# Cyclic and Multilevel Causation in Evolutionary Processes

**Jonathan Warrell · Mark Gerstein**

**Abstract** Many models of evolution are implicitly causal processes. Features such as causal feedback between evolutionary variables and evolutionary processes acting at multiple levels, though, mean that conventional causal models miss important phenomena. We develop here a general theoretical framework for analyzing evolutionary processes drawing on recent approaches to causal modeling developed in the machine-learning literature, which have extended Pearl's 'do'-calculus to incorporate cyclic causal interactions and multilevel causation. We also develop information-theoretic notions necessary to analyze causal information dynamics in our framework, introducing a causal generalization of the Partial Information Decomposition framework. We show how our causal framework helps to clarify conceptual issues in the contexts of complex trait analysis and cancer genetics, including assigning variation in an observed trait to genetic, epigenetic and environmental sources in the presence of epigenetic and environmental feedback processes, and variation in fitness to mutation processes in cancer using a multilevel causal model respectively, as well as relating causally-induced to observed variation in these variables via information theoretic bounds. In the process, we introduce a general class of multilevel causal evolutionary processes which connect evolutionary processes at multiple levels via coarse-graining relationships. Further, we show how a range of 'fitness models' can be formulated in our framework, as well as a causal analog of Price's equation (generalizing the probabilistic 'Rice equation'), clarifying the relationships between realized/probabilistic fitness and direct/indirect selection. Finally, we consider the potential relevance of our

J. Warrell
Program in Computational Biology and Bioinformatics,
Yale University, New Haven, CT, 06520, USA.
E-mail: jonathan.warrell@yale.edu

M. Gerstein
Program in Computational Biology and Bioinformatics,
Yale University, New Haven, CT, 06520, USA.
E-mail: mark.gerstein@yale.edu

framework to foundational issues in biology and evolution, including supervenience, multilevel selection and individuality. Particularly, we argue that our class of multilevel causal evolutionary processes, in conjunction with a minimum description length principle, provides a framework in which identification of multiple levels of selection may be addressed as a model selection problem.

**Keywords** Causality · Information Theory · Multilevel selection · Price's equation · Supervenience

## 1 Introduction

Causality is typically invoked in accounts of evolutionary processes. For instance, for a variant to be subject to direct selection, it is necessary that it has a causal impact on fitness. The role of causality is made explicit in axiomatic accounts of evolution [31]. Further, the formal framework of Pearl's 'do'-calculus [28] has been used explicitly in analyzing Mendelian Randomization [21], the relationship between kin and multilevel selection [26], and information derived from genetic and epigenetic sources in gene expression [12]. A number of features of evolutionary processes however limit the potential for direct formalization in the 'do'-calculus framework, which requires causal relationships to be specified by a directed acyclic graph (DAG), and cannot represent causal processes at multiple levels. In contrast, cyclical causal interactions are ubiquitous in natural processes, for instance in regulatory and signaling networks which lead to high levels of epistasis in the genotype-phenotype map [36]. Further examples of cyclical causal interactions arise through environmental feedback, both in the generation of traits, leading to an extended genotype-environment-phenotype map [15], and across generations in the form of niche construction [20]. Hierarchy is also ubiquitous in evolution, and many phenomena, such as multicellularity and eusociality, seem to require a multilevel selection framework for analysis, implicitly invoking causal processes at multiple levels [25]. Such a framework would also seem necessary in analyzing major transitions in evolution [6].

A number of frameworks have been proposed in the machine-learning literature for extending Pearl's 'do'-calculus to allow for cyclic causal interactions. These include stochastic models with discrete variables [16], and deterministic [23] and stochastic [32] models with continuous variables. Further, approaches have been introduced for analyzing causal processes at multiple levels using the 'do'-calculus [7,32]. In [32], both of these phenomena are related through the notion of a *transformation*, which is a mapping between causal models which preserves causal structure. Coarse-graining is a particular kind of transformation, special cases of which involve mapping a causal model over micro-level variables into one over macro-level variables, and mapping a directed causal model which is extended across time into a cyclical model which summarizes its possible equilibrium states (subject to interventions).

In addition, the 'do'-calculus has been combined with information theory in order to define notions of information specifically relevant to causal models,

such as *information flow* [3], *effective information* [14], *causal specificity* [12] and *causal strength* [18]. Although not explicitly cast in causal terms, there has also been much interest in defining non-negative multivariate decompositions of the mutual information between a dependent variable and a set of independent variables, which may be collectively described as types of *Partial Information Decomposition* (PID) [37, 4, 12]. Such definitions however can only be applied in causal models with a DAG structure, leaving open the question of how causal information should be defined and decomposed in a system with cyclic interactions.

Motivated by the above, we propose a general causal framework for formulating models of evolutionary processes which allows for cyclic interactions between evolutionary variables and multiple causal levels, drawing on the transformation framework of [32] as described above. Further, we propose a causal generalization of the Partial Information Decomposition (Causal Information Decomposition, or CID) appropriate for such cyclic causal models, and show that our definition has a number of desirable properties and can be related to previous measures of causal information. We analyze a number of specific evolutionary models within our framework, including first, a model with epigenetics and environmental feedback, and second, a model of multilevel selection which we apply to the particular cases of group selection and selection between mutational processes in cancer. We analyze the CID in the context of both models, and demonstrate in both cases conditions under which bounds can be derived between components of the CID and components of the PID derived from the observed distribution. Finally, we discuss the causal interpretation of Price's equation and related results in our causal framework. In general, our analysis is intended both to help clarify conceptual issues regarding the role of causation in the models analyzed, as well as to aid in the interpretation of data when the assumptions of these (or similar) models are adopted, via the bounds introduced.

We begin in Sec. 2 by introducing the model of discrete cyclic causal systems which will form the basis of all subsequent analyses, and introduce the CID in this context. Sec. 3 then outlines our general framework for causal evolutionary processes. Sec. 4 and Sec. 5 analyze specific models of epigenetics with environmental feedback and multilevel selection within this framework respectively, and Sec. 6 provides a causal interpretation of Price's equation and related results. Sec. 7 then concludes with a discussion, including the potential relevance of our framework to foundational issues in the philosophy of biology and evolution.

## 2 Cyclic and Multilevel Causality in Biology

We begin by outlining and motivating some of the basic concepts that will be used to develop our framework. We do so here in an informal way: technical definitions and proofs are given in Appendix A.

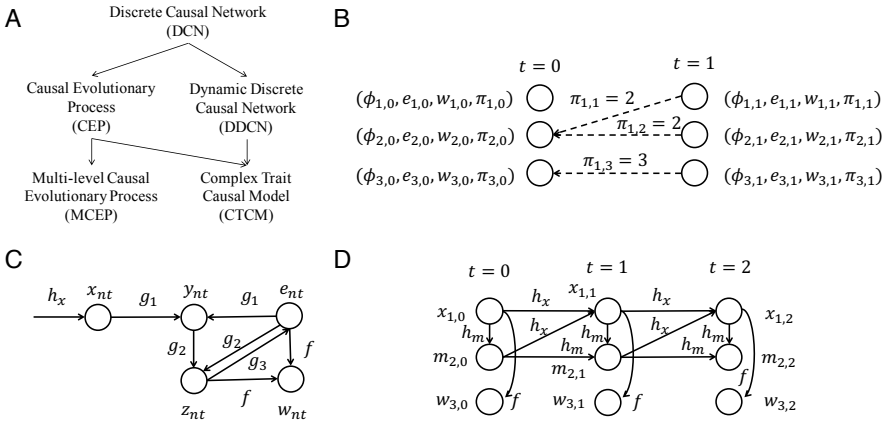


**Fig. 1** Summary of models. (A) shows the relationships between the main models defined in the paper, where the arrows point from larger to smaller model classes, the latter being a special case (subset) of the former. (B) shows a schematic of a Causal Evolutionary Process, showing a population of size $N = 3$ at two time steps (nodes representing all variables associated with an individual at time $t$). $\phi$, $e$, $w$ and $\pi$ represent phenotype, environment, fitness and parental map respectively, where the latter is also represented explicitly by the dotted arrows, which connect each individual at $t = 1$ to its parent at $t = 0$. (C) and (D) illustrate the CTCM* and $\text{MCEP}_{\text{mut-proc}}$ models described in Secs. 4 and 5 of the paper respectively. The first is a model of a complex trait, with genetics ($X$), epigenetics ($Y$) and feedback between behavior ($Z$) and environment ($e$), while the second is a model of multilevel selection in cancer with genetics ($X$) and mutational processes ($m$). Solid arrows here represent the directed graphical structure of the underlying discrete causal network (the $Pa$ relation, which is distinct from the $\pi$ evolutionary variable in (B)). Nodes and edges are labeled with variable and kernel names respectively as defined in these models.

**Cyclic Causality.** The *do*-calculus provides a compelling formalization of the mathematical structure of causation [28]. However, a requirement of this framework is that causal relationships between variables must form a Directed Acyclic Graph (DAG). A DAG is any graph (a collection of nodes and edges) which contains no loops. For instance, if gene A regulates gene B, and gene B regulates genes C and D (for example, the genes are transcription factors, and regulatory relationships are established via promoter binding), variables corresponding to the expression levels of these genes can be arranged in a graph with no loops. Then, performing manipulations on gene B (*do*-operations) will affect the expression of genes C and D but not A. Pearl's calculus provides exact rules for deducing the distribution of all variables after an intervention, which will always return a valid distribution; formally, the graph is altered by cutting all incoming edges to the manipulated node, and the joint distribution is recalculated using a delta distribution to represent the intervention.

However, such well-ordered networks are the exception rather than the rule in gene regulatory networks (GRNs). For instance, we could add to the above network a feedback interaction by assuming that gene D regulates A. Here, although A causes B's expression locally, B also causes A's expression via D. In fact, in this case no harm is done assuming the system has a solution as a

whole, since any intervention will either split the cycle (genes A, B and D), or have no effect on it (gene C), and hence all distributions are well defined. Alternatively though, we could consider starting with the original graph, and in addition let C regulate D and D regulate C. Now, by intervening on B we are not guaranteed to find a solution for every intervention, even if the original system has a solution. Recent work has investigated the extension of Pearl's calculus to graphs with cycles (where the graph manipulations are identical to the acyclic case), characterizing the situations in which solutions exist, and allowing systems to be defined which have a restricted set of interventions allowed [16,23,32]. A particular example which has a clear solution is a deterministic network governed by linear differential equations: $\dot{x_i} = f_i(x_1, ..., x_N)$, where the $x_i$'s may be gene expression values for instance. Here, 'solutions' are taken to be the equilibrium points of the system, and these exist for any intervention provided the original system of equations $f_1, ...f_N$ is *contractive*, meaning that it maps a given region of state-space to one with a smaller volume with time [32]. Stochasticity can be added as long as the functions are contractive almost surely.

In these examples, we have considered feedback processes in gene regulatory networks. However, similar feedback processes can occur at many levels. For instance, consider a psychological trait, such as depression. This trait may be ultimately caused by numerous aspects of brain structure and gene expression patterns; however, since we have pharmacological interventions which can control the severity of depression, these can introduce a feedback process from the environment to the molecular layer. In general, the expression of any trait may be subject to environmental feedback in this way.

We now note some features about the above. First, we have phrased both the GRN and the environmental feedback example in terms of a process in time. One way to deal with such cyclic structures is to 'unroll' them over time, that is, consider a discretized set of time points, and repeat all variables at each time, while connecting a given variable, say a gene, to its regulatory parents at the previous time-step. This will automatically generate an acyclic causal graph. The cyclic causal system corresponding to this temporal process can be considered to be that formed by the equilibrium distributions (assuming they exist) after certain 'macro' interventions are applied, which fix a particular variable to a given value across all times. The cyclic system can thus be considered a *coarse-graining* of the temporal acyclic system. Not all cyclical causal systems can be formed this way (see [32]), but our emphasis will be on such systems, given their prevalence in biology (although the framework is agnostic to the system's origins). For convenience, in setting up our framework, we use a 'Discrete Causal Network' model (DCN, Appendix A; see also Fig. 1A for the relationships between all models of the paper), which allows cycles and variables which take discrete values, and is parameterized by a set of *probability kernels*, (one for each variable) specifying the conditional distribution of the variable on the values of its graphical parents. We further introduce dynamic-DCN and equilibrium-DCN models (Appendix A,

Def. 2.6), corresponding respectively to an underlying temporal process and coarse-grained equilibrium cyclic causal model as discussed above.

