ABSTRACT

Title of Dissertation: DYNAMICS, NETWORKS, AND INFORMATION: METHODS FOR NONLINEAR INTERACTIONS IN BIOLOGICAL SYSTEMS

Jesse Milzman
Doctor of Philosophy, 2021

Dissertation Directed by: Professor Doron Levy
Department of Mathematics

In this dissertation, we investigate complex, non-linear interactions in biological systems. This work is presented as two independent projects. The mathematics and biology in each differ, yet there is a unity in that both frameworks are interested in biological responses that cannot be reduced to linear causal chains, nor can they be expressed as an accumulation of binary interactions.

In the first part of this dissertation, we use mathematical modeling to study tumor-immune dynamics at the cellular scale. Recent work suggests that LSD1 inhibition reduces tumor growth, increases T cell tumor infiltration, and complements PD1/PDL1 checkpoint inhibitor therapy. In order to elucidate the immunogenic effects of LSD1 inhibition, we create a delay differential equation model of tumor growth under the influence of the adaptive immune response in order to investigate the anti-tumor cytotoxicity of LSD1-mediated T cell dynamics. We fit our model to the B16 mouse model data

from Sheng *et al.* [107]. Our results suggest that the immunogenic effect of LSD1 inhibition accelerates anti-tumor cytoxicity. However, cytotoxicity does not seem to account for the slower growth observed in LSD1 inhibited tumors, despite evidence suggesting immune-mediation of this effect.

In the second part of this dissertation, we consider the partial information decomposition (PID) of response information within networks of interacting nodes, inspired by biomolecular networks. We specifically study the potential of PID synergy as a tool for network inference and edge nomination. We conduct both numeric and analytic investigations of the $I_\cap^{\min}$ and $I_\cap^{\mathrm{PM}}$ PIDs, from [125] and [45], respectively. We find that the $I_\cap^{\mathrm{PM}}$ synergy suffers from issues of non-specificity, while $I_\cap^{\min}$ synergy is specific but somewhat insensitive. In the course of our work, we extend the $I_\cap^{\mathrm{PM}}$ and $I_\cap^{\min}$ PIDs to continuous variables for a general class of noise-free trivariate systems. The $I_\cap^{\mathrm{PM}}$ PID does not respect conditional independence, while $I_\cap^{\min}$ does, as demonstrated through asymptotic analysis of linear and non-linear interaction kernels. The technical results of this chapter relate the analytic and information-theoretic properties of our interactions, by expressing the continuous PID of noise-free interactions in terms of the partial derivatives of the interaction kernel g.

DYNAMICS, NETWORKS, AND INFORMATION: METHODS FOR NONLINEAR INTERACTIONS IN BIOLOGICAL SYSTEMS

by

Jesse M. W. Milzman

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2021

Advisory Committee:
Professor Doron Levy, Chair/Advisor
Professor Vince Lyzinski, Co-Chair/Advisor
Professor Radu Balan
Professor Wojciech Czaja
Professor Michelle Girvan

Acknowledgments

I wish to thank the innumerable people who made my graduate experience worthwhile.

First, I would like to thank my advisor, Professor Doron Levy. As a mentor and a friend, I could not have asked for a better guide into mathematical oncology. His patience and generosity saw me through many set-backs, and I am very grateful. I would like to thank my co-advisor, Professor Vince Lyzinski. His unceasing dedication and aid made the latter part of this work possible.

I would like to thank Professor Michelle Girvan. I am not the only COMBINE Fellow who gained much from her energetic leadership. Dr. Girvan has been enthusiastically supportive of my progress since I joined the COMBINE fellowship, and has helped me grow as an interdisciplinary researcher.

I would like to thank Professors Radu Balan and Wojciech Czaja, for serving on my committee. I would like to thank Dr. Balan for the many good conversations, and for my time participating in his group's seminar in previous years.

I would like to thank the COMBINE Program at the University of Maryland, for supporting me during my research, introducing me to network science, and developing my interdisciplinary research capacities. I owe much

of my current research communication skills to the environment provided by COMBINE, and continue to rely on the scale-free network of colleagues that I met through the program. I look forward to opportunities to give back to the COMBINE Fellowship as an alumnus. I would like to thank Dr. Daniel Serrano, who worked with me frequently in my research for the COMBINE Program, and was also invaluable in his aid to the COMBINE student committee. I thank Glynis Smith, for tirelessly coordinating for all COMBINE fellows and our events.

I want to thank M. Cristina Garcia, the Graduate Program Coordinator for the Department of Mathematics, for all the help she has provided to me over the years on matters great and small.

I want to thank the friends and colleagues that I have made during my time at the University of Maryland. I owe particular thanks to Liam Fowl, Stavros Papathanasiou, S. Gilles, Stephen Sorokanich, Dina Genkina, Troy Sewell, and Spiros Lentas for their irreplaceable friendship.

Finally, I want to thank my family. My father, David Milzman, and my mother, Colette Magnant, for teaching me perseverance through caffeine dependency. My brothers, Matt and Danny Milzman, have always kept things interesting.

Table of Contents

List of Tables

Tables for Chapter 2

Tables for Chapter 3

List of Figures

Figures for Chapter 2

Figures for Chapter 3

Chapter 1
Introduction

The aim of this dissertation is to present two perspectives on the inference of complex, non-linear interactions in biological systems. The mathematics and biology in each differ, yet there is a unity in that both frameworks are interested in biological responses that cannot be reduced to linear causal chains, nor can they be expressed as an accumulation of binary interactions. One perspective aims to represent population-level interactions between cancer and immune cells with a model characterized by full system feedback, in the sense that none of the state variables are source or sink nodes. The other represents a lengthy investigation into the multivariate decomposition of synergistic interactions within networks, first using simulations within networks proper, and then with a formal analysis of trivariate systems. These demonstrate a diversity of approaches, utilizing both mathematical and computational tools, to uncover multivariate interactions within living systems. Or, to be more precise, we investigate the inference of multivariate interaction models, which better represent the flux of real-life observables, as compared to simpler unidirectional models [9].

Since the mathematics and biology of each work differ, they are to be presented in self-contained chapters. The first part (Chapter 2) develops a delay differential equations (DDE) model for tumor-immune dynamics, and uses standard modeling approaches to explore potential consequences of LSD1 inhibition on tumor growth, using experiment data from [107]. We both extend and simplify previous models of tumor-immune dynamics [47, 64] in order to investigate the impact of the inhibition of the LSD1 gene on tumor growth.

The following chapter (Ch. 3), which comprises the bulk of this dissertation,

inspects quantities of multivariate synergy and redundancy within model networks, inspired by molecular biology. The aim of that work is to explore synergy measures as instruments for the inference of interactions between genes, with respect to a target response. To that end, we extend information-theoretic metrics from the partial information decomposition (PID) literature to continuous interactions in order to understand their behavior. To do so, we extend the $I_\cap^{\min}$ and $I_\cap^{\mathrm{PM}}$ PIDs developed in [125] and [45], respectively, to continuous variables. We are able to distinguish specific and non-specific information metrics by their treatment of a given interaction kernel, determined by its analytic properties. The results of this investigation demonstrate a coherence of analytic and probabilistic perspectives.

This work is structured as follows. The rest of this introductory section will be devoted to a non-technical discursion on the relationship between interactions and biological complexity. This narrative will serve to frame the main works of this dissertation within a common intuition of multivariate interactions.[1] Then, we will present the two body chapters in turn. Each can be read as a stand-alone work, as they include all the prerequisite material within themselves, including mathematical preliminaries and literature review.

1.1 Interacting Systems: Models and Biology

Humans best understand the systems that they themselves build. It is understandable that we prefer to recast biological phenomenon in the language of artificial constructs: brains as computers, bodies as machines, and societies as networks. The issue with such metaphors is that artificial systems are implemented models, and conform to feed-forward flows of both material and

[1] Although this framework draws on rigorous work identifying hierarchies of complexity [5, 60], that is not its purpose here. Rather, we aim to present an intuition that guides the work in this dissertation.

information, as is more natural to human understanding. In addition, human-built systems are likely to minimize unnecessary interactions between parallel processes, i.e. between distinguished modules, in order to minimize the possibility of unforeseen, synergistic effects that may constitute system 'failure.' By contrast, natural systems emerge dynamically, and are characterized by cybernetic feedback loops, which enable their adaptive reconstitution. Moreover, at every scale of organization, there is significant cross-talk and 'inter-modular'[2] contamination.[9]

Thus, in applying models to biological phenomena, we can take for granted that the models are at a lower level of complexity. It is far from clear that any natural system would have an upper-bound on complexity, in the sense of an upper bound on the complexity of models beyond which we could not improve, given perfect data. Nonetheless, we never have perfect data, and it is well-known that beyond a crude level of complexity, models become overfit to their data, and their generalizability suffers as a result. In statistical modeling, it is typical to include an 'Ockham factor' [97], usually an information criterion, such as the Akaide Information Criterion (AIC) or Bayesian Information Criterion (BIC), introduced by Akaide in [1] and Schwarz [102], respectively. Using such measures, a modeler attempts to maximize the 'fitness' (log-likelihood), with a penalty for complexity (also called 'flexibility' in [97]).

Describing model complexity, as a more general mathematical task, is a non-trivial matter. There are many distinct approaches, arguably based upon completely different understandings of what is meant by the word 'complexity.' Nihat Ay *et al.* argue that there are three main meanings employed for math-

[2]It is worth noting that, as words, both 'network' and 'module' have an etymology suggesting artificiality. The frequent use of the word 'community' in modern network science seems more appropriate, given the greater focus on social and biological systems in contemporary networks literature.

ematical complexity: the difficulty in describing or generating an object, the difficulty in describing the regularities of an object, and the extent to which a whole object is "more than the sum of the parts" [5].

Central to this third, "Aristotelian" notion of complexity, developed in [5, 60], are the interactions between components. Indeed, we might refer to this approach as 'interaction complexity'. The works of this dissertation are concerned with inferring or investigating complex interactions, and quantifying the balance of information between interacting components of biological systems. To that end, we now offer a prefatory formalization of our intuition of interacting components within a complex system.

1.2 Interacting Systems and Higher Order Complexity

We will first consider the formulation in [60]. Suppose we have a system with a finite collection of parts V, $|V| = N$, and that the system and each of its part can be in one of a finite number of configurations, i.e. we have the full configuration space

$$\mathcal{X}_V = \prod_{v \in V} \mathcal{X}_v \tag{1.2.1}$$

where $\mathcal{X}_v$ is the finite alphabet of each component v. Real-valued functions of the system states $f : \mathcal{X}_V \to \mathbb{R}$ are equivalent to indexed vectors $f \in \mathbb{R}^{\mathcal{X}_V} = \mathbb{R}^{|\mathcal{X}_V|}$. Imagining this system as a collection of random variables, we may associate to it a probability mass function $p \in \mathcal{P}(\mathcal{X}_V)$, where

$$\mathcal{P}(\mathcal{X}_V) = \left\{ p \in [0,1]^{\mathcal{X}_V} \,\middle|\, \sum_{x \in \mathcal{X}_V} p(v) = 1 \right\}. \tag{1.2.2}$$

Using the exponential map, every function $f \in \mathbb{R}^{\mathcal{X}_V}$ can be non-orthogonally projected onto $\mathcal{P}(\mathcal{X}_V)$ as a Boltzmann distribution:

$$\exp : \mathbb{R}^{\mathcal{X}_V} \to \mathcal{P}(\mathcal{X}_V) \qquad f \mapsto \frac{e^f}{\sum_{x \in \mathcal{X}_V} e^{f(x)}} \tag{1.2.3}$$

For each $k = 0, 1, ..., N$, we may define the k-interactions on this system to be the set of real functions that depend on only k of the components, i.e.

$$\mathcal{I}_k = \bigcup_{A \subset V, |A| = k} \mathcal{I}_A \tag{1.2.4}$$

$$\mathcal{I}_A = \left\{ f \in \mathbb{R}^{\mathcal{X}_V} | f(x_A, x_{V \setminus A}) = f(x_A, x'_{V \setminus A}) \text{ for all } x'_{V \setminus A} \in \mathcal{X}_{V \setminus A} \right\} \tag{1.2.5}$$

Note that $\mathcal{I}_k \subset \mathcal{I}_{k+1}$ for all k. Via the exponential map, we may associate these interaction spaces to an exponential family $\mathcal{E}_k := \exp(\mathcal{I}_k)$. We then have a hierarchy of distributions:

$$\mathcal{E}_1 \subset ... \subset \mathcal{E}_N = \mathcal{P}(\mathcal{X}_V) \tag{1.2.6}$$

We see, then, that any probability mass function $p \in \mathcal{P}$ must be a N-interaction, $p \in \mathcal{E}_N$. Consider, on the other hand, if $p \in \mathcal{E}_1$. Then p factors into a product of single-variable marginals, and we have that the entire set of N system variables are independent. Note that $\mathcal{E}_0$ is the singleton containing the uniform distribution $p(x) = \frac{1}{|\mathcal{X}_V|}$. If, for a given p, we set $k^* = \min\{k | p \in \mathcal{E}_k\}$, then we may say that the probabilistic system $(\mathcal{X}_V, p)$ is a k^*-interaction system, or has a maximal interaction complexity of k^*. In [60], the authors went beyond a single metric, and quantified interaction complexity as a vector successive 'distances' (divergences) within the model hierarchy $\mathcal{E}_1 \subset ... \subset \mathcal{E}_N$. To make this more precise, for any $p \in \mathcal{P}(\mathcal{X}_V)$ and $\mathcal{E}_k$, we may consider the

Kullback-Liebler divergence

$$D(p||\mathcal{E}_k) = \inf_{q \in \mathcal{E}_k} D(p||q). \tag{1.2.7}$$

These divergences may be thought of as measuring the distance of the true distribution from a k-interaction. For each $k > 1$, define the k complexity component:

$$\mathcal{I}^{(k)}(p) = D(p||\mathcal{E}_{k-1}) - D(p||\mathcal{E}_k) \tag{1.2.8}$$

This quantity represents how much 'closer' to p one can get by increasing complexity in order to allow for k-interactions. The authors then propose that the non-negative vector $\mathcal{I}(p) = (\mathcal{I}^{(k)}(p))_k \subset \mathbb{R}^N_+$ as a complexity measure, quantifying the amount of system information that can be captured by each level of interaction complexity. They demonstrate the utility of this approach with multiple discrete dynamical systems. Intuitively, their approach partitions system information[3] into ascending levels of interaction complexity. [4]

The theme here that we want to emphasize is identification of model complexity with the order of the interactions composing the model. Let us now sketch what the generalization of such a framework might look like.

We again begin by identifying our system as having N components collected in $V = \{1, ..., N\}$, with associated configuration spaces $\mathcal{X}_1, ..., \mathcal{X}_N$. For our purpose, it is enough to require that each is a linear subspace $\mathcal{X}_v \subset \mathcal{B}_v$ for some Banach space $\mathcal{B}_v$. For instance, we may consider a finite time ODE of N

[3] More precisely, 'system information' refers to $D(p|\mathcal{E}_0)$, the 'distance' from the maximal entropy distribution.

[4] For instance, while studying a coupled tent map on a fully-connected network (i.e. where the dynamics of every $v \in V$ are coupled to every other $v' \in V$), Kahle *et al.* tune the coupling parameter ϵ. For $\epsilon > 0.45$, the system enters a phase of synchronized chaos, in which all system information is captured by pairwise interactions. Here, $\mathcal{I}^{(2)}$ dominates and $\mathcal{I}^{(k)} = 0$ for $k > 2$. On the other hand, on the 'edge of synchronized chaos', i.e. roughly $0.3 < \epsilon < 0.45$, more complex dynamics emerge, and $\mathcal{I}^{(k)}$ is nonzero for $k > 2$. This information collapses down into $\mathcal{I}^{(1)}$ and $\mathcal{I}^{(2)}$ as $\epsilon \to 0.45^-$ and synchronization is reached.

variables, in which case we have that each component is $\mathcal{X}_v = C^\infty([0, t_{\max}])$. Or we may consider a system of N real-valued random variables of bounded variance, in which case $\mathcal{X}_v \in L^2(\Omega, \mu)$. We will speak of a **realized system** as a point in the full configuration space:

$$\boldsymbol{x} = (x_v)_{v=1}^N \in \prod_{v=1}^N \mathcal{X}_v = \mathcal{X} \tag{1.2.9}$$

This would be a (vector-valued) orbit for the ODE example, or the specification of well-defined random variables in $L^2(\Omega, \mu)$. Systems are specified by the introduction of **laws** or constraints. We may identify a law with a operator $C : \mathcal{X} \to \mathcal{Y}$, for some Banach space $\mathcal{Y}$. For a given system, we would restrict the operators to an allowable collection $\mathcal{C}$. If we have a finite subset of laws $\{C_1, ..., C_M\} \subset \mathcal{C}$, we specify a submanifold $\mathcal{S}$ of $\mathcal{X}$ to be those $\boldsymbol{x} \in \mathcal{X}$ satisfying

$$\begin{aligned} C_1(\boldsymbol{x}) &= 0 \\ \vdots &= \vdots \\ C_M(\boldsymbol{x}) &= 0 \end{aligned}$$

where equality is within $\mathcal{Y}$. So, for instance, if our components are random variables, we could set covariances as laws, and $\mathcal{Y} = \mathbb{R}$. On the other hand, the laws of an ODE might specify that each $\dot{x}_k - F_k(\boldsymbol{x}) \equiv 0$ in C^∞. If there exists a unique $\boldsymbol{x} \in \mathcal{X}$ satisfying these laws, we may say that the laws $C_1, ..., C_M$ specify $\boldsymbol{x}$.[5]

In this context, we may now speak of interactions as those elements that comprise a law. Given a subset of components $A \subset V$, A-interactions $\mathcal{I}_A$ are

[5]This is similar to well-posedness, but that term carries an implication of continuity of $\boldsymbol{x}$ upon the laws.

those maps:

$$\mathcal{I}_A = \{f : \mathcal{X} \to \mathcal{Y} | f(\boldsymbol{x}_A, \boldsymbol{x}_{V\setminus A}) = f(\boldsymbol{x}_A, \boldsymbol{x}'_{V\setminus A}),\ \forall \boldsymbol{x}'_{V\setminus A} \in \mathcal{X}_{V\setminus A}\}. \tag{1.2.10}$$

We may then define k-interactions for $k = 0, ..., N$:

$$\mathcal{I}_k = \bigcup_{A\subset V,\, |A|=k} \mathcal{I}_A. \tag{1.2.11}$$

Analogous to the exponential families $\mathcal{E}_k$ from [60], we may consider families of systems:

$$\mathcal{S}_k = \{\boldsymbol{x} \in \mathcal{X} |\ \boldsymbol{x} \text{ is well specified by } \{C_\alpha\}_\alpha \subset \operatorname{span} \mathcal{I}_k\}. \tag{1.2.12}$$

If we are given a dissimilarity measure $\rho : \mathcal{X} \times \mathcal{X} \to [0, \infty]$, then for any specified $\boldsymbol{x} \in \mathcal{X}$, we might quantify interaction complexity with a vector similar to that in Eq. 1.2.8:

$$\mathcal{I}(\boldsymbol{x}) = (\mathcal{I}^{(1)}(\boldsymbol{x}), ..., \mathcal{I}^{(N)}(\boldsymbol{x})), \tag{1.2.13}$$

$$\mathcal{I}^{(k)}(\boldsymbol{x}) = \rho(\boldsymbol{x}, \mathcal{S}_k) = \inf_{\boldsymbol{y}\in\mathcal{S}_k} \rho(\boldsymbol{x}, \boldsymbol{y}). \tag{1.2.14}$$

The DDE system under investigation in Chapter 2 would fall within $\mathcal{S}_2$ for the correct framing of $\mathcal{X}$ and constraints C_α, i.e. it can be characterized with laws that depend upon binary interactions.[6] In Chapter 3, we investigate network-response systems numerically, and noise-free trivariate systems $T = g(X, Y)$ analytically. If we limit the allowable laws $\mathcal{C}$ to be joint distributions and moments, then these latter interacting systems will fall within $\mathcal{S}_3$. This

[6] Since we defined interactions as maps upon the full components (i.e. the full trajectories of each state variable, representing a cell population), the delayed terms in Eqs. (2.2.1) do not complicate matters here.

difference between the projects can be grasped visually. The DDE system in Ch. 2 can be represented as a network of binary interactions, as in Fig. 2.2. On the other hand, to represent the systems under investigation in Ch. 3, either the simulated response networks or noise-free trivariate systems, we would need to draw multiedges, since the interactions under investigation represent the effect of two gene expression predictors X and Y upon a response T, mediated by a kernel $g(X, Y)$. We are, however, able to represent the correlational structure relating gene predictors as standard networks (Figs. 3.1 & 3.5), since these correspond to laws of binary interactions (covariances) in $\mathcal{S}_2$.

From this perspective, it is apparent that interaction complexity is merely one dimension of complexity among many. For instance, by leaving the ODE framework and expanding allowable laws in $\mathcal{C}$ to include delayed terms, it is arguable that we have allowed more complexity than the we would have be limiting ourselves to non-delayed ODEs while allowing higher-order interactions from $\mathcal{I}_3$. In Ch. 3, of the two noise-free interaction systems under study, one of which utilizes a linear kernel and the other utilizing a non-linear sigmoidal function, both are specified within $\mathcal{S}_3$ and not $\mathcal{S}_2$. The interdependence between the three variables is much less simple for the non-linear function, and we would likewise call this 'more complex.'

This perspective parallels the discussion of model complexity as 'flexibility' in [97]. In that work, Rougier and Priebe discuss the decomposition of log-evidence, the statistical objective, as a difference of 'fitness' (log-likelihood) and a complexity penalty, which they term 'flexibility.' For observations y^{obs}, the evidence for the fitted parameter vector $\hat{\theta}$ is given by

$$\log E(\hat{\theta}) = \log f(y^{\text{obs}}; \hat{\theta}) - \underbrace{\log \frac{\pi^{\star}(\hat{\theta})}{\pi(\hat{\theta})}}_{\text{Flexibility}} \tag{1.2.15}$$

where $\hat{\theta} = \text{argmax}_\theta \log E(\theta)$, π is the prior distribution, and π^* is the posterior. They describe a model as flexible if the latter term is large, i.e. $\pi^*(\hat{\theta}) >> \pi(\hat{\theta})$. This occurs when θ has many degrees of freedom *and* the prior distribution is unconstrained. To our way of thinking, interaction complexity is analogous to a kind of operator dimensionality, while the prior restriction of allowable laws in $\mathcal{C}$ is analogous to a strong prior distribution. In order to ensure that a model maintains enough rigidity to generalize, a modeler may restrict one or both of these qualities.

We are now ready to present our contribution to the study of complex interactions in biological systems. In the two works presented in this dissertation, we keep the interactions under investigation simple, in terms of their k-order as described above ($\mathcal{I}_2$ and $\mathcal{I}_3$, respectively, for Ch. 2 and Ch. 3). Our work with DDE modeling of tumor-immune dynamics in Ch. 2 must be somewhat limited in both the interaction complexity and function complexity (that is, the allowable laws $\mathcal{C}$ for the system equations), due to data sparsity. Our investigation of the partial information decomposition of gene networks in Ch. 3 is more focused upon the information-theoretic behavior of network models themselves, rather than the inference of a model from limited data. In that investigation, interaction complexity is instead limited for reasons of computational tractability and straight-forward interpretability. For higher-order interactions, PID information atoms become more difficult to interpret, i.e. they mix the very notions of synergy and redundancy that PID is meant to disentangle [125], and for some measures intuitive sign behavior (i.e. non-negativity) will break down [96], as was the case in pre-PID measures of multivariate information such as [10].

Chapter 2

Modeling LSD1-Mediated Tumor Stagnation

Abstract

LSD1 (KDMA1) has gained attention in the last decade as a cancer biomarker and drug target. Recent work suggests that LSD1 inhibition reduces tumor growth, increases T cell tumor infiltration, and complements PD1/PDL1 checkpoint inhibitor therapy. In order to elucidate the immunogenic effects of LSD1 inhibition, we develop a mathematical model of tumor growth under the influence of the adaptive immune response. In particular, we investigate the anti-tumor cytotoxicity of LSD1-mediated T cell dynamics, in order to better understand the synergistic potential of LSD1 inhibition in combination immunotherapies, including checkpoint inhibitors. To that end, we formulate a nonspatial delay differential equation model, and fit to the B16 mouse model data from Sheng *et al.* [107]. Our results suggest that the immunogenic effect of LSD1 inhibition accelerates anti-tumor cytoxicity. However, cytotoxicity does not seem to account for the slower growth observed in LSD1 inhibited tumors, despite evidence suggesting immune-mediation of this effect. This chapter was previously published in [85].

2.1 Introduction

The gene coding for the histone lysine-specific demethylase LSD1 (KDMA1) has gained attention in the last decade as a cancer cell biomarker. It has been shown to mediate disease progression in multiple cancers, including acute myeloid leukemia [54, 90, 100] as well as carcinomas of the breast[75, 104],

liver [98], prostate, bladder, colon, and lung [55], among others. Thus, LSD1 has become a promising drug target. By suppressing LSD1 transcription in cancer cells, LSD1 inhibitors have demonstrated preclinical benefit, first in leukemia [44, 66] and more recently in carcinomas [130]. Moreover, LSD1 inhibitors have demonstrated benefit in combination therapies [14], including immunotherapies.

In particular, LSD1 inhibitors have shown preclinical potential in overcoming resistance to anti-PD1/PDL1 immune checkpoint inhibitors (ICIs) [31, 95, 107]. ICIs are among the most promising developments in cancer research of the past decade, as recognized by the 2018 Nobel Prize in Medicine. Despite this potential, the clinical reality is that typically up to 60% of patients show no response to single-agent ICI therapy [109, 129]. Overlapping factors contributing to resistance include a lack of T cells at the tumor site, immunosupressive mechanisms within the tumor microenvironment (TME), and tumor-intrinsic features that enable immunoescape [109]. Moreover, in clinical combination therapies, ICIs are frequently administered concurrently with other treatments, with little regard to the dynamics of the immune response [129]. Sheng *et al.* demonstrated that LSD1 inhibition induces a type 1 interferon response, increasing T cell infiltration into the TME (Fig. 2.1). By knocking out LSD1 *in vivo*, they were able to overcome the poor immunogenecity of the B16-F10 melanoma cell line, increasing tumor infiltrating lymphocyte counts (TILs) and sensitizing the tumors to anti-PD1 treatment [107]. Similarly, Qin *et al.* combined clinical LSD1 inhibitors with anti-PD1 treatment in xenograft models of triple negative breast cancer [95]. They likewise found that LSD1 inhibition overcame the resistance observed in anti-PD1 treatment alone.

Modeling the dynamics of tumor growth and therapeutic response is a

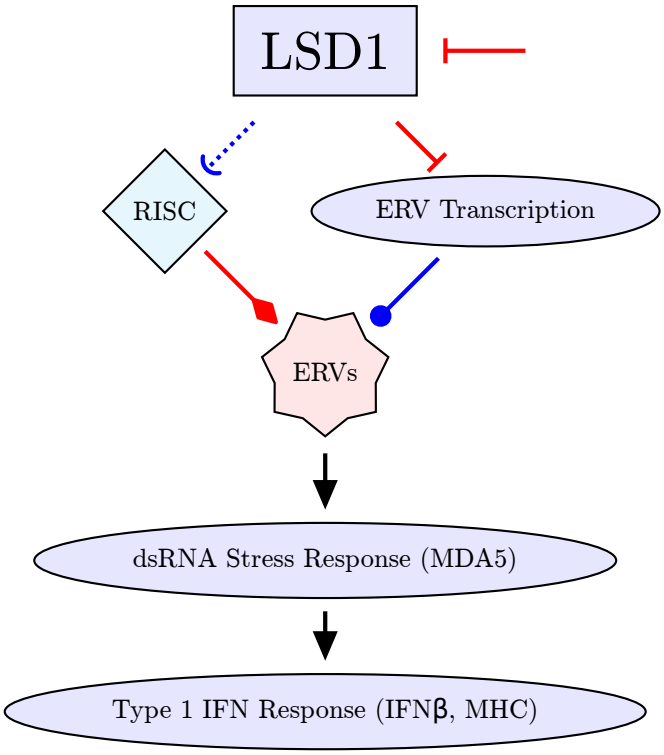


Figure 2.1: **The Proposed LSD1-interferon mechanism from [107].** LSD1 both suppresses transcription of endogenous retroviruses (ERVs) and regulates the RNA-induced silencing complex (RISC), which typically cleaves ERVs. Through both mechanisms, LSD1 inhibition enables ERV transcription and double-stranded RNA (dsRNA) accumulation. Through MDA5 and other sensors, dsRNA stress activates a type 1 interferon response in the cell, leading to tumor immunogenicity.

central focus of mathematical oncology, in which biological and biophysical knowledge is used to construct a formal model amenable to quantitative investigation. Among other possibilities, these models may offer a mechanistic explanation of existing data and generate new hypotheses, allow for the optimization of some experimental or clinical procedure, focus the scope or direction of future experiments, or inform a speculative, theoretical understanding of cancer-immune biology[40]. We refer to [40] for a recent review of mathematical models of immunology, and [39] for non-spatial modeling of tumor-immune dynamics. A broader overview of mathematical oncology can be found in [67, 126]. We briefly review the models that inspire our current work. Aligning with the data from [107], we specifically considered non-spatial

ODE models.

The classic model developed by de Pillis *et al.* in [35] simulates immune-mediated tumor-growth as the interaction of three populations: tumor cells, natural killer (NK) cells, and CD8+ cytotoxic T cells. With only three equations, this model is powerful in its versatility in capturing the dynamics of tumor growth under the influence of both innate and adaptive immune cytotoxicity. However, our focus in this work is on T cell dynamics specifically, since these are the populations most relevant to PD1/PDL1 checkpoint therapies and experimentally observed in [107]. Since the publication of the model [35] in 2005, much has been learned about the complex interplay between different CD4+ and CD8+ T cells populations. In particular, we wish to account for the complex management of T cell cytotoxicity by CD4+Foxp3+ regulatory T cells (Tregs).

Kim and Levy developed a model of the regulated adaptive immune response to antigen in [64] and [65]. Their model includes both naïve and mature compartments for antigen-presenting cells (APCs), CD8+ cytotoxic T cells, and CD4+ helper and regulatory T cells. A key feature of their model is the use of constant delay terms to account for the proliferative dynamics of T cells. The immune dynamics in our model are inspired by this work, although we exclude both APCs and naïve cell populations.

Part of our work is motivated by that of Gadhamsetty *et al.* in [47, 48]. In these works, they used cellular Potts models to investigate cytotoxic T cell killing dynamics. Gadhamsetty *et al.* analytically derived their killing term for simple, monogamous killing regimes, and demonstrated *in silico* that this function extends to joint and mixed killing regimes [48]. The precise way in which we use this work is explained in Section 2.2.2.

The model presented in this work simulates the regulated T cell response

to normal and LSD1-inhibited tumor growth, in order to further investigate the immunogenic and anti-tumor effects of LSD1 inhibition observed in [107]. This immunogenicity underlies the synergistic potential of LSD1-inhibitors combined with PD1/PDL1 ICIs.

The structure of this paper is as follows. In Section 2.2 we introduce our model, its underlying biological assumptions, the data we are using to fit it, and the alternative models considered. More technical detail is found in Appendix 2.B. The results of our modeling are presented in Section 2.3. Our work suggests that LSD1 inhibition accelerates the anti-tumor T cell response, but does not necessarily enhance T cell cytotoxicity. Rather, LSD1 inhibition seems to reduce tumor tumor through other immune-mediated mechanisms. We provide diagnostics and validation for our model in Section 2.4, comparing it favorably to simpler alternatives. We also explore our model's robustness to the removal of data points. Concluding remarks are provided in Section 2.5.

2.2 Model and Methods

2.2.1 Model Overview

We model T cell-mediated tumor growth as a system of delayed differential equations, representing cancer and immune cell populations within the tumor microenvironment (Fig. 2.2). Our model has five state variables, (C, H, K, R, P), corresponding to tumor cells, helper and cytotoxic "killer" T cells, regulatory T cells (Tregs), and pro-immune cytokine. The model equations are:

$$\dot{C} = \underbrace{aC(1-C/\mu)}_{\text{Intrinsic Tumor Growth}} - \underbrace{k\psi(C,K)}_{\text{Cytotoxicity}}, \tag{2.2.1a}$$

$$\dot{H} = \underbrace{2^{m_H} s_H C^{\sigma_H}}_{\text{Recruitment}} - \underbrace{k\pi(C,H) + 2k\pi(C^{\rho_H}, H^{\rho_H})}_{\text{Proliferation}} - \underbrace{(d_H + r)H}_{\text{Death+Differentiation}} - \underbrace{kRH}_{\text{Regulation}}, \tag{2.2.1b}$$

$$\dot{K} = \underbrace{2^{m_K} s_K C^{\sigma_K}}_{\text{Recruitment}} - \underbrace{kPK + 2kP^{\rho_K} K^{\rho_K}}_{\text{Proliferation}} - \underbrace{d_K K}_{\text{Death}} - \underbrace{kRK}_{\text{Regulation}}, \tag{2.2.1c}$$

$$\dot{R} = \underbrace{rH}_{\text{Differentiation}} \underbrace{-kPR + 2kR^{\rho_H} P^{\rho_H}}_{\text{Proliferation}} - \underbrace{d_H R}_{\text{death}}, \tag{2.2.1d}$$

$$\dot{P} = \underbrace{p_H H + p_K K}_{\text{Cytokine Secretion}} - \underbrace{d_P P}_{\text{Decay}} - \underbrace{kP(R+K)}_{\text{Consumption}}, \tag{2.2.1e}$$

$$\psi(C,K) = \frac{\ell CK}{C+K+1}, \tag{2.2.1f}$$

$$\pi(C,H) = \frac{CH}{C+H+1}. \tag{2.2.1g}$$

In all equations, the superscripts correspond to the delay notation:

$$X^{\delta} = X(t-\delta).$$

The dimensionality of the populations in our model reflects the data provided by [107], which measures tumor volume (in mm^3). Tumor volume is proportional to tumor population, and there is little benefit in estimating cell numbers in the absence of more immune data. Rather, for any of the immune quantities present, we assume a scale comparable to the tumor volume. The population scales should be understood as approximating an effective proportionality, rather than absolute cell counts.

Note that our model has only a single compartment for pro-immune cytokine signalling, following the example in [64]. This variable primarily models

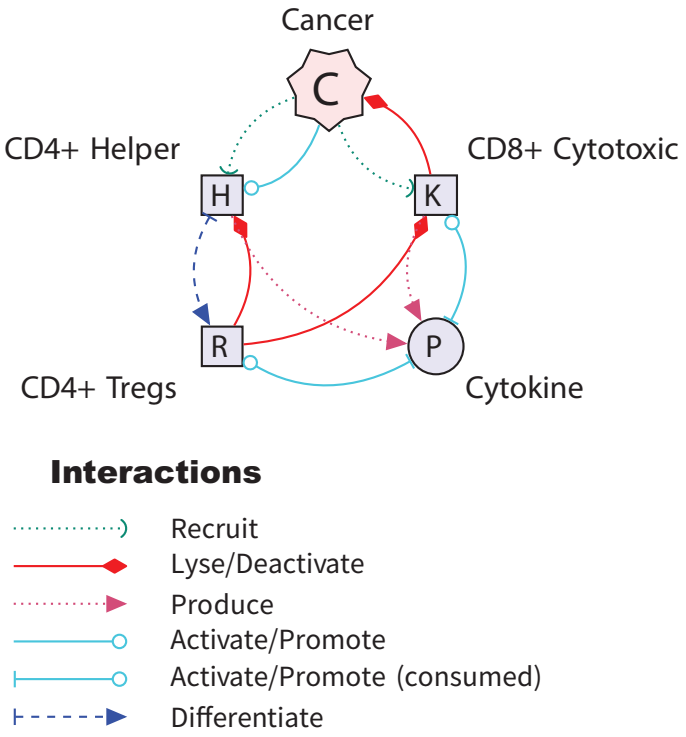


Figure 2.2: **Nonspatial population model of immunosurveilled tumor growth.** Our constant-delay ODE model of T cell-medaited tumor growth. The populations modeled are tumor cells (T), CD4+ helper T cells (H), CD8+ cytotoxic T cells (K), CD4+FOXP3+ regulatory T cells (R), and a simple pro-immune cytokine compartment (P).

the known functions of IL-2, e.g. from [83]. We acknowledge that many other pathways are involved in typical T cell dynamics. However, this simplification both limits model complexity while also allowing us to remain agnostic regarding the topology of extraneous signalling pathways.

Our construction of this model was not agnostic. Nonetheless, we compared it to simpler alternatives. We will consider the following three alternative models for tumor growth:

$$\left.\begin{aligned}\dot{C} &= \alpha C(1-\frac{C}{\mu})\end{aligned}\right\} \quad \text{Logistic Growth} \tag{2.2.2}$$

$$\left.\begin{aligned}\dot{C} &= \alpha C \log\left(\frac{\mu}{C}\right)\end{aligned}\right\} \quad \text{Gompertz Growth} \tag{2.2.3}$$

$$\left.\begin{aligned}\dot{C} &= \alpha C(1-\frac{C}{\mu}) - CI \\ \dot{I} &= \beta C + \gamma IC - \theta I\end{aligned}\right\} \quad \text{Two Compartment} \tag{2.2.4}$$

For our results in Section 2.3, we simulate model (2.2.1) by solving Eqs. (2.2.1a)-(2.2.1e) numerically. The parameters used for model (2.2.1) are given in Table 2.1. A detailed discussion of parameter fitting and sensitivity is included in Section 2.4 and Appendix 2.B.

2.2.2 Model Features and Assumptions

1. **Tumors intrinsically exhibit logistic growth.** The first term in Eq. (2.2.1a) models intrinsic tumor growth as logistic. We considered exponential, Gompertzian, and logistic forms of the growth term, and additionally considered a distinct linear death term for each. We fit these terms to the tumor growth data in immunodeficient (TCRα KO) mice, from [107], in order to get a sense of the growth inherent to the tumor independent of the immune dynamics we seek to model. While other immunosuppressive mechanisms may still be active in TCRα KO tumors, those are beyond the scope of our model. We found that the immunodeficient tumors were best modeled by logistic growth, with no distinct death term.

2. **Delayed T cell recruitment to tumor site following T cell development program.** In our model, both CD4+ helper and CD8+ cyto-

toxic T cells are recruited to the tumor microenvironment in proportion to the cancer population, represented in the first terms of Eqs. (2.2.1b) and (2.2.1b). The recruitment is delayed to account for the process of antigen-presenting cell (APC)-induced program of T cell development and proliferation, as developed in [64, 65]. In [64], antigen stimulation activates APC cell maturation. APCs, in turn, migrate to the lymph node to activate the primary adaptive response, which develops according to a program of minimal development followed by APC-dependent expansion. Since we do not have any time series data for immune populations, we do not model this full process, excluding both APCs and naïve T cells. Nonetheless, for each population $X = H$ or K, we disentangle the parameters s_X, encompassing naïve T cell availability and stimulation rate, from the expansion multipliers 2^{m_X}. Here, m_X is the fixed number of divisions in the T cell development program. Since the time delay σ_X also depends on the number of divisions m_X, it is desirable to allow direct manipulation of the length and magnitude of the T cell development program in the model independently of the other influences on supply and recruitment dynamics. For our current work, we fix m_X to the values from [64].

3. **Helper T cells proliferate in dual-saturated response to tumor.** The second and third terms of Eq. (2.2.1b) model CD4+ helper cell proliferation dynamics. Our model assumes that CD4+ helper cells proliferate in response to APCs in the TME, a process which we simplify to a more direct cell-tumor interaction function π similar to the lysis function ψ, discussed further below. This saturation in the proliferative term has the added benefit of mimicking acquired immune resistance, including mechanisms mediated by the PD1 and CTLA4 immune check-

points. Unlike CD8+ cells, IL-2 concentration does not seem to significantly modulate the proliferative TCR response in CD4+ cells [3]. Thus, in our model, the proliferative dynamics of CD4+ helper cells are uncoupled from the pro-immune signalling compartment. To account for proliferation time, the third term utilizes a constant delay ρ_H, set to 11/24, corresponding to an 11 hour cell cycle for CD4+ helper cells.

4. **Cytotoxic and Regulatory T cells proliferate in mass action response to pro-immune signalling.** The second and third terms in Eqs. (2.2.1c) and (2.2.1d) represent the proliferative dynamics of activated CD8+ and Treg populations. Our model assumes that, unlike CD4+ helper cells, CD8+ cytotoxic and CD4+ regulatory cells proliferate via cytokine signalling, according to a simplified mass action law. IL-2 modulates the proliferation of the activated CD8+ population, without significant dependency on further stimulation [3, 61]. By contrast, low concentrations of IL-2 instead promote the differentiation of memory phenotype in naïve CD8+ cells, observed both *in vivo* during viral infections [61] and *in vitro* during chimeric antigen-receptor (CAR) T cell expansion [59]. We acknowledge that despite many *in vitro* experiments suggesting the necessity of IL-2 for the expansion of the CD8+ cytotoxic cell response, it has been observed that there seem to be redundant mechanisms for CD8+ proliferation *in vivo* for IL-2R KO mice [61]. As for Tregs, IL-2 has been well-documented as essential for peripheral Treg function and expansion [30]. The induction of Tregs by low-dose IL-2 has emerged in recent years as a promising new treatment for autoimmune disease [131]. As an additional modeling benefit, the structure we use captures cytokine competition between CD8+ cells and Tregs as a distinct mechanism of immunosuppression, due to the fourth

and fifth terms in Eq. (2.2.1d). This is supported by the the work of Chinen *et al.* , which suggests that IL-2 competition is a significant component of Treg-driven control of CD8+ populations, but not CD4+ populations [30]. Further, by using an identical, non-saturated proliferative mechanism for CD8+ and Treg populations, by design allowing them to proliferate at a greater rate than CD4+ helper cells post tumorigenesis, our model dynamics align with the proliferative patterns observed experimentally in [107]. Sheng *et al.* found both Tregs and cytotoxic cells to be more proliferative than helper T cells. At day 14, they found that up to 70% of Tregs and 60% of CD8+ cells expressed the proliferative marker Ki67+, compared to only 30% of helper CD4+ cells. We assume CD4+ regulatory cells have the same 11-hour proliferation time as helper cells (ρ_H), while CD8+ cells have a proliferation time (ρ_K) of 8 hours [3].

5. **Helper and cytotoxic T cells produce inflammatory cytokine.** Our model assumes that helper T cells are primarily responsible for pro-immune signalling, although CD8+ cytotoxic cells also produce pro-inflammatory signals. Both populations produce cytokine at fixed linear rates p_H, p_K. For our present study, we fix these values to those from [64]. Thus, for the moment, the dynamics of cytokine signalling in our model is rigidly contingent on those of the other compartments. However, we expect the interferon response induced by LSD1 inhibition to alter the dynamics of the signalling compartment. We leave this for future work.

6. **Differentiation of helper T cells into Tregs.** In our model, CD4+ helper cells differentiate into CD4+ Tregs at a fixed rate r, similar to the Treg dynamics from [64]. Our model does not distinguish between

peripheral iTregs and thymal nTregs. We note the mechanisms of peripheral Treg differentiation are still unclear, as is the degree of plasticity between helper and regulatory CD4+ cell lineages [71].

7. **Decay of immune populations.** We assume that CD4+ helper and regulatory cells and CD8+ cytotoxic cells deactivate at fixed linear rates: d_H for both CD4+ populations and d_K for CD8+ populations. Further, positive growth signal decays at rate d_P. We take these rates from [64].

8. **Cytotoxic T cells lyse tumor cells with double-saturated sigmoidal dynamics.** The second term in Eq. (2.2.1a) represents anti-tumor cytoxicity from CD8+ T cells. Saturated kill terms for T cell cytotoxicity are standard in the literature [39]. more generally, sigmoidal functional forms are typical for immune response to antigen, and have been used, for instance, to faithfully model Potts-type lattice simulations of TCR-pMHC binding dynamics [136]. It is desirable to use a function that saturates with respect to both tumor and immune populations, as such a property allows the function to handle the dynamics of both tumor growth and collapse. We take our function from the work of Gadhamsetty *et al.* , which simulates T cell cytotoxicity in a Potts model framework [47, 48]. They heuristically derive a similar function when T cells follow a monogamous killing regime, and demonstrated *in silico* that the function extends to joint and mixed killing regimes [47].

9. **Kinetic coefficient.** All terms representing cell interactions in the TME are multiplied by a kinetic coefficient, k. Adjusting this coefficient affects the speed of population transitions in the transient immune dynamics. For our current study, we fix k to a constant value of 10.

Our model has several limitations worth highlighting. We excluded these

dynamics due to their practical unidentifiability in this current work. In particular, as discussed above, we excluded T cell exhaustion, checkpoint-mediated immune tolerance, the innate immune response, and myeloid cell dynamics.

Table 2.1: **Model parameters.** Parameter values replaced with an asterisk (*) were estimated individually for each mouse model.

	Parameter Name	**Description**	**Value**	**Reference**
a	Tumor growth rate	Controls tumor-intrinsic logistic growth	*	estimated
d_H	CD4+ death rate	Linear death rate for CD4+ helper and regulatory T cells	0.23	[64]
d_K	CD8+ death rate	Linear death rate for CD8+ cytotoxic T cells	0.4	[64]
d_P	Cytokine decay rate	Linear decay rate for IL-2	5.5	[64]
k	Kinetic Coefficient	Controls rate of immune interactions in the TME	10	fixed
ℓ	Immune-tumor lysis parameter	Controls CD8+ T cell cytotoxicity in kill function $\psi(C, K)$	*	estimated
μ	Tumor carrying capacity	Limits tumor-intrinsic logistic growth	*	estimated
m_H	CD4+ developmental divisions	Number of CD4+ cell divisions in APC-driven development program in lymph node	2	[64]
m_K	CD8+ developmental divisions	Number of CD8+ cell divisions in APC-driven development program in lymph node	7	[64]
p_H	CD4+ cytokine secretion	Controls production of IL-2 by CD4+ helper T cells	100	[64]
p_K	CD8+ cytokine secretion	Controls production of IL-2 by CD8+ T cells	1	[64]
r	Treg differentiation rate	Fractional rate at which CD4+ helper cells differentiate into Tregs	*	estimated
ρ_H	CD4+ division time	Length of cell cycle for proliferating CD4+ helper and regulatory T cells	11 hr	[34]

s_H	Supply rate of CD4+ cells	Controls delayed supply of CD4+ cells to TME in response to tumor antigen	*	estimated
s_K	Supply rate of CD8+ cytotoxic cells.	Controls delayed supply of CD8+ cells to TME in response to tumor antigen	*	estimated
σ_H	CD4+ development time	Length of APC-driven CD4+ T cell development program in lymph node (divisions $\times$ doubling time)	1.46 days	[64]
σ_K	CD4+ development time	Length of APC-driven CD4+ T cell development program in lymph node (divisions $\times$ doubling time)	3 days	[64]

2.2.3 Experimental Data

For data, we used three of the experimental data sets from [107]. First, our target for modeling was the experimental data set corresponding to [107, Fig. 5E] which measured the tumor growth in 28 individual B16 murine xenografts. These are divided into 4 experimental conditions of CRISPR gene silencing: LSD1 KO, MDA5 KO, LSD1+MDA5 DKO, and a scramble control ($N = 7, 7, 6, 8$, respectively). Note that, as described in Fig. 2.1, MDA5 is an important mediating component for the pro-immunity interferon response produced by LSD1 inhibition, and the mechanism of focus in [107]. Thus, our work, we are looking for consistent differences between the LSD1 KO condition and both the control and LSD1+MDA5 DKO conditions, which would implicate the LSD1-IFN axis. In addition to this target data set, we also make use of some of the growth data from [107, Fig. 5C] corresponding to scramble and LSD1 KO tumors within immunodeficient TCRα KO mice. Finally, in Fig. 2.4, we use the flow cytometry T cell counts from [107, Fig. 6A] in Fig. 2.4 in order to provide circumstantial evidence for the earlier onset of the immune response in LSD1 KO tumors.

2.3 Results

2.3.1 Immune Response Accounts for Interrupted Tumor Growth

Many of the mouse tumors from [107] have irregular growth in the second or third week, usually in the 10-15 day range. A similar pattern can be observed in other subcutaneous B16 models in [58, 69, 127]. After a week or two of tumorigenesis and steady growth, the tumor stagnates or even regresses for a few days. Then, in typical B16 tumors, growth resumes, often at an accelerated rate. By design, our model hypothesizes that we can account for this irregularity via the tumor's interaction with the primary adaptive immune response.

As described in Section 2.2 and Appendix 2.B, we opted to parameterize independently for each tumor growth time series. In most tumor-specific simulations of the immune response, a stereotypical script emerges, as seen in Fig 2.3. Both helper and cytotoxic T cells are recruited to the tumor site, although helper T cells are usually recruited more quickly. Helper T cells proliferate when stimulated by the tumor cells, and release pro-immune cytokines, including IL-2. These cytokines stimulate cytotoxic T cell proliferation, inducing significant cytotoxicity that interrupts steady tumor growth. Note that this pattern conforms to the B16 immune data from [93], which saw CD4+ helper cells peak a few days before CD8+ cytotoxic cells. The helper T cell population differentiates into Tregs at a fixed rate, and regulatory cells likewise proliferate in the presence of pro-immune signalling. The regulatory T cells deactivate both the helper and cytotoxic populations while consuming most of the remaining pro-immune signal for their own proliferation. Once Tregs dominate the immune populations, tumor stagnation ends and growth

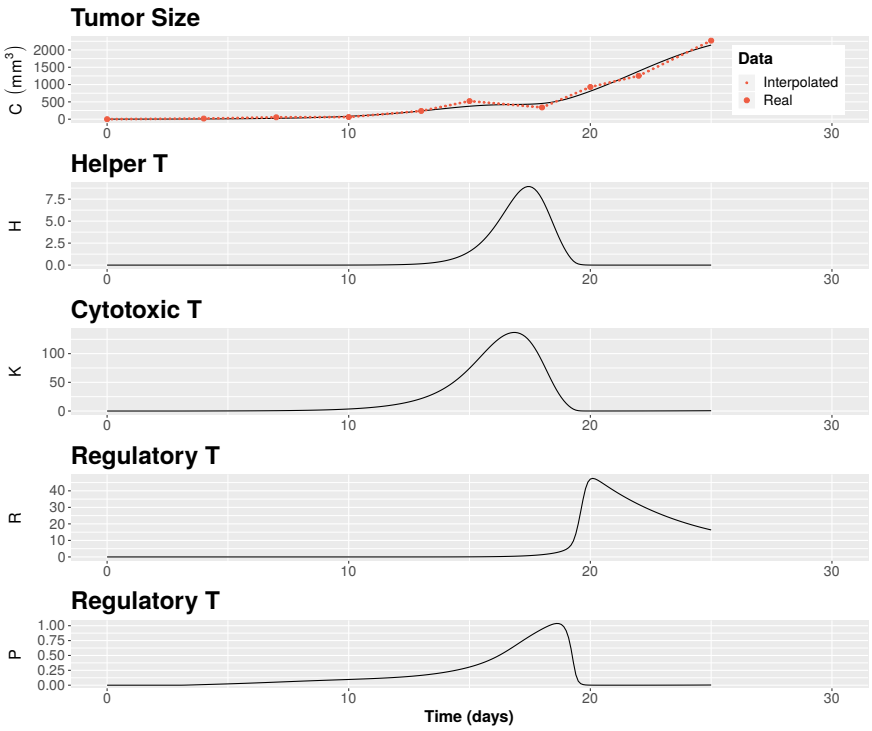


Figure 2.3: **Simulations of immune-mediated tumor growth.** Our model (black) infers an underlying adaptive immune response as responsible for temporary tumor growth stagnation, from linearly interpolated tumor growth data (red) from [107]. The graphs, from top to bottom, show simulated dynamics over 30 days for tumor volume, CD4+ helper T cells, CD8+ cytotoxic T cells, CD4+FOXP3+ regulatory T cells, and pro-immune cytokine.

resumes.

2.3.2 Simulated Immune Response Activates and Peaks Earlier in LSD1 KO Tumors

In our simulations, both the CD4+ helper and CD8+ cytotoxic cell populations tended to reach their peak earlier in the LSD1 KO condition, compared to both control and LSD1/MDA5 DKO tumors. On average, helper cells reached their maximal helper (cytotoxic) T cell population at day 14.01 (14.2), while control and DKO tumors peaked at day 17.11 (16.68) and 18.53 (17.62), respectively (Fig. 2.4, **A-B**). However, after adjusting for multiple testing, only the difference in timing between LDS1 KO and LSD1/MDA5 DKO tumors

rises to statistical significance in our small sample. Even so, this agrees well with a separate experiment from [107] under similar experimental conditions, in which flow cytometry revealed elevated levels CD4+ and CD8+ T cells in LSD1 KO tumors on day 14, relative to control and DKO tumors. (Fig. 2.4**D**, or Fig. 6**A** from [107]).

We emphasize that this difference appears only in LSD1 KO tumors where the dsRNA sensor MDA5 has not likewise been knocked out. Thus, our model suggests that the LSD1-dsRNA-IFN axis under investigation in [107] accelerates the anti-tumor T cell response.

Interestingly, our model not demonstrate an increase in effective cytotoxicity under LSD1 inhibition. At the time of peak immune response, we simulated no difference in the fractional kill rate ($C^{-1}\psi$ in our model, Eqs. 2.2.1a,2.2.1f), as seen in Fig. 2.4**D** and Fig. 2.5. This coincides with unpublished experimental evidence collected by Sheng et al. that LSD1 inhibition does not enhance the cytotoxicity of individual T cells. Moreover, as noted in [107], RNAseq revealed that PD-L1 was upregulated in LSD1 KO tumors, possibly suppressing the anti-tumor immunity of CD8+ T cells. This agreement between model and experiment furhter supports the notion that LSD1 inhibition does not directly enhance anti-tumor cytotoxicity. The significantly retarded tumor growth in LSD1 inhibited tumors, as observed in [107] and [95], must be attributed to other factors. Nonetheless, the process is likely to be immune-mediated, given the rescuing effect of MDA5 inhibition. Further, if LSD1-inhibition reduces T cell cytotoxicity through a mechanism besides the upregulation of PD-L1, then countering that mechanism and restoring full CD8+ functionality would maximize the effect of combination anti-PD1 therapy.

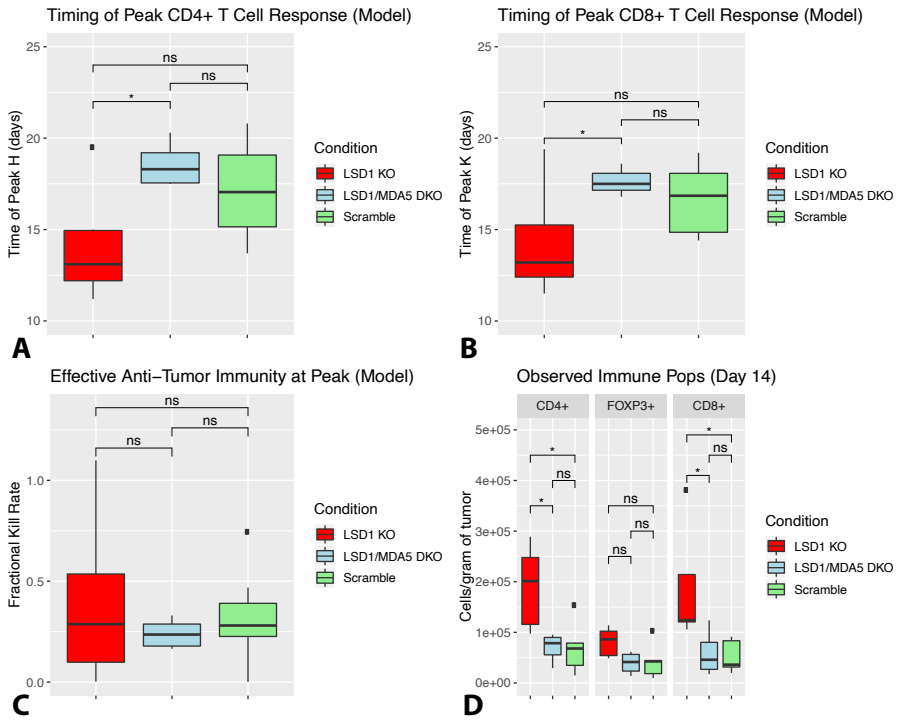


Figure 2.4: **Inferred dynamics of T cell response.** In subfigures (A) and (B), we present the time (days post-tumor injection) at which the CD4+ helper and CD8+ cytotoxic T cells reach their maximum population within our model, grouped by experimental condition.

2.3.3 Growth Dynamics

As noted in Section 2.3.2, the simulated cytotoxicity in our model suggests that LSD1 inhibited tumors see a quicker onset of the adaptive immune response, accounting for the increased number of tumor infiltrating lymphocytes. However, this does not convincingly account for the decreased tumor growth observed LSD1 KO tumors.

