of the system described in Eq. 5.1, and draw conclusions about the relative likelihood of stability under the restricted and less-restricted diet regimes. Consider the ubiquitous A upregulates B and C motif. If we fix $\alpha = 1, h = 1$ for example, then the parameter space of this system can be visualized in three-dimensions: $[\delta, k_{12}, k_{13}]$ (Fig. 5.2). This parameter space can then be imagined as partitioned into regions of topologically distinct behaviors. For example, low δ and very high k values can cause concentrations to reach infinity mathematically, which would biologically represent cell death through over-consumption of resources. Intermediate values form a region with a single stable equilibrium. Other combinations cause this equilibrium to bifurcate into two stable equilibria or a limit cycle. If all three turnover rates were known, it would constrain the solution to a single point in this 3-space which unambiguously belongs to just one behavior regime. However, when one of the proteins' turnover rate is unknown, then the possible solutions form a line through parameter space; any point on this line could generate the two known turnover rates, for different values of the third unknown rate. However, different lines through parameter space spend more or less time in the stable regime, and thus the known rates can be used to determine the relative likelihood of stability (Fig. 5.3).

5.3.3 Bayesian inference of behavior across parameter space

We use a Bayesian approach to measure the posterior probability of stable vs. unstable behaviors across parameter values in our model. The likelihood was determined by the log-unnormalized density function

$$\log \mathscr{L} = \sum \log[\mathscr{N}(\dot{P}_i, 0, 0.1)] + \sum \log[\mathscr{N}(C_i, T_{obs,i}, 0.1)] + \sum \log[\mathscr{N}(A_i, T_{obs,i}, 0.1)]$$

where $\mathcal{N}(x,\bar{x},\sigma)$ is the density of the normal distribution at x for a given mean and standard deviation, $C_i = \alpha_i + \sum_{k_{i,j}>0} k_{i,j}[P_j][P_i]^h$ is the total catabolism of protein i, $A_i = \delta[P_i] + \sum_{k_{i,j}<0} k_{i,j}[P_j][P_i]^h$ is the total anabolism of protein i, and $T_{obs,i}$ is the observed turnover rate. That is, the first term penalizes departure from steady state, while the second and third terms penalizes proteins that are being built up / torn down at rates different than observed, respectively. We augment this likelihood function with uniform priors on log parameters in the range $\theta_i \in [10^{-3}, 10^{1}], \forall i$.

We use the Metropolis random walk algorithm to construct the Monte Carlo Markov Chains themselves. Specifically, we use the formulation proposed in [295] based on the philosophy of MCMC outlined in the introduction of [296], and implemented in the v9.7 MCMC package of R 4.3.1. In order to allow the exact observed turnover rates to be mathematically possible, parameters were expanded to include a separate α_i for each protein, while the universal degradation rate δ, and non-zero k_{ji}'s were selected to correspond to the median of the 1-fixed point behavior space. The initial protein concentrations were set to $[P_i] = 1 = \log 0$. In order to provide a good-but-not-exact initial guess, this point was then allowed to move towards the local maximum likelihood using 25 steps of Nelder-Mead optimization or until step-size was $< .001$.

Each of the 4,532 MCMC runs (2 diet conditions $\times$ 2,266 networks of three proteins where turnover rates were known under both diet conditions for two of the proteins) was run for 25,000 steps, then sampled every 25 steps for a final sampling of 1000 points. The success of an MCMC run was determined by the acceptance rate, length of the burn-in, and the autocorrelation spectrum. The analysis was performed again if acceptance rates fell outside of the 10-30% range with adjusted hyperparameters, such as coarser subsampling or relaxed error tolerance from the arbitrarily chosen $\sigma = 0.1$ in Eq. 5.3.3. If burn-in lasted over 25 steps (2.5% of the run), the run was performed again at new initial parameters. Finally, if autocorrelation extended further than the 25 steps of subsampling for more than one parameter, the run was adjusted and rerun.

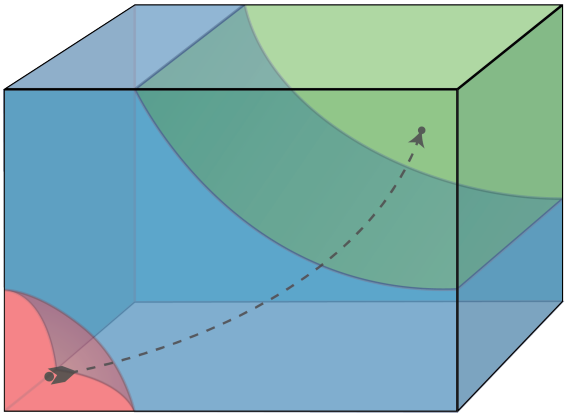

Figure 5.2 A conceptual model of a 3-dimensional parameter space, where values consistent with observed turnover rates and the steady-state assumption (dashed line) pass through three distinct behavioral regimes (colored regions of the space). Changing turnover rates moves the curved line, causing it to spend more or less time in the three regions, and therefore be considered more or less stable. MCMC samples are constrained to be near the curved line by the cost/log-likelihood function.

The behavior space of each network had been characterized in Ch. 3 by sampling 1,000 points across the d-dimensional parameter space and determining the number of fixed points there. For example, one parameter vector might produce 0 fixed points (the protein concentrations run to zero or infinity), another might have one fixed point (the system tends toward equilibrium no matter the initial protein concentrations), and a third might have four fixed points (multiple equilibria are possible, the final state depends on the initial concentrations). For each topology represented by the 2,266 3-motifs with two known turnover rates, the 1,000 known behaviors were supplemented with 4,000 additional solved points in the parameter space. This created a 5,000 point "Behavior Set" database for each topology. The membership of each point in the MCMC run to a behavioral regime was assigned by cross-matching each MCMC point to its $d+1$ nearest neighbors in the Behavior Set, which form a simplex around it. For example, a point in the MCMC run surrounded by points in the Behavior Set who all have 1 fixed point can be confidently asserted to belong to the 1-fixed point region. However, a point whose nearest neighbors show a mix of 0- and 1-fixed point behaviors is likely to be close to the stability boundary in the parameter space.

5.4 Observed Stability Increases

While previous studies indicated that on a CR diet (1) overall bulk protein turnover rate decreases, and (2) that a majority of individual proteins decrease their turnover, it was not known if (3) proteins that decreased that turnover were central in a regulatory context. That is, it is possible that toughened proteins are primarily end points of regulatory chains, while the relatively few embrittled proteins on average have many downstream regulatory targets. If this were the case, it would imply that CR disproportionately effects central regulators, which would be expected to destabilize regulatory networks.

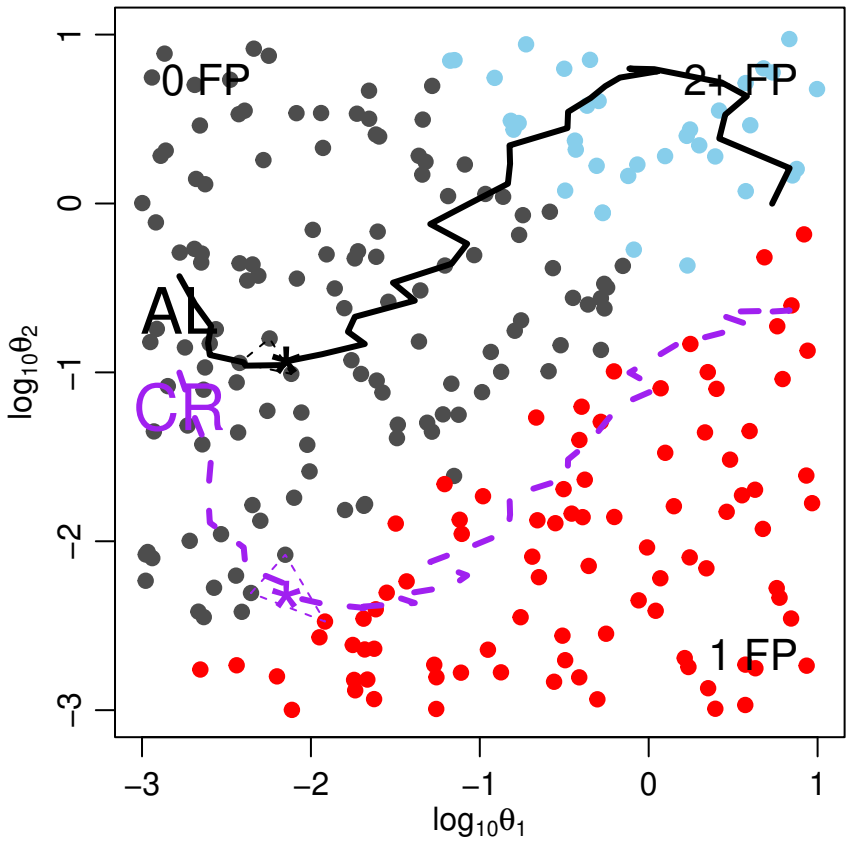

Figure 5.3 The number of fixed points is determined for a set of random parameter values (circles, grey = 0 fixed points, red = 1, light blue = 2 or more) as a proxy for behavior. In this example figure, we use θ_1 , θ_2, but actual networks are higher dimensional. Two MCMC paths then traverse this parameter space using the observed turnover rates under the *ad libitum* (black) and calorie restricted (purple) diets. At each point along the run, the $d+1$ nearest behaviors are tallied (dashed triangles around * example points). In this example, the percent of neighbors showing 0, 1, or 2+ fixed point behavior is 56, 0 and 44% for the AL diet, but 24, 75, and 1% for the more stable CR diet.

This was not the case. Of the 196 proteins found in both databases, 57.1% of them were toughened (in keeping with facts 1 and 2 above), and these toughened proteins were over-represented

across network motifs. Out of our set of 2,266 3-motifs with two known turnover rate changes, one would expect 739.7 where both toughened and 416.4 where both embrittled if choosing two proteins from the set of 196 randomly, but instead we observe 981 and 318 respectively ($\chi^2 = 56.9,\ df = 2,\ p < .0001$; see Fig. 5.4). This over-representation of toughened proteins indicates that CR disproportionately toughens important regulators, which is consistent with the hypothesis that CR leads to network stability.

Of all the embrittled proteins, all but eight appear in fewer than 40 motifs; the remaining eight appear in over 100. These proteins are all involved in fatty acid metabolism: three members of the Cytochrome p450 family (Cyp-4a10, -4a14, and -8b1), Apolipoprotein A1 (Apoa1), diazepam binding inhibitor (Dbi, which has long-chain fatty acyl-CoA binding activity and modulate autophagy) [297], fatty acid binding protein 1 (Fabp1), sterol carrier protein 2 (Scp2), and acyl-Coenzyme A oxidase 1 (Acox1). Many of these proteins are also involved in regulating peroxisome and mitochondrial activity via autophagy. Many previous studies have found that fatty acid metabolism and autophagy are processes strongly affected by calorie restriction. [298] Detailed study of this process indicates that CR mice endogenously synthesize ∼3x the fatty acid of controls in their subcutaneous and adipose tissues (but not the liver), to compensate for their ∼4x greater rates of fatty acid oxidation elsewhere in the body. Researchers have argued that this shift in metabolism is caused at the molecular level by an increased expression of genes such as Fabp1 and Acox1 in the five hours after daily feeding, and extremely low expression in the 19 hours thereafter. [299] This is

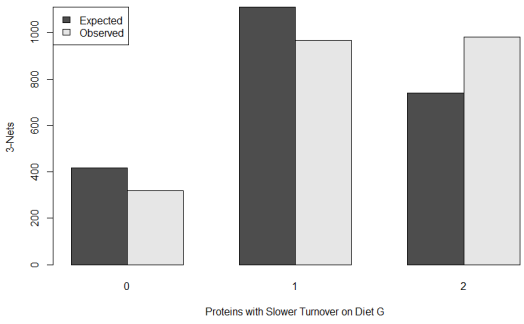

Figure 5.4 Toughened proteins (whose turnover rate decreases under CR) are over-represented in real-life metabolic networks. This indicates that, not only are most proteins toughened by CR, the toughened proteins are more central than the embrittled ones in a regulatory context.

in contrast to mice (or humans) on a more regular diet, where feeding has a relatively minor effect on synthesis rates. [300] Our sampling regime was not sufficiently detailed to pick out this peak, but suggests that even the most central embrittled proteins are in fact turning over very slowly most of the time, mobilizing only briefly in response to sharp changes in the energy environment of the cell. It also suggests that exceptions to the "regulators slow down turnover" rule are permitted only if they are crucial to meeting the cells' constrained energy budgets.

A more direct test of the link between diet and stability comes from the MCMC analysis outlined above. Each 3-motif generated a set of 1,000 parameter vectors along the MCMC path, and each of these points was matched to its nearest neighbors in parameter space in the behavior set. To be sure our results were consistent, we used four different stability metrics, and compared the differences in stability in four different ways. Across all sixteen of these tests, a CR resulted in a modest increase in stability (average across tests 1.22%, range 0.93% to 1.7%; ensemble t-test: $t = 18.4,\ p < 0.0001$; Fig. 5.5). This stability increase was significant for all tests where stability was measured by the simplex of points around the MCMC run, and significant for only some of the tests where stability was measured by the single nearest neighbor to the MCMC run. Tests were fairly consistent, whether stability metrics were compared using paired or unpaired means, and whether the mean was calculated by weighting the parameter vector by its likelihood or not.

Though relatively modest, it should be kept in mind that these percentage differences represent alternate paths through parameter space due to incomplete turnover information, not actual stability solutions for fully determined motifs. It is entirely possible that full information would greatly enhance the signficance of our results by greatly decreasing the error bars around each point. It should also be noted that even minor differences in probability become exaggerated when experienced continuously through through time, and may be able to account for the moderate lengthening of lifespan caused by a change in diet.

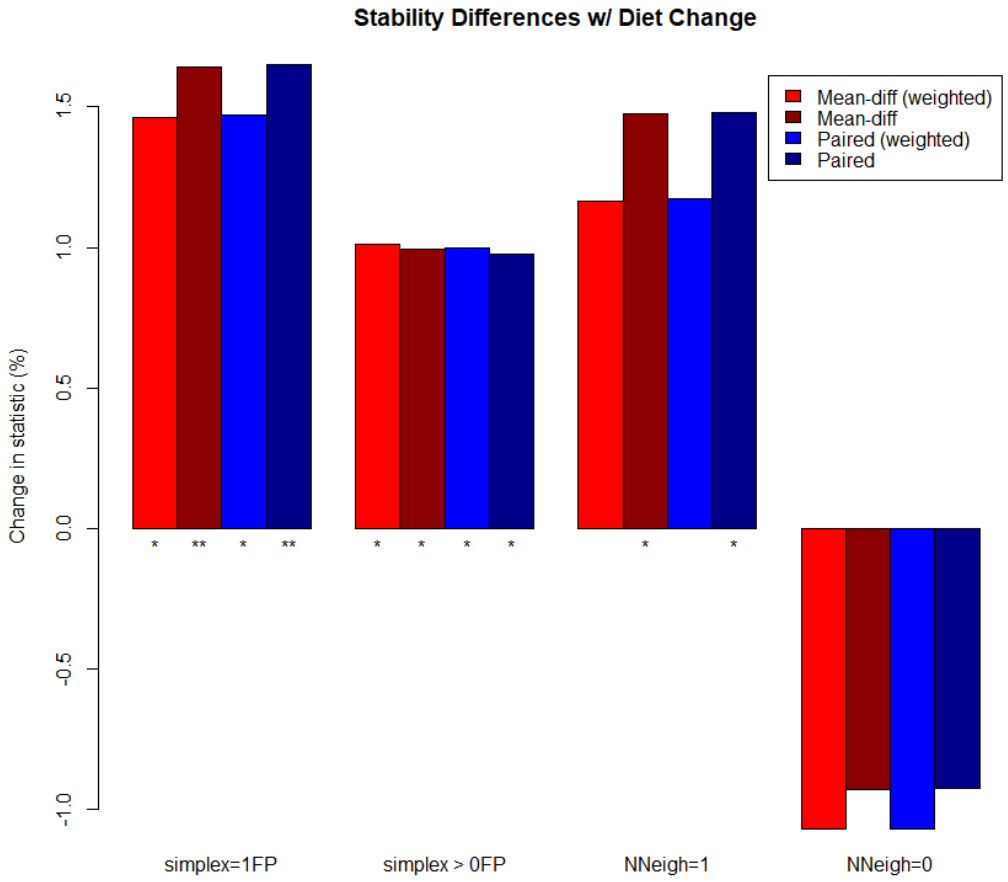

Figure 5.5 A set of 16 different of statistical measures indicate that all testable 3-motifs in mice are between 1.0 and 1.5% more stable (variously defined) under CR conditions. While the magnitude of these differences is consistent, the significance of the differences varied based on the consistency of the stability metric. For stability defined by the simplex of nearest neighbors that either had one equilibrium (simplex=1FP) or had any equilibria at all (simplex >0FP), all comparisons were significant at the $\alpha = .05$ level. By contrast, those tests where stability was determined only by a single nearest neighbor, whether that neighbor was stable (NNeigh=1) or unstable (NNeigh=0) all showed increases in stability, but this increase was typically less significant. Contrasts either used the ensemble mean stability of the CR and AL diets or the mean difference between the diets for each motif (mean diff vs paired), where these means were either weighted by their log-likelihood or not.