**Multilevel Causality.** A further limitation of straightforward applications of Peal's *do*-calculus is its seeming reliance on a single level of causal analysis. For instance, in analyzing the causes of an action, it seems appropriate to identify causes at multiple levels, such as nerves firing, muscles contracting, psychological beliefs and desires, past learning. Indeed, in analyzing group selection using causal graphs, a recent approach has distinguished between *causal* and *supervenient* relationships between variables, while maintaining a DAG graphical structure overall [26]. Recent approaches have formalized the idea that a causal system can be described at multiple levels, by introducing coarse-graining mappings between causal structures at different scales [32]. Such approaches capture the relation of *supervenience*, since multiple *fine-grained* interventions in one model may be mapped to the same intervention in another provided the causal structure is preserved by the mapping.

The possibility of multiple levels of causation is arguably of central importance in evolution. In particular, we argue that multilevel selection should be seen as a special case of multilevel causality, and introduce a 'Multilevel Causal Evolutionay Process' model (MCEP) as a general framework for analyzing such processes. In particular, *fitness* is treated as a causal variable at each level of the MCEP, allowing coarse-grained fitness to supervene on lower-level fitness values (and other evolutionary variables). A particular case we consider is cancer evolution. As has been recently demonstrated [1,8,33], tumors not only acquire particular sets of mutations (with positive, negative and neutral effects on growth) over their development, but also acquire prototypical *mutational processes*. These processes are caused by factors such as disruption of the DNA repair machinery or other cellular mechanisms such as DNA methylation, or environmental effects such as carcinogens, which cause particular mutations to become more prevalent depending on local sequence characteristics or chromosomal position for example. The fact that such processes introduce bias into the way variation is acquired in the tumor means that they can contribute towards the tumors' evolution. At a fine-grained level, it is the individual mutations themselves which are responsible for fitness variations among cells, and the mutational processes are simply a source of variation. We show however, that by considering a multilevel model, fitness across larger time-scales can be driven by a combination of individual mutations and mutational processes, potentially even primarily by the latter as suggested by recent results [8,33].

**Causal Transformations.** The models we develop in response to the above (cyclic and multilevel causation) both rely on the technical apparatus of a *transformation* between causal systems, as introduced in [32]. In general, this can be thought of as a structure preserving map between causal systems, or a *morphism* in a category theoretic sense [2], and is thus directly analogous to other kinds of morphism, such as homomophisms between groups, or other

algebraic structures. A causal transformation requires that variables and interventions in one causal system are mapped to those in another, while preserving all causal relationships in the first system as seen 'from the viewpoint' of the second. Like a group homomorphism, a transformation of causal systems is not necessarily one-to-one or onto, and so the mapping may embed the first causal system in the second, or map many variables in the first onto a single variable in the second. The transformations we consider typically correspond to the latter possibility, and thus can be seen as forms of *coarse-graining*. However, it is important to stress that transformations are not limited to coarse-graining relationships, and are a general mechanism for relating causal systems. We explicitly define the notion of transformation we need for the special case of discrete causal networks in Appendix A, Def. 2.2.

**Information in Causal Systems.** An advantage of framing evolutionary models in explicitly causal terms is that it becomes possible to make distinctions between different ways in which evolutionary variables may interact, which are difficult to make otherwise. For instance, it is common to trace variation in a particular trait (such as height) to genetic and environmental sources. With genetics, the (broadly justified) assumption is made that variation in the trait is in response to genetic variation, and thus intervention is not required to assess the causal impact (having controlled for confounders such as population structure). However, the situation is less clear when variation at other levels such as epigentics (transcriptomics, DNA methylation), or environmental factors are considered in relation to high-level trait variation (height, depression). Here, we would like to be able to trace variation to sources which may be involved in cyclic interactions. In general, interventions may be required to assess the causal impact of one variable on another, but it may also be possible to combine observations and assumptions to infer aspects of the causal structure.

For this reason, we also consider how to define a general notion of causal impact in cyclic causal systems, by generalizing the Partial Information Decomposition framework (PID, [37]), which cannot handle cyclic interactions, to a Causal Information Decomposition framework (Appendix A, Def. 2.4). We show that bounds may be derived in this framework that potentially allow direct causal relations between variables to be inferred from observational data by observing non-zero unique information, and differences in observed and causal information between variables to be predicted given assumptions about feedback and interference. We provide technical background for these bounds in Appendix A, Theorems 2.10 and 4.3, and draw connections with alternative definitions of causal impact and related bounds (Prop. 2.5), and summarize the implications as they apply to models of epigenetics and multilevel selection in the relevant sections (Th. 4.3 and Prop. 5.5). These results are intended both to motivate the application of models using PID and CID frameworks in analyzing data, and also contextualize the issues underlying existing approaches such as GWAS and TWAS, even when not couched explicitly in information-theoretic terms.

**Deterministic, Stochastic and Causal Models.** Finally, we wish to emphasize the intrinsic differences between mathematical structures underlying deterministic, stochastic (or probabilistic) and causal models. These kinds of models can be seen as strictly nested inside one another: deterministic models are simply stochactic models whose probabilities are all taken to be either 0 or 1, while stochastic models are causal models which are not subject to any interventions (i.e. subject to the null intervention). In this sense, causal models contain strictly more information than stochastic models, since they represent a *family of distributions* parameterized by all possible interventions, rather than a single distribution. Alternatively, we can say that causal models contain counterfactual as well as probabilistic information. As stressed in axiomatic accounts of evolution [31], we view causal structure as intrinsic to the definition of an evolutionary process, and thus causal models as the appropriate mathematical structure for a complete description of such a process. In Sec. 6, we briefly consider this viewpoint in relation Price's Equation and a related information-theoretic result based on the Kullback-Leiber divergence (which is a quasi-distance measure between probability distributions, here the trait distribution at two time-points), discussing their analogues in stochastic and causal models and stressing how the causal viewpoint offers a more complete picture (albeit, one implicit in other modeling frameworks).

## 3 Causal Evolutionary Processes

We now introduce a general model of a *causal evolutionary process*. In its general form, the model is a formalized 'phenotype-based theory' of evolution (see [31]), which is agnostic about underlying mechanisms. As argued in [31] (and as we will be elaborated in subsequent sections), such a perspective naturally embeds traditional population genetics models as a special case, since genotypes may be treated as special kinds of discrete phenotypes, while offering a more general viewpoint. For notational convenience, we introduce all definitions and examples below in the context of an asexual population of constant population size, although the model naturally generalizes to mating populations and varying population sizes. Fig. 1B illustrates the model definition below.

**Definition 3.1.** (Causal Evolutionary Process (CEP)): *A CEP is a Discrete Causal Network (DCN) over the variables $\phi_{nt}$, $e_{nt}$, $w_{nt}$, $\pi$ (all variables discretized, and $\pi_{nt} \in \{1...N\}$), representing the phenotype, environment, fitness and parent of the n'th individual in the population at time t respectively, where 'fitness' and 'parent' are to be understood in a structural sense to be defined, and $n \in \{1...N\}$, $t \in \{0...T\}$. We write $\phi_t$, $e_t$, $w_t$, $\pi_t$ for the collective settings of these variables at t, and for convenience use identical notation for names and values taken by random variables. Further, $\phi_{nt}$ and $e_{nt}$ may be viewed as a collection of sub-phenotypes and sub-environmental variables, in which case we write $\phi_{nst}$ and $e_{nst}$ for the value of sub-phenotype (resp. environment) s of*

*individual $n$ at time $t$. We set $Pa(\phi_0) = Pa(e_0) = Pa(\pi_0) = \{\}$ (noting that $Pa$ stands for the 'graphical parents' of a variable in the causal graph, while $\pi_{nt}$ is the evolutionary variable representing the parent of individual $n$ at time $t$, whose values are indices of individuals at $t-1$). For all other variables, we set $Pa(w_t) = \{\phi_t, e_t\}$, $Pa(\phi_t) = \{\phi_{t-1}, e_t, \pi_t\}$, $Pa(e_t) = \{e_{t-1}, \phi_t, \pi_t\}$, $Pa(\pi_t) = \{w_{t-1}\}$ . A model is specified by defining the following kernel forms; here, we write $(:=)$ for an optional factorization, $\dot{\prod}$ to represent products over kernels, and set the underlying variables of the DCN to correspond to the lowest-level factorization consistent with the model kernels (further details on these notations are given in Appendix E):*

Fitness kernel:

$$\begin{aligned} K(w_t|Pa(w_t)) &= f(w_t|\phi_t, e_t) \\ &(:=) \dot{\prod_n} f_n(w_{nt}|\phi_t, e_t), \end{aligned} \tag{1}$$

Heritability kernel:

$$\begin{aligned} K(\phi_t|Pa(\phi_t)) &= h(\phi_t|\phi_{t-1}, e_t, \pi_t) \\ &(:=) \dot{\prod_{ns}} h_{ns}(\phi_{nst}|\phi_{(\pi_t(n),t-1)}, e_{nt}, \phi_{n\bar{s}t}), \end{aligned} \tag{2}$$

Structure kernel:

$$\begin{aligned} K(\pi_t|Pa(\pi_t)) &= i(\pi_t|w_{t-1}) \\ &(:=) \dot{\prod_n} i_n(\pi_{nt}|w_{(n,t-1)}). \end{aligned} \tag{3}$$

*For the environmental variables, we have the following alternative kernel forms:*

$$\begin{aligned} &K(e_t|Pa(e_t)) \\ &\quad = K_e(e_t|e_{t-1}, \phi_t, \pi_t) \\ &\quad (:=) \begin{cases} \dot{\prod}_n P_e(e_{nt}) & \text{(a)} \\ \dot{\prod}_{ns} K_e(e_{nst}|\phi_{nt}, e_{n\bar{s}t}) & \text{(b)} \\ \dot{\prod}_{ns} \hbar_{ns}(e_{nst}|e_{(\pi_t(n),t-1)}, \phi_{nt}, e_{n\bar{s}t}). & \text{(c)} \end{cases} \end{aligned} \tag{4}$$

*In Eq. 4, we refer to factorizations (a) and (b) as independent and interactive environmental kernels respectively, and (c) as an environmental heritability kernel (represented by the symbol $\hbar$). Additionally, kernels must be given over the remaining variables for which $Pa(.) = \{\}$ to completely specify a CEP model.*

The interaction of the fitness and structure kernels ($f$ and $i$ resp.) give rise to different possible fitness models. We summarize some of these possibilities below:

**Definition 3.2.** (Fitness models): *We define the following CEP fitness models:* Classical fitness representation:

$$f(w_t|\phi_t, e_t) = P(\{w_1, ..., w_N\}|\phi_t, e_t)$$
$$i(\pi_t|w_{t-1}) \propto \prod_n [(\sum_m [\pi_t(m) = n]) = w_n], \quad (5)$$

Multinomial model ($\omega_{nt} = w_{nt}/(\sum_n w_{nt})$ denotes normalized fitness):

$$f(w_t|\phi_t, e_t) = \prod_n P(w_{nt}|\phi_t, e_t)$$
$$i(\pi_t|w_{t-1}) = \text{Mult}(\{(\sum_m [\pi_t(m) = n])|n = 1...N\}|\{\omega_{(n=1...N,t-1)}\}), \quad (6)$$

Moran model:

$$f(w_t|\phi_t, e_t) = \prod_n P(w_{nt}|\phi_t, e_t)$$
$$i(\pi_t|w_{t-1}) \propto \begin{cases} w_{(n^*,t-1)} & \text{if } \pi_t \in S_{n^*} \\ 0 & \text{otherwise.} \end{cases} \quad (7)$$

*where $S_{n^*}$ is the set of all vectors in $N^N$ containing exactly two values $n^*$, and all other values appear at most once.*

In the classical fitness representation, the variable $w_{nt}$ directly represents the number of descendants of individual $n$ at time $t$ in the following generation, and the structure kernel $i$ simply ensures the parent map $\pi$ is consistent with these values. In the multinomial model, the values $w_{nt}$ determine the relative fitnesses of individuals at time $t$ and the actual numbers of descendants are determined by multinomial sampling, implemented by $i$ (as in the Wright-Fisher model with selection [9]). In contrast, in the Moran model, the $w_{nt}$'s determine the probability that an individual is chosen to reproduce, and $i$ implements the constraint that only one individual reproduces and dies per generation, with the latter being chosen uniformly. Although both Multinomial and Moran models could be represented in the classical fitness representation, this would be at the expense of using a non-factorized form of the $f$ kernel; hence we argue that the representations in Eqs. 6 and 7 are more natural parameterizations of these models (so that, in general, 'fitness' is not exclusively interpreted as the number of offspring at time $t+1$, but rather a set of sufficient statistics for generating the parental map at time $t+1$). Further, we note that, as in the case of the Moran model, the time steps $t$ need not correspond to discrete generations.