Beyond the immune dynamics incorporated into our model, LSD1 inhibition appears to slow tumor growth via other mechanisms. We see that, despite comparable distributions for the carrying capacity parameter μ, the tumor growth rate parameter α is lower for LSD1 KO tumors, compared to both control and LSD1/MDA5 DKO tumors. In our model, this difference manifests in early tumorigenesis, when tumor volume is well below capacity

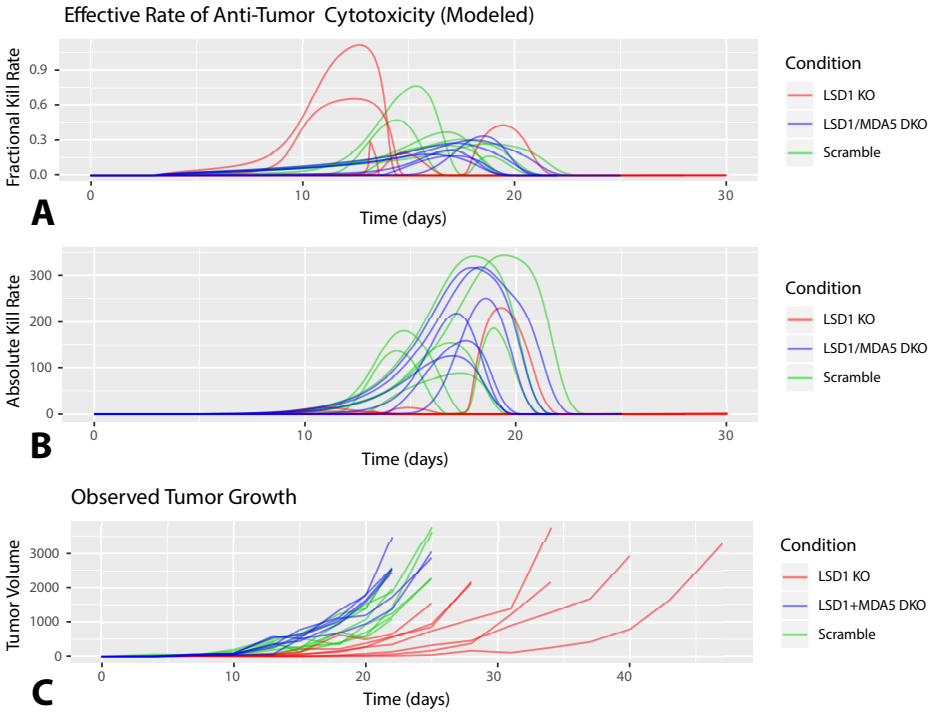


Figure 2.5: **Effective anti-tumor immunity in model.** In subfigures (**A**) and (**B**), we present the simulated fractional and absolute kill rates (in our model, $C^{-1}\psi(C,K)$ and $\psi(C,K)$, respectively) over time. We do not see any consistent and appreciable difference indicating that LSD1 inhibition increases anti-tumor immunity. For comparison, we include the observed tumor growth data from [107], which demonstrates an appreciable reduction of tumor-growth (**C**).

and growth is approximately exponential with rate α. Thus, when we remove T cell dynamics, our model still has LSD1 KO tumors growing more slowly *in silico* (see Fig. 2.6). Interestingly, this pattern does not appear in

the immunocompromised mice from [107]. When Sheng *et al.* used mice without functioning T cell receptor alpha (TCRα), they observed no difference in growth between LSD1 KO and control tumors. Similarly, when we fit the logistic growth model to the TCRα KO data (Fig. 2.6D, Table 2.3), we do not observe the difference captured in our T cell free model for immunocompetent mice.

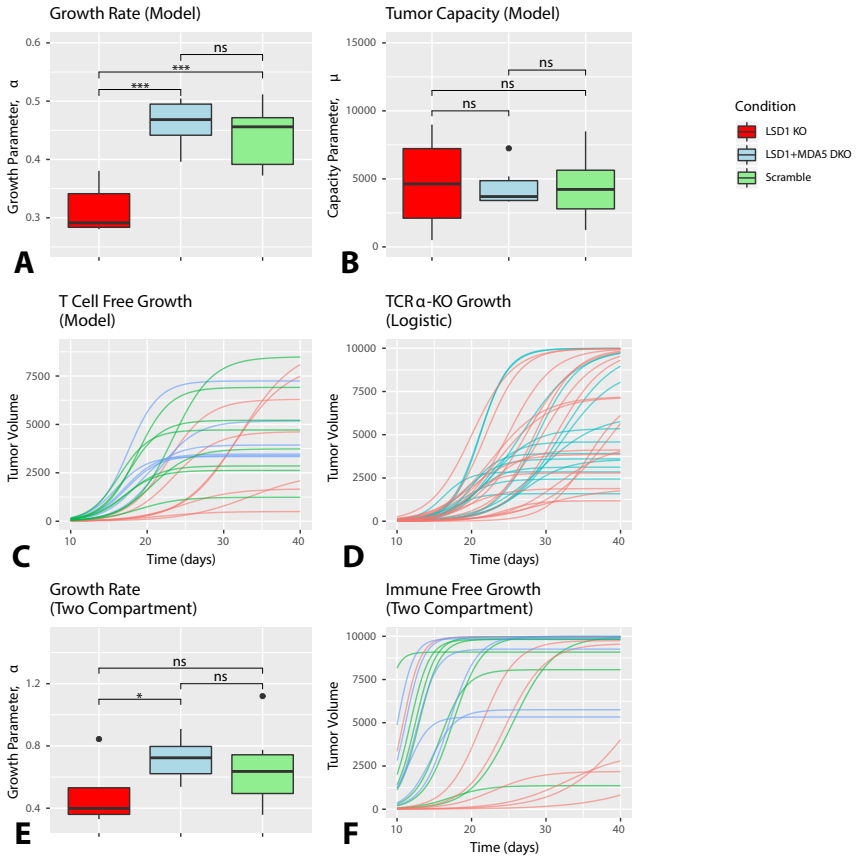


Figure 2.6: **T cell dynamics alone do not fully account for slowed tumor growth in LSD1 KO tumors.** We present the T cell-independent growth dynamics of our model. In the top two figures, we present the values of the model parameters for intrinsic growth (**A**) and tumor capacity (**B**). In (**C**), we simulate T cell free tumor growth by eliminating immune cell dynamics from our parameterized model (i.e. reducing our model to logistic growth with parameters a, M_C). We see that, despite comparable carrying capacities, the LSD1 KO tumors seem to exhibit slower 'intrinsic' tumor growth, separate from the model's T cell dynamics. This suggests that the anti-tumor effects of LSD1 inhibition include the innate immune response and/or tumor-intrinsic factors. Since this pattern is not observed in TCRα KO tumors (**D**, fit to simple logistic growth curves, see Table 2.3), the T cell response is still implicated as a mediator.

Assuming that our model accurately captures T cell dynamics, we consider two non-mutually-exclusive explanations for the 'intrinsic' reduction in tumor growth within our model. First, our system does not comprehensively model tumor-immune cytotoxicity. For instance, natural killer (NK) cells and memory T cells are not accounted for in our model. Another possibility is that tumor growth mechanisms themselves may be slowed by LSD1 inhibition, mediated by the MDA5-dsRNA-immune stress response.

It is possible that the reduction in the growth parameter is at least partially accounted for by the presence of continuous cytotoxic immunity not otherwise included in our model. Indeed, RNAseq analysis in [107] found the innate immune response to be upregulated. Innate cytotoxicty from, e.g., NK cells, is not included in our model. The nature of model (2.2.1) is such as to specifically capture the rapid-onset, well-regulated adaptive response to antigen. Although regulatory T cells do suppress NK cells in the TME[28], the timing for this will not align with the immune dynamics of our model. Moreover, T cell cytotoxicity itself is not limited to the sharp onset-regulation dynamic of our model. Our work emphasizes peak T cell response, at the cost of both ongoing and memory T cell dynamics, which are less clearly understood [128].

To explore this possibility, we examine the two compartment model (2.2.4), which, unlike our primary model, allows for sustained cytotoxicity. We do see that that the difference in the growth parameter α between LSD1 KO and control tumors is no longer significant in the two compartment model, despite a greater magnitude (Fig. 2.6E). Nonetheless, we have reason to suspect the ability of the two compartment model to capture T cell-independent tumor growth. Inspection reveals that the two compartment model has a tendency to estimate relatively high growth rates that do not, qualitatively, resemble tumor growth in TCRα KO mice (Fig. 2.6F). In particular, the two compartment

model tends to predict a higher rate of unencumbered tumor growth than we observe both when we fit a logistic growth model to immune-compromised (TCRα KO) mice, and when we fit our other models to the primary dataset of immunocompetent mice (Fig. 2.9).

We consider also tumor-intrinsic explanations for the reduced growth parameter. We considered EMT- and CSC-related pathways, looking at the RNAseq data from [107], and found no convincing evidence for them. Alternatively, we consider that a major limiting factor on growth is proper tumor angiogenesis. There is evidence that LSD1 regulates angiogenesis [62]. The RNAseq data from [107] suggests that the Notch pathway is activated in LSD1 KO tumors (Fig. 2.7). Notch regulates angiogenesis via the VEGF pathway, balancing tip and stalk cell populations in the fomration of new blood vessels. Its effect on tumor angiogenesis is context-dependent, e.g. [73] vs [103]. Recently, Augurt *et al.* found that LSD1 inhibitors reactivate the Notch pathway in small cell lung cancer, inhibiting tumor growth [4]. In the *in vivo* data collected by Sheng *et al.* , we see that the Notch activation in LSD1 KO tumors is less pronounced in LSD1/MDA5 DKO tumors. Moreover, Notch was not activated *in vitro*. This suggests the possibility that the LSD1-dsRNA-interferon axis induces Notch activation, which could restrain tumor growth. However, the significance of Notch activation to anti-tumor LSD1 inhibition remains hypothetical.

2.4 Model Fitting and Diagnostics

2.4.1 Model Fitting and Validation

For the DDE model described in Section 2.2.2, Eqs. (2.2.1a)–(2.2.1g), we adapted many of the parameters from [64], and fixed the kinetic coefficient k

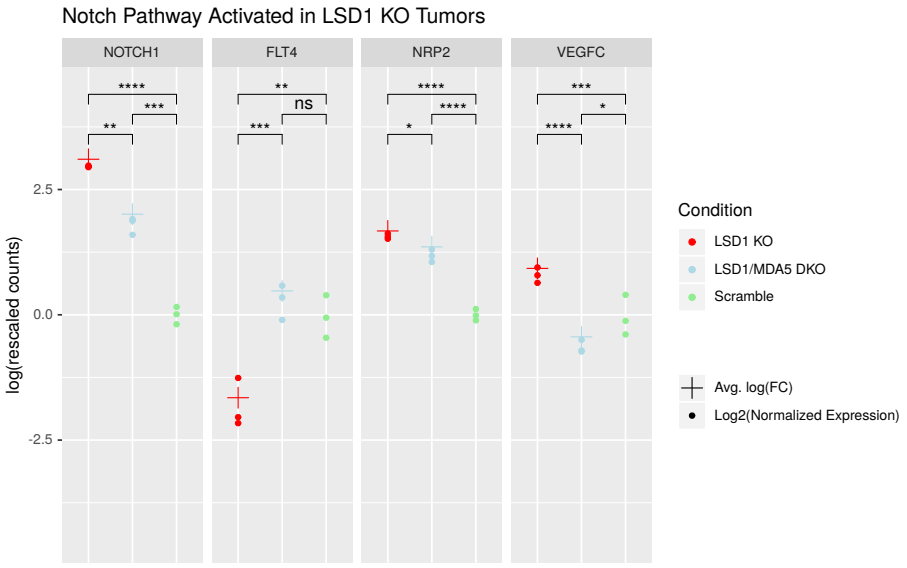


Figure 2.7: **Notch pathway upregulated *in vivo*.** NOTCH1 is upregulated *in vivo*, but not *in vitro* (not shown here), for LSD1 KO tumors, and VEGF-r4/FLT4 was downregulated. This could potentially affect angiogenic sprouting, as in [108, 112]. Dots represent pre-processed $\log_2$ counts from [107], rescaled by the average for the scramble control, and pluses represent the estimated average $\log_2$FC, after smoothing and normalization. We assessed differential gene expression among all significantly expressing genes, as described in Appendix 2.A

to an arbitrary value. Thus, for the purposes of model fitting and validation, we limit our degrees of freedom to six parameters: tumor growth parameters α, μ, immune recruitment parameters s_K, s_H, the cytotoxicity parameter ℓ, and the rate of Treg differentiation r. The full list of model parameters is given in Table 2.1.

For the one-dimensional models (2.2.2) and (2.2.3), we allowed our initial condition C_0 to vary as a third parameter. For our main model (2.2.1) and the two compartment model (2.2.4), we used a uniform initial condition for all tumors. All immune populations were initialized at 0. For the main DDE model (2.2.1), we assumed all state variables were 0 for $t < 0$.

For the one-dimensional curves, our models become the closed-form solutions of the logistic and Gompertz equations. For our main model (2.2.1) and the two compartment model (2.2.4), we solved our equations numerically.

We fit our model parameters using a Markov chain Monte Carlo, taking the point-estimates from the posterior distribution. We linearly interpolated our data prior to fitting. More technical details are provided in Appendix 2.B, and paramter estimates in Appendix 2.C. Figure 2.12 in Appendix 2.D provides a visualization of the prior and posterior distributions for one of our fittings. As can be seen, although the posteriors are sometimes multimodal, the model is locally identifiable.

We compare our model to the alternatives by computing the Bayesian Information Criterion (BIC) for each model, using the point-estimated parameters (Fig. 2.8). A lower BIC indicates a more explanatory and/or parsimonious model. Using a signed rank test, we found that our model was generally favored over all three of the alternatives (Fig. 2.8). Unsurprisingly, model (2.2.1) was heavily favored over the logistic and Gompertz models, with a mean BIC improvement of -40.2 and -40.3, respectively. More importantly, for 17 of the 28 tumors, model (2.2.1) was preferred to the two compartment model (2.2.4), albeit with a more modest mean difference of -3.4. For 5 tumors, the magnitude of BIC improvement of model (2.2.1) was greater than 10, while for only one tumor was (2.2.4) preferred by more than 4.

As an additional point of comparison, we consider the model-inferred rates of intrinsic tumor growth. Both models (2.2.1) and (2.2.4) assume that, in the absence of competitive immunity, tumor growth is logistic. Thus, parameterizing each involves estimating a rate of 'intrinsic' growth apart from the model's immune dynamic. To validate these estimates, we can use the tumor growth data from immunodeficient (TCRα KO) as a rough experimental proxy of immune-free growth. As previously discussed in Section 2.3.3 and Fig. 2.6F, we fit logistic growth to the TCRα KO mice tumors. In Fig. 2.9, we compare the growth rates of models (2.2.1), (2.2.2), and (2.2.4), fit to our

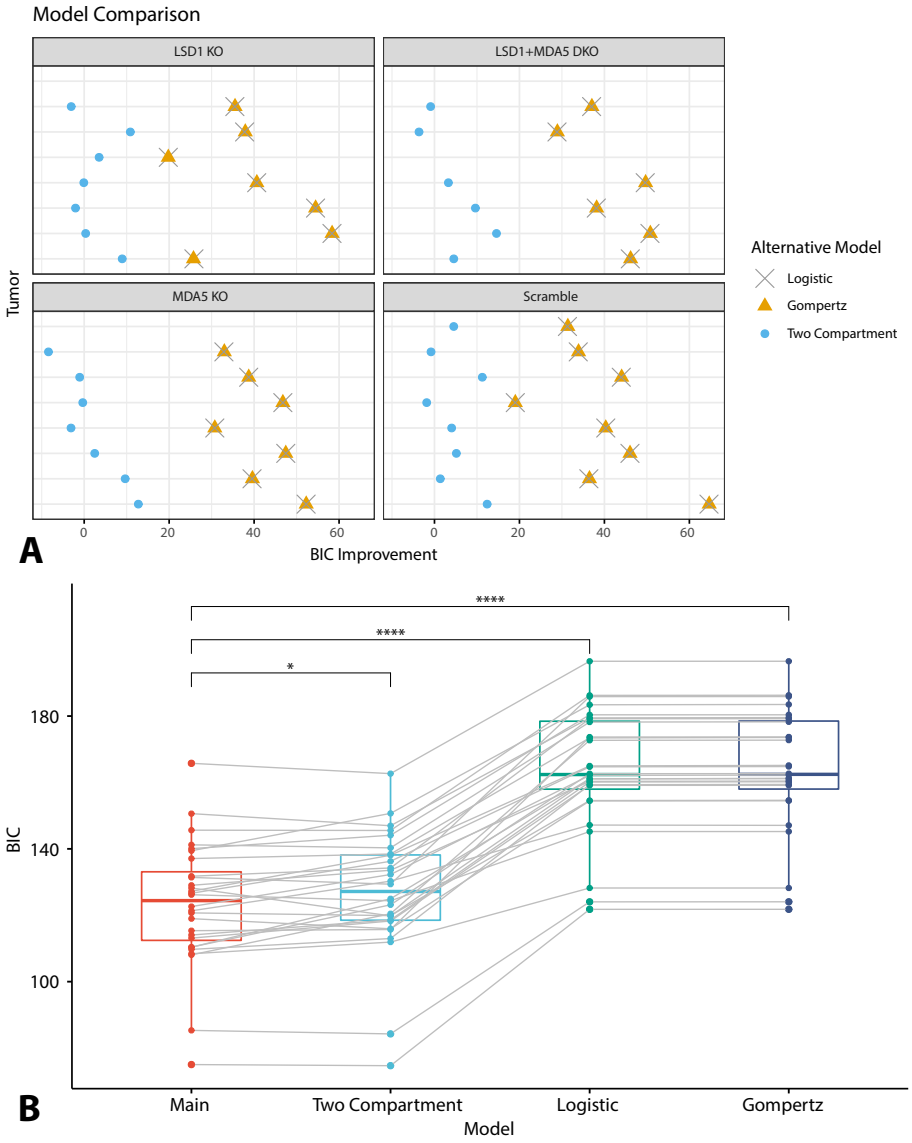


Figure 2.8: **Model comparison.** For each of the 28 tumors, we computed the the BIC to evaluate the relative performance of our model versus simpler alternatives. In **A**, for each tumor time series from [107], we compare our model (2.2.1) to each alternative, i.e. $\mathrm{BIC}_{\mathrm{Alt}} - \mathrm{BIC}_{\mathrm{Main}}$. In **B**, we compare the alternatives to (2.2.1) as paired populations, using the Wilcox signed rank test and adjusting for multiple testing. We see that our primary model significantly improves on the three alternatives, although the improvement is typically modest when we consider the two compartment alternative.

primary data set, to the growth rates estimated for the TCRα KO data. We see that the two compartment model tends to estimate much higher growth rates than those estimated for the immunodeficient control. By contrast, the primary model estimated growth rates comparable to those of the TCRα KO

tumors. That is to say, they were not statistically differentiable.

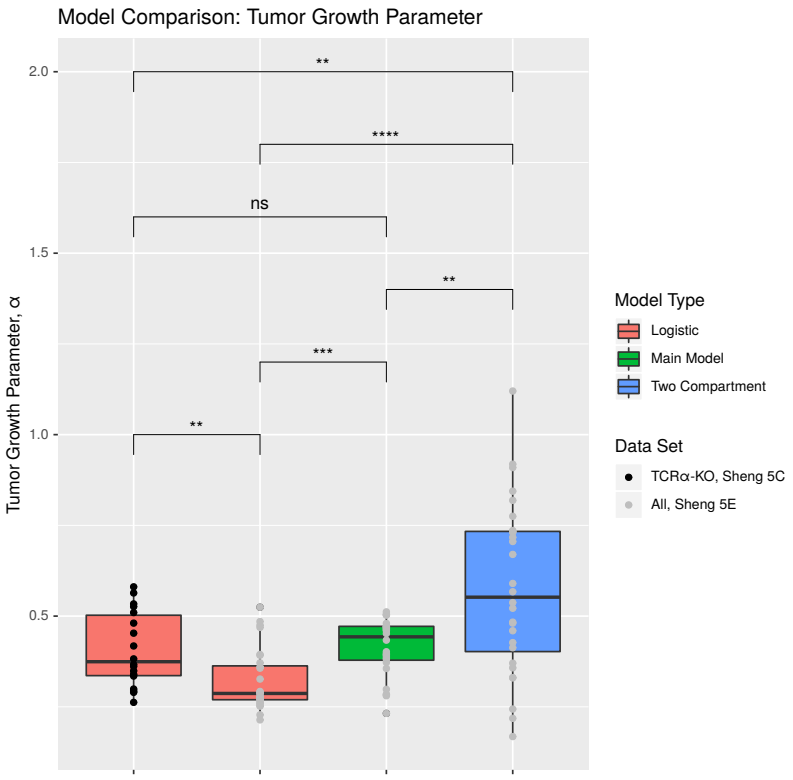


Figure 2.9: **Tumor growth rate estimates in immunodeficient and immunocompetent mice.** We compare the estimated value of the tumor growth parameter α for immunocompetent mice, using models (2.2.1), (2.2.2), and (2.2.4), to those rates estimated from the tumors in TCRα KO mice, using model (2.2.2).

2.4.2 Model Sensitivity (Data Removal)

We examined the robustness of our model fitting to the exclusion of key data points. To that end, we examined two specific, tumor growth time series from our data set: one from a control tumor and another from an LSD1 KO tumor. For each, we considered two modifications to the data.

First, for each time series, we removed an intermediate interval of data corresponding to a single irregular observation. We chose the removed observation to be one that suggests a period of particularly irregular growth and

stagnation. This allows us to consider the possibility that an interruption in growth, which our modeling attributes to T cell cytotoxicity, is instead due to a 'blip' of errorful measurement. This is key to the validity of our current work.

Second, for each time series, we removed the last two observation (and the associated interpolated points). This allows us to examine the forecasting potential of our model and its ability to recover the dynamics of immunoescape and tumor recovery from early growth data. Forecasting future growth was not the objective of our present modeling work. Nonetheless, it is an informative test of the limitations of our model.

In Fig. 2.10, we present the result of this first modification to our model fitting. For both data sets, we see that removing an interval of irregular growth does not affect the timing of the immune response. The intensity of the immune reaction decreases, somewhat, as a smoother, more regular trajectory is fit to the missing interval. This is seen with particular clarity in the LSD1 KO tumor data. The timing of the peak T cell reaction is robust to the removal of the observation at day 18, indicating that more than a single outlying observation is suggestive of significant tumor stagnation. Though the biological mechanism assumed by our model is hypothetical, our inference of immune dynamics seems robust.

In Fig. 2.11, we present the result of the second modification to model fitting, in which we truncate the final few days of tumor growth. The two data sets under consideration demonstrate qualitatively different sensitivities to this exclusion. For the scramble control tumor, the exclusion of the final data points leads to a fitting in which the final remaining data point is close to capacity. Thus, the projected future growth levels out. This a significant divergence from both the actual data and the original fit. In contrast, for the

LSD1 KO time series considered, the projected future growth still reasonably resembles future data. We additionally note that, for the control time series, the entire timing of the estimated immune response is shifted to an earlier window. For the LSD1 KO tumor, the timing of the immune response remains the same. Taken together, our model does not show itself to be a reliable tool for future growth forecasting.

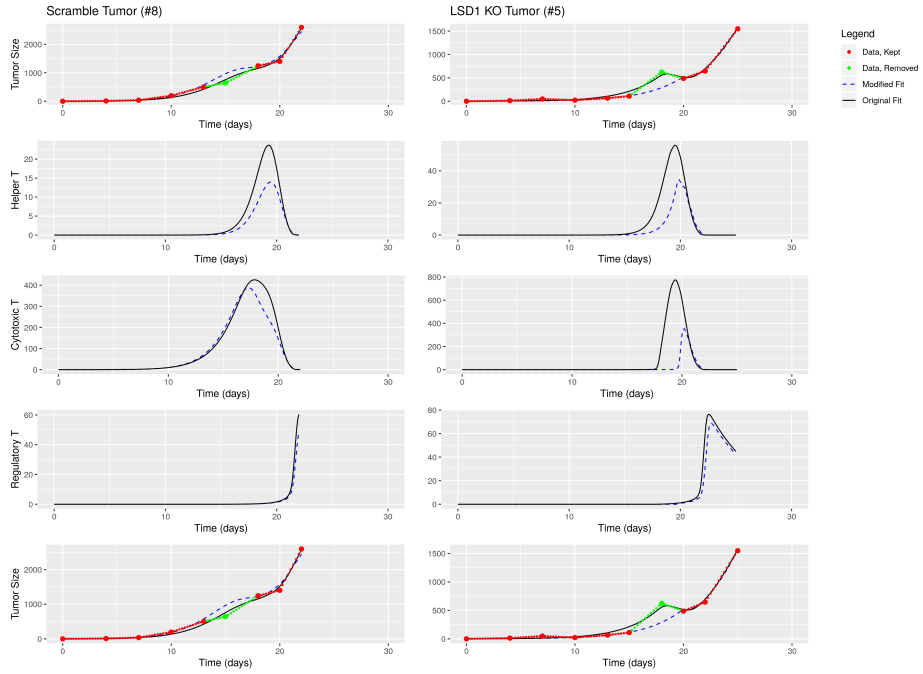


Figure 2.10: **Model refitting after data exclusion (intermediate interval).** We present the original and modified fittings of model (2.2.1) for two of our time series. For the modified fitting, we excluded one intermediate real observation and the associated interpolated points, shown in green to distinguish from the remaining (red) data points. For the scramble control (LSD1 KO) tumor, we excluded all points between days 13 and 18 (15 and 20).

2.5 Discussion

We created a simple mathematical model of the adaptive immune response to tumor growth in order to infer the potential effects of LSD1 inhibition on T cell dynamics. Our model suggests that LSD1 inhibition accelerates tumor

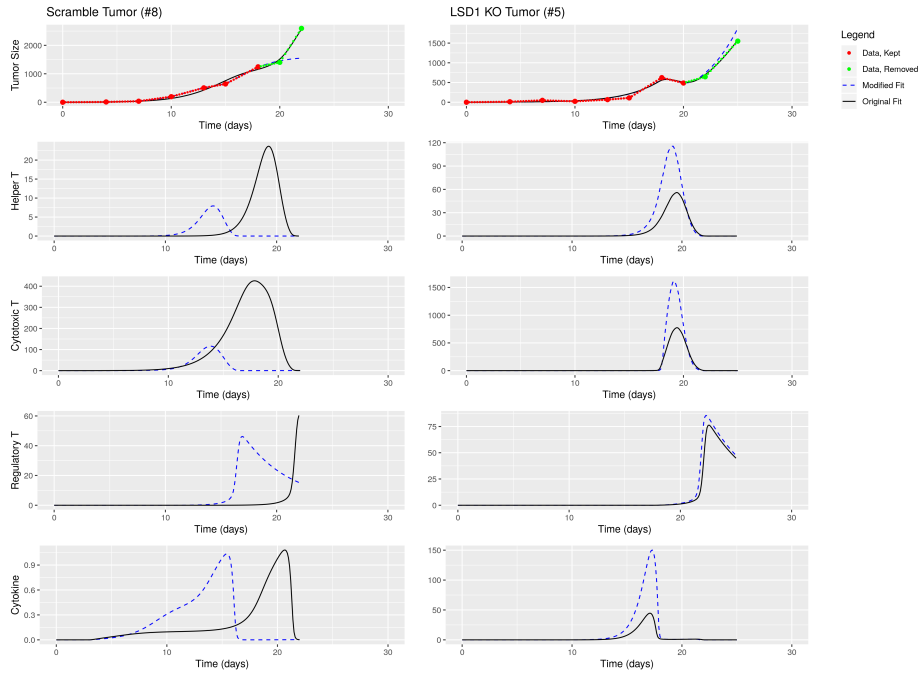


Figure 2.11: **Model refitting after data exclusion (final data).** We present the original and modified fittings of model (2.2.1) for two of our time series. For the modified fitting, we excluded the final two real observations and the associated interpolated points, shown in green to distinguish from the remaining (red) data points. For the scramble control (LSD1 KO) tumor, we excluded all points after day 18 (20).

infiltration of T cells, via the MDA5-mediated interferon response studied in [107]. We further found that, despite increased levels of tumor-infiltrating T cells observed at day 14 *in vivo*, our model does not imply that LSD1 inhibition alone increases the instantaneous rate of T cell cytotoxicity in TME. This suggests that the synergistic effect of combination anti-LSD1/anti-PD1 treatment observed *in vivo* in [95, 107] is not due to additive anti-tumor effects. Rather, our modeling suggests that, if there is any benefit of LSD1 inhibition to effective anti-tumor T cell cytotoxicity, it only occurs if PD1/PDL1 is likewise targeted. However, as previously observed in [107] and [95] and discussed in Section 2.3.3, we still have that LSD1 inhibition alone reduces tumor growth. Moreover, this reduction is eliminated when MDA5 is also knocked out, implicating the interferon response from [107] and, by extension, the immune system.

It is unclear why the increased level of T cells in LSD1-inhibited tumors may not translate to increased cytotoxicity. As we have noted, [107] found the PDL1 was upregulated in LSD1 KO tumors. This could explain part of the discrepancy, as PDL1 would deactivate T cells in the absence of anti-PD1/PDL1 ICIs. There are other plausible explanations. LSD1 inhibition both induces an IFN-β response and upregulates TGF-β. IFN-β is commonly used to reduce autoimmunity in multiple sclerosis [63, 133], inhibiting memory T cell activation. It is possible that the type 1 inteferon response sustaining T cell activity in the TME of LSD1 KO tumors is simultaneously suppressing elements of anti-tumor immunity. Alternatively, the upregulation of TGF-β may also reduce activation, proliferation, and/or cytotoxic function in CD8+ cells [82, 91]. Despite its potential to increase and sustain T cell infiltration, if LSD1 inhibition can simultaneously undermine anti-tumor immunity in the TME, the mechanism responsible needs to be identified for effective therapeutic targeting. Given two tumors with distinct immune profiles, it is entirely plausible that LSD1 inhibitors could sensitize one to checkpoint therapies, while disabling immunosurveillance in the other.

In future work, we intend to investigate and model the effect of LSD1 inhibition on many tumor-immune mechanisms currently excluded from the model. The effect of LSD1 inhibition on mechanisms of PD1-mediated immune tolerance is still unclear. Additionally, our current work does not consider myeloid cell dynamics. Recent work suggests that LSD1 promotes immunosuppressive myeloid cells, and that its inhibition reduces the differentiation of these populations in the TME. Condamine *et al.* found that LSD1 inhibitors reduced the differentiation of myeloid derived suppressor cells (MDSCs) and polymorphonuclear (PMN) cells *in vitro* and *in vivo*, synergistically enhancing the effect of anti-PDL1 agent [31]. With the appropriate data, we hope to later

explore the dynamic interplay between myeloid and T cell populations in the TME under the effect of LSD1 inhibitors, clarifying the tumor-immune signatures favorable to anti-LSD1/anti-PD1 combination therapy. Finally, more investigation is warranted into T cell exhaustion in LSD1 inhibited tumor-immune systems. While LSD1 inhibition increases T cell tumor infiltration, complementing anti-PD1 treatment in the short-term, we cannot discount the possibility that an LSD1-mediated interferon response might also accelerate other (non-PD1) forms of T cell exhaustion. This would have serious implications for the long-term efficacy of LSD1 inhibitors as compliments to immunotherapy.

Acknowledgements

We would like to thank Heyrim Cho and Asia Wyatt for helpful discussions in drafting this paper. The work of JM was supported in part by the COMBINE Fellowship under NSF award DGE-1632976. The work of DL was supported in part by the National Science Foundation under Grant Number DMS-1713109.

2.A Statistical Comparisons in Figures

For the pairwise comparisons in Figs. 2.4, 2.6, 2.8, and 2.9, we used an unpaired Student's t-test. For each panel, we applied a Benjamini-Hochberg adjustment for multiple-testing. In order to validate the chocie of a t-test, we used the Shapiro-Wilk test for normality. We cannot reject normality for the samples in Figs. 2.4, 2.6, and 2.8 ($P > 0.05$). For Fig. 2.9, two of the samples were somewhat non-normal. Left to right in Fig. 2.9, the S-W statistic had values $0.93, 0.85, 0.90, 0.98$ corresponding to $P = 0.17, 8.8e-4, 9, 0.010.81$. We used the same comparison for Fig. 2.4**D** as in [107].

For the differential gene expression (DGE) analysis for Fig. 2.7, we used the gene counts from [107] (data accessible at NCBI GEO database, accession GSE112230), and the R package `edgeR`. We normalized the counts using the TMM method, and removed minimally expressing genes, leaving us with 12305 genes remaining. We fit a linear model to compare gene expression between the three experimental tumor conditions: scramble control, LSD1 KO, and LSD1/MDA5 DKO tumors. Our workflow was based upon the tutorial in [32]. When adjusting for multiple testing, we used the BH method as before, and adjusted for all three pairwise contrasts (between our experimental conditions) for the full set of expressing genes, together.

2.B Description of Statistical Model and MCMC Fitting

Consider the tumor growth data for tumor i as a time series $\boldsymbol{y}_i = (y_i(t_j))_{j=1}^{n_i}$. We assume a statistical model of the form

$$y_i(t) = f(\boldsymbol{p}_i, t) + \epsilon_i(t) \tag{2.B.1}$$

$$\epsilon_i(t_1), ..., \epsilon_i(t_n) \sim_{\text{iid}} N(\mu_i, \sigma_i) \tag{2.B.2}$$

where $f(\boldsymbol{p}_i, \cdot)$ is a deterministic model and $\epsilon_i(\cdot)$ is the measurement noise, parameterized by $\boldsymbol{p}_i, \mu_i, \sigma_i$ individually for each tumor time series i. The model $f(\boldsymbol{p}_i, \cdot)$ is the solution to either our main DDE model (2.2.1) described in Section 2.2.1, (2.2.1), or one of the alternative ODE models in Section 2.2.1 (2.2.2, 2.2.3, 2.2.4). We use $\boldsymbol{f}(\boldsymbol{p}_i)$ to denote $(f(\boldsymbol{p}_i, t_j))_j$, i.e. the estimated time series from our model, corresponding the data $\boldsymbol{y}_i$. For our measurement noise, we ideally would have $\mu_i = 0$ for each tumor. We make this assumption for model fitting, but estimate μ_i for the purposes of model validation below, in order to strengthen the likelihood of simple alternative models.

Given a model f and parameters $\boldsymbol{p}_i, \mu_i, \sigma_i$, the conditional log-likelihood is given by

$$\begin{aligned}\log L(\boldsymbol{y}_i \mid f, \boldsymbol{p}_i, \mu_i, \sigma_i) = &- n_i \log(\sqrt{2\pi}\sigma_i) \\ &- \frac{1}{2}\sum_{j=1}^{n_i}\left(\frac{y_i(t_j) - f(\boldsymbol{p}_i, t_j) - \mu_i}{\sigma_i}\right)^2\end{aligned} \tag{2.B.3}$$

For the purposes of model fitting, we estimated $\hat{\boldsymbol{p}}_i$ for fixed σ_i using a Markov chain Monte Carlo,[1] under the assumption that $\mu_i = 0$. We linearly interpolated our data $\boldsymbol{y_i} \mapsto \tilde{\boldsymbol{y}}_{\boldsymbol{i}}$ so that we had 5 data points per day, in order to ensure smooth fits. Per standard practice, we employ the ℓ_2 error as the target function g, for which $e^g \propto L$:

$$\begin{aligned}g(\tilde{\boldsymbol{y}}_{\boldsymbol{i}} \mid f, \hat{\boldsymbol{p}}_i, \mu_i, \sigma_i) &= -||\tilde{\boldsymbol{y}}_{\boldsymbol{i}} - \boldsymbol{f}(\hat{\boldsymbol{p}}_i)||^2_{\ell^2} \\ &= -\sum_{j=1}^{n_i}\left(\tilde{y}_i(t_j) - f(\hat{\boldsymbol{p}}_i, t_j)\right)^2\end{aligned} \tag{2.B.4}$$

To find $\boldsymbol{f}(\boldsymbol{p}_i)$, we need to solve systems (2.2.1, 2.2.2, 2.2.3, 2.2.4). The one-dimensional systems 2.2.2, 2.2.3 have well-known closed-form solutions:

$$C(t) = \mu \exp\left[e^{-\alpha t}\log\left(\frac{C_0}{\mu}\right)\right] \tag{2.B.5}$$

$$C(t) = \frac{\mu C_0}{C_0 + (\mu - C_0)e^{-\alpha t}} \tag{2.B.6}$$

For our main model (2.2.1) and the two compartment model (2.2.4), we solved our equations numerically using the R package `diffeqr`, which is a convenient wrapper for the Julia suite `DifferentialEquations.jl`.

To validate our model, we compared it to each of the alternative choices of

[1] We use the MCMC implementation from the R package `BayesianTools`, employing the differential evolution sampler 'DEzs'.

f, for each tumor i, using the standard Bayesian Information Criterion (BIC):

$$BIC_i = N_p \log(n_i) - 2 \log \hat{L} \tag{2.B.7}$$

where N_p is the number of free parameters for our model, i.e. the length of $(\boldsymbol{p}_i, \sigma_i)$. In particular, these are $7, 4, 4, 6$ for models (2.2.1),(2.2.2),(2.2.3), and (2.2.4), respectively. Substituting (2.B.3) into (2.B.7) and using our estimate $\boldsymbol{p}_i$, we have

$$BIC_i = N_p \log(n_i) + 2n_i \log \sqrt{2\pi}\hat{\sigma}_i + \frac{1}{\hat{\sigma}_i^2} ||\boldsymbol{y_i} - \boldsymbol{f}(\hat{\boldsymbol{p}}_i)||_{\ell^2}^2 \tag{2.B.8}$$

$$BIC_i = N_p \log(n_i) + 2n_i \log(\hat{\sigma}_i) + n_i(1 + \log(2\pi)) \tag{2.B.9}$$

$$\text{where } \hat{\sigma}_i^2 = \frac{||\boldsymbol{y_i} - \boldsymbol{f}(\hat{\boldsymbol{p}}_i)||_{\ell^2}^2}{n_i} \tag{2.B.10}$$

2.C Supplementary Tables

Table 2.2: **Estimated model parameter values.** The estimated values of the (non-fixed) model parameters for each mouse tumor (from the experiment for Fig. 5E in [107]), based on tumor volume time series data.

Tumor	α	μ	ℓ	r	s_H	s_K
Scramble 1	4.60E-01	2.62E+03	1.50E-01	5.27E-02	1.50E-07	5.13E-04
Scramble 2	4.69E-01	6.91E+03	1.43E-01	1.83E-02	3.17E-06	2.74E-04
Scramble 3	4.52E-01	2.86E+03	1.45E-01	2.04E-02	3.94E-06	4.20E-05
Scramble 4	3.94E-01	3.74E+03	2.69E-02	4.79E-04	4.72E-06	1.67E-05
Scramble 5	4.80E-01	5.21E+03	7.77E-02	1.49E-02	1.01E-08	9.99E-04
Scramble 6	3.85E-01	8.49E+03	1.21E-01	1.10E-02	2.22E-08	1.89E-06
Scramble 7	3.72E-01	1.24E+03	9.75E-03	9.07E-02	5.35E-04	1.36E-08
Scramble 8	5.12E-01	4.71E+03	1.10E-01	2.26E-02	3.94E-08	9.59E-04
LSD1 KO 1	2.98E-01	1.67E+03	5.20E-02	7.06E-02	1.03E-06	2.80E-05
LSD1 KO 2	2.81E-01	8.99E+03	1.89E-02	6.80E-02	1.07E-04	1.12E-08
LSD1 KO 3	2.32E-01	2.55E+03	3.99E-02	7.95E-02	7.51E-05	4.99E-06
LSD1 KO 4	3.80E-01	6.30E+03	1.22E-01	7.92E-02	3.03E-05	9.23E-04
LSD1 KO 5	3.56E-01	4.63E+03	7.19E-02	1.73E-03	5.54E-08	1.74E-06
LSD1 KO 6	2.83E-01	8.16E+03	1.02E-03	2.77E-03	6.24E-04	2.56E-04
LSD1 KO 7	2.84E-01	5.03E+02	6.68E-02	1.56E-02	7.29E-05	3.13E-04
LSD1 MDA5 DKO 1	4.65E-01	3.39E+03	1.17E-01	2.38E-02	2.05E-07	1.71E-04
LSD1 MDA5 DKO 2	4.72E-01	3.47E+03	7.70E-02	4.00E-02	1.61E-07	9.65E-04
LSD1 MDA5 DKO 3	3.96E-01	5.18E+03	1.34E-01	1.14E-02	4.86E-08	1.32E-05
LSD1 MDA5 DKO 4	4.34E-01	3.94E+03	2.48E-02	9.29E-04	1.44E-06	8.81E-06
LSD1 MDA5 DKO 5	5.04E-01	3.34E+03	9.27E-02	2.71E-02	1.04E-08	9.64E-04
LSD1 MDA5 DKO 6	5.03E-01	7.25E+03	1.11E-01	2.76E-02	3.21E-08	9.55E-04

Table 2.3: **Logistic growth for immune deficient tumors.** To account for the possibility of immune-independent mechanisms of inhibited growth in LSD1 KO tumors, we estimated the logistic growth model (2.2.2) for TCRα KO and TCRα/LSD1 DKO tumors (from Fig. 5C in [107]). We compare the means of parameters α, μ, and C_0 between the two conditions via unpaired t-test. Even pre-FDR adjustment, we find no significant difference between μ and C_0 of our samples ($t = 0.73, 1.36$, and $P = 0.476, 0.193$, respectively), and marginal evidence that the LSD1 KO condition has a *higher* immune-deficient growth rate than the control scramble ($t = -2.35$, $P = 0.038$). This latter difference vanishes after adjustment. Thus, for immunodeficient mice, we see no evidence for any anti-tumor effects of LSD1 inhibition.

Tumor	$\boldsymbol{\alpha}$	$\boldsymbol{\mu}$	$\boldsymbol{C_0}$
Scramble TCRα KO 1	0.563743699	3912.194739	0.052569707
Scramble TCRα KO 2	0.480629132	4581.405315	0.137067466
Scramble TCRα KO 3	0.533601886	9990.79188	0.124822197
Scramble TCRα KO 4	0.262316687	9999.356249	9.160142923
Scramble TCRα KO 5	0.452943779	2437.152773	0.551289523
Scramble TCRα KO 6	0.349057835	5360.823555	1.827053661
Scramble TCRα KO 7	0.509650981	9987.564364	0.214161265
Scramble TCRα KO 8	0.580677922	2852.511648	0.229292158
LSD1 KO TCRα KO 1	0.334763587	9978.038937	2.40658147
LSD1 KO TCRα KO 2	0.289841785	7179.317676	7.514285629
LSD1 KO TCRα KO 3	0.339625003	3870.760716	4.572794932
LSD1 KO TCRα KO 4	0.299030515	9999.841561	2.850807521
LSD1 KO TCRα KO 5	0.36148731	9979.685307	7.152464226
LSD1 KO TCRα KO 6	0.366512912	9979.305856	0.294670011
LSD1 KO TCRα KO 7	0.293537226	7208.25365	8.984451724
LSD1 KO TCRα KO 8	0.382124921	1886.942171	1.05053347
LSD1 KO TCRα KO 9	0.526204655	2787.05914	0.088720404
LSD1 KO TCRα KO 10	0.417661279	9996.276858	1.062130233
Scramble TCRα (Avg)	0.46657774	6140.225065	1.537049863
LSD1 KO TCRα (Avg)	0.361078919	7286.548187	3.597743962

2.D Marginal Parameter Densities

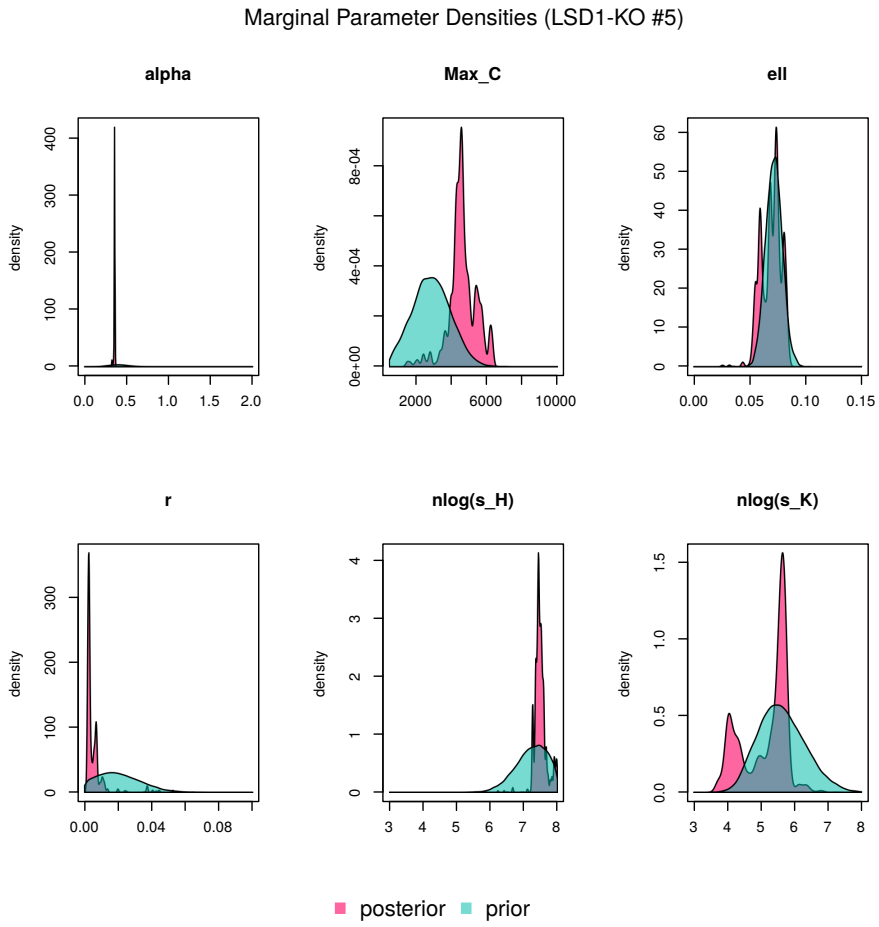


Figure 2.12: **Prior and posterior parameter distribution for main model fitting (LSD1 KO #5).** Presented are the prior and posterior distribution of the parameters $\boldsymbol{p}_i$ of model 2.2.1, taken from the MCMC, where i is one of the LSD1 KO tumors. In order from left to right, the panels here correspond to $\alpha, \mu, \ell, r, \log_{10}(s_H), \log_{10}(s_K)$.

Chapter 3

Signed and Unsigned Partial Information Decompositions of Continuous Network Interactions

Abstract

Motivated by the task of drug-specific network inference in cancer biology, we investigate the potential of the partial information decomposition (PID) framework as a tool for network inference and edge nomination. In contrast to previous metrics of 'synergy' based upon interaction information, which frames synergistic and redundant information as opposing quantities, the PID framework offers a means of disentangling the two. Thus, PID may be able to identify pairs of expressing genes that affect a response synergistically, even in the presence of redundant information due to network dependencies.

To that end, we conduct both numeric and analytic investigations of the $I_\cap^{\min}$ and $I_\cap^{\mathrm{PM}}$ PIDs, from [125] and [45], respectively. In the course of our work, we extend the $I_\cap^{\mathrm{PM}}$ and $I_\cap^{\min}$ PIDs to continuous variables for a general class of noise-free trivariate systems. We examine how each PID apportions information into redundant, synergistic, and unique information atoms within the source-bivariate framework. Although the synergy of the $I_\cap^{\mathrm{PM}}$ PID is quite sensitive to interactions, both our simulation experiments and analytic inquiry uncover that $I_\cap^{\mathrm{PM}}$ is non-specific, and cannot distinguish interacting pairs from univariate signals. By contrast, the $I_\cap^{\min}$ PID is quite specific, perhaps to a fault, as it is not sensitive to interacting gene pairs when multiple interactions determine the network response. We see that the $I_\cap^{\mathrm{PM}}$ PID does not respect conditional independence, while $I_\cap^{\min}$ does, demonstrated through

asymptotic analysis of linear and non-linear interaction kernels.

The main technical contribution of our paper is the work extending the $I_\cap^{\min}$ and $I_\cap^{\mathrm{PM}}$ PIDs to noise-free interactions of the form $T = g(X, Y)$ for a well-behaved kernel g and jointly Gaussian predictors X and Y. We provide straight-forward computations of the $I_\cap^{\min}$ and $I_\cap^{\mathrm{PM}}$ PIDs for noise-free linear interactions, which serve as an intuitive foundation. We then demonstrate a general form of the unique information, for each $I_\cap^{\min}$ and $I_\cap^{\mathrm{PM}}$, for arbitrary kernel g. This form is the expectation of a function that incorporates ratios of the partial derivatives of g. In this way, we are able to connect the analytic and information-theoretic behavior of our trivariate systems. We conclude with an application of our framework to a non-linear sigmoidal switch kernel, employed also in our network simulations.

3.1 Introduction

The development of high-throughput molecular assay technologies brought molecular biology into the era of 'big data', providing high-dimensional multiomic measurements of cellular samples. Network science has emerged as a core framework through which this data may be organized in a comprehensive vision of cellular biology and disease pathology [70, 89]. The inference of molecular networks from multiomics data has become a common method of investigation. The emergence of single-cell sequencing technologies in the past decade has been especially fruitful, as this allows investigators to account for biological heterogeneity within sample. However, network inference methods often grapple with unfavorable statistical power and noisy data.

Although less common than other methods, information theoretic tools have been widely applied to gene network analysis. Information-theoretic

methods are non-parametric and well-suited to capturing complex, non-linear dependency between variables. Mutual information (MI) has commonly been employed in place of Pearson correlation to create association networks ([18, 19, 25]). However, much like Pearson correlation and other crude pairwise methods, MI cannot distinguish between direct and indirect associations. In gene regulatory networks (GRNs) and protein-protein interaction (PPI) networks, we expect a highly interdependent structure, in which indirect associations are not immediately distinguishable from direct via any pairwise comparison. Thus, these methods have an unacceptably high false discovery rate; any networks inferred with them will include too many spurious associations for the biologically meaningful ones to stand out [110]. Thus, the next generation of information methods moved beyond MI in two keys ways. First, they attempted to remove indirect associations by accounting for redundancy [116], e.g. [74, 80, 84, 110, 118, 134, 135]. Second, attempts were made to specifically identify interacting genes by quantifying synergistic information that can be acquired from neither gene alone [2, 123, 124].

Meanwhile, an initially parallel development within information theory has emerged in the past two decades that addresses both concerns. First the neuroscience ([15, 49]) and later the more general network communities rediscovered the formalism of interaction information, whose first known use was by McGill in [81] (see also [10, 117]). This quantity, initially termed synergy in the networks literature [2, 15, 49], is a signed extension of MI, distinguishing synergistic and redundant information as opposites. A positive interaction information indicates synergy, while a negative value indicates redundancy. However, this quantity assumes that redundancy and synergy are opposing possibilities, and provides no framework for the possibility of the simultaneous presence of redundant and synergistic information between predictor variables [116]. This

would be a major concern in any highly interdependent network, such as a GRN or the human brain, where we might expect a large amount of synergy and redundancy to each occur in distinct contexts for the same pair of nodes.

This lead into the development of the Partial Information Decomposition (PID) framework, in order to addresses this limitation. Disentangling the synergy and redundancy of interaction information into distinct atoms [125], PID is also able to highlight the presence of unique information atoms, which would normally be cancelled out in the computation of interaction information. For the purposes of network inference, the PID framework offers the possibility of controlling the redundancy in interdependent gene expression data without neglecting the useful information remaining, be it unique to a single predictor or synergistic among multiple. The PID framework was met with much enthusiasm within a community of researchers meeting at the intersection of information theory, networks, complex systems, and neuroscience. However, there is still dispute and no clear consensus over the correct definition of 'redundancy', from which all other PID atoms follow. Multiple alternatives to the original redundancy measure(function? metric?) $I_\cap^{\min}$ of [125] have been proposed (e.g. [12, 45, 53, 56], see [77] for others). It seems likely that there may not be any definition of redundancy that will be universally appropriate to all or even most applications.

The PID framework offers a clarification of the confused concepts of synergy and redundancy within information theory and theoretical(computational?) neuroscience [116]. It took a few years for this work to migrate into the gene networks literature. The PIDC method for single-cell GRN inference ([24]) was the first to apply the PID paradigm to the task of gene network inference (see also [22, 23]), and has inspired [20]. This work used the unique information atoms of the original $I_\cap^{\min}$ PID (see Definition 14) to remove redundant

information from edge inference. The PID framework has not seen much use, however, in the building of synergistic gene networks, similar to the interaction information-based method of Watkinson *et al.* in [124]. Unlike the networks inferred by [24], a synergy network is specifically interested in the network by which genes create a particular biological response or phenotype. For instance, Watkinson *et al.* considered a Boolean cancer/not-cancer variable [124]. In [27], Chatterjee *et al.* construct synergy networks using a definition of synergy identical to that of the minimal mutual information (MMI) PID of [6], which under appropriate conditions encapsulates many non-negative PIDs including the original of [125]. To our knowledge, no previous work has examined the potential of the $I_\cap^{\mathrm{PM}}$ PID to construct response-specific synergy networks, transcriptomic or otherwise.

Moreover, there is still relatively little research into the effective use of PID for continuous variables. Part of the original appeal of MI and information theory for the construction of relevance networks lay in the ready adaptability of MI to representations of gene expression data as continuous distributions, via standard kernel methods [18]. To our knowledge, there are three other efforts aimed at extending the PID framework to continuous variables ([6, 92], and another ongoing work building upon [78]). We will be covering these efforts in greater depth. Overall, the conceptual scaffolding for a continuous partial information decomposition is still under construction [77].

The contribution of this chapter, then, is two-fold. First, we explore the potential of PID as a tool for the inference of synergy networks, particularly the PM PID of [45] (see Definition 15 for the relevant definition of PID in this setting). The $I_\cap^{\mathrm{PM}}$ PID is unusual in that it allows for signed information atoms. We explore the implications of this in both simulation and analysis, and find that it allows for counter-intuitive behavior that undermines the specificity

of the synergy defined by the $I_\cap^{\mathrm{PM}}$ PID. We see this in simulation, and also analytically for a degenerate system of random variables (X, Y, T), for which $T = g(X, Y)$ is a deterministic response of the jointly Gaussian predictors X and Y. Our second contribution, then, is to be the first work to extend the both the $I_\cap^{\mathrm{min}}$ and $I_\cap^{\mathrm{PM}}$ PIDs to such a system. We highlight that in Theorem 4 of Section 3.8, we are able to provide an explicit integral formula for the unique information in such systems for a general class of kernels g. This formula depends upon the relative magnitude of the partial derivatives of g, connecting the analytic properties of g to the probabilistic properties of (X, Y, T). We note that our work parallels the recent work done by Pakman *et al.* in [92] for the I^{BROJA} PID [12], in that both our work and theirs is concerned with computing the unique information atom for continuous variables. There is strong resemblance between the noise-free interaction that we are considering (Eq. E4) and the example neural models (III.1) and (III.2) in [92]. Besides our focus on different PIDs ($I_\cap^{\mathrm{min}}$ and $I_\cap^{\mathrm{PM}}$ rather than I^{BROJA}), our approach, focus, and conclusions are also distinct.

The layout of this chapter is as follows.

In the rest of Sec. 3.1, we will review the previous work in the literature relevant to our current work. We will review the application of information theoretic tools to gene network inference, the development of the concepts of redundancy and synergy in the networks literature, and the development of the PID framework. We will review previous attempts to apply PID to gene network inference as well as previous work on the application of PID to continuous variables.

In Section 3.2, we will present the idealized gene network inference problem, which motivates our investigation of the $I_\cap^{\mathrm{min}}$ and $I_\cap^{\mathrm{PM}}$ PIDs to follow. We are interested in examining their synergies as a tool for edge nomination,

with respect to correlational networks and a response variable paired to them (Def. 1).

In Section 3.3, we present three experiments, in which we simulated and discretize network data in order to evaluate the (discrete) PID synergies as inference tools. These experiments will provide a baseline understanding of the distinguishing behaviors of the $I_\cap^{\min}$ and $I_\cap^{\mathrm{PM}}$ PIDs, in the context of our network inference problem.

In Section 3.5, we present the information theory that we will be needing for our mathematical investigation of continuous interactions.

In Section 3.6, we present the conceptual framework for our investigation of continuous PIDs. We will present a mathematical formulation of bivariate interactions $T = g(X, Y)$, which simplify the network interaction framework and will serve as our primary object of investigation. We also introduce the definition of the $I_\cap^{\min}$ and $I_\cap^{\mathrm{PM}}$ PIDs in the continuous setting.

In Section 3.7, we present a fully worked-out example of the continuous PID concepts introduced in Sec. 3.6, using a linear interaction kernel. From this simple example, we will already be able to gain insight into some of the unusual behavior that the $I_\cap^{\mathrm{PM}}$ PID exhibits in the experiments in Sec. 3.3.

In Section 3.8, we present a general integral formula for the unique information of noise-free interactions in Theorem 4, for both the $I_\cap^{\min}$ and $I_\cap^{\mathrm{PM}}$ PIDs. Although the integrals set up by this formula tend to be non-trivial, it nonetheless provides us with a deeper insight into the relationship between the analytic properties of an interaction kernel g, and the informational relationship between predictors X, Y and target T under the constraint $T = g(x, Y)$. Besides allowing us to examine asymptotic conditional independence of predictors using Corollary 3 (applied later in Prop. 3.9.1), it also allows us to relate the behavior of the $I_\cap^{\mathrm{PM}}$ PID observed in this work to the concepts of

informative and misinformative probability mass exclusions, as originally developed in [46]. These concepts grounded the development of $I_\cap^{\mathrm{PM}}$ in [45], and thus, in Sec. 3.8.2, we are able to tie our own investigations into this literature.