5.5 Well-defined Behavior Regions

All of the foregoing analysis has assumed that parameter space is partitioned cleanly into regions of distinct behaviors, with boundaries that are relatively smooth and continuous, as depicted conceptually in Fig. 5.2. Approaching this problem from the perspective of bifurcation theory, this seems like a reasonable assumption. Classically, equations that exhibit changes from one behavior to another are said to have undergone a bifurcation, or non-homeomorphic transitions in their phase-space to adopt the language dynamical systems theory. For example, the Hopf equation $\dot{z} = z(a + b|z|^2)$ has a stable fixed point at the origin of the complex plane if $a < 0$ and a stable limit cycle of radius $\sqrt{-a/\Re(b)}$ if $a > 0, \Re(b) < 0$, and is unstable otherwise. The changes from one behavior to another are unambigious and occur exactly at $a = 0$ and $\Re(b) = 0$. Similarly, the standard bifurcation text goes, the logistic map $N_{t+1} = rN_t(1 - N_t)$ transitions from a single equilibrium where $r < 3$, to a cycle of period 2 as r increases, then period 4, then period 8, and so forth until reaching chaos when $r \approx 3.57$. However, many bifurcations partition their parameter space in ways that are far less cut-and-dried. With the exception of "normal-form bifurcations", the boundaries of behavioral regions frequently depend on nonlinear combinations of parameters, and so divide parameter space with boundaries that are slanted and curved, but otherwise smooth. [68] Worse, transitions into chaos are sometimes characterized by boundaries that are periodic or fractal, and the chaotic regions contain isolated islands of stability. [301–303] That is, parameter space is not always partitioned neatly into distinct regions like Neopolitan ice cream, as one would expect from normal form bifurcations, but instead can consist of irregular boundaries and unexpected islands of stability in a matrix of instability, like marble cake.

All this led us to question our assumption that the equations governing network dynamics created nicely partitioned behaviors across parameter space. To quantify the apparent smoothness of topological boundaries, we determined the behavior of 1,000 random parameter vectors, then asked what fraction of these points' nearest neighbors had the same behavior. We call this the

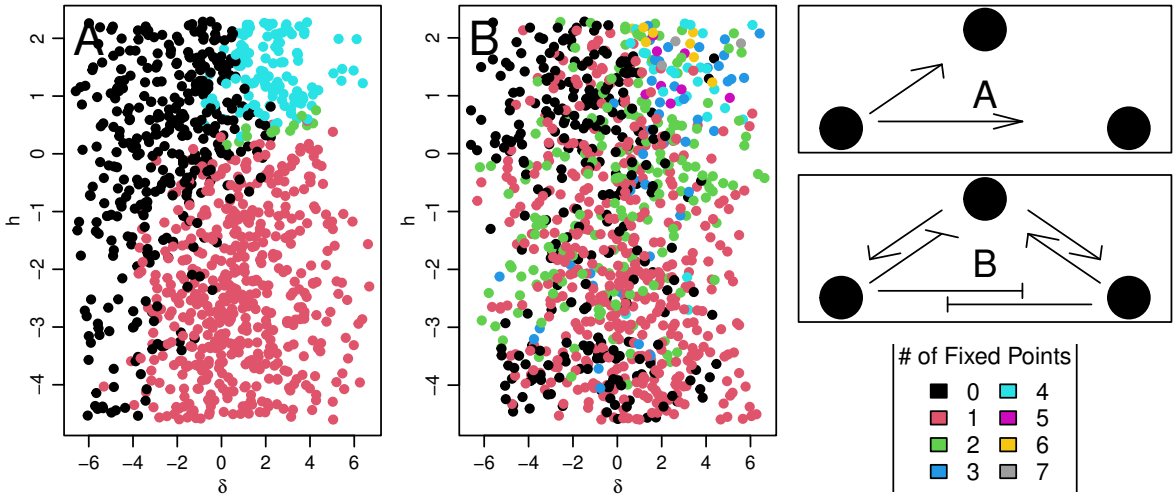

Figure 5.6 The assumption that parameter space is neatly partitioned into well-bounded behavior spaces is more true for some motifs than others. The clumpiness index (defined in the text) ranges from (A) $C = 0.898$ for the motif where protein X coregulates Y and Z, ubqiutious in real life, to (B) $C = 0.553$ for a more complicated network that has not yet been observed in nature. Because panels A and B represent two-dimensional projections of higher dimensional spaces, some of the overlap can be due to behavior boundaries being tilted in other dimensions.

clumpiness index, C. For motifs where behaviors are tidily clumped together, this index would approach 1, as only points lying very close to each other across a small surface-area boundary would be different. On the other hand, motifs where boundaries are fjord-like or even fractal, or have many behavioral enclaves within regions of different behaviors, would approach the null expectation $C_0 = \sum p_i^2$ where p_i is the proportion of parameter vectors displaying each behavior. (For example: if 30% of simulations have 0 fixed points, 50% have 1 fixed point, 20% have 2 fixed points, then $C_0 = .3^2 + .5^2 + .2^2 = .38$) Note that C_0 thus defined is also called Simpson's Diversity in ecology, and the Herfindahl–Hirschman index in economics.

The clumpiness of observed networks was in fact quite variable across motifs, ranging from about 50% to 90% for 3-motifs, and from 35% to 100% for 4-motifs. This was further simplified by considering only networks with a global equilibrium (1 fixed point) stable, and those with 0 or multiple equilibria as "less stable". We then use Fisher's exact test to calculate the probability of the

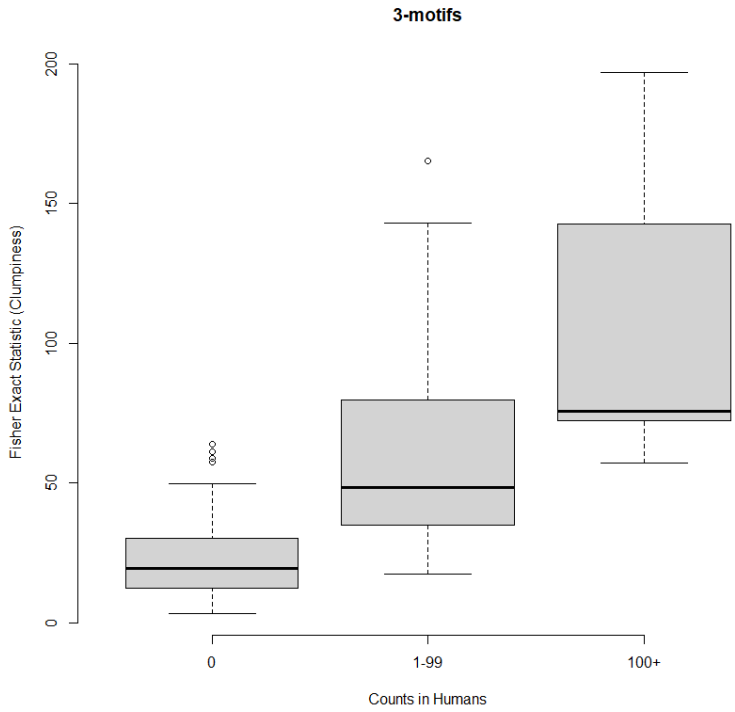

Figure 5.7 The degree of clumpiness relative to expected was a strong predictor of the abundance of 3- and 4-motifs in humans and mice. We show human 3-motifs here, but quantitatively similar results were obtained for the other three conditions.

observed number of nearest neighbors having the same stability state. This index strongly predicted the observed count of each network in both humans and mice for both 3- and 4-motifs (Fig. 5.7). This is in contrast to previous attempts, which failed to find strong correlations between abundance and any of twelve different measures of stability.

5.6 Conclusions

The evidence presented here indicates that CR mice do in fact have more stable PPINs than non-CR mice, and there is a mechanistic reason to believe this is a major contributor to the increase in lifespan. Not only are individual proteins, on balance, toughened by CR, but the minority of proteins that are embrittled are under-represented in the regulatory networks we have data for.

Summarizing a 70-year career in 2006, J. C. Waterlow concluded his masterful textbook on protein turnover by saying, "It seems that synthesis of the proteins that make up the bulk of our tissues, and the regulation of that synthesis, involve a vast array of proteins that themselves undergo turnover, which is presumably regulated by other proteins, and so on *ad infinitum*. This is not a very satisfactory conclusion." [304] The specifics of the vast protein regulatory network presented above hopefully provide a more satisfactory insight into the nature of this process.

The data presented here also provide an explanation for why some motifs are more common than others. Natural selection appears not to favor networks with a larger parameter space, but rather networks whose stable regions (of whatever size) have a more clearly defined stable. Mechanistically, a small and uniformly stable region is more robust to environmental fluctuations and chemical stochasticity than even a large stable region riddled with pockets of instability. Such networks could easily provide a constant background of disregulation that consumes the cells resources bringing them back in line, and possibly even tips them into alternate stable states that are oncogenic or senescent for example.

There are basal metazoan organisms that seem to escape aging altogether, generally by cloning or keeping a high proportion of toti- and pluri-potent stem cells within the organism. [305] Even within the complex bilatarians, germline cells can be thought of as functionally immortal; less benignly, so can tumors. Thus, it is apparently possible for individual cells to continue carrying out their living functions as long as repair mechanisms outpace the rate of decay. The geroscience hypothesis states that if the deterioration mechanisms of somatic tissues can be understood and overcome therapeutically (perhaps using pathways similar to the functionally immortal germ line) then the course of aging can be permanently altered. [263] We know it is possible for the proteostasis network to be hijacked in a negative way; *e.g.*, by parasites that produce enzymes which activate the degradation pathway to cause tissue necrosis in leishmaniasis. [306] Is it possible to hijack the network in the positive direction as well? An RNA-Seq study of BJ fibroblasts as they shift from

natural to immortal to transformed and finally to metastasized suggests that while relatively few genes change on the last step, the barrier between immortal and transformed is fairly high with 856 of the 1,357 differentially expressed cells (63%) changing at that step, suggesting it may be fairly easy to push cells into immortality without continuing on to cancer, though this has proven difficult so far. Intriguingly, nearly 80% of the mRNAs were downregulated by transformation, the majority on cell membranes, implying a despecialization of immortal cells, as one would expect from comparative evolution. [307] However, multi-omics studies teach us to be wary of assuming changes to mRNA imply changes to proteins, either concentration or kinetics. [48]

While experiments in humans are, of course, ethically problematic, comparative studies of populations with diets that promote naturally low protein turnover are also correlated with long lifespans: Okinawa, Japan; Sardinia, Italy; and among Seventh-Day Adventists in Loma Linda, California. [308–311] These diets are either low protein, low calorie, or both. This intriguing hypothesis is in need of further confirmation in humans. Improvements to analytical methods and software make collecting protein turnover information far faster, more sensitive, and less invasive, so large-scale turnover studies are now possible in humans. [312]

Unfortunately, CR in humans is extremely difficult to maintain for extended periods of time. When the physiologist Ancel Keys using conscientious objectors to simulate starvation conditions in prisoner-of-war camps as World War II, the participants described feelings of hatred to people eating regular portions, resentment “as strong as anything [they had] ever experienced”, marked depression, loss of sexual drive, and unsustainable behavior modifications to conserve energy. [313, 314] Meanwhile, Europeans undergoing similar (but involuntary) deprivations experienced the negative psychological side effects, but also a 34% reduction in death rates in Denmark during WWI, and 30% in Norway during WWII. [315,316] A more recent study with the modest goal of a 20% CR diet or an equivalent amount of exercise, despite significant interventional effort by experimenters, was able to achieve only 11.5% CR over the course of a year; even this modest decrease was associated

with health improvements. [317] A review of the many attempts at CR experiments concluded there were both physiological (low blood pressure, decreased sex hormones, bone thinning, muscle wasting, slow wound healing, and decreased innate immunity) and psychological (depression, low libido, mood swings, and social isolation) side effects that were strong enough to warrant extreme caution in adopting this diet. [318] Despite these challenges, enough compliance with large scale studies such as CALERIE-1 and -2, Biosphere 2, and Keys' Wisconsin Starvation study indicate there are many health benefits observable in the short- and long-term to humans able to maintain this diet. [311]

For these reasons among others, CR is unlikely to be a viable therapeutic option for most humans. However, this study indicates that the benefits of CR accrue through toughened proteins and slower turnover, so it may be possible to achieve this via pharmaceuticals and avoid the compliance problems of a long-term CR diet. Many longevity treatments already focus on decreasing protein turnover, [269, 319, 320] and stabilizing regulatory bionetworks. [321] Dozens of possible Calorie Restriction Mimetics (CRMs) have already been tested in animal models since 2-deoxy-D-glucose was proposed in 1998, some with promising extensions to lifespan; clinical trials of these drugs in humans has already begun, though with mixed results. [322–325] If these interventions succeed in mimicking the results of $\sim$40% CR reduction in other species, then perhaps the $\sim$30% increase in lifespan resulting from lifelong CR is achievable in humans as well. Growing experimental and mathematical arguments suggest that humans "stop aging" in their late 90s in the sense that risk of mortality plateaus, an effect observed in other species. [326] If turnover-slowing interventions enable humans to reach that plateau, lifespan may indeed be greatly extended. We hope that this research clarifies that the mechanism by which CR promotes longevity is moving PPRNs back from the edge of instability, and future therapies will be able to provide the same buffer to human life.

Chapter 6

Conclusion

This thesis has made the following contributions:

- Developed TWIG, a method for characterizing bifurcations that offers some benefits over currently established methods. For better or worse, one of those is that it offers insights into the nature of even very complicated bifurcations that otherwise would require substantial mathematical training to realize.
- Leveraged this understanding of bifurcations to uncover that even models of simple networks exhibit a large number of behaviors, many of them harmful when translated back into their biological meaning.
- Demonstrated that, counter-intuitively, natural selection does not appear to systematically favor network structures that maximize stability *per se*, at least as it is usually measured.
- However, real life bionetworks are largely shaped by favoring topologies where the stable region has well-defined boundaries.
- Finally, we demonstrated that the link between longevity and slow protein turnover is mediated by increased stability across the protein regulatory network.

While a link between protein network stability and longevity has been suspected since antiquity, it is in only in the last decade that tools to prove it have been developed in several widely separated subdisciplines, and this first attempt to link them all together, while ultimately successful, is based on less complete data and less well-established tools than we might like; in the case of TWIG, we were forced to create the tools ourselves. We hope this early success will inspire further refinement of these methods, and spur studies that analyze the regulatory details of protein-protein interactions, rather than merely cataloguing their undifferentiated and unexamined presence. While we would like to believe that the modest increases in stability recorded by comparing the relatively likelihood of MCMC runs across incompletely specified parameter space will snap into something stronger with complete turnover information for thousands, if not millions, of real life motifs, only time will tell if this is the case.

As the gaps in protein network interaction databases fill in, there is also the hazard that the relative frequencies of motifs may begin to shift. There are reasons to think many of the sparse motifs are solid, even with gappy data: they appear in data from different taxa and even across kingdoms, and they are overrepresented not in an absolute sense, but relative to null expectations. For example, the A co-regulates B and C motif being far more common than A->B->C passthrough motif is unlikely to change because random assembly should make passthrough twice as frequent as coregulation. However, conclusions about well-connected motifs are still sparse because so few examples exist. As more regulatory links are discovered, the sample size is likely to grow to the point that we can say more interesting things about the under-/over-representation of topological cliques and near cliques.

Another hope is that with more data, it will become possible to distinguish between different network assembly rules with a fair degree of certainty. Ecological networks are shaped by one of perhaps as few of three biological principles that approaches the status of a law, niche exclusion, [4]

which places significant constraints on the types of networks that can be formed. Are future quasi-laws of network formation waiting to be discovered for proteins as well?

Of course, the most commercially interesting implication of this study is that it may be possible to extend human lifespan by stabilizing protein networks. However, comparative biology suggests that immortality, be it among basal metazoans or stem cells or cancers, appears to derive from a lack of the specializations necessary for complex life. Is it possible to maintain this necessary individuality while also gaining immortality? The quasi-theological sound to this question suggests that it may be a long time before we have a definite answer. However, this thesis suggests that calorie-restriction mimetics can gain the life-extending benefits of CR by focusing on decreasing post-transcriptional protein turnover, possibly of only a few near-tipping point networks, rather than trying to match all the many and varied physiological, psychological, and chemical effects of CR. [240, 324, 325] Given our reasonable understanding of the mechanisms of protein synthesis and degradation, some level of control seems feasible in the near term. [319] For example, turnover rates can be slowed globally by inhibiting ribosomal initiation factors; more targeted siRNA therapies for vulnerable networks could also be devised.