Finally, we define a *transformation* between CEPs:

**Definition 3.3.** (Transformation between CEPs): A *transformation* between CEPs is defined as a transformation between their underlying DCNs in the sense of Def. 2.2 (Appendix A).

In relation to Def. 3.3, we note that a transformation between CEPs need only preserve the *causal* structure; hence, it may (for instance) map environmental onto phenotypic variables, or a large population onto a small population by merging individuals. All that is necessary is that the resulting causal structure may be interpreted as an evolutionary process *in some way*. We shall give examples in the following sections of transformations with characteristics such as above.

## 4 Model 1: Genetics, Epigenetics and Environmental Feedback

The CEP model as introduced in Sec. 3 does not include a model of genetics. However, as noted, the genotype may be regarded as a special type of discrete phenotype, and the process of genetic transmission with mutation can be naturally modeled in the heritability kernel. Here, we describe a type of CEP which includes both genetics and epigenetics, along with potential effects from and impacts on the environment (*environmental feedback*), for instance via behavior or drugs used in treating diseases. The model thus formalizes a *gene-environment-phenotype* map (G-E-P map) of the kind described informally in [15] (for simplicity, we refer to any 'intermediate phenotype' as epigenetic, including for instance the transcriptome). Our purpose is to provide a general model appropriate for analyzing the causal factors underlying complex traits, such as psychiatric disorders. As we describe, using this model, the causal information decomposition (CID) described in Sec. 2 provides a principled framework for breaking down the variation in a complex trait due to genetic, epigenetic and environmental factors; the model is more general than other models with similar goals, for instance TWAS (see [38]), since it aims to model both genetic and other causes of a trait, allowing feedback with the environment at the epigenetic level, and uses an information theoretic framework to decompose the variation and hence is appropriate in the context of arbitrary (non-linear) dependencies.

We first define a general CEP model with the above characteristics:

**Definition 4.1.** (Complex Trait Causal Model (CTCM)): *We define a CTCM as a CEP with the following special structure. In terms of phenotypes, we require three sub-phenotypes which we denote $X, Y$ and $Z$, hence $\phi_{nt} = \{x_{nt}, y_{nt}, z_{nt}\}$, which represent genotype, epigenome (including transcriptome), and observed trait(s) respectively. Environmental variables are referred to collectively as $e$. Further, we use the factorized form of the heritability kernel in Eq. 2, and require the following special forms for the sub-kernels:*

$$h_n(\phi_{nt}|\phi_{(n,t-1)}, e_{nt}, \pi_t) := h_x(x_{nt}|x_{(\pi_t(n),t-1)}) \cdot g_1(y_{nt}|x_{nt}, e_{nt}) \cdot g_2(z_{nt}|y_{nt}, e_{nt}), \tag{8}$$

*where $g_1$ and $g_2$ are referred to collectively as the* genotype-phenotype map*, and independently as the* genetic-epigenetic *and* epigenetic-observed *kernels*

*respectively, while $h_x$ is referred to as the* genetic transmission *kernel. Further, the CTCM uses an environmental kernel having either an independent factorization, or an interactive form (Eq. 4 (a) and (b) resp.); the latter takes the form:*

$$K_e(e_{nt}|\phi_{nt}) = g_3(e_{nt}|z_{nt}), \tag{9}$$

*and $g_1$, $g_2$ and $g_3$ are referred to collectively (when all present) as the* genotype-environment-phenotype map*. [We note that 'map' here and following Eq. 8 refers in general to a stochastic map.]*

We next define a special class of CTCMs which embed a DDCM over the $\phi_{nt}$ and $e_{nt}$ variables at each time-point; hence, we model the genetic, epigenetic, observed trait and environmental interactions by an embedded dynamic causal process. We can the coarse-grain this process to a cyclical CTCM over these variables at equilibrium (see Fig. 1C for a related schematic). For conciseness, the full definition of the CTCM* is given in Appendix B, Def. 4.2.

**Definition 4.2.** (CTCM with embeded DDCM (CTCM*)): See Appendix B.

The CTCM* model gives us a convenient way of decomposing the variation/information in a trait into components which depend on unique, redundant and synergistic combinations of genetic, epigenetic and environmental factors. We propose that, for a population at time $t > 0$, this is achieved by selecting an arbitrary individual $n$, and calculating the backward causal information decomposition $CID(S \rightarrow \underline{Z_{nt}})$, where $S \subset \{X_{nt}, Y_{nt}, e_{nt}\}$, in the eq-CTCM* associated with the original CTCM* (assuming it exists, and that the structure kernel $i$ is invariant to permutations of the population indices). Since we specify in Def. 4.2 that the embedded DDCM kernels have the self-separability property, Theorem 2.9 implies that this is a strict decomposition of the variation in $Z_{nt}$, i.e. $CID(S \rightarrow \underline{Z_{nt}}) \leq H(Z_{nt})$. Further, if $g_3$ is an independent rather than an interactive environmental kernel, Theorem 2.10 implies that we can lower-bound the forward-CIDs $CID(\underline{X_{nt}} \rightarrow Z_{nt})$, $CID(\underline{e_{nt}} \rightarrow Z_{nt})$, and $CID(\underline{Y_{nt}} \rightarrow \{Z_{nt}, X_{nt}, e_{nt}\})$, using the observed unique information between each variable and $Z$ (i.e. the observed unique information is predictive of the consequences with respect to an observed trait of performing manipulations on each variable). We note that the PID of the observed distribution in this case is identical to the backward-CID as above.

In the case that $g_3$ is an interactive kernel, we have environmental feedback from the observed trait to the epigenetic levels, making it harder to relate the observed phenotype distribution to the proposed causal decomposition. However, we can outline a number of possible relationships. For this purpose, we introduce an alternative representation of a CTCM*. We consider that all transition kernels share a common parameter $\alpha$ from the self-separable representation, Eq. 28 (with $K_1$ set to the identity). Since Eq. 28 has the form of a mixture distribution, an equivalent representation of a CTCM* is formed by introducing latent variables, $C_Y(\tau)$, $C_Z(\tau)$, $C_e(\tau)$, which are Bernoulli variables (or collections of Bernoulli variables if $Y, Z$ or $e$ are factorized) with

mean $1-\alpha$. If $C_V(\tau) = 0$, variable $V \in \{Y, Z, e\}$ does not update at time-step $\tau$, otherwise $V$ updates according to the $K_2$ component of the self-separable representation in Eq. 28. If $\alpha$ is set large enough with respect to the number of variables, we can ensure that with probability $1-\epsilon$, with $\epsilon$ arbitrarily small, $\sum_V C_V(\tau) \leq 1$, i.e. at most one variable updates at a given time-step. Conceptually, we can view an increase in $\alpha$ as effectively a reduction in the duration of the time-step $\tau$. We also introduce $C^*(\tau) \in \{X, Y, Z, e, \emptyset\}$, writing $C^*(\tau) = V$ when $V$ is the most recent variable for which $C_V(\tau' < \tau) = 1$ assuming $V$ is unique, and $C^*(\tau) = \emptyset$ when $V$ is not unique (hence $P(C^*(\tau) = \emptyset) = \epsilon$). We can then make the following observation (see Appendix B for the proof):

**Theorem 4.3.** (Backward-CID bounds): *For a CTCM* represented as above with latent factors $C$, and associated eq-CTCM*, where $II$ is the interaction information, $S \subset \{X, Y, e\}$, $V \in \{X, Y, Z, e\}$, and $\overline{(.)}_V$ denotes the mean over values of $V$, in the limit $\alpha \to 1$ we have that:*

$$[II(S; Z; C^*) \leq 0] \wedge [CID(S \to \underline{Z}) \leq \overline{(CID(S : Z|Q_V))}_V] \implies PID(S : Z) \geq CID(S \to \underline{Z}), \tag{10}$$

*and similarly:*

$$[II(S; Z; C^*) \geq 0] \wedge [CID(S \to \underline{Z}) \geq \overline{(CID(S : Z|Q_V))}_V] \implies PID(S : Z) \leq CID(S \to \underline{Z}). \tag{11}$$

*where all $II$, $CID$ and $PID$ quantities are evaluated in the eq-CTCM* model (at a given $n$ and $t$, where $C^*$ is treated as an additional phenotype). Further, $Q_V = P_{eq}(\neg V)K_2^V(V|\neg V)$ (unrelated to the notation $Q_{X_A}$ used in Def. 2.4) with $K_2^V$ the second component of $V$'s kernel, as in Eq. 28, and we assume $Y, Z$ and $e$ are not factorized. For the case that $Y, Z$ or $e$ are factorized, $S$ and $V$ are subsets and elements of the sets of relevant factorized variables respectively, and Eqs. 10 and 11 hold identically.*

*Proof* See Appendix B.

Theorem 4.3 shows that we can identify certain situations in which the observed $PID$ components at equilibrium consistently under or over-estimate the equivalent components of the $CID$. The LHS of Eqs. 10 and 11 each contain two conditions, the first depending on the sign of an Interaction Information (II) term, and the second comparing two $CID$ terms. Broadly, the latter condition implies that if the feed-forward interaction between $S$ and $Z$ is strong compared to any feedback interactions, the $PID$ will tend to underestimate the $CID$ (Eq. 11), and vice-versa if the feedback is stronger (Eq. 10). However, the first condition makes this dependent on the type of feedback interactions present: if these tend to interfere with the feedforward interactions so the the net effect is to reduce the mutual information between $S$ and $Z$ (a synergistic interaction), the II term will be negative as in Eq. 11, while non-interfering

interactions will tend to increase the mutual information as in Eq. 10. Potentially, the situation in Eq. 11 may apply when $Z$ is a disease trait, and $S$ is the transcriptome in a relevant tissue, where the feedforward interaction is strong (the trait is strongly determined by $S$), and the feedback (in terms of treatment) reduces the severity of the disease by directly interfering with the underlying mechanisms. In contrast, the situation in Eq. 10 will apply if the feedforward effects are weak, and there is not a strong interference with feedback interactions at the level of $S$ (for instance, a disease treatment which targets symptoms in a different tissue, inducing variation in $S$ orthogonal to the causal factors for the disease).