Penultimately, In Section 3.9, we use Thoerem 4 from the previous section to analytically bound the unique information U_X^{PM} of the sigmoidal interaction kernel used for the experiments in Sec. 3.3. We are able to demonstrate that this unique information, U_X^{PM}, is negative, much as it was in all of our simulations in Sec. 3.3.

We conclude this chapter in Section 3.10, with a thematic summary, some final thoughts, and directions for future investigations.

3.1.1 Research Biography

Before discussing previous work in detail, it may be helpful to the reader to understand the problem space within which this current work emerged. The reader may skip to Sec. 3.1.2 without missing anything of essential importance to the rest of this work.

Our work in the partial information decomposition began from earlier efforts to identify interacting genes pairs in a small network of candidate genes with limited data. Discussions with collaborators in cellular and cancer biology made it apparent that they expected gene-drug interactions to be heavily context-dependent. A single gene and its expressed proteins perform multiple functions. Biological response can be hormetic and otherwise non-linear. Moreover, a single gene may contribute to distinct pathways that simultaneously facilitate and inhibit drug action. Standard, low-complexity regression was ineffective at identifying the relationships between candidate genes in our data set. Insofar as our data was already statistically under-powered, more complex regressions would have been useless. We thus turned to information

theoretic methods, as many of these are non-parametric and do not require model assumptions that are untestable without further experiments. Our initial aim was to identify collections of genes worth investigating together, as potential interacting agents. Once these pairs are properly identified, siRNA gene knockdown experiments could be used to test for a causal relationship between the two genes and the response. The question of formulating and testing a particular interaction model would have been the follow-up direction for methodological and experimental research.

The collaborative effort did not materialize prior to the 2020 COVID epidemic, at which point experimental collaboration became impossible. During the preparation for experimental collaboration, however, we ran many simulated experiments to investigate the information theoretic tools available, and determine those appropriate for evaluating in-lab experimental data. These simulations serve as the pseudo-experimental foundation of our current investigation. Given the current state of affairs, these will suffice for the moment.

3.1.2 Previous Work

In this section, we review the literature relevant to our current work. Although we will pay detailed attention to those works most pertinent to our own, we are also interested in providing a coherent narrative for the intersection of two distinct branches of research: the PID literature, and the 'omics network literature. Although ours is not the first work at the intersection of these fields ([13, 20, 24, 50], the overlap is still relatively sparse. Most experimental applications of PID have been in neuroscience [77], which falls within the broader field of network science, but is distinguished from the mainstream 'omics network research.

We begin by reviewing the history of the application of information-

theoretic tools to the task of network inference, with an emphasis on the inference of gene regulatory networks (GRNs) from bulk or single-cell gene expression data (Sec. 3.1.2.1). We will often refer to 'gene networks' as a more generic label, which includes both GRNs, which merely describe inter-gene relationships, and also synergistic networks, in which associations between genes represent a synergistic interaction by which the two genes affect a target biological response. We will discuss the early uses of mutual information as an alternative statistic of association between genes. We will then discuss subsequent attempts to remove confounding redundant information, analogous to the use of partial correlation in place of pairwise correlation for associational networks. We will also cover the subsequent interest, first in neuroscience and then in gene networks, of quantifying 'synergy' within a network structure.

This discussion will lead us in the foundation of PID by [125], and the subsequent developments within that literature (Sec.3.1.2.2). We will discuss previous attempts to apply PID to gene network inference, and also the work that has been done to extend PID to continuous variables.

We will then devote subsections to three previous works in need of attention. We will review the PIDC methodology of PID-based GRN inference as developed in [24] (Sec. 3.1.2.3). We will review the continuous MMI PID for Gaussian variables developed in [6] (Sec. 3.1.2.4). Finally, we will cover a brief evaluation of entropy-based gene-gene interaction methods [72], which highlights many of the same themes that we will emphasize in our work, including non-specificity and conditional (in)dependence of predictor variables (Sec. 3.1.2.5).

3.1.2.1 Information Theory and Network Inference

Multiple information theoretic approaches to network inference have been developed over the past two decades. Our current work focuses upon the Partial Information Decomposition (PID) framework, which can be understood as a branch within the broader computational information theory literature. We will first sketch the development of information theoretic methods as they have been applied to the inference of molecular networks, in order motivate the emergence of the PID paradigm. A recent overview of information theoretic methods in computational biology can be found in [25]. A comprehensive (as of 2015) enumeration and categorization of the use of information methods for molecular network inference can be found in the two-part series [88] and [87]. The PID literature began in 2011 with with [125]. Thus, for our purposes, reviewing the development of generic (non-PID) information theoretic methods, up to that point in time, is sufficient background to motivate the application of PID to biological network inference.

Mutual information (MI) has commonly been used to build associational networks, as it offers a non-linear, non-parametric alternative to correlation and other traditional statistical options. One of the early applications of MI to gene network inference was in the development of relevance networks from gene expression data by Butte and Kohane in [16]. Originally, they used pairwise correlation [17], before employing mutual information in [16]. In choosing MI, they claimed, among other advantages [18]:

1. MI can better handle irregularly distributed variables.

2. MI can handle more complex interactions, including negative regulation and mid-range expression activity.

3. MI can easily extend to models where expression levels are modeled with

noise, i.e. as distributions.

4. MI can naturally incorporate other, qualitatively distinct variables (e.g. phenotype).

However, MI suffers from many of the same issues as Pearson correlation for network inference. Neither can distinguish direct from indirect interactions, which is especially problematic for dense gene networks [80, 110]. For cancer-specific gene networks, in particular, there is reason to suspect more cross-talk, i.e. more active network edges, within molecular networks, and thus more indirect interaction. When networks are inferred via isolated pairwise comparisons, type II errors become a structural concern, and not merely a statistical problem to be overcome with more data.

Thus, the next generation of information-theoretic inference methods aimed at controlling FDR due to indirect associations. A head-to-head comparison of a few of these, compared to traditional MI relevance networks, can be found in [41]. Some of the more popular methods from this next generation include:

- The ARACNE method [7, 80] attempts to eliminate indirect associations in a two-part process. First, it nominates edges using MI, taking all gene pairs with MI above a cutoff I_0, determined from an empirical MI distribution estimated via data shuffling. Then, it examines all nominated triangles, and removes the edge with the lowest MI. They motivate this work with the data processing inequality (DPI). If X, Y, Z are a triplet of genes and $X \to Y \to Z$ is a Markov chain, then the DPI gives us $I(X;Z) \leq \min[I(X;Y), I(Y;Z)]$. Hence, their triplet pruning is guaranteed to remove any edge between conditionally independent genes.

- MRNET [84] takes a sequential approach to edge nomination in MI.

Each potential neighbor to a target gene Y is evaluated on the difference between its MI with the target, $I(X_j, Y)$, and the cumulative redundancy $\sum_k I(X_j, X_k)$ for each previously selected neighbor X_k of Y. This is an application of the minimum redundancy-maximum relevance (MRMR) method from [37].

- The MIDER method [118] is also a two-part process, with the first step nominating edges via MI. In the second step, indirect associations are sequentially filtered out by requiring nominated edges to meet a threshold in entropy reduction, conditioned on already accepted neighbors. This is mathematically synonymous with a relative conditional mutual information (CMI) threshold.

- The work in [110] is likely the first to explicitly consider CMI in network inference. Soranzo *et al.* offer a head-to-head comparison of many Pearson correlational and information methods of inference, including the ARACNE framework.

- Many other works [74, 76, 134, 135] also use CMI to control for indirect associations.

What unifies all these of works is their concern with the information contained in indirect associations, which accounts for the high FDR of isolated pairwise comparisons. We may say that they all aim, directly or indirectly, to exclude gene pairs with high MI but low CMI. The ARACNE method, although it does not directly employ CMI or a related quantity, is essentially an indirect method of eliminating edges with high MI but low CMI when conditioned on a mutual neighbor of both genes. The authors of MRNET understand their instantiation of the MRMR criterion to be an approximation of CMI [84]. The MIDER method's entropy reduction criterion is synony-

mous with normalized CMI. Thematically, what ties together this literature is a wariness of what we may term **redundancy**. The highly complex dependency structure within GRNs leads to an information landscape characterized by **redundant information**, or the sharing of uncertainty among more than two variables.

While this branch of research aimed at limiting redundancy, a parallel effort aimed to capture and employ the equally elusive quality of synergy. Whereas redundancy characterizes indirect associations that ought to be discarded, a synergy ought to locate truly multivariate associations between variables that are not reducible to pairwise interactions. The neuroscience literature introduced synergy into the networks literature in [15, 49]. Within a signal, neural spikes convey less information in isolation than when they are taken together in their succession. Brenner *et al.* [15] introduced[1] the following definition: the synergy level of X_1 and X_2 with respect to the variable Y is

$$\text{Syn}_Y(X_1; X_2) = I(X_1; X_2|Y) - I(X_1; X_2).$$

A positive value is said to indicate synergy, while a negative value indicates redundancy. This formula for synergy was previously defined in the information literature as interaction information [81] which we will turn to presently. Dimitris Anastassiou applied this definition of synergy framework to gene expression data in [2]. John Watkinson, a student of Anastassiou's, was to our knowledge the first to apply synergy to gene network inference in [124] and [123], both as part his dissertation research [122]. In [124], Watkinson *et al.* inferred gene 'synergy networks' for a specific cancer phenotype, formalized as a Boolean variable. In [123], the same group used synergy to augment the

[1] Brenner *et al.* introduced the notion of synergy in an NEC technical note in 1998, cited in [49]. The author has not located this original note, and the work in [15] does not reference it.

CLR algorithm, an extension of MI relevance networks previously developed in [42].

Interaction information, equivalent to Brenner's synergy for three variables, was first introduced by McGill in [81], and is formalized in contemporary notation[2] as an extension of mutual information:

$$I(T;X;Y) = I(T;X|Y) - I(T;X). \tag{3.1.1}$$

where T is the received signal and X and Y are the transmitted signals[3] In this definition, the three variables are said to have a positive interaction when knowledge of one variable increases the information flow between the others, i.e. when $I(T;X|Y) > I(T:X)$. As discussed above, this phenomenon has come to be understood as signifying synergy [116, 125]. By contrast, a negative interaction occurs when knowledge of one variables reduces the information shared between the other two, i.e. $I(T;X|Y) \leq I(T;X)$. This is understood as signifying redundancy [116, 125].

Let us consider an example relevant to gene networks. Suppose X, Y, and T signify the expression levels of genes, where there is cross-talk between X and Y, and both are suspected of activating a cellular process involving T. It may be that, by gaining knowledge of activation of Y, X becomes more informative about T because the activation of both X and Y would suggest a broader

[2]McGill's original formulation uses the terminology of transmission information in place of mutual information, and assumed empirical distributions. His definition of interaction information for three variables can be found in Eq. 13-14 on p. 101 of [81]. Interaction information is defined for more than three variables, however. Furthermore, up to a difference of sign convention, it is equivalent to the quantity developed by Hu Kuo Ting in [117] and Anthony Bell in [10]. Ting defines this alternative quantity as an additive set function, said to quantify multivariate amounts of information, but does not give it a particular name [117]. Bell designates the quantity as co-information. In this work, we discuss only interaction information in the three variable case, taking the sign convention used in [81]. This is the setting in which the PID framework is situated [125].

[3]The quantity is symmetric in the three arguments, and thus the designation of target signal is arbitrary.

functional context, e.g. anaerobic respiration. Thus, $I(T;X|Y) \geq I(T;X)$ because X informs on T within a specific context that Y can affirm, reducing the uncertainty in the relationship between X and T. However, it is also quite possible that the relationship that X has to the activation of T is largely mediated through Y. Here, any information that X provides about T is shared or redundant with that information in Y, and thus $I(T;X|Y) < I(T;X)$, i.e. we have a negative interaction. In an idealized, biologically implausible version of this scenario, we might have that $X \to Y \to T$ forms a Markov chain, and so $I(T;X|Y) = 0$. Thus, interaction information is pure redundancy: $I(T;X,Y) = -I(T;X)$. In general, if X and Y are independent, then the interaction information is strictly non-negative, exactly equal to the conditional mutual information $I(X;Y|T)$, and so the interaction is non-negative (synergistic or zero). If either X or Y is conditionally independent of T, given the other, then the interaction is non-positive (zero or redundant). Schneidman refers to the condition $I(T;X;Y) = 0$ as 'information independence', as a third type of 'independence' after typical and conditional, i.e. $I(X;Y) = 0$ and $I(X;Y|T) = 0$, respectively [101].

Identifying genes that act synergistically is indeed a fundamental problem in molecular biology. For instance, it is a commonplace that transcription factors behave synergistically [99, 119]. It is not difficult to see the shortcoming of the interaction information approach, which classifies triplets as **net synergistic** or **net redundant**. In truth, synergy and redundancy are not incompatible opposites. In reality, we expect the two to often co-occur within the same variables. When there is significant cross-talk and dependency between X and Y, then we expect some quantity of redundant information about any significant causal relationship $X \to T$ or $Y \to T$. On the other hand, we expect, for many gene pairs of interest, some element of synergy. GRNs are

generally characterized by a good deal of pathway redundancy and robustness, and this becomes especially pronounced in cancer cells[4] Thus, between alternative feedback mechanisms and topological cyclicity involving genes outside $\{X, Y, T\}$, we do not generally expect a situation where $X \to Y \to T$ is a Markov chain. To the contrary, the knowledge of both X and Y together will often better inform us regarding activated processes than either alone, providing synergistic information about T. Synergy should estimate that information as an explicit quantity beside, and not in opposition to, redundancy induced by cross-talk between the genes. The PID framework attempts to disentangle these two quantities of synergy and redundancy [116, 125]., thus offering an attractive framework for the inference of response-specific synergy networks. We will now turn to the development of PID and its current state in the literature.

3.1.2.2 Partial Information Decomposition (PID)

The PID paradigm was introduced by Williams and Beer in [125]. Their work was motivated by interaction information, and aimed to address the conflation of synergy and redundancy within $I(T; X; Y)$. In particular, they sought to decompose the information provided about T by predictors X and Y into four atoms of information: redundancy, synergy, and unique information for each

[4]See, for instance, literature on network entropy [36, 79] and on signaling entropy [113, 114, 115]. The former is more theoretical, while the latter is grounded within computational biology and transcriptomic data analysis.

predictor. The atoms would fulfill the following equations:

$$I(T;X,Y) = \underbrace{R(T;X,Y)}_{\text{Redundant Info.}} + \underbrace{U_X(T;X,Y) + U_Y(T;X,Y)}_{\text{Unique Infos.}} + \underbrace{S(T;X,Y)}_{\text{Synergistic Info.}} \tag{E1}$$

$$I(T;X) = R(T;X,Y) + U_X(T;X,Y) \tag{E2}$$

$$I(T;Y) = R(T;X,Y) + U_Y(T;X,Y) \tag{E3}$$

This type of decomposition is generally referred to as 'bivariate', since it includes only two predictor variables. The framework in [125] is defined for an arbitrarily large (though finite) number of predictors. A PID can be uniquely defined by its redundancy function $I_\cap$. Strictly speaking, it is an order-preserving function on a lattice of ascending collections of **sources**, the designation for subsets of predictor variables. Williams and Beer cited, among others, the information lattice structure from [10], in developing their own lattice formalism in [125]. We will not be utilizing this formalism in our current work except in passing, as we are restricting ourselves to the bivariate case, with the four PID elements R, U_X, U_Y, S corresponding to the lattice elements $\{X\}\{Y\}, \{X\}, \{Y\}$, and $\{X,Y\}$ respectively [125].

If one of these four atoms is defined, then the other three will follow from Eqs E1-E3. Williams and Beer introduced one measure of redundancy, the $I_\cap^{\min}$ function that we will present in Def. 14 (typically denoted $I_{\min}$ in [125] and elsewhere). This redundancy function, in turn, defines all four of these atoms $R^{\min}, S^{\min}, U_X^{\min}$, and $U_Y^{\min}$. With their definition, Williams and Beer achieved the aim of decomposing interaction information:

$$I(T;X;Y) = S^{\min}(T;X,Y) - R^{\min}(T;X,Y) \tag{3.1.2}$$

Their work opened a novel approach to multivariate information. It offered

a welcome break from the previous perspective on synergy and redundancy, which held them to be mutually exclusive quantities as captured by the sign of interaction information [77, 116]. This is especially attractive to our interest in gene network analysis, where we expect both redundancy and synergy in any pair of interacting genes.

However, the $I_\cap^{\min}$ redundancy function was not widely accepted as the optimal choice for quantifying redundant information [77], and multiple alternatives have been proposed [12, 45, 53, 56]. One of the most common complaints is that the $I_\cap^{\min}$ function does not distinguish the same information from the same *amount* of information. Specifically, regarding each outcome $T = t$ of the target variable, $I_\cap^{\min}$ compares the amount of information each predictor provides about that outcome, regardless of the degree of overlap of the outcomes in which each predictor informs upon T. Harder *et al.* constructed a particularly striking example of the draw-back to this approach [53] (see also [52]), later dubbed the 'two-bit copy' problem [56, 77]. When $T = (X, Y)$ is a two-bit copy of i.i.d. one-bit binary variables X and Y — i.e. $H(T) = 2$ and $H(X) = H(Y) = 1$ — the $I_\cap^{\min}$ function assigns one bit each of redundant and synergistic information to the decomposition of $I(T; X, Y)$. For any $t = (x, y) \in \{0, 1\}^2$, we have that X and Y provide the same amount of information about that outcome, on average. Thus, all the information contained within $I(T; X)$ (and $I(T; Y)$) is assigned to redundancy, and the synergy atom follows. Much of the PID community believes that it is more intuitive for a PID to designate $R = S = 0$ and $U_X = U_Y = 1$. A more thorough review of the objections and alternatives to the $I_\cap^{\min}$ PID can be found in the review [77], the overview of a 2018 special issue of *Entropy* on the PID framework. This cover paper also serves as a helpful gateway to the emergent PID canon, as it was co-written by notable senior researchers in the field.

Despite the rich development of the PID framework over the past decade, there are two key areas of research that remain relatively unexplored. First, the literature on the application of PID to the task of gene network inference has been relatively sparse. The most visible effort has been in the development by Chan *et al.* of the PID and context (PIDC) methodology for network inference [22, 24], which we will cover in greater detail presently (Sec. 3.1.2.3). More recently, Cang and Qin applied a similar approach to spatial single-cell sequencing data [20]. We will describe this in greater detail below.

Second, the PID literature has largely been confined to the study of discrete or discretized variables [77]. Adam Barrett contributed an early foray into the PID analysis of continuous variables in [6], in which he considered the extension of the PID framework to jointly Gaussian variables, both static and dynamic. In the specific context of jointly Gaussian variables, Barrett demonstrated that for a large class of potential PIDs (including those in [125], [52], and [12], but excluding [45] and [56]), the PID of a trivariate jointly Gaussian system must necessarily take the same form, which they term the minimal mutual information (MMI) PID. We will review this work in greater detail in Sec. 3.1.2.4, as there is significant overlap with our own analysis of the $I_{\cap}^{\min}$ PID, particularly in Sec. 3.7. We present a stand-alone proof that aligns with the narrative focus upon interaction kernels $g(X, Y)$ that we will be exploring in this effort. Besides MMI PID for Gaussians, there are other, much more recent works relevant to extending PID to continuous variables. Ari Pakman has extended the I^{BROJA} PID ([12]) to continuous variables [92]. Abdullah Makkeh and his colleagues have made a major stride is developing a PID redundancy function that is differentiable with respect to the underlying probability mass function in [78]. This development is a crucial step for their continuous extension of the PID framework, and we anticipate forthcoming

work from the same group.

Besides the PIDC, we are aware of only the one other applications of the partial information framework to gene network inference [20], and handful of other related works [13, 50]. We briefly summarize these in turn. Zixuan Cang and Qing Nie developed an ambitious methodology for inferring inter-cellular signaling networks from limited spatial scRNAseq data [20], using optimal transport methods. They employed the unique information atom of the $I_\cap^{\min}$ PID in order to infer intercellular signaling within a spatial neighborhood, with a formalism similar to that from the PIDC method.

In a biostatistical application of PID unrelated to gene networks as such, Granada *et al.* used PID as part of the information-theoretic analysis of cellular predictors of the response to cisplatin, a common chemotherapeutic agent [50]. As discrete predictors, they examined both proliferative history and cell cycle state, in addition to cisplatin dosage. An MI-based analysis indicated that cisplatin dose was the most informative sole predictor. However, employing the $I_\cap^{\mathrm{PM}}$ PID from [125], they found synergistic and redundant information for the combination of cisplatin dosage and proliferative history as predictors. They noted that for low and medium cisplatin dosage, proliferative history seemed to shift the likelihood of cell arrest versus cell death, with a higher proliferative index favoring non-lethal arrest.

Finally, Ayan Biswas conducted a PID-based investigation of a two-step cascade, modeling the transcriptional motif $S \to X \to Y$ as a system of stochastic differential equations [13]. Using Barrett's continuous extension [6] for Gaussian variables, the author examined relative PID components under a varying 'fitness' parameter controlling the expected population of the source species S. There is some resemblance between this work and the simulation analysis we conduct in Section 3.3 and Section 3.7.3, at least in terms of the

visualizations used to examine relative PID components.

3.1.2.3 PID for GRN Inference: The PIDC Method

Our work applies the PID framework to the study of gene regulatory networks, which is a surprisingly sparse niche of the literature. To our knowledge, the first application of PID to GRN inference lies in the development of the PIDC algorithm [24], central to the dissertation work [22]. In [24], Chan *et al.* develop a similarity score designated as the proportional unique contribution (PUC). For the collection of genes in their network $\mathcal{V}$ (denoted S in their paper), for each candidate pair (X, Y) they define the PUC (originally denoted $\mu_{X,Y}$) as

$$\mathrm{PUC}_{X,Y} = \sum_{Z \in \mathcal{V} \setminus \{X,Y\}} \frac{U_X^{\min}(Y; X, Z)}{I(X;Y)} + \frac{U_Y^{\min}(X; Y, Z)}{I(X;Y)}. \tag{3.1.3}$$

The idea behind this quantity is as follows. Consider the first term in the summand, $U_X^{\min}(Y; X, Z)/I(X;Y)$. This considers the $I_\cap^{\min}$ PID of the bivariate system in which Y is the target variable, and X is paired with every other gene Z as potentially overlapping (i.e. redundant) predictors of Y. The fraction quantifies the proportion of mutual information $I(Y;X)$ that is *not* redundant with the other gene Z. This proportion is computed and summed over every other gene Z. Likewise, for the symmetric scenario in which X is the the target and Y is paired with each Z as a predictor, we have the second term in the summand. The statistic $PUC_{X,Y}$ is then compared to the empirical distribution of scores for each gene to arrive at a confidence scoring.

This approach is more interested in using PID to eliminate redundancy, rather than discover synergy, as compared to what we will be presenting in this paper. We see it as akin to the earlier methods of reducing redundancy

that we have previously discussed, including ARACNE, MRNET, and MIDER [80, 84, 118]. PIDC uses the $I_{\cap}^{\min}$ redundancy function, but any question of which PID redundancy function to use reduces to the question of how a given redundancy chooses to partition the MI between candidate genes $I(X;Y)$ into unique and redundant information for each candidate confounder Z. If the relationship between X and Y is conditionally independent of every other gene, then $\mathrm{PUC}_{X,Y}$ achieves its maximal value of 2.

PIDC has had mixed results in benchmarking exercises. Chen and Mar evaluated PIDC along with four other methods for GRN inference from single cell expression data in [29]. Two other information-based methods were also examined: the ARACNE method [80] described above, and the CLR method from [42]. Using simulated networks and also experimental data sets matched to the STRING network (STRINGv10, [111]) they found none of the methods examined to be particularly powerful. For one simulation of 100 genes (Sim1 in Fig. 2a in [29]), the ROC curve is indistinguishable from random guessing. Pratapa *et al.* conducted another recent benchmarking exercise [94] that found that PIDC was relatively favored, compared to many other methods. PIDC performed well for some experimental data sets (with the STRING network taken as ground truth) and Boolean models curated from the literature. The authors ultimately recommend PIDC along with tree-based methods GENIE3 [57] and the related GRNBoost2 [86], citing both the accuracy and stability of PIDC and GENIE3. One of the interesting results from [94] was that methods that performed well for synthetic data sets created with GeneNetWeaver (similar to those in [24]) did not perform as well for literature-curated Boolean models or experimental data sets (paired to the STRING network). Taken together, these two exercises suggest that PIDC appears at least as reliable as other popular methods of GRN inference from single-cell data, and performs

reasonably well for models and data based on the current understanding of biological reality. However, we do want to emphasize that the methods examined in [24] and [94] are still a limited subset of those popular in the literature.

GRN inference and synergy network inference are distinct tasks: the former is interested only in the relationships between genes, while the latter is interested in how these relationships relate to a biological response. The former task, by its nature, may only be interested in PID synergy insofar as it is interested in discovering multiedges. Our work in this paper, in motivation, is more interested in the latter task. Even so, much of our theoretical investigation, particularly that in Sections 3.8 & 3.9, is examining unique information. If there is a stronger connection between our work and the PIDC method besides a high-level interest in gene network inference using PID, it is not apparent to us. We leave the exploration of that possibility to future work.

3.1.2.4 PID of Gaussian Variables and the Minimal Mutual Information (MMI) PID

Adam Barrett made the first effort at extending the PID framework to continuous variables in [6]. He studied three jointly Gaussian variables of mean zero and unit variance, which we represent here:

$$(X, Y, T) \sim N(\mathbf{0}, \Sigma) \tag{3.1.4}$$

$$\Sigma = \begin{bmatrix} 1 & \rho_{X,Y} & \rho_{X,T} \\ \rho_{X,Y} & 1 & \rho_{Y,T} \\ \rho_{X,T} & \rho_{Y,T} & 1 \end{bmatrix}$$

Using the I^{BROJA} PID developed in [12] and [53], the bivariate PID was

computed for this triplet, which was proven to be identical for a large class of potential PIDs, including the original from [125]. When Bertschinger *et al.* introduced the I^{BROJA} PID in [12], they introduced the condition $(\star)$ for a PID. This new condition requires that unique information, i.e. U_X and U_Y in any PID of $I(T; X, Y)$, must depend only upon the marginals $p_{X,T}, p_{Y,T}$, and $p_{X,Y}$. The $I_\cap^{\min}$ PID fulfills this condition, as does I^{BROJA}. For both of these PIDs, then, we may present one of Barrett's main results (the WB Axioms, labeled equations (M), (P), (SR), (S), appear in Section 3.6.2).

Theorem 1 (Barrett 2014)**.** *Let $I_\cap$ be a PID redundancy function (Def. 13) that satisfies the WB axioms, and for which the induced unique information atoms in the bivariate PID also satisfy condition $(\star)$ from [12], i.e. they depend only upon the marginal distributions $p_{X,Y}, p_{X,T}$, and $p_{X,T}$. Then for variables (X, Y, T) as in Eq. 3.1.4, $I_\cap$ induces the following bivariate PID:*

$$
\begin{aligned}
R(T; X, Y) &= \min(I(T; X), I(T; Y)) \\
U_X(T; X, Y) &= I(T; X) - \min(I(T; X), I(T; Y)) \\
U_Y(T; X, Y) &= I(T; Y) - \min(I(T; X), I(T; Y)) \\
S(T; X, Y) &= I(T; X, Y) - I(T; X) - I(T; Y) + \min(I(T; X), I(T; Y))
\end{aligned}
$$

If we assume $|\rho_{X,T}| \leq |\rho_{Y,T}|$*, then we have that*

$$
\begin{aligned}
R(T;X,Y) &= I(T;X) = \frac{1}{2}\log\frac{1}{1-\rho_{X,T}^2} \\
U_X(T;X,Y) &= 0 \\
U_Y(T;X,Y) &= I(T;Y) - I(T;X) = \frac{1}{2}\log\frac{1-\rho_{X,T}^2}{1-\rho_{Y,T}^2} \\
S(T;X,Y) &= I(T;X,Y) - I(T;Y) \\
&= \frac{1}{2}\log\frac{(1-\rho_{X,Y}^2)(1-\rho_{Y,T}^2)}{1-(\rho_{X,Y}^2+\rho_{X,T}^2+\rho_{Y,T}^2)+2\rho_{X,Y}\rho_{X,T}\rho_{Y,T}}
\end{aligned}
$$

In particular, this decomposition holds for both the $I_\cap^{min}$ *and* I^{BROJA} *PIDs developed in [125] and [12, 53] respectively.*

The demonstration of this decomposition in [6] proceeds as follows. First, the I^{BROJA} PID from [12] is extended to continuous Gaussian variables (X, T, T), and the above decomposition is demonstrated for that PID. Then, by invoking Lemma 3 from [12], it follows that this decomposition provides an upper bound on unique information for any decomposition induced by a redundancy $I_\cap$ satisfying the (bivariate) WB axioms and $(\star)$ conditions. If $U_X^{\text{BROJA}} = 0$, it follows that $U_X = 0$ for any such $I_\cap$. Thus, the same PID follows, regardless. Barrett terms this the minimal mutual information (MMI) PID.

Our investigation to follow builds upon this work, particularly in Section 3.7. In that section, we are considering a noise-free linear interaction T of jointly Gaussian predictors X and Y. With appropriate renormalization, our variables (X, Y, T) in Sec. 3.7 are identical to a singular limit of the (X, Y, T) above, and our result in Theorem 2 is essentially identical to that in Theorem 1. Indeed, Barrett anticipates this limit (pg. 7 of [6]). Our approach to Theorem 2 is distinguished from that in [6] in that it (a) directly

employs $I_\cap^{\min}$ in its continuous form, whereas [6] uses I^{BROJA} and (b) explores how $I_\cap^{\min}$ behaves in noise-free (i.e. degenerate) interactions. We present it as a stand-alone theorem because of these features, plus the narrative continuity it provides in conjunction with Theorem 3 and the more general Theorem 4 in the following section.

3.1.2.5 Lee *et al.* Investigate the Non-Specificity of Entropy-Based Synergy Inference Methods

Finally, we conclude our review of the literature by noting that our investigation of PID synergy builds upon many of the themes highlighted in the concise article [72]. In this work, Lee *et al.* considered multiple entropy and MI-based methods for detecting gene-gene interactions for a binary trait. They were working within the elementary framework in which a binary phenotype Y may be synergistically influenced by discrete SNP predictors X_1 and X_2. They considered the entropy methods from [38], [43], and [132], with their analysis of the latter applicable to the similar measure proposed in [26]. They highlighted the lack of specificity in entropy-based methods of edge nomination. Thematically, this is related to a central topic within our own work as well: the non-specificity of the $I_\cap^{\text{PM}}$ PID.

The authors observed that only the method from [38] was guaranteed to assign a zero score to a system in which X_2 is conditionally independent of Y, given X_1. Similarly, in our own work, we will continually highlight how, unlike the $I_\cap^{\min}$ PID or any PID satisfying the WB axioms, the $I_\cap^{\text{PM}}$ PID will not necessarily assign zero synergy to a pair of predictor genes even if one of the genes is conditionally independent of the target variables. Ironically, it is by assigning too much redundancy to such cases that the $I_\cap^{\text{PM}}$ PID also assigns too much synergy (Eq. 3.6.33).

3.2 Motivation: Gene Network Inference Problem

Our current work is motivated by the application of partial information methods to the task of network inference, with the specific goal of developing tools for edge nomination in cancer biology applications. To create clinically useful models of drug action and sensitivity, it is important to identify key molecular agents and interactions, and to distinguish them from auxiliary pathways. The Partial Information Decomposition (PID) extension to information theory offers an agnostic, non-parametric approach to identifying the most informative combinations of predictor variables. Consider, for example, a drug response metric, e.g., the half inhibitory concentration (IC_{50}), as a target and the expression levels of active genes as candidate predictors. The PID framework then gives us a means of identifying and coupling *synergistic* biomarkers while eliminating redundant pathway information.

Going forward, we will restrict our terminology in order to distinguish between 'edges' and 'interactions' in a network, where 'edges' will correspond to a meaningful relationship between the node variables (in our case, correlations), and 'interactions' will refer to the joint effect that the node variables have upon a response variable.[5] In this work, our motivation is in developing the appropriate metrics for the task of inferring a response-specific network of interacting genes from transcriptomic data. The interactions are taken to represent the joint contribution of two expressing genes to a drug response. Although we focus on gene-interaction networks as our main motivation, these methods can be applied to many of the biological (and social) interaction networkse. To frame this problem mathematically, we consider a simple model

[5]This is in contradistinction to, for instance, a protein-protein interaction (PPI) network, which has no associated response variable and by 'interaction' indicates a biomolecular relationship. Rather, our use of 'interaction' is more akin to the edges in synergy networks [124].

of a gene interaction network paired to a drug response, in which genes are modeled via standard normal variables. Gene expression is often modeled as either log-normal or Gamma distributed [8]. Assuming log-normal distributions, our predictors can be interpreted as expression levels that have been log transformed and normalized.

Definition 1 (Gene Interaction Network)**.** *Let* $(\mathcal{V}, \mathcal{E})$ *be an undirected graph of* $n = |\mathcal{V}|$ ***genes****, and* $\text{Sgn} : \mathcal{E} \to \{\pm 1\}$ *an edge attribute signifying positive or negative regulation. Let* $\boldsymbol{X} \sim N(\mathbf{0}, \Sigma)$ *be an* n*-dimensional Gaussian vector, where each* X_i *signifies the* ***expression of gene*** $\boldsymbol{i}$*. For a given constant* $\rho \in (0, 1)$*,* $\boldsymbol{X}$ *has covariance structure:*

$$\Sigma_{i,j} = \begin{cases} 1, & i = j \\ 0, & \{i, j\} \notin \mathcal{E} \\ \pm\rho, & \text{Sgn}(\{i, j\}) = \pm 1 \end{cases} \tag{3.2.1}$$

Let $\mathcal{E}' \subset \{(i, j) | \{i, j\} \in \mathcal{E}\}$ *be a subset of directed edges, called* ***interactions****. The* ***(drug) response*** T *is the real-valued variable of the form*

$$T = \sum_{(i,j) \in \mathcal{E}'} g(X_i, X_j) \tag{3.2.2}$$

where $g : \mathbb{R}^2 \to \mathbb{R}$ *is the* ***interaction kernel.*** *The interacting genes* $\mathcal{E}'$ *model the subset of the total interactions that significantly alter the response* T*. We refer to this full collection of objects as a* ***gene interaction network****, denoted* $\mathcal{N}$*.*

We may refer to a ***mixed gene interaction network*** *when, for some subset of vertices* $\mathcal{S} \subset \mathcal{V}$*, and associated coefficient vector* $\boldsymbol{\beta} \in \mathbb{R}^{|\mathcal{S}|}$*, the re-*

sponse T takes the form

$$T = \sum_{(i,j)\in\mathcal{E}'} g(X_i, X_j) + \sum_{s\in\mathcal{S}} \beta_s X_s \tag{3.2.3}$$

for a non-linear kernel g.

Given a gene interaction network $\mathcal{N}$, our goal is to identify the pairs of interacting genes $\mathcal{E}'$ from sample data drawn from $(\boldsymbol{X}, T)$. Here, we seek a statistic that discriminates interacting pairs from non-interacting pairs, i.e., between the hypotheses $(i,j) \in \mathcal{E}'$ and $(i,j) \notin \mathcal{E}'$. For this network inference task, we will use measures of synergistic information from the PID literature to infer membership (or lack thereof) in $\mathcal{E}'$. Given the expression of two genes X_i and X_j and the response T, the PID framework allows us to decompose the mutual information between predictors and response into four atoms of information:

$$I(T;X,Y) = \underbrace{R(T;X,Y)}_{\text{Redundant Info.}} + \underbrace{U_X(T;X,Y) + U_Y(T;X,Y)}_{\text{Unique Infos.}} + \underbrace{S(T;X,Y)}_{\text{Synergistic Info.}} \tag{E1}$$

$$I(T;X) = R(T;X,Y) + U_X(T;X,Y) \tag{E2}$$

$$I(T;Y) = R(T;X,Y) + U_Y(T;X,Y) \tag{E3}$$

We discuss these atoms in more detail in Section 3.6.2.1. While the original PID framework from [125] defines redundant information as in Def. 14, multiple other PIDs have been proposed [12, 45, 56]. In this paper, we are mainly concerned with the decompositions proffered in [125] and [45], and to distinguish amongst these in the subsequent motivating experiments and theory, we shall adopt the following notation: we will let $I_\cap^{\min}$ denote the general PID framework of [125] with $S^{\min}$, $R^{\min}$, $U_X^{\min}$, and $U_Y^{\min}$ denoting the respective synergistic, redundant, and unique informations in the $I_\cap^{\min}$ framework; simi-

larly, we use $I_\cap^{\mathrm{PM}}$ to denote the general PID framework of [45] with S^{PM}, R^{PM}, U_X^{PM}, and U_Y^{PM} denoting the respective synergistic, redundant, and unique informations in the $I_\cap^{\mathrm{PM}}$ framework.

From each of the above PID frameworks, we will attempt to infer membership in $\mathcal{E}'$ via an estimate (denoted $\hat{S}$) of the synergistic information atom $S(T; X_i, X_j)$ (i.e, $S^{\min}(T; X_i, X_j)$ and $S^{\mathrm{PM}}(T; X_i, X_j)$) derived from the empirical distribution. To this end, if two genes i and j contribute to the response in Eq. (3.2.2), then for some fixed $c > 0$ we would expect synergistic information to satisfy

$$S(T; X_i, X_j) > S(T; X_{i'}, X_{j'}) + c$$

for some $c > 0$ and any $(i', j') \notin \mathcal{E}'$, and hence

$$\mathbb{E}\hat{S}(T; x_i, x_j) > \mathbb{E}\hat{S}(T; x_{i'}, x_{j'}) + c.$$

We explore this further in three illustrative examples considered in the next section.

3.3 Synergistic Information Discrimination in Network Simulation Experiments

In order to evaluate the multiple competing definitions of PID synergy, we first simulate gene interaction networks and evaluate the performance of sample synergy as a discriminator of true interactions from random pairs of genes, i.e. direct vs indirect associations as discussed in Sec. 3.1.2.1. These simple experiments will serve to illustrate that not all PID frameworks are equally useful for a given interaction inference task. To this end, we consider a sig-

moidal switch interaction kernel of the following form:

$$g(X, Y) = \frac{Y}{1 + e^{\alpha - X}} \tag{E4}$$

Sigmoidal functions are a typical choice, among others (including Hill functions), when modeling biomolecular activation in transcriptional networks [120, 121]

In Experiment I, we will naively compare four competing definitions of PID synergy, evaluating their performance at identifying four intercorrelated interactions on the same gene network hub (Section 3.3.1). We will see that the $I_\cap^{\mathrm{PM}}$ PID's synergy, $\hat{S}^{\mathrm{PM}}$, seems to outperform the other synergies at this task. In particular, this experiment will highlight the **sensitivity** of the $I_\cap^{\mathrm{PM}}$ PID for our network inference task.

In Experiment II, we will use a mixed interaction network to demonstrate the limitation of MI as a network inference methodology, as it fails to distinguish between truly interacting gene pairs and those where only one gene contributes to the response signal. Moreover, we will see that the $I_\cap^{\mathrm{PM}}$ PID shares this limitation, whereas the original $I_\cap^{\min}$ PID behaves in a more intuitive and desirable manner. This experiment will demonstrate the **specificity** of the $I_\cap^{\min}$ PID, that is: its tendency to avoid type II errors in edge nomination.

In Experiment III, we revisit the network inference task in Experiment I, now analyzing the effect of the parameter α in Eq. (E4) upon the $I_\cap^{\mathrm{PM}}$ and $I_\cap^{\min}$ PIDs. While we expect the joint information that X and Y provide regarding $g(X, Y)$, and thus toward the total response T, to remain similar, altering α ought to affect the balance of information, so-to-speak, *between* the variables X and Y with regards to T. In the analysis of this experiment, we will also take the opportunity to present some preliminary heuristics to suggest the

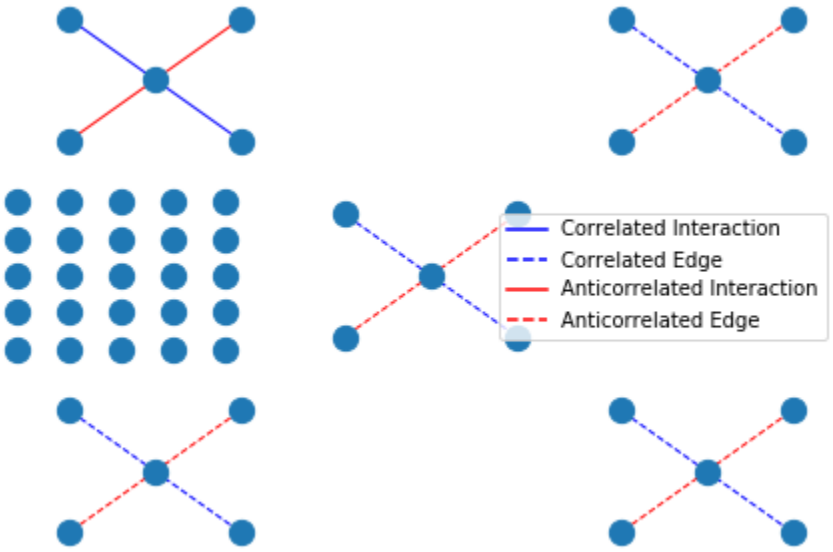


Figure 3.1: **Gene interaction network topology for Experiment I** We simulate a network of $n = 50$ genes, as described in Def. 1 with 25 edges ($\mathcal{E}$) edges arranged in 5 4-stars. The drug response (Eq. 3.3.1) is computed on the edges of one of the stars, using the interaction kernel (E4).

relationship between the analytic properties of the interaction kernel (E4) and the information decomposition of predictor information.

Taken together, these three experiments will demonstrate that although $I_\cap^{\mathrm{PM}}$ produces PID atoms sensitive to the overall informative potential of a gene pair, it is the $I_\cap^{\mathrm{min}}$ that will behave more intuitively toward less synergistic pairs of predictors. In heavily interdependent and noisy molecular interaction networks, edge nomination methodologies must anticipate 'false' interactions and indirect associations that may become nominated through a combination of the true informative importance of one predictor and the spurious relevance of another.

3.3.1 Experiment I

We first simulate a gene network of $n = 50$ nodes, with 25 isolated nodes, and 25 nodes arranged into 5 disjoint 4-stars, so that we have 20 degree-1

'spokes', 5 degree-4 hubs (Fig 3.1). On each star, two edges are correlated and two are anti-correlated (i.e., $\Sigma_{i,j} < 0$). For our interactions, we choose the 4 edges of one of the stars. In other words, given the interaction kernel (E4), our response T is the sum

$$T = \sum_{i=1}^{4} \frac{Y}{1 + e^{-X_i}} \tag{3.3.1}$$

where, for convenience, we let Y denote the gene expression of the hub, and $X_1, ..., X_4$ denote the expressions of the spoke genes. We ran 20,000 simulations of this network, broken into 100 batches of 200 replicates (simulating the small batch size typical of clinical experiments). Each batch of the data is represented by a data matrix D where each row $k = 1, ..., 200$ is a data point $D_k = (\boldsymbol{X}(k), T(k))$. For each batch, we discretized D column-wise into 3 equal-width data bins, discretizing each variable individually.

In order to evaluate the performance, for each simulation and for each pair of nodes (i, j), we computed the discrete PID of the $\hat{I}(T; X_i, X_j) = I(\hat{T}; \hat{X}_i, \hat{X}_j)$ into its four component atoms $(\hat{R}, \hat{U}_{X_i}, \hat{U}_{X_j}, \hat{S})$, using the PID definitions from [12, 45, 56, 125]. For each simulation, we ranked pairs (i, j) according to their synergistic information under these definitions. In particular, we are interested in the ranking of our true interaction pairs, as synergy is only a useful quantity if it statistically distinguishes true interactions from null pairs.

We present the ranked synergy distributions from this first experiment in Fig 3.2, the full ranked PIDs of interacting pairs in Fig 3.3, and the distributions of PID atoms in Fig 3.4. The $I_\cap^{\mathrm{PM}}$ PID of [45] produces the only decomposition to consistently rank true interactions in the top 5% of synergistic information in the simulations (Fig 3.2). That is, $\hat{S}^{\mathrm{PM}}$ out-performs $\hat{S}^{\mathrm{min}}$, $\hat{S}^{\mathrm{CCS}}$, and $\hat{S}^{\mathrm{BROJA}}$ at discriminating interacting gene pairs from null pairs.

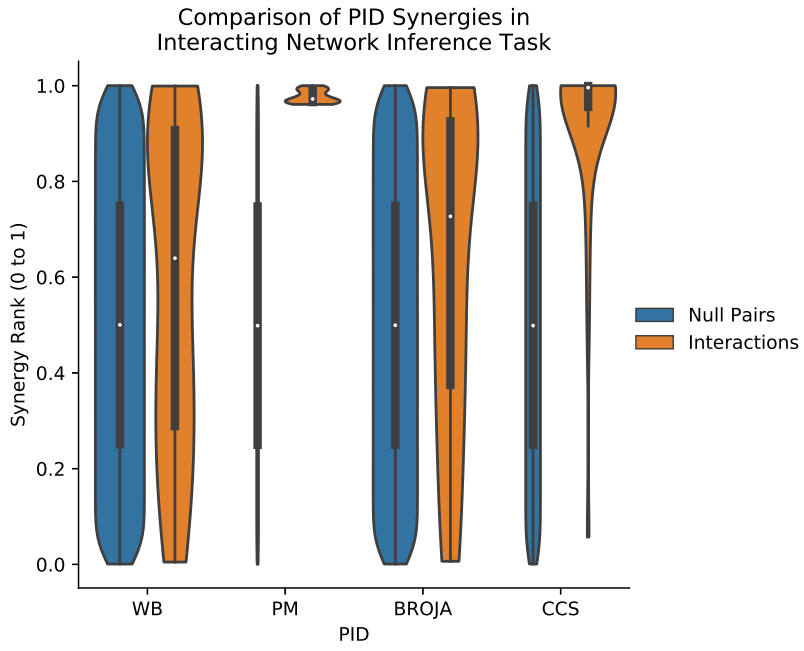


Figure 3.2: **Performance comparison of PID synergies for Experiment I** We compare the performance of different PID synergies as discriminators of network interactions, as part of our first experiment with the network in Fig 3.1. For each batch of 200 replicates, we computed the synergy atom of the bivariate PID (Eq. E1) for each pair of genes in our network, and then converted these values into ranked scores from 0 to 1. Shown here are the distributions of these scores, for interacting pairs and non-interacting pairs, the latter serving to approximate an empirical null distribution. As can be seen, only S^{PM} consistently ranks true interactions in the top 5% of gene pair synergies.

Taking this result in isolation, we might conclude the the $I_\cap^{\mathrm{PM}}$ PID is the appropriate choice of PID for gene network inference. As we shall see, this is not the conclusion that we will ultimately reach.

To better understand the differences between our candidate PID synergies, we examine the distributions of the full 4-atom PIDs, as ranked scores in Fig 3.3 and as (pre-ranked) information atoms in Fig 3.4. We see that, compared to the other PIDs, $I_\cap^{\mathrm{PM}}$ assigns more synergistic information and less unique information to the 'switch' gene X, the first argument in the kernel (E4); see Fig 3.3. In fact, when we look at the information distributions in Fig 3.4, we see that $I_\cap^{\mathrm{PM}}$ alone is consistently assigning negative unique information $\hat{U}_X^{\mathrm{PM}}$. In Eqs E1-E3, there is a trade-off between the synergistic and

redundant information atoms on the one hand, and unique informations on the other. The monotonicity property of the $I_\cap^{\min}$ PID in [125] guarantees

$$R(T; X_i, X_j) = I_\cap^{\min}(T; X_i, X_j) \leq I(T; X_i) \leq I(T; X_i, X_j)$$

and thus, in turn, guarantees that the other information atoms U_X, U_Y are non-negative as $U_X = I(T; X) - R(T; X, Y)$ (and similarly for U_Y). By contrast, $I_\cap^{\mathrm{PM}}$ does not guarantee monotonicity, and thus allows for negative information atoms, as we see in our simulations. We note that the $I_\cap^{CCS}$ PID from [56] also allows for negative information atoms, and is second only to $I_\cap^{\mathrm{PM}}$ in discriminating true interactions from null pairs in this experiment. This is a strange, unintuitive feature of the $I_\cap^{\mathrm{PM}}$ PID. Later, we will explore the natural question: What does negative information indicate about the relationship between the predictor X and the target variable T?

The simulations from our first experiment suggest that S^{PM} is better suited to our network inference task than the other synergies. Moreover, every information atom of the PM PID distinguishes true interactions from null pairs, while synergy does not for the WB PID, as can be seen in Table 3.1 and Figs 3.3-3.4. However, we see that mutual information itself strongly suggests the true pairs (Fig. 3.3), which is a limitation of this particular simulation. If it were the case that MI can distinguish interacting pairs as well as any PID atom, then what PID might offer in place of MI is potential interpretability. By itself, this does not make for a compelling case for the application of PID to network inference. In our next two experiments, we explore the distinction between mutual information and PID atoms, and will make the case for the indispensability of the latter. This will highlight the drawbacks that $I_\cap^{\mathrm{PM}}$

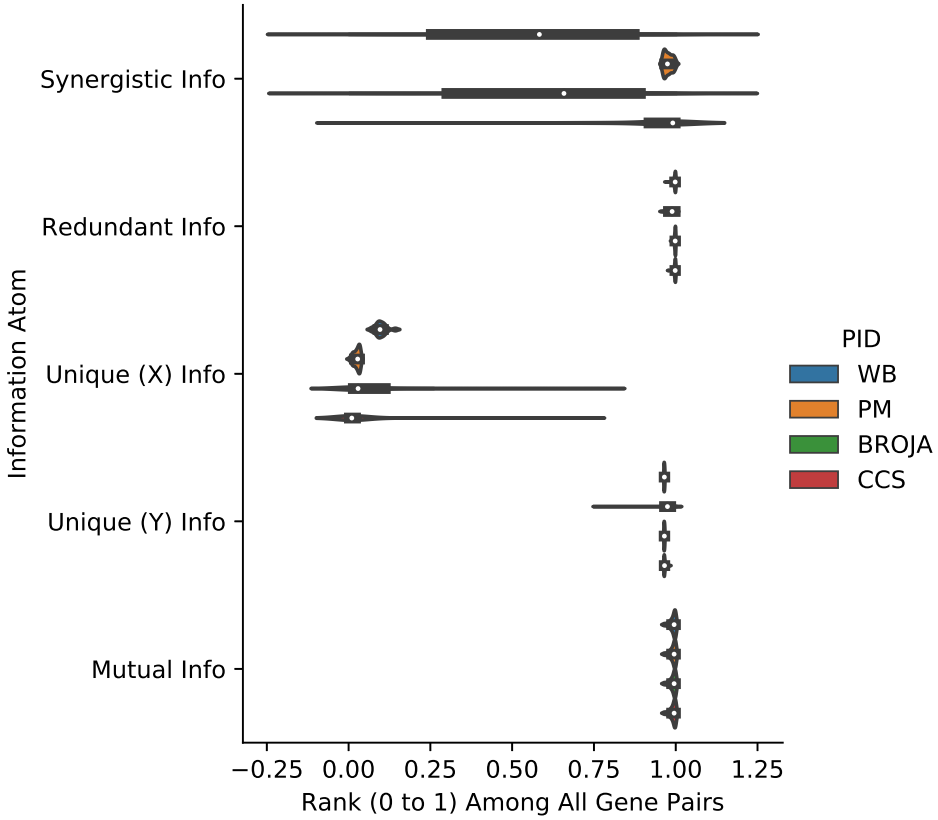


Figure 3.3: **Ranked bivariate PID atoms for interactions in Experiment I** For the simulated data from Experiment I, we present the ranked scores of the four bivariate PID information atoms S, R, U_X, U_Y and the mutual information $I(T; X_i, X_j)$ for true interaction pairs $(i, j) \in \mathcal{E}'$, as computed from the empirical distributions of each batch of simulated data. Although synergy only distinguishes true interactions for the $I_\cap^{\text{PM}}$ PID (as in Fig. 3.3), mutual information itself discriminates them. Thus, other atoms (R and U_Y especially) distinguish true interactions for all four PIDs.

shares with MI for this particular type of task.

3.3.2 Experiment II

The benefit of a PID approach over a mutual information approach to gene network inference becomes apparent when we instead consider mixed gene interaction networks (Def. 1). In our next experiment, we run a second set of simulations in which we consider a mixed gene interaction network, again of $n = 50$ nodes, this time with 20 edges (Fig 3.5). Our network topology is comprised of two disjoint 10-stars, with hub predictors we will denote as Y_1

PID	Info. Atom	KS Stat	pval	Mean Rank (Intxn)	Mean Rank (Null)
WB	$\mathrm{S}^{\min}$	0.232	2.769e-02	0.585	0.500
WB	$\mathrm{R}^{\min}$	0.997	7.897e-35	0.998	0.499
WB	$\mathrm{U}_X^{\min}$	0.800	1.447e-22	0.102	0.502
WB	$\mathrm{U}_Y^{\min}$	0.966	9.069e-33	0.965	0.499
WB	$I(T;X,Y)$	0.966	9.069e-33	0.992	0.499
PM	S^{PM}	0.963	1.501e-32	0.977	0.499
PM	R^{PM}	0.965	1.167e-32	0.986	0.499
PM	$\mathrm{U}_X^{\mathrm{PM}}$	0.966	9.069e-33	0.025	0.502
PM	$\mathrm{U}_Y^{\mathrm{PM}}$	0.911	3.898e-29	0.969	0.499
PM	$I(T;X,Y)$	0.966	9.069e-33	0.992	0.499

Table 3.1: Kolmogorov-Smirnov comparison of PID atoms, as ranked scores, between true interactions and non-interacting null pairs from our first network simulation experiment (Figs 3.1-3.4).

and Y_2, and with spokes $X_1, ..., X_{10}$ and $X_{11}, ..., X_{20}$ respectively. This network was paired to the response

$$T = \frac{Y_1}{1 + e^{-X_1}} + \beta Y_2 \tag{3.3.2}$$

for a given real parameter β. Put into words, the response is the sum of a paired interaction on the first star and a univariate signal from the second hub. Moreover, we also experimented with the number of interactions k on the first hub. For $1 \leq k \leq 10$, then, we have the more general response:

$$T = \sum_{i=1}^{k} \frac{Y_1}{1 + e^{X_k}} + \beta Y_2. \tag{3.3.3}$$

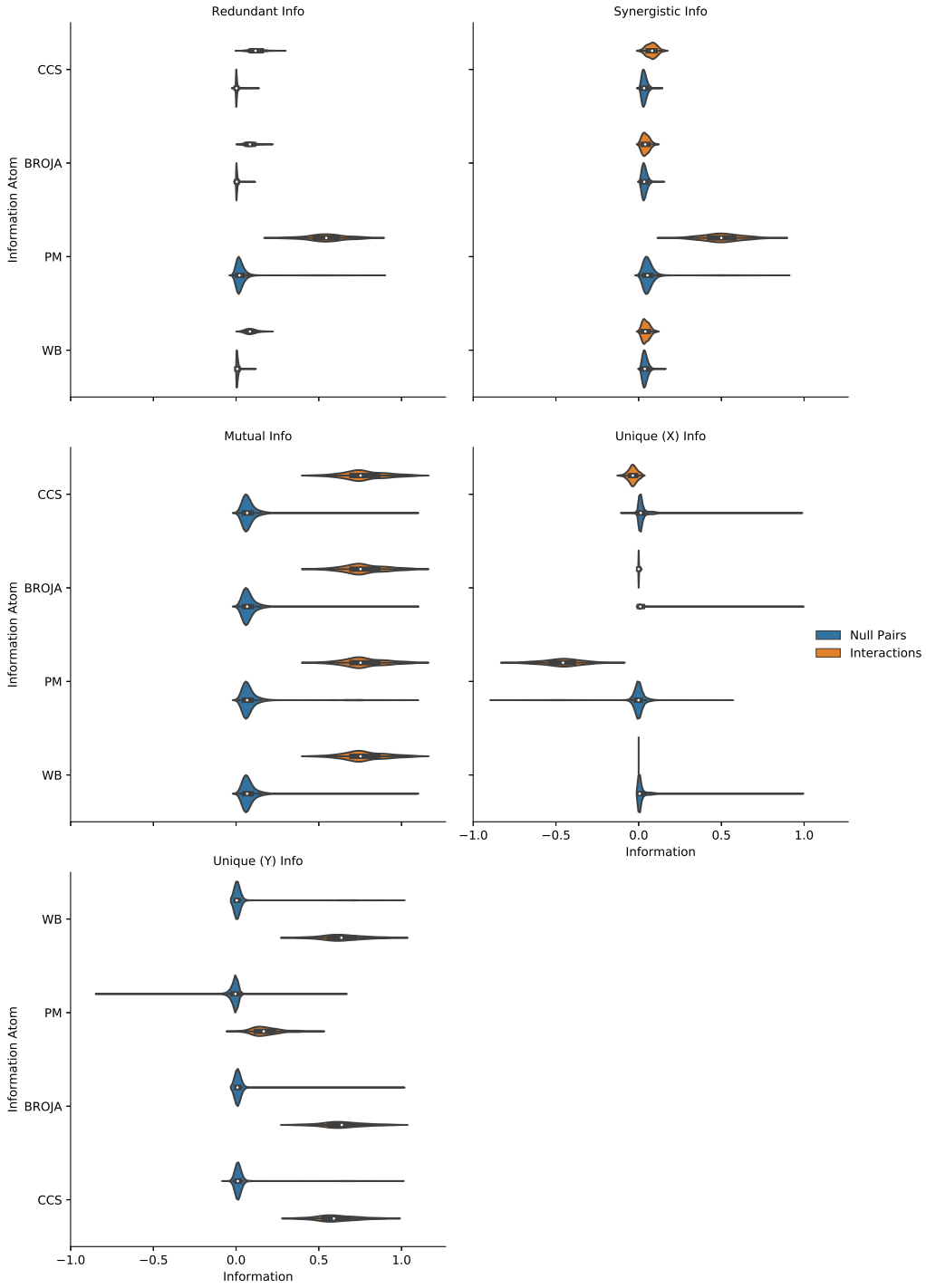


Figure 3.4: **Bivariate PID atoms for interactions in Experiment I** For the same bivariate PID data as in Fig 3.2 & 3.3, we present the full distributions of all PID atoms and mutual information, for both true interactions $(i, j) \in \mathcal{E}'$ and null pairs $(i, j) \notin \mathcal{E}'$. We can see that the $I_\cap^{\mathrm{PM}}$ PID assigns more synergy S^{PM} to the interactions and negative unique information U_X^{PM} to the switch gene.