This thesis focused on longevity as a convenient test case, but the approach also holds promise for cancer, schizophrenia, and other partially inherited but "omnigenic" conditions, where it appears practically all genes in the genome contribute to the condition's heritability. Many of these conditions can be thought of as an extremely high-dimensional steady state of the regulatory network, existing in its own basin of attraction separate from the typical basin of attraction enjoyed by most people's regulatory network. While there is no hope of reducing such a system to a normal form, TWIG holds out some promise for characterizing the high-dimensional separatrix between the two states, and therapies that tweak reaction rate constants or temporarily jostle key protein concentrations hold out further promises of tipping cells over that ridge and restoring normal function. This of course relies on the hope that characterizing turnover of the entire proteome will gives us a clearer

picture of key pathways than characterizing the RNA and protein concentrations has so far done. However, the faster protein turnover rate in cancer cells [327] makes this likely, and indeed has already been exploited by protease inhibitors like bortezomib, [328] which causes protein buildup faster in tumors than regular tissue, triggering apoptosis, and successful treatment in some forms of myeloma. [329]

Swinging from the most applied to the most basic of implications, we also hope that TWIG analysis will contribute to bifurcation research. First, we expect it to place bifurcation analysis within the range of more scientists lacking the extensive mathematical background usually needed to contribute to bifurcation theory. Second, we hope the method will be extended so that mapping of the bifurcation surface (separatrix) can be automated and optimized to deal with high-dimensional systems. Finally, we envision a future where TWIG's exploitation of information geometry to pluck driving parameters out of a complex system enables a lower-cost lower-effort method for systems or network pharmacology. Currently, much of drug discovery is phenomenological: a substance is found that seems to have a beneficial effect, and is placed into a testing pipeline with little understanding of how or why it is working. Such drugs frequently fail out of the system, but at a cost both financial and in lives of test animals. [330] A tool that can identify key pathways is urgently needed, [280, 321] and a general bifurcation analysis package as already been called for as a potential solution. [172]

With the accumulation of Big Data about diseases and biological systems, it appears clear that "the answer" to many medical mysteries already lies on hard drives around the world. What is needed is better tools to dig the understanding out of this ocean of information. This thesis has made a few tentative paddles toward this horizon, and anticipates a future of far grander and beneficial explorations to come.

Appendix A

Fisher Information Matrix Derived for Normal Form Bifurcations

A.1 FIM of Saddle-Node Bifurcations

The normal form of the saddle-node bifurcation is

$$\frac{dy}{dt} = r + y(t)^2 + \alpha_1 y(t)^3 + \alpha_2 y(t)^4 + \dots \tag{A.1}$$

This differential equation can be solved locally when all parameters $\overrightarrow{\theta} = 0$, which happens to be the bifurcation point of the system. At that point:

$$\frac{dy}{dt} = y^2 \rightarrow \frac{dy}{y^2} = dt$$

Integrating both sides yields

$$-\frac{1}{y}\Big|_{y_0}^{y(t)} = t\Big|_0^t$$

$$\frac{1}{y_0} - \frac{1}{y(t)} = t$$

$$y(t) = \frac{y_0}{1 - y_0 t} \tag{A.2}$$

This implies there is a singularity at $t = 1/y_0$, so a proper coarse-graining procedure will involve taking data from $t = 0$ to some value near $1/y_0$, say $0.99/y_0$. We avoided this singularity by using negative values for y_0 and were therefore able to run simulations to large values of t_{max}. As noted in Eq. 2.6, to find the FIM of a system it is only necessary to find the Jacobian, so we need only find the first partial derivative of these data with respect to each parameter in the model.

A.1.1 Partial derivative of r

Let the α_i's=0. The derivative of the normal form w.r.t. r becomes:

$$\begin{aligned} &\frac{\partial}{\partial r}\left(\frac{\partial y}{\partial t} = r + y^2\right) \\ &\frac{\partial^2 y}{\partial r \partial t} = 1 + 2y\frac{\partial y}{\partial r} \end{aligned} \tag{A.3}$$

We let $w = \frac{\partial y}{\partial r}$, and this becomes $\frac{\partial w}{\partial t} = 1 + 2yw$, which requires the use of an integration factor to solve [331]. If $p_1 x' + p_0 x = q$ then

$$x = \frac{1}{\mu p_1}\left[C + \int \mu q dt\right] \text{ where } \mu = p_1^{-1}\exp\left(\int \frac{p_0}{p_1} dt\right) \tag{A.4}$$

Allowing $p_1 = 1$, $p_0 = -2y$, $q = 1$ implies that

$$\begin{aligned} \mu &= 1^{-1}\exp\left(\int \frac{-2y}{1} dt\right) \\ &= \exp\left(-\int \frac{2y_0 dt}{1 - y_0 t}\right) \\ &= \exp(2\ln(1 - y_0 t)) \\ &= (1 - y_0 t)^2 \end{aligned}$$

Therefore,

$$\begin{aligned} w &= \frac{C + \int (1 - y_0 t)^2 dt}{(1 - y_0 t)^2} \\ &= \frac{C - \frac{(1-y_0 t)^3}{3y_0}\big|_0^t}{(1 - y_0 t)^2} \\ &= \frac{C + \frac{1-(1-y_0 t)^3}{3y_0}}{(1 - y_0 t)^2} \end{aligned}$$

Recall this function is being evaluated at the initial condition, where the partial derivative $w = \frac{\partial y}{\partial r} = 0$ (i.e., changes to r do not change y_0). This implies that $C = -\frac{1-(1-y_0 t)^3}{3y_0}$; when $t = 0$ this further reduces to $C = 0$. Therefore.

$$\frac{\partial y}{\partial r} = \frac{1 - (1 - y_0 t)^3}{3y_0 (1 - y_0 t)^2} \tag{A.5}$$

A.1.2 Partial derivative of α_1

Using the same procedure as above,

$$\begin{aligned} \frac{\partial}{\partial \alpha_1}\left(\frac{\partial y}{\partial t}\right) &= \frac{\partial}{\partial \alpha_1}\left(y^2 + \alpha_1 y^3\right) \\ \frac{\partial^2 y}{\partial \alpha_1 \partial t} &= 2y\frac{\partial y}{\partial \alpha_1} + y^3 + \cancel{3y^2 \alpha_1 \frac{\partial y}{\partial \alpha_1}} \\ \frac{\partial w}{\partial t} &= 2yw + y^3 \end{aligned} \tag{A.6}$$

Note on the second line, we are able to cancel the third term because we are evaluating the slope where α_1 is zero. On the last line, note that p_0 and p_1 are the same as for r, so as above $\mu = (1 - y_0 t)^2$, but since now $q = y^3$:

$$
\begin{aligned}
w &= \frac{C + \int y^3 (1 - y_0 t)^2 dt}{(1 - y_0 t)^2} \\
&= \frac{C + \int \left(\frac{y_0}{1 - y_0 t}\right)^3 (1 - y_0 t)^2 dt}{(1 - y_0 t)^2} \\
&= \frac{C + \int \frac{y_0^3 dt}{1 - y_0 t}}{(1 - y_0 t)^2} \\
&= \frac{C - y_0^2 \log(1 - y_0 t)|_0^t}{(1 - y_0 t)^2} \\
&= \frac{C - y_0^2 \log(1 - y_0 t)}{(1 - y_0 t)^2}
\end{aligned}
$$

Again, assuming $w = t = 0 \rightarrow C = 0$, so

$$
\frac{\partial y}{\partial \alpha_1} = -\frac{y_0^2 \log(1 - y_0 t)}{(1 - y_0 t)^2} \tag{A.7}
$$

A.1.3 Partial derivatives of higher-order α's

Higher order terms in the series are of the form $\alpha_n y^{n+2}$ and so

$$
\begin{aligned}
&\frac{\partial}{\partial \alpha_n}\left(\frac{\partial y}{\partial t} = y^2 + \alpha_n y^{n+2}\right) \\
\frac{\partial^2 y}{\partial \alpha_n \partial t} &= 2y\frac{\partial y}{\partial \alpha_n} + y^{n+2} + \cancel{(n+2) y^{n+1} \alpha_n \frac{\partial y}{\partial \alpha_n}} \\
&\frac{\partial w}{\partial t} = 2yw + y^{n+2}
\end{aligned} \tag{A.8}
$$

As above, we are able to cancel $(n+2) y^{n+1} \alpha_n \frac{\partial y}{\partial \alpha_n}$ because we are solving for slopes about the point $\alpha_n = 0$. With the same value of μ, we use integration factors to demonstrate:

$$
\begin{aligned}
w &= \frac{C + \int \left(\frac{y_0}{1-y_0 t}\right)^{n+2} (1-y_0 t)^2 dt}{(1-y_0 t)^2} \\
&= \frac{C + \int \frac{y_0^{n+2} dt}{(1-y_0 t)^n}}{(1-y_0 t)^2} \\
&= \frac{C - \frac{y_0^{n+1}}{1-n}(1-y_0 t)^{1-n}|_0^t}{(1-y_0 t)^2} \\
&= \frac{C + \frac{y_0^{n+1}}{n-1}(1-(1-y_0 t)^{1-n})}{(1-y_0 t)^2}
\end{aligned}
$$

Which again implies that $C = 0$ at the initial condition $t = 0$, and so for $n > 1$ we can say

$$
\frac{\partial y}{\partial \alpha_n} = \frac{y_0^{n+1}(1-(1-y_0 t)^{1-n})}{(1-n)(1-y_0 t)^2} \tag{A.9}
$$

Recall that the Jacobian of our system is

$$
J = \begin{bmatrix}
\frac{\partial y_0}{\partial r} & \frac{\partial y_0}{\partial \alpha_1} & \frac{\partial y_0}{\partial \alpha_2} & \cdots \\
\frac{\partial y_1}{\partial r} & \frac{\partial y_1}{\partial \alpha_1} & \frac{\partial y_1}{\partial \alpha_2} & \cdots \\
\cdots & \cdots & \cdots & \cdots
\end{bmatrix} \tag{A.10}
$$

Because the Fisher information matrix $\mathscr{I} = J^T J$, we can see that element $\mathscr{I}_{1,1} = \left(\frac{\partial y}{\partial r}\right)^2$ will be $\mathscr{O}(t^2)$ because $\frac{\partial y}{\partial r}$ is $\mathscr{O}(t^1)$; all other elements will be a lower order of t. Thus, at long time scales, the FIM's element (1,1) will grow faster than all other elements, and therefore the most relevant parameter is clearly r.

In the case where $\mathscr{I}$ is being derived from data (or from noise added to a non-/normal form equation), the importance of r can be evaluated by increasing $\sigma^2 \propto y_0^{-3}$. Since, by the central limit theorem standard error $\sigma^2 \propto n$, then the number of time points sampled should decrease as $n \propto y_0^{-3}$.

A.2 FIM of Transcritical Bifurcations

These have a similar normal form as the saddle-node bifuractions above:

$$\frac{dy}{dt} = ry(t) - y(t)^2 + \alpha_1 y(t)^3 + \alpha_2 y(t)^4 + ...$$

However, the change of sign in the second term causes the solution to the differential equation to also have a changed sign:

$$\begin{aligned} \frac{dy}{dt} = -y^2 \rightarrow -\frac{dy}{y^2} = dt \rightarrow \frac{1}{y}\Big|_{y_0}^{y(t)} = t\Big|_0^t \\ \frac{1}{y(t)} - \frac{1}{y_0} = t \rightarrow y(t) = \frac{y_0}{1+y_0 t} \end{aligned} \tag{A.11}$$

Now the singularity occurs at $t = -\frac{1}{y_0}$, which generally only complicates the coarse-graining if initial conditions are negative.

A.2.1 Partial derivative of r

The full solution to the partial derivative of r is somewhat complicated because it depends on y:

$$\begin{aligned} \frac{\partial}{\partial r}\left(\frac{\partial y}{\partial t} = ry - y^2\right) \\ \frac{\partial^2 y}{\partial r \partial t} = r\frac{\partial y}{\partial r} + y - 2y\frac{\partial y}{\partial r} \\ \frac{\partial w}{\partial t} = w(r-2y) + y \end{aligned} \tag{A.12}$$

where $w = \frac{\partial y}{\partial r}$. Recall that the derivative is being evaluated where $r = 0$, and so we can argue that

$$\begin{aligned} \frac{\partial w}{\partial t} + 2yw = y \rightarrow \\ \mu &= \exp\left(\int \frac{2y_0 dt}{1+ty_0}\right) \\ &= \exp[2\log(1+ty_0)] \\ &= (1+ty_0)^2 \end{aligned} \tag{A.13}$$

Using our integration factors, we see:

$$\begin{aligned} w &= \frac{C + \int (1+ty_0)^2 \frac{y_0}{1+ty_0} dt}{(1+ty_0)^2} \\ &= \frac{C + y_0 t(1+\frac{y_0 t}{2})}{(1+ty_0)^2} \rightarrow C = 0 \\ &= \frac{y_0 t(2+y_0 t)}{2(1+ty_0)^2} = \frac{\partial y}{\partial r} \end{aligned} \tag{A.14}$$

Note that in the limit that $t \rightarrow \infty$, this expression is order 0 for t; therefore, unlike the other bifurcation classes, transcriticals are expected to have a *relevant*, rather than a hyperrelevant, leading eigenvalue. This was confirmed with simulations (see Fig. 2.7).

A.2.2 Partial derivative of α_1

The derivative can be set up as:

$$\begin{aligned} \frac{\partial}{\partial \alpha_1}\left(\frac{\partial y}{\partial t} = -y^2 + \alpha_1 y^3\right) \\ \frac{\partial^2 y}{\partial \alpha_1 \partial t} = -2y\frac{\partial y}{\partial \alpha_1} + 3\alpha_1 y^2 \frac{\partial y}{\partial \alpha_1} + y^3 \\ \frac{\partial w}{\partial t} = -2yw + y^3 \end{aligned} \tag{A.15}$$

Since we already know that $\mu = (1+ty_0)^2$, it follows that

$$\begin{aligned} w &= \frac{C + \int (1+ty_0)^2 \left(\frac{y_0}{1+ty_0}\right)^3}{(1+ty_0)^2} \\ &= \frac{C + y_0^2 \int \frac{y_0}{1+ty_0}}{(1+ty_0)^2} \\ &= \frac{C + y_0^2 \log(1+ty_0)}{(1+ty_0)^2} \rightarrow C = 0 \\ \frac{\partial y}{\partial \alpha_1} &= \frac{y_0^2 \log(1+ty_0)}{(1+ty_0)^2} \end{aligned} \tag{A.16}$$

A.2.3 Partial derivative of higher-order α's

Using similar arguments, we arrive at the conclusion that for α_n where $n > 1$

$$\frac{\partial y}{\partial \alpha_n} = \frac{y_0^{n+1}((1+ty_0)^{1-n}-1)}{(1-n)(1+ty_0)^2} \tag{A.17}$$

Plots of the sensitivities suggest that r is the dominant parameter for values of $y_0 < 1$, though exactly where this transition occurs is probably worth investigating.

The top-left entry in the FIM is

$$\begin{aligned}\mathscr{I}_{1,1} &= \left(\frac{\partial y}{\partial r}\right)^2 \\ &= \left(\frac{y_0 t(y_0 t+2)}{2(y_0 t+1)^2}\right)^2 \\ &= \frac{y_0^2 t^2 (y_0 t+1)^2}{4(y_0 t+1)^4}\end{aligned} \tag{A.18}$$

In the limit $t \to \infty$, this approaches $\frac{t^4}{t^4}$ which is order $\mathscr{O}(t^0)$. This implies that the leading eigenvector of transcritical bifurcations will be relevant, not hyperrelevant like for all other forms of bifurcations considered here. It is tempting to speculate that the topological interpretation of this quirk in the algebra stems from the unique flow-field around transcritical bifurcations. For $r < 0$, the vector field has a negative-positive-negative pattern; for $r > 0$ this negative-positive-negative pattern is duplicated, just with an unstable equilibrium at $y = 0$ which had been stable before. Only at the critical value itself ($r = 0$) is there a topological inhomogeneity. The other bifurcations have fundamentally different flow-fields on either side of the critical value, and thus, perhaps, their bifurcation parameters acquire hyper-relevance rather than simply relevance. Further study is needed to prove this conjecture.

Because $\frac{\partial y}{\partial \alpha_1} \to \mathscr{O}(\log(t)-2)$ and $\frac{\partial y}{\partial \alpha_n} \to \mathscr{O}(t^{-1-n})$, simple multiplication shows that all the other entries in the FIM will be of lower order than the top-left.

A.3 FIM of Pitchfork Bifurcations

In the supercritical case, the normal form is

$$\frac{dy}{dt} = ry(t) - y(t)^3 + \alpha_1 y(t)^4 + \alpha_2 y(t)^5 + ... \tag{A.19}$$

and the subcritical case is the same except the sign on the cubic term changes. At the critical value of $\theta_i = 0$, the system reduces to:

$$\begin{aligned} \frac{dy}{dt} = -y^3 \rightarrow -\frac{dy}{y^3} = dt \rightarrow \frac{1}{2y^2}\Big|_{y_0}^{y(t)} = t\Big|_0^t \\ \frac{1}{y(t)^2} - \frac{1}{y_0^2} = 2t \rightarrow \frac{1}{y(t)^2} = 2t + \frac{1}{y_0^2} \\ \rightarrow y(t) = \frac{y_0}{\sqrt{1+2ty_0^2}} \end{aligned} \tag{A.20}$$

Following the same logic, the formula for the subcritical case is

$$y(t) = \frac{y_0}{\sqrt{1-2ty_0^2}} \tag{A.21}$$

Note that this creates a potentially-problematic singularity at $t = \frac{1}{2y_0^2}$.