## 5 Model 2: Multilevel Selection

Multilevel selection has been identified as an important component in a number of evolutionary contexts, such as eusociality in insects [24], bacterial plasmid evolution [27], and group selection [35]. It has also been proposed that multilevel selection is a driving force behind major evolutionary transitions, such as the transition to multicellularity [25][6]. Here, we propose a basic definition of a multilevel causal evolutionary process (MCEP) in the framework introduced above, which naturally connects the notion of evolution occurring at multiple levels to coarse-graining transformations in the sense of Defs. 2.2 and 3.3. We then show how two types of multilevel evolutionary process are special cases of our model (group selection, and selection acting on mutational processes in cancer). Our general definition takes the following form:

**Definition 5.1.** (Multilevel Causal Evolutionary Process (MCEP)): *An* M-level MCEP *is a collection of causal evolutionary processes,* $E_1, E_2, ..., E_M$*, such that each pair of adjacent processes forms a* 2-level MCEP *in the following sense. CEPs* $E$ *and* $F$ *form a 2-level CEP iff there exists a transformation from* $E$ *to* $F$ *(denoted* $(\tau, \omega)$*) in the sense of Def. 3.3, along with a partial map* $\mu$ *of time-points in* $F$ *to time-points in* $E$ *(which may depend on* $(\phi, e, w, \pi)$ *across all variables in* $E$*), and the following conditions apply.* **(1)** *We have that* $|F|_{N(t)} \leq |E|_{N(\mu(t))}$ $\forall t$*, and* $|F|_T \leq |E|_T$*, where we write* $|A|_{N(t)}$ *for the 'actual population size' of process* $A$ *at time* $t$*, and* $|A|_T$ *for the 'actual number of time-points' in process* $A$*. Each of these may be different from the values of* $N$ *and* $T$ *in* $A$*, since we will allow a* null *phenotype value to be declared in each CEP (whose parents are arbitrary, and whose offspring are all* null*): any individuals having* $\phi_{nt} =$ null *will not count towards* $|A|_{N(t)}$*, and time-points for which all individuals are* null *do not count towards* $|A|_T$*, these being the only time-points excluded from the domain of* $\mu$*; further, time-points beyond* $\max_t\{t|\exists t'\ \mu(t') = t\}$ *do not count towards* $|E|_T$*.* **(2)** *We require at least one of the inequalities in (1) to be strict.* **(3)** *We require that for any time-point* $t$ *in* $E$ *for which* $\mu(t') = t$*, the projection of the map* $\tau$ *onto* $\phi_{t'}$ *in* $F$ *is not independent of* $\phi_{nt}$ *in* $E$ *for any (non-null) individual* $n$ *(i.e. it does not take the same value for all settings of* $\phi_{nt}$ *given a joint setting of all other variables*

*in $E$), so that no individual's phenotype is entirely 'projected out' at these time-points by $\tau$.*

We now show how the group-selection model of [35] can be represented as an MCEP:

**Example 5.2.** (Group Selection model ($\text{MCEP}_{\text{group}}$)): *We fix an $N$ and $T$ for process $E$. For all individuals in $E$, $\phi_{nt} \in \{C, D, \text{null}\}$, where $C$ and $D$ represent cooperators and defectors respectively. $e_{nt} \in \{1...M\}$ represents the group membership of an individual (where $M$ is the maximum number of groups), and fitness $w_{nt}$ is determined by the expected pay-off for an individual when interacting with other members of the same group according to a fixed game matrix (see [35]). A maximum group-size is fixed at $N_G$, such that $N = N_G M$. The heritability and environmental kernels ($h$ and $\hbar$) enforce strict inheritance of phenotype and group membership (with the exception noted below), while structure kernel $i$ implements Moran dynamics (Eq. 7), so long as doing so will not allow a group to exceed $N_G$; otherwise, with probability $(1-q)$ a random individual from the same group dies, and with probability $q$ the group divides (implemented by the environmental kernel as a random partition) while all members of another uniformly chosen group die. Since the population number may fluctuate below $N$,* null *values are used to 'pad' the population as required.*

*For process $F$, we set the population size to be $M$ and the number of time-steps to be $T$. We map $t = 1$ in $F$ to the first time-point in $E$, and subsequent time-points in $F$ to the times at which the 1st, 2nd, 3rd... group divisions occurred in $E$. The 'individuals' in $F$ correspond to the groups in $E$; hence, we let $\phi_{nt} = (\nu_n, C_n)$, where $\nu_n$ is the number of individuals in group $n$ at time $\mu(t)$ in $E$, and $C_n$ is the proportion of cooperators in group $n$ (we set $\phi_{nt} =$* null *if $\nu_n = 0$). In $F$, $e_{nt} = \{\}$. We can naturally specify the parental map $\pi_{nt}$ on $F$ by mapping an individual at $t$ to the group at $t-1$ when the groups they correspond to at $\mu(t)$ and $\mu(t-1)$ in $E$ are either the same or split from one another. The fitness model can be specified by letting $w_{nt}$ be the probability that group $n$ will split first (in the context of all other groups). The structure kernel $i$ then simply needs to implement Moran dynamics, unless the number of groups is less than $N_G$, in which case a* null *group is chosen for replacement in place of uniform sampling. The heritability kernel in $F$ then needs to implement a conditional distribution over $\phi_t$ corresponding to the joint distribution over the sizes and cooperator prevalences in group at $t$, given that a particular group from the previous generation divided first (we note that since, in general, dependencies will be induced between the group phenotypes by the intervening dynamics in $E$, the unfactorized version of the heritability kernel in Eq. 2 must be used). Time-points in $F$ following the last time-point for which $\mu(t)$ is assigned are padded with* null *phenotype values (clearly, $|F|_T < T$, since at most $T-1$ group divisions can occur in $E$).*

*By construction, the pair of CEPs $E$ and $F$ above form a 2-level $MCEP$. For the required transformation, we simply take $\tau$ to map a configuration in $E$ to the configuration in $F$ which consistently represents the sizes and pro-*

*portions of cooperators in each group a at the times* $\mu(0), \mu(1)...$ *by* $(\nu_{a0}, C_{a0})$, $(\nu_{a1}, C_{a1}), ....$ *For the mapping* $\omega$ *we must be careful to restrict the interventions allowed on* $E$ *to those which fix all phenotypes and environments at a given time* $t$*. With this restriction, these can be mapped many-to-one onto interventions in* $F$ *which match the induced group characteristics. The first two conditions in Def. 5.1 are satisfied by construction, while the third follows since a change in phenotype of a individual in* $E$ *at a time point* $\mu(t)$ *necessarily induces a change in the proportion of cooperators in one group, and hence changes* $\phi_t$ *in* $F$*.*

We note that the $\text{MCEP}_{\text{group}}$ example above illustrates how the division and complexity of interactions between individuals and environment may depend on the level at which an evolutionary process is viewed (as well as the particular representation): In process $E$, group indices are considered environmental variables, which induce complex inter-dependencies in the fitnesses and heritabilities (phenotypic and environmental) between individuals; however, in process $F$, the groups are themselves considered individuals with their own properties, and much of the complexity at the underlying level is folded into the heritability of group phenotypes, along with a simpler fitness model.

Before outlining our final example, we introduce a general kind of MCEP over multiple temporal levels:

**Definition 5.3.** (Regular MCEP with multiple time-scales ($\text{MCEP}_{\text{temp}}$)): *An* $MCEP_{temp}$ *is an MCEP over processes* $E$ *and* $F$ *with total time-steps* $T_E$ *and* $T_F$*, where* $T_E = T_F T_S$ *(with* $T_S > 1$ *a 'temporal scaling factor'), and* $\mu(t) = t \cdot T_S$*. The transformation* $\tau$ *involves the projection of all phenotype and environmental variables in* $E$ *onto their values at* $\{\mu(0), \mu(1), ...\}$*, while the variable* $\pi_{nt}$ *in* $F$ *is set to the ancestor of* $n$ *at time-step* $\mu(t-1)$ *in* $E$*. The fitness variable* $w_{nt}$ *in* $F$ *is set to the absolute number of offspring of* $n$ *at* $t+1$*, and a classical fitness model is used as in Eq. 5. In general, the fitness, heritability and environmental kernels in* $F$ *will need to take unfactorized forms to capture the complex dependencies induced by the low-level dynamics in* $E$*. Further, we restrict interventions in* $E$ *to interventions on the phenotypes and environments of variables at* $\{\mu(0), \mu(1), ...\}$*, and map these to corresponding interventions in* $F$*.*

We note that any CEP may be converted to an $\text{MCEP}_{\text{temp}}$ by simply fixing a temporal scaling $T_S$, setting the original CEP as $E$, and following the construction above to form $F$. In this context, the values $w_{nt}$ represent a limited form of 'inclusive fitness' over the period $\mu(t)$ to $\mu(t+1)$ in $E$, with respect to genealogical relatedness relative to a base population at $\mu(t)$ (see [26] for a discussion of genealogical relatedness and genetic similarity based definitions of inclusive fitness, the former corresponding to Hamilton's formulation). As our final example, we use the above to illuminate the interaction of mutational processes and selection in cancer. As cancers evolve, subclones acquire not only distinct sets of mutations, but also distinct *mutational processes* governing the

random process by which mutations are generated (see [1]). For instance, by disrupting the DNA repair machinery, certain mutations may increase the mutation rate, or make it more likely that specific mutations (e.g. in particular trimer or pentamer contexts) are acquired in the future. Recent evidence has emerged that cancer driver mutations are differentially associated with the presence of particular mutational processes, and that the prevalence of particular mutational processes change in prototypical ways across the development of particular cancers [8,33]. To analyse the interaction of mutational processes with subclonal fitness, we introduce the following MCEP model (see Fig. 1D for related schematic):

**Example 5.4.** (MCEP with mutational processes ($\text{MCEP}_{\text{mut-proc}}$)): *We build an $MCEP_{mut\text{-}proc}$ model by introducing a CEP model of mutational processes as $E$, and forming $F$ directly by applying the $MCEP_{temp}$ definition in Def. 5.3. For $E$, we set the phenotype variables as $\phi_{nt} = \{x_{nt}, m_{nt}\}$, where $x$ and $m$ represents the genotype and mutational processes acting in cell $n$ at time $t$ respectively. We use the factorized form of heritability kernel in Eq. 2, setting $h_n := h_x(x_{nt}|x_{(\pi_t(n),t-1)}, m_{(\pi_t(n),t-1)}) \cdot h_m(m_{nt}|x_{nt}, m_{(\pi_t(n),t-1)})$. We note that this incorporates a genetic transmission kernel $h_x$ which is influenced by the mutational processes operating in the parent cell, and a mutational process kernel $h_m$ which allows for potential epigenetic inheritance of mutational processes across generations (as well as determination from the genotype). Further, we use a factorized fitness kernel of the form $f_n(w_{nt}|x_{nt})$; hence we assume that the genotype acts as a 'common cause' to the mutation processes and fitness of a given cell, but that the latter two variables are not directly causally linked. The structure kernel can be of arbitrary form, and all environments are empty.*

The $\text{MCEP}_{\text{mut-proc}}$ example above illustrates the following points. First, we note that for any individual in the lower-level process $E$, the mutational processes $m_{nt}$ are causally indendent of fitness $w_{nt}$; that is, intervening on $m_{nt}$ will not affect $w_{nt}$ (by definition). However, this is no longer the case in the higher-level process $F$; here, because of the intervening lower-level dynamics, there is a feedback between the mutational processes and fitness in $E$ across multiple time-steps, meaning that $w_{nt}$ for an individual in $F$ is affected by interventions on both $x_{nt}$ and $m_{nt}$. In fact, in $F$ we have the following:

**Proposition 5.5.** (Unique Information bounds for $\text{MCEP}_{\text{mut-proc}}$): *For an individual $n$ at time $t$ in the high-level component process ($F$) of an $MCEP_{mut\text{-}proc}$ as above, we have that $UI(w_{nt} : x_{nt} \backslash m_{nt}) \leq CID(\underline{x_{nt}} \rightarrow w_{nt})$, and $UI(w_{nt} : m_{nt} \backslash x_{nt}) \leq CID(\underline{m_{nt}} \rightarrow w_{nt}) + C$, with $C$ defined as in Th. 2.10. (Appendix A)*

The proof of Prop. 5.5 follows directly from Th. 2.10, along with the $\text{MCEP}_{\text{mut-proc}}$ definition, which implies that $x_{nt}$ causally influences, but is not influenced by $m_{nt}$ (in both $E$ and $F$), and both influence and are not influenced by $w_{nt}$ (in $F$). We note that Prop. 5.5 implies that the unique

information components of the observed distribution PID over $x_{nt}, m_{nt}, w_{nt}$ across multiple generations are informative about the potential contributions of $x_{nt}$ and $m_{nt}$ on subclone fitness. Particularly, $UI(w_{nt} : m_{nt} \backslash x_{nt}) > 0$ implies that there is a generic impact on fitness from the mutational processes across a particular time-scale, providing a lower-bound up to the additive constant $C$. Further, we note that while the $\text{MCEP}_{\text{mut-proc}}$ model above postulates no intra-generational feedback of the mutational processes on the genotype (only inter-generational feedback), the analysis above can be elaborated to include such feedback within generations, and can be expected to hold so long as the intra-generational feedback is weaker than that across generations.