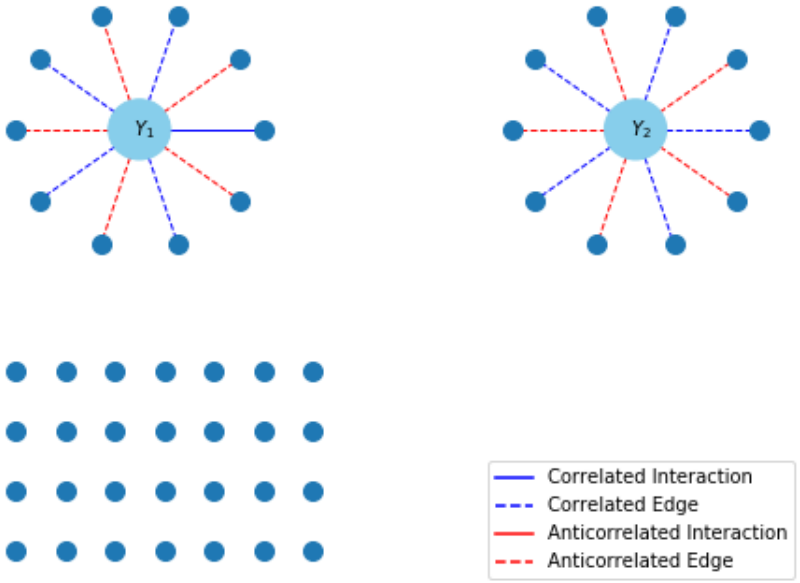


Figure 3.5: **G**ene interaction network topology for Experiment II For our second experiment, we simulated a network of the same size as before ($N = 50$), this time with with 20 edges, arranged in two 10-stars. The drug response is determined by the sum of a sigmoidal switch interaction on the first star and a univariate signal βY_2 from the hub of the second star, for a given parameter β (Eq. (3.3.2)).

For this experiment and the next (Exp. III), we used the same batch size and discretization method as in Experiment I. We used 30 simulated batches for each parameter choice, as we observed no meaningful change after this.

The goal of this type of network is to disentangle the PID atoms from mutual information, that is, $R(T; X_i, Y_j), U_X(T; X_i, Y_j), U_Y(T; X_i, Y_j)$ and $S(T; X_i, Y_j)$ from the total mutual information $I(T; X_i, Y_j)$, for a given pair (X_i, Y_j). Mutual information does not necessarily distinguish between a true interaction hub-spoke pair (X_1, Y_1) on the first star and a 'fake' interaction hub-spoke pair (X_j, Y_2) on the second star, since the latter hub contributes linearly to the response in Eq. (3.3.2). As discussed in Sec. 3.1.2.1, unadjusted

MI-based GRNs are known to struggle with type II errors. By adjusting the parameter β, we can make the MI of an interaction pair $I(T; X_1, Y_1)$ indistinguishable from the MI $I(T; X_j, Y_2)$ for any spoke $X_j, 11 \leq j \leq 20$ on the second star. For our experiments, we consider $\beta = 0.54$, as this was sufficient to demonstrate the desired effect. Data-adaptively choosing an optimal β to allow for maximal discrimination ability of the PID framework is a natural next step, though we do not pursue it further here.

From our simulations, we discover that $\hat{S}^{\min}$ is able to distinguish the true interaction (X_1, Y_1) from any 'false' interaction (X_j, Y_2), independent of mutual information. By contrast, $\hat{S}^{\mathrm{PM}}$ only distinguishes when the mutual information is able do so. In Fig 3.6 and Table 3.2, we compare the ranked scores of synergy and mutual information between true and false interactions, for $\beta = 0.54$. In Fig 3.7, we vary β, and see that the difference in S^{PM} between true and false pairs tracks closely with the difference in MI. For β sufficiently large, the univariate signal βY_2 is a stronger informer of the response T then the true interaction, and thus the false pairs on the Y_2 star are ranked above the true interactions on the Y_1 star, in terms of both MI and S^{PM}. By contrast, $S^{\min}$ consistently separates true and false interactions, with the former mostly located in the top decile of synergistic pairs, and the latter hovering around the distributional median. What's most striking is that this synergy difference appears largely uncoupled from the ratio between the empirical mutual informations $\hat{I}(T; X_1, Y_1)$ and $\hat{I}(T; X_j, Y_2)$. At higher values of β, true pairs continue to demonstrate more synergy in the $I_\cap^{\min}$ lattice than false pairs despite the greater contribution of Y_2 to the value of T than X_1 and Y_1 combined.

Thus the crucial difference that we observe in this experiment is that the atoms of the $I_\cap^{\mathrm{PM}}$ PID track closely with mutual information, while the $I_\cap^{\min}$

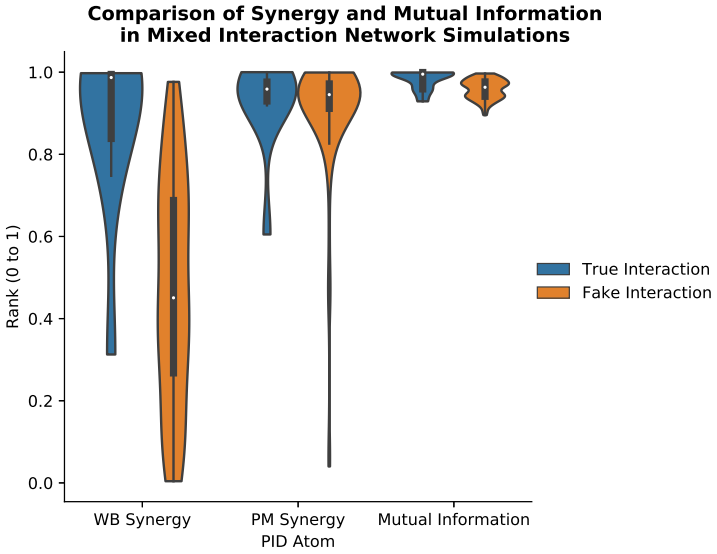


Figure 3.6: **Synergy ($S^{\min}$ and S^{PM}) and mutual information of mixed interactions in Experiment II** We compare the percentile distributions of PID synergy and mutual information for the network in Fig 3.5, with response T as in Eq. 3.3.2, with parameter $\beta = 0.54$. We examine the tendency of these statistics to distinguish between the true interaction (X_1, Y_1) and false interactions (X_j, Y_2). Since Y_2 contributes directly to the response, the mutual information of true and false interactions are comparably high, relative to the null distribution (RIGHT). We see that $S^{\min}$ (LEFT) seems to distinguish true and false interacting gene pairs, while S^{PM} does not (MIDDLE). See also Table 3.2.

atoms do not (Fig 3.8). Moreover, as can be seen in Fig 3.8, every atom of the $I_\cap^{\mathrm{PM}}$ PID tracks closely to mutual information. For false interactions on Y_2, R^{PM}, S^{PM}, and $U_{Y_2}^{\mathrm{PM}}$ all grow with mutual information, and $U_{X_j}^{\mathrm{PM}}$ shrinks, becoming more negative. By contrast, for the $I_\cap^{\min}$ PID of false interactions, $R^{\min}$ and $S^{\min}$ retain the same low value, $U_{X_j}^{\min} \equiv 0$, and only $U_{Y_2}^{\min}$ grows linearly with MI. This is what we would intuitively expect, as mutual information is growing with β, and only Y_2 contributes to the βY_2 term in Eq. (3.3.2).

Although the $I_\cap^{\min}$ PID assigns less synergy to true interactions than the $I_\cap^{\mathrm{PM}}$ PID, it nonetheless assigns a consistently low value of synergy to fake interactions $S^{\min}(T; X_j, Y_2)$, regardless of the mutual information value $I(T; X_j, Y_2)$. We will demonstrate in Proposition 7 that $S^{\min}$ must be zero for such false interactions, regardless of the dependency structure between X and

Y.

PID	Info. Atom	KS Stat	pval	Mean Rank (Intxn)	Mean Rank (Null)
WB	$\mathrm{S}^{\min}$	0.748	8.882e-16	0.918	0.468
WB	$\mathrm{R}^{\min}$	0.652	1.507e-11	0.985	0.808
WB	$\mathrm{U}_X^{\min}$	0.178	3.299e-01	0.205	0.133
WB	$\mathrm{U}_Y^{\min}$	0.715	3.708e-14	0.936	0.967
WB	Total MI	0.356	1.577e-03	0.975	0.965
PM	S^{PM}	0.152	5.241e-01	0.942	0.940
PM	R^{PM}	0.219	1.345e-01	0.951	0.957
PM	$\mathrm{U}_X^{\mathrm{PM}}$	0.422	7.807e-05	0.093	0.041
PM	$\mathrm{U}_Y^{\mathrm{PM}}$	0.237	8.397e-02	0.883	0.933
PM	Total MI	0.356	1.577e-03	0.975	0.965

Table 3.2: Kolmogorov-Smirnov comparison of PID atoms, as ranked scores, between true and fake interactions from our second network simulation experiment (Figs 3.5-3.7). Here, the 'true' interaction is the spoke-hub pair (X_1, Y_1) that contributes the sigmoidal term $g(X_1, Y_1)$ to the response in Eq (3.3.2), while the 'fake' interaction is any spoke-hub pair (X_j, Y_2) on the second star. Note that, since Y_2 contributes a univariate signal to (3.3.2), the mutual informations of the two types of pairs are comparable, while the WB synergy is not.

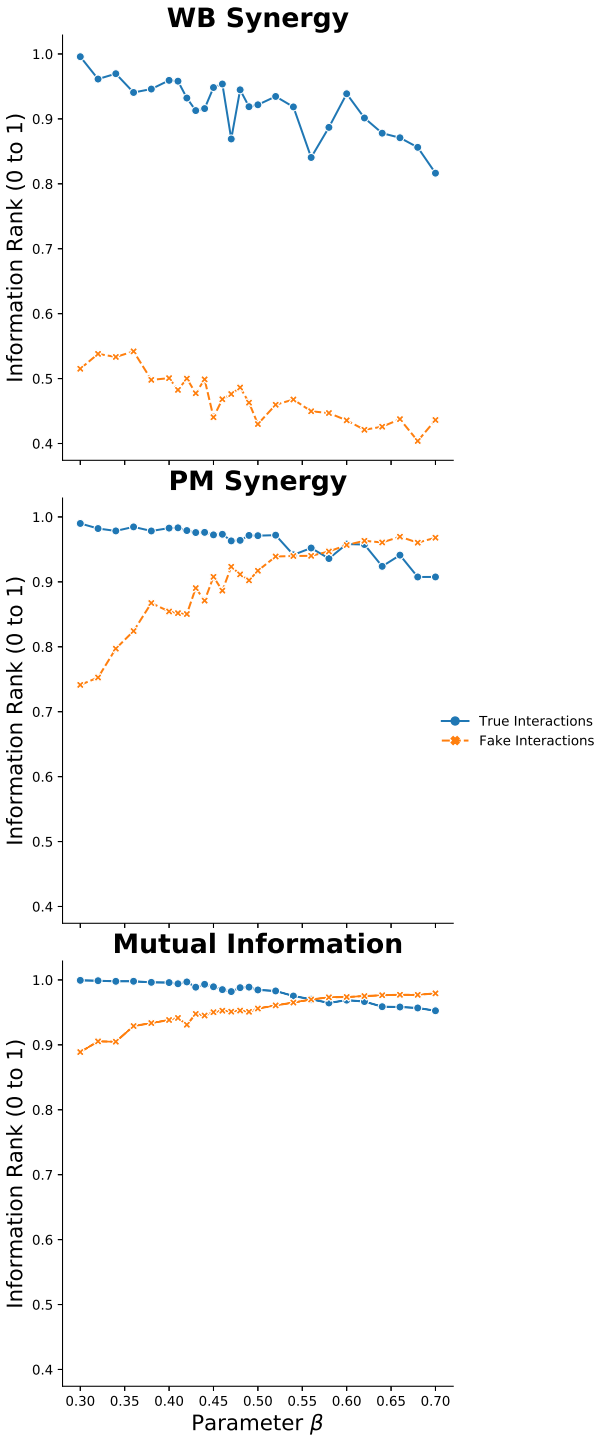


Figure 3.7: **Relative synergy ($S^{\min}$ and S^{PM}) of true and false interactions in Exp. II as a function of parameter β in Eq. (3.3.2)** Expanding upon our analysis of the network in Fig 3.5 & 3.6, we examine the ranked scores of synergy and mutual information as we vary the parameter β (Eq. (3.3.2)).

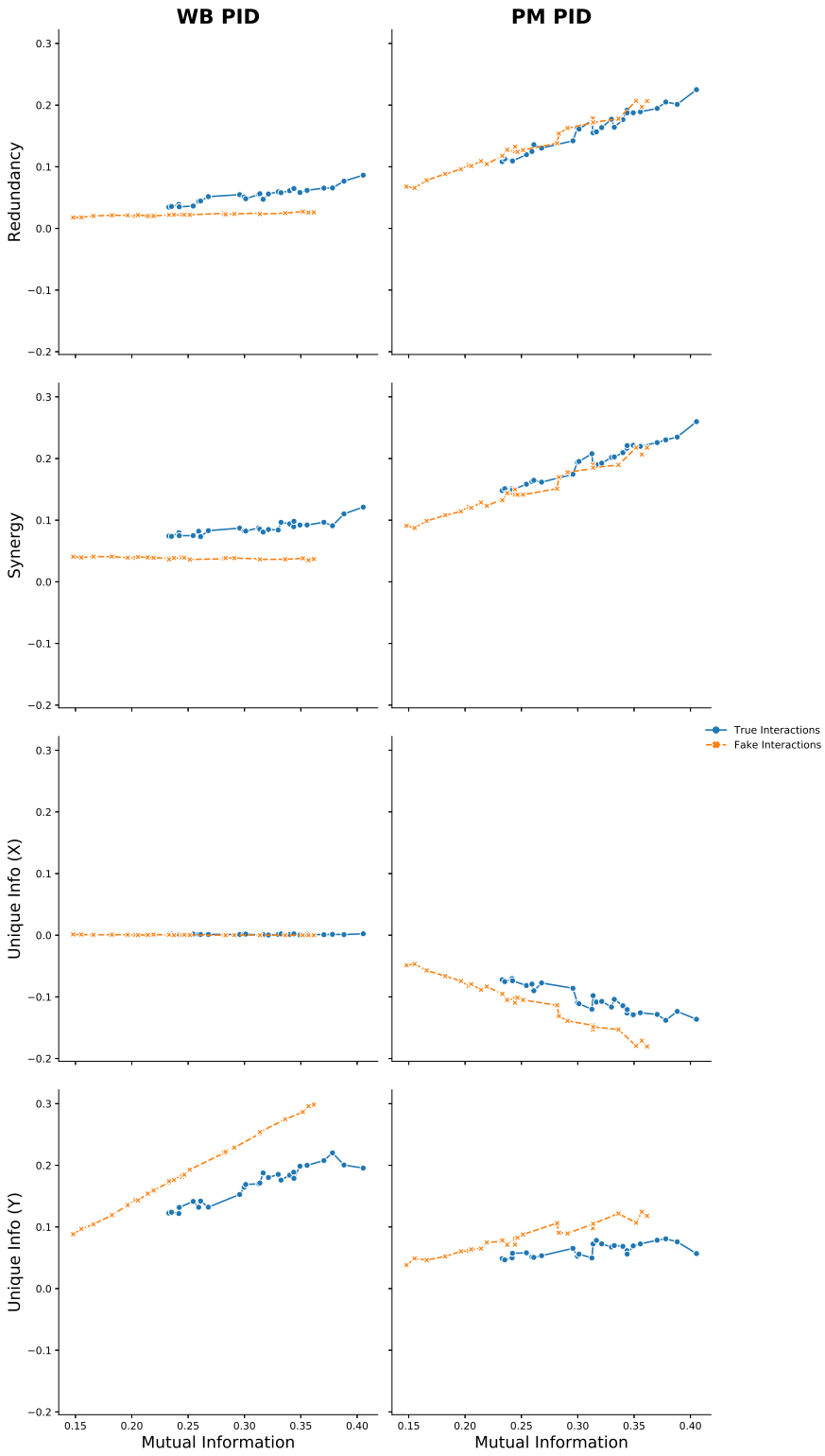


Figure 3.8: **Relationship between mutual information and bivariate $I_\cap^{\min}$ and $I_\cap^{\mathrm{PM}}$ atoms, for mixed interactions in Experiment II .** In this figure, we display the relationship between mutual information and the (unranked) PID information atoms for Experiment B (Figs 3.5-3.7). Each connected scatter plot maps mutual information against a PID atom, for true and false interactions.

The inability of MI and S^{PM} to distinguish between real and fake interactions suggest their limitations as methods of network analysis. Consider the application in which we wish to infer a mechanistic gene network for a drug's effectiveness. Many oncogenes are major hubs in the underlying human gene network, and activate multiple, highly interdependent pathways. It is quite likely that many such hubs correspond to the Y_2 in our simulations, in the sense that they partially predict or facilitate drug response. Even if we were to limit the candidate edges of our network inference task to known edges (X_j, Y_2) from the literature for a given hub, mutual information could falsely implicate many such edges despite the irrelevance of X_j to the response. Similarly, although S^{PM} demonstrates a sensitivity to interactions in Experiment I (Section 3.3.1), this experiment demonstrates that it lacks the specificity to distinguish between true interacting gene pairs and those with only one relevant gene.

Considering S^{PM} as a method of network inference, the previous experiment provided a point in its favor due to apparent sensitivity, while this experiment provides a point against due to the lack of specificity. In our final experiment to follow, we will return to the previous network-interaction topology in order to better examine the apparent sensitivity of the $I_{\cap}^{\mathrm{PM}}$ PID.

3.3.3 Experiment III

Our choice of a sigmoidal switch interaction kernel corresponds to an approximation of boolean or otherwise discrete logic circuits within the 'omic regulatory system of a biological organism. The gene X in Eq. (E4) is the 'switch' gene, with higher values of X turning the interaction on, and lower values turning it off. We allow the parameter α to recenter this transition, which is equivalent to recentering X. In Experiment I , we chose $\alpha = 0$, which allows

both the 'on' and 'off' regime to be observed with high probability. However, by choosing another α, the interaction can be made to default to one state or another.

In our final experiment, we again simulate the network in Fig 3.1, with the same response T as in Eq. 3.3.1. This time, we allow α to vary. As before, we see that even as we vary α, S^{PM} consistently places the interactions in the top 5% of synergistic pairs, while S^{min} does not (Fig 3.9). However, insofar as we expect a change in α to affect the balance of information between genes X and Y in the PID of $I(T; X, Y)$, the behavior of S^{PM} is counterintuitive. If we accept that S^{PM} is tracking the MI of the interaction itself, i.e. $I(T; X, Y)$, as was the case in the previous experiment, then S^{PM}'s behavior makes sense.

A clearer picture emerges when we examine the individual atoms of each bivariate PID in Fig 3.10. We see that the proportional amount of positive and negative information assigned to each atom of the $I_{\cap}^{\mathrm{PM}}$ PID remains relatively constant across $\alpha \in [-4, 4]$. In the $I_{\cap}^{\mathrm{min}}$ PID framework, by contrast, we will see that as α increases, S^{min} also increases modestly whereas U_Y^{min} decreases. This aligns with the expected behavior of our interactions. Indeed, consider the second-order Taylor expansion of our kernel (E4) about the mean $(X, Y) = (0, 0)$:

$$g(x, y) \approx \underbrace{\frac{1}{1 + e^{\alpha}}}_{\partial_y g} y + \underbrace{\frac{e^{\alpha}}{(1 + e^{\alpha})^2}}_{\partial_{x,y} g} xy. \tag{3.3.4}$$

The term $\partial_y g$ dominates $\partial_{x,y} g$ for lower α, but the terms converge as α in-

creases. That is

$$(\partial_y g, \partial_{x,y} g) \to c(\alpha) \left(\frac{1}{2}, \frac{1}{2}\right) \text{ as } \alpha \to \infty, \tag{3.3.5}$$
$$\text{where } c(\alpha) = |\partial_y g| + |\partial_{x,y} g|.$$

Of course, our response in Eq. (3.3.1) is the sum of four such interactions: $T = f(X_1, X_2, X_3, X_4, Y) = \sum_i g(X_i, Y)$. Expanding this as above gives us

$$f(x_1, x_2, x_3, x_4, y) \approx \underbrace{\frac{4}{1+e^\alpha}}_{\partial_y} y + \sum_{i=1}^{4} \underbrace{\frac{e^\alpha}{(1+e^\alpha)^2}}_{\partial_{x_i,y} g} x_i y \tag{3.3.6}$$

and thus

$$(\partial_y f, \partial_{x_i,y} f) \to c(\alpha) \left(\frac{1}{2}, \frac{1}{8}\right) \text{ as } \alpha \to \infty, \tag{3.3.7}$$
$$\text{where } c(\alpha) = |\partial_y g| + \sum_{i=1}^{4} |\partial_{x_i,y} g|.$$

We might expect, then, that for any of our four interactions (X_i, Y), most of the mutual information $I(T; X_i, Y)$ will be located within a PID in the unique information U_Y in proportion to magnitude of $\partial_y f$ relative to that of $\partial_{xy} f$. We see in Fig 3.10 that $I_\cap^{\min}$ *and not* $I_\cap^{PM}$ locates most of the information in $U_Y^{\min}$. In Fig 3.11, we directly compare $U_Y^{\min}$ and U_Y^{PM} to the relative value of $\partial_y f$ as above $(\partial_y f / c(\alpha))$, and see that $U_Y^{\min}$ nearly tracks this expression, unlike U_Y^{PM}. This exposition suggests that the $I_\cap^{\min}$ framework demonstrates the more intuitive behavior for a PID.

It not immediately clear whether the joint sensitivity of T to both X and Y, captured in $\partial_{xy} f$, ought to be accounted for as synergistic or as redundant information. On the one hand, the contribution of either predictor to this

term, and thus to T, depends on and modifies the value of the other, suggesting synergy. On the other hand, for our unit variance, zero centered predictors X and Y, their correlation is $\rho_{X,Y} = \mathbb{E}XY$, the monomial associated to $\partial_{xy}g$ in Eq. 3.3.6. We expect any joint information found in the correlation between X and Y to be accounted for in redundancy, not synergy. After all, if two variables are near-perfectly correlated, we almost certainly know one from the other, and thus any information provided by either regarding T will follow from knowledge of only one. To get around this conceptual uncertainty, we look at the sum of synergy and redundancy in Fig 3.11, both $S^{\min} + R^{\min}$ and $S^{\mathrm{PM}} + R^{\mathrm{PM}}$, and compare these to the relative value of $\partial_{xy}f$ as above $(\partial_{xy}f/(c(\alpha))$. Much in the same way that $U_Y^{\min}$ tracks with $\partial_y f$, we see that the joint synergy-redundancy of the $I_\cap^{\min}$ PID loosely tracks with the second-order term $\partial_{xy}f$. By contrast, the $I_\cap^{\mathrm{PM}}$ PID assigns a good deal more synergistic and redundant information than the relative magnitude of this joint sensitivity would suggest.

Moreover, as in the previous experiments, U_X^{PM} is consistently negative in more-or-less constant proportion to the total MI of the interaction (Fig 3.10, as compared to Figs 3.8 & 3.4 for previous experiments). We hypothesize that this feature is important to the non-specific sensitivity of the $I_\cap^{\mathrm{PM}}$ atoms, and the synergy S^{PM} in particular, to any pairs of genes providing information on the response. As we have seen in this experiment, the sensitivity of S^{PM} completely ignores the relative sensitivity of the response to the predictors X and Y, which we posit is the analytic analogue of ignoring the balance of information between X and Y concerning T.

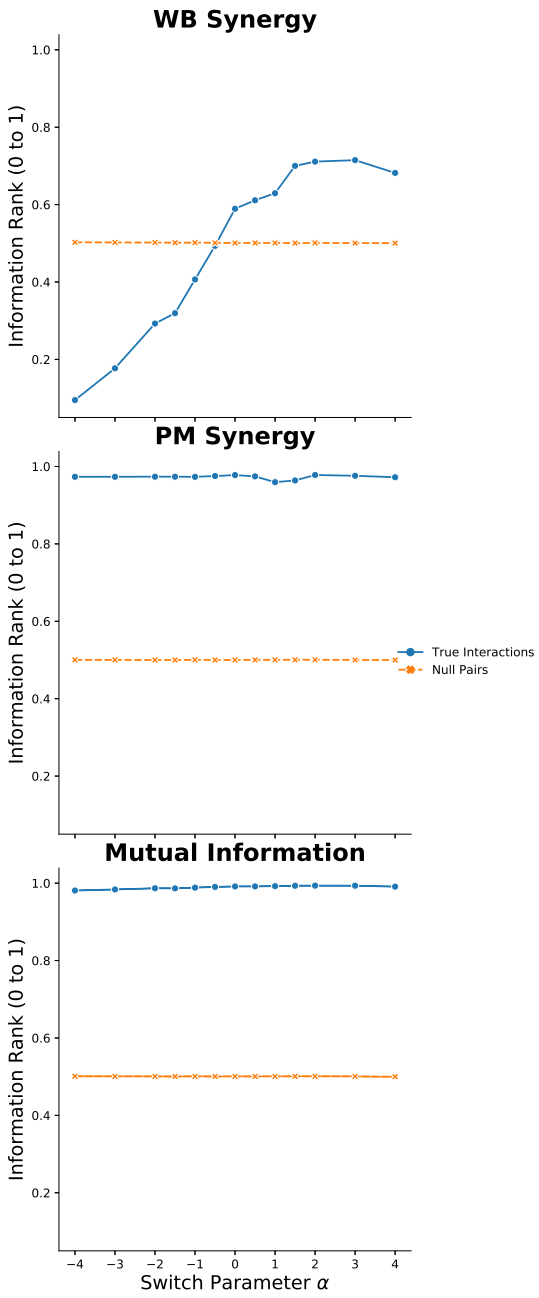


Figure 3.9: **Synergy and mutual information of interactions as a function of switch parameter α, from Experiment III** From the simulations in Experiment III , we compare the mean ranked scores of PID synergy and MI for both interacting and non-interacting pairs, as we vary parameter α for the network response in Eq. (3.3.1). We see that, for the four interacting pairs $(X_i, Y), i = 1, ..., 4$, both MI and S^{PM} will rank true interactions in the top 5% of synergistic pairs, while $S^{\min}$ will not. The insensitivity of these two metrics to changes in α is not necessarily desirable, and certainly compatible with the non-specificity observed in Experiment II . For values of α far below zero, the dependence of T upon the switch genes X_i should be significantly decreased, even though the MIs of $I(T; X_i)$ and $I(T; Y)$ ought to be comparable.

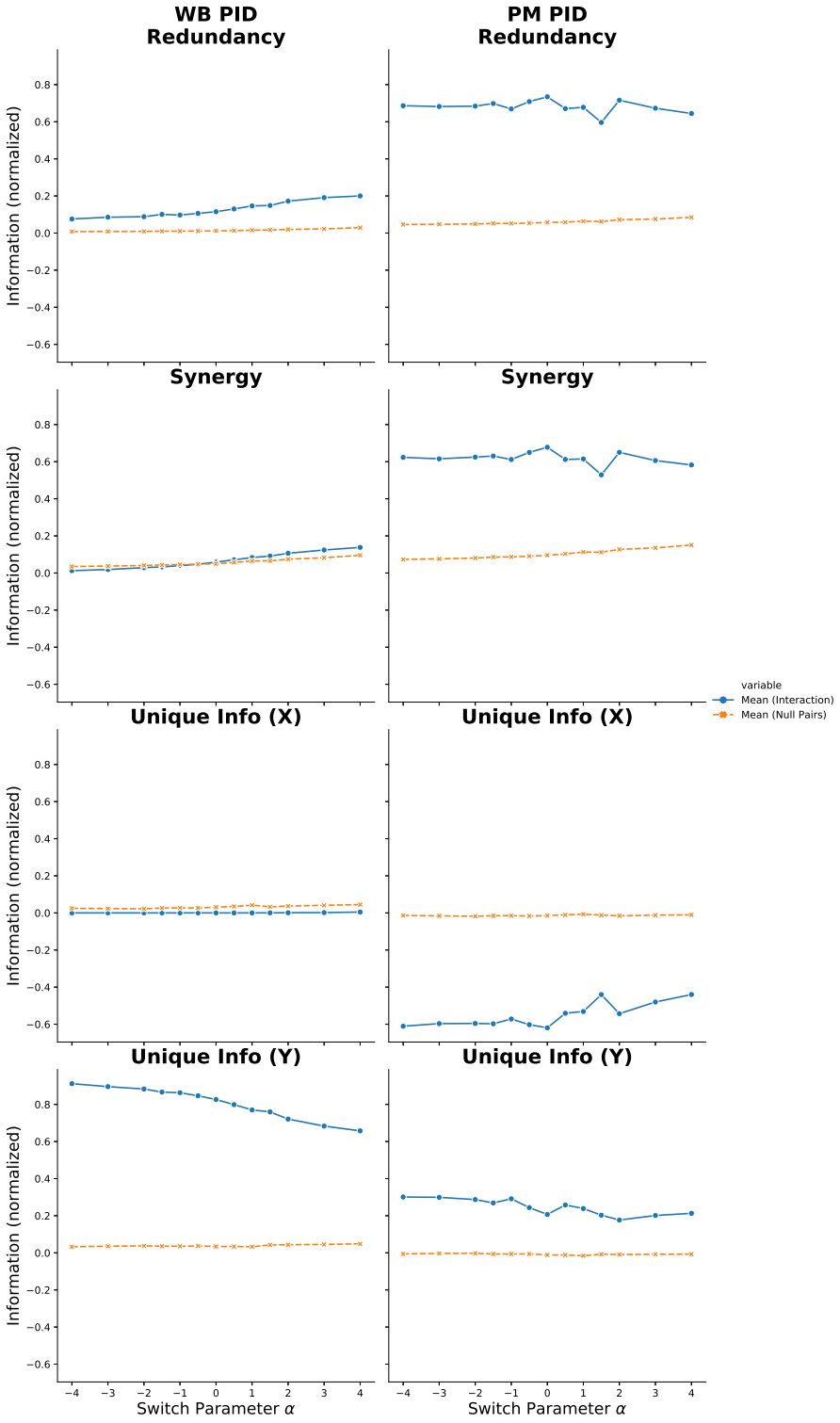


Figure 3.10: **Bivariate PID atoms (normalized) as a function of the switch parameter α from Experiment III** We present the four PID atoms of the $I_\cap^{\min}$ and $I_\cap^{\mathrm{PM}}$ PIDs for Exp. III , in which we vary the parameter α in the response (Eq. 3.3.1). Here, each atom has been normalized by the average MI of the interaction pairs, so as to discount for the effect of total MI varying with α.

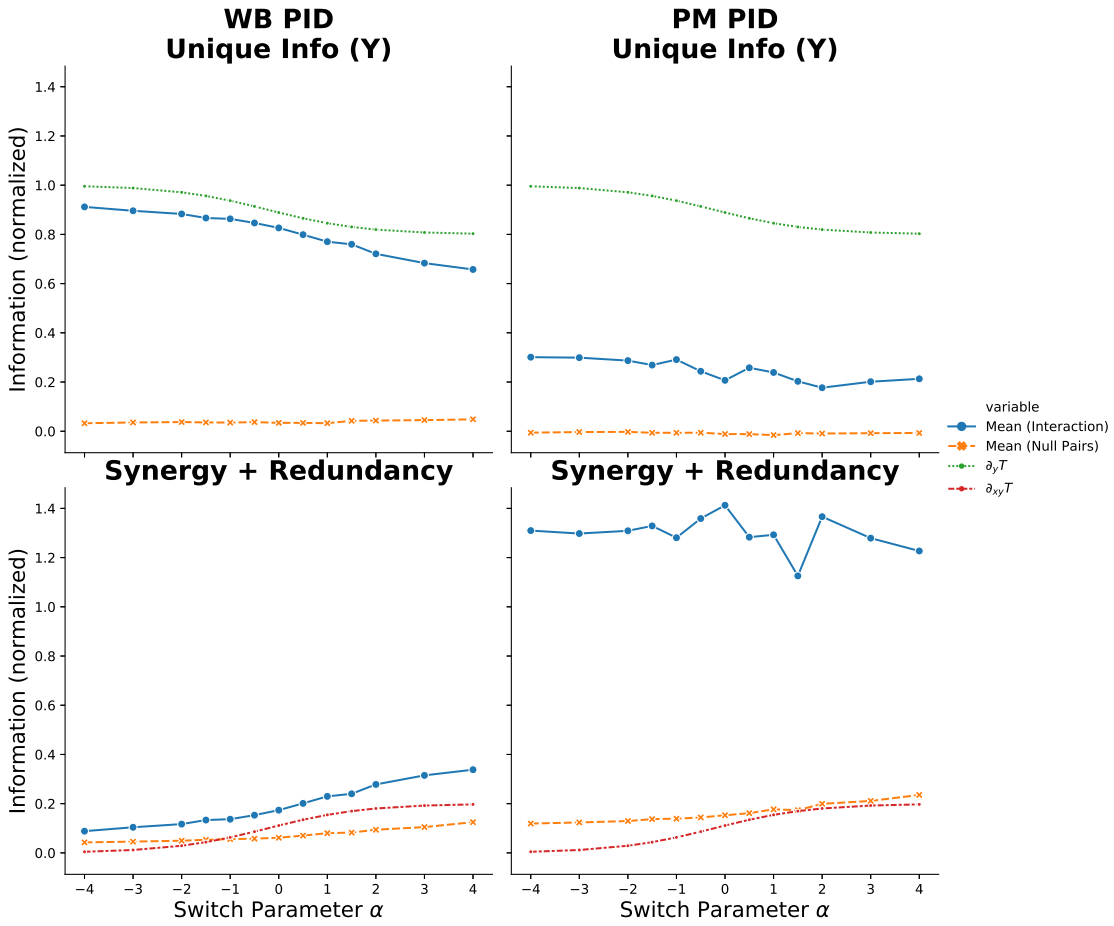


Figure 3.11: **Relationship between bivariate PID atoms and sigmoidal kernel derivatives in Experiment III** Continuing our analysis of Exp. III, we examine both the unique info U_X of the hub gene Y and the sum of synergy and redundancy $S + R$ for our interaction pairs (X_i, Y), in light of the expansion of the response in Eq. (3.3.6). We see that, as we vary α, $U_Y^{\min}$ tracks the normalized term $\partial_y f / c(\alpha)$, where $c(\alpha) = |\partial_y f| + |\partial_{xy} f|$, while $S^{\min} + R^{\min}$ tracks $\partial_{xy} f / c(\alpha)$. The former term corresponds to the linearized sensitivity of the response T to Y alone near $X = Y = 0$, while the latter tracks the sensitivity to the product XY.

3.3.4 Experimental Summary: Sensitivity and Specificity of PID Synergies

The PID problem is the task of quantifying—within the context of a statistical model—the redundant, synergistic, and unique information provided by sub-

sets of predictor variables about a target. Our interest in the problem stems from the idea that the PID toolbox holds promise for the task of network inference. However, which PID framework is appropriate to this task, and how it ought to be employed, are both unclear. This is still an active area of research, and we do not presume to provide either a tentative answer or prospective methodology in this chapter. We rather hope to illuminate desirable (and undesirable) features that a PID might possess to be appropriate (or inappropriate) for the task of gene network inference.

In Experiment I , we investigated the potential of four bivariate synergies, $S^{\min}, S^{\mathrm{PM}}, S^{\mathrm{BROJA}}$, and S^{CCS}, as metrics of synergy network inference, using the network in Fig. 3.1. At first glance, S^{PM} outperformed the others in this respect, as it was the only synergy to consistently rank interacting pairs above the 95% percentile using the empirical distribution of all gene pairs from the simulated experiment. However, this was a network in which MI itself was highly predictive of interactions, independent of any PID.

In Experiment II , we created a network with a simpler response, composed of one sigmoidal interaction $g(X_1, Y_1)$ and one univariate signal βY_2 (Eq. (3.3.2)). This allowed us to simulate a situation in which MI is an unreliable metric for inferring interactions, as pairs that include the univariate contributor Y_2 will have comparable total information about the response as (X_j, Y_2). In this situation, we saw that S^{PM} was **sensitive but non-specific** for interactions. By contrast, we saw that $S^{\min}$ ranked false interactions (X_j, Y_2) around the mean for all pairs within the network, while ranking (X_1, Y_1) highly. Although $S^{\min}$ has not demonstrated the desired sensitivity to network interactions that might hope for, it nonetheless has a quality of **specificity** to its synergy. In highly interdependent networks, such as gene-regulatory networks, specificity is crucial, as many genes are involved in and

activate multiple overlapping pathways. Identifying pairs that are jointly important to a response requires a metric that can distinguish joint and unique contributions to the information of that response.

In Experiment III , we adjust the sigmoidal switch interaction kernel itself, via the parameter α in Eq. (3.3.1). This allows us to explore the relationship between the analytic properties of a given network interaction kernel and the observed PIDs in network simulations using that kernel. Put differently, the suitability of a given PID ($I_\cap^{\min}$ or $I_\cap^{\mathrm{PM}}$ or another) for network inference may depend upon computable properties of the interactions under investigation, or at least the computable properties of a reasonable model of them.

Our network interactions, however, live in continuous probability space, whereas PID has primarily been investigated for discrete variables [77]. In the next few sections, we extend the $I_\cap^{\min}$ and $I_\cap^{\mathrm{PM}}$ PIDs to continuous variables in order to continue our analytic investigation of the potential of these PIDs as tools of network inference. These are first steps, but they already shed considerable light on the peculiarities of the $I_\cap^{\min}$ and $I_\cap^{\mathrm{PM}}$ PIDs, previewed already in the experiments of this section. In the continuous extension of PID, it becomes evident that the specificity issue of $I_\cap^{\mathrm{PM}}$ is structural to the definition. We will see that, whereas the continuous extension of $I_\cap^{\min}$ will have $U_X^{\min} \to 0$ as a predictor X tends toward conditional independence of the target, U_X^{PM} will remain negative. In this way, we take the first steps of an analytic argument for the non-specific nature of the $I_\cap^{\mathrm{PM}}$ PID that we see in these experiments, as the persistence of negative unique information for irrelevant genes inflates the synergistic information in our discrete estimates.

3.4 Mathematical Notation and Formatting

A few notes about the style of this work are in order.

First, we suggest that this document be read within software or on an e-reader amenable to internal hyperlinks, as we make frequent use of them. Due to its length, we aimed to make this paper as interconnected and internally referential as possible. As is typically the case, the author aimed to write a work that he would find enjoyable to read and later review, and perhaps erred on the side of repetition: both themes and internal references will be repeatedly reintroduced in order to maintain the conceptual forest for the trees.

3.4.1 Equation Numbering

Throughout this work, we will generally number equations by section. Equations that are especially important and frequently referenced throughout the paper will be denoted with a capital letter - number combination. Moreover, within proofs, we will sometimes number equations with $\star$'s when they will be used later within that same proof, in order to distinguish them from less interesting intermediate equations. We will typically denote a sequence of multiline, successively equations with a single label.

3.4.2 Logarithms

Throughout this work, we use logarithms to define our information quantities. The choice of base is arbitrary, for the purpose of our investigation. In keeping with standard conventions, we will use bits ($\log_2$) for discrete variables, and nats ($\log_e$) for continuous variables, denoting both as log.

3.5 Information Theory Preliminaries

Before we define the $I_\cap^{\min}$ and $I_\cap^{\mathrm{PM}}$ PIDs for continuous random interactions, we will first present some definitions and concepts from information theory.

Symbol	Description	Definition
(Ω, μ)	Sample space	NA
$X, Y, Z, T \in \mathcal{R}$	Random variables	NA
$\boldsymbol{X}, \boldsymbol{Y} \in \mathcal{P}(\mathcal{R})$	Random sources	Sec. 3.5.1
$(\mathcal{A}_{\boldsymbol{X}}, \mu_{\boldsymbol{X}})$	Probability Space induced by $\boldsymbol{X}$	NA
$p_{\boldsymbol{X}}$	Density (continuous) or mass (discrete) function	NA
X, Y, X_i, Y_i	Predictor variables	NA
T	Target variable	NA
$u(\boldsymbol{X}) \overset{\omega}{=} v(\boldsymbol{Y})$	Almost surely equal as r.v.'s	Def. 3
$D(p_{\boldsymbol{X}} \vert p_{\boldsymbol{Y}})$	KL Divergence (when $\mathcal{A}_{\boldsymbol{X}} = \mathcal{A}_{\boldsymbol{Y}}$)	Def. 8
$H(\boldsymbol{X})$	Entropy (discrete $\boldsymbol{X}$)	Def. 6
$h(\boldsymbol{X})$	Differential entropy (continuous $\boldsymbol{X}$)	Def. 6
$I(\boldsymbol{X}; \boldsymbol{Y})$	Mutual information	Def. 7
$I_\cap$	Generic redundancy function	Def. 13
$I_\cap^{\min}$	WB redundancy function	Def. 14, [125]
$I_\cap^{\mathrm{PM}}$	Pointwise mutual redundancy function	Def. 14, [45]
$I_{\boldsymbol{X}} : \mathcal{A}_T \to \mathbb{R}$	Specific information function	Def. 9, [125]
$I_\cap^{\mathrm{PM},+}$	Redundant specificity function	Def. 15, [45]
$I_\cap^{\mathrm{PM},-}$	Redundant ambiguity function	Def. 15, [45]
$\mathfrak{d} = (R, S, U_X, U_Y)$	Bivariate PID of $I(T; X, Y)$	Eqs. E1-E3

Table 3.3: **Table of commonly used notations.**

Readers familiar with the fundamentals of both discrete and continuous entropies and informations are welcome to move ahead. To more efficiently compartmentalize background knowledge, this section will be split into three parts. In Subsection 3.5.1, we specify our terminology of discrete and continuous random sources. In Subsection 3.5.2, we introduce and define information theoretical quantities for discrete sources and continuous sources that admit densities. Finally, in Subsection 3.5.3, we provide a more general framework to handle edge cases and, at times, to simplify computations.

3.5.1 Random Variables and Random Sources

Let $(\Omega, \mathcal{B}, \mu)$ be our assumed probability space. A random variable X is a measurable mapping $\Omega \to \mathcal{A}_X$, where $\mathcal{A}_X$ is the alphabet of X. In this paper, all alphabets are subsets of $\mathbb{R}^k$ for some k. We say that X is a discrete variable if $\mathcal{A}_X$ is countable. We say that X is continuous if it admits a density p_X on $\mathbb{R}$. More precisely, we say that X is continuous if it induces a measure μ_X on $\mathbb{R}$ that is absolutely continuous with respect to Lebesgue measure λ (denoted $\mu_X \ll \lambda$), in which case the density is the Radon-Nikodym derivative $p_X = \frac{d\mu_X}{d\lambda}$.

It is more useful to define the notion of a **random source,** which we take to be a finite collection of random variables. One can think of a random source $\boldsymbol{X} = \{X_1, ..., X_N\}$, with associated alphabets $\mathcal{A}_{X_1}, ..., \mathcal{A}_{X_N}$ as a random variable $\boldsymbol{X}$ on the product alphabet $\mathcal{A}_{\boldsymbol{X}} = \bigtimes_i \mathcal{A}_{X_i}$. Nonetheless, we distinguish the two concepts for clarity, and reserve 'random variable' to refer to univariate real-valued variables. Moreover, we may elect to treat a random source $\boldsymbol{X} = \{X_i\}_i$ as a random vector $\boldsymbol{X} = (X_i)_i$ when convenient, e.g. when we want to admit a density $p_{\boldsymbol{X}}$ on $\mathbb{R}^k$.[6] We similarly use $(\mathcal{A}_{\boldsymbol{X}}, \mu_{\boldsymbol{X}})$ to denote the

[6] We implicitly assume throughout that every function of random sources that makes use

induced probability space. We will denote the image of an event $E \subset \Omega$ in $\mathcal{A}_{\boldsymbol{X}}$ as $\pi_{\boldsymbol{X}}(E)$, where $\pi_{\boldsymbol{X}}$ is here synonymous with the random vector $\boldsymbol{X}$ as a mapping $\Omega \to \mathcal{A}_X$.

Definition 2 (Random Sources)**.** *A finite collection of random variables* $\boldsymbol{X} = \{X_1, ..., X_k\}$ *on the space* (Ω, μ) *is called a* ***random source****.*

- *If every variable* X_i *is discrete, we say that* $\boldsymbol{X}$ *is a discrete source.*
- *If every variable* X_i *is continuous, we say that* $\boldsymbol{X}$ *is* ***marginally continuous****.*
- *We say that* $\boldsymbol{X}$ *is* ***(jointly) continuous*** *if the joint distribution of* $X_1, ..., X_k$ *admits a joint density* $p_{\boldsymbol{X}}$*. Equivalently,* $\boldsymbol{X}$ *is jointly continuous if it induces a measure* $\mu_{\boldsymbol{X}}$ *on* $\mathbb{R}^k$ *such that* $\mu_{\boldsymbol{X}} \ll \lambda^k$*, in which case* $p_{\boldsymbol{X}} = \frac{d\mu_{\boldsymbol{X}}}{d\lambda^k}$*.*
- *We say that* $\boldsymbol{X}$ *is* ***non-degenerate*** *if it is discrete or continuous. It is* ***degenerate*** *otherwise.*

We will focus upon non-degenerate sources for most of this section. For the density of a continuous source, we may represent it as $p(\boldsymbol{X}) := p_{\boldsymbol{X}}(\boldsymbol{X})$ when the subscript is obvious from context. Moreover, for a continuous source, we exclude from the alphabet $\mathcal{A}_{\boldsymbol{X}}$ all points $\boldsymbol{x}$ for which $p_{\boldsymbol{X}}(\boldsymbol{x}) = 0$. We may still allow singular sets of measure zero that are possible, in the sense that arbitrary neighborhoods have positive probability.[7]

We highlight a few immediate properties.

---

of their joint densities is invariant under permutation of labels. For instance, when we define entropy $H(\{X_1, ..., X_k\})$, we use the density $p_{X_1,...,X_k}(x_1, ..., x_k)$, but the entropy would be the same if we used $p_{X_{\sigma(1)},...,X_{\sigma(k)}}(x_{\sigma(1)}, ..., x_{\sigma(k)})$ instead.

[7] For instance, in Section 3.9, we will have that the random source $\{Y, T\}$ admits a density $p_{Y,T}$ almost everywhere, but not when $Y = 0$, an event with zero probability. Nonetheless, any neighborhood of $(Y, T) = (0, 0)$— i.e. any event $E = \pi^{-1}_{Y,T} B_\epsilon(0, 0)$, where B_ϵ denotes an open ϵ-disk — will have positive probability. Thus, we include $(0, 0) \in \mathcal{A}_{X,T}$.

Proposition 1 (Remarks on Continuous Sources)**.** *For random sources* $\boldsymbol{X}, \boldsymbol{Y}$*, we have the following*

1. *If* $\boldsymbol{X}$ *is continuous, then it is marginally continuous.*
2. *If* $\boldsymbol{X}$ *is jointly (respectively, marginally) continuous, then every subset* $\boldsymbol{X}' \subset \boldsymbol{X}$ *is also jointly (respectively, marginally) continuous.*
3. *If* $\boldsymbol{X}$ *and* $\boldsymbol{Y}$ *are both marginally continuous, then so is* $\boldsymbol{X} \cup \boldsymbol{Y}$*.*
4. *If* $\boldsymbol{X}$ *and* $\boldsymbol{Y}$ *are both continuous, it is not necessary that* $\boldsymbol{X} \cup \boldsymbol{Y}$ *be continuous.*

The first two properties follow from the fact that the integration of a continuous function is continuous. The third property is trivial. The final property is demonstrated by the following counterexample.

Example 1 (Non-Degenerate Bivariate Gaussian)**.** *Suppose* $X, Y \sim N(\mathbf{0}, \Sigma)$ *where* $\Sigma = \begin{pmatrix} \sigma^2 & \sigma^2\rho \\ \sigma^2\rho & \sigma^2 \end{pmatrix}$*. Then* $\{X, Y\}$ *is a continuous source if and only if* $|\rho| < 1$*.*

Example 2 (Degenerate Bivariate Gaussian)**.** *If we allow* $\rho = \pm 1$ *in our previous example, we have identical Gaussians with the joint distribution* $p_{X,Y}(x, y) = \frac{1}{\sqrt{2\pi}}\delta(y \mp x)e^{-x^2/2}$*, which is not a density but a distribution or generalized function. We still have that* $\{X\}$ *and* $\{Y\}$ *are continuous sources, but* $\{X, Y\}$ *is only marginally continuous and degenerate.*

We will typically work within induced probability spaces $(\mathbb{R}^N, \mu_{\boldsymbol{X}})$ for a continuous source $\boldsymbol{X}$. However, in situations when $\boldsymbol{X} \cup \boldsymbol{Y}$ is not jointly continuous, we will instead often have to move between distinct spaces $(\mathbb{R}^N, \mu_{\boldsymbol{X}})$ and $(\mathbb{R}^N, \mu_{\boldsymbol{Y}})$ by a μ-preserving change of variables $\boldsymbol{X} \to \boldsymbol{Y}$. For instance, $\{X, Y, T\}$ will be jointly degenerate for noiseless interactions (Def. 11), but

every subset of this collection will admit a density in $\mathbb{R}$ or $\mathbb{R}^2$, and in Proposition 9 we relate the densities $p_{X,T}$ and $p_{Y,T}$ to $p_{X,Y}$. In Section 3.7, we will easily move between (X,Y)-space $(\mathbb{R}^2, \mu_{X,Y})$, (X,T)-space $(\mathbb{R}^2, \mu_{X,T})$, and (Y,T)-space $(\mathbb{R}^2, \mu_{Y,T})$ by using standard results regarding invertible transformations of Gaussian vectors. We emphasize that expectation $\mathbb{E}$, and all definitions that depend upon it, must properly be understood as an integral on the sample space (Ω, μ):

$$\mathbb{E}f(\boldsymbol{X}) = \int_{\mathcal{A}_{\boldsymbol{X}}} f(\boldsymbol{x}) p_{\boldsymbol{X}}(\boldsymbol{x}) d\boldsymbol{x} = \int_{\Omega} f(\boldsymbol{x}(\omega)) d\mu(\omega). \tag{3.5.1}$$

This is crucial for well-defined information quantities. We will leave this technicality implicit where we can, as we can dispense with it for most of our computations. However, in a few circumstances where we need to highlight the underlying probabilistic context, the following notation and terminology will serve as a helpful shorthand:

Definition 3 (ω-Equivalence)**.** *For the random sources $\boldsymbol{X}$ and $\boldsymbol{Y}$, consider the functions $f : \mathcal{A}_{\boldsymbol{X}} \to \mathbb{R}$ and $g : \mathcal{A}_{\boldsymbol{Y}} \to \mathbb{R}$. We say that $f(\boldsymbol{X})$ and $g(\boldsymbol{Y})$ are **almost surely equal as random variables,** denoted $f(\boldsymbol{X}) \overset{\omega}{=} g(\boldsymbol{Y})$, when the following holds:*

$$f(\boldsymbol{X}(\omega)) = g(\boldsymbol{Y}(\omega)) \text{ for } \mu\text{-a.e. } \omega \in \Omega. \tag{3.5.2}$$

*We may also regard this relationship as an equivalence relationship on function-source pairs. We will synonymously say that $(f, \boldsymbol{X})$ and $(g, \boldsymbol{Y})$ are **$\boldsymbol{\omega}$-equivalent**, denoted $(f, \boldsymbol{X}) \overset{\omega}{=} (g, \boldsymbol{Y})$:*

$$f(\boldsymbol{X}) \overset{\omega}{=} g(\boldsymbol{Y}) \iff (f, \boldsymbol{X}) \overset{\omega}{=} (g, \boldsymbol{Y}) \iff f(\boldsymbol{x}(\omega)) = g(\boldsymbol{y}(\omega)) \text{ a.e. } [\mu]. \tag{3.5.3}$$

What we refer to as ω-equivalence or a.s. equality of $f(\boldsymbol{X})$ and $g(\boldsymbol{Y})$ when considered as random variables refers to the essential equality of the pullback functions $\pi_{\boldsymbol{X}}^{\star} f$ and $\pi_{\boldsymbol{Y}}^{\star} g$ on Ω. We want to take care to distinguish this from (essential) equality of our functions on the Euclidean alphabets of $\boldsymbol{X}$ and $\boldsymbol{Y}$. Almost sure equality $f(\boldsymbol{X}) \overset{\omega}{=} g(\boldsymbol{Y})$ is a distinct property from, and neither sufficient nor necessary for, equality or near-equality of f and g as functions on coinciding alphabets $\mathcal{A}_{\boldsymbol{X}} = \mathcal{A}_{\boldsymbol{Y}}$.

Almost sure equality as random variables is more than sufficient for equality in expectation:

$$f(\boldsymbol{X}) \overset{\omega}{=} g(\boldsymbol{Y}) \Longrightarrow \mathbb{E} f(\boldsymbol{X}) = \mathbb{E} g(\boldsymbol{Y}). \tag{3.5.4}$$

Consider the following example.

Example 3. *Let $X \sim N(0,1)$ and $Y = 2X$, i.e. $X, Y \sim N(\mathbf{0}, \Sigma)$ where $\Sigma = \left(\begin{smallmatrix} 1 & 2 \\ 2 & 4 \end{smallmatrix}\right)$. Then $\{X, Y\}$ is marginally continuous and degenerate, with joint distribution:*

$$p_{X,Y}(x,y) = \delta(y-2x) p_X(x) = \frac{1}{\sqrt{2\pi}} \delta(y-2x) e^{-x^2/2} \tag{3.5.5}$$

$$= \delta(x - \frac{y}{2}) p_Y(y) = \frac{1}{2\sqrt{2\pi}} \delta(x - \frac{y}{2}) e^{-y^2/8}. \tag{3.5.6}$$

Consider the marginal distributions as random functions. In this case, $p_{\boldsymbol{X}}(\boldsymbol{X}) \overset{\omega}{\neq} p_{\boldsymbol{Y}}(\boldsymbol{Y})$, but instead $p_{\boldsymbol{X}}(\boldsymbol{X}) \overset{\omega}{=} 2 p_{\boldsymbol{Y}}(\boldsymbol{Y})$.

We may now define conditional sources for non-degenerate distributions, using their mass and density functions.

Definition 4 (Conditional Random Sources)**.** *Let $\boldsymbol{X}, \boldsymbol{Y}$ be random sources, either both discrete or jointly continuous. Let $\boldsymbol{y} \in \mathcal{A}_{\boldsymbol{Y}}$ such that $p_{\boldsymbol{Y}}(\boldsymbol{y}) > 0$.*

The conditional source $\boldsymbol{X}$ *given* $\boldsymbol{Y} = \boldsymbol{y}$ *is denoted* $\boldsymbol{X}_{|\boldsymbol{Y}=\boldsymbol{y}} = \boldsymbol{X}_{|\boldsymbol{y}}$*, and is the discrete (continuous) random vector with alphabet* $\mathcal{A}_{\boldsymbol{X}|\boldsymbol{y}} \subset \mathcal{A}_{\boldsymbol{X}}$*, with mass function (density):*

$$p(\boldsymbol{x}_i|\boldsymbol{y}_j) := \frac{p(\boldsymbol{x}_i, \boldsymbol{y}_j)}{p(\boldsymbol{y}_j)} \qquad \textit{(discrete sources)} \tag{3.5.7}$$

$$p_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{x}|\boldsymbol{y}) := \frac{p_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x}, \boldsymbol{y})}{p_{\boldsymbol{Y}}(\boldsymbol{y})} \qquad \textit{(continuous sources).}$$

Pointwise information is measured as a specific transformation of the of the distribution functions, typically referred to as the surprisal.