A.3.1 Partial derivative of r

Let the α_i's=0. The derivative of the normal form w.r.t. r becomes:

$$\begin{aligned} \frac{\partial}{\partial r}\left(\frac{\partial y}{\partial t} = ry - y^3\right) \\ \frac{\partial^2 y}{\partial r \partial t} = r\cancel{\frac{\partial y}{\partial r}} + y - 3y^2\frac{\partial y}{\partial r} \\ \frac{\partial w}{\partial t} = y - 3y^2 w \end{aligned} \tag{A.22}$$

where $w = \frac{\partial y}{\partial r}$. Using integration factors $p_1 = 1, p_0 = 3y^2, q = y$, we see that

$$\begin{aligned}\mu &= \exp\left(\int -3y^2 dt\right) \\ &= \exp\left(-\int \frac{3y_0^2 dt}{1+2y_0^2 t}\right) \\ &= \exp\left(\frac{3}{2}\ln(1+2y_0^2 t)\right) \\ &= (1+2y_0^2 t)^{3/2}\end{aligned}$$

Therefore,

$$\begin{aligned}w &= \frac{C+\int \mu y(t) dt}{\mu} \\ &= \frac{\not{C}+\int \frac{y_0}{\sqrt{1+2ty_0^2}}(1+2y_0^2 t)^{3/2} dt}{(1+2y_0^2 t)^{3/2}} \\ \frac{\partial y}{\partial r} &= \frac{y_0 t(1+y_0^2 t)}{(1+2y_0^2 t)^{3/2}}\end{aligned} \tag{A.23}$$

Following the same logic for the subcritical case eventually brings us to

$$\frac{\partial y}{\partial r} = \frac{t y_0\left(1-t y_0^2\right)}{\left(1-2t y_0^2\right)^{3/2}} \tag{A.24}$$

A.3.2 Partial derivative of α's

When $r = 0$, and all $\alpha_{i\neq n} = 0$, then the normal form reduces to

$$\frac{dy}{dt} = -y(t)^3 + \alpha_n y(t)^{n+3} \tag{A.25}$$

which conveniently allows us to use the same μ integration factor as above. Using the integration scheme outlined there, after many steps we reach the conclusion that

$$\frac{\partial y}{\partial \alpha_n} = \frac{y_0^{n+1}}{2-n}\frac{(1+2ty_0^2)^{1-n/2}-1}{\mu} \tag{A.26}$$

This produces an obvious problem when $n = 2$, but in that case the integration step simplifies and we find that

$$\frac{\partial y}{\partial \alpha_2} = \frac{y_0^3 \ln(1 + 2t y_0^2)}{2\mu} \tag{A.27}$$

All this indicates that in the FIM, the entry corresponding to $(\partial y/\partial r)^2$ is $\mathcal{O}(t^1)$, while all other entries are lower order, so r will be the only hyperrelevant direction.

A.4 FIM of Hopf Bifurcations

Analysis of the Hopf bifurcation in either the complex or Cartesian formulation is complicated, because the introduction of nuisance parameters to the normal form equations tends to alter the period of limit cycles. This means standard trigonometric functions would also need to be altered with time-dependent terms to dilate/expand the period for a closed form solution of the trajectories $z(t)$ or $x(t), y(t)$ respectively.

However, reparameterizing the equation into polar coordinate form simplifies matters greatly. The system $\dot{r} = r(\mu - r^2);\ \dot{\theta} = -1$ should look familiar, as the equation for r is simply the normal form for a supercritical pitchfork bifurcation. Therefore, deriving the elements of its Fisher information matrix has already been performed in Appendix A.3, albeit with different variable and parameter names.

Appendix B

Derivation of log-transformed motif ODE

We begin by dividing Eq. 3.2 by the protein concentration x_i:

$$\begin{aligned}
\frac{dx_i}{dt}\frac{1}{x_i} &= \frac{\alpha}{x_i} - \delta + \frac{1}{x_i}\sum_{i\neq j} K_{i,j}x_j x_i^h \\
\frac{d}{dt}\log x_i &= \frac{\alpha}{x_i} - \delta + \sum_{i\neq j} K_{i,j}x_j x_i^{h-1} \\
\text{Let } X_i = \log x_i &\rightarrow x_i = e^{X_i} \\
\dot{X}_i &= \alpha e^{-X_i} - \delta + \sum_{i\neq j} K_{i,j} e^{X_j + X_i(h-1)} \\
&= -\delta + e^{-X_i}\left(\alpha + \sum_{i\neq j} K_{i,j} e^{X_j + hX_i}\right)
\end{aligned}$$

which is equivalent to Eq. 3.3 in the main text.

Appendix C

Additional bionetwork figures

Below, find more complete counts of the motifs in the mouse and human data. These parallel the figures presented in the main text.

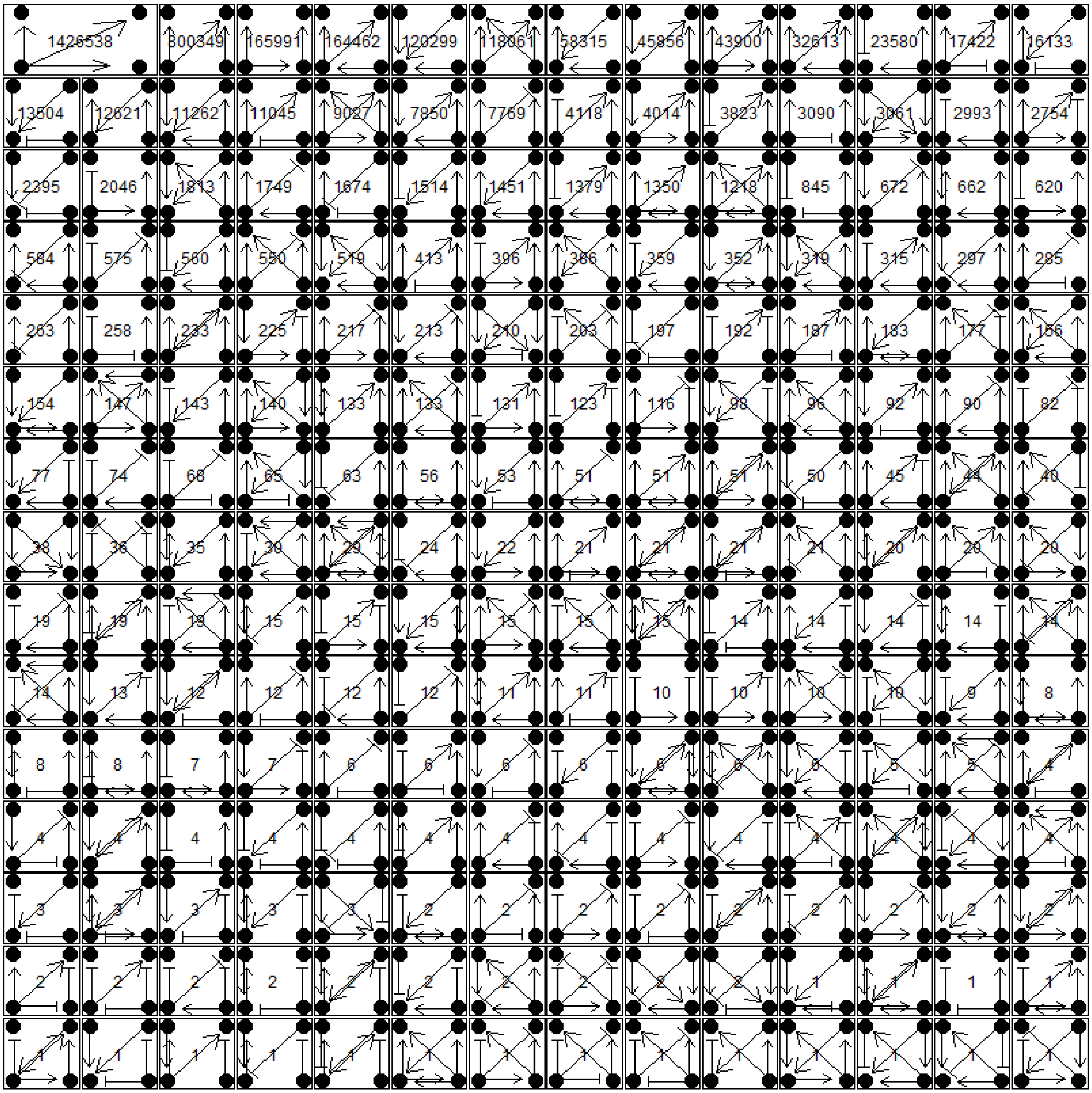

Figure C.1 Counts of all of the 4-motifs in humans.

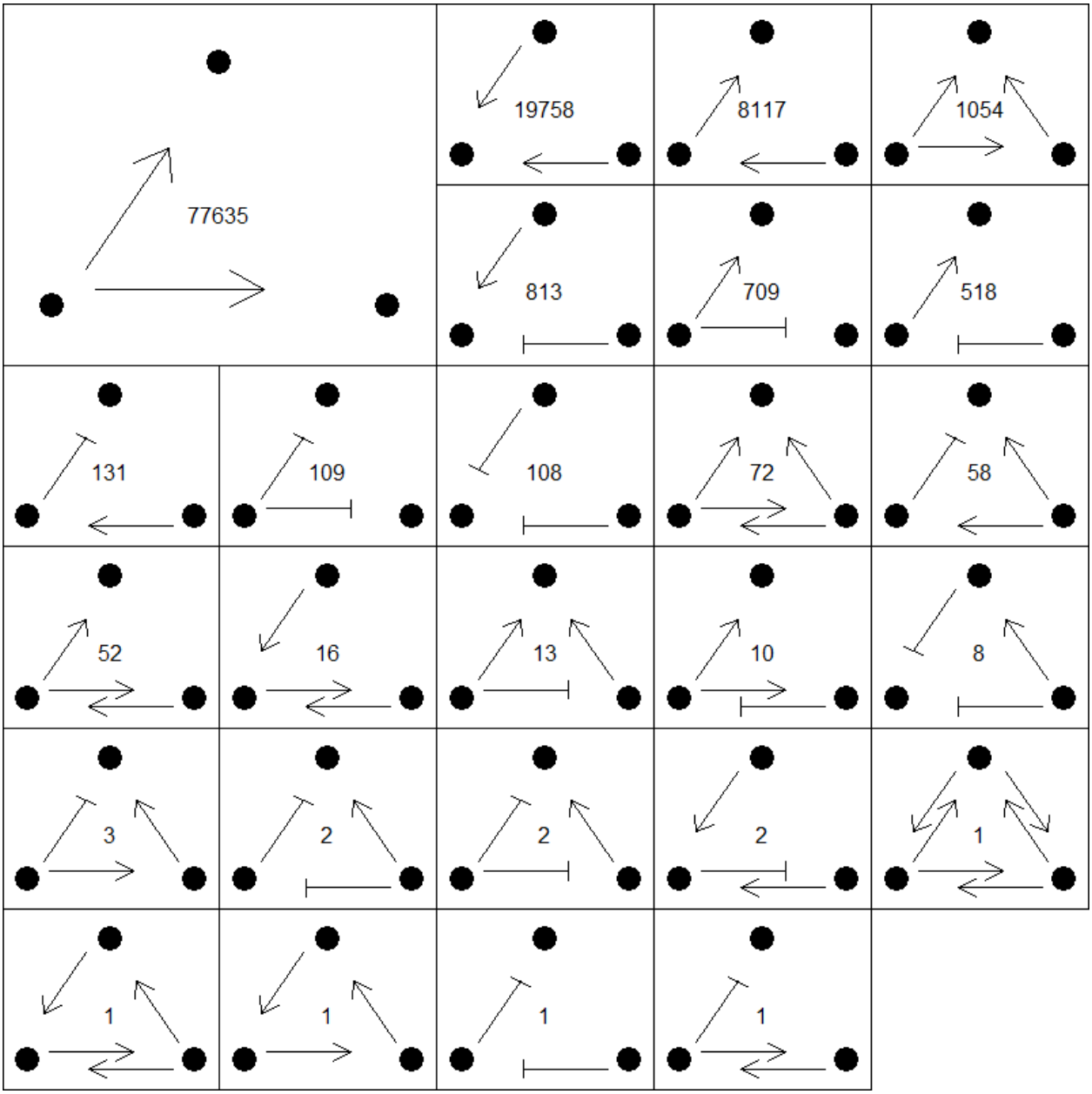

Figure C.2 Counts of 3-motifs in mice

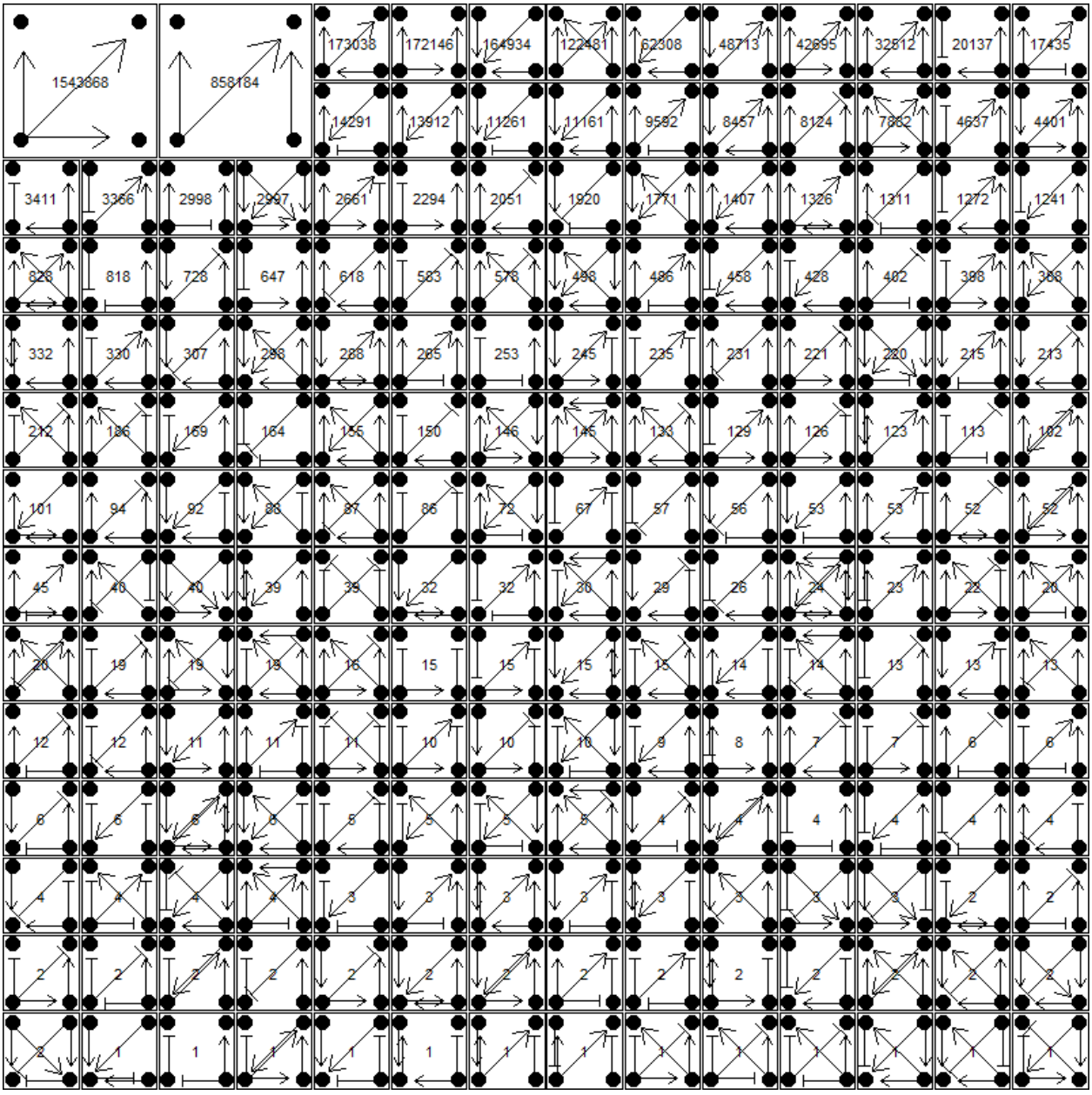

Figure C.3 Counts of all of the 4-motifs in mice

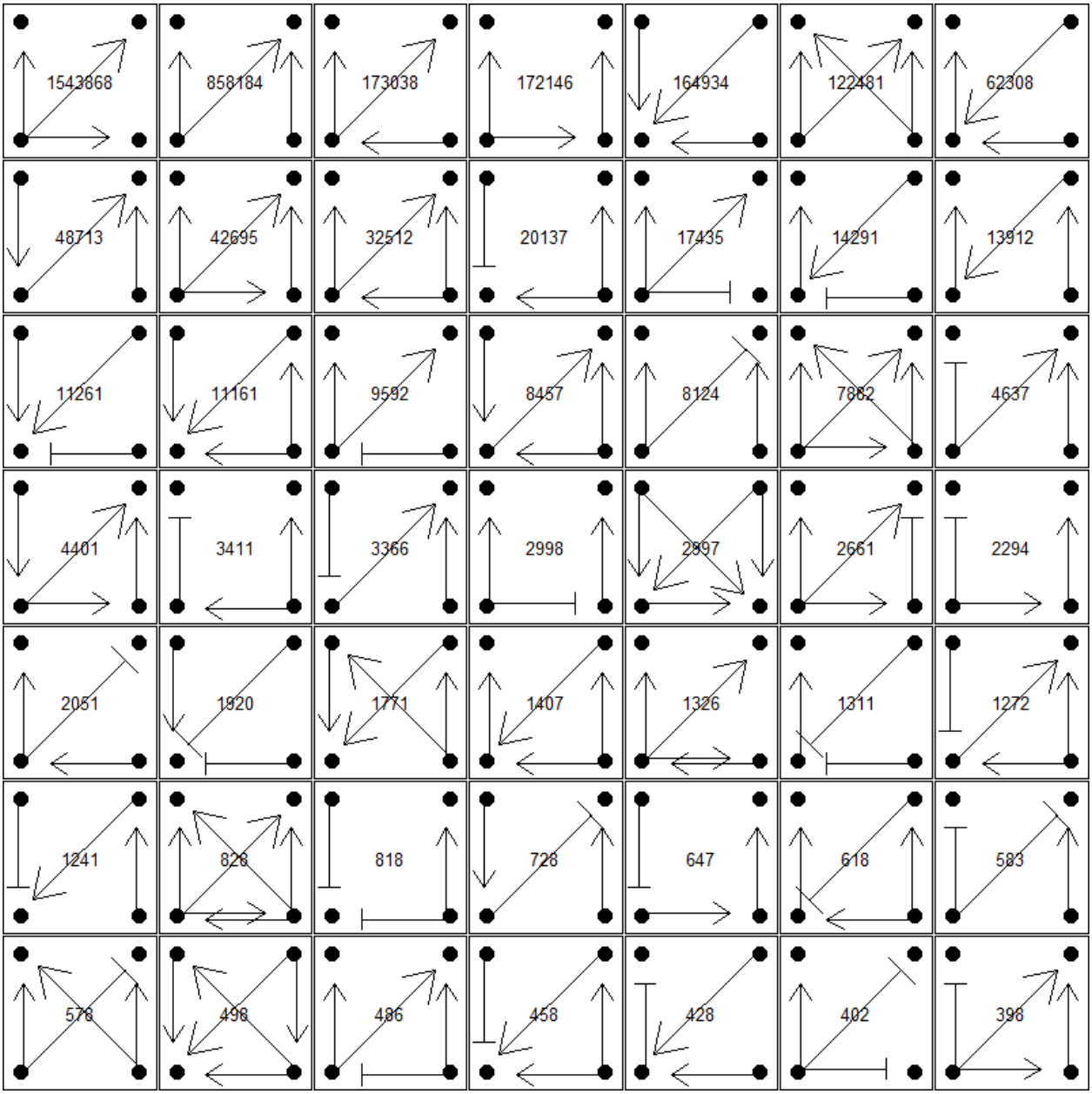

Figure C.4 Counts of the 4-motifs in mice that made up over 0.01% of the total motifs counted.