## 6 Causal Interpretation of Price's Equation and Related Results

We finish by outlining a number of relationships which can be shown to hold in our CEP framework, including Price's equation an a number of analogous results. First, we can show:

**Theorem 6.1.** (Price's Equation with probabilistic and causal analogues (assuming perfect transmission)): *In a CEP, with empty environmental contexts and perfect transmission (hence, the heritability kernel factorizes and takes the form* $h_n(\phi_{nt}|\phi_{(\pi_t(n),t-1)}) = \delta(\phi_{nt}|\phi_{(\pi_t(n),t-1)})$*, where* $\delta(.|a)$ *is a delta distribution centered on* $a$*), and a classical fitness model as in Def. 3.2, we have:*
(a) Price's Equation:

$$\Delta\bar{\phi} = \frac{1}{\bar{w}}\,\text{Cov}(\phi, w), \tag{12}$$

(b) Probabilistic Price Equation (Rice's Equation [31,6]):

$$\Delta\hat{\phi} = \text{Cov}(\phi, \hat{\Omega}), \tag{13}$$

(c) Causal Price Equation:

$$\Delta_{\text{do}(\phi=\phi_0)}\hat{\phi} = \text{Cov}_{\text{do}(\phi=\phi_0)}(\phi_0, \hat{\Omega}), \tag{14}$$

*where we write* $\bar{a}$ *for the average of* $a$ *across individuals in a single, observed population, and* $\hat{a}$ *for the expected average of* $a$ *across the ensemble of populations modeled by the CEP (following [31]),* $\Omega_n = (w_n/\bar{w}|\bar{w} \neq 0)$ *is relative fitness (see [31]);* $\text{Cov}(.|.)$ *is the covariance; and the subscripts* $\text{do}(\phi = \phi_0)$ *indicate that a given quantity is evaluated under the distribution after intervening on* $\phi$ *(setting* $\phi$ *for all individuals at a given time-step).*

*Proof* For proofs of (a) and (b) see [25] and [31], which can be applied directly since no interventions are specified. For (c), we note that, having applied the operation $\text{do}(\phi = \phi_0)$, we produce a derived CEP whose underlying distribution is $P_{\text{do}(\phi=\phi_0)}$. Eq. 14 then follows directly by applying (b) to this derived CEP. □

We note that the distinctions between the original, probabilistic and causal versions of Price's equation in Theorem 6.1 allow us to make fine distinctions corresponding to direct and indirect selection on traits. For instance, although $\phi + \Delta\bar{\phi}$ and $\phi + \Delta\hat{\phi}$ will vary with $\phi$ for any trait which covaries with fitness or expected fitness, for $\phi_0 + \Delta_{\mathrm{do}(\phi=\phi_0)}\hat{\phi}$ this will only be the case for traits which have a causal impact on fitness (provided the intervention $\phi_0$ does not fix all individuals to a single phenotype).

Finally, we note an information-theoretic analogue of the Price equation based on the KL-divergence, which we believe has not been previously observed:

**Theorem 6.2.** (Analogue of Price's Equation based on KL-divergences): *In a CEP with restrictions and notation as in Th. 6.1, we have:*

$$\widehat{KL}(P_\phi||P'_\phi) = \widehat{KL}(P_\phi||\Omega^\dagger) + H(P_\phi), \tag{15}$$

*where $P_\phi$ and $P'_\phi$ are the observed (sample-level) distributions across trait $\phi$ at arbitrary time-points $t$ and $t+1$ resp., $KL(A||B) = \sum_i A_i \log(A_i/B_i)$ is the KL-divergence between (possibly unnormalized) distributions $A$ and $B$, and $\Omega^\dagger$ is a vector of relative fitness values for each value of the phenotype.*

*Proof* From the replicator equation, we have:

$$P'_i = P_i\left(\frac{w_i}{\bar{w}}\right) = P_i\Omega^\dagger_i, \tag{16}$$

where the subscript $i$ ranges across values of the phenotype. The result follows by substituting Eq. 16 into the LHS of Eq. 15 and rearranging:

$$\begin{aligned}
\widehat{KL}(P_\phi||P'_\phi) &= \mathbb{E}[\sum_i P_i \log(P_i/P'_i)] \\
&= -H(P_\phi) - \mathbb{E}[\sum_i P_i \log(P'_i)] \\
&= -H(P_\phi) - \mathbb{E}[\sum_i P_i \log(P_i\Omega^\dagger_i)] \\
&= \widehat{KL}(P_\phi||\Omega^\dagger) + H(P_\phi).
\end{aligned} \tag{17}$$

□

Using a similar argument to Th. 6.1 part (c), we can also state a causal analogue to Eq. 15:

**Corollary 6.3.** (Causal analogue of Theorem 6.2): *Using the notation of 6.2, and writing* $\mathrm{do}(\phi_0)$ *for* $\mathrm{do}(\phi = \phi_0)$*:*

$$\widehat{KL}_{\mathrm{do}(\phi_0)}(P_{\phi_0}||P'_{\phi_0}) = \widehat{KL}_{\mathrm{do}(\phi_0)}(P_{\phi_0}||\Omega^\dagger) + H(P_{\phi_0}). \tag{18}$$

Eqs. 15 and 18 are similar in form to the Price equation, since they determine the distance a trait will move (measured using displacement of its mean or KL divergence over the population-level distribution, for the Price equation and KL-analogue respectively) based on the similarity between the distributions of the trait and relative fitness (measured using the covariance or KL divergence respectively). For complex traits and evolutionary dynamics, the KL-analogues may be more informative, since they model the change in the whole trait distribution, as opposed to only its mean. For instance, at an evolutionary fixed point, we require not only that $\Delta\bar{\phi} = 0$ but also $\widehat{KL}(P_\phi||P'_\phi) = 0$ (assuming a large population). We also note the following properties of Eq. 15: Unlike the Price equation, Eq. 15 includes a dependency on the trait's entropy; further, the KL-'distance' moved by a trait's distribution increases as the (unnormalized) KL distance between trait and fitness distributions increases, or the entropy increases; the KL divergence thus need not go to 0 to reach a fixed point, but may be balanced by the entropy term, which may occur since the unnormalized KL divergence is not strictly positive, although it is bounded below by $-H(P_\phi)$ since the LHS of Eq. 15 is non-negative (for instance, the uniform distribution on a trait whose values all have equal fitness is a fixed point for which $\widehat{KL}(P_\phi||\Omega^\dagger) = -H(P_\phi)$). Finally, we briefly note that the relationship in Th. 6.2 differs from the associations between Price's equation and information theory that have been drawn in [10]: There, the mean change in a trait associated with Price's equation is re-expressed in terms of the Fisher Information between the trait and the environment (or an equivalent form involving the Shannon information), and it is shown that Fisher's Fundamental Theorem (FFT) arises by maximizing the information captured by the population; in contrast, Th. 6.2 does not rederive Price's equation or FFT, but rather relates two KL divergences involving analogous quantities to those in the former, leading to distinct conditions for evolutionary fixed-points as discussed.

## 7 Discussion

The framework of Causal Evolutionary Processes introduced in this paper provides a principled way to formulate evolutionary models, allowing both for cyclical interactions between evolutionary variables, and the analysis of evolutionary processes at multiple levels. We have developed a technical apparatus appropriate for this analysis in the form of Discrete Causal Networks and the Causal Information Decomposition, and have shown how a diverse range of evolutionary phenomena can be captured in our framework, including complex traits produced by feedback processes acting between epigenetic, behavioral and environmental levels, and multilevel selection models, including the selection of mutational processes in cancer. We have explored the properties of these models and our general framework, showing that under certain circumstances the causal impact of a given variable on another (for instance, a variant's impact on a trait) can be bounded by observed information-theoretic

quantities, and that a number of generalizations of Price's equation hold in our framework.

Our framework may be extended in various ways. For convenience, we have restricted our attention to discrete models in the above analysis (having both discrete time and discrete evolutionary variables). Our current framework may be formulated in the more general context of Cyclical Structural Causal Models [5] (see Appendix C), allowing for a measure-theoretic analysis including continuous variables and time. Further, we have restricted attention to the case of evolutionary processes with asexual reproduction; generalization to processes involving sexual reproduction, as well as lateral gene transfer, are straightforwardly handled by altering the structure of the parental map $\pi$ so that individuals are mapped to subsets of individuals in the previous generation as opposed to single individuals, while processes such as recombination and assortative mating can be modeled by using particular forms of heritability and structure kernels.

Further, we have not considered the problem of learning the causal structure and kernel forms from data. Methods relying on Mendelian randomization (e.g. [21]) can estimate the causal effects of variants on a trait assuming a linear relationship, but in general we may be interested in the causal effects of variables above the genetic level (e.g. epigenetics) and environmental factors on a trait, as well as non-linear models. In general, this is a hard problem, but general methods have been proposed, for instance the multi-level approach of [7], or the information-theoretic approach of [18], which may be imported into our framework. Further, we intend the unique information and backward-CID bounds in Th. 2.10 and 4.3 to be relevant for approximating the causal impacts of variables when certain assumptions are made, and in general these may be seen in the context of a host of bounds which relate various kinds of causal effect to observables (without interventions) under a range of assumptions (see [11]).

Finally, we intend our framework to be useful in clarifying foundational conceptual issues regarding causation and evolution. For instance, identifying a realized evolutionary process in nature requires providing criteria for identifying 'individuals' on which the process acts, and separating these from an 'environment'; attempts have been made to cast such criteria in information-theoretic terms (see for instance [22]), and our framework provides a natural 'language' for expressing such 'connecting principles'. We may for instance declare that, to a first approximation, a single level evolutionary process requires environmental variables which are independent of an individual's identity given its generation, acting as a 'thermal-bath' to the system; features such as spatial population structure, behavior-environmental feedback and niche construction (leading to more complex forms of heritability kernel) would then be taken as second-order principles which, if strong enough, may disrupt the 'individuality' of the entities in the original system (and hence the system's 'existence' qua system). Such considerations may also help sharpen questions regarding multilevel selection, whose role has been called into question in explanations of evolutionary processes (see [25, 26] for a summary of the issues). Potentially, a

model such as the multilevel CEP we outline, along with principles concerning which types of kernels are more or less 'preferred' at each level (for instance, in terms of description length, where lower-level kernels may inherit structure from kernels at higher-levels), could allow us to perform model selection among MCEPs with different numbers of levels. The problem of identifying multilevel selection can thus be cast in the more general framework of identifying causality at multiple levels, where we may have multple levels of variables which supervene on one another (for instance, see [14, 26, 7]); the existence of selection at multiple levels is the particular case of this problem when the causal relationships are constrained to have a particular structure (such as an MCEP). In summary, we believe that consideration of explicit causal models of the kind we have outlined will be useful when approaching both computational and conceptual issues in models of evolution.