Definition 5. *Let* $\boldsymbol{X}$ *be a discrete or continuous random source. Then the pointwise surprisal* $s_{\boldsymbol{X}}$ *is defined as the function*

$$s_{\boldsymbol{X}}(\boldsymbol{x}) = \log \frac{1}{p(\boldsymbol{x})} \tag{3.5.8}$$

$$= \begin{cases} \log \frac{1}{p(\boldsymbol{x}_i)} & \textit{(discrete source)} \\ \log \frac{1}{p_{\boldsymbol{X}}(\boldsymbol{x})} & \textit{(jointly continuous source)} \end{cases} \tag{3.5.9}$$

For discrete sources, $s_{\boldsymbol{X}}$ *takes* $\mathcal{A}_{\boldsymbol{X}}$ *as its domain. For continuous sources,* $s_{\boldsymbol{X}}$ *is defined where* $\mu_{\boldsymbol{X}}$ *admits a nonzero density* $p_{\boldsymbol{X}}$*.*

If $\boldsymbol{X}$ *and* $\boldsymbol{Y}$ *are both discrete or jointly continuous, we define the conditional surprisal of* $\boldsymbol{x}$ *given* $\boldsymbol{Y} = \boldsymbol{y}$ *as*

$$s_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{x}|\boldsymbol{y}) = \log \frac{1}{p(\boldsymbol{x}|\boldsymbol{y})} \tag{3.5.10}$$

$$= \begin{cases} \log \frac{1}{p(\boldsymbol{x}_i|\boldsymbol{y}_j)} & \textit{(discrete source)} \\ \log \frac{1}{p_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{x}|\boldsymbol{y})} & \textit{(jointly continuous source)} \end{cases} \tag{3.5.11}$$

Surprisal can be thought of as the unlikelihood of a particular outcome. For discrete sources, these quantities are always positive, as probability mass

$p(\boldsymbol{x}_i)$ is always in $(0, 1)$. For continuous sources, this is no longer true (indeed, the conditional surprisal may be negative).

It is worth noting that conditional surprisal is the difference between the joint and individual surprisals, that is,

$$s_{\boldsymbol{X}|\boldsymbol{Y}} = s(\boldsymbol{x}, \boldsymbol{y}) - s(\boldsymbol{x}) \tag{3.5.12}$$

In the discrete case, it is clear that this difference is positive, as the intersection event $\boldsymbol{x} \cap \boldsymbol{y}$ must be more unlikely than $\boldsymbol{x}$. In the continuous case, this is quantity is not necessarily positive.

3.5.2 Discrete and Density Information Theory

The Shannon entropy, originally defined in [106], quantifies the uncertainty in the value of a discrete variable X. We introduce the general definition for a random source that admits a tractable pointwise 'probability', whether a proper probability mass function (if $\boldsymbol{X}$ is discrete) or density function (if $\boldsymbol{X}$ is continuous).

Note that we notationally distinguish between discrete and continuous entropy H and h, but not mutual information I. This is typical, and reflects the sense in which continuous mutual information is the limit of discrete mutual information, and the two are the same object within a more general definition (Sec. 3.5.3), whereas discrete and continuous entropies meaningfully diverge.[8]

[8] Kolmogorov's definition of MI (Sec. 3.5.3) covers both the discrete and continuous case, the relationship between them, and distributions that are not quite either, e.g. contain point-masses and smooth portions alike. In a simpler setting, for successively finer discretizations of continuous random variables, the discrete MI converges to the continuous. By contrast, discrete entropy H diverges for successively finer discretizations of a continuous variable. Moreover, insofar as continuous entropy h is extended to deal with point masses, we have that $h(X) = -\infty$ for any discrete variable X. Thus, an infinite gulf separates H and h on the same objects, whereas MI is essentially the same object. See [33, ch. 7] for a more extended discussion of these limits.

Definition 6 (Shannon Entropy for Random Sources)**.** *Let* $\boldsymbol{X}$ *be a random source, either discrete or continuous. Then the entropy of* $\boldsymbol{X}$ *is defined as:*

$$\begin{aligned} H(\boldsymbol{X}) = \mathbb{E}s_{\boldsymbol{X}}(\boldsymbol{x}) &= -\sum p(\boldsymbol{X}_i)\log p(\boldsymbol{X}_i) && \textit{(discrete source)} \\ h(\boldsymbol{x}) = \mathbb{E}s_{\boldsymbol{X}}(\boldsymbol{x}) &= -\int p_{\boldsymbol{X}}(\boldsymbol{x})\log p_{\boldsymbol{X}}(\boldsymbol{x})d\boldsymbol{x} && \textit{(continuous source)} \end{aligned} \tag{3.5.13}$$

If $\boldsymbol{Y}$ *is another source, such that either both* $\boldsymbol{X}$ *and* $\boldsymbol{Y}$ *are discrete or* $\boldsymbol{X} \cup \boldsymbol{Y}$ *is continuous, then the conditional entropy of* $\boldsymbol{X}$ *given* $\boldsymbol{Y}$ *is*

$$\begin{aligned} H(\boldsymbol{X}|\boldsymbol{Y}) = \mathbb{E}s_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{x}|\boldsymbol{y}) &= -\sum_{i,j} p(\boldsymbol{x}_i, \boldsymbol{y}_j)\log\frac{p(\boldsymbol{x}_i, \boldsymbol{y}_j)}{p(\boldsymbol{y}_j)} && \textit{(discrete source)} \\ h(\boldsymbol{x}|\boldsymbol{y}) = \mathbb{E}s_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{x}|\boldsymbol{y}) &= -\int p_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y})\log\frac{p_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y})}{p_{\boldsymbol{Y}}(\boldsymbol{y})}d(\boldsymbol{x},\boldsymbol{y}) && \textit{(continuous source)} \end{aligned} \tag{3.5.14}$$

Recall that the surprisal $s_{\boldsymbol{X}}(\boldsymbol{x})$ (also sometimes called the information content) quantifies the point-wise unlikelihood of an outcome. Entropy is the expectation of this quantity: a distribution with high entropy is likely to have an unlikely outcome. In practice, a distribution that maximizes entropy is one that spreads out its probability mass as evenly as possible among outcomes.

The following example comes from [33, Theorem 8.4.1].

Example 4 (Entropy of a Gaussian Vector)**.** *Let* $\boldsymbol{X} \sim N(\boldsymbol{0}, \Sigma)$ *be a* k*-dimensional Gaussian random vector, with a positive definite covariance matrix* Σ*. Then*

$$h(\boldsymbol{X}) = \frac{k}{2}\log(2\pi e) + \frac{1}{2}\log|\Sigma| \tag{3.5.15}$$

where $|\Sigma|$ *designates the determinant of the covariance matrix. In particular,*

when $X \sim N(\mu, \sigma)$*, we have*

$$h(X) = \log(\sqrt{2\pi e}) + \log(\sigma) \tag{3.5.16}$$

Thus, we see that $h(X)$ for a Gaussian variable can be any real number, positive or negative, depending upon its variance σ^2. Notice that $h(X) \to -\infty$ as $\sigma \to 0$ and $h(X) \to \infty$ as $\sigma \to \infty$. This makes the continuous version of Shannon entropy somewhat less intuitive than the discrete version. For a discrete variable with a single, certain outcome, $H(X) = 0$. For a 'continuous' variable approaching an analogous situation, $h(X) \to -\infty$.[9] Thus, the situation $h(X) = 0$ itself has no recognized significance, nor does the sign of differential entropy.

If the entropy of a random source $\boldsymbol{X}$ quantifies its uncertainty, then the mutual information between two random sources $\boldsymbol{X}$ and $\boldsymbol{Y}$ quantifies the shared uncertainty between them. Here, we provide the definition for discrete and jointly continuous sources. We will provide the general form later, in Def. 10 of Sec. 3.5.3.

Definition 7 (Mutual Information for Random Sources)**.** *Let* $\boldsymbol{X}$ *and* $\boldsymbol{Y}$ *be random sources such that either both are discrete or* $\boldsymbol{X} \cup \boldsymbol{Y}$ *is continuous. Then the mutual information of* $\boldsymbol{X}$ *and* $\boldsymbol{Y}$ *is given by*

$$\begin{aligned} I(\boldsymbol{X};\boldsymbol{Y}) &= \mathbb{E}\log\frac{s_{\boldsymbol{X}}(\boldsymbol{X})s_{\boldsymbol{Y}}(\boldsymbol{Y})}{s_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{X},\boldsymbol{Y})} = \mathbb{E}\log\frac{p(\boldsymbol{X},\boldsymbol{Y})}{p(\boldsymbol{X})p(\boldsymbol{Y})} \\ &= \begin{cases} \displaystyle\sum_{i,j} p(\boldsymbol{x}_i,\boldsymbol{y}_j)\log\frac{p(\boldsymbol{x}_i,\boldsymbol{y}_j)}{p(\boldsymbol{x}_i)p(\boldsymbol{y}_j)} & \textit{(discrete source)} \\ \displaystyle\int p_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y})\log\frac{p_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y})}{p_{\boldsymbol{X}}(\boldsymbol{x})p_{\boldsymbol{Y}}(\boldsymbol{y})}d(\boldsymbol{x},\boldsymbol{y}) & \textit{(continuous source)} \end{cases} \end{aligned} \tag{3.5.17}$$

[9] Moreover, if we represent a discrete distribution as the sum of point masses, and attempt to apply the definition of differential entropy (using the theory of distributions or a similar method, and extending differential entropy continuously to degenerate distributions), we will likewise have $h(X) = -\infty$.

If $\boldsymbol{Z}$ is another source and either $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}$ are all discrete or $\boldsymbol{X} \cup \boldsymbol{Y} \cup \boldsymbol{Z}$ is continuous, then the conditional mutual information of $\boldsymbol{X}$ and $\boldsymbol{Y}$ given $\boldsymbol{Z}$ is

$$I(\boldsymbol{X}; \boldsymbol{Y}|\boldsymbol{Z}) = \mathbb{E}_{(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z})} \log \frac{p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{z})}{p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{y}|\boldsymbol{z})} \tag{3.5.18}$$

$$= \begin{cases} \sum_{i,j,k} p(\boldsymbol{x}_i, \boldsymbol{y}_j, \boldsymbol{z}_k) \log \frac{p(\boldsymbol{x}_i, \boldsymbol{y}_j|\boldsymbol{z}_k)}{p(\boldsymbol{x}_i|\boldsymbol{z}_k)p(\boldsymbol{y}_j|\boldsymbol{z}_k)} & \textit{(discrete source)} \\ \int p_{\boldsymbol{X},\boldsymbol{Y}\boldsymbol{Z}}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) \log \frac{p_{\boldsymbol{X},\boldsymbol{Y}|\boldsymbol{Z}}(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{z})}{p_{\boldsymbol{X}|\boldsymbol{Z}}(\boldsymbol{x}|\boldsymbol{z})p_{\boldsymbol{Y}|\boldsymbol{Z}}(\boldsymbol{y}|\boldsymbol{z})} d(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) & \textit{(cont. source)} \end{cases}$$

Shannon originally referred to mutual information as the "rate of transmission" of a communications channel, in which $\boldsymbol{X}$ is a transmitted signal and $\boldsymbol{Y}$ is the received signal. Another intuition for the meaning of mutual information comes from the following property:

$$I(\boldsymbol{X}; \boldsymbol{Y}) = H(\boldsymbol{X}) - H(\boldsymbol{X}|\boldsymbol{Y}) \tag{3.5.19}$$

In this sense, mutual information is the reduction in uncertainty of $\boldsymbol{X}$ gained from learning $\boldsymbol{Y}$. This is the sense in which mutual information is shared uncertainty.

We make use of some common conventions for notation that allow us to treat entropy and information as functions of variables and lists of variables

as well as sources. The following conventions hold:

$$
\begin{aligned}
H(X) &:= H(\{X\}) \\
I(X;Y) &:= I(\{X\};\{Y\}) \\
H(X_1,...,X_k) &:= H(\{X_1,...,X_k\}) \\
I(X_1,...,X_k;Y_1,...,Y_\ell) &:= I(\{X_i\};\{Y_j\}) \\
H(\boldsymbol{X}_1,...,\boldsymbol{X}_k) &:= H(\cup_i \boldsymbol{X}_i) \\
I(\boldsymbol{X}_1,...,\boldsymbol{X}_k;\boldsymbol{Y}_1,...,\boldsymbol{Y}_\ell) &:= I(\cup_i \boldsymbol{X}_i;\cup_j \boldsymbol{Y}_j)
\end{aligned}
$$

We review common properties of discrete and continuous information.

Proposition 2 (Properties of Entropy and Mutual Information)**.** *Let $\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}$ be random sources. Then the following properties hold, where the use of H (respectively, h) implies the combined source is discrete (respectively, continuous):*

1. ***Non-Negativity***

   - $H(\boldsymbol{X}) \geq 0$*, where* $H(\boldsymbol{X}) = 0$ *if and only if* X *is degenerate, i.e.* $P(\boldsymbol{X} = \boldsymbol{x}) = 1$ *for some* $\boldsymbol{x}$*.*
   - $H(\boldsymbol{X}|\boldsymbol{Y}) \geq 0$*, where* $H(\boldsymbol{X}|\boldsymbol{Y}) = 0$ *if and only if* X *is conditionally degenerate with respect to* $\boldsymbol{Y}$*, i.e. for every* $\boldsymbol{y}$*,* $P(X = \boldsymbol{x}|\boldsymbol{Y} = \boldsymbol{y}) = 1$ *for some* $\boldsymbol{x}$*.*
   - $I(\boldsymbol{X},\boldsymbol{Y}) \geq 0$*, where* $I(\boldsymbol{X},\boldsymbol{Y}) = 0$ *if and only if* $\boldsymbol{X}$ *and* $\boldsymbol{Y}$ *are independent.*
   - $I(\boldsymbol{X},\boldsymbol{Y} \mid \boldsymbol{Z}) \geq 0$*, where* $I(\boldsymbol{X},\boldsymbol{Y} \mid \boldsymbol{Z}) = 0$ *if and only if* $\boldsymbol{X}$ *and* $\boldsymbol{Y}$ *are conditionally independent with respect to* $\boldsymbol{Z}$*.*

2. ***Chain Rules (Entropy)***

   - $H(\boldsymbol{X},\boldsymbol{Y}) = H(\boldsymbol{X}) + H(\boldsymbol{Y}|\boldsymbol{X})$

- $H(\boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{Z}) = H(\boldsymbol{X}|\boldsymbol{Z}) + H(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z})$
- $h(\boldsymbol{X}, \boldsymbol{Y}) = h(\boldsymbol{X}) + h(\boldsymbol{Y}|\boldsymbol{X})$
- $h(\boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{Z}) = h(\boldsymbol{X}|\boldsymbol{Z}) + h(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z})$

For an arbitrary number of sources:

- $H(\boldsymbol{X}_1, ..., \boldsymbol{X}_k) = \sum_i H(\boldsymbol{X}_i; \boldsymbol{X}_{i-1}, ..., \boldsymbol{X}_1)$
- $h(\boldsymbol{X}_1, ..., \boldsymbol{X}_k) = \sum_i h(\boldsymbol{X}_i; \boldsymbol{X}_{i-1}, ..., \boldsymbol{X}_1)$

3. ***Chain Rule (MI)***

- $$I(\boldsymbol{X}; \boldsymbol{Y}, \boldsymbol{Z}) = I(\boldsymbol{X}; \boldsymbol{Y}) + I(\boldsymbol{X}; \boldsymbol{Z}|\boldsymbol{Y}) \tag{3.5.20}$$

For an arbitrary number of sources:

- $$I(\boldsymbol{Y}; \boldsymbol{X}_1, ..., \boldsymbol{X}_k) = \sum_i I(\boldsymbol{Y}; \boldsymbol{X}_i|\boldsymbol{X}_{i-1}, ..., \boldsymbol{X}_1) \tag{3.5.21}$$

4. ***Identities for Entropy and Information***

- $$I(\boldsymbol{X}; \boldsymbol{Y}) = H(\boldsymbol{X}) - H(\boldsymbol{X}|\boldsymbol{Y}) \tag{3.5.19}$$
- $$I(\boldsymbol{X}; \boldsymbol{Y}) + H(\boldsymbol{X}, \boldsymbol{Y}) = H(\boldsymbol{X}) + H(\boldsymbol{Y}) \tag{3.5.22}$$

We may now extend Example 4 to MI and use it to demonstrate the common properties of entropy and MI.

Example 5 (Information of a Bivariate Gaussian Variable)**.** *Let X and Y be*

jointly Gaussian, i.e. $(X, Y) \sim N(\boldsymbol{\mu}, \Sigma)$ *and*

$$\sigma = \begin{bmatrix} \sigma_X^2 & \sigma_X \sigma_Y \rho \\ \sigma_X \sigma_Y \rho & \sigma_Y^2 \end{bmatrix}.$$

We then have that

$$h(X, Y) = \log(2\pi e) + \log(\sigma_X) + \log(\sigma_Y) - \log(\frac{1}{\sqrt{1-\rho^2}}) \tag{3.5.23}$$

$$I(X; Y) = \log(\frac{1}{\sqrt{1-\rho^2}}) \tag{3.5.24}$$

Suppose $\sigma = \sigma_X = \sigma_Y$ *as in Example 1. Then*

$$h(X) = h(Y) = \frac{1}{2}\log(2\pi e) + \log \sigma, \tag{3.5.25}$$

$$h(X|Y) = \frac{1}{2}\log(2\pi e) + \log \sigma + \frac{1}{2}\log(1-\rho^2), \tag{3.5.26}$$

$$h(X, Y) = \log(2\pi e) + 2\log \sigma + \frac{1}{2}\log(1-\rho^2), \tag{3.5.27}$$

$$I(X; Y) = -\frac{1}{2}\log(1-\rho^2). \tag{3.5.28}$$

Note that the Chain Rule (3.5.20) and Eq (3.5.19) hold. Further, as the variables become more strongly correlated/anti-correlated, i.e. as $\rho \to \pm 1$*, we see that entropy becomes arbitrarily small and mutual information arbitrarily large.*

Note that, for Eq (3.5.23), the joint entropy is maximized when $\rho = 0$, i.e. X and Y are independent. On the other hand, it approaches $-\infty$ as $\rho \to \pm 1$. Similarly, mutual information is zero when $\rho = 0$, and $I(X; Y) \to \infty$ as $\rho \to \pm 1$. This example demonstrates some of the issues that can arise when dealing with degeneracies in random variables with non-finite ("continuous") alphabets.

An early application of Shannon information theory to statistics was made by Solomon Kullback and Richard Liebler in 1951. In a concise paper [68], they used Shannon ('Shannon-Wiener') information formalism to measure the 'information for discrimination' between hypotheses, which they give as the difference in hypothesis-specific surprisals for observed data. This work firmly connected (while distinguishing) Shannon-Wiener information with Fisher information, and used Shannon-Wiener formalism to characterize sufficient statistics. In doing so, they defined a divergence function that serves, among other purposes, as a pseudo-distance between probability distributions.

Definition 8 (Kullback-Liebler Divergence)**.** *Suppose $\boldsymbol{X}$ and $\boldsymbol{Y}$ are random sources, either both discrete or individually continuous, on the same alphabet $\mathcal{A} = \mathcal{A}_{\boldsymbol{X}} = \mathcal{A}_{\boldsymbol{Y}}$. Then the Kullback-Liebler (KL) divergence between their distributions is given by*

$$\begin{aligned} D(p_{\boldsymbol{X}}|p_{\boldsymbol{Y}}) &= \mathbb{E}\left[s_{\boldsymbol{Y}}(\boldsymbol{X}) - s_{\boldsymbol{X}}(\boldsymbol{X})\right] \\ &= \begin{cases} \displaystyle\sum_{\boldsymbol{x}_i \in \mathcal{A}} p_{\boldsymbol{X}}(\boldsymbol{x}_i) \log \frac{p_{\boldsymbol{X}}(\boldsymbol{x}_i)}{p_{\boldsymbol{Y}}(\boldsymbol{x}_i)} & \textit{(discrete source)} \\ \displaystyle\int p_{\boldsymbol{X}}(\boldsymbol{x}) \log \frac{p_{\boldsymbol{X}}(\boldsymbol{x})}{p_{\boldsymbol{Y}}(\boldsymbol{x})} d\boldsymbol{x} & \textit{(continuous source)} \end{cases} \end{aligned} \tag{3.5.29}$$

We use D to represent KL divergence between distributions, and use $\mathcal{D}$ when we taking the random variables as arguments:

$$\mathcal{D}(\boldsymbol{X}|\boldsymbol{Y}) = D(p_{\boldsymbol{X}}|p_{\boldsymbol{Y}}) \tag{3.5.30}$$

KL divergence is often described as measuring the 'distance' or dissimilarity between probability distributions on a shared alphabet. It is not a distance, as it is not symmetric and does not obey the triangle inequality. Moreover, KL

divergence can be infinite, if $\mu_{\boldsymbol{X}}$ and $\mu_{\boldsymbol{Y}}$ are not both absolutely continuous with respect to each other.

Proposition 3 (Properties of KL Divergence)**.** *KL Divergence has the following properties:*

- $D(p_{\boldsymbol{X}}|p_{\boldsymbol{Y}}) \geq 0$ *with equality if and only if* $p_{\boldsymbol{X}}(\boldsymbol{X}) \overset{\omega}{=} p_{\boldsymbol{Y}}(\boldsymbol{X})$.
- *Suppose* $\boldsymbol{X}$ *and* $\boldsymbol{Y}$ *are discrete or jointly continuous random sources. Then*

$$I(\boldsymbol{X};\boldsymbol{Y}) = D(p_{\boldsymbol{X},\boldsymbol{Y}}|p_{\boldsymbol{X}}p_{\boldsymbol{Y}}). \tag{3.5.31}$$

  In other words, mutual information is the divergence between the joint distribution $p_{\boldsymbol{X},\boldsymbol{Y}}$ *and the product distribution* $p_{\boldsymbol{X}}p_{\boldsymbol{Y}}$ *on the product alphabet* $\mathcal{A}_{\boldsymbol{X}} \times \mathcal{A}_{\boldsymbol{Y}}$.

Proposition 4. *Let* p *and* q *be two Gaussian densities,* $p \sim N(\mu_1, \sigma_1^2)$ *and* $q \sim N(\mu_2, \sigma_2^2)$. *Then the Kullback-Liebler (KL) Divergence between them is given by*

$$D(p||q) = \log\left(\frac{\sigma_2}{\sigma_1}\right) + \frac{1}{2}\left(\frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{\sigma_2^2}\right) - \frac{1}{2}.$$

To define the Williams-Beer PID later, it is useful to introduce the concept of **specific information**. For two sources $\boldsymbol{X}$ and $\boldsymbol{Y}$, specific information is an intermediate quantity between mutual information $I(\boldsymbol{X};\boldsymbol{Y})$, which is defined in expectation, and the pointwise mutual information function $\log \frac{p(\boldsymbol{x},\boldsymbol{y})}{p(\boldsymbol{x})p(\boldsymbol{y})}$. Specific information is a pointwise function of one of the two sources, taken in expectation of the other source.

Definition 9 (Specific Information)**.** *Let* $\boldsymbol{X}$ *and* $\boldsymbol{Y}$ *be discrete or jointly continuous. Then for any* $\boldsymbol{y} \in \mathcal{A}_{\boldsymbol{Y}}$ *such that* $p_{\boldsymbol{Y}}(\boldsymbol{y}) > 0$, *we define the specific*

information of $\boldsymbol{X}$ *about* $\boldsymbol{Y} = \boldsymbol{y}$ *to be*

$$I_{\boldsymbol{X}}(\boldsymbol{y}) = D(p_{\boldsymbol{X}|\boldsymbol{Y}=\boldsymbol{y}} || p_{\boldsymbol{X}}). \tag{3.5.32}$$

Moreover, we let $I_{\boldsymbol{X}}(\boldsymbol{y}) = 0$ *whenever* $p_{\boldsymbol{Y}}(y) = 0$.

This concludes most of our preliminary review of classical information theory. We have covered everything in the contexts in which we are working with discrete distributions or densities. In the section that follows, we will introduce enough of the general case to handle the degenerate distributions of noise-free interactions in Sec. 3.6.

3.5.3 General Information Theory

We have covered the usual definitions of information quantities, particularly in how they are typically applied by computational scientists and mathematicians alike. Recall, for instance, from Example 5, that mutual information between correlated Gaussians is given by:

$$I(X, Y) = \log \frac{1}{\sqrt{1 - \rho^2}}. \tag{3.5.24}$$

Clearly, $I(X, Y) \to \infty$ as $\rho \to \pm 1$. This may seem more a matter of mathematical interest than practical concern. For the continuous, noise-free interaction model that we introduce in Sec. 3.6, however, we will have such an idealized situation. Our predictor genes X and Y will perfectly determine a response T, and we will be taking the PID of $I(T; X, Y) = \infty$. Thus, it is worthwhile for us to briefly cover the more general case of mutual information. Our discussion draws from the references [33, ch. 7] and [51, ch. 7].

We need a few more definitions. If $\boldsymbol{X}$ is a random source with alphabet

$\mathcal{A}_{\boldsymbol{X}}$, let $\mathscr{P}$ be a (countable) partition of $\mathcal{A}_{\boldsymbol{X}}$, i.e.

$$\mathcal{A} = \bigsqcup_{P_i \in \mathscr{P}} P_i.$$

Then let $[\boldsymbol{X}]_{\mathscr{P}}$ be the random variable with alphabet $\mathscr{P}$ such that

$$[\boldsymbol{X}]_{\mathscr{P}}(\omega) = P_i$$
$$\Updownarrow$$
$$\boldsymbol{X}(\omega) \in P_i.$$

In other words, $[\boldsymbol{X}]_{\mathscr{P}}$ is the variable that takes value P_i whenever $\boldsymbol{X} = \boldsymbol{x}$ for some $\boldsymbol{x} \in P_i$. This variable $[\boldsymbol{X}]_{\mathscr{P}}$ is called the discretization of $\boldsymbol{X}$ with respect to the partition $\mathcal{P}$. This definition does not just cover continuous variables. If X is a discrete variable, we can represent it as a sum of point-masses, i.e. $p_X(x) = \sum_i \delta_{x_i}(x)$. For a sufficiently fine discretization $[\boldsymbol{X}]_{\mathscr{P}}$ such that no P_i contains within it more than one point x_i, we will have that $\boldsymbol{X}$ and $[\boldsymbol{X}]_{\mathscr{P}}$ are almost surely equal as random variables. Assuming we index such that $x_i \in P_i$, we would have that $P(X = x_i) = P([\boldsymbol{X}]_{\mathscr{P}} = P_i)$.

Definition 10. *Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be two random sources, with alphabets $\mathcal{A}_{\boldsymbol{X}}$ and $\mathcal{A}_{\boldsymbol{Y}}$. The mutual information between $\boldsymbol{X}$ and $\boldsymbol{Y}$ is given by*

$$I(\boldsymbol{X}; \boldsymbol{Y}) = \sup_{\mathscr{P}, \mathscr{Q}} I([\boldsymbol{X}]_{\mathscr{P}}; [\boldsymbol{Y}]_{\mathscr{Q}}) \tag{3.5.33}$$

where the supremum is taken over all discretizations of $\boldsymbol{X}$ and $\boldsymbol{Y}$.

The supremum term is monotonically increasing for successively finer partitions $\mathscr{P}$ and $\mathscr{Q}$. Thus, this definition can be approached with standard methods from real analysis and measure theory. For our purposes, we use this

definition only to demonstrate the following proposition.

Proposition 5. *Let $\boldsymbol{X}, \boldsymbol{Y}$ be two random sources, each individually continuous, but such that $\boldsymbol{X} \cup \boldsymbol{Y}$ is not continuous. Then $I(\boldsymbol{X}; \boldsymbol{Y}) = \infty$.*

The proof is conceptually identical to that of Lemma 7.4 in [51].

Proof. Let $n = |\boldsymbol{X}|$ (the dimension of $\boldsymbol{X}$), and $m = |\boldsymbol{Y}|$. Since $\boldsymbol{X}, \boldsymbol{Y}$ are each continuous, their induced measures $\mu_{\boldsymbol{X}}$ and $\mu_{\boldsymbol{Y}}$ are absolutely continuous with respect to Lebesgue measure λ^n and λ^m. Thus, $\mu_{\boldsymbol{X}} \times \mu_{\boldsymbol{Y}} \ll \lambda^{n+m}$. If $\mu_{\boldsymbol{X},\boldsymbol{Y}} \ll \mu_{\boldsymbol{X}} \times \mu_{\boldsymbol{Y}}$, then we would have $\mu_{\boldsymbol{X},\boldsymbol{Y}} \ll \lambda^{n+m}$, which would imply $\boldsymbol{X} \cup \boldsymbol{Y}$ continuous. Thus, $\mu_{\boldsymbol{X},\boldsymbol{Y}} \not\ll \mu_{\boldsymbol{X}} \times \mu_{\boldsymbol{Y}}$.

This implies there is a set $P \subset \mathbb{R}^{n+m}$ such that

$$\mu_{\boldsymbol{X},\boldsymbol{Y}}(P) > 0 \text{ and } (\mu_{\boldsymbol{X}} \times \mu_{\boldsymbol{Y}}(P)) = 0.$$

For the partition $\mathscr{P} = \{P, P^C\}$, we have that $I(\boldsymbol{X}_{\mathscr{P}}; \boldsymbol{Y}_{\mathscr{P}}) = \infty$, and our result follows from Def. 10. □

Corollary 1. *Let $\boldsymbol{X}, \boldsymbol{Y}$ be two random sources with joint distribution*

$$p_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{M} \delta(g_i(\boldsymbol{x}, \boldsymbol{y})) p_{\boldsymbol{X}}(\boldsymbol{x}) \tag{3.5.34}$$

where $\{g_i\}$ is a collection of real-valued functions with isolated zeros. Then

$$I(\boldsymbol{X}, \boldsymbol{Y}) = \infty. \tag{3.5.35}$$

In such a situation, we may say that $\boldsymbol{X}$ provides **perfect** (or **infinite**) **information** information about $\boldsymbol{Y}$. We will use the two words, perfect and infinite, interchangeably, depending on whether we want to emphasize either

the collapse of uncertainty or the quantitative irregularity (i.e. infinite-valued information functions).

This concludes our review of information theory. We may now introduce our extension of the $I_\cap^{\min}$ and $I_\cap^{\mathrm{PM}}$ PIDs to continuous interactions of random variables.

3.6 PID for Bivariate Interactions

3.6.1 Bivariate Interaction Problem

Part of the aim of this work is to extend the partial information framework of Williams and Beer [125], and the follow-up work by Finn and Lizier in [45], to interactions of continuous variables. In this section, we will introduce the PID framework within the continuous setting in which we are interested, using the discrete case as a starting point from which we will develop our continuous extension. We direct those interested in a detailed background of the discrete theory to [12, 45, 125] and to [77] for a recent overview of the current state of PID research. For simplicity, we only present the bivariate setting in which two predictors inform upon a single target variable.

Consider three random variables: two predictors $\{X, Y\}$ and a target T. We consider two types of situations:

- (Noise-free) We model T as a deterministic function of X and Y via

$$T = g(X, Y) \tag{3.6.1}$$

  where g is a non-random function. The conditional entropy of T given the predictors is $H(T|X, Y) = 0$ in the discrete case and $h(T|X, Y) = -\infty$ in the continuous setting.

- (Noisy) We model T as a noisy realization of $g(X, Y)$, where the noise model is given by

$$T = g(X, Y) + Z \tag{3.6.2}$$

  where Z is a zero-mean variable independent of (X, Y). In this case, the conditional entropy (depending on the distribution of Z) may be

non-degenerate.

Consider the noisy model. Assuming the regularity of these variables, one can think of them as having the joint density:

$$p_{X,Y,T}(x, y, t) = p_{X,Y}(x, y)p_Z(t - g(x, y)) \tag{3.6.3}$$

If the noise Z tends to be relatively small in both magnitude and variance compared to $g(X, Y)$, then we can intuitively understand that X and Y provide most of the information available regarding T. For a given realization, observing X and Y substantially constrains the uncertainty in T.

We can understand this formally. The mutual information between our target T and the collection $\{X, Y\}$ is the nonnegative quantity

$$\begin{aligned} I(T; X, Y) &= h(T) - h(T \mid X, Y) \\ &= h(g(X, Y) + Z) - h(Z). \end{aligned}$$

This quantity represents the gap between the full uncertainty of T and the reduced uncertainty of T given knowledge of X and Y. If $h(g(X, Y)) \gg h(Z)$, then $h(T) \approx h(g(X, Y))$. Moreover, as the noise becomes arbitrarily small, the information becomes arbitrarily large:

$$I(T; X, Y) \to \infty \text{ as } \operatorname{Var}(Z) \to 0 \text{ since } h(Z) \to -\infty \tag{3.6.4}$$

In the noise-free case, the three variables will instead have a degenerate joint distribution:

$$p_{X,Y,T}(x, y, t) = p_{X,Y}(x, y)\delta(t - g(x, y)) \tag{3.6.5}$$

Applying Corollary 1 of Proposition 5, we have that $I(T; X, Y) = \infty$. This can

be seen to follow from (3.6.4) if one understands (3.6.5) as the limit distribution of (3.6.3).

Before moving on, we present formal definitions of noisy and noise-free bivariate interactions.

Definition 11 (Bivariate Interaction (Noise-free))**.** *Let X, Y, T be real-valued random variables and g be a continuous function such that Eq. (3.6.1) holds for every realization, i.e.*

$$T(\omega) = g(X, Y)(\omega) \text{ for every } \omega \in \Omega. \tag{3.6.6}$$

*We call the triplet (X, Y, T) a **noise-free bivariate interaction**, and denote it $X, Y \to T$. If $(X, Y) \sim N(\boldsymbol{\mu}, \Sigma)$, we refer to the interaction as **Gaussian**, denoted $X, Y \to_\Sigma T$. If $\boldsymbol{\mu} = \mathbf{0}$ and $\Sigma = (1, \rho; \rho, 1)$, we denote it $X, Y \to_\rho T$.*

Definition 12 (Bivariate Interaction (Noisy))**.** *Let X, Y, Z, T be real-valued random variables and g be a continuous function such that Eq. (3.6.2) holds. Further, assume Z is independent of X and Y. We call the tuple (X, Y, Z, T) a **noisy bivariate interaction**, and denote it $X, Y \overset{Z}{\to} T$. If X, Y are jointly Gaussian, we refer to the interaction as Gaussian. Analogous to the notation in Definition 11, we use the notation $X, Y \overset{Z}{\to}_\Sigma T$ and $X, Y \overset{Z}{\to}_\rho T$.*

Now that we have an understanding of continuous interactions as jointly random variables, we may extend the PID framework to this class of objects.

3.6.2 Partial Information Decomposition

3.6.2.1 The PID Framework

For the moment, consider a nondegenerate (i.e., non-deterministic Z) noisy interaction $X, Y \overset{Z}{\to}_\Sigma T$, so that $I(T; X, Y) < \infty$. If $I(T; X, Y)$ is the total

information about T found in X and Y together, we might ask the following questions (see [77] for related discussions):

1. How much information about T is contained in X and Y individually?
2. How much of that information about T can be learned from either X or Y?
3. How much information about T can be learned from only X and not Y? From Y and not X?
4. How much information about T can be learned only from both X and Y, yet neither individually?

The answer to the first question is found in the MIs $I(T;X)$ and $I(T;Y)$. The PID framework was developed to answer questions 2-4 for discrete variables. Respectively, these information atoms are referred to as redundant information $R(T;X,Y)$, unique informations $U_X(T;X,Y), U_Y(T;Y,X)$, and synergistic information $S(T;X,Y)$. The total and individual MIs $I(T;X,Y), I(T;X), I(T;Y)$ decompose into these atoms according to Eqs E1-E3, reproduced here:

$$I(T;X) = R(T;X,Y) + U_X(T;X,Y) \tag{E1}$$

$$I(T;Y) = R(T;X,Y) + U_Y(T;X,Y) \tag{E2}$$

$$I(T;X,Y) = R(T;X,Y) + U_X(T;X,Y) + U_Y(T;X,Y) + S(T;X,Y) \tag{E3}$$

Per standard practice [125], we can represent this decomposition visually as a Venn diagram (Fig. 3.12). When convenient, we will abbreviate these four atoms as R, S, U_X, U_Y, denoted collectively as the tuple $\mathfrak{d}$. All these quantities become well-defined if a redundancy function $I_\cap^{(\alpha)}$ is provided (defined later in

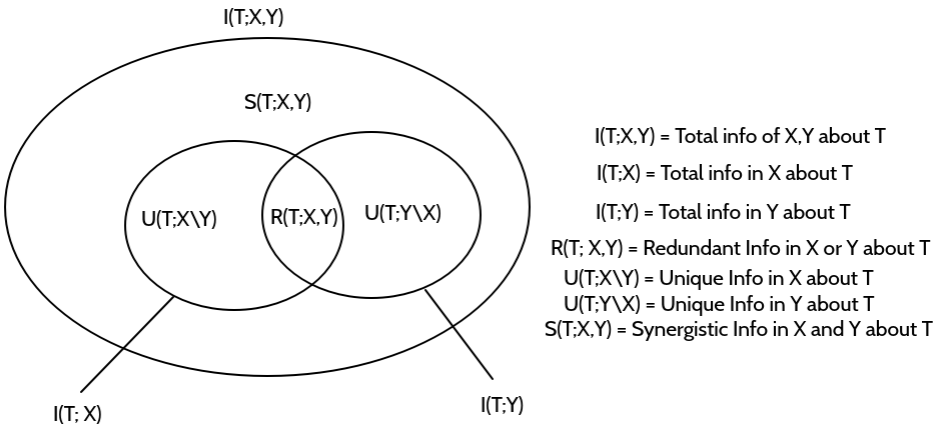


Figure 3.12: **Bivariate PID diagram** Venn diagram representation of the PID decomposition of total MI $I(T; X, Y)$.

Def. 13), for which α is an arbitrary label, e.g. 'min' for $I_\cap^{\min}$. We then denote the labeled PID:

$$\mathfrak{d}_\alpha = (R^{(\alpha)}, S^{(\alpha)}, U_X^{(\alpha)}, U_Y^{(\alpha)}),$$

where

$$\begin{aligned}
R^{(\alpha)} &= I_\cap^{(\alpha)}(T; X, Y), \\
U_X^{(\alpha)} &= I(T; X) - I_\cap^{(\alpha)}(T; X, Y), \\
U_Y^{(\alpha)} &= I(T; Y) - I_\cap^{(\alpha)}(T; X, Y), \\
S^{(\alpha)} &= I(T; X, Y) - I(T; X) - I(T; Y) + I_\cap^{(\alpha)}(T; X, Y).
\end{aligned}$$

We focus upon two choices for the redundancy function: $I_\cap^{\min}$ and $I_\cap^{\mathrm{PM}}$.

Before we introduce these functions, we present a common example for the PID of discrete variables (X, Y, T), to provide a helpful intuition for the atoms in a PID.

Example 6 (Bivariate XOR)**.** *Consider the following distribution* $T = X \oplus Y$*:*

X	Y	T	p
0	0	0	1/4
0	1	1	1/4
1	0	1	1/4
1	1	0	1/4

Then any partial information decomposition $\mathfrak{d}$ *of* $I(T; X, Y)$ *satisfying Eqs E1-E3 will be of the following form:*

$$R = -U_X = -U_Y \tag{3.6.7}$$

$$S - R = 1 \tag{3.6.8}$$

This follows readily from the observation that $I(T; X, Y) = 1$ *bit and* $I(T; X) = I(T; Y) = 0$. *If we further require that every information atom is nonnegative (as is the case for* $I_\cap^{min}$*), then* $R = 0$ *and* $S = 1$, *i.e.* X *and* Y *have one bit of synergistic information about* T *and zero redundant and unique information.*

As stated above, a PID $\mathfrak{d}_\alpha$ is defined by its redundancy function $I_\cap^{(\alpha)}$. Before we move on, we need introduce a general definition of a redundancy function and the associated axioms. We will first present a definition of redundancy functions that accords with what have become known as the Williams-Beer (WB) Axioms. These were originally stated as properties of $I_\cap^{\min}$ in [125], but were later taken as axiomatic by the community [12]. We follow the enumeration from [11], which includes nonnegativity (there, referred to as global positivity).

0. **Nonnegativity.** Redundant information is a nonnegative quantity.

$$I_\cap(T; \boldsymbol{X}_1, ..., \boldsymbol{X}_n) \geq 0 \tag{P}$$

1. **Symmetry.** Redundant information is invariant under permutations of sources:

$$I_\cap(T; \boldsymbol{X}_1, ..., \boldsymbol{X}_n) = I_\cap(T; \boldsymbol{X}_{\sigma(1)}, ..., \boldsymbol{X}_{\sigma(n)}) \tag{S}$$

2. **Monotonicity.** Every additional source included can only decrease the amount of redundant information.

$$I_\cap(T; \boldsymbol{X}_1, ..., \boldsymbol{X}_n) \leq I_\cap(T; \boldsymbol{X}_1, ..., \boldsymbol{X}_{n-1}) \tag{M}$$

3. **Self-Redundancy.** The redundant information of a single source is the mutual information between that source and the target.

$$I_\cap(T; \boldsymbol{X}) = I(T; \boldsymbol{X}) \tag{SR}$$

Many, though not all, of the redundancy functions in the literature accept this framework, There are notable exceptions (including $I_\cap^{\mathrm{PM}}$ and $I_\cap^{CCS}$). Thus, we will then present our own definition that encapsulates these. From our perspective, self-redundancy is the indispensable axiom. Insofar as PID is meant to extend mutual information, this axiom must hold. The symmetry axioms, although uncontested in the literature, may be superfluous. The lattice framework in [125] defines the redundancy function $I_\cap^{\min}$ as acting on collections of sources, not ordered tuples. In this framework, symmetry must be implicit in order for any redundancy $I_\cap$ to be well-defined. Thus, we will dispense with it in our definition.

Definition 13 (Redundancy Function)**.** *Let $\mathfrak{R}$ be a finite collection of predictor variables for a target variable T, and assume all variables are discrete or jointly continuous. Using $\mathcal{P}$ to denote the powerset, let $\mathfrak{S} = \mathcal{P}(\mathcal{P}(\mathfrak{R}))$ be the set of all collections of subsources of $\mathfrak{R}$. A* ***redundancy function***

$I_\cap : \mathfrak{S} \to \mathbb{R}$, *must have the following property*

$$I_\cap(\boldsymbol{X}) = I(T; \boldsymbol{X}) \qquad \textit{Self-Redundancy} \qquad \text{(SR)}$$

It may also have the following optional properties:

$$I_\cap(\boldsymbol{X}_1, ..., \boldsymbol{X}_m) \geq 0 \qquad \textit{Nonnegativity} \qquad \text{(P)}$$

$$I_\cap(\boldsymbol{X}_1, ..., \boldsymbol{X}_{m-1}) \geq I_\cap(\boldsymbol{X}_1, ..., \boldsymbol{X}_m) \qquad \textit{Monotonicity (M)} \qquad \text{(M)}$$

Associated with a redundancy function $I_\cap$ is a Möbius inverse, Π, that computes the PID atoms associated with ascending a lattice of sources [125]. In the bivariate case, for instance:

$$R = \Pi(T; \{X\}, \{Y\}) = I_\cap(T; \{X\}\{Y\})$$

$$U_X = \Pi(T; \{X\}) = \underbrace{I_\cap(T; \{X\})}_{I(T;X)} - I_\cap T; \{X\}\{Y\}$$

$$U_Y = \Pi(T; \{Y\}) = \underbrace{I_\cap(T; \{Y\})}_{I(T;Y)} - I_\cap T; \{X\}\{Y\}$$

$$S = \Pi(T; \{X, Y\}) = \underbrace{I_\cap(T; \{X, Y\})}_{I(T;X,Y)} - \sum_\beta \Pi(\beta)$$

$$\text{where we range } \beta = \{X\}, \{Y\}, \{X\}\{Y\}.$$

Since we are concerned with networks and the bivariate PID in this work, we do not treat the full lattice framework here. For the development of PID lattices, see the Appendix of [125]. [10]

It immediately follows that if a redundancy function $I_\cap$ fulfills all three axioms (SR), (P), and (M), then the four atoms of the bivariate PID $\mathfrak{d}$ will be

[10] Information lattices enjoy a broad treatment in the information theory literature. The earliest treatment of an 'information lattice' of which we are aware can be found in a brief preliminary sketch by Shannon [105].

non-negative. More generally, this property is equivalent to the non-negativity of the Π function [125].

We now turn to the two redundancy functions that will be the focus of our study. $I_\cap^{\min}$ will fulfill both axioms, and $I_\cap^{\mathrm{PM}}$ will only fulfill (SR).

3.6.2.2 Redundancy Functions $I_\cap^{\min}$ and $I_\cap^{\mathrm{PM}}$

In this section, we present the redundancy functions that we will be studying. Both of these measures were originally defined for discrete sources. We will present a general definition for discrete and continuous sources.

Williams and Beer defined the redundancy function $I_\cap^{\min}$ in [125]. For a collection of sources $\boldsymbol{X}_1, ...\boldsymbol{X}_m$, they defined the redundant information among them regarding T to be:

$$I_\cap^{\min}(T; \boldsymbol{X}_1, ..., \boldsymbol{X}_m) = \sum_j p(t_j) \min_k I(T = t_j; \boldsymbol{X}_k) \tag{3.6.9}$$

where

$$I(T = t; \boldsymbol{X}_k) := \sum_i p(\boldsymbol{x}_i|t) \left(\log \frac{1}{p(t)} - \log \frac{1}{p(t|\boldsymbol{x}_i)} \right) \tag{3.6.10}$$

Note that in the case where there is only one predictor source, this definition is identical to the mutual information $I(T, \boldsymbol{X})$. Written in this form, one can see that the pointwise redundant information at $T = t$ is the minimal reduction in surprisal $\log \frac{1}{p(t)}$ caused by conditioning on a source. We expand their definition to our context.

Definition 14 (Definition of $I_\cap^{\min}$)**.** *Let T be a target variable and $\boldsymbol{X}_1, ...\boldsymbol{X}_m$ be a collection of sources. Suppose further either that all are discrete, or each*

X_i is jointly continuous with T, i.e. $\boldsymbol{X}_i \cup \{T\}$ is continuous for each i. Then the redundant information provided by the sources $\{\boldsymbol{X}_i\}$ about T is given by:

$$I_{\cap}^{\min}(T; \boldsymbol{X}_1, ..., \boldsymbol{X}_m) = \mathbb{E}_T \min_k I_{\boldsymbol{X}_k}(T). \tag{3.6.11}$$

Here, $I_{\boldsymbol{X}_k}(t)$ denotes the specific information function (Def. 9):

$$\begin{aligned} I_{\boldsymbol{X}}(t) &:= D(p_{\boldsymbol{X}|T=t} || p_{\boldsymbol{X}}) && (3.6.12) \\ &= \mathbb{E}_{\boldsymbol{X}_k|T=t} \log \frac{p_{\boldsymbol{X},T}(\boldsymbol{X}, t)}{p_{\boldsymbol{X}}(\boldsymbol{x}) p_T(t)} && (3.6.13) \\ &= \begin{cases} \sum_i p(\boldsymbol{x}_i|t) \log \dfrac{p(\boldsymbol{x}_i, t)}{p(\boldsymbol{x}_i) p(t)} & \textit{(discrete source)} \\ \int \dfrac{p_{\boldsymbol{X},T}(\boldsymbol{x}, t)}{p_T(t)} \log \dfrac{p_{\boldsymbol{X},T}(\boldsymbol{x}, t)}{p_{\boldsymbol{X}}(\boldsymbol{x}) p_T(t)} d\boldsymbol{x} & \textit{(continuous source)} \end{cases} \end{aligned}$$

Intuitively, we can be think about this definition of redundant information in the following way. At every pointwise instantiation of T, we consider how much uncertainty in T is eliminated, locally near a specific outcome $T = t$, by each $\boldsymbol{X}_i$ individually. Then, when we take the minimum, we take the value corresponding to the minimal reduction in local uncertainty possible when learning one of the sources. As has been frequently discussed in the literature [11, 45, 77] this approach conflates variables containing the same information about a target outcome $T = t$ with the same *amount* of information about that outcome. We discuss this and the 'two-bit copy' problem in Sec. 3.1.2.2, above.

From Def. 14 and Eqs. E1-E3, the definitions of the atoms of the bivariate

$I_\cap^{\min}$ PID $\mathfrak{d}_{\min}$ follow:

$$R^{\min}(T;X,Y) = I_\cap^{\min}(T;X,Y) \tag{3.6.14}$$

$$U_X^{\min}(T;X,Y) = I(T;X) - I_\cap^{\min}(T;X,Y) \tag{3.6.15}$$

$$U_Y^{\min}(T;X,Y) = I(T;Y) - I_\cap^{\min}(T;X,Y) \tag{3.6.16}$$

$$S^{\min}(T;X,Y) = I(T;X,Y) - I(T;X) - I(T;Y) + I_\cap^{\min}(T;X,Y) \tag{3.6.17}$$

We emphasize that the $I_\cap^{\min}$ PID fulfills all of the properties discussed in Definition 13.

Proposition 6. *The redundancy function $I_\cap^{min}$ fulfills the properties (SR), (P), and (M) in Def. 13. Moreover, in the continuous bivariate setting $I_\cap^{min}$, U_X^{min}, and U_Y^{min} are all nonnegative. In the discrete bivariate setting $I_\cap^{min}$, U_X^{min}, U_Y^{min}, and S^{min} are all nonnegative.*

The proofs of these properties in the discrete case can be found in the Appendix of [125]. The only subtlety in our case in that we are working with continuous variables.

Proof. Self-redundancy (SR) follows from Def. 14 and Def. 7. For non-negativity (P), we note that since $I_{\boldsymbol{X}}(t)$ is a KL-divergence, it is nonnegative on its domain. Thus, $\min_k I_{\boldsymbol{X}_k}(t)$ is likewise non-negative, and nonnegativity of $R^{\min}(T;X,Y)$ follows. Monotonicity (M) follows from pointwise monotonicity:

$$\min_{k=1,\ldots,m}(I_{\boldsymbol{X}_k}(t)) \leq \min_{k=1,\ldots,m-1}(I_{\boldsymbol{X}_k}(t)).$$

In the bivariate setting, the nonnegativity of $U_X^{\min}$ (and analogously, that of $U_Y^{\min}$) follows from the monotonicity that we have just demonstrated:

$$R^{\min}(T;X,Y) = I_\cap^{\min}(T;X,Y) \leq I_\cap^{\min}(T;X) = I(T;X).$$

□

As was hinted in Section 3.3.2, the $I_\cap^{\min}$ PID has a particular virtue that makes it well-suited to our network inference problem. Consider the case where a pair of predictor genes (X, Y) provides a significant amount of information about a response T simply because one of them is important to the response and the other is not (i.e., $I(T; X, Y) > I(T; Y) \gg I(T; X|Y)$). We contend that synergy ought to distinguish a pair of predictors that jointly influence a response T from a pair where one predictor is conditionally independent of the response. The conditional independence is important, as dependency structure may cause both $I(T; X)$ and $I(T; Y)$ to stand out, with each gene identified as a relatively strong predictor.

More specifically, consider a gene interaction network (Def. 1), where T is an interaction on a network of predictors $\boldsymbol{X}$. Let X_i be one of the 'true' predictors of T, i.e. $(i, i') \in \mathcal{E}'$ for some other gene i', and so $g(X_i, X_{i'})$ appears in Eq. (3.2.2). Let j be some other gene that is not a true predictor, i.e. $\{(j, k), (k, j)|k \in [n]\} \cap \mathcal{E}' = \emptyset$. Since X_j plays no role in determining T, we would want our inference statistic to distinguish between $(X_i, X_{i'})$ and (X_i, X_j), and to *not* distinguish between (X_i, X_j) and (X_j, X_k) for any other non-predicting gene k. In other words, we would want a statistic that demonstrates *specificity* in expectation. The following proposition demonstrates that this specificity follows for any non-negative PID.

Proposition 7. *Let X, Y, T be jointly discrete or continuous random variables. Suppose further that X is conditionally independent of T given Y, i.e. $X \perp T|Y$. Let $\mathfrak{d}$ be a bivariate PID of $I(T; X, Y)$, induced by a redundancy function*

$I_\cap$ that induces a non-negative Π *function. Then we have that:*

$$I(T; X, Y) = I(T; Y)$$
$$S(T; X, Y) = U_X(T; X, Y) = 0$$

Proof. By the MI chain rule, Eq. (3.5.20), we have that

$$I(T; X, Y) = I(T; Y) + I(T; X|Y). \tag{3.6.18}$$

Since X and T are conditionally independent, we have that $I(T; X|Y) = 0$, and thus

$$I(T; X, Y) = I(T; Y). \tag{3.6.19}$$

Usings Eqs. (E1)-(E3), we have

$$R + U_X + U_Y + S = R + U_Y. \tag{3.6.20}$$

Since all atoms are non-negative, it follows that $U_X = S = 0$. □

We turn now to the other PID that we consider in this work: the $I_\cap^{\mathrm{PM}}$ PID from [45]. Let us again consider the separation of mutual information into entropic components, as we did in Eq. (3.6.10). The pointwise MI function can be split into a difference of surprisals in more than one way:

$$\log \frac{p(\boldsymbol{x}, t)}{p(\boldsymbol{x})p(t)} = \log \frac{1}{p(t)} - \log \frac{1}{p(t|\boldsymbol{x})} \tag{3.6.21}$$
$$= \log \frac{1}{p(\boldsymbol{x})} - \log \frac{1}{p(\boldsymbol{x}|t)} \tag{3.6.22}$$

Such decompositions are the pointwise analogue of the relationship between MI and (conditional) entropy. Note that (3.6.21) is the pointwise counterpart

to (3.6.10). The $I_\cap^{\mathrm{PM}}$ PID is concerned instead with (3.6.22), as this is the decomposition of the two that fulfills the axiomatic approach in [46].

Whereas the surprisal functions (pointwise entropy) are always nonnegative for discrete sources, pointwise mutual information can be negative even for discrete sources, if $p(t|\boldsymbol{x}) < p(t)$ (equivalently, $p(\boldsymbol{x}|t) < p(\boldsymbol{x})$). In Eq. (3.6.22), pointwise mutual information is decomposed into two positive quantities: the surprisal of a predictor, and the surprisal of the predictor conditioned on the target. Using this decomposition, Finn and Lizier [45] provided an alternative definition of redundant information that relies upon a pointwise axiomatic approach developed in [46]. In their definition, redundancy minimizes *each entropic component separately*. This has the consequence of allowing the signed nature of pointwise mutual information to carry over to their PID atoms in expectation. Their framework allows each information atom in the PID in Fig 3.12 to be decomposed into the difference of two components, which they call **specificity** and **ambiguity**. We first define their redundancy function, I^{PM}.

Definition 15. *Let T be a target variable and $\boldsymbol{X}_1, ...\boldsymbol{X}_m$ be a collection of sources, such that the conditions of Definition 14 hold. Then the redundant information provided by the sources $\{\boldsymbol{X}_i\}$ about T is defined as the difference of the redundant **specificity** $I_\cap^{PM,+}$ and the redundant **ambiguity** $I_\cap^{PM,-}$:*

$$I_\cap^{PM}(T; \boldsymbol{X}_1, ..., \boldsymbol{X}_m) = I_\cap^{PM,+}(T; \boldsymbol{X}_1, ..., \boldsymbol{X}_m) - I_\cap^{PM,-}(T; \boldsymbol{X}_1, ..., \boldsymbol{X}_m) \quad (3.6.23)$$

where

$$I_{\cap}^{PM,+} := \mathbb{E} \min_{k=1,\ldots,m} s_{\boldsymbol{X}_k}(\boldsymbol{X}_k) \tag{3.6.24}$$

$$I_{\cap}^{PM,-} := \mathbb{E} \min_{k=1,\ldots,m} s_{\boldsymbol{X}_k|T}(\boldsymbol{X}_k|T) \tag{3.6.25}$$

In fact, this latter definition allows one to decompose every atom of information from our bivariate PID into its specificity and redundancy:

$$R^{\mathrm{PM}}(T;X,Y) = R^{+}(T;X,Y) - R^{-}(T;X,Y) \tag{3.6.26a}$$

$$U_X^{\mathrm{PM}}(T;X \setminus Y) = U^{+}(T;X \setminus Y) - U^{-}(T;X \setminus Y) \tag{3.6.26b}$$

$$U_Y^{\mathrm{PM}}(T;Y \setminus X) = U^{+}(T;Y \setminus X) - U^{-}(T;Y \setminus X) \tag{3.6.26c}$$

$$S^{\mathrm{PM}}(T;X,Y) = S^{+}(T;X,Y) - S^{-}(T;X,Y) \tag{3.6.26d}$$

where

$$R^{\pm}(T;X,Y) = I_{\cap}^{\mathrm{PM},\pm}(T;X,Y) \tag{3.6.26e}$$

$$U^{\pm}(T;X \setminus Y) = I_{\cap}^{\mathrm{PM},\pm}(T;X) - I_{\cap}^{\mathrm{PM},\pm}(T;X,Y) \tag{3.6.26f}$$

$$U^{\pm}(T;Y \setminus X) = I_{\cap}^{\mathrm{PM},\pm}(T;Y) - I_{\cap}^{\mathrm{PM},\pm}(T;X,Y) \tag{3.6.26g}$$

$$\begin{aligned} S^{\pm}(T;Y \setminus X) = I_{\cap}^{\mathrm{PM},\pm}(T;\{X,Y\}) - I_{\cap}^{\mathrm{PM},\pm}(T;X) \\ - I_{\cap}^{\mathrm{PM},\pm}(T;Y) + I_{\cap}^{\mathrm{PM},\pm}(T;X,Y) \end{aligned} \tag{3.6.26h}$$

To more clearly see the difference between these two definitions, consider the following discrete example.