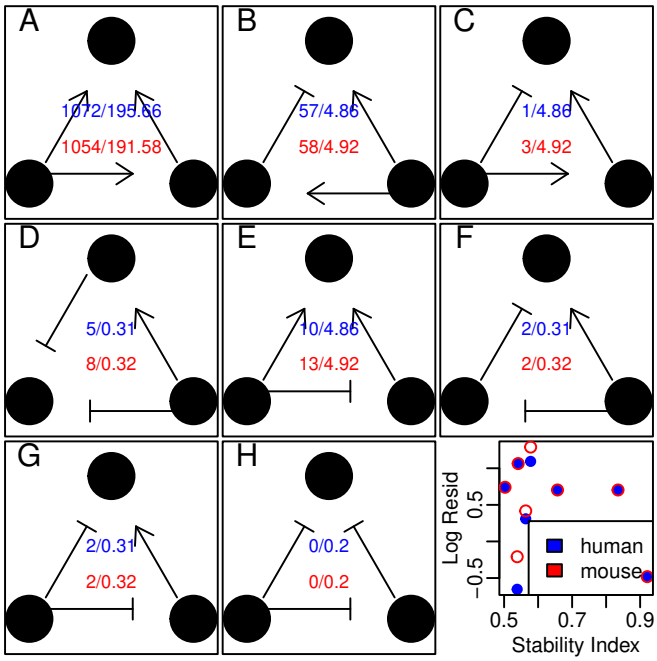

Figure C.5 On average, FFL motifs occur about 3x more often than expected (observed/expected, blue=human, red=mouse), with the largest enrichments occurring in FFL A and B as in other organisms. [3] As before, stability is a poor predictor of departure from expectations.

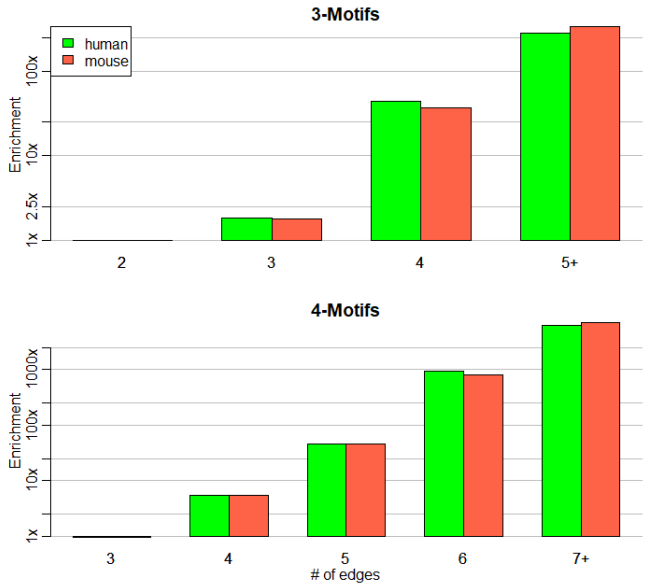

Figure C.6 Because connectivity is $< 1\%$, densely connected motifs are expected to be rare. While this was the case, motifs with relatively large numbers of edges for their size were far more common than predicted by the null model.

Index

Bibliography

[1] R. M. May, "Biological Populations with Nonoverlapping Generations: Stable Points, Stable Cycles, and Chaos," Science **186,** 645–647 (1974).

[2] E. E. Sel'kov, "Self–Oscillations in Glycolysis 1. A Simple Kinetic Model," European Journal of Biochemistry **4,** 79–86 (1968).

[3] S. Mangan and U. Alon, "Structure and function of the feed-forward loop network motif," Proceedings of the National Academy of Sciences of the United States of America **100,** 11980–11985 (2003).

[4] E. O. Voit, *Systems Biology: A Very Short Introduction* (Oxford University Press, 2020), especially chapter 3.

[5] K. S. Brown, C. C. Hill, G. A. Calero, C. R. Myers, K. H. Lee, J. P. Sethna, and R. A. Cerione, "The statistical mechanics of complex signaling networks: Nerve growth factor signaling," Physical Biology **1,** 184–195 (2004).

[6] M. K. Transtrum, B. B. Machta, and J. P. Sethna, "Geometry of nonlinear least squares with applications to sloppy models and optimization," Physical Review E **83,** 36701 (2011).

[7] A. White, M. Tolman, H. D. Thames, H. R. Withers, K. A. Mason, and M. K. Transtrum, "The Limitations of Model-Based Experimental Design and Parameter Estimation in Sloppy Systems," PLOS Computational Biology **12,** e1005227 (2016).

[8] M. K. Transtrum, B. B. Machta, K. S. Brown, B. C. Daniels, C. R. Myers, and J. P. Sethna, "Perspective: Sloppiness and emergent theories in physics, biology, and beyond," The Journal of Chemical Physics **143,** 010901 (2015).

[9] B. B. Machta, R. Chachra, M. Transtrum, and J. P. Sethna, "Parameter Space Compression Underlies Emergent Theories and Predictive Models," Science **342,** 604–606 (2013).

[10] J. J. Waterfall, F. P. Casey, R. N. Gutenkunst, K. S. Brown, C. R. Myers, P. W. Brouwer, V. Elser, and J. P. Sethna, "Sloppy-model Universality Class and the Vandermonde Matrix," Physical Review Letters 97 (2006).

[11] E. P. Wigner, "The unreasonable effectiveness of mathematics in the natural sciences," Communications on Pure and Applied Mathematics **13,** 1–14 (1960).

[12] G. C. Brown, *The Energy of Life: The Science of what Makes Our Minds and Bodies Work* (Free Press, 2000), see discussion of Greek ideas of homeostasis on p.197.

[13] D. Schäfer, "Aging, Longevity, and Diet: Historical Remarks on Calorie Intake Reduction," Gerontology **51,** 126–130 (2005).

[14] F. Magendie and E. (trans) Milligan, *An Elementary Compendium of Physiology for the Use of Students*, 3 ed. (Longmans Green, 1829), quotes from p. 18 and 468.

[15] J. C. Waterlow, *Protein Turnover* (CAB International, 2006), pp. 1–301.

[16] E. L. Huttlin *et al.*, "Dual proteome-scale networks reveal cell-specific remodeling of the human interactome," Cell **184,** 3022–3040.e28 (2021).

[17] Z.-P. Liu, C. Wu, H. Miao, and H. Wu, "RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse," Database **2015,** bav095 (2015).

[18] M. Mann, "Fifteen years of stable isotope labeling by amino acids in cell culture (SILAC)," Methods in Molecular Biology **1188,** 1–7 (2014).

[19] S. P. Gygi, B. Rist, S. A. Gerber, F. Turecek, M. H. Gelb, and R. Aebersold, "Quantitative analysis of complex protein mixtures using isotope-coded affinity tags," Nature Biotechnology **17,** 994–999 (1999).

[20] S.-E. Ong, B. Blagoev, I. Kratchmarova, D. B. Kristensen, H. Steen, A. Pandey, and M. Mann, "Stable Isotope Labeling by Amino Acids in Cell Culture, SILAC, as a Simple and Accurate Approach to Expression Proteomics," Molecular & Cellular Proteomics **1,** 376–386 (2002).

[21] L. M. F. de Godoy, J. V. Olsen, J. Cox, M. L. Nielsen, N. C. Hubner, F. Fröhlich, T. C. Walther, and M. Mann, "Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast," Nature **455,** 1251–1254 (2008).

[22] J. C. Price, S. Guan, A. Burlingame, S. B. Prusiner, and S. Ghaemmaghami, "Analysis of proteome dynamics in the mouse brain," Proceedings of the National Academy of Sciences of the United States of America **107,** 14508–14513 (2010).

[23] S. Guan, J. C. Price, S. B. Prusiner, S. Ghaemmaghami, and A. L. Burlingame, "A Data Processing Pipeline for Mammalian Proteome Dynamics Studies Using Stable Isotope Metabolic Labeling," Molecular & Cellular Proteomics **10,** M111.010728 (2011).

[24] B. Mohar, J. B. Grimm, R. Patel, T. A. Brown, P. W. Tillberg, L. D. Lavis, N. Spruston, and K. Svoboda, "Brain-wide measurement of protein turnover with high spatial and temporal resolution," BioRxiv (2023).

[25] A. Raju, B. B. Machta, and J. P. Sethna, "Information loss under coarse graining: A geometric approach," Physical Review E **98,** 052112 (2018).

[26] B. Hu, "Introduction to real-space renormalization-group methods in critical and chaotic phenomena," Physics Reports **91,** 233–295 (1982).

[27] J. Hu, Ph.D. thesis, City University of New York, New York, 1995.

[28] R. E. DeVille, A. Harkin, M. Holzer, K. Josić, and T. J. Kaper, "Analysis of a renormalization group method and normal form theory for perturbed ordinary differential equations," Physica D: Nonlinear Phenomena **237,** 1029–1052 (2008).

[29] P. Hartman, *Ordinary Differential Equations* (SIAM, Philadelphia, 1982), p. 632.

[30] J. D. Crawford, "Introduction to bifurcation theory," Reviews of Modern Physics **63,** 991–1037 (1991), quote from p. 1021.

[31] I. M. Sobol', "Sensitivity Estimates for Nonlinear Mathematical Models," Mathematical Modeling in Civil Engineering **1,** 407–414 (1993).

[32] R. Gul and S. Bernhard, "Parametric uncertainty and global sensitivity analysis in a model of the carotid bifurcation: Identification and ranking of most sensitive model parameters," Mathematical Biosciences **269,** 104–116 (2015).

[33] N. Alon, R. Yuster, and U. Zwick, "Finding and counting given length cycles," Algorithmica 1997 17:3 **17,** 209–223 (1997).

[34] S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon, "Network motifs in the transcriptional regulation network of Escherichia coli," Nature Genetics 2002 31:1 **31,** 64–68 (2002).

[35] S. Mangan, A. Zaslaver, and U. Alon, "The Coherent Feedforward Loop Serves as a Sign-sensitive Delay Element in Transcription Networks," Journal of Molecular Biology **334,** 197–204 (2003).

[36] S. Mangan, S. Itzkovitz, A. Zaslaver, and U. Alon, "The Incoherent Feed-forward Loop Accelerates the Response-time of the gal System of Escherichia coli," Journal of Molecular Biology **356,** 1073–1081 (2006).

[37] U. Alon, "Simplicity in biology," Nature **446,** 497 (2007).

[38] U. Alon, *An Introduction to Systems Biology: Design Principles of Biological Circuits* (Chapman & Hall, 2007), all directed unsigned graphs appear in Figs 4.1 (3-motifs) and 5.5 (4-motifs).

[39] L. Stone, D. Simberloff, and Y. Artzy-Randrup, "Network motifs and their origins," PLOS Computational Biology **15,** e1006749 (2019).

[40] L. Bunimovich and B. Webb, in *Isospectral Transformations: A New Approach to Analyzing Multidimensional Systems and Networks* (Springer, 2014), Chap. 3, pp. 53–89.

[41] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, "Network motifs: Simple building blocks of complex networks," Science **298,** 824–827 (2002).

[42] R. Itzhack, Y. Mogilevski, and Y. Louzoun, "An optimal algorithm for counting network motifs," Physica A: Statistical Mechanics and its Applications **381,** 482–490 (2007).

[43] T. Fushimi, K. Saito, and H. Motoda, "Efficient analytical computation of expected frequency of motifs of small size by marginalization in uncertain network," In , Proceedings of the

2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining pp. 1–8 (ACM, 2021).

[44] C. López-Otín, M. A. Blasco, L. Partridge, M. Serrano, and G. Kroemer, "The Hallmarks of Aging," Cell **153,** 1194–1217 (2013).

[45] R. Weindruch, R. L. Walford, S. Fligiel, and D. Guthrie, "The Retardation of Aging in Mice by Dietary Restriction: Longevity, Cancer, Immunity and Lifetime Energy Intake," The Journal of Nutrition **116,** 641–654 (1986).

[46] J. L. Barger, R. L. Walford, and R. Weindruch, "The retardation of aging by caloric restriction: its significance in the transgenic era," Experimental Gerontology **38,** 1343–1351 (2003).

[47] B. Agarwal and J. A. Baur, "Resveratrol and life extension," Annals of the New York Academy of Sciences **1215,** 138–143 (2011).

[48] R. H. Carson, Ph.D. thesis, Brigham Young University, 2019.

[49] E. Brown, J. Gao, P. Holmes, R. Bogacz, M. Gilzenrat, and J. D. Cohen, "Simple Neural Networks that Optomize Decisions," International Journal of Bifurcation and Chaos **15,** 803–826 (2005).

[50] S. Feng, P. Holmes, A. Rorie, and W. T. Newsome, "Can Monkeys Choose Optimally When Faced with Noisy Stimuli and Unequal Rewards?," PLOS Computational Biology **5,** e1000284 (2009).

[51] B. Bettonvil and J. P. Kleijnen, "Searching for important factors in simulation models with many factors: Sequential bifurcation," European Journal of Operational Research **96,** 180–194 (1997).

[52] T. Homma and A. Saltelli, "Importance measures in global sensitivity analysis of nonlinear models," Reliability Engineering and System Safety **52,** 1–17 (1996).

[53] A. Berezhkovskii and A. Szabo, "One-dimensional reaction coordinates for diffusive activated rate processes in many dimensions," Journal of Chemical Physics 122 (2005).

[54] M. J. Feigenbaum, "Quantitative Universality for a Class of Nonlinear Transformations," Journal of Statistical Physics **19,** 25–52 (1978).

[55] M. J. Feigenbaum, "The universal metric properties of nonlinear transformations," Journal of Statistical Physics **21,** 669–706 (1979).

[56] M. Widom and L. P. Kadanoff, "Renormalization group analysis of bifurcations in area-preserving maps," Physica D: Nonlinear Phenomena **5,** 287–292 (1982).

[57] B. Hu and J. Rudnick, "Exact solutions to the feigenbaum renormalization-group equations for intermittency," Physical Review Letters **48,** 1645–1648 (1982).

[58] D. Hathcock and J. P. Sethna, "Reaction rates and the noisy saddle-node bifurcation: Renormalization group for barrier crossing," Physical Review Research 3 (2021).

[59] L.-N. Chang and N.-P. Chang, "Bifurcation and Dynamical Symmetry Breaking in a Renormalization-Group-Improved Field Theory," Physical Review Letters **54,** 2407 (1985).

[60] A. Raju, C. B. Clement, L. X. Hayden, J. P. Kent-Dobias, D. B. Liarte, D. Z. Rocklin, and J. P. Sethna, "Normal Form for Renormalization Groups," Physical Review X **9,** 021014 (2019).

[61] M. K. Transtrum, B. B. Machta, and J. P. Sethna, "Why are nonlinear fits to data so challenging?," Physical Review Letters **104,** 2–5 (2010).

[62] M. K. Transtrum and P. Qiu, "Bridging Mechanistic and Phenomenological Models of Complex Biological Systems," PLOS Computational Biology **12,** e1004915 (2016).

[63] K. N. Quinn, M. C. Abbott, M. K. Transtrum, B. B. Machta, and J. P. Sethna, "Information geometry for multiparameter models: new perspectives on the origin of simplicity," Reports on Progress in Physics **86,** 035901 (2022).

[64] M. K. Transtrum and P. Qiu, "Model Reduction by Manifold Boundaries," Physical Review Letters **113,** 098701 (2014).

[65] H. H. Mattingly, M. K. Transtrum, M. C. Abbott, and B. B. Machta, "Maximizing the information learned from finite data selects a simple model," Proceedings of the National Academy of Sciences **115,** 1760–1765 (2018).