## Appendix A Discrete Cyclic Causal Systems and Causal Information

We present here a technical summary of the basic concepts we need for our framework. In the process, we introduce the *Causal Information Decomposition* (CID), and summarize a number of its properties. First, we define a model of discrete cyclic causal systems, a *Discrete Causal Network* (DCN), which we will use as our basic model throughout the paper (see Fig. 1A for a summary of the relationships between the main models of the paper). We focus on discrete models to avoid the need to use differential entropy when defining information theoretic quantities, and leave generalization of our framework to the continuous case for future work.

**Definition 2.1.** (Discrete Causal Network (DCN)): *Let $\mathcal{X} = \{X_i\}$ be a set of discrete random variables indexed by $i \in \mathbb{I} = \{1...I\}$, each taking values in the set $\mathcal{V} = 1...V$, and $Pa : \mathbb{I} \to \mathcal{P}(\mathbb{I})$ be a function which returns a set of parents for each index (where $\mathcal{P}(.)$ denotes the powerset, and the underlying graph of $Pa$ may contain cycles). Then, a* DCN *over $\mathcal{X}$ consists of a collection of probability kernels $K_i(X_i = x_i | X_{Pa(i)} = x_{Pa(i)})$ specifying the conditional distribution of each variable on its parents, and a partial ordering $(\mathcal{I}, \leq)$ where $\mathcal{I}$ is a subset of all perfect interventions on $\mathcal{X}$ with the inherited ordering. Further, a* solution *to a DCN is a set of joint distributions $P_{\mathrm{do}(\iota \in \mathcal{I})}(\mathcal{X})$ such that the conditional distributions of all non-intervened variables $X_i$ on $X_{Pa(i)}$ match $K_i$, and the marginals of all other variables are delta distributions at their respective intervened values.*

We note that our Def. 2.1 can be viewed as a *Causal generalized Bayesian network* as introduced in [16] or a special case of a *Structural Causal Model* (SCM) as in [5], with an additional restriction in each case to a subset of interventions $(\mathcal{I}, \leq)$ (for an equivalent SCM formulation, see Appendix C; also note that for convenience we assume all variables in Def. 2.1 have a common discrete codomain, $\mathcal{V}$, which can be assumed without loss of generality, since $V$ may be taken large enought to embed all codomains if they are differently sized). By adding the restriction on the interventions considered, we are able to define a notion of *transformation* between DCMs, following the notion of transformations between Structural Equation Models in [32]:

**Definition 2.2.** (Transformations between DCNs): *Suppose we have two DCNs over variables $\mathcal{X} = \{X_{i=1...I}\}$, $\mathcal{Y} = \{Y_{j=1...J}\}$ taking values in $\{1...V_1\}^I$ and $\{1...V_2\}^J$ respectively, with kernels $\{K_i^1\}$ and $\{K_j^2\}$ (resp.) and intervention posets over $\mathcal{I}$ and $\mathcal{J}$ (resp.). Let $\tau$ be a map from $V_1^I$ to $V_2^{J'}$, where $J' = |A|$ for $A \subseteq \mathbb{J}$, and $\omega$ be an order preserving surjective map from $\mathcal{I}$ to $\mathcal{J}$. Then $(\tau, \omega)$ is a* transformation of DCMs *iff there exists a pair of solutions for which:*

$$P_{\mathrm{do}(\iota)}(X \in \tau^{-1}(y_A)) = P_{\mathrm{do}(\omega(\iota))}(Y_A = y_A) \quad \forall \iota, y_A, \tag{19}$$

*where $\tau^{-1}(y)$ is the pre-image of $y$ under $\tau$.*

Further, we wish to be able to analyse the information shared between DCN variables, and how this is affected by interventions. For this purpose, we first summarize the *Partial Information Decomposition* (PID) framework, using for convenience the formulation in [12]. As originally formulated [37], the PID decomposes the mutual information between a set of predictors $X_{1...I}$ and a dependent variable $Y$ , such that every collection of subsets of predictors is assigned an amount of (non-negative) *redundant* information. As noted by [12], this is equivalent to defining the *union information* for subsets of predictors, which we summarize as:

**Definition 2.3.** (Partial Information Decomposition (PID)): *Given random variables $\mathcal{X} = \{X_i\}$ indexed by $i \in \mathbb{I} = \{1...I\}$, with joint distribution $P(\mathcal{X})$, a collection of (possibly overlapping) subsets $\{S_1, S_2, ..., S_J\}$, $\forall j : S_j \subset I$, and a subset $T \subset I$ disjoint from all $S_j$'s, we use the symbol $PID$ to denote the* union information*, defined as:*

$$PID(\{X_{S_1}, X_{S_2}, ..., X_{S_J}\} : X_T) = \min_{Q \in \Delta} I_Q(\{X_{S_1}, X_{S_2}, ..., X_{S_J}\} : X_T) \quad (20)$$

*where $I_Q(X : Y)$ is the mutual information of $X$ and $Y$ under the distribution $Q$, and $\Delta$ is the set of all distributions over $\{X_{S_1}, X_{S_2}, ..., X_{S_J}, X_T\}$ whose pairwise marginals over $\{X_{S_j}, X_T\}$ match those of $P$, i.e. $Q(\{X_{S_j}, X_T\}) = P(\{X_{S_j}, X_T\})$, $\forall j$.*

Following [4], a number of further quantities may be defined in terms of the PID in the case that $J = 2$. These include the *shared* or *redundant information*, $SI(X_{S_1}; X_{S_2} : X_T) = I(X_{S_1} : X_T) + I(X_{S_2} : X_T) - PID(\{X_{S_1}, X_{S_2}\} : X_T)$, the *co-information* or *synergy*, $CI(X_{S_1}; X_{S_2} : X_T) = I(\{X_{S_1}, X_{S_2}\} : X_T) - PID(\{X_{S_1}, X_{S_2}\} : X_T)$, and the *unique information*, $UI(X_{S_1} \backslash X_{S_2} : X_T) = I(X_{S_1} : X_T) - SI(X_{S_1}; X_{S_2} : X_T)$. Analogues of these quantities may be defined for $J > 2$ as in [12].

To define a causal analogue to Def. 2.3, we include also a dependency on an *interventional distribution*. Hence, we set:

**Definition 2.4.** (Causal Information Decomposition (CID)): *Given a DCN as in Def. 2.1, subsets over indices $\{S_1, S_2, ..., S_J\}$ and $T$ as in Def. 2.3, and a distribution over interventions, $P_{\mathcal{I}}$, we define the CID as:*

$$CID(\{X_{S_1}, X_{S_2}, ..., X_{S_J}\} : X_T | \operatorname{do}(\iota) \sim P_{\mathcal{I}}) = \min_{Q \in \Delta(P_{\mathcal{I}})} I_Q(\{X_{S_1}, X_{S_2}, ..., X_{S_J}\} : X_T), \quad (21)$$

*where:*

$$\Delta(P_{\mathcal{I}}) = \{Q | Q(\{X_{S_j}, X_T\}) = P_{\operatorname{do}(\iota) \sim P_{\mathcal{I}}}(\{X_{S_j}, X_T\}), \forall j\}. \quad (22)$$

*Further, we use the short-hand notations:*

$$CID(X_{A_1}, ..., X_{A_J} \rightarrow \underline{X_B}) = CID(\{X_{A_1}, ..., X_{A_J}\} : X_B | \operatorname{do}(\iota) \sim Q_{\mathcal{X} \backslash X_B}), \quad (23)$$

and

$$CID(\underline{X_A} \to X_{B_1}, ..., X_{B_J}) = CID(\{X_{B_1}, ..., X_{B_J}\} : X_A | \operatorname{do}(\iota) \sim Q_{X_A}), \tag{24}$$

*when $\mathcal{I}$ contains a single intervention for each configuration of $X_B$ and $X_A$ in Eqs. 23 and 24 respectively, and $Q_{\mathcal{X}\backslash X_B}, Q_{X_A}$ weight these according to the marginals $P(\mathcal{X}\backslash X_B), P(X_A)$ (resp.), while assigning* 0 *to all other interventions (hence these intervention distributions capture the* actual variation *in $\mathcal{X}\backslash X_B$ and $X_A$ resp. in the sense of [13]). We refer to Eqs. 23 and 24 as the* backward- *and* forward- *CIDs respectively.*

The CID may be viewed as both a generalization of the PID and the *effective information* (EI) [34,14]. The EI can be defined as: $EI(P_{\mathcal{I}}(X_A) \to X_B) = I_{\operatorname{do}(\iota)\sim P_{\mathcal{I}_A}}(X_A; X_B)$, where $X_A$ and $X_B$ are disjoint, and $P_{\mathcal{I}_A}$ is an intervention distribution over $X_A$ (i.e., for any intervention $\iota$ affecting a variable in $\mathcal{X}\backslash X_A$, $P_{\mathcal{I}_A}(\iota) = 0$). We thus have:

**Proposition 2.5.** (Basic CID identities): *Letting $Q_{X_S}$ be as in Def. 2.4, and writing $\emptyset$ for the null intervention with $\delta_\emptyset(.)$ the intervention distribution which places probability* 1 *on $\emptyset$:*

$$(a)\ \ CID(\{X_{S_{1...J}}\} : X_T | \operatorname{do}(\iota) \sim \delta_\emptyset(.)) = PID(\{X_{S_{1...J}}\} : X_T) \tag{25}$$
$$(b)\ \ CID(\mathcal{X}\backslash X_B \to \underline{X_B}) = EI(Q_{\mathcal{X}\backslash X_B} \to X_B) \tag{26}$$
$$(c)\ \ CID(\underline{X_A} \to X_B) = EI(Q_{X_A} \to X_B). \tag{27}$$

*Proof* For (a), setting $P_{\mathcal{I}} = \delta_\emptyset(.)$ in Eq. 22 makes $\Delta(P_{\mathcal{I}})$ identical to the set $\Delta$ in Eq. 20, and hence the identity follows. For (b) and (c), when $J = 1$ in Eqs. 23 and 24, the CID reduces to the mutual information between variable subsets under the intervention distributions $Q_{\mathcal{X}\backslash X_B}$ and $Q_{X_A}$ respectively. Hence, setting the EI intervention distributions identically leads to the proposition. □

We now introduce a particular DCM model which will be important in later sections. This is a causal analogue of a Dynamic Bayesian Network [19], which we refer to as a Dynamic DCN (DDCN):

**Definition 2.6.** (Dynamic DCN (DDCN)): *A dynamic DCN is a DCN whose variables and kernel functions have a restricted structure. Particularly, we have $\mathcal{X} = \{X_{(i,t)}\}$ where $i \in \{1...I\}$ and $t \in \{0...T\}$, so that $X_{(i,t)}$ represents an observation of a quantity $i$ at time $t$. Also, for all $t > 0$, $Pa(i,t) = (Pa'(i), t-1)$ and $K_{(i,t)}(X_{(i,t)}|X_{Pa(i,t)}) = K'_i(X_{(i,t)}|X_{Pa(i,t)})$, where $Pa'(.)$ $K'_i(.|.)$ are auxiliary functions, and for $t = 0$, $Pa(i,t) = \{\}$. Further, it will be useful to consider a restricted DDCN (r-DDCN) with a constrained set of interventions, where we take $\mathcal{I}$ to include $\emptyset$, along with interventions of the form $\iota(i, v_i) = \operatorname{do}(X_{(i,t=0...T)} = v_i)$ and all combinations of such interventions. For an r-DDCN, we assume that the network converges to a unique steady-state under*