Example 7 (PIDs of AND and OR Distributions)**.** *Consider the follow distributions* $T = X \vee Y$ *and* $T = X \wedge Y$

X	Y	T *(AND)*	T *(OR)*	p
0	0	0	0	1/4
0	1	0	1	1/4
1	0	0	1	1/4
1	1	1	1	1/4

For both of these, we have mutual informations (where log *is base 2 in the below computations):*

$$I(T;X,Y) = 2 - \frac{3}{4}\log 3 \approx 0.8113, \tag{3.6.27}$$

$$I(T;X) = I(T;Y) = \frac{3}{2} - \frac{3}{4}\log 3 \approx 0.3113. \tag{3.6.28}$$

Then any partial information decomposition of $I(T;X,Y)$ *satisfying Eqs E1-E3 will be of the following form:*

$$U_Y = U_X, \tag{3.6.29}$$

$$R = \left(\frac{3}{2} - \frac{3}{4}\log 3\right) - U_X, \tag{3.6.30}$$

$$S = \frac{1}{2} - U_X. \tag{3.6.31}$$

For the $I_\cap^{min}$ *and* $I_\cap^{PM}$ *PIDs, we have*

PID	R	U_X	U_Y	S
$I_\cap^{min}$	$\frac{3}{2} - \frac{3}{4}\log 3$	0	0	$\frac{1}{2}$
$I_\cap^{PM}$	$\frac{7}{4} - \frac{3}{4}\log 3$	$-\frac{1}{4}$	$-\frac{1}{4}$	$\frac{3}{4}$

This example demonstrates an important principle for finite PIDs. Namely, that given two PIDs $\mathfrak{d}_1, \mathfrak{d}_2$ of a distribution, we have that $R^{(1)} \leq R^{(2)}$ if and

only $S^{(1)} \leq S^{(2)}$. In fact, looking at Eqs E1-E3, we see that

$$S(T;X,Y) - R(T;X,Y) = I(T;X,Y) - I(T;X) - I(T;Y) \tag{3.6.32}$$

Although S and R depend on the PID used, the right-hand side contains only mutual informations, which are defined independently of our choice of PID. Thus, in the case when all the atoms are finite, we have that

$$R^{(2)} - R^{(1)} = S^{(2)} - S^{(1)} = -(U_X^{(2)} - U_X^{(1)}) = -(U_Y^{(2)} - U_Y^{(1)}) \tag{3.6.33}$$

This means that synergistic and redundant information must increase or decrease by the same amount when one changes their PID definition.

We have now introduced both the notion of a continuous interaction and defined $I_\cap^{\min}$ and $I_\cap^{\mathrm{PM}}$ for both discrete and continuous variables. We will now proceed to a simple example of a continuous interaction in order to demonstrate the explicit computation and use of the continuous PIDs.

3.7 Continuous PID of Linear Interactions

Now that we have defined bivariate interactions, the PID framework, and our two redundancy functions $I_\cap^{\min}$ and $I_\cap^{\mathrm{PM}}$, we may present a fully worked out example of the PID of a continuous interaction. A simple linear interaction kernel can be solved analytically without much difficulty (Theorems 2 & 3). These computations will prove quite instructive in elucidating the difference between these two measures — that is, between $I_\cap^{\min}$ and $I_\cap^{\mathrm{PM}}$. As noted in Sec. 3.1.2.4, our results in this section for the $I_\cap^{\min}$ PID overlap with those in [6], although the approach is different. In the following example, we will see how the continuous $I_\cap^{\min}$ induces a fully non-negative PID $\mathfrak{d}_{\min}$, while $I_\cap^{\mathrm{PM}}$ allows for negative atoms. Crucially, the limiting properties of these PIDs will analytically support the specificity (and lack thereof) in $S^{\min}$ (S^{PM}) that we observed in our experiments in Sec. 3.3.

Consider the following noiseless Gaussian interaction $X, Y \to_\rho T$:

$$T = aX + bY \qquad \text{where } 0 < a < b \tag{3.7.1}$$

It immediately follows that

$$T \sim N(0, a^2 + b^2 + 2\rho ab) \tag{3.7.2}$$

Since this interaction is noiseless, we know that $I(T; X, Y) = \infty$ (Sec. 3.6.1). We affirm that $I(T; X)$ and $I(T; Y)$ are finite, however.

Proposition 8. *Let $X, Y \to_\rho T$ be the linear interaction as in Eq. (3.7.1).*

Then the mutual informations between T and each predictor separately are

$$I(T;X) = -\log b + \log \sigma_T + I(X;Y) \tag{3.7.3}$$
$$= \log \frac{\sqrt{a^2+b^2+2ab\rho}}{b\sqrt{1-\rho^2}}$$
$$I(T;Y) = -\log a + \log \sigma_T + I(X;Y) \tag{3.7.4}$$
$$= \log \frac{\sqrt{a^2+b^2+2ab\rho}}{a\sqrt{1-\rho^2}}$$

Proof. Consider $(X,Y) \to (X,T)$ as defined by (3.7.1) as an invertible linear interaction from one bivariate Gaussian vector to another. Per Rule 1 in the Appendix, we have that

$$\boldsymbol{\mu}_{X,T} = (0,0),$$
$$\Sigma_{X,T} = \begin{bmatrix} 1 & a+\rho b \\ a+\rho b & \sigma_T^2 \end{bmatrix},$$
$$\sigma_T^2 = a^2 + b^2 + 2\rho ab.$$

Moreover, we have the correlation coefficient:

$$\rho_{X,T} = \frac{a+\rho b}{\sigma_T}. \tag{3.7.5}$$

From Example 5, the MI between two correlated Gaussians (U,V) with correlation ρ is given by:

$$I(U;V) = -\frac{1}{2}\log(1-\rho^2). \tag{3.7.6}$$

Using this formula and Eq. (3.7.5), we arrive at Eq. (3.7.3). The argument for Eq. (3.7.4) is identical, with a and b permuted in the expressions for $\Sigma_{Y,T}$ and $\rho_{Y,T}$. □

We see immediately that, since $a < b$, we have $I(T;Y) > I(T;X)$. In fact, we will see that, under the continuous extension of the $I_\cap^{\min}$ PID, the redundant information $R^{\min}$ coincides with $I(T;X)$, while $I(T;Y) = I(T;X) + U_Y$ for the unique atom U_Y. Compare with the work in [6] and the MMI PID (Sec. 3.1.2.4).

Before we move on to the continuous PIDs of this interaction, we examine the behavior of mutual information to anticipate what we might intuitively expect from a PID. In particular, what happens as a becomes arbitrarily small relative to b, and X approaches conditional independence of T? Observe that, as $a \to 0^+$, we have $I(T;Y) \to \infty$. This makes sense, as we are approaching $T = bY$, in which case Y and T have a degenerate joint distribution (see Proposition 5 and Corollary 1). However, we do not necessarily have that $I(T;X) \to 0$ as $a \to 0^+$. When $\rho \neq 0$ and our predictors are dependent, we instead have that

$$\lim_{a \to 0^+} I(T;X) = I(X;Y) = \log \frac{1}{\sqrt{1-\rho^2}}. \tag{3.7.7}$$

Even if X does not directly affect T, the dependency between X and Y still leads to positive mutual information between X and T. Only when $\rho = 0$ and X informs on neither Y nor T can we claim $I(T;X) = 0$. On the other hand, as $|\rho| \to 1$, this limit tends toward ∞, as X tends toward perfect information of Y, which in turn provides perfect information of T. Stepping back to the intermediate case $0 < |\rho| < 1$, the existence of positive MI despite conditional independence (of X and T given Y when $a = 0$) hints at what we might hope to see in the continuous PID. Since $I(T;X) = R + U_X$, any positive information must be located in one of these two atoms. We would hope, if the information is purely mediated through the correlated information between X and Y, that

exactly this information would be located in R, while U_X would be zero. As we shall see, $I_\cap^{\mathrm{PM}}$ does not behave in such a manner.

3.7.1 Computation of Continuous PIDs $\mathfrak{d}_{\mathrm{min}}$ and $\mathfrak{d}_{\mathrm{PM}}$

In this section, we compute the $I_\cap^{\mathrm{min}}$ and $I_\cap^{\mathrm{PM}}$ PIDs — Theorems 2 & 3, respectively — for the continuous interaction in Eq. (3.7.1). The structure of this section is as follows. First, we state both theorems without their proofs (proofs to follow). Next, since the $I_\cap^{\mathrm{min}}$ PID depends upon the computation of the specific information functions $I_X(t)$ and $I_Y(t)$, we present these in Lemma 1, followed by its proof. We then provide the proof for Theorem 2, which follows immediately from Lemma 1 and the MIs from Prop. 8. It is worth highlighting that this computation is particularly simple due to the convenient property that $I_Y(t) \geq I_X(t)$ for all t, which is a necessary and sufficient condition for the coincidence of $I_\cap^{\mathrm{min}}$ redundancy and minimal predictor MI, i.e. $I(T;X) = R^{\mathrm{min}}(T;X,Y)$. We then turn to the $I_\cap^{\mathrm{PM}}$ PID. Since this PID is the difference of the decomposition of specificity $I^+(T;X,Y)$ and ambiguity $I^-(T;X,Y)$, we compute each of those lattices in turn, in Lemmas 2 & 3. With those proven, we conclude this section on the computation of our PIDs.

Theorem 2 (PID $\mathfrak{d}_{\mathrm{min}}$ for Linear Interaction)**.** *Let $X, Y \to_\rho T$ be a linear interaction as in Eq. (3.7.1). Then, using $I_\cap^{\mathrm{min}}$ as defined in Def. 14, it has*

the following PID $\mathfrak{d}_{min}$*:*

$$R = -\log b + \log \sigma_T + I(X;Y) \tag{3.7.8}$$

$$= \log \frac{\sqrt{a^2 + b^2 + 2ab\rho}}{b\sqrt{1-\rho^2}} \tag{3.7.9}$$

$$U_X = 0 \tag{3.7.10}$$

$$U_Y = \log \frac{b}{a} \tag{3.7.11}$$

$$S = \infty \tag{3.7.12}$$

Theorem 3 (PID $\mathfrak{d}_{\mathrm{PM}}$ for Linear Interaction)**.** *Let* $X, Y \to_\rho T$ *be a linear interaction as in Eq. (3.7.1). Then, using* I_{PM} *as defined in Def. 15, it has the following PID* $\mathfrak{d}_{PM} = \mathfrak{d}^+_{PM} - \mathfrak{d}^-_{PM}$*:*

$$R = \log \frac{\sqrt{a^2 + b^2 + 2ab\rho}}{a\sqrt{1-\rho^2}} - \frac{1}{\pi}\sqrt{1-\rho^2} \tag{3.7.13}$$

$$U_X = \frac{1}{\pi}\sqrt{1-\rho^2} - \log\left(\frac{b}{a}\right) \tag{3.7.14}$$

$$U_Y = \frac{1}{\pi}\sqrt{1-\rho^2} \tag{3.7.15}$$

$$S = \infty \tag{3.7.16}$$

We begin with the $I_\cap^{\min}$ lattice $\mathfrak{d}_{\min}$. In order to demonstrate Theorem 2, we need to first compute the specific information functions for each predictor variable.

Lemma 1 (Specific Information Functions for Linear Interaction)**.** *Let* $X, Y \to_\rho T$ *be a linear interaction as in Eq. (3.7.1). Then the specific information functions* $I_X(t)$ *and* $I_Y(t)$ *as given in Def. 9 are given as the following:*

$$I_X(t) = -\log b + \frac{1}{2\sigma_T^4}\left[b^2(1-\rho^2)\sigma_T^2 + (a+\rho b)^2 t^2\right] + c \tag{3.7.17a}$$

$$I_Y(t) = -\log a + \frac{1}{2\sigma_T^4}\left[a^2(1-\rho^2)\sigma_T^2 + (b+\rho a)^2 t^2\right] + c \tag{3.7.17b}$$

where

$$c = \log \sigma_T - \log \sqrt{1-\rho^2} - \frac{1}{2} \tag{3.7.17c}$$

$$\sigma_T^2 = a^2 + b^2 + 2\rho ab \tag{3.7.17d}$$

Moreover, their difference $I_Y - I_X$ is positive, given by:

$$I_Y(t) - I_X(t) = \log \frac{b}{a} + \frac{1}{2\sigma_T^4}\left[(1-\rho^2)(b^2-a^2)(t^2-\sigma_T^2)\right] \tag{3.7.17e}$$

Proof of Lemma 1. As in the proof of Prop 8, we have that

$$\begin{aligned}
\boldsymbol{\mu}_{X,T} &= (0,0), \\
\Sigma_{X,T} &= \begin{bmatrix} 1 & a+\rho b \\ a+\rho b & \sigma_T^2 \end{bmatrix}, \\
\sigma_T^2 &= a^2 + b^2 + 2\rho ab, \\
\rho_{X,T} &= \frac{a+\rho b}{\sigma_T}.
\end{aligned}$$

Using Rule 2, this gives us a conditional distribution $X_{T=t} = (X|T=t)$ for any choice of t:

$$X_{T=t} \sim N(\mu_{X|T=t}, \sigma^2_{X|T=t})$$

where

$$\mu_{X|T=t} = \frac{(a+\rho b)t}{\sigma_T^2} \tag{$*$}$$

$$\sigma^2_{X|T=t} = \frac{b^2(1-\rho^2)}{\sigma_T^2} \tag{$**$}$$

This allows us to compute the specific information at $T = t$ as the KL divergence between $X|_{T=t}$ and X. Both are univariate Gaussians, so we use Prop. 4, and have that

$$\begin{aligned} I_X(t) &= D(X_{T=t}||X) \\ &= \log\left(\frac{1}{\sigma_{X|T=t}}\right) + \frac{1}{2}\left(\sigma^2_{X|T=t} + \mu^2_{X|T=t}\right) - \frac{1}{2}. \end{aligned}$$

We plug in Eqs ($*$) and ($**$) to arrive at the form in Eq. 3.7.17a. In particular, with c as in Eq. 3.7.17c, note that

$$\begin{aligned} \log\left(\frac{1}{\sigma_{X|T=t}}\right) - \frac{1}{2} &= -\log b + c, \\ \sigma^2_{X|T=t} + \mu^2_{X|T=t} &= \frac{1}{\sigma_T^4}\left[b^2(1-\rho^2)\sigma_T^2 + (a+\rho b)^2 t^2\right]. \end{aligned}$$

The proof for Eq. 3.7.17b is identical under the transposition of a and b.

From Eq. 3.7.17a and Eq. 3.7.17b, we have that Eq. 3.7.17e follows by simply expanding terms and cancelling. To demonstrate that this difference is strictly positive, it suffices to consider the worst-case scenario of (3.7.17e) where $t = 0$:

$$I_Y(0) - I_X(0) = \log\frac{b}{a} - \frac{1}{2\sigma_T^2}\left[(1-\rho^2)(b^2-a^2)\right]$$

We make the substitution $b = \gamma a$, and the above becomes a function f of γ (and ρ, implicitly):

$$f(\gamma) := \log \gamma - \frac{1(1-\rho^2)(\gamma^2-1)}{2(\gamma^2+2\rho\gamma+1)} \tag{3.7.18}$$

Note that $\gamma > 1$ since $0 < a < b$. In Computation 4, we demonstrate this function is strictly increasing for $\gamma \geq 1$, under the assumption $\rho \in [-1, 1]$. Since $f(1) = 0$, it follows that $f(\gamma) > 0$ for any $\gamma > 1$ and $\rho \in [-1, 1]$. □

We are now ready to prove Theorem 2.

Proof of Theorem 2. The form of $I_\cap^{\min}$ follows from $I_X(t) \leq I_Y(t)$ for all t and hence $I_\cap^{\min} = \mathbb{E}(I_X(T))$ where $I_X(T)$ has the form of Eq. (3.7.17c). It is immediate that $U_X^{\min} = 0$ and the form of $U_Y^{\min}$ follows from combining equations Eq. (3.7.4) and Eq. (3.7.17c). Lastly, $S^{\min} = \infty$ follows from $I(T; X, Y) = \infty$ and Eq. (3.6.17). □

We now turn to the proof of the $I_\cap^{\mathrm{PM}}$ PID in Theorem 3. This PID follows readily from computing the specificity and ambiguity lattices, which we do in Lemmas 2 and 3, respectively. The specificity lattice does not depend on the interaction itself, only the distribution of the predictors X and Y. Thus

Lemma 2 (I_{PM} Specificity Lattice for Gaussian PID)**.** *Let $X, Y \to_\rho T$ be a Gaussian bivariate interaction, noisy or noiseless. Then the specificity lattice $\mathfrak{d}_{PM}^+$ as defined by Def. 15 has the following form:*

$$R^+ = \log(\sqrt{2\pi e}) - \frac{1}{\pi}\sqrt{1-\rho^2} \tag{3.7.19}$$

$$U_X^+ = \frac{1}{\pi}\sqrt{1-\rho^2} \tag{3.7.20}$$

$$U_Y^+ = \frac{1}{\pi}\sqrt{1-\rho^2} \tag{3.7.21}$$

$$S^+ = \log(\sqrt{2\pi e}) + \log(\sqrt{1-\rho^2}) - \frac{1}{\pi}\sqrt{1-\rho^2} \tag{3.7.22}$$

Proof. We begin by computing the redundant specificity

$$R^+ = \mathbb{E}_{X,Y} r^+(X,Y)$$

where

$$r^+(x,y) = \min(-\log(p_X(x)), -\log(p_Y(y))).$$

We take the expectation over $X \times Y$ space. Since $p_X = p_Y \sim N(0,1)$, we have that

$$p_X(u) = p_Y(u) = \frac{1}{\sqrt{2\pi}} e^{\frac{-u^2}{2}}$$

and so

$$-\log p_X(x) \leq -\log p_Y(y) \Leftrightarrow x^2 \leq y^2.$$

This yields that (where we use the fact that $2\min(x,y) = x + y - |x-y|$

$$\begin{aligned}
\mathbb{E}_{X,Y} r^+(X,Y) &= \log(\sqrt{2\pi}) + \frac{1}{2}\mathbb{E}(\min(X^2, Y^2)) \\
&= \log(\sqrt{2\pi}) + \frac{1}{4}\left(\mathbb{E}X^2 + \mathbb{E}Y^2 - \mathbb{E}(|X^2 - Y^2|)\right) \\
&= \log(\sqrt{2\pi}) + \frac{1}{4}\left(2 - \mathbb{E}(|X-Y|)\mathbb{E}(|X+Y|)\right) \\
&= \log(\sqrt{2\pi}) + \frac{1}{4}\left(2 - \frac{2}{\pi}\sqrt{4-4\rho^2}\right) \\
&= \log(\sqrt{2\pi}) + \frac{1}{2} - \frac{1}{\pi}\sqrt{1-\rho^2},
\end{aligned}$$

where we used above the fact that $X - Y \sim \text{Norm}(0, 2-2\rho)$ is independent of $X + Y \sim \text{Norm}(0, 2+2\rho)$ (indeed, the covariance is $\text{Cov}(X-Y, X+Y) = 0$), and for $W \sim \text{Norm}(0, \sigma^2)$, $\mathbb{E}(|W|) = \sigma\sqrt{2/\pi}$.

We now turn to the unique specificities U_X^+ and U_Y^+. The specificity components $I^+(T;X)$ and $I^+(T;Y)$ of the mutual informations are just the entropies $h(X)$ and $h(Y)$, by definition. Since X and Y are standard normal variables in their marginal distributions, we have that

$$I^+(T;X) = I^+(T;Y) = \log(\sqrt{2\pi e}).$$

Thus, unique specificity becomes

$$\begin{aligned} U_X^+ = U_Y^+ &= I^+(T;Y) - R^+ \\ &= \frac{1}{\pi}\sqrt{1-\rho^2} \end{aligned}$$

Finally, we compute synergistic specificity. Again, the specificity component of the mutual information $I(T;X,Y)$ is the joint entropy:

$$I^+(T;\{X,Y\}) = h(X,Y) = \log(2\pi e) + \log(\sqrt{1-\rho^2}).$$

So, by Eq. (3.6.26h), we have that

$$S^+ = \log(\sqrt{2\pi e}) + \log(\sqrt{1-\rho^2}) - \frac{1}{\pi}\sqrt{1-\rho^2}$$

□

Whereas the specificity lattice depends only upon the distribution of the predictors, the ambiguity lattice depends upon the actual interaction. This is also the component where the synergistic component blows up for a noiseless interaction.

Lemma 3 (I_{PM} Ambiguity Lattice for Linear Interaction)**.** *Let $X, Y \to_\rho T$ be a linear interaction as in Eq. (3.7.1). Then the ambiguity lattice $\mathfrak{d}_{PM}^-$, as*

defined by Def. 15 has the following form:

$$R^- = \log(\sqrt{2\pi e}) + \log\sqrt{1-\rho^2} + \log\left(\frac{a}{\sqrt{a^2+b^2+2ab\rho}}\right) \tag{3.7.23}$$

$$U_X^- = \log\left(\frac{b}{a}\right) \tag{3.7.24}$$

$$U_Y^- = 0 \tag{3.7.25}$$

$$S^- = -\infty \tag{3.7.26}$$

Proof. We begin with redundant ambiguity. A change of variables reveals that, for a given allowable triplet $(X, Y, T) = (x, y, t)$:

$$\frac{p_{X,T}(x,t)}{p_T(t)} = \frac{1}{b}\frac{p_{X,Y}(x,y)}{p_T(t)} \tag{$*$}$$

$$\frac{p_{Y,T}(y,t)}{p_T(t)} = \frac{1}{a}\frac{p_{X,Y}(x,y)}{p_T(t)} \tag{$**$}$$

We see that, since $0 < a < b$, we have that

$$-\log\frac{p_{Y,T}(y,t)}{p_T(t)} \leq -\log\frac{p_{X,T}(x,t)}{p_T(t)}$$

and so, using Def. 15:

$$\begin{aligned} R^- &= \mathbb{E}\left(-\log\frac{p_{Y,T}(y,t)}{p_T(t)}\right) \\ &= \log a + h(X,Y) - h(T) \end{aligned} \tag{$***$}$$

Using the formulae in Eqs 3.5.16 and 3.5.27, as well as Eq. 3.7.2, we have that

$$h(T) = \log\sqrt{2\pi e} + \log(\sigma_T)$$

$$h(X,Y) = \log(2\pi e) + \log(\sqrt{1-\rho^2})$$

Thus, redundant ambiguity becomes

$$R^- = \log \sqrt{2\pi e} + \log \sqrt{1-\rho^2} + \log \left(\frac{a}{a^2 + b^2 + 2\rho ab} \right).$$

We can now turn to the unique ambiguity atoms U_X^- and U_Y^-. By Def. 15, Eqs (3.6.26f) and (3.6.26g), we have that

$$\begin{aligned} U_X^- &= \mathbb{E}\left(-\log \frac{p_{X,T}(x,t)}{p_T(t)} \right) - R^- \\ U_Y^- &= \mathbb{E}\left(-\log \frac{p_{Y,T}(y,t)}{p_T(t)} \right) - R^- \end{aligned}$$

By plugging in Eq. $(***)$, we arrive at

$$\begin{aligned} U_Y^- &= 0 \\ U_X^- &= \mathbb{E}\left(-\log \frac{p_{X,T}(x,t)}{p_T(t)} \right) - \mathbb{E}\left(-\log \frac{p_{Y,T}(y,t)}{p_T(t)} \right) \end{aligned}$$

We expand U_X^- with Eqs $(*)$ and $(**)$:

$$\begin{aligned} U_X^- &= \mathbb{E}\left(-\log \frac{p_{X,T}(x,t)}{p_T(t)} \right) - \mathbb{E}\left(-\log \frac{p_{Y,T}(y,t)}{p_T(t)} \right) \\ &= (\log b + h(X,Y) - h(T)) - (\log a + h(X,Y) - h(T)) \\ &= \log \left(\frac{b}{a} \right) \end{aligned}$$

We have computed finite values for the atoms R^-, U_X^-, and U_Y^-. Since $I(T;X,Y) = \infty$, we note that $I^-(T;X,Y) = -\infty$, and thus Eq. (3.6.26h), when rewritten in the following form:

$$I^- = R^- + S^- + U_X^- + U_Y^-$$

makes it determinate that $S^- = -\infty$.

□

Thus, we have the exact values of the continuous PIDs $\mathfrak{d}_{\min}$ and $\mathfrak{d}_{\mathrm{PM}}$ for the linear interaction. We now turn the question of which PID, $\mathfrak{d}_{\mathrm{PM}}$ or $\mathfrak{d}_{\min}$, is preferable (and in what sense) for this interaction (3.7.1).

3.7.2 Properties, Limits, and Insights of the PIDs of Linear Interactions

Now that we have computed the PIDs of the linear interaction, we may ask if their behavior sheds light on the issues of comparative specificity that we saw in Section 3.3. Recall Experiment II , in which we considered a network of two 10-stars centered on hubs Y_1 and Y_2, and a response of the form:

$$T = g(X_1, Y_1) + \beta Y_2$$

where g is the kernel from Eq (E4), and β a positive constant. We recall that $S^{\min}$ could distinguish between the true interaction (X_1, Y_1) and a 'false' one (X_j, Y_2) for any X_j on the second star. Moreover, the $I_\cap^{\min}$ PID tended to properly locate most of the information of a false pair $I(T; X_j, Y)$ in the unique atom $U_Y^{\min}$, whereas the $I_\cap^{\mathrm{PM}}$ PID behaved less intuitively (Fig. 3.8).

Putting aside the WB axioms, for the moment, we would like to identify behavior that we consider intuitive for the bivariate PIDs of 'true' and 'false' interactions. Many of these are mathematically trivial, relative to the strength of the axioms from which they follow. Our purpose is to emphasize the full implications of what is lost in the absence of the WB axioms: core properties of information formalism that strongly align with technical intuition. Consider the following expected properties:

1. **Monotonicity Bound on Redundant Information**

   For any X_i, Y_j, we expect

   $$R(T; X_i, Y_j) \leq \min(I(T; X_i), I(T; Y_J)).$$

   Put into words, the *redundant* information in both predictors is no greater than the total information in either predictor, with respect to the target.[11]

2. **Data Processing Inequality of Redundant Information**

   If (X_i, Y_j) is a false interaction, then

   $$R(T; X_i, Y_j) \leq I(X_i; Y_J)$$

   In words, given a Markov chain $X_i \to Y_j \to T$, where Y_j is a true predictor and X_i is independent of T except for its dependency on Y, then the redundant information of both predictors is no greater than the information between $X_i \leftrightarrow Y_j$. [12]

3. **Specificity of Unique Information:**

   If X_i, Y_j are a false interaction, with $X_i \perp T | Y_j$, then we expect $U_{X_i}(T; X_i, Y_j) = 0$. In words, given a Markov chain $X_i \to Y_j \to T$, there should be no information overlap or 'flow' of any kind between X_i and T, if we have accounted entirely for Y_j.

The first property is just a special case of the monotonicity property (M) in Def. 13. The second property follows from the first and the Data Processing Inequality. The third is Proposition 7 from Section 3.6. The $I_\cap^{\min}$ PID (at least

[11] Compare with the MMI PID for jointly Gaussian variables in [6], discussed more in Sec. 3.1.2.4.

[12] Compare with (3.7.7) in the introduction to the current Sec. 3.7.

in discrete settings), and similarly every PID that induces only non-negative information atoms, will fulfill these properties. The $I_\cap^{\mathrm{PM}}$ PID will not. We saw in Section 3.3 that such counter-intuitive behavior arguably detracts from the the *utility* of S^{PM} as a statistic for edge nomination.

The linear interaction under consideration in this section is different, in many respects, from the experiments we conducted in Section 3.3. First, in this section, we are considering a continuous extension of PID, applied to the theoretical distributions of the predictors and targets, whereas all the experiments in Section 3.3 used discretized data, i.e. used empirical distributions for a discretization of all variables. In addition, in this section, we consider only two predictors, and a single interaction of a linear form. This discretization alone has major consequences, as we have seen, since continuous MI and PID information can be infinite. Nonetheless, the differences between how the $I_\cap^{\min}$ and $I_\cap^{\mathrm{PM}}$ PIDs treat both true and false interactions have as much to do with how redundant and unique information atoms are assigned, since synergy is entirely determined by the balance of information in $I(T;X)$ and $I(T;Y)$ in the atoms R, U_X, U_Y. For instance, in Fig. 3.8 and Fig. 3.14, we see that the elevated $\hat{S}^{\mathrm{PM}}$ of both true and false interactions is more or less the mirror of the negative $\hat{U}_X^{\mathrm{PM}}$. The following limiting properties of the linear interaction in this section conform to this phenomenon.

Corollary 2. *Let $\{X, Y \to_\rho T\}_a$ be the collection of linear interactions as in Eq (3.7.1), where $b > 0$ and $0 < \rho < 1$ are fixed and only $a \in (0, b)$ varies. Then as $a \to 0^+$, we have the following limits:*

$$R^{min} \to I(X;Y) \qquad (3.7.27a)$$
$$= -\log\sqrt{1-\rho^2}$$
$$U_X^{min} = 0 \qquad (3.7.27b)$$
$$U_Y^{min} \to \infty \qquad (3.7.27c)$$
$$S^{min} = \infty \qquad (3.7.27d)$$

$$R^{PM} \to \infty \qquad (3.7.28a)$$
$$U_X^{PM} \to -\infty \qquad (3.7.28b)$$
$$U_Y^{PM} = \frac{1}{\pi}\sqrt{1-\rho^2} \qquad (3.7.28c)$$
$$S^{PM} = \infty \qquad (3.7.28d)$$

where equality denotes a fixed value for all $a \in (0, b)$ (and hence trivial limit). Moreover, the following ratios have limits:

$$\frac{U_Y^{min}}{I(T;Y)} \to 1 \qquad (3.7.29a)$$
$$\frac{R^{min}}{I(T;Y)} \to 0 \qquad (3.7.29b)$$

$$\frac{U_X^{PM}}{I(T;Y)}, \frac{U_X^{PM}}{R^{PM}} \to -1 \qquad (3.7.30a)$$
$$\frac{R^{PM}}{I(T;Y)} \to 1 \qquad (3.7.30b)$$

Combined, this provides

$$\frac{U_X^{PM}}{U_Y^{min}} \to -1 \qquad (3.7.31)$$
$$\frac{U_Y^{min}}{R^{PM}} \to 1 \qquad (3.7.32)$$

Proof. The limits are readily computable with calculus, given the expressions in Theorems 2 & 3. The ratio limits are most easily computed via disentangling

the logarithms:

$$\begin{aligned} U_X^{\mathrm{PM}} &= \log a + c_1(b, \rho) \\ U_Y^{\min} &= -\log a + c_2(b) \\ R^{\mathrm{PM}} &= -\log a + c_3(a, b, \rho) \\ I(T;Y) &= -\log a + c_4(a, b, \rho) \end{aligned}$$

where the c_i are either fixed terms of b and ρ or functions of a, b, ρ that converge to a finite term as $a \to 0^+$. □

We will now attempt to provide an interpretation of these limits according to our understanding. We begin with the $I_\cap^{\min}$ PID. As $a \to 0^+$, the redundant information $R^{\min}$ approaches the total shared information between X and Y (3.7.27a). Since T becomes asymptoptically determined by Y (that is, $I(T:Y) \to \infty$), as $a \to 0^+$, this makes intuitive sense: any shared uncertainty between X and Y is exactly X's shared uncertainty with any variable exactly determined by Y. It is also intuitive that $U_Y^{\min} \to \infty$ (3.7.27c). Since $|\rho| < 1$, the mutual uncertainty X shares with Y (and thus T) is not all of either's uncertainty: X and Y share finite, rather than perfect, information of each other. The gap from finite to perfect information is infinite, formally rendered $I(T;Y) - R^{\min}(T;X,Y) = \infty$, and so $U_Y^{\min} = \infty$ (3.7.27c), i.e. if Y has perfect information regarding T, then it must be unique to Y. Less intuitive, on the other hand, is that $U_X^{\min} = 0$ even when $a > 0$ (3.7.27b). Though we might desire that $U_X^{\min} \to 0$ as $a \to 0^+$, so long as $a > 0$ and $|\rho| < 1$, there is some portion of uncertainty that X shares with T and not with Y. This is another manifestation of the common critique of $I_\cap^{\min}$, which is that it does not distinguish the same information contained in the predictors from the same amount of information given by the predictors regarding a given outcome

$T = t$. Since the expected total information that X gives about any $T = t$ will always be less than that of Y (as $a < b$), $I_\cap^{\min}$ will assign the full uncertainty that X shares with T to redundant information.

Let us now consider the $I_\cap^{\mathrm{PM}}$ PID $\mathfrak{d}_{\mathrm{PM}}$. To contextualize the limits of unique information, we note that this PID pseudobounds[13] U_X^{PM} and U_Y^{PM} above by the unique specificities $U_X^{\mathrm{PM},+}$ and $U_Y^{\mathrm{PM},+}$. Since Y has zero unique ambiguity, $U_Y^{\mathrm{PM},-} = 0$, we have that $U_Y^{\mathrm{PM}} = U_Y^{\mathrm{PM},+}$, a fixed quantity depending only upon the marginal distribution p_Y. We see that, as $a \to 0^+$ and $I(T;Y) \to \infty$, the $I_\cap^{\mathrm{PM}}$ PID treats the perfect information that Y holds regarding T as shared between the predictors (i.e., captured by the redundancy; see Eq. 3.7.30b), and it treats the finite information in $I(T;X)$, which approaches zero, as what one may imprecisely think of as an *infinite shortcoming* (see Eq. 3.7.30a) of X as a predictor. Explicitly, for the expression for U_X^{PM} in Eq. (3.7.14), this shortcoming is quantified as the term $-\log \frac{b}{a}$, which grows without bound in magnitude in proportion to the discrepancy between the information in X and Y.

In Section 3.8, we will generalize this understanding of unique information to well-behaved kernels, providing a framework to link the analytic and information properties of our variables. This will allow us to see the relationship between the motivation underlying the $I_\cap^{\mathrm{PM}}$ PID developed in [45, 46], and the curious phenomenon of asymmetric negative unique information that we have seen in bivariate interactions. In the discrete context in which PID is typically applied to data, there would not of course be any infinite discrepancies between information atoms. Nonetheless, we contend that these limits demonstrate the intrinsic orientation of the PIDs towards the concepts of redundancy, synergy,

[13] In the discrete case, redundancy and ambiguity atoms are non-negative, and U_X^{PM} is indeed bound above by $U_X^{\mathrm{PM},+}$. This is no longer true in the continuous case, e.g. consider $|\rho| \to 1$ in Lemmas 2 & 3.

and uniqueness of information. We now return to the discrete PID analysis of network simulations in order to further relate the computational analysis of this section with the motivating experiments in Section 3.3.

3.7.3 Network Simulation Analysis of Linear Interactions

We return to simulating network data, this time to bridge our theoretic investigation of the continuous $I_\cap^{\min}$ and $I_\cap^{\mathrm{PM}}$ PIDs in this section with our earlier experimental investigation in Section 3.3. In this section, we run analogous simulation experiments to those previously presented, except now we will utilize the linear interaction kernel (3.7.1). For this kernel, we run into the issue that multiple 'interactions' are non-identifiable (i.e., all genes are contributing to the interaction at the individual and not the paired level). The total response will be a linear combination of some subset of genes, with no meaningful pair distinction. Nonetheless, we will be able recreate—and better understand—similar PID behavior to that observed in Section 3.3, when we were considering more paired/identifiable interactions. We use the same network and procedure as in Experiments I & III (Secs. 3.3.1, 3.3.3). Our response will similarly be sum of four identical interactions on the same hub. Applying the kernel in (3.7.1), this becomes the (pairwise undifferentiated) response:

$$T = \sum_{i=1}^{4} aX_i + 4bY. \tag{3.7.33}$$

Note that, in this random experiment, the pairwise contribution of X_i and Y to the response is now $aX_i + 4bY$, and Y is the more informative gene in the pair whenever $a < 4b$, rather than $a < b$ as in the isolated bivariate system in Secs.3.7.1-3.7.2.

We present the values of the bivariate $I_\cap^{\mathrm{PM}}$ and $I_\cap^{\min}$ PIDs in Fig. 3.13, nor-

malized by the mean $\hat{I}(X_i, Y)$ as in previous visualizations. We see a strong resemblance to Experiment III and Fig. 3.10, albeit with more pronounced limiting behavior. Notably, when a/b is small and the information is concentrated highly in Y, then we see that $\hat{\mathfrak{d}}_{\min}$ locates almost all information in U_Y. As $a \to 4b^-$, this atom's share of $\hat{I}(X_i, Y)$ recedes as the information is redistributed. By contrast, $\hat{\mathfrak{d}}_{\mathrm{PM}}$ locates a relatively large amount of positive information in synergy and redundancy, and less in U_Y even in the a/b small case. Moreover, we see U_X^{PM} takes a negative value to offset the inflated synergy and redundancy. As in Fig. 3.10, the negative of the estimated U_X^{PM} (i.e., $-\hat{U}_X^{\mathrm{PM}}$) resembles the estimated R^{PM} and S^{PM} when $a < 4b$ (corresponding to $I(T;X) < I(T;Y)$).

We have repeatedly highlighted this relationship between significant negative information in U_X^{PM} and the inflated values of S^{PM} and R^{PM}, as this phenomenon is responsible for the non-specificity of the $I_\cap^{\mathrm{PM}}$ PID for our motivating problem of edge nomination (Sec. 3.3, particularly 3.3.2). In order to connect our analytic and simulated explorations, we examined the estimated ratio $\hat{U}_X/\hat{R}$ in Fig. 3.14, for both our current linear interaction simulations and the results from Experiment III , earlier presented in Fig. 3.10. As reference, we also include the theoretical ratio U_X/R for a linear interaction, using the PIDs we just computed in Theorems 2-3 [14]. We see that, as a/b becomes small, the estimated ratio $\hat{U}_X^{\mathrm{PM}}/\hat{R}^{\mathrm{PM}}$ approaches a value near -1, similar to the asymptotic behavior from Corollary 2 in Sec. 3.7.2.

Taken together, this provides evidence that the discrete $I_\cap^{\mathrm{PM}}$ PID, as applied to network data, exhibits the same pathologies as the continuous $I_\cap^{\mathrm{PM}}$ PID explored in this section, at least in the case where the network response T is a linear interaction. Neither the continuous nor discrete $I_\cap^{\mathrm{PM}}$ PID respect

[14] In order to make the isolated interaction more comparable to the four-fold sum, we used a substitution $b \to 4b$ in the analytic expressions, as plotted in all figures.

asymptotic independence in the way we would hope for, as both will locate a large amount of target-predictor MI $I(T;X,Y)$ in the redundancy atom $R^{\mathrm{PM}}(T;X,Y)$. If $I(T;X)$ is small, both the discrete and continuous $I_{\cap}^{\mathrm{PM}}$ PID will account for a large redundancy atom, $R^{\mathrm{PM}} \gg I(T;X)$ (and so the synergy also increases problematically in the discrete case as well), by assigning a proportionate share of negative information to U_X^{PM}. In the next section, we will see that this phenomenon will hold for a more generic class of continuous interactions and kernels g.

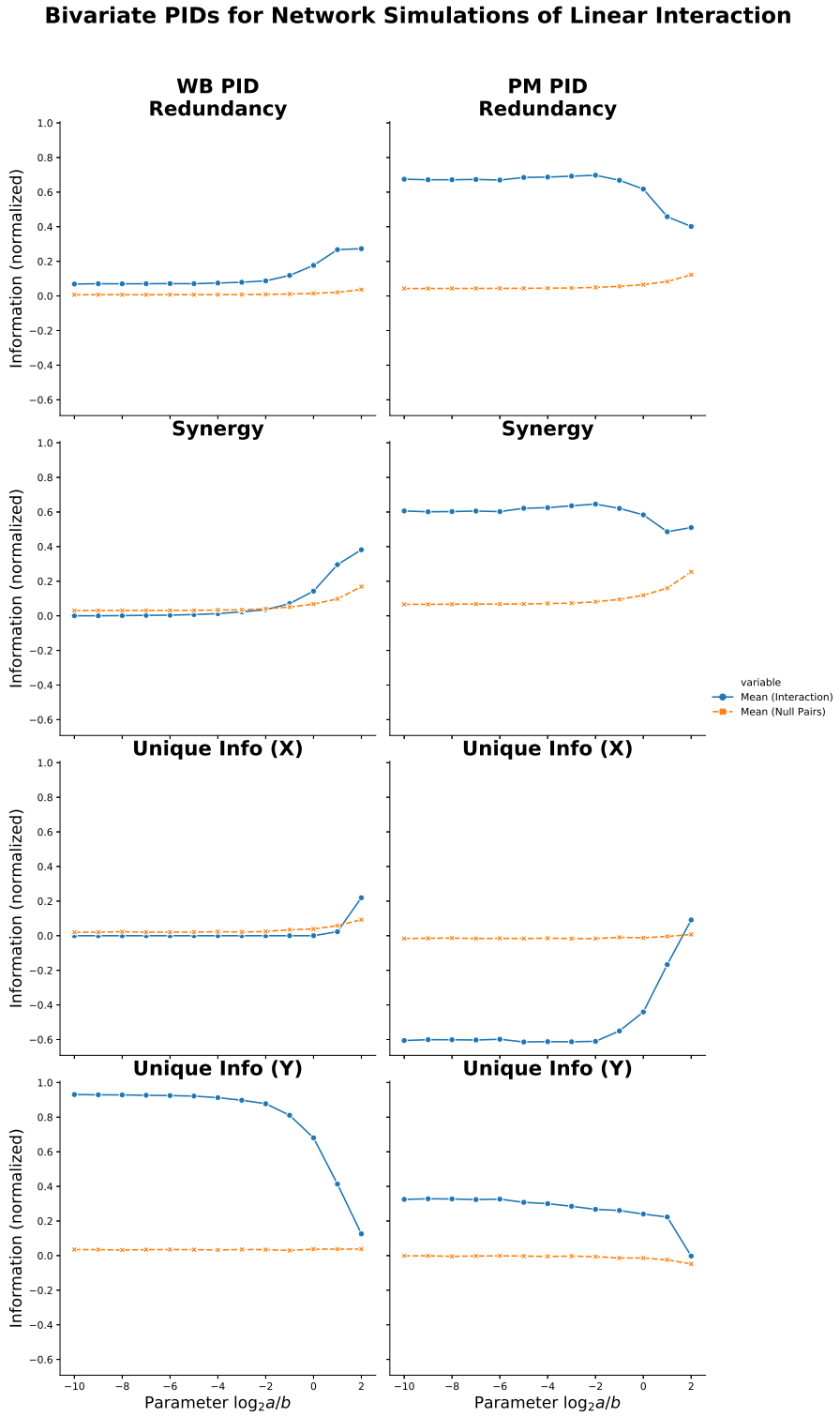


Figure 3.13: **Bivariate PID atoms (normalized) for linear kernel simulations, as a function of coefficient ratio** $\log a/b$ For the linear interaction kernel (Eq 3.7.1), we ran a fourth network simulation experiment on the same network topology as Exps. I and III (Fig. 3.1), and computed the PIDs of the four 'interactions' and all null pairs, as in Section 3.3.

Figure 3.14: **Ratio of unique (X) and redundant information of the interactions in simulations of sigmoidal and linear interaction kernels** From the experiments in Secs. 3.3.3 & 3.7.3, we plot the ratio of unique information of X to redundant information, $\hat{U}_X/\hat{R}$, for both a sigmoidal (Experiment III , Sec. 3.3.3) and linear (Sec. 3.7.3) interaction kernel.

3.8 Unique Information for a Generic Interaction Kernel

We have seen that the difference between the $I_\cap^{\text{min}}$ and $I_\cap^{\text{PM}}$ PID frameworks can be better understood by examining how each treats unique information. As we saw in Eqs (3.6.33), for discrete bivariate PIDs there is a trade-off between the positive information assigned to redundancy and synergy and the unique information atoms. We have emphasized that, since the $I_\cap^{\text{PM}}$ PID does not fulfill the (M) WB axiom, its redundancy and synergy atoms can behave non-intuitively for predictors with asymmetric relationships to the target, e.g., if one is conditionally independent of the target, given the other. In this section, we have a two-fold purpose. First, we wish to expand the analytic analysis of the continuous $I_\cap^{\text{min}}$ and $I_\cap^{\text{PM}}$ PIDs to a large class of interactions, including the sigmoidal switch kernel in Eq. (E4). This will provide a more general principle to explain $I_\cap^{\text{PM}}$'s non-specificity for edge nomination, including negative unique information and inflated redundancy. We demonstrate a general form for unique information that is approachable with calculus.

Second, from this more general principle, we will show in Sec. 3.8.2 that the $I_\cap^{\text{PM}}$ remains loyal the principles underlying the specificity-ambiguity information decomposition that the work of Finn and Lizier takes as its starting point (the entropic decomposition in Eq. 3.6.21). Up to this point, we have mostly highlighted the shortcomings of the $I_\cap^{\text{PM}}$ PID in a specific network inference context and under certain assumptions about how a PID 'ought' to behave with respect to response-irrelevant predictors. The $I_\cap^{\text{PM}}$ PID was developed as the result of thoughtful reflections upon the nature of pointwise information, and the probability mass exclusions that it quantifies [45]. To our knowledge, we are the first to investigate the continuous $I_\cap^{\text{PM}}$ PID and its response to the analytic relationships between target and predictor variables. While we show

$I_\cap^{\mathrm{PM}}$ is not as suitable for edge nomination, we here tie our continuous extension back to the important motivating roots and contributions of the original $I_\cap^{\mathrm{PM}}$ framework in [45] and [46].

3.8.1 Unique Information for a Generic Kernel

In the previous section, we were able to compute every atom of the $I_\cap^{\mathrm{min}}$ and $I_\cap^{\mathrm{PM}}$ PIDs explicitly for a linear interaction. We were able to do so because, as a linear combination of Gaussian variables, T itself had a normal distribution. It is less clear what distribution T takes on when we instead use, for instance, the sigmoidal switch kernel in Eq (E4). By limiting our attention to unique information, we may sidestep the question of T's distribution.

We will demonstrate that for well-behaved interaction kernels g, the unique information for the $I_\cap^{\mathrm{PM}}$ lattice becomes the positive expectation of the log-ratio between the its partial derivatives — that is, the expectation of $\log |\partial_y g|/|\partial_x g|$ where it is positive. This will make it clear that, if the relative sensitivity of the kernel g upon the predictor X is much smaller than that of Y, $U_X^{\mathrm{PM},-}$ will become large. Given $|\partial_y g|$ bounded away from 0, we will see that

$$U_X^{\mathrm{PM},-} \to -\infty \text{ as } |\partial_x g| \to 0$$

which in turn implies that

$$U_X^{\mathrm{PM}} \to \infty \text{ as } |\partial_x g| \to 0.$$

We will make this more precise in Corollary 3. In this way, we generalize the limiting behavior of the $I_\cap^{\mathrm{PM}}$ PID that we saw in Section 3.7.2. We thereby also give a theoretical explanation for the negative unique information estimated for conditionally independent, non-predicting variables in Section 3.3. Recall

that this behavior coincided with inflated values for estimated redundancy and synergy, $\hat{R}^{\mathrm{PM}}$ and $\hat{S}^{\mathrm{PM}}$, that undermined the ability of the $I_{\cap}^{\mathrm{PM}}$ PID to distinguish true interactions from univariate signals in Exp II .

We now provide the statement of our main result for this section.

Theorem 4 (Unique Information for Continuous $I_{\cap}^{\min}$ and $I_{\cap}^{\mathrm{PM}}$ PIDs)**.** *Let $(X, Y) \to T$ be a noise-free bivariate interaction with associated kernel g, and let us further assume that there exists an open set $U \subset \mathcal{A}_{X,Y}$ of measure $\mu_{X,Y}(U) = 1$ such that g satisfies the following conditions:*

i. The function g is continuously differentiable on U, i.e. $g \in C^{(1)}(U)$;

ii. The partials $|\partial_y g(x, y)| > 0$, $|\partial_x g(x, y)| > 0$ for $(x, y) \in U$.

Then the unique informations for the decompositions $\mathfrak{d}_{min}$ and $\mathfrak{d}_{PM}$ are given by

$$U_X^{min}(T; X, Y) = \mathbb{E}\left[\mathbb{1}_A \left(\log \frac{p_Y(Y)}{p_X(X)} - \log \frac{|\partial_y g|}{|\partial_x g|}\right)\right] \tag{3.8.1}$$

$$U_X^{PM}(T; X, Y) = \underbrace{\mathbb{E}\left[\mathbb{1}_B \log \frac{p_Y(Y)}{p_X(X)}\right]}_{(U_X^{PM})^+} - \underbrace{\mathbb{E}\left[\mathbb{1}_C \log \frac{|\partial_y g|}{|\partial_x g|}\right]}_{(U_X^{PM})^-} \tag{3.8.2}$$

where $\mathbb{1}$ is the indicator function for the events:

$$A = \{I_X(T) \geq I_Y(T)\}$$

$$B = \{p_X(X) \leq p_Y(Y)\}$$

$$C = \{|\partial_x g|(X, Y) \leq |\partial_y g|(X, Y)\}$$

In particular, if $(X, Y)_\rho \to T$ is a normalized Gaussian interaction, then

we have that

$$U_X^{min}(T;X,Y) = \mathbb{E}\left[\mathbb{1}_A\left(\frac{\log(e)}{2}(X^2-Y^2) - \log\frac{|\partial_y g|}{|\partial_x g|}\right)\right], \tag{3.8.3}$$

$$U_X^{PM}(T;X,Y) = \frac{\log(e)}{2}\sqrt{1-\rho^2} - \mathbb{E}\left[\mathbb{1}_C \log\frac{|\partial_y g|}{|\partial_x g|}\right]. \tag{3.8.4}$$

Before we prove this result, we state the following general change of variables formula.

Proposition 9. *Let $(X,Y) \to T$ be a noise-free bivariate interaction under the same conditions as in Theorem 4. Then the densities on the induced probability spaces $(\mathcal{A}_{X,T}, \mu_{X,T})$ and $(\mathcal{A}_{Y,T}, \mu_{Y,T})$ are defined a.e. (where we write $\partial_y g$ for $\partial_y g(x,\tilde{y})$ and similarly for $\partial_x g$)*

$$p_{X,T}(x,t) = \frac{1}{|\partial_y g|} p_{X,Y}(x,\tilde{y}) \ \ a.e.\ [\mu_{X,T}], \tag{3.8.5}$$

$$p_{Y,T}(y,t) = \frac{1}{|\partial_x g|} p_{X,Y}(\tilde{x},y) \ \ a.e.\ [\mu_{Y,T}], \tag{3.8.6}$$

where

$$\tilde{y}(x,t) = (g(x,\cdot))^{-1}(t), \tag{3.8.7}$$

$$\tilde{x}(y,t) = (g(\cdot,y))^{-1}(t). \tag{3.8.8}$$

Moreover, we may regard these densities as random variables that are almost surely equal (where we write $\partial_y g$ for $\partial_y g(X,Y)$ and similarly for $\partial_x g$):

$$p_{X,T}(X,T) \overset{\omega}{=} \frac{1}{|\partial_y g|} p_{X,Y}(X,Y) \tag{3.8.9}$$

$$p_{Y,T}(Y,T) \overset{\omega}{=} \frac{1}{|\partial_x g|} p_{X,Y}(X,Y) \tag{3.8.10}$$

Proof. Let $U \subset \mathcal{A}_{X,Y}$ be the full measure open set as in the statement of

Theorem 4. Consider the transformation $\Phi : U \to \mathbb{R}^2$ given by $\Phi : (x, y) \mapsto (x, g(x, y))$. Since $g \in C^1(U)$, the Jacobi matrix J_Φ is defined, and we have that $|\det J_\Phi| = |\partial_y g| > 0$ on U. By the Inverse Function Theorem, it follows that Φ is a diffeomorphism on $U \xrightarrow{\cong} \Phi(U)$. We arrive at our density $p_{X,T}$ by a standard change of variables (see, for instance, Section 4.3 of Casella and Berger [21]).

□

We now proceed to prove our main result.

Proof of Theorem 4. Let g and U be given as in the statement of Theorem 4. On the almost sure event $E = \pi_{X,Y}^{-1}(U)$, note that via Proposition 9, the densities $p_{X,Y}, p_{X,T}$, and $p_{Y,T}$ are well-defined. Moreover, since we can choose U to ensure $p_T(t) > 0$ over $(x, y) \in U$, the conditional densities $p_{X|T}$ and $p_{X|T}$ are likewise defined.

We begin with (3.8.1). We recall from Def. 14 that $I_\cap^{\min}(T; X, Y) = \mathbb{E} \min(I_X(T), I_Y(T))$, so

$$\begin{aligned} U_X^{\min} &= I(T; X) - I_\cap^{\min}(T; X, Y) \\ &= \mathbb{E} I_X(t) - \mathbb{E} \min(I_X(T), I_Y(T)) \\ &= \mathbb{E} \left[\mathbb{1}_A (I_X(T) - I_Y(T)) \right] \end{aligned}$$

where A is the event in which $I_X(t) - I_Y(t) > 0$. If $p_T(t) = 0$, $I_X(t) = I_Y(t) = 0$,

so we may assume $p_T(t) > 0$. Then observe that

$$(I_X(T) - I_Y(T)) = \mathbb{E}\left(\log \frac{p_{X,T}(X,T)}{p_X(X)p_T(T)} \,\middle|\, T\right) - \mathbb{E}\left(\log \frac{p_{Y,T}(Y,T)}{p_Y(Y)p_T(T)} \,\middle|\, T\right) \tag{3.8.11}$$

$$= \mathbb{E}\left(\log \frac{p_{X,T}(X,T)}{p_X(X)} - \log \frac{p_{Y,T}(Y,T)}{p_Y(Y)} \,\middle|\, T\right) \tag{3.8.12}$$

$$= \mathbb{E}\left(\log \frac{p_Y(Y)}{p_X(X)} - \log \frac{p_{Y,T}(Y,T)}{p_{X,T}(X,T)} \,\middle|\, T\right) \tag{3.8.13}$$

By Proposition 9, we have that

$$\log \frac{p_Y(Y)}{p_X(X)} - \log \frac{p_{Y,T}(Y,T)}{p_{X,T}(X,T)} \overset{\omega}{=} \log \frac{p_Y(Y)}{p_X(X)} - \log \frac{|\partial_y g|}{|\partial_x g|}$$

Thus, considered as random functions of T,

$$I_X(T) - I_Y(T) \overset{\omega}{=} \mathbb{E}\left(\log \frac{p_Y(Y)}{p_X(X)} - \log \frac{|\partial_y g|}{|\partial_x g|} \,\middle|\, T\right). \tag{3.8.14}$$

Now, $\mathbb{1}_A$ is a function of T so that

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{1}_A \mathbb{E}\left(\log \frac{p_Y(Y)}{p_X(X)} - \log \frac{|\partial_y g|}{|\partial_x g|} \,|\, T\right)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left(\mathbb{1}_A \left(\log \frac{p_Y(Y)}{p_X(X)} - \log \frac{|\partial_y g|}{|\partial_x g|} \,|\, T\right)\right)\right] \\
&= \mathbb{E}\left[\mathbb{1}_A \log \frac{p_Y(Y)}{p_X(X)} - \log \frac{|\partial_y g|}{|\partial_x g|}\right]
\end{aligned}$$

and Eq. 3.8.1 follows.

The expression for unique specificity $U_X^{\mathrm{PM},+}$ in Eq. (3.8.1) follows easily from Def. 15 and Eq. (3.6.26f).

To find the expression for unique ambiguity in Eq. 3.8.2, we have that

$$\begin{aligned} U_X^{\mathrm{PM},-} &= I_\cap^{\mathrm{PM},-}(T;X) - I_\cap^{\mathrm{PM},-}(T;X,Y) \\ &= \mathbb{E}\log\frac{1}{p_X(X)} - \mathbb{E}\min(\frac{1}{p_{X|T}(X|T)}, \frac{1}{p_{Y|T}(Y|T)}) \\ &= \mathbb{E}\left[\mathbb{1}_C \log\frac{p_{Y|T}(Y|T)}{p_{X|T}(X|T)}\right] \end{aligned}$$

where $C \subset E$ is the event

$$C = \{p_{X|T}(X|T) > p_{Y|T}(Y|T)\} \cap E.$$

Note that C may be empty or probability zero, in which case $U_X^{\mathrm{PM},-} = 0$. Prop. 9 gives us the almost sure equality of the random variables

$$\log\frac{p_{X|T}(X|T)}{p_{Y|T}(Y|T)} \overset{\omega}{=} \log\frac{|\partial_x g(X,Y)|}{|\partial_y g(X,Y)|}. \tag{3.8.15}$$

Thus, our expression in Eq. (3.8.2) follows, as does our definition of the event C stated in the theorem.

For the specific case where $(X,Y)_\rho \to T$, then we have that both marginals are the standard normal Gaussian density $p_X(z) = p_Y(z) = \frac{1}{\sqrt{2\pi}}e^{-z^2/2}$. Thus,

$$\log\frac{p_Y(y)}{p_X(x)} = \frac{\log(e)}{2}(x^2 - y^2)$$

and Eq. (3.8.3) follows. The expression for $U_X^{\mathrm{PM},+}$ was demonstrated in Lemma 2 of Section 3.7.1.

□

As we mentioned earlier in this section, Theorem 4 allows us to understand

how the $I_\cap^{\min}$ PID comes to assign negative unique information to a particular predictor X, versus a more informative predictor Y. We can see, explicitly in Eqs. (3.8.1-3.8.4), how the amount of unique information assigned to a predictor X shrinks as the the relative sensitivity of the kernel function to X, quantified in $|\partial_x g|/|\partial_y g|$, similarly shrinks. It is easy to see how $U_X^{\min}$ could be zero and U_X^{PM} negative when this ratio becomes very small.