[66] J. E. Jeong, Q. Zhuang, M. K. Transtrum, E. Zhou, and P. Qiu, "Experimental design and model reduction in systems biology," Quantitative Biology **6,** 287–306 (2018).

[67] E. Roesch and M. P. Stumpf, "Parameter inference in dynamical systems with co-dimension 1 bifurcations," Royal Society Open Science 6 (2019).

[68] S. H. Strogatz, *Nonlinear dynamics and chaos: With applications to physics, biology, chemistry, and engineering*, 2ed ed. (Westview Press, 2015), p. 513.

[69] O. Jiménez–Ramírez, E. J. Cruz–Domínguez, M. A. Quiroz–Juárez, J. L. Aragón, and R. Vázquez–Medina, "Experimental detection of Hopf bifurcation in two-dimensional dynamical systems," Chaos, Solitons & Fractals: X **6,** 100058 (2021).

[70] J. P. C. Kleijnen, B. Bettonvil, and F. Persson, "Screening for the Important Factors in Large Discrete-Event Simulation Models: Sequential Bifurcation and Its Applications," Screening: Methods for Experimentation in Industry, Drug Discovery, and Genetics pp. 287–307 (2006).

[71] S. Waldherr and F. Allgower, "A feedback approach to bifurcation analysis in biochemical networks with many parameters," Proceedings of the FOSBE 2007 pp. 479–484 (2007).

[72] A. F. Brouwer and M. C. Eisenberg, available on arXiv.org (unpublished).

[73] X. Song, B. A. Bryan, A. C. Almeida, K. I. Paul, G. Zhao, and Y. Ren, "Time-dependent sensitivity of a process-based ecological model," Ecological Modelling **265,** 114–123 (2013).

[74] A. Alexanderian, P. A. Gremaud, and R. C. Smith, "Variance-based sensitivity analysis for time-dependent processes," Reliability Engineering & System Safety **196,** 106722 (2020).

[75] T. Sumner, E. Shephard, and I. D. Bogle, "A methodology for global-sensitivity analysis of time-dependent outputs in systems biology modelling," Journal of The Royal Society Interface **9,** 2156–2166 (2012).

[76] I. Bendixson, "Sur les courbes définies par des équations différentielles," Acta Mathematica **24,** 1–88 (1901).

[77] B. L. Francis and M. K. Transtrum, "Unwinding the model manifold: Choosing similarity measures to remove local minima in sloppy dynamical systems," Physical Review E **100,** 012206 (2019).

[78] I. J. Pérez-Arriaga, G. C. Verghese, and F. C. Schweppe, "Selective modal analysis with applications to electric power systems, Part I: Heuristic introduction," IEEE Transactions on Power Apparatus and Systems **PAS-101,** 3117–3125 (1982).

[79] F. Garofalo, L. Iannelli, and F. Vasca, "Participation Factors and their Connections to Residues and Relative Gain Array," IFAC Proceedings Volumes **35,** 125–130 (2002).

[80] S. H. Strogatz, *Nonlinear dynamics and chaos: With applications to physics, biology, chemistry, and engineering*, 2ed ed. (Westview Press, Boulder, CO, 2015), p. 513, the system is described in section 3.5, but has the same form as an overdamped pendulum in 4.4.

[81] A. N. Tikhonov, "On the dependence of the solutions of differential equations on a small parameter," Matematicheskii Sbornik. Novaya Seriya **22,** 193–204 (1948).

[82] S. H. Strogatz, *Nonlinear dynamics and chaos: With applications to physics, biology, chemistry, and engineering*, 2ed ed. (Westview Press, Boulder, CO, 2015), p. 513, figure 7.3.7.

[83] B. Chance, G. Williamson, I. Lee, L. Mela, D. DeVault, A. Ghosh, and E. Pye, "Synchronization Phenomena in Oscillations of Yeast Cells and Isolated Mitochondria," in *Biological and Biochemical Oscillators* (Academic Press, 1973), pp. 285–300.

[84] J. Briggs and F. D. Peat, *Turbulent Mirror* (Harper and Row, 1989), p. 222.

[85] P. Holmes, "Poincaré, celestial mechanics, dynamical-systems theory and "chaos"," Physics Reports **193,** 137–163 (1990).

[86] G. Sugihara and R. M. May, "Nonlinear forcasting as a way of distinguishing chaos from measurement error in time series," Nature **344,** 734–741 (1990).

[87] C. hao Hsieh, C. Anderson, and G. Sugihara, "Extending nonlinear analysis to short ecological time series.," The American Naturalist **171,** 71–80 (2008).

[88] S. H. Strogatz, *Nonlinear dynamics and chaos: With applications to physics, biology, chemistry, and engineering*, 2ed ed. (Westview Press, Boulder, CO, 2015), p. 513, especially Ch 10.6.

[89] D. L. Hitzl and F. Zele, "An exploration of the Hénon quadratic map," Physica D: Nonlinear Phenomena **14,** 305–326 (1985).

[90] O. Rössler, "An equation for continuous chaos," Physics Letters A **57,** 397–398 (1976).

[91] S. Bleher, E. Ott, and C. Grebogi, "Routes to chaotic scattering," Physical Review Letters **63,** 919–922 (1989).

[92] Y.-C. Lai, "Abrupt bifurcation to chaotic scattering with discontinuous change in fractal dimension," Physical Review E **60,** R6283–R6286 (1999).

[93] R. Chachra, M. K. Transtrum, and J. P. Sethna, "Structural susceptibility and separation of time scales in the van der Pol oscillator," Physical Review E **86,** 026712 (2012).

[94] S. N. Rasband, *Chaotic Dynamics of Nonlinear Systems*, 2ed ed. (Wiley, New York, NY, 1990), p. 230.

[95] D. A. Rand, A. Raju, M. Sáez, F. Corson, and E. D. Siggia, "Geometry of gene regulatory dynamics," Proceedings of the National Academy of Sciences 118 (2021).

[96] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations Dynamical Systems, and Bifurcations of Vector Fields* (Springer New York, 1983), p. 497.

[97] P. D. Kirk, T. Toni, and M. P. Stumpf, "Parameter inference for biochemical systems that undergo a Hopf bifurcation," Biophysical Journal **95,** 540–549 (2008).

[98] S. M. McMahon, "Networking Tips for Social Scientists and Ecologists," Science **293,** 1604–1605 (2001).

[99] G. Sugihara and H. Ye, "Complex systems: Cooperative network dynamics," Nature **458,** 979–980 (2009).

[100] N. Alon, P. Dao, I. Hajirasouliha, F. Hormozdiari, and S. C. Sahinalp, "Biomolecular network motif counting and discovery by color coding," Bioinformatics **24,** i241–i249 (2008).

[101] J. M. Diamond, "Assemby of Species Communities," in *Ecology and evolution of communities*, M. L. Cody and J. M. Diamond, eds., (Harvard Unviersity Press, Cambridge, MA, 1975), Chap. 14, pp. 342–444.

[102] E. F. Connor and D. Simberloff, "The Assembly of Species Communities: Chance or Competition?," Ecology **60,** 1132–1140 (1979).

[103] J. M. Diamond and M. E. Gilpin, "Examination of the "null" model of connor and simberloff for species co-occurrences on Islands," Oecologia 1982 52:1 **52,** 64–74 (1982).

[104] R. M. May, "Network structure and the biology of populations," Trends in Ecology & Evolution **21(7),** 394–399 (2006).

[105] G. Box, "Robustness in the Strategy of Scientific Model Building," in *Robustness in Statistics* (Elsevier, 1979), pp. 201–236.

[106] G. E. P. Box, "Science and Statistics," Journal of the American Statistical Association **71,** 791–799 (1976), though not stated, the principle is strongly implied by this earlier work.

[107] N. Goldenfeld and L. P. Kadanoff, "Simple lessons from complexity," Science (1999).

[108] R. B. Laughlin and D. Pines, "The Theory of Everything," Proceedings of the National Academy of Sciences **97,** 28–31 (2000).

[109] P. Truran, "Models: Useful but Not True," in *Practical Applications of the Philosophy of Science* (Springer, New York, 2013), pp. 61–67.

[110] S. Jones and J. M. Thornton, "Principles of Protein-Protein Interactions," Proc Natl Acad Sci USA **93,** 13–20 (1996).

[111] I. M. A. Nooren and J. M. Thornton, "Diversity of protein-protein interactions," The EMBO Journal **22,** 3486–3492 (2003).

[112] J. McFadden and J. Al-Khalili, *Life on the Edge* (Crown, 2016), p. 368, see especially chapters 5 and 10.

[113] K. Balasubramanian and S. P. Gupta, "Quantum Molecular Dynamics, Topological, Group Theoretical and Graph Theoretical Studies of Protein-Protein Interactions," Current Topics in Medicinal Chemistry **19,** 426–443 (2019).

[114] L. Pion-Tonachini *et al.*, "Learning from learning machines: a new generation of AI technology to meet the needs of science," arXiv (2021).

[115] D. P. Minde, A. K. Dunker, and K. S. Lilley, "Time, space, and disorder in the expanding proteome universe," Proteomics **17,** 1600399 (2017).

[116] C. J. Bley *et al.*, "Architecture of the cytoplasmic face of the nuclear pore," Science 376 (2022).

[117] S. Basu, K. Kumbier, J. B. Brown, and B. Yu, "Iterative random forests to discover predictive and stable high-order interactions," Proceedings of the National Academy of Sciences **115,** 1943–1948 (2018).

[118] M. H. Schweitzer, Z. Suo, R. Avci, J. M. Asara, M. A. Allen, F. T. Arce, and J. R. Horner, "Analyses of Soft Tissue from Tyrannosaurus rex Suggest the Presence of Protein," Science **316,** 277–280 (2007).

[119] J. M. Asara, M. H. Schweitzer, L. M. Freimark, M. Phillips, and L. C. Cantley, "Protein Sequences from Mastodon and Tyrannosaurus Rex Revealed by Mass Spectrometry," Science **316,** 280–285 (2007).

[120] J. C. Taggart and G. W. Li, "Production of Protein-Complex Components Is Stoichiometric and Lacks General Feedback Regulation in Eukaryotes," Cell Systems **7,** 580–589.e4 (2018).

[121] G. Boël *et al.*, "Codon influence on protein expression in E. coli correlates with mRNA levels," Nature **529,** 358–363 (2016).

[122] B. Schwanhäusser, D. Busse, N. Li, G. Dittmar, J. Schuchhardt, J. Wolf, W. Chen, and M. Selbach, "Global quantification of mammalian gene expression control," Nature **473,** 337–342 (2011).

[123] O. Borkowski, A. Goelzer, M. Schaffer, M. Calabre, U. Mäder, S. Aymerich, M. Jules, and V. Fromion, "Translation elicits a growth rate-dependent, genome-wide, differential protein production in Bacillus subtilis," Molecular Systems Biology **12,** 870 (2016).

[124] A. K. Sharma, P. Sormanni, N. Ahmed, P. Ciryam, U. A. Friedrich, G. Kramer, and E. P. O'Brien, "A chemical kinetic basis for measuring translation initiation and elongation rates from ribosome profiling data," PLoS computational biology **15,** e1007070 (2019).

[125] R. Schleif, "Control of production of ribosomal protein," Journal of Molecular Biology **27,** 41–55 (1967).

[126] T. V. Pestova, V. G. Kolupaeva, I. B. Lomakin, E. V. Pilipenko, I. N. Shatsky, V. I. Agol, and C. U. T. Hellen, "Molecular mechanisms of translation initiation in eukaryotes," Proceedings of the National Academy of Sciences of the United States of America **98,** 7029–7036 (2001).

[127] A. D. Mathis *et al.*, "Mechanisms of *In Vivo* Ribosome Maintenance Change in Response to Nutrient Signals," Molecular & Cellular Proteomics **16,** 243–254 (2017).

[128] S. Reuveni, I. Meilijson, M. Kupiec, E. Ruppin, and T. Tuller, "Genome-scale analysis of translation elongation with a ribosome flow model," PLoS Computational Biology 7 (2011).

[129] I. Nanikashvili, Y. Zarai, A. Ovseevich, T. Tuller, and M. Margaliot, "Networks of ribosome flow models for modeling and analyzing intracellular traffic," Scientific Reports **9,** 1–14 (2019).

[130] A. B. Ross, J. D. Langer, and M. Jovanovic, "Proteome turnover in the spotlight: Approaches, applications, and perspectives," Molecular and Cellular Proteomics **20,** 100016 (2021).

[131] B. C. Naylor *et al.*, "DeuteRater-H: Using nonequilibrium isotope enrichments to optimize breadth of turnover measurements in single biopsy samples from humans," Bioinformatics (in press).

[132] J. I. DiStefano, *Dynamic Systems Biology Modeling and Simulation*, 2ed ed. (Elsevier, Oxford, UK, 2014), p. 859, see the discussion of the history of nonlinear modeling on p.416-434.

[133] A. R. Ives and S. R. Carpenter, "Stability and Diversity of Ecosystems," Science **317,** 58–62 (2007).

[134] M. Xiao and J. Cao, "Genetic oscillation deduced from Hopf bifurcation in a genetic regulatory network with delays," Mathematical Biosciences **215,** 55–63 (2008).

[135] M. Chaplain, M. Ptashnyk, and M. Sturrock, "Hopf bifurcation in a gene regulatory network model: Molecular movement causes oscillations," Mathematical Models and Methods in Applied Sciences **25,** 1179–1215 (2015).

[136] P. E. Rapp, "Why are so many biological systems periodic?," Progress in Neurobiology **29,** 261–273 (1987).

[137] J. J. Fox and C. C. Hill, "From topology to dynamics in biochemical networks," Chaos: An Interdisciplinary Journal of Nonlinear Science **11,** 809 (2001).

[138] J. T. F. Wong, "On the Steady-State Method of Enzyme Kinetics," J. Am. Chem. Soc. **87,** 1788–1793 (1965).

[139] L. A. Segel, "On the validity of the steady state assumption of enzyme kinetics," Bulletin of Mathematical Biology 1988 50:6 **50,** 579–593 (1988).

[140] W. Bechtel, "Mechanism and Biological Explanation," Philosophy of Science **78,** 533–557 (2011).

[141] B. Schwanhäusser, D. Busse, N. Li, G. Dittmar, J. Schuchhardt, J. Wolf, W. Chen, and M. Selbach, "Corrigendum: Global quantification of mammalian gene expression control," Nature **495,** 126–127 (2013).

[142] F. M. Boisvert, Y. Ahmad, M. Gierliński, F. Charrière, D. Lamont, M. Scott, G. Barton, and A. I. Lamond, "A quantitative spatial proteomics analysis of proteome turnover in human cells," Molecular and Cellular Proteomics **11,** M111.011429 (2012).

[143] W. Ma, A. Trusina, H. El-Samad, W. A. Lim, and C. Tang, "Defining Network Topologies that Can Achieve Biochemical Adaptation," Cell **138,** 760–773 (2009).

[144] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah, "Julia: A fresh approach to numerical computing," SIAM review **59,** 65–98 (2017).

[145] D. Reber and B. Webb, "Intrinsic stability: stability of dynamical networks and switched systems with any type of time-delays," Nonlinearity **33,** 2660 (2020).

[146] K. Ciesielski, "The Poincaré-Bendixson Theorem: From Poincaré to the XXIst century," Central European Journal of Mathematics **10,** 2110–2128 (2012).

[147] N. MacDonald, "Birfurcation theory applied to a simple model of a biochemical oscillator," Journal of Theoretical Biology **65,** 727–734 (1977).

[148] J. Mallet-Paret and H. L. Smith, "The Poincare-Bendixson theorem for monotone cyclic feedback systems," Journal of Dynamics and Differential Equations **2,** 367–421 (1990).

[149] M. B. Elowitz and S. Leibier, "A synthetic oscillatory network of transcriptional regulators," Nature **403,** 335–338 (2000).

[150] S. Müller, J. Hofbauer, L. Endler, C. Flamm, S. Widder, and P. Schuster, "A generalized model of the repressilator," Journal of Mathematical Biology **53,** 905–937 (2006).

[151] G. Zames, "On the input-output stability of time-varying nonlinear feedback systems. Part One: Conditions derived using concepts of loop gain, conicity, and positivity," IEEE Transactions on Automatic Control **11,** 228–238 (1966).

[152] J. C. Price *et al.*, " The Effect of Long Term Calorie Restriction on in Vivo Hepatic Proteostatis: A Novel Combination of Dynamic and Quantitative Proteomics ," Molecular & Cellular Proteomics **11,** 1801–1814 (2012).

[153] M. Visscher *et al.*, "Proteome-wide Changes in Protein Turnover Rates in C. elegans Models of Longevity and Age-Related Disease," Cell Reports (2016).

[154] K. Swovick, K. A. Welle, J. R. Hryhorenko, A. Seluanov, V. Gorbunova, and S. Ghaemmaghami, "Cross-species Comparison of Proteome Turnover Kinetics.," Molecular & cellular proteomics : MCP **17,** 580–591 (2018).

[155] K. C. H. Fearon, D. T. Hansell, T. Preston, J. A. Plumb, J. Davies, D. Shapiro, A. Shenkin, K. C. Calman, and H. J. G. Burns, "Influence of whole body protein turnover rate on resting energy expenditure in patients with cancer," Cancer Research **48,** 2590–2595 (1988).

[156] S. D. Mitchell, "Dimensions of Scientific Law," Philosophy of Science **67,** 242–265 (2000).

[157] J. Woodward, "Law and Explanation in Biology: Invariance is the Kind of Stability That Matters," Philosophy of Science **68,** 1–20 (2001).