*all interventions. Finally, we define a projected DDCN at time $\alpha$ (p-DDCN($\alpha$)) to be a DCN constructed from an r-DDCN, including variables $\{X_i\}$ and $\{\zeta_i\}$, with $i \in \{1...I\}$ ranging across the same indices as the underlying r-DDCN, and the $X_i$'s and $\zeta_i$'s each taking values from the same set as the $X_{(i,t)}$'s, augmented in the case of the $\zeta_i$'s with a null value 0. Writing $(i,0)$ for the index of $X_i$, and $(i,1)$ for the index of $\zeta_i$, we set $Pa(i,0) = (Pa'(i) \backslash i, 0) \cup \{(i = 1...I, 1)\}$, $Pa(i,1) = \{\}$, $K_{(i,1)} = \delta_0(.)$, and let $K_{(i,0)}$ be the conditional distribution of $X_{(i,\alpha)}$ on $X_{(Pa'(i)\backslash i,\alpha)}$ in the underlying r-DDCN under the intervention $\wedge_i \iota(i,\zeta_i)$ (where $\iota(i,0) = \emptyset$, $\forall i$). The set of interventions in the p-DDCN consists of $\emptyset$, along with all combinations of interventions involving $\mathrm{do}(\zeta_i = v_i)$ where $v_i > 0$. By construction, the limiting p-DDCN($\alpha$) as $\alpha \rightarrow \infty$ is well defined, and represents the set of equilibrium distributions (under interventions) of the original r-DDCN, which we denote eq-DDCN.*

We immediately note the following:

**Proposition 2.7.** *For an r-DDCN and a derived p-DDCN($\alpha$), we have a transformation of DCNs $(\tau, \omega)$ from the former to the latter by setting: $\omega(\wedge_i \iota(i,v_i)) = \wedge_i \mathrm{do}(\zeta_i = v_i)$, $\omega(\emptyset) = \emptyset$ and $\tau$ to be the embedding $(x_{(1...I,\{0...T\}\backslash\alpha)}, x_{(1...I,\alpha)}) \mapsto (x_{(1...I),\alpha)}$, where the set $A$ in Def. 2.2 is $A = \{(i = 1...I, 0)\}$.*

*Proof* The proposition follows directly from the definitions, along with the fact that $\omega(.)$ as defined is order preserving, since it simply maps the basic intervention $\iota(i,v_i)$ in the r-DDCN to the basic intervention $\mathrm{do}(\zeta_i = v_i)$ in the p-DDCN, implying that order relations between all combinations will be preserved. □

Further, we introduce a special separability property on DDCN kernel functions which we will make use of in several places below:

**Definition 2.8.** (Self-separable DDCN kernels): *A DDCN kernel function $K(X_{(i,t)}|X_{(Pa'(i),t-1)})$ will be said to be self-separable, if $i \notin Pa'(i)$, or:*

$$K(X_{(i,t)}|X_{(Pa'(i),t-1)}) = \alpha K_1(X_{(i,t)}|X_{(i,t-1)}) + (1-\alpha)K_2(X_{(i,t)}|X_{(Pa'(i)\backslash i,t-1)}). \quad (28)$$

We finish by noting a number of further properties which follow from the definitions above. First, we summarize a number of properties of DDCNs with self-separable kernels as in Def. 2.8:

**Theorem 2.9.** (Properties of Separable DDCNs): *Given an eq-DDCN, derived from an r-DDCN in which all kernels are self-separable, we have (writing $H(.)$ for the entropy):*

$$(a) \quad P(X_i = x_i) = P_{\mathrm{do}(\iota)\sim Q_{\mathcal{X}\backslash X_i}}(X_i = x_i) \quad (29)$$

$$(b) \quad H(X_i) \geq CID(\mathcal{X}\backslash X_i \rightarrow \underline{X_i}) \quad (30)$$

$$(c) \quad H(S \subset \mathcal{X}\backslash X_i) \geq CID(S \rightarrow \underline{X_i}). \quad (31)$$

*Proof* For (a), we note that since P(.) is the equilibrium distribution of the underlying r-DDCN, we can write (letting $Y_i = \mathcal{X}\backslash X_i$):

$$\begin{aligned} P(x_i) &= \sum_{x_i,y_i} P(x_i,y_i)K(x_i|x_i,y_i) \\ &= \alpha \sum_{x_i} P(x_i)K_1(x_i|x_i) + \\ &\quad (1-\alpha)\sum_{y_i} P(y_i)K_2(x_i|y_i), \end{aligned} \tag{32}$$

where the second line uses the self-separable property. Since the terms on the RHS depend only on the marginals $P(X_i)$ and $P(Y_i)$, the theorem follows, since the marginals over $Y_i$ are preserved in the intervention distribution $Q_{\mathcal{X}\backslash X_i}$.

Part (b) follows directly from (a), since $CID(\mathcal{X}\backslash X_i \to \underline{X_i})$ is a mutual information involving $X_i$ under the intervention distribution $Q_{\mathcal{X}\backslash X_i}$. From (a), $H(X_i)$ is preserved under this intervention distribution, and the mutual information between two variables cannot exceed the entropy of either alone.

Part (c) follows by noting that the entropy $H(S)$ is also preserved in the intervention distribution $Q_{\mathcal{X}\backslash X_i}$. Since the RHS is again a mutual information, the inequality must hold. □

We summarize also a number of bounds involving the unique information in DCNs with a more restricted structure, namely a $Pa(.)$ function which forms a DAG (i.e. containing no cycles). Particularly, we focus on the effect of an arbitrary variable $X$ on another $Z$, where $Z$ has no descendants. All other variables are collapsed together as a single variable $Y = \mathcal{X}\backslash\{X,Z\}$. Further, we refer to the *causal strength*, $\mathfrak{C}$ (see [11,18]), where $\mathfrak{C}_{X\to Z} = KL(P(X)||P(Y)P(X|Y)\cdot\ P'(Z|Y))$, writing $KL(.||.)$ for the KL divergence, and $P'(Z|Y) = \sum_x P(X,Y)\cdot\ P(Z|X,Y)$.

**Theorem 2.10.** (Unique information bounds): *For a DCN with* $Pa(.)$ *forming a DAG and* $X, Y, Z$ *as above, we have:*

$$(a)\ \ UI(Z:X\backslash Y) \leq \mathfrak{C}_{X\to Z} \tag{33}$$

$$(b)\ \ UI(Z:X\backslash Y) \leq CID(\underline{X}\to Z), \text{if } Pa(X) = \{\} \tag{34}$$

$$(c)\ \ UI(Z:X\backslash Y) \leq CID(\underline{X}\to \{Y,Z\}) + C, \text{if } Pa(X) \neq \{\}, \tag{35}$$

*where*

$$C = \max_{x,y}(KL(P(Z|x,y)||P'(Z|y))) - \min_{x,y}(KL(P(Z|x,y)||P'(Z|y))), \tag{36}$$

*with* $P'(Z|Y) = \sum_x P(X,Y)P(Z|X,Y)$.

*Proof* For (a), we have from [11] that $\mathfrak{C}_{X\to Z} \geq I(Z:X|Y)$, where $I(.:.|.)$ denotes the conditional mutual information. Further, from [30] we have:

$$UI(Z:X\backslash Y) = \min_{Q\in\Delta} I(Z:X|Y), \tag{37}$$

with $\Delta$ as in Def. 2.3. Since $P \in \Delta$, $UI(Z : X \backslash Y) \leq I(Z : X|Y) \leq \mathfrak{C}_{X \to Z}$.

For (b), from [11] (prior to Lemma 3), we have that $\mathfrak{C}_{X \to Z} = I(X : Z)$ when $Pa(X) = \{\}$. Further, for $Pa(X) = \{\}$ we have that $P_{\mathrm{do}(\iota) \sim Q_X} = P$, and hence $I(X : Z) = CID(\underline{X} \to Z)$. Hence, from (a) $UI(Z : X \backslash Y) \leq \mathfrak{C}_{X \to Z} \leq CID(\underline{X} \to Z)$.

For (c), we note that we may write the causal strength as:

$$\begin{aligned} \mathfrak{C}_{X \to Z} &= KL(P(X)||P(Y)P(X|Y)P'(Z|Y)) \\ &= \sum_{x,y} P(x,y) KL(P(Z|x,y)||P'(Z,y)). \end{aligned} \tag{38}$$

Further, we may write:

$$\begin{aligned} &CID(\underline{X} \to \{Y,Z\}) \\ &= KL(P_{\mathrm{do}(\iota) \sim Q_X}||P(X)P(Y)P'(Z|Y)) \\ &= \sum_{x,y} P(x)P(y) KL(P(Z|x,y)||P'(Z,y)). \end{aligned} \tag{39}$$

Since Eqs. 38 and 39 are both weighted averages over $\cup_{x,y}\{KL(P(Z|x,y)|| P'(Z,y))\}$, we must have:

$$|\mathfrak{C}_{X \to Z} - CID(\underline{X} \to \{Y,Z\})| \leq C, \tag{40}$$

with $C$ as in the theorem. The inequality follows from Eq. 40 and (a). □

Since the RHS's of (a), (b) and (c) in Theorem 2.10 may all be regarded as measures of the impact interventions on $X$ will have on $Z$ (possibly in combination with $Y$), these bounds provide a way of predicting this effect from knowledge of only the observed unique information (i.e. without applying interventions). A corollary of Theorem 2.10 is that if $UI(X : Z) > 0$, $\mathfrak{C}_{X \to Z} > 0$, and necessarily $CID(\underline{X} \to Z) > 0$ if $Pa(X) = \{\}$. We explore Th. 2.10 further through simulations in Appendix D.

## Appendix B Full Proofs and Definitions from Section 4

Below, we give in full the definitions and proofs omitted from Sec. 4.

**Definition 4.2.** (CTCM with embeded DDCM (CTCM*)): *A CTCM* is a CTCM with further structure as follows. We let* $\phi_{nt} = \{x_{nt\tau}, y_{nt\tau}, z_{nt\tau}\}$ *and* $e_{nt} = \{e_{nt\tau}\}$*, where* $\tau$ *is an intra-generational time index, which runs from* $0...T_\tau$*. The kernels of a CTCM* have the form of an embedded DDCM:*

$$\begin{aligned} h_x(x_{nt}|x_{(\pi_t(n),t-1)}) &:= h'_x(x_{nt0}|x_{(\pi_t(n),t-1,0)}) \cdot \dot{\prod}_{\tau>0} \delta(x_{nt\tau}|x_{(n,t,\tau-1)}) \\ g_1(y_{nt}|x_{nt},e_{nt}) &:= g'_1(y_{nt0}) \cdot \dot{\prod}_{\tau>0} g''_1(y_{nt\tau}|x_{(n,t,\tau-1)}, e_{(n,t,\tau-1)}, y_{(n,t,\tau-1)}) \\ g_2(z_{nt}|y_{nt},e_{nt}) &:= g'_2(z_{nt0}) \cdot \dot{\prod}_{\tau>0} g''_2(z_{nt\tau}|y_{(n,t,\tau-1)}, e_{(n,t,\tau-1)}, z_{(n,t,\tau-1)}) \\ g_3(e_{nt}|z_{nt}) &:= g'_3(e_{nt0}) \cdot \dot{\prod}_{\tau>0} g''_3(e_{nt\tau}|z_{(n,t,\tau-1)}, e_{(n,t,\tau-1)}). \end{aligned} \tag{41}$$

*We allow that these variables and kernels can be further factorized, for instance by decomposing $y_{nt\tau}$ into sub-phenotypes representing expression values of individual genes or gene modules, and $e_{nt\tau}$ into different environmental factors, and introducing sub-kernels of $g_1', g_1'', g_3', g_3''$ for each sub-variable. Given a lowest level factorization, we require that the all transition kernels (i.e. the kernels $g_1'', g_2'', g_3''$, or their sub-kernels) are self-separable in the sense of Def. 2.8, where in all cases $K_1(.|.)$ in Eq. 28 is set to a delta function at the identity ($K_1(a|a) = \delta(a|a)$). In analogy with Def 2.6, we can define restricted and projected CTCM*'s by applying these constructions to the embedded DDCMs. In the former case, we restrict interventions over the variables $X, Y, Z$ and $e$ to those which fix the variable in an individual at time $t$ across all values of $\tau$, and in the latter case writing p-CTCM*($\alpha$) for the CTCM* formed by projecting the phenotype/environmental variables onto $\tau = \alpha$ at each $n$ and $t$. By taking the limit $\tau \to \infty$, we write eq-CTCM*=p-CTCM*($\infty$). Finally, we note that there is a subtlety in that, in moving from an r-DDCN to a p-DDCN in Def 2.6, we introduce the 'intervention variables' $\zeta$; these may be conveniently added as extra environmental variables in a p-CTCM*, since $g_1, g_2, g_3$ are all conditioned on $e$.*