We conclude this section with a corollary to Theorem 4 that can be thought of as a generalization of our results in Section 3.7.2. In the context of a noiseless interaction $(X, Y) \to T$, it provides a framework in which we may say that, as X approaches conditional independence of T given Y, $U_X^{\min} \to 0$ while $U_X^{\mathrm{PM}} \to -\infty$. Whereas, in our linear interaction, we formalized this as the limit as $a/b \to 0^+$, here we instead have the more general situation where $|\partial_x g|/|\partial_y g| \to 0$.

Corollary 3. *Let $\{X, Y \to_\rho T_{(n)}\}_{n\in\mathbb{N}}$ be a sequence of noiseless Gaussian interactions, associated to the sequence of kernels $\{g_n\}$. Let us further assume that each g_n satisfies the conditions of Theorem 4, on the same common, full measure open set $U \subset \mathcal{A}_{X,Y}$. Assume also that $U_X^{min}(T_{(1)}; X, Y)$ and $U_X^{PM}(T_{(1)}; X, Y)$ are finite.*

Then the following monotonic limit of random variables

$$\frac{|\partial_x g_{(n)}|}{|\partial_y g_{(n)}|}(X, Y) \downarrow 0 \text{ as } n \to \infty \text{ almost surely} \tag{3.8.16}$$

implies the following information limits:

$$U_X^{min}(T_{(n)}; X, Y) \to 0, \tag{3.8.17}$$

$$U_X^{PM,-}(T_{(n)}; X, Y) \to \infty, \tag{3.8.18}$$

$$U_X^{PM}(T_{(n)}; X, Y) \to -\infty. \tag{3.8.19}$$

Proof. Consider the events

$$E = \left\{ \omega \text{ s.t. } \frac{\partial_x g_n}{\partial_y g_n}(X,Y)(\omega) \downarrow 0 \right\}$$
$$\tilde{E} = E \cap U$$

and note that $\tilde{E}$ has full measure. Note that for $\omega \in \tilde{E}$, there exists an $n(\omega)$ such that for $n > n(\omega)$,

$$\frac{\log(e)}{2}(X(\omega)^2 - Y(\omega)^2) - \log \frac{|\partial_y g_n|}{|\partial_x g_n|}(X,Y)(\omega) < 0$$

and hence $\mathbb{1}_{A_n}(\omega) = 0$. Hence (as the point-wise convergence in E is monotone), for $\omega \in \tilde{E}$, $\mathbb{1}_{A_n}(\omega)\left(\frac{\log(e)}{2}(X(\omega)^2 - Y(\omega)^2) - \log \frac{|\partial_y g_n|}{|\partial_x g_n|}(X,Y)(\omega)\right) \downarrow 0$. As we have integrability when $n = 1$ and the $n = 1$ case upper bounds the other cases pointwise on $\tilde{E}$, the Dominated Convergence Theorem yields the desired convergence to 0 of $U_X^{\min}(T_{(n)}; X, Y)$.

We will now demonstrate Eqs. (3.8.18) & (3.8.19). From Theorem 4, we have

$$U_X^{\mathrm{PM}}(T; X, Y) = \frac{\log(e)}{2}\sqrt{1-\rho^2} - \underbrace{\mathbb{E}\left[\mathbb{1}_C \log \frac{|\partial_y g|}{|\partial_x g|}\right]}_{U_X^{\mathrm{PM},-}} \tag{3.8.4}$$

Consider again the limit in Eq. (3.8.16), which defines our event E. This limit implies, for every $\omega \in \tilde{E} \cap C$:

$$\left[\mathbb{1}_C \log \frac{|\partial_y g_{(n)}|}{|\partial_x g_{(n)}|}\right](\omega) \nearrow \infty.$$

Since this random variable is non-negative (by the indicator), it follows from the Monotone Convergence Theorem that $U_X^{\mathrm{PM},-} \to \infty$, i.e. Eq. (3.8.18). Hence, Eq. (3.8.19) as well. □

It is straight-forward to demonstrate that Eqs (3.7.27b, 3.7.28b) are immediate applications of this corollary. A more interesting example is the sigmoidal switch kernel (E4) used for simulations in Sec. 3.3. We will approach this kernel presently, in Section 3.9.1. For the moment, we turn to a more thorough discussion of Finn and Lizier's work in [45] and [46] to comprehend our result within the narrative context of the $I_\cap^{\mathrm{PM}}$ redundancy measure.

3.8.2 The Curse of Ambiguity: Vanishing Partials, Flat Conditionals, and Negative Information

In this section, we aim to incorporate our current effort into the narrative development of the $I_\cap^{\mathrm{PM}}$ PID in [45] and [46]. Although the $I_\cap^{\mathrm{PM}}$ has not demonstrated itself to be a desirable tool for edge nomination in network inference, we nonetheless affirm that it quantifies exactly the probability mass exclusions that it was meant to, and obeys its own logic in this respect. Although our current application may not be the appropriate one for such a perspective, perhaps a continuous elaboration of the principles of the $I_\cap^{\mathrm{PM}}$ PID will aid in understanding the meaning of the specificity and ambiguity lattices defined in [45].

In section 3.8.2.1, we review the notion of informative and misinformative probability mass exclusions from [46], and how they motivated the construction of the $I_\cap^{\mathrm{PM}}$ PID in [45]. In section 3.8.2.2, we relate these concepts more firmly to our results in Theorem 4, and what we have uncovered, more generally, about the behavior of the $I_\cap^{\mathrm{PM}}$ PID for continuous interactions.

3.8.2.1 Probability Mass Exclusions and the Axiomatic Development of the $I_\cap^{\mathrm{PM}}$ PID

The $I_\cap^{\mathrm{PM}}$ PID, like many others, was developed with the intention of correcting the limited perspective inherent in the $I_\cap^{\mathrm{min}}$ PID. In order to overcome the issue of quantifying the same information versus the same amount of information regarding a target outcome $T = t$, Finn and Lizier examined the pointwise value of information in [45] and [46]. As observed in [46], information theory is built upon the understanding that a *gain in information* is exactly a *reduction in uncertainty*. With this in mind, they closely examined how mutual information accounts for exclusions of probability mass — that is, excludes sources of uncertainty — at every pointwise realization of elementary events.

We give an abbreviated summary of the investigation conducted in [46], without regard for formal rigor. We direct those seeking more detail to the original work. Adopting notation similar to that in [46] and [45], consider a discrete system with a predictor random variable X and target random variable T. Let us fix realized, elementary events x and t. We denote the complementary events as $x^{\complement}$ and $t^{\complement}$. Observe that, by taking intersections, we may divide our probability space into four probability masses:

$$\underbrace{p(t,x) + p(t,x^{\complement})}_{p(t)} + \underbrace{p(t^{\complement},x) + p(t^{\complement},x^{\complement})}_{p(t^{\complement})} = 1. \tag{3.8.20}$$

Pointwise mutual information, denoted i, is given by

$$i(t;x) = \log \frac{p(t,x)}{p(x)p(t)} = \log \frac{p(t|x)}{p(t)} \tag{3.8.21}$$

and we may in turn decompose $p(t|x)$ into the ratio of probability masses:

$$p(t|x) = \frac{p(t) - p(t, x^{\complement})}{1 - p(t, x^{\complement}) - p(t^{\complement}, x^{\complement})}. \tag{3.8.22}$$

Taken together, pointwise mutual information may then be rewritten as:

$$i(t; x) = \log \frac{1 - p(t, x^{\complement})/p(t)}{1 - p(t, x^{\complement}) - p(t^{\complement}, x^{\complement})}. \tag{3.8.23}$$

The conditioned probability mass $p(t|x)$ accounts for two probability mass exclusions that determine the information that the event $X = x$ provides about $T = t$. First, we see that for fixed $p(t)$, pointwise information $i(t; x)$ is a strictly increasing function of the mass $p(t^{\complement}, x^{\complement})$. Finn and Lizier refer to this as an **informative probability mass exclusion**, as it quantifies how much of the probability mass of the complementary target event $T \neq t$ is excluded by conditioning on $X = x$. On the other hand, for fixed $p(t)$, we have that $i(t; x)$ is a strictly decreasing function of $p(x^{\complement}, t)$. They refer to this as a **misinformative probability mass exclusion**. The understanding provided is that, when conditioning on $X = x$, one loses probability mass associated with the target event.

Finn and Lizier observed that pointwise MI $i(t; x)$ is signed, and can be negative when the misinformative exclusion $p(x^{\complement}, t)$ is relatively large and the information exclusion $p(x^{\complement}, t^{\complement})$ relatively small. Although MI, in expectation, is non-negative, the pointwise signed nature reveals that, at every point x, t, information and misinformation are being conflated. Thus, they proposed that $i(t; x)$ be decomposed into the difference of two non-negative components:

$$i(t; x) = i^{+}(t; x) - i^{-}(t; x) \tag{3.8.24}$$

Finn and Lizier provided three axioms (termed 'postulates' in [46]) to characterize this pointwise decomposition:

FL Axiom 1 **(Decomposition)** *The information provided by x about t can be decomposed into two non-negative components, such that $i(t;x) = i^+(t;x) - i^-(t;x)$.*

FL Axiom 2 **(Monotonicity)** *The functions $i^+(t;x)$ and $i^-(t;x)$ should satisfy the following conditions:*

a. For all fixed $p(t,x)$ and $p(t,x^{\complement})$, the function $i^+(t;x)$ is a continuous, increasing function of $p(t^{\complement},x^{\complement})$.

b. For all fixed $p(t^{\complement},x)$ and $p(t^{\complement},x^{\complement})$, the function $i^-(t;x)$ is a continuous, increasing function of $p(t,x^{\complement})$.

c. For all fixed $p(t,x)$ and $p(t^{\complement},x)$, the functions $i^+(t;x)$ and $i^-(t;x)$ are increasing and decreasing functions of $p(t^{\complement},x^{\complement})$, respectively.

FL Axiom 3 **Self-Information**. *An event cannot misinform about itself, hence $i^+(x;x) = i(x;x) = \log 1/p(x)$.*

FL Axiom 4 **Chain Rule**. *The functions i^+ and i^- satisfy a chain rule; i.e.,*

$$\begin{aligned} i_{\pm}(t;x,y) &= i_{\pm}(t;x) + i_{\pm}(t;y|x) \\ &= i_{\pm}(t;y) + i_{\pm}(t;x|y) \end{aligned}$$

where the conditional notation denotes the same function only with conditional probability as an argument.

Axioms 1, 3, & 4 provide a framework for the desired decomposition. Axiom 2 asserts the desired relationship of positive and negative information components to informative and misinformative probability mass exclusions. The positive component ought to increase with informative exclusions, and the negative component with misinformative exclusions. Finn and Lizier demonstrated that, for discrete variables, these axioms are satisfied for unique functions i^+ and i^-.

Theorem 5 (Finn & Lizier, [46])**.** *The unique functions satisfying Axioms 1-3 are the pointwise surprisals:*

$$i^+(t;x) = \log \frac{1}{p(x)} \tag{3.8.25}$$

$$i^-(t;x) = \log \frac{1}{p(x|t)} \tag{3.8.26}$$

The refer to the positive information component i^+ as specificity, and the negative component i^- as ambiguity. In expectation, these become $H(X)$ and $H(X|T)$, respectively. We may rewrite them in terms of the probability masses from Eq. (3.8.20):

$$i^+(t;x) = -\log\left(1 - p(t, x^{\complement}) - p(t^{\complement}, x^{\complement})\right) \tag{3.8.27}$$

$$i^-(t;x) = -\log\left(1 - \frac{p(t, x^{\complement})}{p(t)}\right) \tag{3.8.28}$$

In Fig. 3.15, we provide a visualization of the probability mass exclusions and their relationship to pointwise mutual information.

In [45], Finn and Lizier use this pointwise decomposition to develop the discrete $I_{\cap}^{\mathrm{PM}}$ PID, as we defined back in Section 3.6. This PID decomposes the specificity and ambiguity of predictor sources separately, corresponding to the informative and misinformative probability mass exclusions, respectively

(Eqs. 3.8.25-3.8.26, Def. 15). What we have seen repeatedly throughout this work is that the $I_\cap^{\mathrm{PM}}$ PID frequently assigns negative unique information to less informative predictors.

Pointwise, unique information in the $I_\cap^{\mathrm{PM}}$ PID, for discrete variables, takes the form:

$$u_X^-(t;x,y) = i^-(t;x) - \min(i^-(t;x), i^-(t;y)) \tag{3.8.29}$$

$$= \begin{cases} i^-(t;x) - i^-(t;y) & \text{when } p(t,x) \leq p(t,y) \\ 0 & \text{when } p(t,x) \geq p(t,y) \end{cases} \tag{3.8.30}$$

Assuming that $p(t,y) > p(t,x)$, we use Eq. (3.8.28) to rewrite Eq. (3.8.30), and have that

$$u_X^-(t;x,y) = \log \frac{1 - p(t, y^{\complement})}{1 - p(t, x^{\complement})} \tag{3.8.31}$$

We see that the pointwise unique ambiguity of x, then, grows monotonically with the misinformation that x contributes to the target event t, while it shrinks as the misinformation contributed by y. More precisely, it grows with the misinformative probability mass exclusion $p(t, x^{\complement})$, and shrinks as the exclusion $p(t, y^{\complement})$ grows. Moreover, the condition for which $u_X^- > 0$ is equivalently stated as a comparison of the sizes of these misinformative exclusions:

$$u_X^-(t;x,y) > 0 \iff \frac{p(t,y)}{p(t,x)} > 1 \iff \frac{p(t,x^{\complement})}{p(t,y^{\complement})} > 1. \tag{3.8.32}$$

This concludes our review of the probability mass exclusions that motivated the development of the $I_\cap^{\mathrm{PM}}$ PID in [46] and [45]. We will now relate these ideas to our analytic characterization of unique information in Theorem 4.

3.8.2.2 Unique Information and the Continuous Analogue of Misinformative Probability Mass Exclusions

Consider the case of a continuous interaction $(X, Y) \to_\rho T$ under the conditions of Theorem 4. Let (x_0, y_0) be a point in the open set U, and let $\underline{\mathrm{x}}_\delta, \underline{\mathrm{y}}_\delta$ and $\underline{\mathrm{t}}_\epsilon$ be radius δ and ϵ intervals about x_0, y_0, and $t_0 = g(x, y)$ such that $\underline{\mathrm{x}}_\delta \times \underline{\mathrm{y}}_\delta \subset U$ and $g(\underline{\mathrm{x}}_\delta \times \underline{\mathrm{y}}_\delta) \subset \underline{\mathrm{t}}_\epsilon$. Observe that[15]

$$\frac{1 - p(\underline{\mathrm{t}}_\epsilon, \underline{\mathrm{y}}_\delta^{\complement})}{1 - p(\underline{\mathrm{t}}_\epsilon, \underline{\mathrm{x}}_\delta^{\complement})} = \frac{p(\underline{\mathrm{t}}_\epsilon, \underline{\mathrm{y}}_\delta)}{p(\underline{\mathrm{t}}_\epsilon, \underline{\mathrm{x}}_\delta)} \approx \frac{\epsilon\delta p_{Y,T}(y_0, t_0)}{\epsilon\delta p_{X,T}(x_0, t_0)} = \frac{|\partial_y g|}{|\partial_x g|}(x_0, y_0) \tag{3.8.33}$$

Note that the right-most expression is the pointwise unique information for the continuous PID from Eq. (3.8.2) of Theorem 4. And thus, for this δ-ϵ discretization, we have that

$$\underbrace{u_X^-(\underline{\mathrm{t}}_\epsilon; \underline{\mathrm{x}}_\delta, \underline{\mathrm{y}}_\delta)}_{\text{Discretized } u_X^-} = \underbrace{\log \frac{1 - p(\underline{\mathrm{t}}_\epsilon, \underline{\mathrm{y}}_\delta^{\complement})}{1 - p(\underline{\mathrm{t}}_\epsilon, \underline{\mathrm{x}}_\delta^{\complement})}}_{\text{Function of Relative Misinformation}} \approx \underbrace{\frac{|\partial_y g|}{|\partial_x g|}(x_0, y_0)}_{\text{Continuous } u_X^-} . \tag{3.8.34}$$

Pointwise, we see that the probabilistic discrepancy between the misinformative exclusions — favoring the mass $p(\underline{\mathrm{t}}_\epsilon, \underline{\mathrm{x}}_\delta^{\complement})$ over $p(\underline{\mathrm{t}}_\epsilon, \underline{\mathrm{y}}_\delta^{\complement})$ — corresponds to an analytic discrepancy between the (absolute) partial derivatives of the kernel — favoring $|\partial_y g|$ over $|\partial_x g|$.

This is meaningful in the following sense. If $|\partial_y g| >> |\partial_x g|$, then that means that the kernel g, and thus the response T, is much more sensitive to a perturbation in Y rather than X. Consider the probability mass of our target neighborhood $p(\underline{\mathrm{t}}_\epsilon)$, and split it into the realized event and misinformative

[15] We might have used a discretization approach to define or investigate the continuous $I_\cap^{\min}$ and $I_\cap^{\mathrm{PM}}$ PIDs. This can be done by extending standard methods of demonstrating convergence of discrete and continuous MI and ($\log \epsilon$-corrected) entropy, see Ch. 7 in [33]. We leave this approach to future work.

exclusion of each predictor:

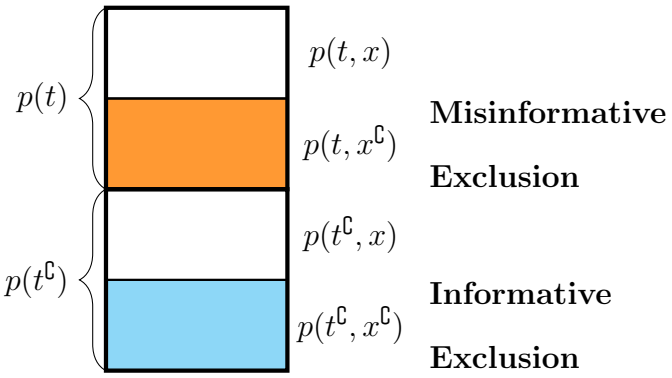
$$p(\underline{\mathrm{t}}_\epsilon) = p(\underline{\mathrm{t}}_\epsilon, \underline{\mathrm{x}}_\delta) + p(\underline{\mathrm{t}}_\epsilon, \underline{\mathrm{x}}_\delta^{\complement}) \tag{3.8.35}$$

$$= p(\underline{\mathrm{t}}_\epsilon, \underline{\mathrm{y}}_\delta) + p(\underline{\mathrm{t}}_\epsilon, \underline{\mathrm{y}}_\delta^{\complement}) \tag{3.8.36}$$

If T is more sensitive to changes in Y than in X, we expect that effective (probable) range of values $T = t$ to diverge from t_0 more quickly with growing perturbations of Y. Assuming nice global properties, significant probability mass will then be concentrated in $p(\underline{\mathrm{t}}_\epsilon, \underline{\mathrm{y}}_\delta)$ relative to $p(\underline{\mathrm{t}}_\epsilon, \underline{\mathrm{y}}_\delta^{\complement})$, since for $y, |y - y_0| > \delta$, it is relatively more likely that $|t - t_0| > \epsilon$ as well. By contrast, since $|\partial_x g|$ is relatively small, then X will have a much smaller impact on the probable value of T. Thus, there is a greater likelihood of $t \in \underline{\mathrm{t}}_\epsilon$ and $x \in \underline{\mathrm{x}}_\delta^{\complement}$, i.e. more of the mass in $p(\underline{\mathrm{t}}_\epsilon)$ will be spread out into the misinformative exclusion.

To summarize, the take-away of this discursion is the following. In the context in which T is determined by X and Y via an interaction kernel g, a relatively small sensitivity of T upon X, quantified analytically as $|\partial_x g|$, corresponds to a larger *misinformative* probability mass exclusion. In the development of the $I_\cap^{\mathrm{PM}}$ PID, this was taken to be the negative component in pointwise mutual information. In this way, we see why the $I_\cap^{\mathrm{PM}}$ PID will assign negative unique information to predictors with little-to-no impact upon the target T. As a measure of the balance between informative and misinformative probability mass exclusions, $I_\cap^{\mathrm{PM}}$ behaves as we would expect.

Probability Mass Exclusions



Mutual Information

$$I(T;X) = H(X) - H(X|T) \qquad \textbf{Entropic Decomposition of MI}$$

$$i(t;x) = i_+(t;x) - i_-(t;x) \qquad \textbf{Pointwise Decomposition}$$

$$i_+(t;x) = -\log\left(1 - p(t, x^{\complement}) - p(t^{\complement}, x^{\complement})\right) \qquad \textbf{Specific Information}$$

$$i_-(t;x) = -\log\left(1 - p(t, x^{\complement})/p(t)\right) \qquad \textbf{Ambiguous Information}$$

Pointwise Mutual (PM) PID

$$I_\cap^{\mathrm{PM}} = I_\cap^{\mathrm{PM},+} - I_\cap^{\mathrm{PM},-} \qquad \textbf{PM Redundancy}$$

$$I_\cap^{\mathrm{PM},+}(T;\{\boldsymbol{X}_k\}) = \mathbb{E}\min_k i_+(t;\boldsymbol{x}_k) \qquad \textbf{Redundant Specificity}$$

$$I_\cap^{\mathrm{PM},-}(T;\{\boldsymbol{X}_k\}) = \mathbb{E}\min_k i_-(t;\boldsymbol{x}_k) \qquad \textbf{Redundant Ambiguity}$$

Figure 3.15: **Probability mass exclusions and the $I_\cap^{\mathrm{PM}}$ PID** We present a probability mass diagram similar to those in [46] and [45], presenting the decomposition of the sample space into four masses as in Eq. (3.8.20). In the PM PID, pointwise mutual information $i(t;x)$ between a target event t and source event x is decomposed into the difference of two non-negative components, the positive component i_+ and the negative i_-, referred to as specificity and ambiguity in [45]. Specificity grows with both the informative and misinformative probability mass exclusions, $p(t^{\complement}, x^{\complement})$ and $p(t, x^{\complement})$ (signified in cyan and orange, respectively, in place of the cross-hatching used in [46]).

3.9 Continuous Unique Information in the Sigmoidal Switch Interaction

We will conclude this work with an analytic investigation into noise-free bivariate interactions with a sigmoidal switch interaction kernel, as in Eq. (E4). This was the kernel that we used in our motivating simulations (Eq. (E4)) back in Sec. 3.3. This section will demonstrate an application of the generic approach developed in Theorem 4.

First, we will apply the asymptotic results from the previous section to complete our analysis of Experiment III in Sec. 3.3.3. Recall that in Experiment III , we conducted network simulations in which we allowed the parameter α to vary, affecting the kernel sensitivity upon x and y and thus altering the balance of information between X and Y. Via Corollary 3, we demonstrate that $U_X^{\mathrm{PM}} \to -\infty$ and $U_X^{\min} \to 0$ as $\alpha \to -\infty$.

Then, we will apply Theorem 4 to directly investigate the unique information of the switch gene X for a noise-free bivariate interaction with the kernel (E4). Specifically, we present an upper bound $\beta(\rho, \alpha)$ on this quantity U_X^{PM}. This is sufficient to demonstrate that, for the continuous PID, $U_X^{\mathrm{PM}} < 0$. This agrees with our simulation experiments in Sec. 3.3, and demonstrates that the generic framework we have developed for the continuous $I_\cap^{\mathrm{PM}}$ PID may shed light on its behavior when applied to discretized simulations of continuous phenomenon.

3.9.1 Limiting Behavior of U_X^{PM} and $U_X^{\min}$

In Experiment III in Sec. 3.3.3, we conducted a series of network simulations with the sigmoidal kernel, in which we allowed the switching parameter α to vary. We then examined the effect this had upon the PID of interacting

pairs of genes. Altering α has the effect of re-centering the predictor X about $X = \alpha$ in the interaction. This has the effect of shifting the relative sensitivity of $g(X, Y)$ upon each argument X and Y near their mean $\mu_X = \mu_Y = 0$, and thus in expectation as well. More precisely, we saw that the expansion of the sigmoidal kernel takes the form

$$g(x, y) \approx \underbrace{\frac{1}{1 + e^{\alpha}}}_{\partial_y g} y + \underbrace{\frac{e^{\alpha}}{(1 + e^{\alpha})^2}}_{\partial_{x,y} g} xy. \tag{3.3.4}$$

for $(x, y) \approx (0, 0)$, which we emphasize is the probable scenario. Thus, as $\alpha \to -\infty$, we have that $g(x, y) \approx y$ and $g(X, Y)$ approaches conditional independence of X given Y. We saw in Fig. 3.10 that U_X^{PM} was consistently negative for a range of α.[16] For a linear kernel, we saw in Sec. 3.7.2 that $U_X^{\mathrm{PM}} \to -\infty$ as X approaches conditional independence. A core contention of ours in this work is that, in such a scenario, we ought to see unique information $U_X \to 0$ as X becomes less conditionally dependent, as otherwise the bivariate PID inflate both redundancy and synergy, R and S. The $I_\cap^{\min}$ PID behaves in this way, but the $I_\cap^{\mathrm{PM}}$ PID does not.

Using Corollary 3, we may demonstrate that, similar to the linear interaction limits in Cor. 2, we have that $U_X^{\min} \to 0$ and $U_X^{\mathrm{PM}} \to -\infty$ as $\alpha \to -\infty$. For the sigmoidal interaction kernel g as in Eq. (E4), we have $g(x, y) \approx y$ as $\alpha \to -\infty$., and thus X is approaching conditional independence of $T = g(X, Y)$, given Y. Thus, the themes explored in Sec. 3.7.2, regarding how the $I_\cap^{\min}$ and $I_\cap^{\mathrm{PM}}$ PIDs treat conditionally independent, 'false' interactions, are also analytically apparent in the sigmoidal interaction kernel as well as in the linear interaction kernel. Moreover, since we are now examining the

[16] The minimal α in Fig. 3.10 is $\alpha = -4$, which corresponds in $\partial_y g$, for instance, to $\frac{1}{1+e^{\alpha - x}} \approx 0.982 + 0.018x$. We consider this sufficient for anticipating the trend as $\alpha \to -\infty$ in simulation.

sigmoidal kernel, the work in this section helps to explain the behavior we saw in our experiments in Sec. 3.3, particularly regarding the behavior of the $\hat{I}_\cap^{\min}$ and $\hat{I}_\cap^{\mathrm{PM}}$ PIDs in Experiment II toward 'false pairs' and in Experiment III as α becomes small.

Proposition 10. *Let $\{X, Y \to_\rho T_{(\alpha)}\}_{\alpha\in\mathbb{R}}$ be the family of noise-free Gaussian interactions with kernel $g_{(\alpha)}$ from Eq. E4, parametrized by the real parameter α.*

Then we have the following limits:

$$U_X^{min}(\alpha) \to 0 \qquad \text{as } \alpha \to -\infty \tag{3.9.1}$$

$$U_X^{PM,-}(\alpha) \to -\infty \qquad \text{as } \alpha \to \infty \tag{3.9.2}$$

$$U_X^{PM}(\alpha) \to -\infty \qquad \text{as } \alpha \to -\infty \tag{3.9.3}$$

Proof. We may apply Corollary 3 to our continuously parametrized family of interactions via the sequential characterization of continuous limits. Observe that

$$\frac{|\partial_x g|}{|\partial_y g|} = \frac{|Y|e^{\alpha - X}}{1 + e^{\alpha - X}}$$

It is easy to see that for any $(x, y) = (X, Y)(\omega)$, $\alpha \leq 0$,

$$\frac{|\partial_x g|}{|\partial_y g|} \downarrow 0 \text{ as } \alpha \to -\infty.$$

We take $\alpha = 0$ as our starting point, so that all that remains to be shown is that $U_X^{\min}(0)$ and $U_X^{\mathrm{PM}}(0)$ are finite.

Consider the following series of expansions

$$\begin{aligned}\frac{1}{2}|x^2 - y^2| + \left|\ln \frac{|\partial_x g|}{|\partial_y g|}\right| &= \frac{1}{2}|x^2 - y^2| + \left|\ln|y| + (\alpha - x) - \log(1 + e^{\alpha - x})\right| \\ &\leq \frac{1}{2}|x^2 - y^2| + |\ln|y|| + \alpha + |x| + \overbrace{\log(1 + e^{\alpha - x})}^{\leq e^{\alpha - x}} \\ &\leq \frac{x^2}{2} + \frac{y^2}{2} + \alpha + |x| + |\ln|y|| + e^{\alpha - x}\end{aligned}$$

Set $\alpha = 0$, and consider the functions f_1 and f_2:

$$\begin{aligned}f_1(x, y) &:= \frac{1}{2}|x^2 - y^2| + \left|\ln \frac{|\partial_x g|}{|\partial_y g|}\right| \\ &\leq \frac{x^2}{2} + \frac{y^2}{2} + \alpha + |x| + |\ln|y|| + e^{\alpha - x} =: f_2(x, y)\end{aligned}$$

All the terms in f_2 have finite expectation when integrated against the marginal

$$p_X(z) = p_Y(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

and so $\mathbb{E}f_1(X, Y) \leq \mathbb{E}f_2(X, Y) < \infty$.

Moreover, from the forms of $U_X^{\min}$ and U_X^{PM} in Theorem 4, we have that

$$\begin{aligned}|U_X^{\min}(0)| &\leq \mathbb{E}f_1(X, Y), \\ |U_X^{\mathrm{PM}}(0)| &\leq \mathbb{E}f_1(X, Y) + c,\end{aligned}$$

for a finite constant $c > 0$. Thus, $U_X^{\min}(0)$ and $U_X^{\mathrm{PM}}(0)$ are finite.

□

3.9.2 Upper Bound on Unique Information (PM PID) for the Switch Gene U_X^{PM} in a Bivariate Sigmoidal Interaction

We will demonstrate the following result:

Theorem 6 (Upper Bound for Unique Information of Switch Gene)**.** *Let* $(X, Y)_\rho \to T$ *be a noiseless bivariate interaction, where* $T = g(x, y)$ *for* g *defined in Eq E4. Assume* $\alpha < 2$*. Then the unique information of the switch gene* U_X^{PM} *is bounded above by the function* β

$$U_X^{PM}(T; X, Y) \leq \beta(\rho, \alpha) = \beta_\rho(\rho) + \frac{\alpha}{4} \tag{3.9.4a}$$

where

$$\begin{aligned}
\beta_\rho(\rho) &= c_1 + c_0\sqrt{1-\rho^2} - \frac{\sqrt{1-\rho}+\sqrt{1+\rho}}{4\sqrt{\pi}} - \frac{-\log(1-\rho^2)}{8} && \text{(3.9.4b)}\\
&\quad + \frac{\sqrt{1+\rho}\,(-\log(1-|\rho|))\,(\Phi(\rho))}{16\sqrt{\pi}(1-\rho)^{1/4}} \\
\Phi(\rho) &= \sqrt{\rho\sqrt{1-\rho}+2\phi(\rho)} && \text{(3.9.4c)}\\
\phi(\rho) &= \arcsin(\rho) + \arctan(\frac{\sqrt{1-\rho}}{\sqrt{1+\rho}}) && \text{(3.9.4d)}\\
c_0 &= \frac{16 - 3(2 * K + \pi * ln(2))}{16\pi} \approx 0.079 && \text{(3.9.4e)}\\
c_1 &= \frac{\ln 2 - \gamma_E}{8} \approx 0.0145 && \text{(3.9.4f)}\\
& && \text{(3.9.4g)}
\end{aligned}$$

The difficult part of this result lies in bounding $U_X^{\mathrm{PM},-}$, which we do in Sec. 3.9.3. Moreover, this is exactly where improvement upon the bound can be made. Once we are satisfied with a bound for unique ambiguity, the rest of the theorem follows readily.

Proof of Theorem 6. Under the conditions of our theorem, from Lemma 4, we have that

$$L(\rho, \alpha) \leq U_X^{\mathrm{PM},-}$$

where this function L is as in Eqs. (3.9.5a-h). Moreover, in Lemma 2 in

Section 3.7.1, we computed

$$U_X^{\mathrm{PM},+} = \frac{1}{\pi}\sqrt{1-\rho^2}.$$

Thus, by combining terms, we arrive at our upper bound:

$$\begin{aligned} U_X^{\mathrm{PM}} &= (U_X^{\mathrm{PM},+} - U_X^{\mathrm{PM},-}) \\ &\leq \frac{1}{\pi}\sqrt{1-\rho^2} - L(\rho,\alpha) \end{aligned}$$

By collecting terms, this bound takes the form of β in Eqs. (3.9.4). □

Our bound for U_X^{PM} is somewhat unwieldly. It is, nonetheless, decidedly negative for $|\rho| < 0.9$. We present a graph of this bound in Fig. 3.16. Notable, our bound is negative for all $\rho < 0.9$. Thus, we have the U_X^{PM} is negative when the predictors are at least mildly distinguished, as variables.

3.9.3 Unique Ambiguity $U_X^{\mathrm{PM},-}$ for the Switch Gene X

As mentioned in the previous section, the 'hard' part of computing (or bounding) unique information of a bivariate interaction under the $I_\cap^{\mathrm{PM}}$ PID lies in computing the ambiguity atom. To establish our upper bound $U_X^{\mathrm{PM}} \leq \beta(\rho,\alpha)$ in the previous section, we must establish a lower bound $L(\rho,\alpha) \leq U_X^{\mathrm{PM},-}$. We do so in the following lemma, and devote the rest of this subsection to the proof and its component computations.

Lemma 4 (Lower Bound for Unique Ambiguity of Switch Gene)**.** *Let* $(X,Y)_\rho \to T$ *be a noiseless bivariate interaction, where* $T = g(x,y,\alpha)$ *for* g *defined in Eq. E4. Assume* $\alpha < 2$*. Then the unique ambiguity of the switch*

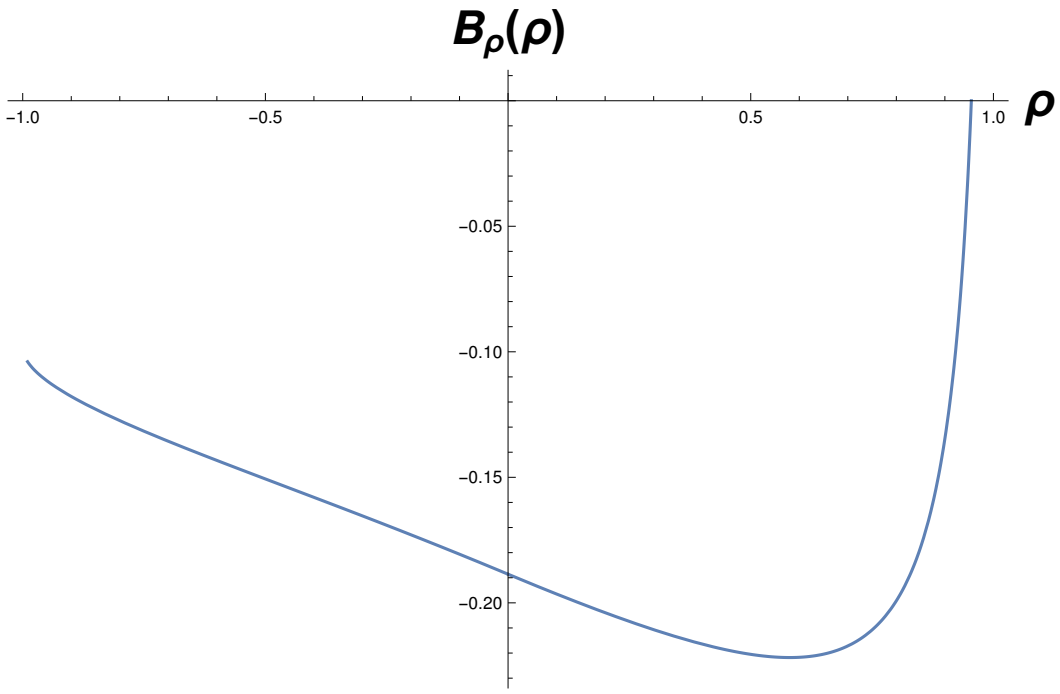


Figure 3.16: **Upper bound for unique information of the switch gene in a noise-free, sigmoidal switch interaction** Here, we plot the function $\beta_\rho(\rho)$ from Theorem 6, which bounds-above the unique information U_X^{PM} of the switch gene (in nats), under the assumption $\alpha \leq 0$. For $\alpha < 2$, we have the modified bound $U_X^{\mathrm{PM}} \leq \beta_\rho(\rho) + \frac{\alpha}{4}$.

gene $U_X^{PM,-}$ is bounded below

$$L(\rho, \alpha) \leq U_X^{PM,-}(T; X, Y) \tag{3.9.5a}$$

where

$$L(\rho,\alpha) = L_1(\rho) + L_2(\alpha) \tag{3.9.5b}$$

$$L_1(\rho) = \frac{\sqrt{1-\rho}+\sqrt{1+\rho}}{4\sqrt{\pi}} + \frac{-\log(1-\rho^2)}{8} - c_1 + \frac{3c_2\sqrt{1-\rho^2}}{16\pi} - \frac{\sqrt{1+\rho}\,(-\log(1-|\rho|))\,(\Phi(\rho))}{16\sqrt{\pi}(1-\rho)^{1/4}} \tag{3.9.5c}$$

$$L_2(\alpha) = -\frac{\alpha}{4} \tag{3.9.5d}$$

$$\Phi(\rho) = \sqrt{\rho\sqrt{1-\rho}+2\phi(\rho)} \tag{3.9.5e}$$

$$\phi(\rho) = \arcsin(\rho) + \arctan(\frac{\sqrt{1-\rho}}{\sqrt{1+\rho}}) \tag{3.9.5f}$$

$$c_1 = \frac{\ln 2 - \gamma_E}{8} \approx 0.0145 \tag{3.9.5g}$$

$$c_2 = 2K + \pi\log(2) \approx 4.0095 \tag{3.9.5h}$$

where γ is Euler's constant and K is the Catalan constant.

Proof of Lemma 4. From Eq. 3.8.2 in Theorem 4, we have that

$$U_X^{\mathrm{PM},-} = \mathbb{E}\left[\mathbb{1}_C\left(\log\frac{|\partial_y g|}{|\partial_x g|}\right)\right]. \tag{3.9.6}$$

For $\alpha < 2$, we have that

$$\tilde{C} := \{(x,y)|x>0, |y|<x\} \subset \pi_{X,Y}(C). \tag{3.9.7}$$

We demonstrate this claim in Computation (3) using elementary methods of little interest.

For the rest of our computations in this section, we no longer take care to distinguish events $E \subset \Omega$ from their image $\pi_{X,Y}(E) \in \mathcal{A}_{X,Y}$. Via Theorem 4, we have all the tools we need to work in the induced probability space $(\mathcal{A}_{X,Y}, \mu_{X,Y})$, and so no confusion need arise.

For our computations, we will typically divide $\tilde{C}$ into its components in the upper and lower half-plane of (X, Y)-space:

$$\tilde{C} = C_+ \sqcup C_-$$

$$C_+ = \{(x, y) \mid 0 < y \leq x\} \tag{3.9.8}$$

$$C_- = \{(x, y) \mid 0 < -y \leq x\} \tag{3.9.9}$$

Since $\log \frac{|\partial_y g|}{|\partial_x g|}$ is nonnegative on $C \supset \tilde{C}$, we have that

$$\begin{aligned} U_X^{\mathrm{PM},-} &= \mathbb{E}\left[\mathbb{1}_C \left(\log \frac{|\partial_y g|}{|\partial_x g|}\right)\right] \\ &\geq \mathbb{E}\left[\mathbb{1}_{\tilde{C}} \left(\log \frac{|\partial_y g|}{|\partial_x g|}\right)\right]. \end{aligned}$$

We expand $\log \frac{|\partial_y g|}{|\partial_x g|}$:

$$\begin{aligned} \log \frac{|\partial_y g|}{|\partial_x g|} &= \log \frac{1 + e^{\alpha - x}}{|y| e^{\alpha - x}} \\ &= \log(1 + e^{\alpha - x}) - \log(|y|) + (x - \alpha) \\ &\geq (x - \alpha) - \log(|y|) \end{aligned}$$

We have that

$$\begin{aligned} U_X^{\mathrm{PM},-} &\geq \mathbb{E}\mathbb{1}_{\tilde{C}} X - \mathbb{E}\mathbb{1}_{\tilde{C}} \log |Y| - \alpha \mu_{X,Y}(\tilde{C}) \\ &\geq \mathbb{E}\mathbb{1}_{\tilde{C}} X - \mathbb{E}\mathbb{1}_{\tilde{C}} \log |Y| - \frac{\alpha}{4} \end{aligned}$$

We compute $\mathbb{E}\mathbb{1}_{\tilde{C}} X$ explicitly in Computation 1 and find an upper bound for $\mathbb{E}\left[\mathbb{1}_{\tilde{C} \ln |Y|}\right]$ in Computation 2. Eqs (3.9.5a-h) follow readily by combining

terms. □

In our computations, we will exploit the following symmetry property, which will allow us to take expectations in the symmetric regions C_+ and C_- by merely flipping the sign on the correlation ρ between X and Y. Let $\mathbb{H}^+$ denote the open upper half-plane, and H^- the lower half-plane, i.e.

$$\mathbb{H}^+ = \{(x, y)|y > 0\}$$
$$\mathbb{H}^- = \{(x, y)|y < 0\}.$$

Proposition 11 (Symmetry Property)**.** *Let* $(X, Y) \sim N(\mathbf{0}, \Sigma)$ *where* $\Sigma = [1, \rho; \rho, 1]$. *By* $\mathbb{E}\mathbb{1}_A f(X, Y)(\rho)$, *we denote the expectation of a function of* X *and* Y *over region* $A \subset \mathbb{R}^2$ *given correlation* ρ. *Let* $A \subset \mathbb{H}^+$ *be an open set in the upper half-plane, and* $A^\dagger$ *denote its reflection into the lower half-plane, i.e.* $(x, y) \in A \iff (x, -y) \in A^\dagger$. *Let* $r : A \cup A^\dagger \to \mathbb{R}$ *be a function for which* $r(x, y) = r(x, -y)$ *for* $(x, \pm y) \in A \cup A^\dagger$. *Then we have that, for any* $\rho \in (-1, 1)$,

$$\mathbb{E}\mathbb{1}_{A^\dagger} r(X, Y)(\rho) = \mathbb{E}\mathbb{1}_A r(X, Y)(-\rho). \tag{3.9.10}$$

In particular, for V_+ *and* V_- *as defined as in Eqs. (3.9.8) & (3.9.9), we have that*

$$\mathbb{E}\mathbb{1}_{V_+} r(X, Y)(\rho) = \mathbb{E}\mathbb{1}_{V_-} r(X, Y)(-\rho). \tag{3.9.11}$$

Proof. For $(X, Y) \sim N(\mathbf{0}, \Sigma)$ as above, we have joint density (parametrized by ρ)

$$p_{X,Y}(x, y)(\rho) = \frac{1}{2\pi\sqrt{1-\rho^2}} e^{\frac{1}{2(1-\rho^2)}(x^2+y^2-2\rho xy)}.$$

Thus, we see that

$$p_{X,Y}(x, y)(-\rho) = p_{X,Y}(x, -y)(\rho). \tag{3.9.12}$$

Applying a change of variables $(x, y) \to (x, -y)$ to the integral

$$\int_{A^\dagger} \left[r(x,y) p_{x,y}(x,y) \right] (\rho) d(x,y),$$

our result follows. □

Computation 1 (Expectation of X on $\tilde{C}$)**.** *Let* $(X, Y) \sim N(\mathbf{0}, \Sigma)$ *where* $\Sigma = [1, \rho; \rho, 1]$*, and* $\tilde{C}$ *be as in Eq. (3.9.7). Then*

$$\mathbb{E}_{\tilde{C}} X = \frac{1}{4\sqrt{\pi}} (\sqrt{1-\rho} + \sqrt{2}\rho). \tag{3.9.13}$$

Proof. For convenience, let $a = (2\pi\sqrt{1-\rho^2})^{-1}$ and $b = (2(1-\rho^2))^{-1}$, so that we may represent the joint density of X and Y as

$$p_{X,Y}(x,y) = a e^{-b(x^2+y^2-2\rho xy)}. \tag{3.9.14}$$

We first compute our expectation on V_+:

$$\begin{aligned}
\mathbb{E}\mathbb{1}_{C_+} X &= a \int\int x e^{-b(x^2+y^2-2\rho xy)} \\
&= a \int_0^{\pi/4} \cos(\theta) \int_0^\infty r^2 e^{-br^2(1-\rho\sin(2\theta))} dr d\theta \\
&= a \int_0^{\pi/4} \sqrt{\pi} \left[\frac{\cos(\theta)}{4b^{3/2}(1-\rho\sin(2\theta))^{3/2}} \right] d\theta \\
&= \frac{a\sqrt{\pi}}{4b^{3/2}} \int_0^{\pi/4} \left[\frac{\cos(\theta)}{(1-\rho\sin(2\theta))^{3/2}} \right] d\theta \\
&= \frac{a\sqrt{\pi}}{8b^{3/2}} \left[\frac{\sqrt{2}\sqrt{1-\rho} + 2\rho}{1-\rho^2} \right] \\
&= \frac{\sqrt{2}}{8\sqrt{\pi}} (\sqrt{2}\sqrt{1-\rho} + 2\rho) \\
&= \frac{1}{4\sqrt{\pi}} (\sqrt{1-\rho} + \sqrt{2}\rho)
\end{aligned}$$

Using the symmetry property from Prop. 11,

$$\begin{aligned}\mathbb{E}\mathbb{1}_{C_-}X(\rho) &= \mathbb{E}\mathbb{1}_{V_+}X(-\rho)\\ &= \frac{1}{4\sqrt{\pi}}(\sqrt{1+\rho}-\sqrt{2}\rho)\end{aligned}$$

We conclude by taking the sum

$$\begin{aligned}\mathbb{E}\mathbb{1}_{\tilde{C}}X &= \left(\mathbb{E}\mathbb{1}_{C_-}X\right)+\left(\mathbb{E}\mathbb{1}_{C_+}X\right)\\ &= \frac{1}{4\sqrt{\pi}}(\sqrt{1-\rho}+\sqrt{1+\rho}).\end{aligned}$$

□

Computation 2 (Expectation of $\log(|Y|)$ on $\tilde{V}$)**.** *The expectation* $\mathbb{E}\left[\mathbb{1}_{\tilde{C}}\log(|Y|)\right]$*, for the region* $\tilde{C}$ *defined in (3.9.7), is bounded above by*

$$\begin{aligned}\mathbb{E}\mathbb{1}_{\tilde{C}}\log(|Y|) \leq{}& \frac{1}{8}\left(-(\gamma-\log(2))+\log(1-\rho^2)\right) \\ &+ \frac{\sqrt{1+\rho}\,(-\log(1-|\rho|))}{16\sqrt{\pi}(1-\rho)^{1/4}}\sqrt{\rho\sqrt{1-\rho}+2\phi(\rho)} \\ &+ \frac{-3\sqrt{1-\rho^2}}{16\pi}\left(2K+\pi\log(2)\right)\end{aligned} \tag{3.9.15}$$

$$\phi(\rho) = \arcsin(\rho)+\arctan\left(\frac{\sqrt{1-\rho}}{\sqrt{1+\rho}}\right) \tag{3.9.16}$$

Proof. By applying the symmetry property (Prop. 11)

$$\begin{aligned}\mathbb{E}\mathbb{1}_{\tilde{C}}\log(|Y|)(\pm|\rho|) &= \mathbb{E}\mathbb{1}_{C_+}\log(|Y|)(\pm|\rho|)+\mathbb{E}\mathbb{1}_{C_-}\log(|Y|)(\pm|\rho|)\\ &= \mathbb{E}\mathbb{1}_{C_+}\log(|Y|)(\pm|\rho|)+\mathbb{E}\mathbb{1}_{C_+}\log(|Y|)(\mp|\rho|)\end{aligned}$$

Thus, we will proceed accordingly. We fix $|\rho| \in (0,1)$. We consider the expectation on C_+, first with $\rho = |\rho| \in (0,1)$ and then with $\rho = -|\rho| \in (-1,0)$,

and find upper bounds:

$$\mathbb{E}\left[\mathbb{1}_{C_+}\log(Y)(|\rho|)\right] \leq I,$$
$$\mathbb{E}\left[\mathbb{1}_{C_+}\log(Y)(-|\rho|)\right] \leq \bar{I}.$$

By summing these two bounds, respectively, we have an upper bound on the full expectation on $\tilde{C}$ for $\rho = \pm|\rho|$.

We first break our expectation up into manageable integrals. As in our previous proof for Computation 1, we again use the shorthand $a = (2\pi\sqrt{1-\rho^2})^{-1}$ and $b = (2(1-\rho^2))^{-1}$.

$$\begin{aligned}
\mathbb{E}\left[\mathbb{1}_{C_+}\log(Y)\right] &= \int_0^\infty \int_0^x \log(y) p_{X,Y}(x,y)\,dy\,dx \\
&= a\int_0^\infty \int_0^x \log(y) e^{-b(x^2+y^2-2\rho xy)}\,dy\,dx \\
&= a\int_0^{\pi/4} \int_0^\infty r\log(r\sin\theta) e^{-br^2(1-\rho sin(2\theta))}\,dr\,d\theta \\
&= a\int_0^{\pi/4} \underbrace{\left[\int_0^\infty r\log(r) e^{-br^2(1-\rho sin(2\theta))}\,dr\right]}_{\zeta_1(\theta)} \\
&\quad + \log(\sin\theta)\underbrace{\left[\int_0^\infty r e^{-br^2(1-\rho sin(2\theta))}\,dr\right]}_{\zeta_2(\theta)}\,d\theta \qquad (\star)
\end{aligned}$$

We compute the inner integrals to arrive at $\zeta_1(\theta)$ and $\zeta_2(\theta)$:

$$\zeta_1(\theta) = -\frac{\gamma + \log(b) + \log(1-\rho\sin 2\theta))}{4b(1-\rho\sin 2\theta)},$$
$$\zeta_2(\theta) = \frac{1}{2b(1-\rho\sin 2\theta)}.$$

We substitute these formulas into ($\star$), and factor out $(4b)^{-1}$, and our expectation becomes

$$\mathbb{E}\left[\mathbb{1}_{C_+}\log(Y)\right] = \frac{\sqrt{1-\rho^2}}{4\pi}\int_0^{\pi/4} \underbrace{\left[-\frac{\gamma+\log(b)}{1-\rho\sin 2\theta}\right]}_{J_1} + \underbrace{\left[\frac{-\log(1-\rho\sin 2\theta)}{1-\rho\sin 2\theta}\right]}_{J_2} + \underbrace{\left[\frac{2\log(\sin\theta)}{1-\rho\sin 2\theta}\right]}_{J_2} d\theta$$

We have thus split up the integral I into three terms:

$$\begin{aligned} I &= J_1 + J_2 + J_3 \\ J_1 &= \frac{\sqrt{1-\rho^2}}{4\pi}\int_0^{\pi/4} -\frac{\gamma+\log(b)}{1-\rho\sin 2\theta}d\theta \\ J_2 &= \frac{\sqrt{1-\rho^2}}{4\pi}\int_0^{\pi/4} \frac{-\log(1-\rho\sin 2\theta)}{1-\rho\sin 2\theta}d\theta \\ J_3 &= \frac{\sqrt{1-\rho^2}}{4\pi}\int_0^{\pi/4} \frac{2\log(\sin\theta)}{1-\rho\sin 2\theta}d\theta \end{aligned}$$

We begin with the first term, J_1. Observe that

$$\begin{aligned} \int_0^{\pi/4}\left[\frac{1}{1-\rho\sin 2\theta}\right]d\theta &= \frac{\phi(\rho)}{\sqrt{1-\rho^2}} \\ \phi(\rho) &= \arcsin(\rho) + \arctan(\frac{\sqrt{1-\rho}}{\sqrt{1+\rho}}) \end{aligned}$$

Thus, our first term simplifies

$$J_1 = \frac{\sqrt{1-\rho^2}}{4\pi} \int_0^{\pi/4} \left[-\frac{\gamma + \log(b)}{1 - \rho \sin 2\theta} \right] d\theta$$
$$= \frac{\phi(\rho)}{4\pi} \left(-(\gamma - \log(2)) + \log(1-\rho^2) \right)$$

For $\bar{J}_1$, we have that

$$\bar{J}_1 = \frac{\phi(-\rho)}{4\pi} \left(-(\gamma - \log(2)) + \log(1-\rho^2) \right)$$

Since $\phi(\rho) + \phi(-\rho) = \pi/2$,

$$J_1 + \bar{J}_1 = \frac{1}{8} \left(-(\gamma - \log(2)) + \log(1-\rho^2) \right)$$

We will now examine the second term, J_2.

$$J_2 = \frac{\sqrt{1-\rho^2}}{4\pi} \left[\int_0^{\pi/4} \frac{-\log(1 - \rho \sin 2\theta))}{1 - \rho \sin 2\theta} d\theta \right]$$

We use Holder's Inequality:

$$\| \frac{-\log(1 - \rho \sin(2\theta))}{1 - \rho \sin(2\theta)} \|_1 \leq \left(\| \frac{1}{1 - \rho \sin(2\theta)} \|_2 \right) (\| -\log(1 - \rho \sin(2\theta)) \|_2)$$

We find our first L_2 norm by computing the integral via a computer algebra

system:

$$\left\|(1-\rho\sin(2\theta))^{-1}\right\|_2 = \left(\frac{\rho\sqrt{1-\rho^2}+2\phi(\rho)}{2\left(1-\rho^2\right)^{3/2}}\right)^{1/2}$$
$$= \frac{\sqrt{\rho\sqrt{1-\rho}+2\phi(\rho)}}{\sqrt{2}(1-\rho)^{3/4}}$$

We will now bound the second L_2 norm. We use the triangle inequality to bound it above by the following series:

$$\|-\log(1-\rho\sin 2\theta))\|_2 = \left\|\sum_{k=1}^{\infty}\frac{(\rho\sin 2\theta)^k}{k}\right\|_2$$
$$\leq \sum_{k=1}^{\infty}\frac{\left\|(\rho\sin 2\theta)^k\right\|_2}{k}.$$

We bound this term:

$$\left\|(\rho\sin(2\theta))^k\right\|_2 = \sqrt{\int_0^{\pi/4}(\rho\sin(2\theta))^{2k}d\theta}$$
$$= \sqrt{\frac{\sqrt{\pi}\rho^{2k}\Gamma\left(k+\frac{1}{2}\right)}{4\Gamma(k+1)}}$$
$$= \frac{\sqrt{\pi}|\rho|^k}{2}\sqrt{\frac{(2k-1)!!}{(2k)!!}}$$
$$\leq \frac{\sqrt{\pi}|\rho|^k}{2^{3/2}}.$$

Thus, we have bounded our second L_2 norm:

$$\begin{aligned}\|-\log(1-\rho\sin 2\theta))\|_2 &\leq \frac{\sqrt{\pi}}{2^{3/2}}\sum_{k=1}^{\infty}\frac{|\rho|^k}{k}\\ &= -\frac{\sqrt{\pi}}{2^{3/2}}\log(1-|\rho|)\end{aligned}$$

Putting our two L_2 norms back together, we get:

$$\begin{aligned}\left\|\frac{-\log(1-\rho\sin 2\theta))}{1-\rho\sin(2\theta)}\right\|_1 &\leq \left(\left\|\frac{1}{1-\rho\sin(2\theta)}\right\|_2\right)\left(\|-\log(1-\rho\sin 2\theta))\|_2\right)\\ &= \frac{\sqrt{\pi}\,(-\log(1-|\rho|))}{4(1-\rho)^{3/4}}\sqrt{\rho\sqrt{1-\rho}+2\phi(\rho)}\end{aligned}$$

So J_2 is bounded above:

$$\begin{aligned}J_2 &= \frac{\sqrt{1-\rho^2}}{4\pi}\left[\int_0^{\pi/4}\frac{-\log(1-\rho\sin 2\theta))}{1-\rho\sin 2\theta}d\theta\right]\\ &= \frac{\sqrt{1-\rho^2}}{4\pi}\left\|\frac{-\log(1-\rho\sin 2\theta))}{1-\rho\sin(2\theta)}\right\|_1\\ &\leq \frac{\sqrt{1+\rho}\,(-\log(1-|\rho|))}{16\sqrt{\pi}(1-\rho)^{1/4}}\sqrt{\rho\sqrt{1-\rho}+2\phi(\rho)}.\end{aligned}$$

On the other hand, the term $\bar{J}_2 \leq 0$ is of neglible magnitude, relative to J_2, Thus, we will ignore it for our upper bound, and assume the worst case scenario (i.e. $\bar{J}_2 = 0$).

We will now examine our third term, J_3:

$$J_3 = \frac{\sqrt{1-\rho^2}}{4\pi}\left[\int_0^{\pi/4}\frac{2\log(\sin\theta)}{1-\rho\sin 2\theta}d\theta\right] = \frac{\sqrt{1-\rho^2}}{2\pi}\left[\int_0^{\pi/4}\frac{\log(\sin\theta)}{1-\rho\sin 2\theta}d\theta\right]$$

Observe that $J_3 \leq 0$, so our upper bound for this term will ultimately have a lower magnitude.

$$\begin{aligned}\int_0^{\pi/4} \frac{\log(\sin\theta)}{1-\rho\sin 2\theta} d\theta &\leq \int_0^{\pi/4} \log(\sin\theta) d\theta \\ &= \frac{-1}{4}\left(2K + \pi\log(2)\right) \\ &\approx -1.00238\end{aligned}$$

where K is Catalan's constant. So,

$$J_3 \leq \frac{-\sqrt{1-\rho^2}}{8\pi}\left(2K + \pi\log(2)\right)$$

On the other hand, when we repeat the computations for for $\bar{J}_3$, we have

$$\begin{aligned}\bar{J}_3 &= \frac{\sqrt{1-\rho^2}}{2\pi}\int_0^{\pi/4} \frac{\log(\sin\theta)}{1+\rho\sin 2\theta} d\theta \\ &\leq \frac{\sqrt{1-\rho^2}}{4\pi}\int_0^{\pi/4} \log(\sin\theta) d\theta \\ &= \frac{-\sqrt{1-\rho^2}}{16\pi}\left(2K + \pi\log(2)\right)\end{aligned}$$

So

$$J_3 + \bar{J}_3 \leq \frac{-3\sqrt{1-\rho^2}}{16\pi}\left(2K + \pi\log(2)\right)$$

Collecting our bounds on $I + \bar{I} = J_1 + \bar{J}_1 + J_2 + J_3 + \bar{J}_3$, we arrive at the expression in (3.9.15).