[158] J. Woodward, "Causation in biology: stability, specificity, and the choice of levels of explanation," Biology & Philosophy 2010 25:3 **25,** 287–318 (2010).

[159] F. C. Boogerd, F. J. Bruggeman, J. H. S. Hofmeyr, and H. V. Westerhoff, "Towards philosophical foundations of Systems Biology: Introduction," in *Towards philosophical foundations of Systems Biology: Introduction*, F. C. Boogerd, F. J. Bruggeman, J.-H. S. Hofmeyr, and H. V. Westerhoff, eds., (Elsevier Science, 2007).

[160] M. P. H. Stumpf, W. P. Kelly, T. Thorn, and C. Wiuf, "Evolution at the system level: the natural history of protein interaction networks," Trends in Ecology & Evolution **22,** 366–373 (2007).

[161] C. J. Cain *et al.*, "What Systems Biology Is (Not, Yet)," Science **320,** 1013–1014 (2008).

[162] A. L. Barabási, "Scale-free networks: A decade and beyond," Science **325,** 412–413 (2009).

[163] N. Barkai and S. Leibler, "Robustness in simple biochemical networks," Nature **387,** 913–917 (1997).

[164] A. Eldar, R. Dorfman, D. Weiss, H. Ashe, B.-Z. Shilo, and N. Barkai, "Robustness of the BMP morphogen gradient in Drosophila embryonic patterning," Nature **419,** 304–308 (2002).

[165] A. Eldar, D. Rosin, B.-Z. Shilo, and N. Barkai, "Self-Enhanced Ligand Degradation Underlies Robustness of Morphogen Gradients," Developmental Cell **5,** 635–646 (2003).

[166] A. Eldar, B.-Z. Shilo, and N. Barkai, "Elucidating mechanisms underlying robustness of morphogen gradients," Current Opinion in Genetics & Development **14,** 435–439 (2004).

[167] R. U. Ibarra, J. S. Edwards, and B. O. Palsson, "Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth," Nature 420 (2002).

[168] E. Dekel, S. Mangan, and U. Alon, "Environmental selection of the feed-forward loop circuit in gene-regulation networks," Physical Biology 2 (2005).

[169] J. J. Hopfield, "Origin of the genetic code: a testable hypothesis based on tRNA structure, sequence, and kinetic proofreading," Proceedings of the National Academy of Sciences of the United States of America 75 (1978).

[170] M. A. Savageau, "Accuracy of proofreading with zero energy cost," Journal of Theoretical Biology 93 (1981).

[171] T. M. Yi, Y. Huang, M. I. Simon, and J. Doyle, "Robust perfect adaptation in bacterial chemotaxis through integral feedback control," Proceedings of the National Academy of Sciences of the United States of America 97 (2000).

[172] H. Kitano, "Systems biology: A brief overview," Science **295,** 1662–1664 (2002).

[173] E. D. Sontag, "Molecular Systems Biology and Control," European Journal of Control **11,** 396–435 (2005).

[174] D. Bray, "Protein molecules as computational elements in living cells.," Nature **376,** 307–312 (1995).

[175] S. Bandyopadhyay *et al.*, "Rewiring of genetic networks in response to DNA damage," Science **330,** 1385–1389 (2010).

[176] A. Califano, "Rewiring makes the difference," Molecular Systems Biology **7,** 463 (2011).

[177] N. M. Luscombe, M. M. Babu, H. Yu, M. Snyder, S. A. Teichmann, and M. Gerstein, "Genomic analysis of regulatory network dynamics reveals large topological changes," Nature 2004 431:7006 **431,** 308–312 (2004).

[178] C. H. Waddington, "Canalization of development and the inheritance of acquired characters," Nature **150,** 563–565 (1942).

[179] J. Hermisson and G. P. Wagner, "Evolution of phenotypic robustness," Robust design: a repertoire for biology, ecology and engineering (2005).

[180] A. V. Spirov, M. A. Sabirov, and D. M. Holloway, "Systems Evolutionary Biology of Waddington's Canalization and Genetic Assimilation," Evolutionary Physiology and Biochemistry - Advances and Perspectives (2018).

[181] L. Vuillon and C. Lesieur, "From local to global changes in proteins: A network view," Current Opinion in Structural Biology 31 (2015).

[182] A. Thompson, *Hardware evolution : automatic design of electronic circuits in reconfigurable hardware by Artificial Evolution* (Springer, 1998), p. 115.

[183] A. Thompson, "On the automatic design of robust electronics through artificial evolution," Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **1478,** 13–24 (1998).

[184] J. Gerhart and M. Kirschner, *Cells, embryos, and evolution : toward a cellular and developmental understanding of phenotypic variation and evolutionary adaptability* (Wiley, 1997), p. 642, especially Ch 9.].

[185] N. Kashtan and U. Alon, "Spontaneous evolution of modularity and network motifs," Proceedings of the National Academy of Sciences of the United States of America 102 (2005).

[186] L. Fang, Y. Li, L. Ma, Q. Xu, F. Tan, and G. Chen, "GRNdb: decoding the gene regulatory networks in diverse human and mouse conditions," Nucleic Acids Research **49,** D97–D103 (2021).

[187] V. A. Huynh-Thu, A. Irrthum, L. Wehenkel, and P. Geurts, "Inferring Regulatory Networks from Expression Data Using Tree-Based Methods," PLOS ONE **5,** e12776 (2010).

[188] E. Altermann and T. R. Klaenhammer, "PathwayVoyager: pathway mapping using the Kyoto Encyclopedia of Genes and Genomes (KEGG) database," BMC genomics **6,** 1–7 (2005).

[189] B. Lehne and T. Schlitt, "Protein-protein interaction databases: keeping up with growing interactomes.," Human genomics **3,** 291–297 (2009).

[190] D. Szklarczyk *et al.*, "The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets," Nucleic acids research **49,** D605–D612 (2021).

[191] S. L. Salzberg, "Open questions: How many genes do we have?," BMC Biology **16,** 1–3 (2018), note that Gencode, Ensembl, and RefSeq report 1,000 to 1,400 fewer protein-encoding genes than CHESS in Tab 1.

[192] M. K. Burke, J. P. Dunham, P. Shahrestani, K. R. Thornton, M. R. Rose, and A. D. Long, "Genome-wide analysis of a long-term evolution experiment with Drosophila," Nature **467,** 587–590 (2010).

[193] H. H. McAdams and L. Shapiro, "System-level design of bacterial cell cycle control," FEBS Letters **583,** 3984–3991 (2009).

[194] E. A. Boyle, Y. I. Li, and J. K. Pritchard, "An Expanded View of Complex Traits: From Polygenic to Omnigenic," Cell **169,** 1177–1186 (2017).

[195] J. Alber, J. Gramm, and R. Niedermeier, "Faster exact algorithms for hard problems: A parameterized point of view," Discrete Mathematics **229,** 3–27 (2001).

[196] M. I. Saunders, C. J. Brown, M. M. Foley, C. M. Febria, R. Albright, M. G. Mehling, M. T. Kavanaugh, and D. D. Burfeind, "Human impacts on connectivity in marine and freshwater ecosystems assessed using graph theory: A review," Marine and Freshwater Research 67 (2016).

[197] N. E. Baldwin, E. J. Chesler, S. Kirov, M. A. Langston, J. R. Snoddy, R. W. Williams, and B. Zhang, "Computational, Integrative, and Comparative Methods for the Elucidation of Genetic Coexpression Networks," Journal of Biomedicine and Biotechnology **2005,** 172–180 (2005).

[198] J. D. Eblen, C. A. Phillips, G. L. Rogers, and M. A. Langston, "The maximum clique enumeration problem: algorithms, applications, and implementations," BMC Bioinformatics **13,** S5 (2012).

[199] S. P. Farris and M. F. Miles, "Ethanol Modulation of Gene Networks: Implications for Alcoholism," Neurobiology of disease **45,** 115 (2012).

[200] L. Lovász, "Graph minor theory," Bulletin of the American Mathematical Society **43,** 75–86 (2005).

[201] N. Robertson and P. Seymour, "Graph Minors. XX. Wagner's conjecture," Journal of Combinatorial Theory, Series B **92,** 325–357 (2004), theorem 10.5 in an unpublished 2017 update to this paper.

[202] M. Axenovich, A. Giraõ, R. Snyder, and L. Weber, "Strong complete minors in digraphs," Combinatorics, Probability and Computing **31,** 489–506 (2022).

[203] T. Mészáros and R. Steiner, "Complete directed minors and chromatic number," Journal of Graph Theory (2022).

[204] L. Gishboliner, R. Steiner, and T. Szabó, "Oriented Cycles in Digraphs of Large Outdegree," Combinatorica (2022).

[205] K. I. Kawarabayashi and S. Kreutzer, "Towards the graph minor theorems for directed graphs," Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **9135,** 3–10 (2015).

[206] V. Campos, R. Lopes, A. K. Maia, and I. Sau, "Adapting the Directed Grid Theorem into an FPT Algorithm,", 2020.

[207] C. Ma, R. Cheng, L. V. S. Lakshmanan, T. Grubenmann, Y. Fang, and X. Li, "LINC," Proceedings of the VLDB Endowment **13,** 155–168 (2019).

[208] X. Hu, Y. Tao, and C. W. Chung, "Massive graph triangulation," Proceedings of the ACM SIGMOD International Conference on Management of Data pp. 325–336 (2013).

[209] A. Kara, H. Q. Ngo, M. Nikolic, D. Olteanu, and H. Zhang, "Counting Triangles under Updates in Worst-Case Optimal Time," In , P. Barcelo and M. Calautti, eds., 22nd International Conference on Database Theory pp. 4:1–4:18 (Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2019).

[210] P. Ribeiro, P. Paredes, M. E. P. Silva, D. Aparicio, and F. Silva, "A Survey on Subgraph Counting," ACM Computing Surveys **54,** 1–36 (2022).

[211] S. Wernicke and F. Rasche, "FANMOD: a tool for fast network motif detection," Bioinformatics **22,** 1152–1153 (2006).

[212] T. Hočevar and J. Demšar, "A combinatorial approach to graphlet counting," Bioinformatics **30,** 559–565 (2014).

[213] I. Melckenbeeck, P. Audenaert, D. Colle, and M. Pickavet, "Efficiently counting all orbits of graphlets of any order in a graph using autogenerated equations," Bioinformatics **34,** 1372–1380 (2018).

[214] M. Gonen and Y. Shavitt, "Approximating the Number of Network Motifs," Internet Mathematics **6,** 349–372 (2009).

[215] D. Marcus and Y. Shavitt, "Efficient counting of network motifs," Proceedings - International Conference on Distributed Computing Systems pp. 92–98 (2010).

[216] A. Pinar, C. Seshadhri, and V. Vishal, "ESCAPE," In , Proceedings of the 26th International Conference on World Wide Web pp. 1431–1440 (International World Wide Web Conferences Steering Committee, 2017).

[217] P. Ribeiro, F. Silva, and L. Lopes, "Parallel Calculation of Subgraph Census in Biological Networks.," Proceedings on the International Conference on Bioinformatics pp. 56–65 (2010).

[218] A. Milinković, M. S, and L. Lazić, "A contribution to acceleration of graphlet counting," Proceedings of the Infoteh Jahorina Symposium Infoteh Jahorina **14,** 741–745 (2015).

[219] S. Shahrivari and S. Jalili, "Fast Parallel All-Subgraph Enumeration Using Multicore Machines," Scientific Programming **2015,** 1–11 (2015).

[220] A. N. Eddin and P. Ribeiro, "Scalable subgraph counting using MapReduce," In , Proceedings of the Symposium on Applied Computing **Part F128005,** 1574–1581 (ACM, 2017).

[221] R. J. Prill, P. A. Iglesias, and A. Levchenko, "Dynamic Properties of Network Motifs Contribute to Biological Network Organization," PLoS Biology **3,** e343 (2005).

[222] S. Pawar, "Community assembly, stability and signatures of dynamical constraints on food web structure," Journal of Theoretical Biology **259,** 601–612 (2009).

[223] T. Johnson, N. Robertson, P. D. Seymour, and R. Thomas, "Directed tree-width," Journal of Combinatorial Theory. Series B **82,** 138–154 (2001), butterfly minors are defined, but not called by that name, in section 5.

[224] I. Adler, "Directed tree-width examples," Journal of Combinatorial Theory, Series B **97,** 718–725 (2007).

[225] J. Bensmail, V. Campos, M. Correia, A. K. Maia, N. Nisse, and A. Silva, available at http://www-sop.inria.fr/members/Nicolas.Nisse/cylindrical.pdf (unpublished).

[226] M. Hatzel, R. Rabinovich, and S. Wiederrecht, "Cyclewidth and the Grid Theorem for Perfect Matching Width of Bipartite Graphs," Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **11789 LNCS,** 53–65 (2019).

[227] A. L. Barabási and Z. N. Oltvai, "Network biology: understanding the cell's functional organization," Nature Reviews Genetics 2004 5:2 **5,** 101–113 (2004).

[228] T. I. Lee *et al.*, "Transcriptional regulatory networks in Saccharomyces cerevisiae," Science **298,** 799–804 (2002).

[229] U. Alon, in *An Introduction to Systems Biology: Design Principles of Biological Circuits* (Chapman & Hall, 2007), Chap. 5, p. 91, all directed unsigned graphs appear in Figs 4.1 (3-motifs) and 5.5 (4-motifs). See Ref. [38].

[230] H. Yu, N. M. Luscombe, J. Qian, and M. Gerstein, "Genomic analysis of gene expression relationships in transcriptional regulatory networks," Trends in Genetics **19,** 422–427 (2003).

[231] M. T. Laub, H. H. McAdams, T. Feldblyum, C. M. Fraser, and L. Shapiro, "Global analysis of the genetic network controlling a bacterial cell cycle," Science **290,** 2144–2148 (2000).

[232] A. Zaslaver, A. E. Mayo, R. Rosenberg, P. Bashkin, H. Sberro, M. Tsalyuk, M. G. Surette, and U. Alon, "Just-in-time transcription program in metabolic pathways," Nature Genetics 2004 36:5 **36,** 486–491 (2004).

[233] M. Z. Ali and R. C. Brewster, "Controlling gene expression timing through gene regulatory architecture," PLOS Computational Biology **18,** e1009745 (2022).

[234] J. Zhao, A. Pokhilko, O. Ebenhöh, S. J. Rosser, and S. D. Colloms, "A single-input binary counting module based on serine integrase site-specific recombination," Nucleic Acids Research **47,** 4896–4909 (2019).

[235] S. M. Vallina and C. L. Quéré, "Stability of complex food webs: Resilience, resistance and the average interaction strength," Journal of Theoretical Biology **272,** 160–173 (2011).

[236] J. Doyle and M. Cseste, "Motifs, Control and Stability," PLoS Biology **3,** e392 (2005).

[237] M. A. Savageau, "Design of molecular control mechanisms and the demand for gene expression (positive compared to negative control/evolution/ecological niche)," Genetics **74,** 5647–5651 (1977).

[238] G. C. Conant and A. Wagner, "Convergent evolution of gene circuits," Nature Genetics 2003 34:3 **34,** 264–266 (2003).

[239] A. Warmflash, P. Francois, and E. D. Siggia, "Pareto evolution of gene networks: an algorithm to optimize multiple fitness objectives," Physical Biology **9,** 056001 (2012).

[240] P. J. Ingram, M. P. Stumpf, and J. Stark, "Network motifs: Structure does not determine function," BMC Genomics **7,** 1–12 (2006).

[241] V. Spirin and L. A. Mirny, "Protein complexes and functional modules in molecular networks," Proceedings of the National Academy of Sciences of the United States of America **100,** 12123–12128 (2003).

[242] H. H. McAdams and L. Shapiro, "A bacterial cell-cycle regulatory network operating in time and space," Science **301,** 1874–1877 (2003).

[243] W. Filipowicz, S. N. Bhattacharyya, and N. Sonenberg, "Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight?," Nature Reviews Genetics 2008 9:2 **9,** 102–114 (2008).

[244] G. Sun, M. Li, G. Li, Y. Tian, R. Han, and X. Kang, "Identification and abundance of miRNA in chicken hypothalamus tissue determined by Solexa sequencing," Genetics and Molecular Research **11,** 4682–4694 (2012).

[245] W. K. Paik, D. C. Paik, and S. Kim, "Historical review: the field of protein methylation," Trends in Biochemical Sciences **32,** 146–152 (2007).

[246] L. I. Hu, B. P. Lima, and A. J. Wolfe, "Bacterial protein acetylation: the dawning of a new age," Molecular Microbiology **77,** 15–21 (2010).

[247] R. H. Newman, J. Zhang, and H. Zhu, "Toward a systems-level view of dynamic phosphorylation networks," Frontiers in Genetics 5 (2014).

[248] H. Nishi, A. Shaytan, and A. R. Panchenko, "Physicochemical mechanisms of protein regulation by phosphorylation," Frontiers in Genetics **5,** 270 (2014).

[249] L. Hicke, "Protein regulation by monoubiquitin," Nature Reviews Molecular Cell Biology 2001 2:3 **2,** 195–201 (2001).

[250] N. L. Anderson and N. G. Anderson, "The Human Plasma Proteome," Molecular & Cellular Proteomics **1,** 845–867 (2002).

[251] V. Chubukov, I. A. Zuleta, and H. Li, "Regulatory architecture determines optimal regulation of gene expression in metabolic pathways," Proceedings of the National Academy of Sciences of the United States of America **109,** 5127–5132 (2012).