□

Computation 3. *Let g be the interaction kernel from Eq. (E4), on $\mathcal{A}_{X,Y} \subset \mathbb{R}^2$, and let V be the image in $\mathcal{A}_{X,Y}$ of the event as in Theorem 4, i.e.*

$$C = \{(x, y) | \frac{\partial_y g}{\partial_x g}(x, y) \geq 1\}.$$

Let $\tilde{C} = C_+ \sqcup C_-$ be as in Eqs. (3.9.7-3.9.9). Then for $\alpha < 2$, we have that $\tilde{C} \subset C$.

Proof. Observe that

$$\frac{|\partial_y g|}{|\partial_x g|}(x, y) = \frac{1 + e^{\alpha - x}}{|y| e^{\alpha - x}}$$

and so $|\partial_y g| \geq |\partial_y g|$ if and only if

$$|y| \leq \underbrace{\frac{1 + e^{\alpha - x}}{e^{\alpha - x}}}_{r(x)}.$$

Thus, we can think of C as the region between the curves $y = r(x)$ and $y = -r(x)$, for the positive real function r.

The region $\tilde{C}$ is bounded above and below in the (X, Y)-plane by the lines $y = \pm x$. If $x < r(x)$ for all $x \in \mathbb{R}$, then it follows that $\tilde{C} \subset C$, since the identity line $y = x$ will always be beneath $y = r(x)$, and similarly for the lower bound. Thus, it suffices to demonstrate that the following real function is strictly positive for $\alpha < 2$:

$$h(x) = r(x) - x.$$

It is clear that h is continuous on $\mathbb{R}$, and that $h(0) = r(0) > 0$. It suffices to show that h is non-vanishing.

Let $\tilde{h}(x) = e^{\alpha - x} r(x) = 1 + (1 - x)e^{\alpha - x}$, and observe that $h(x) = 0 \iff \tilde{h}(x) = 0$. Using the first derivative $\tilde{h}'(x)$, we see that $\tilde{h}$ has a global minimum at $x = 2$, where

$$\tilde{h}(2) = 1 - e^{\alpha - 2}.$$

If $\alpha < 2$, then $\tilde{h}(2) > 0$, and so $\tilde{h} > 0$. This in turn implies that h is non-vanishing.

□

3.10 Concluding Thoughts: Perspective on the Specificity of Edge Nomination

In this chapter, we have evaluated the appropriateness of the $I_\cap^{\min}$ and $I_\cap^{\mathrm{PM}}$ PIDs as tools of network inference. We have also extended these PIDs to continuous interactions, in order to better understand their behavior. In order to address the problem of gene network inference, we initially paid special attention to the bivariate synergies $S^{\min}$ and S^{PM}, which hypothetically might serve to nominate response-relevant edges in gene networks or other multiomic frameworks. To better understand the behavior that we saw in our simulations in Sec. 3.3, we have taken a multifaceted approach, examining both discrete and continuous PIDs, the former in network simulations, and the latter in noise-free systems of three variables. This chapter may better be characterized as an extended discursion around the central problem (Sec. 3.2), rather than a single straight-forward attempt to develop and prescribe a methodology. If we had taken this latter approach, we would have instead found ourselves proposing a theoretical 'solution' to what may be considered a bioinformatics question.[17] Instead, we developed a perspective with which to better understand the behavior of PIDs as applied to numerically or analytically tractable models of interacting gene networks. We have examined the discrete PIDs of simple networks, and have also extended them to continuous, noise-free interactions, in order to understand how they might capture non-linear, synergistic relationships between interacting gene pairs and a biological response variable, such as drug effectiveness.

[17] As observed in [94], simulated network and those 'known' to the biological literature do not always or even typically align, in the sense that different methods may be appropriate for synthetic and real biological networks (see Sec. 3.1.2.3 and the accompanying discussion of the PIDC method). This was the case for synthetic networks built upon models much less elementary than our own. Thus, we do not aim in this work to develop nor to prescribe a network inference methodology.

To that end, we may now summarize the conclusions that we have drawn about the $I_\cap^{\mathrm{PM}}$ and $I_\cap^{\min}$ PIDs, as they are applied to the network inference problem formulated in Section 3.2. We will also highlight the major findings of our computational and analytic investigation as they relate to this problem, so that the reader may more easily weigh the strength of our evidence and conclusions.

1. **The $I_\cap^{\mathbf{PM}}$ PID is ill-suited to synergy-network inference, as $S^{\mathbf{PM}}$ is not specific.**

   (a) Our experiments in Section 3.3 suggest that $\hat{S}^{\mathrm{PM}}$ has a non-specific sensitivity to pairs (X, Y) with high mutual information toward T, independent of the balance of information between the predictors. This non-specificity extends also to inflated redundancy R^{PM} and negative unique information U_X^{PM} for the less informative predictor.

   - As we saw in the experiments in Section 3.3, although $\hat{S}^{\mathrm{PM}}$ is fairly sensitive to interactions that are informative of the response variable, this tracks closely with the total MI $I(T; X, Y)$ of the pair (X, Y). The $\hat{S}^{\mathrm{PM}}$ synergy cannot distinguish between pairs demonstrating synergy, in which $\hat{I}(T; X, Y)$ cannot be explained by $\hat{I}(T; X)$ or $\hat{I}(T; Y)$, and 'false interactions' in which only one predictor is informative of the response, i.e. when $\hat{I}(T; X, Y) \approx \hat{I}(T; Y)$. Experiment II was particularly striking in this regard, as a false interaction (X_i, Y_2) was assigned greater synergy than a true interaction (X_1, Y_1) when the univariate signal from βY_2 was stronger than the joint signal from $g(X_1, Y_1)$.
   - Expanding upon these observations, we also saw in our experi-

ments that the inflated value of $\hat{S}^{\mathrm{PM}}$ relative to other synergies (for both 'true' and 'false' interactions) mirrored the inflated value of $\hat{R}^{\mathrm{PM}}$ for these pairs, in the sense that $\hat{R}^{\mathrm{PM}} > \hat{I}(T;X)$, which is impossible for any PID obeying the monotonicity WB axiom (M). Moreover, this inflated value of R^{PM} is accounted for (in the sense of Eq. (E1)) by a correspondingly negative value of $\hat{U}_X^{\mathrm{PM}}$ for the less informative switch gene X. See Figs. 3.4, 3.8, 3.10, specifically. See also Eq. (3.6.33), in which we noted that between any two bivariate PIDs of finite information,

$$\Delta S = \Delta R = -\Delta U_X.$$

Thus, we suggest that the non-specificity of the $I_\cap^{\mathrm{PM}}$ is due to its allotment of too much positive information to R^{PM}, accounted for with negative unique information U_X^{PM} for the less relevant predictor in a pair.

- This investigative turn toward redundancy R and unique information U_X^{PM}, as opposed to the synergy S, allows us to approach noise-free continuous interactions, in which $I(T;X,Y) = \infty$, and thus $S = \infty$ even when $I(T;X)$ and $I(T;Y)$ are finite. We may nonetheless demonstrate the inflated redundancy (and negative unique information) for less relevant predictors in such a case, in order to indirectly suggest a structural mechanism of non-specificity for $\hat{S}^{\mathrm{PM}}$ in our experiments, manifested in a continuous information and PID context.

(b) When the $I_\cap^{\mathrm{PM}}$ PID is extended to continuous, noise-free interactions, its non-specificity can be demonstrated analytically, in the

sense that R^{PM} assigns positive redundant information and U_X^{PM} assigns negative unique information to the non-informative predictor X. **In particular, we demonstrate that the continuous $I_\cap^{\mathbf{PM}}$ PID does not respect conditional independence.**

- For the continuous $I_\cap^{\mathrm{PM}}$ PID of the noise-free linear interaction

$$T = aX + bY \qquad \text{where } 0 < a < b, \tag{3.7.1}$$

  we saw that $U_X^{\mathrm{PM}} \to -\infty$ and $U_X^{\mathrm{PM}}/R^{\mathrm{PM}} \to -1$ as $a/b \to 0^+$ (Corollary 2), i.e. as X and T approached conditional independence. Note that, by contrast, $U_X^{\min} \to 0$.

- For the continuous $I_\cap^{\mathrm{PM}}$ PID of the noise-free sigmoidal switch interaction used in our earlier experiments (Sec. 3.3), given by the kernel:

$$T = \frac{Y}{1 + e^{\alpha - X}} \tag{E4}$$

  we saw that $U_X^{\mathrm{PM}} \to -\infty$ as $\alpha \to -\infty$, i.e. for α small enough that $T \approx Y$ (Prop. 10). By contrast, $U_X^{\min} \to 0$.

- More generally, in Section 3.8 we consider the continuous $I_\cap^{\mathrm{PM}}$ PID of a noise-free interaction with a kernel g that satisfies the requirements of Theorem 4. Under appropriate conditions, $U_X^{\mathrm{PM}} \to -\infty$ as $|\partial_x g|/|\partial_y g| \to 0$ (Cor. 3). As discussed in Sec. 3.8.2, this increasingly negative unique information can be understood as the unbounded growth (in relative terms) of a **misinformative probability mass exclusion**, as described in the theoretical foundation of the $I_\cap^{\mathrm{PM}}$ PID in [46] and [45].

2. **The $I_\cap^{\mathbf{min}}$ treats false interactions in an intuitive way, and has**

many desirable properties with respect to specificity. Many of these follow from the Monotonicity WB axiom (M). However, our experiments in Sec. 3.3 suggest that $S^{\mathbf{min}}$ **may be insufficiently sensitive for edge nomination.**

(a) Our experiments in Section 3.3 suggest that $\hat{S}^{\mathbf{min}}$ **distinguishes true interactions from false ones, and is sensitive to changes in the balance of information between predictors** X **and** Y**.** Thus, for instance, when one predictor X has small conditional information about the target T given the other predictor Y (i.e. approaching conditional independence), the empirical $I_{\cap}^{\min}$ PID will locate most of the information in $\hat{I}(T; X, Y)$ in the unique information atom $\hat{U}_Y^{\min}$, and little of it in $\hat{S}^{\min}$ or $\hat{U}_X^{\min}$. **However,** $\hat{S}^{\mathbf{min}}$ **is not very sensitive,** in the sense that $\hat{S}^{\min}$ does not reliable rank interacting pairs highly compared to random (null) pairs, even when the MI $\hat{I}(T; X, Y)$ is greater than most other pairs.

- Experiment I demonstrated the insensitivity of $S^{\min}$ to asymmetric, non-linear interactions. In this experiment, we examined a network with a response that was the sum of four sigmoidal switch interactions on the same 4-star, i.e.

$$T = \sum_{i=1}^{4} \frac{Y}{1 + e^{-X_i}}. \tag{3.3.1}$$

  $\hat{S}^{\min}$ did not consistently rank true interactions above most random gene pairs in our network, in terms of the synergistic information provided by the pair regarding the response. Indeed, S^{PM} was the only synergy to consistently locate true interactions in the top 95% of the empirical distribution of S over all

pairs. This is striking, as the MI $\hat{I}(T; X, Y)$ itself ranked true pairs in in the top 5%. Thus, only $\hat{S}^{\text{PM}}$ was *as* useful as MI, let alone more useful, for this experiment.

- In Experiment II , on the other hand, we saw that $S^{\min}$ was somewhat more sensitive and *specific* in its ability to nominate a sigmoidal interaction within a simpler response, which took the form:

$$T = \frac{Y_1}{1 + e^{-X_1}} + \beta Y_2. \tag{3.3.2}$$

  By adjusting β, we engineer a situation in which MI cannot distinguish the true pairs from a false pairs, and neither could $\hat{S}^{\text{PM}}$, as discussed above. $\hat{S}^{\min}$ reliably ranked the true interaction highly, while ranking false pairs that included Y_2 near the mean of random pairs, see Fig. 3.6. Thus, we saw that whatever its sensitivity shortcomings for multiple interactions, $\hat{S}^{\min}$ respects conditional independence in simulation (and in theory, Prop. 7).

- Much in the same way that the negative value of U_X^{PM} for an irrelevant predictor X can be used to account for the non-specificity of the $I_\cap^{\text{PM}}$ PID, we may consider $U_Y^{\min}$ to demonstrate a mirror phenomenon. In Exp II , for a false pair (X_j, Y_2), we saw that that the $I_\cap^{\min}$ PID properly assigns most of the information from the term βY_2 to the unique information $\hat{U}_Y^{\min}$, and that $\hat{U}_Y^{\min}$ increases linearly with $\hat{I}(T; X_i, Y_2)$,, as we increase β (Fig. 3.8). The amount of redundant information $R^{\min}$ constant, reflecting the constant level of dependency between X_i and Y_2.

- **The $\hat{I}_\cap^{\min}$ PID demonstrates that it is sensitive to the balance of information between variables X and Y, unlike the $I_\cap^{\mathrm{PM}}$ PID.** We saw this most clearly in Experiment III , in which we considered the $I_\cap^{\min}$ and $I_\cap^{\mathrm{PM}}$ PIDs, normalized by MI, as we adjusted the α paramter in the sigmoidal switch kernel (E4). We saw that the proportion of MI assigned to $\hat{U}_Y^{\mathrm{PM}}$, on the one hand, and the combined joint information $\hat{S}^{\min} + \hat{R}^{\min}$, on the other, followed similar patterns to the balance between the relative magnitudes of the Taylor coefficients $|\partial_y g|$ and $|\partial_{xy} g|$, respectively (Fig. 3.11).

(b) When the $I_\cap^{\min}$ PID is extended to continuous, noise-free interactions, its specificity can be demonstrated in that it accounts for the balance of information between predictors X and Y by resembling the MMI PID (from [6]), assigning no unique information ($U_X^{\min} = 0$) and only redundant information ($R^{\min} = I(T;X)$) for a minimally informative predictor X. **In particular, we demonstrate that the continuous $I_\cap^{\min}$ PID *does* respect conditional independence, unlike the $I_\cap^{\mathrm{PM}}$ PID.**

- For the continuous $I_\cap^{\mathrm{PM}}$ PID of the noise-free linear interaction (3.7.1) as above, we had that $U_X^{\min} = 0$ when $a < b$ and $I(T;X) < I(T;Y)$, as in [6]. As we let the response approach a univariate function, i.e. $a/b \to 0^+$, we had that $U_Y^{\min}/I(T;Y) \to 1$ and $R^{\min} \to I(X;Y)$ (Cor. 2), i.e. as X and T approached conditional independence.
- For the continuous $I_\cap^{\min}$ PID of the noise-free sigmoidal switch interaction (E4), given above, we have that $U_X^{\min} \to 0$ as $\alpha \to -\infty$, i.e. when $T \approx Y$ (Prop. 10) .

- More generally, in Section 3.8 we consider the continuous $I_\cap^{\min}$ PID of a noise-free interaction with a kernel g that satisfies the conditions of Theorem 4. Under appropriate conditions, $U_X^{\min} \to 0$ as as $|\partial_x g|/|\partial_y g| \to 0$ (Cor. 3).

Without discounting the importance of measure sensitivity, specificity is of central importance to any gene network inference task. For cellular biologists, a false negative may be less costly and disruptive than a false positive, and the complexity of gene networks heightens the risk of spurious associations. **To that end, we consider it crucial that a bivariate PID respect conditional independence,** so that indirect and spurious associations might be discounted in expectation.

The approach that we have begun developing in Cor. 2 & 3 for the $I_\cap^{\min}$ and $I_\cap^{\mathrm{PM}}$ PIDs offers a means of evaluating the specificity of these PIDs for a particular interaction. By continuously or sequentially altering kernel parameters, we have been able to explore the specificity of the PIDs as we alter the analytic balance of information contained within source variables about the target.

There are three directions in which we would aim to expand our current effort. First, we would expand our analysis to noisy bivariate interactions. Although exact computations of PID atoms may not prove tractable, the marginals $p_{X,T}$ and $p_{Y,T}$ for noisy interactions would resemble convolutions of the noise-free marginals in Prop. 9. Thus, we may be able to bound the unique information atoms around formulas like those in Theorem 4.

Second, there are many other PIDs worth considering in the synergistic network context. The recent work extending I^{BROJA} to continuous variables is promising [92]. However, Experiment I in Section 3.3 casts doubt on the sensitivity of I^{BROJA}. It seems probable that I^{BROJA} will resemble $I_\cap^{\min}$ in

demonstrating specificity without sufficient sensitivity to synergistic interactions. We suspect that both $I_\cap^{\min}$ and I^{BROJA} tend to revert to the MMI PID (Sec. 3.1.2.4) for many non-linear interactions. The $I_\cap^{CCS}$ PID, on the other hand, did almost as well in Experiment I as the $I_\cap^{\text{PM}}$ PID. It would be helpful if future work could establish whether or not the $I_\cap^{CCS}$ PID has similar issues with specificity as the $I_\cap^{\text{PM}}$ PID. Both of them violate the monotonicity WB axiom (Axiom M) [45, 56]. It is not clear, however, that this necessarily implies that the $I_\cap^{CCS}$ PID likewise does not respect conditional independence, as does the $I_\cap^{\text{PM}}$ PID.

Finally, and perhaps more ambitiously, it would be worthwhile to extend a similar analysis to simple continuous-time stochastic processes. Barrett already considered discrete time MVAR processes in [6]. Consider the following general model of a GRN. For a network of time-varying gene expressions $X_1(t), ..., X_n(t)$, consider, for instance, a system of the form

$$\dot{X}_1 = \sum_{(i,j)\in\mathcal{E}_1'} g_{i,j}(X_i, X_j|\boldsymbol{\theta}_{i,j}) + \epsilon_1 \tag{3.10.1}$$

$$\vdots \qquad\qquad \vdots$$

$$\dot{X}_n = \sum_{(i,j)\in\mathcal{E}_n'} g_{i,j}(X_i, X_j|\boldsymbol{\theta}_{i,j}) + \epsilon_n \tag{3.10.2}$$

where $\mathcal{E}_k'$ represent the X_k-targeting synergistic interactions in the network, in which two genes X_i and X_j jointly impact X_k, $g_{i,j}$ represent specific kernels dependent on parameters $\boldsymbol{\theta}_{i,j}$, and ϵ_i are mean zero, unit variance noise terms, with $\mathbb{E}\epsilon_i\epsilon_j = \pm\rho$ when $(i,j) \in \mathcal{E}$, i.e. associational edges. Such synergistic activity is typical in genetic circuits. For instance, translated transcription factors are known to activate other genes synergistically [99, 119]. Could we use PID synergy to infer the transcriptional circuits in such a network? We

would consider the simpler model

$$\dot{X} = s(t) - p_Y X - d_X X + \epsilon_X$$
$$\dot{Y} = p_Y X - d_Y Y + \epsilon_Y$$
$$\dot{T} = g(X, Y|\boldsymbol{\theta}) - d_T T + \epsilon_T$$

If we took a sequence of parameters $\boldsymbol{\theta}_i$ such that $|\partial_x g|/|\partial_y g| \to 0$, we would be approaching a system in which $T(t)$ and $X(t)$ are conditionally independent as processes, given the data $Y(t)$. Although it is not immediately apparent to us how PID ought best be applied to such a system, any measure of 'synergistic' regulation within a dynamical GRN ought to approach zero in such a limit.

Ours is among a handful of works applying PID to continuous variables, and the only one (to our knowledge) extending the definitions of $I_\cap^{\min}$ and $I_\cap^{\mathrm{PM}}$ directly. [18] Our emphasis has been on the source bivariate PIDs of trivariate model systems, in which the target is a noise-free realization of a smooth function of the predictors. This is the preliminary form of our central perspective, which seeks to understand a response variable as a (potentially noisy) realization of a regression on the source variables. This perspective is distinct, at least in emphasis, from the other work on continuous PIDs [6, 92], and has its limitations. The designation of source and target variables may be arbitrary for real systems evaluated from an objective perspective, as complex biological systems are structurally characterized by feedback and loops rather than feed-forward information flows [9]. Nonetheless, the perspective is arguably the correct one for experimentalists interested in manipulating a system, e.g. killing tumor cells. It is also the appropriate framework for merging continuous PIDs into statistical inference and regression analysis, both in a classical

[18] The work in [6] computes $I_\cap^{\min}$ indirectly, using I^{BROJA}.

setting and within complex networks.

3.A Auxiliary Proofs

3.A.1 Useful Rules for Gaussian Interactions

Rule 1. *In general, if* $\boldsymbol{U} \sim N(\boldsymbol{\mu_U}, \Sigma_{\boldsymbol{U}})$ *and* $\boldsymbol{V} \sim N(\boldsymbol{\mu_V}, \Sigma_{\boldsymbol{V}})$ *are* k*-dimensional Gaussian vectors such that* $\boldsymbol{V} = A\boldsymbol{U}$ *for an* $n \times n$ *real, nonsingular transformation* A*, then*

$$\boldsymbol{\mu}_V = A\boldsymbol{\mu}_U,$$
$$\Sigma_{\boldsymbol{V}} = A\Sigma_{\boldsymbol{U}}A^T.$$

Rule 2. *Let* X_1, X_2 *be two Gaussians with correlation* ρ*, means* μ_1, μ_2*, standard deviations* σ_1, σ_2*. Then*

$$X_1|_{X_2=x^{(2)}} \sim N(\mu, \sigma^2),$$

where

$$\mu = \mu_X + \rho\frac{\sigma_1}{\sigma_2}(x^{(2)} - \mu_2),$$
$$\sigma^2 = \sigma_1^2(1 - \rho^2).$$

3.A.2 Computations

In this section, we keep details for computations that use elementary methods of little technical interest.

Computation 4 (The function $f(\gamma)$ in Eq 3.7.18 is increasing.)**.** *For any* $\rho \in (-1, 1)$*, the function* $f(\gamma)$ *from Eq 3.7.18, reproduced below, is increasing in* γ *on* $[1, \infty)$*:*

$$f(\gamma) = \log \gamma - \frac{\log(e)(1-\rho^2)(\gamma^2-1)}{2(\gamma^2+2\rho\gamma+1)} \tag{3.7.18}$$

Proof. When we take the derivative of this function with respect to γ, we have

$$f'(\gamma) = \frac{1}{\gamma} - \frac{2(\gamma^2+2\rho\gamma+1)2\gamma(1-\rho^2) - (1-\rho^2)(\gamma^2-1)2(2\rho+2\gamma)}{4(\gamma^2+2\rho\gamma+1)^2}$$
$$= \frac{4(\gamma^2+2\rho\gamma+1)^2 - 4\gamma^2(\gamma^2+2\rho\gamma+1)(1-\rho^2) + \gamma(1-\rho^2)(\gamma^2-1)2(2\rho+2\gamma)}{4\gamma(\gamma^2+2\rho\gamma+1)^2}$$

Positivity of the derivative follows then from

$$4(\gamma^2+2\rho\gamma+1)^2 - 4\gamma^2(\gamma^2+2\rho\gamma+1)(1-\rho^2)$$
$$= (\gamma^2+2\rho\gamma+1)(4+8\rho\gamma+4\gamma^2-4\gamma^2+4\gamma^2\rho^2)$$
$$= (\gamma^2+2\rho\gamma+1)4(1+\rho\gamma)^2 >$$

which is > 0 as desired. □

Chapter 4

Concluding Remarks

In this dissertation, we have studied complex, non-linear interactions in real and simulated biological systems. In the body composed of Chapters 2 and 3, we have examined both dynamical DDE systems and static random network models, corresponding to the cellular and molecular scales in biology, respectively. Each chapter has a more extensive denouement, in Secs. 2.5 and 3.10, respectively, in which we discussed the significant of our results and directions of future work. We emphasize the discussion at the end of Sec. 3.10, as it is here that we hint at a direction that lies at the intersection of both works presented: the use of continuous PIDs to track multivariate information flows in continuous time dynamical systems and stochastic processes.

Bibliography

[1] Hirotugu Akaike. A new look at the statistical model identification. IEEE transactions on automatic control, 19(6):716–723, 1974.

[2] Dimitris Anastassiou. Computational analysis of the synergy among multiple interacting genes. Molecular systems biology, 3(1):83, 2007.

[3] Byron B. Au-Yeung, Geoffrey Alexander Smith, James L. Mueller, Cheryl S. Heyn, Rebecca Garrett Jaszczak, Arthur Weiss, and Julie Zikherman. IL-2 Modulates the TCR Signaling Threshold for CD8 but Not CD4 T Cell Proliferation on a Single-Cell Level. The Journal of Immunology, 198(6):2445–2456, 3 2017.

[4] Arnaud Augert, Emily Eastwood, Ali H. Ibrahim, Nan Wu, Eli Grunblatt, Ryan Basom, Denny Liggitt, Keith D. Eaton, Renato Martins, John T. Poirier, Charles M. Rudin, Francesca Milletti, Wei-Yi Cheng, Fiona Mack, and David MacPherson. Targeting NOTCH activation in small cell lung cancer through LSD1 inhibition. Science Signaling, 12(567):eaau2922, 2 2019.

[5] Nihat Ay, Eckehard Olbrich, Nils Bertschinger, and Jürgen Jost. A geometric approach to complexity. Chaos: An Interdisciplinary Journal of Nonlinear Science, 21(3):037103, 2011.

[6] Adam B. Barrett. Exploration of synergistic and redundant information sharing in static and dynamical gaussian systems. Phys. Rev. E, 91:052802, May 2015.

[7] Katia Basso, Adam A Margolin, Gustavo Stolovitzky, Ulf Klein, Riccardo Dalla-Favera, and Andrea Califano. Reverse engineering of regulatory networks in human b cells. Nature genetics, 37(4):382–390, 2005.

[8] Jacob Beal. Biochemical complexity drives log-normal variation in genetic expression. Engineering Biology, 1(1):55–60, June 2017.

[9] Anthony J Bell. Levels and loops: the future of artificial intelligence and neuroscience. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 354(1392):2013–2020, 1999.

[10] Anthony J Bell. The co-information lattice. In Proceedings of the Fifth International Workshop on Independent Component Analysis and Blind Signal Separation: ICA, volume 2003. Citeseer, 2003.

[11] Nils Bertschinger, Johannes Rauh, Eckehard Olbrich, and Jürgen Jost. Shared information—new insights and problems in decomposing information in complex systems. In Proceedings of the European Conference on Complex Systems 2012, pages 251–269. Springer, 2013.

[12] Nils Bertschinger, Johannes Rauh, Eckehard Olbrich, Jürgen Jost, and Nihat Ay. Quantifying unique information. Entropy, 16(4):2161–2183, Apr 2014.

[13] Ayan Biswas. Multivariate information processing characterizes fitness of a cascaded gene-transcription machinery. Chaos: An Interdisciplinary Journal of Nonlinear Science, 29(6):063108, June 2019.

[14] T Boulding, R D McCuaig, A Tan, K Hardy, F Wu, J Dunn, M Kalimutho, C R Sutton, J K Forwood, A G Bert, G J Goodall, L Malik, D Yip, J E Dahlstrom, A Zafar, K K Khanna, and S Rao. LSD1 activation promotes inducible EMT programs and modulates the tumour microenvironment in breast cancer. Scientific Reports, 8(1):73, 2018.

[15] Naama Brenner, Steven P Strong, Roland Koberle, William Bialek, and Rob R de Ruyter van Steveninck. Synergy in a neural code. Neural computation, 12(7):1531–1552, 2000.

[16] Atul J Butte and Isaac S Kohane. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. In Biocomputing 2000, pages 418–429. World Scientific, 1999.

[17] Atul J Butte and Isaac S Kohane. Unsupervised knowledge discovery in medical databases using relevance networks. In Proceedings of the AMIA Symposium, page 711. American Medical Informatics Association, 1999.

[18] Atul J Butte and Isaac S Kohane. Relevance networks: a first step toward finding genetic regulatory networks within microarray data. In The analysis of gene expression data, pages 428–446. Springer, 2003.

[19] Atul J Butte, Pablo Tamayo, Donna Slonim, Todd R Golub, and Isaac S Kohane. Discovering functional relationships between rna expression and chemotherapeutic susceptibility using relevance networks. Proceedings of the National Academy of Sciences, 97(22):12182–12186, 2000.

[20] Zixuan Cang and Qing Nie. Inferring spatial and signaling relationships between cells from single cell transcriptomic data. Nature communications, 11(1):1–13, 2020.

[21] George Casella and Roger Berger. Statistical Inference. Thomson Learning, Pacific Grove, CA, 2 edition, 2002.

[22] Thalia E Chan. Learning large-scale gene regulatory networks from single cell transcriptomic data using multivariate information theory. 2018.

[23] Thalia E. Chan, Ananth V. Pallaseni, Ann C. Babtie, Kirsten R. McEwen, and Michael P.H. Stumpf. Empirical bayes meets information theoretical network reconstruction from single cell data. February 2018.

[24] Thalia E Chan, Michael PH Stumpf, and Ann C Babtie. Gene regulatory network inference from single-cell data using multivariate information measures. Cell systems, 5(3):251–267, 2017.

[25] Pritam Chanda, Eduardo Costa, Jie Hu, Shravan Sukumar, John Van Hemert, and Rasna Walia. Information theory in computational biology: Where we stand today. Entropy, 22(6):627, 2020.

[26] Pritam Chanda, Aidong Zhang, Daniel Brazeau, Lara Sucheston, Jo L Freudenheim, Christine Ambrosone, and Murali Ramanathan. Information-theoretic metrics for visualizing gene-environment interactions. The American Journal of Human Genetics, 81(5):939–963, 2007.

[27] Prantik Chatterjee and Nikhil Ranjan Pal. Construction of synergy networks from gene expression data related to disease. Gene, 590(2):250–262, 2016.

[28] Belal Chaudhary and Eyad Elkord. Regulatory T Cells in the Tumor Microenvironment and Cancer Progression: Role and Therapeutic Targeting. 2016.

[29] Shuonan Chen and Jessica C Mar. Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. BMC bioinformatics, 19(1):1–21, 2018.

[30] Takatoshi Chinen, Arun K. Kannan, Andrew G. Levine, Xiying Fan, Ulf Klein, Ye Zheng, Georg Gasteiger, Yongqiang Feng, Jason D. Fontenot,

and Alexander Y. Rudensky. An essential role for the IL-2 receptor in T reg cell function. Nature Immunology, 17(11):1322–1333, 10 2016.

[31] Thomas Condamine, Steve Wang, Melody Diamond, Leslie Hall, Huiqing Liu, Antony Chadderton, Jin Lu, Chunhong He, Liangxing Wu, Timothy Burn, Wenqing Yao, Gregory Hollis, Reid Huber, Bruce Ruggeri, Peggy Scherle, Holly Koblish, and Sang Hyun Lee. Abstract 4635: The LSD1 Specific Inhibitor INCB059872 enhances the activity of immune checkpoint blockade by reshaping the myeloid compartment in the syngeneic 4T1 mouse mammary tumor model. 2017.

[32] UC Davis Bioinformatics Core. Differential gene expression analysis in r, Mar 2019.

[33] Thomas M. Cover and Joy A. Thomas. Elements of Information Theory. Wiley, April 2005.

[34] Rob J. De Boer, Dirk Homann, and Alan S. Perelson. Different Dynamics of CD4 + and CD8 + T Cell Responses During and After Acute Lymphocytic Choriomeningitis Virus Infection . The Journal of Immunology, 2003.

[35] Lisette G De Pillis, Ami E Radunskaya, and Charles L Wiseman. A Validated Mathematical Model of Cell-Mediated Immune Response to Tumor Growth. 2005.

[36] Lloyd Demetrius and Thomas Manke. Robustness and network evolution—an entropic principle. Physica A: Statistical Mechanics and its Applications, 346(3-4):682–696, 2005.

[37] Chris Ding and Hanchuan Peng. Minimum redundancy feature selection

from microarray gene expression data. Journal of bioinformatics and computational biology, 3(02):185–205, 2005.

[38] Changzheng Dong, Xun Chu, Ying Wang, Yi Wang, Li Jin, Tieliu Shi, Wei Huang, and Yixue Li. Exploration of gene–gene interaction effects using entropy-based methods. European Journal of Human Genetics, 16(2):229–235, 2008.

[39] Raluca Eftimie, Jonathan L Bramson, and David J D Earn. Interactions Between the Immune System and Cancer: A Brief Review of Non-spatial Mathematical Models. Bulletin of Mathematical Biology, 73(1):2–32, 2011.

[40] Raluca Eftimie, Joseph J Gillard, and Doreen A Cantrell. Mathematical Models for Immunology: Current State of the Art and Future Research Directions. Bull Math Biol, 78:2091–2134, 2016.

[41] Frank Emmert-Streib, Galina V. Glazko, Gökmen Altay, and Ricardo de Matos Simoes. Statistical inference and reverse engineering of gene regulatory networks from observational expression data. Frontiers in Genetics, 3, 2012.

[42] Jeremiah J Faith, Boris Hayete, Joshua T Thaden, Ilaria Mogno, Jamey Wierzbowski, Guillaume Cottarel, Simon Kasif, James J Collins, and Timothy S Gardner. Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles. PLoS biol, 5(1):e8, 2007.

[43] R Fan, M Zhong, S Wang, Y Zhang, A Andrew, M Karagas, H Chen, CI Amos, M Xiong, and JH Moore. Entropy-based information gain approaches to detect and to characterize gene-gene and gene-environment

interactions/correlations of complex diseases. Genetic epidemiology, 35(7):706–721, 2011.

[44] Zizhen Feng, Yuan Yao, Chao Zhou, Fengju Chen, Fangrui Wu, Liping Wei, Wei Liu, Shuo Dong, Michele Redell, Qianxing Mo, and Yongcheng Song. Pharmacological inhibition of LSD1 for the treatment of MLL-rearranged leukemia. Journal of Hematology & Oncology, 9(1):24, 2016.

[45] Conor Finn and Joseph Lizier. Pointwise partial information decomposition using the specificity and ambiguity lattices. Entropy, 20(4):297, April 2018.

[46] Conor Finn and Joseph Lizier. Probability mass exclusions and the directed components of mutual information. Entropy, 20(11):826, October 2018.

[47] Saikrishna Gadhamsetty, Athanasius F M Marée, Joost B Beltman, and Rob J de Boer. A General Functional Response of Cytotoxic T Lymphocyte-Mediated Killing of Target Cells. Biophysical Journal, 106(8):1780–1791, 2014.

[48] Saikrishna Gadhamsetty, Athanasius F.M. Marée, Joost B. Beltman, and Rob J. de Boer. A Sigmoid Functional Response Emerges When Cytotoxic T Lymphocytes Start Killing Fresh Target Cells. Biophysical Journal, 112(6):1221–1235, 3 2017.

[49] Itay Gat and Naftali Tishby. Synergy and redundancy among brain cells of behaving monkeys. Advances in neural information processing systems, pages 111–117, 1999.

[50] Adrián E. Granada, Alba Jiménez, Jacob Stewart-Ornstein, Nils Blüthgen, Simone Reber, Ashwini Jambhekar, and Galit Lahav. The

effects of proliferation status and cell cycle phase on the responses of single cells to chemotherapy. Molecular Biology of the Cell, 31(8):845–857, April 2020.

[51] Robert M Gray. Entropy and information theory. Springer Science & Business Media, 2011.

[52] Virgil Griffith and Christof Koch. Quantifying synergistic mutual information. In Guided self-organization: inception, pages 159–190. Springer, 2014.

[53] Malte Harder, Christoph Salge, and Daniel Polani. Bivariate measure of redundant information. Physical Review E, 87(1):012130, 2013.

[54] William J. Harris, Xu Huang, James T. Lynch, Gary J. Spencer, James R. Hitchin, Yaoyong Li, Filippo Ciceri, Julian G. Blaser, Brigit F. Greystoke, Allan M. Jordan, Crispin J. Miller, Donald J. Ogilvie, and Tim C.P. Somervaille. The Histone Demethylase KDM1A Sustains the Oncogenic Potential of MLL-AF9 Leukemia Stem Cells. Cancer Cell, 21(4):473–487, 4 2012.

[55] Shinya Hayami, John D. Kelly, Hyun-Soo Cho, Masanori Yoshimatsu, Motoko Unoki, Tatsuhiko Tsunoda, Helen I. Field, David E. Neal, Hiroki Yamaue, Bruce A.J. Ponder, Yusuke Nakamura, and Ryuji Hamamoto. Overexpression of LSD1 contributes to human carcinogenesis through chromatin regulation in various cancers. International Journal of Cancer, 128(3):574–586, 2 2011.

[56] Robin A. A. Ince. The partial entropy decomposition: Decomposing multivariate entropy and mutual information via pointwise common surprisal, 2017.

[57] Alexandre Irrthum, Louis Wehenkel, Pierre Geurts, et al. Inferring regulatory networks from expression data using tree-based methods. PloS one, 5(9):e12776, 2010.

[58] Vikram R. Juneja, Kathleen A. McGuire, Robert T. Manguso, Martin W. LaFleur, Natalie Collins, W. Nicholas Haining, Gordon J. Freeman, and Arlene H. Sharpe. PD-L1 on tumor cells is sufficient for immune evasion in immunogenic tumors and inhibits CD8 T cell cytotoxicity. The Journal of Experimental Medicine, 214(4):895–904, 4 2017.

[59] Tanja Kaartinen, Annu Luostarinen, Pilvi Maliniemi, Joni Keto, Mikko Arvas, Heini Belt, Jonna Koponen, Angelica Loskog, Satu Mustjoki, Kimmo Porkka, Seppo Ylä-Herttuala, and Matti Korhonen. Low interleukin-2 concentration favors generation of early memory T cells over effector phenotypes during chimeric antigen receptor T-cell expansion. Cytotherapy, 19(6):689–702, 6 2017.

[60] Thomas Kahle, Eckehard Olbrich, Jürgen Jost, and Nihat Ay. Complexity measures from interaction structures. Physical Review E, 79(2):026201, 2009.

[61] Vandana Kalia and Surojit Sarkar. Regulation of Effector and Memory CD8 T Cell Differentiation by IL-2-A Balancing Act, 2018.

[62] Vasundhra Kashyap, Shafqat Ahmad, Emeli M Nilsson, Leszek Helczynski, Sin ead Kenna, Jenny Liao Persson, Lorraine J Gudas, and Nigel P Mongan. The lysine specific demethylase-1 (LSD1/KDM1A) regulates VEGF-A expression in prostate cancer. 2013.

[63] Nadia Kavrochorianou, Melina Markogiannaki, and Sylva Haralambous.

IFN-β differentially regulates the function of T cell subsets in MS and EAE. Cytokine & Growth Factor Reviews, 30:47–54, 8 2016.

[64] Peter S Kim, Peter P Lee, and Doron Levy. A Theory of Immunodominance and Adaptive Regulation. Bulletin of Mathematical Biology, 73(7):1645–1665, 2011.

[65] Peter S Kim, Peter P Lee, and Doron Levy. Basic Principles in Modeling Adaptive Regulation and Immunodominance. In Mathematical Biosciences, volume 257, pages 33–57. 2013.

[66] Maria Kleppe, Kaitlyn Shank, Papalexi Efthymia, Hugh Riehnhoff, and Ross L. Levine. Lysine-Specific Histone Demethylase, LSD1, (KDM1A) As a Novel Therapeutic Target in Myeloproliferative Neoplasms. Blood, 126(23), 2015.

[67] Yang Kuang, John D. Nagy, and Steffen E. Eikenberry. Introduction to Mathematical Oncology. Chapman and Hall/CRC, 9 2018.

[68] Solomon Kullback and Richard A Leibler. On information and sufficiency. The annals of mathematical statistics, 22(1):79–86, 1951.

[69] Melissa G Lechner, Saman S Karimi, Keegan Barry-Holson, Trevor E Angell, Katherine A Murphy, Connor H Church, John R Ohlfest, Peisheng Hu, and Alan L Epstein. Immunogenicity of murine solid tumor models as a defining feature of in vivo behavior and response to immunotherapy.

[70] Laurel Yong-Hwa Lee and Joseph Loscalzo. Network medicine in pathobiology. The American Journal of Pathology, 189(7):1311–1326, July 2019.

[71] Wonyong Lee and Gap Ryol Lee. Transcriptional regulation and development of regulatory T cells, 2018.

[72] Woojoo Lee, Arvid Sjölander, and Yudi Pawitan. A critical look at entropy-based gene-gene interaction measures. Genetic epidemiology, 40(5):416–424, 2016.

[73] Ji-Liang Li, Richard C.A. Sainson, Wen Shi, Russell Leek, Laura S. Harrington, Matthias Preusser, Swethajit Biswas, Helen Turley, Emily Heikamp, Johannes A. Hainfellner, and Adrian L. Harris. Delta-like 4 Notch Ligand Regulates Tumor Angiogenesis, Improves Tumor Vascular Function, and Promotes Tumor Growth In vivo. Cancer Research, 67(23):11244–11253, 12 2007.

[74] Kuo-Ching Liang and Xiaodong Wang. Gene regulatory network reconstruction using conditional mutual information. EURASIP Journal on Bioinformatics and Systems Biology, 2008:1–14, 2008.

[75] Soyoung Lim, Andreas Janzer, Astrid Becker, Andreas Zimmer, Roland Schüle, Reinhard Buettner, and Jutta Kirfel. Lysine-specific demethylase 1 (LSD1) is highly expressed in ER-negative breast cancers and a biomarker predicting aggressive biology. Carcinogenesis, 31(3):512–520, 3 2010.

[76] Fei Liu, Shao-Wu Zhang, Wei-Feng Guo, Ze-Gang Wei, and Luonan Chen. Inference of gene regulatory network based on local bayesian networks. PLoS computational biology, 12(8):e1005024, 2016.

[77] Joseph Lizier, Nils Bertschinger, Juergen Jost, and Michael Wibral. Information decomposition of target effects from multi-source interactions:

Perspectives on previous, current and future work. Entropy, 20(4):307, April 2018.

[78] Abdullah Makkeh, Aaron J Gutknecht, and Michael Wibral. Introducing a differentiable measure of pointwise shared information. Physical Review E, 103(3):032149, 2021.

[79] Thomas Manke, Lloyd Demetrius, and Martin Vingron. An entropic characterization of protein interaction networks and cellular robustness. Journal of The Royal Society Interface, 3(11):843–850, 2006.

[80] Adam A Margolin, Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Riccardo Dalla Favera, and Andrea Califano. Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. In BMC bioinformatics, volume 7, pages 1–15. Springer, 2006.

[81] William McGill. Multivariate information transmission. Transactions of the IRE Professional Group on Information Theory, 4(4):93–111, 1954.

[82] Susan C. McKarns and Ronald H. Schwartz. Distinct Effects of TGF-β1 on CD4 + and CD8 + T Cell Survival, Division, and IL-2 Production: A Role for T Cell Intrinsic Smad3. The Journal of Immunology, 174(4):2071–2083, 2 2005.

[83] Alice McNally, Geoffrey R. Hill, Tim Sparwasser, Ranjeny Thomas, and Raymond J. Steptoe. CD4+CD25+ regulatory T cells control CD8+ T-cell effector differentiation by modulating IL-2 homeostasis. Proceedings of the National Academy of Sciences of the United States of America, 108(18):7529–7534, 5 2011.

[84] Patrick E Meyer, Kevin Kontos, Frederic Lafitte, and Gianluca Bontempi. Information-theoretic inference of large transcriptional regulatory networks. EURASIP journal on bioinformatics and systems biology, 2007:1–9, 2007.

[85] Jesse Milzman, Wanqiang Sheng, and Doron Levy. Modeling lsd1-mediated tumor stagnation. Bulletin of Mathematical Biology, 83(2):1–29, 2021.

[86] Thomas Moerman, Sara Aibar Santos, Carmen Bravo González-Blas, Jaak Simm, Yves Moreau, Jan Aerts, and Stein Aerts. Grnboost2 and arboreto: efficient and scalable inference of gene regulatory networks. Bioinformatics, 35(12):2159–2161, 2019.

[87] Zaynab Mousavian, José Díaz, and Ali Masoudi-Nejad. Information theory in systems biology. part ii: protein–protein interaction and signaling networks. In Seminars in cell & developmental biology, volume 51, pages 14–23. Elsevier, 2016.

[88] Zaynab Mousavian, Kaveh Kavousi, and Ali Masoudi-Nejad. Information theory in systems biology. part i: Gene regulatory and metabolic networks. In Seminars in cell & developmental biology, volume 51, pages 3–13. Elsevier, 2016.

[89] Giovanna Nicora, Francesca Vitali, Arianna Dagliati, Nophar Geifman, and Riccardo Bellazzi. Integrated multi-omics analyses in oncology: A review of machine learning methods and tools. Frontiers in Oncology, 10, June 2020.

[90] Dennis Niebel, Jutta Kirfel, Viktor Janzen, Tobias Höller, Michael Ma-

jores, and Ines Gütgemann. Lysine-specific demethylase 1 (LSD1) in hematopoietic and lymphoid neoplasms. Blood, 124(1):151–2, 7 2014.

[91] Soyoung A Oh and Ming O Li. TGF-β: guardian of T cell function. Journal of immunology (Baltimore, Md. : 1950), 191(8):3973–9, 10 2013.

[92] Ari Pakman, Dar Gilboa, and Elad Schneidman. Estimating the unique information of continuous variables. arXiv preprint arXiv:2102.00218, 2021.

[93] Marine Potez, Verdiana Trappetti, Audrey Bouchet, Cristian Fernandez-Palomo, Esra Güç, Witold W. Kilarski, Ruslan Hlushchuk, Jean Laissue, and Valentin Djonov. Characterization of a B16-F10 melanoma model locally implanted into the ear pinnae of C57BL/6 mice. PLoS ONE, 2018.

[94] Aditya Pratapa, Amogh P. Jalihal, Jeffrey N. Law, Aditya Bharadwaj, and T. M. Murali. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. Nature Methods, 17(2):147–154, January 2020.

[95] Ye Qin, Shauna N. Vasilatos, Lin Chen, Hao Wu, Zhishen Cao, Yumei Fu, Min Huang, Anda M. Vlad, Binfeng Lu, Steffi Oesterreich, Nancy E. Davidson, and Yi Huang. Inhibition of histone lysine-specific demethylase 1 elicits breast tumor immunity and enhances antitumor efficacy of immune checkpoint blockade. Oncogene, page 1, 8 2018.

[96] Johannes Rauh, Nils Bertschinger, Eckehard Olbrich, and Jürgen Jost. Reconsidering unique information: Towards a multivariate information decomposition. In 2014 IEEE International Symposium on Information Theory, pages 2232–2236. IEEE, 2014.

[97] Jonathan Rougier and Carey E Priebe. The exact form of the "ockham factor" in model selection. The American Statistician, pages 1–6, 2020.

[98] Akihisa Sakamoto, Shinjiro Hino, Katsuya Nagaoka, Kotaro Anan, Ryuta Takase, Haruka Matsumori, Hidenori Ojima, Yae Kanai, Kazunori Arita, and Mitsuyoshi Nakao. Lysine Demethylase LSD1 Coordinates Glycolytic and Mitochondrial Metabolism in Hepatocellular Carcinoma Cells. Cancer Research, 75(7):1445–1456, 4 2015.

[99] Frank Sauer, Stig K Hansen, and Robert Tjian. Multiple tafiis directing synergistic activation of transcription. Science, 270(5243):1783–1788, 1995.

[100] Tino Schenk, Weihsu Claire Chen, Stefanie Göllner, Louise Howell, Liqing Jin, Katja Hebestreit, Hans-Ulrich Klein, Andreea C Popescu, Alan Burnett, Ken Mills, Robert A Casero, Laurence Marton, Patrick Woster, Mark D Minden, Martin Dugas, Jean C Y Wang, John E Dick, Carsten Müller-Tidow, Kevin Petrie, Arthur Zelent, and Arthur Zelent. Inhibition of the LSD1 (KDM1A) demethylase reactivates the all-trans-retinoic acid differentiation pathway in acute myeloid leukemia. Nature medicine, 18(4):605–11, 3 2012.

[101] Elad Schneidman, William Bialek, and Michael J Berry. Synergy, redundancy, and independence in population codes. Journal of Neuroscience, 23(37):11539–11553, 2003.

[102] Gideon Schwarz et al. Estimating the dimension of a model. Annals of statistics, 6(2):461–464, 1978.

[103] Marta Segarra, Cassin Kimmel Williams, Maria De La, Luz Sierra, Marcelino Bernardo, Peter J Mccormick, Dragan Maric, Celeste Regino,

Peter Choyke, and Giovanna Tosato. Dll4 activation of Notch signaling reduces tumor vascularity and inhibits tumor growth.

[104] Nuran Serce, Annette Gnatzy, Susanne Steiner, Henning Lorenzen, Jutta Kirfel, and Reinhard Buettner. Elevated expression of LSD1 (Lysine-specific demethylase 1) during tumour progression from pre-invasive to invasive ductal carcinoma of the breast. BMC clinical pathology, 12:13, 8 2012.

[105] C. Shannon. The lattice theory of information. Transactions of the IRE Professional Group on Information Theory, 1(1):105–107, February 1953.

[106] Claude E Shannon. A mathematical theory of communication. The Bell system technical journal, 27(3):379–423, 1948.

[107] Wanqiang Sheng, Martin W LaFleur, Thao H Nguyen, Sujun Chen, Ankur Chakravarthy, Jake Ryan Conway, Ying Li, Hao Chen, Henry Yang, Pang-Hung Hsu, Eliezer M Van Allen, Gordon J Freeman, Daniel D De Carvalho, Housheng Hansen He, Arlene H Sharpe, and Yang Shi. LSD1 Ablation Stimulates Anti-tumor Immunity and Enables Checkpoint Blockade. Cell, 174(3):549–563, 7 2018.

[108] Arndt F Siekmann, Nathan D Lawson, A F Siekmann, and N D Lawson. Extra View Notch Signalling and the Regulation of Angiogenesis Addendum to: Notch Signalling Limits Angiogenic Cell Behavior in Developing Zebrafish Arteries. Technical Report 2, 2007.

[109] Mengjia Song, Xinfeng Chen, Liping Wang, and Yi Zhang. Future of anti-PD-1/PD-L1 applications: Combinations with other therapeutic

regimens. Chinese journal of cancer research = Chung-kuo yen cheng yen chiu, 30(2):157–172, 4 2018.

[110] Nicola Soranzo, Ginestra Bianconi, and Claudio Altafini. Comparing association network algorithms for reverse engineering of large-scale gene regulatory networks: synthetic versus real data. Bioinformatics, 23(13):1640–1647, 2007.

[111] Damian Szklarczyk, Andrea Franceschini, Stefan Wyder, Kristoffer Forslund, Davide Heller, Jaime Huerta-Cepas, Milan Simonovic, Alexander Roth, Alberto Santos, Kalliopi P Tsafou, et al. String v10: protein–protein interaction networks, integrated over the tree of life. Nucleic acids research, 43(D1):D447–D452, 2015.

[112] Tuomas Tammela, Georgia Zarkada, Elisabet Wallgard, Aino Murtomäki, Steven Suchting, Maria Wirzenius, Marika Waltari, Mats Hellström, Tibor Schomber, Reetta Peltonen, Catarina Freitas, Antonio Duarte, Helena Isoniemi, Pirjo Laakkonen, Gerhard Christofori, Seppo Ylä-Herttuala, Masabumi Shibuya, Bronislaw Pytowski, Anne Eichmann, Christer Betsholtz, and Kari Alitalo. Blocking VEGFR-3 suppresses angiogenic sprouting and vascular network formation. Nature, 454(7204):656–660, 7 2008.

[113] Andrew E Teschendorff, Christopher RS Banerji, Simone Severini, Reimer Kuehn, and Peter Sollich. Increased signaling entropy in cancer requires the scale-free property of proteininteraction networks. Scientific reports, 5(1):1–9, 2015.

[114] Andrew E Teschendorff and Tariq Enver. Single-cell entropy for accurate

estimation of differentiation potency from a cell's transcriptome. Nature communications, 8(1):1–15, 2017.

[115] Andrew E Teschendorff, Peter Sollich, and Reimer Kuehn. Signalling entropy: A novel network-theoretical framework for systems analysis and interpretation of functional omic data. Methods, 67(3):282–293, 2014.

[116] Nicholas Timme, Wesley Alford, Benjamin Flecker, and John M Beggs. Synergy, redundancy, and multivariate information measures: an experimentalist's perspective. Journal of computational neuroscience, 36(2):119–140, 2014.

[117] Hu Kuo Ting. On the amount of information. Theory of Probability & Its Applications, 7(4):439–447, 1962.

[118] Alejandro F Villaverde, John Ross, Federico Morán, and Julio R Banga. Mider: network inference with mutual information distance and entropy reduction. PloS one, 9(5):e96732, 2014.

[119] Chi-Chung Wang, Meng-Feng Tsai, Ting-Hao Dai, Tse-Ming Hong, Wing-Kai Chan, Jeremy JW Chen, and Pan-Chyr Yang. Synergistic activation of the tumor suppressor, hlj1, by the transcription factors yy1 and activator protein 1. Cancer research, 67(10):4816–4826, 2007.

[120] Haixin Wang, Lijun Qian, and Edward Dougherty. Modeling genetic regulatory networks by sigmoidal functions: A joint genetic algorithm and kalman filtering approach. In Third International Conference on Natural Computation (ICNC 2007). IEEE, 2007.

[121] Yin Wang, Rudong Li, Chunguang Ji, Shuliang Shi, Yufan Cheng, Hong Sun, and Yixue Li. Quantitative dynamic modelling of the gene reg-

ulatory network controlling adipogenesis. PLoS ONE, 9(10):e110563, October 2014.

[122] John Watkinson. Synergistic Associations in Systems Biology. Columbia University, 2011.

[123] John Watkinson, Kuo-ching Liang, Xiadong Wang, Tian Zheng, and Dimitris Anastassiou. Inference of regulatory gene interactions from expression data using three-way mutual information. Annals of the New York Academy of Sciences, 1158(1):302–313, 2009.

[124] John Watkinson, Xiaodong Wang, Tian Zheng, and Dimitris Anastassiou. Identification of gene interactions associated with disease from gene expression data using synergy networks. BMC systems biology, 2(1):1–16, 2008.

[125] Paul L. Williams and Randall D. Beer. Nonnegative decomposition of multivariate information. CoRR, abs/1004.2515, 2010.

[126] Dominik Wodarz and Natalia L Komarova. Dynamics of Cancer. WORLD SCIENTIFIC, 6 2014.

[127] Seng-Ryong Woo, Mercedes B Fuertes, Leticia Corrales, Stefani Spranger, Michael J Furdyna, Michael Y K Leung, Ryan Duggan, Ying Wang, Glen N Barber, Katherine A Fitzgerald, Maria-Luisa Alegre, and Thomas F Gajewski. STING-Dependent Cytosolic DNA Sensing Mediates Innate Immune Recognition of Immunogenic Tumors. Immunity, 41(5):830–842, 2014.

[128] Asia Wyatt. Mathematical Models of Acute and Chronic Immunology. PhD thesis, University of Maryland, 2019.

[129] Yiyi Yan, Anagha Bangalore Kumar, Heidi Finnes, Svetomir N Markovic, Sean Park, Roxana S Dronca, and Haidong Dong. Combining Immune Checkpoint Inhibitors With Conventional Cancer Therapy. Frontiers in immunology, 9:1739, 2018.

[130] Guan-Jun Yang, Pui-Man Lei, Suk-Yu Wong, Dik-Lung Ma, and Chung-Hang Leung. Pharmacological Inhibition of LSD1 for Cancer Treatment. Molecules (Basel, Switzerland), 23(12), 12 2018.

[131] Congxiu Ye, David Brand, and Song G. Zheng. Targeting IL-2: an unexpected effect in treating immunological diseases, 12 2018.

[132] Jaeyong Yee, Min-Seok Kwon, Taesung Park, and Mira Park. A modified entropy-based approach for identifying gene-gene interactions in case-control study. PloS one, 8(7):e69321, 2013.

[133] Marina Zafranskaya, Patrick Oschmann, Rosel Engel, Andreas Weishaupt, Johannes M van Noort, Hassan Jomaa, and Matthias Eberl. Interferon-beta therapy reduces CD4+ and CD8+ T-cell reactivity in multiple sclerosis. Immunology, 121(1):29–39, 5 2007.

[134] Xiujun Zhang, Juan Zhao, Jin-Kao Hao, Xing-Ming Zhao, and Luonan Chen. Conditional mutual inclusive information enables accurate quantification of associations in gene regulatory networks. Nucleic acids research, 43(5):e31–e31, 2015.

[135] Wentao Zhao, Erchin Serpedin, and Edward R Dougherty. Inferring connectivity of genetic regulatory networks using information-theoretic criteria. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 5(2):262–274, 2008.

[136] Huan Zheng, Bo Jin, Sarah E Henrickson, Alan S Perelson, U H von Andrian, and Arup K Chakraborty. How Antigen Quantity and Quality Determine T-Cell Decisions in Lymphoid Tissue. Molecular and Cellular Biology, 28(12):4040–4051, 2008.