[252] P. François and E. D. Siggia, "A case study of evolutionary computation of biochemical adaptation," Physical Biology **5,** 12 (2008).

[253] M. D. Schroeder, M. Pearce, J. Fak, H. Q. Fan, U. Unnerstall, E. Emberly, N. Rajewsky, E. D. Siggia, and U. Gaul, "Transcriptional Control in the Segmentation Gene Network of Drosophila," PLOS Biology **2,** e271 (2004).

[254] D. O. Conover and S. B. Munch, "Sustaining Fisheries Yields Over Evolutionary Time Scales," Science **297,** 94–96 (2002).

[255] A. W. Walters and D. M. Post, "An Experimental Disturbance Alters Fish Size Structure but not Food Chain Length in Streams," Ecology **89,** 3261–3267 (2008).

[256] A. B. Neuheimer and C. T. Taggart, "Can changes in length-at-age and maturation timing in Scotian Shelf haddock (Melanogrammus aeglefinus) be explained by fishing?," Canadian Journal of Fisheries and Aquatic Science **67,** 854–865 (2010).

[257] M. R. Rose, H. B. Passananti, and M. Matos, *Methuselah Flies* (WORLD SCIENTIFIC, 2004).

[258] V. Spirin, M. S. Gelfand, A. A. Mironov, and L. A. Mirny, "A metabolic network in the evolutionary context: Multiscale structure and modularity," Proceedings of the National Academy of Sciences of the United States of America **103,** 8774–8779 (2006).

[259] S. Uchida, B. Drossel, and U. Brose, "The structure of food webs with adaptive behaviour," Ecological Modelling **206,** 263–276 (2007).

[260] S. Uchida and B. Drossel, "Relation between complexity and stability in food webs with adaptive behavior," Journal of Theoretical Biology **247,** 713–722 (2007).

[261] L. Berec, J. Eisner, and V. Krivan, "Adaptive foraging does not always lead to more complex food webs," Journal of Theoretical Biology **266,** 211–218 (2010).

[262] M. M. Waldrop, *Complexity: The Emerging Science at the Edge of Order and Chaos* (Simon & Schuster, 1992), quoted on p. 12-13.

[263] P. J. Hornsby, in *The nature of aging and the geroscience hypothesis* (Elsevier, 2021), pp. 69–76.

[264] M. A. Rose, "Antagonistic Pleiotropy, Dominance, and Genetic Variation," Heredity **48,** 63–78 (1982).

[265] D. P. Shanley and T. B. L. Kirkwood, "Calorie Restriction and Aging: A Life-History Analysis," Evolution **54,** 740–750 (2000).

[266] C. M. McCay, M. F. Crowell, and L. A. Maynard, "The Effect of Retarded Growth Upon the Length of Life Span and Upon the Ultimate Body Size: One Figure," The Journal of Nutrition **10,** 63–79 (1935).

[267] R. Weindruch and R. S. Sohal, "Caloric Intake and Aging," New England Journal of Medicine **337,** 986–994 (1997).

[268] J. Labbadia and R. I. Morimoto, "The Biology of Proteostasis in Aging and Disease," https://doi.org/10.1146/annurev-biochem-060614-033955 **84,** 435–464 (2015).

[269] E. T. Powers, R. I. Morimoto, A. Dillin, J. W. Kelly, and W. E. Balch, "Biological and chemical approaches to diseases of proteostasis deficiency," Annual Review of Biochemistry **78,** 959–991 (2009).

[270] R. Hrdlickova, M. Toloue, and B. Tian, "RNA-Seq methods for transcriptome analysis," WIREs RNA 8 (2017).

[271] R. Aebersold and M. Mann, "Mass spectrometry-based proteomics," Nature **422,** 198–207 (2003).

[272] I. Subramanian, S. Verma, S. Kumar, A. Jere, and K. Anamika, "Multi-omics Data Integration, Interpretation, and Its Application,", 2020.

[273] G. Nicora, F. Vitali, A. Dagliati, N. Geifman, and R. Bellazzi, "Integrated Multi-Omics Analyses in Oncology: A Review of Machine Learning Methods and Tools," Frontiers in Oncology 10 (2020).

[274] N. Ishii *et al.*, "Multiple High-Throughput Analyses Monitor the Response of E. coli to Perturbations," Science **316,** 593–597 (2007).

[275] L. Boisen and P. Kristensen, "Confronting cellular heterogeneity in studies of protein metabolism and homeostasis in aging research," Advances in Experimental Medicine and Biology **694,** 234–244 (2010).

[276] D. Dai *et al.*, "Altered proteome turnover and remodeling by short-term caloric restriction or rapamycin rejuvenate the aging heart," Aging Cell **13,** 529–539 (2014).

[277] P. P. Karunadharma *et al.*, "Subacute calorie restriction and rapamycin discordantly alter mouse liver proteome homeostasis and reverse aging effects," Aging Cell **14,** 547–557 (2015).

[278] N. Basisty *et al.*, "Mitochondrial-targeted catalase is good for the old mouse proteome, but not for the young: 'reverse' antagonistic pleiotropy?," Aging Cell **15,** 634–645 (2016).

[279] N. Basisty, J. G. Meyer, and B. Schilling, "Protein Turnover in Aging and Longevity," PROTEOMICS 18 (2018).

[280] N. Basisty, A. Holtz, and B. Schilling, "Accumulation of "Old Proteins" and the Critical Need for MS-based Protein Turnover Measurements in Aging and Longevity," PROTEOMICS 20 (2020).

[281] N. B. Basisty, Y. Liu, J. Reynolds, P. P. Karunadharma, D.-F. Dai, J. Fredrickson, R. P. Beyer, M. J. MacCoss, and P. S. Rabinovitch, "Stable Isotope Labeling Reveals Novel Insights Into Ubiquitin-Mediated Protein Aggregation With Age, Calorie Restriction, and Rapamycin Treatment," The Journals of Gerontology: Series A **73,** 561–570 (2018).

[282] T. Mathieson *et al.*, "Systematic analysis of protein turnover in primary cells," Nature Communications **9,** 689 (2018).

[283] I. Dhondt, V. A. Petyuk, S. Bauer, H. M. Brewer, R. D. Smith, G. Depuydt, and B. P. Braeckman, "Changes of Protein Turnover in Aging Caenorhabditis elegans," Molecular and Cellular Proteomics **16,** 1621–1633 (2017).

[284] I. Spector, "Animal longevity and protein turnover rate," Nature **249,** 66 (1974).

[285] A. C. Thompson *et al.*, "Reduced in vivo hepatic proteome replacement rates but not cell proliferation rates predict maximum lifespan extension in mice," Aging Cell **15,** 118–127 (2016).

[286] D. Schäfer and P. Baker, *Old Age and Disease in Early Modern Medicine* (Routledge, 2015), see especially discussion of Grecco-Roman writings in Ch. 1.1-1.4.

[287] W. J. Darby, "Early concepts on the role of nutrition, diet and longevity," Progress in Clinical and Biological Research **326,** 1–20 (1990).

[288] H. N. Munro, in *Mammalian Protein Metabolism*, H. N. Munro and J. B. Allison, eds., (Elsevier Science, 1964), Chap. 1.

[289] G. Lis, L. I. Wassenaar, and M. J. Hendry, "High-Precision Laser Spectroscopy D/H and 18 O/ 16 O Measurements of Microliter Natural Water Samples," Analytical Chemistry **80,** 287–293 (2008).

[290] J. C. Price, W. E. Holmes, K. W. Li, N. A. Floreani, R. A. Neese, S. M. Turner, and M. K. Hellerstein, "Measurement of human plasma proteome dynamics with 2H2O and liquid chromatography tandem mass spectrometry," Analytical Biochemistry **420,** 73–83 (2012).

[291] M. Vaudel, H. Barsnes, F. S. Berven, A. Sickmann, and L. Martens, "SearchGUI: An open-source graphical user interface for simultaneous OMSSA and X!Tandem searches," PROTEOMICS **11,** 996–999 (2011).

[292] H. Barsnes and M. Vaudel, "SearchGUI: A Highly Adaptable Common Interface for Proteomics Search and de Novo Engines," Journal of Proteome Research **17,** 2552–2555 (2018).

[293] M. Vaudel, J. M. Burkhart, R. P. Zahedi, E. Oveland, F. S. Berven, A. Sickmann, L. Martens, and H. Barsnes, "PeptideShaker enables reanalysis of MS-derived proteomics data sets," Nature Biotechnology 2015 33:1 **33,** 22–24 (2015).

[294] B. C. Naylor, M. T. Porter, E. Wilson, A. Herring, S. Lofthouse, A. Hannemann, S. R. Piccolo, A. L. Rockwood, and J. C. Price, "DeuteRater: A tool for quantifying peptide isotope precision and kinetic proteomics," Bioinformatics **33,** 1514–1520 (2017).

[295] L. T. Johnson and C. J. Geyer, "Variable transformation to obtain geometric ergodicity in the random-walk Metropolis algorithm," https://doi.org/10.1214/12-AOS1048 **40,** 3050–3076 (2012).

[296] C. J. Geyer, in *Introduction to Markov Chain Monte Carlo*, S. Brooks, A. Gelman, G. Jones, and X.-L. Meng, eds., (CRC Press, 2011), pp. 3–48.

[297] A. Joseph *et al.*, "Effects of acyl-coenzyme A binding protein (ACBP)/diazepam-binding inhibitor (DBI) on body mass index," Cell Death and Disease 12 (2021).

[298] S. X. Cao, J. M. Dhahbi, P. L. Mote, and S. R. Spindler, "Genomic profiling of short- and long-term caloric restriction effects in the liver of aging mice," Proceedings of the National Academy of Sciences **98,** 10630–10635 (2001).

[299] M. D. Bruss, C. F. Khambatta, M. A. Ruby, I. Aggarwal, and M. K. Hellerstein, "Calorie restriction increases fatty acid synthesis and whole body fat oxidation rates," American Journal of Physiology - Endocrinology and Metabolism 298 (2010).

[300] J. C. Waterlow, in *Protein Turnover* (CAB International, 2006), Chap. 9, pp. 120–130.

[301] H. Hochstadt, "A Special Hill's Equation With Discontinuous Coefficients," The American Mathematical Monthly **70,** 18–26 (1963).

[302] S. Mandal, S. Ray, S. Roy, and S. E. Jørgensen, "Order to chaos and vice versa in an aquatic ecosystem," Ecological Modelling **197(3-4),** 498–504 (2006).

[303] L. Stone and D. He, "Chaotic oscillations and cycles in multi-trophic ecological systems," Journal of Theoretical Biology **248,** 282–290 (2007).

[304] J. C. Waterlow, in *Protein Turnover* (CAB International, 2006), Chap. 18, p. 276.

[305] R. S. Petralia, M. P. Mattson, and P. J. Yao, "Aging and longevity in the simplest animals and the quest for immortality," Ageing Research Reviews **16,** 66–82 (2014).

[306] S. Besteiro, R. A. Williams, G. H. Coombs, and J. C. Mottram, "Protein turnover and differentiation in Leishmania," International Journal for Parasitology **37,** 1063–1075 (2007).

[307] F. Danielsson, M. Skogs, M. Huss, E. Rexhepaj, G. O'Hurley, D. Klevebring, F. Pontén, A. K. Gad, M. Uhlén, and E. Lundberg, "Majority of differentially expressed genes are

down-regulated during malignant transformation in a four-stage model," Proceedings of the National Academy of Sciences of the United States of America **110,** 6853–6858 (2013).

[308] M. Poulain, G. M. Pes, C. Grasland, C. Carru, L. Ferrucci, G. Baggio, C. Franceschi, and L. Deiana, "Identification of a geographic area characterized by extreme longevity in the Sardinia island: the AKEA study," Experimental Gerontology **39,** 1423–1429 (2004).

[309] D. C. Willcox, B. J. Willcox, H. Todoriki, J. D. Curb, and M. Suzuki, "Caloric restriction and human longevity: what can we learn from the Okinawans?," Biogerontology **7,** 173–177 (2006).

[310] D. Buettner and S. Skemp, "Blue zones: lessons from the world's longest lived," American journal of lifestyle medicine **10,** 318–321 (2016).

[311] J. Most, V. Tosti, L. M. Redman, and L. Fontana, "Calorie restriction in humans: An update," Ageing Research Reviews **39,** 36–45 (2017).

[312] B. C. Naylor *et al.*, "Utilizing Nonequilibrium Isotope Enrichments to Dramatically Increase Turnover Measurement Ranges in Single Biopsy Samples from Humans," Journal of Proteome Research **21,** 2703–2714 (2022).

[313] A. Keys, J. Brožek, A. Henschel, O. Mickelsen, and H. L. Taylor, *The biology of human starvation* (University of Minnesota Press, 1950), Vol. 1 and 2.

[314] S. Tara, *The Secret Life of Fat: The Science Behind the Body's Least Understood Organ and What It Means for You* (WW Norton, 2016), Chap. 11: Mind Over Fat.

[315] M. Hindhede, "The Effect of Food Restriction During War on Mortality in Copenhagen," Journal of the American Medical Association **74,** 381 (1920).

[316] A. Strøm, R. Jensen, M. Oslo, and M. Oslo, "Mortality from Circulatory Diseases in Norway, 1940-1945," The Lancet **257,** 126–129 (1951).

[317] S. B. Racette, E. P. Weiss, D. T. Villareal, H. Arif, K. Steger-May, K. B. Schechtman, L. Fontana, S. Klein, and J. O. Holloszy, "One Year of Caloric Restriction in Humans: Feasibility and Effects on Body Composition and Abdominal Adipose Tissue," The Journals of Gerontology: Series A **61,** 943–950 (2006).

[318] A. J. Dirks and C. Leeuwenburgh, "Caloric restriction in humans: Potential pitfalls and health concerns," Mechanisms of Ageing and Development **127,** 1–7 (2006).

[319] L. D. Fricker, "Proteasome Inhibitor Drugs," Annual Review of Pharmacology and Toxicology **60,** 457–476 (2020).

[320] J. Guo, X. Huang, L. Dou, M. Yan, T. Shen, W. Tang, and J. Li, "Aging and aging-related diseases: from molecular mechanisms to interventions and treatments," Signal Transduction and Targeted Therapy **7,** 391 (2022).

[321] A. L. Hopkins, "Network pharmacology: The next paradigm in drug discovery,", 2008.

[322] M. A. Lane, D. K. Ingram, and G. S. Roth, "2-Deoxy-D-Glucose Feeding in Rats Mimics Physiologic Effects of Calorie Restriction," Journal of Anti-Aging Medicine **1,** 327–337 (1998).

[323] D. K. Ingram, M. Zhu, J. Mamczarz, S. Zou, M. A. Lane, G. S. Roth, and R. deCabo, "Calorie restriction mimetics: an emerging research field," Aging Cell **5,** 97–108 (2006).

[324] F. Madeo, D. Carmona-Gutierrez, S. J. Hofer, and G. Kroemer, "Caloric Restriction Mimetics against Age-Associated Disease: Targets, Mechanisms, and Therapeutic Potential,", 2019.

[325] S. J. Hofer, S. Davinelli, M. Bergmann, G. Scapagnini, and F. Madeo, "Caloric Restriction Mimetics in Nutrition and Clinical Trials," Frontiers in Nutrition 8 (2021).

[326] M. R. Rose, L. F. Greer, K. H. Phung, G. A. Rutledge, M. A. Phillips, C. N. K. Anderson, and L. D. Mueller, in *A Hamiltonian Demography of Life History* (Cambridge University Press, 2017), p. 40.

[327] K. C. H. Fearon, D. T. Hansell, T. Preston, J. A. Plumb, J. Davies, D. Shapiro, A. Shenkin, K. C. Calman, and H. J. G. Burns, "Influence of whole body protein turnover rate on resting energy expenditure in patients with cancer," Cancer Research **48,** 2590–2595 (1988).

[328] C. Montagut, A. Rovira, P. Gascon, J. Ross, and J. Albanell, "Preclinical and clinical development of the proteasome inhibitor bortezomib in cancer treatment," Drugs of Today **41,** 299 (2005).

[329] F. Demarchi and C. Brancolini, "Altering protein turnover in tumor cells: New opportunities for anti-cancer therapies," Drug Resistance Updates **8,** 359–368 (2005).

[330] P. K. Sorger *et al.*, "Quantitative and Systems Pharmacology in the Post-genomic Era: New Approaches to Discovering Drugs and Understanding Therapeutic Mechanisms,", 2011.

[331] S. Hassani, *Mathematical Physics: A Modern Introduction to Its Foundations*, 2ed ed. (Springer, Heidelberg, 2013), p. 1205, equation 14.2.1.

[332] D. Smith and B. Webb, "Hidden symmetries in real and theoretical networks," Physica A: Statistical Mechanics and its Applications **514,** 855–867 (2019).

[333] V. Jovovic, "Sequence A053517," in *The On-Line Encyclopedia of Integer Sequences* (published electronically at https://oeis.org, 2000).

[334] M. Swat, A. Kel, and H. Herzel, "Bifurcation analysis of the regulatory modules of the mammalian G1/S transition," Bioinformatics **20,** 1506–1511 (2004).

[335] C. N. K. Anderson and M. K. Transtrum, "Sloppy model analysis identifies bifurcation parameters without Normal Form analysis,", 2022.

[336] N. Varghese, S. Werner, A. Grimm, and A. Eckert, "Dietary Mitophagy Enhancer: A Strategy for Healthy Brain Aging?," Antioxidants **9,** 932 (2020).