**Theorem 4.3.** (Backward-CID bounds): *For a CTCM* represented as above with latent factors $C$, and associated eq-CTCM*, where $II$ is the interaction information, $S \subset \{X, Y, e\}$, $V \in \{X, Y, Z, e\}$, and $\overline{(.)}_V$ denotes the mean over values of $V$, in the limit $\alpha \to 1$ we have that:*

$$[II(S; Z; C^*) \leq 0] \wedge [CID(S \to \underline{Z}) \leq \overline{(CID(S : Z|Q_V))}_V] \implies PID(S : Z) \geq CID(S \to \underline{Z}), \tag{42}$$

*and similarly:*

$$[II(S; Z; C^*) \geq 0] \wedge [CID(S \to \underline{Z}) \geq \overline{(CID(S : Z|Q_V))}_V] \implies PID(S : Z) \leq CID(S \to \underline{Z}). \tag{43}$$

*where all $II$, $CID$ and $PID$ quantities are evaluated in the eq-CTCM* model (at a given $n$ and $t$, where $C^*$ is treated as an additional phenotype). Further, $Q_V = P_{eq}(\neg V)K_2^V(V|\neg V)$ (unrelated to the notation $Q_{X_A}$ used in Def. 2.4) with $K_2^V$ the second component of $V$'s kernel, as in Eq. 28, and we assume $Y, Z$ and $e$ are not factorized. For the case that $Y, Z$ or $e$ are factorized, $S$ and $V$ are subsets and elements of the sets of relevant factorized variables respectively, and Eqs. 42 and 43 hold identically.*

*Proof* For Eq. 42, we begin by considering the case that, in the underlying CTCM*, at index $\tau$ we have $C^*(\tau) = V \neq \emptyset$. Since all other variables $\neg V$ are arbitrarily sampled and $V$ has just updated according to $g_V''$ (i.e. letting $g_Y'' = g_1''$, $g_Z'' = g_2''$, $g_e'' = g_3''$), the distribution at time $\tau$ is $P_{eq}(\neg V)g_V''(V|\neg V) = Q_V$. Hence, the mutual information between $Z$ and $S$ at $\tau$ is $CID(S : Z|Q_V)$. Since we stipulate a common $\alpha$ for all transition kernels, the average of this quantity

across samples drawn from the equilibrium distribution is approximately the conditional mutual information (neglecting the case in which $C^*(\tau) = \emptyset$):

$$I(S;Z|C^*) = \sum_{C^*} P(C^*) I_{P(.|C^*)}(S;Z) \approx \overline{(CID(S:Z|Q_V))}_V. \quad (44)$$

Further, since $II(S;Z;C^*) = I(S;Z|C^*) - I(S;Z)$, in the limit $\alpha \to 1$ and for $II(S;Z;C^*) \leq 0$ we have:

$$\overline{(CID(S:Z|Q_V))}_V \leq I(S;Z) = PID(S:Z). \quad (45)$$

$PID(S:Z) \geq CID(S \to \underline{Z})$ then follows from the second line of Eq. 42. For Eq. 43 the proof is similar, with the direction of the inequalities reversed, and the generalization to factorized $Y, Z$ or $e$ is straightforward. □

## Appendix C Representing DCNs as Structural Causal Models

In [5], a *Structural Causal Model* (SCM) is defined as a tuple, $< \mathcal{I}, \mathcal{J}, \mathcal{X}, \mathcal{E}, \mathbf{f}, \mathbb{P}_{\mathcal{E}} >$, where $\mathcal{I}, \mathcal{J}$ are finite index sets of endogenous and exogenous variables respectively, $\mathcal{X} = \prod_{i \in \mathcal{I}} \mathcal{X}_i$ and $\mathcal{E} = \prod_{j \in \mathcal{J}} \mathcal{E}_j$ are products of codomains of endogenous and exogenous variables respectively, where each codomain is a measurable space, $\mathbf{f} : \mathcal{X} \times \mathcal{E} \to \mathcal{X}$ is a measurable function, and $\mathbb{P}_{\mathcal{E}} = \prod_{j \in \mathcal{J}} \mathbb{P}_{\mathcal{E}_\mathbf{j}}$ is a product of probability measures over the exogenous variables. A *solution* to an SCM is a pair of random variables $(X, E)$ taking values in $\mathcal{X}$ and $\mathcal{E}$ resp., such that the distribution of $E$ matches $\mathbb{P}_{\mathcal{E}}$, and the structural equations $X = \mathbf{f}(X, E)$ are satisfied almost surely.

We may represent a DCN as an SCM as follows (where we assume a DCN solution exists, and construct from this an SCN solution). We let $\mathcal{I}$ contain indices $(0, i)$ for each variable $X_i$ in the original DCN, along with index $(1, i)$ for a *mirror variable* $\zeta_i$ corresponding to each original variable (these collectively form the $X$'s of the SCM as defined above). We set the codomains $\mathcal{X}_{(0,i)}$ to be $\{1...V\}$ for the $X$'s, and $\{0, 1...V\}$ for the $\zeta$'s. We then set $\mathcal{J} = \mathcal{I}_{DCN}$, i.e. the intervention set in the original DCN, and the codomains $\mathcal{E}_j$ are all set to $V^I$. The probability measure $\mathbb{P}_{\mathcal{E}_\iota}$ is set so that, if joint configuration $[x_1, ..., x_I]$ occurs with probability $p$ under $\iota$ in the original DCN, the measure assigned to $[x_1, ..., x_I]$ under $\mathbb{P}_{\mathcal{E}_\iota}$ is $p$. We then set $\mathbf{f}$ so that $f_{(0,i)}(X, \zeta, E) = v$ if $\zeta$ corresponds to intervention $\iota$ in the original DCN and $E_\iota(i) = v$; $f_{(1,i)}$ is the constant 0, and all other values of $\mathbf{f}$ are set arbitrarily. The intervention $\iota$ in the original DCN corresponds to making a joint setting of the $\zeta$ mirror variables in the SCM to the desired intervention values (with 0 corresponding to no intervention). By construction, a joint setting of the endogenous variables surely exists under any intervention in the SCM constructed, since $\mathbf{f}$ simply 'copies' the joint settings from $E_j$ to $X$, where $j$ corresponds to the relevant intervention represented by $\zeta$.

We note that, in Def 2.1, we use the term *solution* in a slightly different sense to [5]. In our sense, the conditional distributions are specified under each possible intervention, and a set of joint distributions must be found which match these. In [5] however, the full joint distribution over the exogenous variables is specified by the model, and a solution consists of specifying the conditional distribution over the endogenous variables under each possible intervention and setting of $E$ which respects the constraints imposed by $\mathbf{f}$. Further, we note that while the SCM construction given above is fully general in the sense that any DCN can be represented in the form given, it is also purely 'formal' in the sense that the $f_i$'s do not directly correspond to causal mechanisms in the original DCN (represented by the kernels). Clearly, particular DCNs may have more compact representations as SCMs with a stronger correspondence in this sense; for instance, for acyclic DCNs the $\zeta$'s are not required, and each $X_i$ may be associated with an $E_i \in [0\ 1]$ which is sampled independently and uniformly, so that $f_i(X, E_i) = g_i(E_i|X_{Pa(i)})$, where $g_i(.|.)$ is the inverse of the cumulative distribution function of the kernel $K_i(X_i|X_{Pa(i)})$, and hence the $f_i$'s correspond directly to the DCN kernels. However, even if such direct correspondences cannot be drawn, the general SCM construction above ensures that for any DCN an SCM exists whose behavior is identical on all interventions.

## Appendix D Simulation study of the Unique Information bound

We explore the behavior of the bound in Th. 2.10 both in conditions when its assumptions are and are not satisfied through simulations. The results are shown in Fig. 2. Here, we run simulations in three DDCN models over the variables $X, Y, Z$, with the connectivity of each model shown on the left (defining the $Pa$ map). Each variable can take 4 values ($V = 4$), and we use self-separable DDCN models for all kernels (Def. 2.8). For the kernel parameters, we set $\alpha = 1 - 10^{-\gamma}$, $K_1$ to be the identity, $K_2$ by sampling each transition kernel entry uniformly at random and normalizing so that all conditional distributions sum to one, and we set the initial distributions similarly by uniform sampling. This parameterization lets $\gamma$ act as a 'stability' parameter, which we sweep between 0 (low stability) and 5 (high stability), where former implies the identity kernel is never chosen for updates, while the latter implies it almost always is. We first run 10 simulations of each model for $T = 500$ time-steps under no interventions, where a simulation involves sampling the parameters as above, building the full transition matrix $\mathbb{T}$ over the $4^3$ system states, and analytically calculating $p_T = p_0\mathbb{T}^T$. From $p_T$, we then calculate all marginal distributions, and use these to calculate $CID(\underline{V} \to \neg V)$, $CID(V_1 \to \underline{V_2})$ and $CID(\neg V \to \underline{V})$ for all variables $V$ and variable pairs $V_1, V_2$ ($V, V_1, V_2 \in \{X, Y, Z\}$) in the projected DDCN at time $T = 500$, approximating the equilibrium DDCN (see Def. 2.6). We calculate these quantities by running further simulations under the required intervention models, with the intervention distributions set using the marginals calculated. The latter two

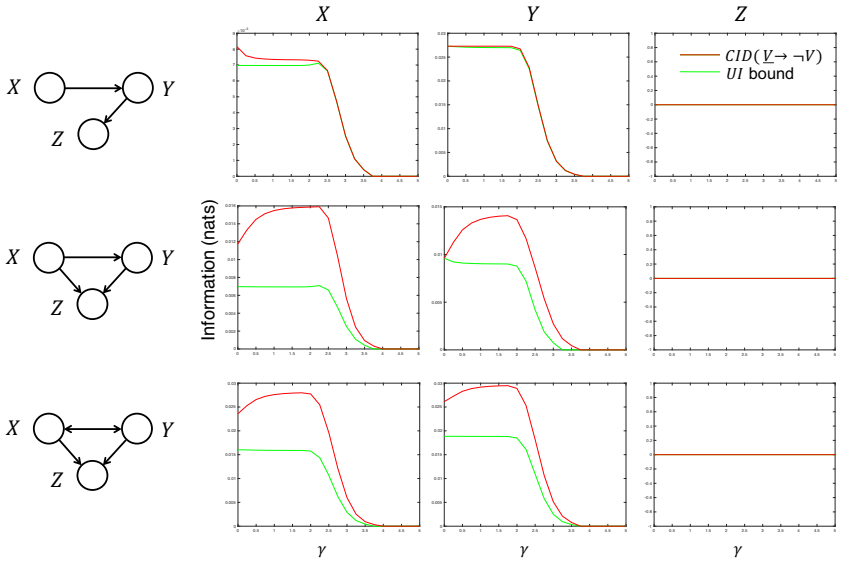


**Fig. 2** Simulation of the Unique Information bound. Figure shows results of the simulations described in Appendix D. Rows correspond to simulations of the model shown on the left, and columns show the forward-CID (a measure of causal effect) and its unique information lower-bound calculated for each variable. See text for full details.

quantities allow us to calculate the unique information $UI(V_1 \backslash V_2 : V_3)$ for all variable settings (Def. 2.3 and following). The figure shows, for each variable, a plot which compares the quantity $CID(\underline{V} \rightarrow \neg V)$ (the 'forward'-CID, which may be taken to measure the causal effect of the variable), with the maximum value of $UI(V \backslash V_1 : V_2)$, where $V_1, V_2 \in \neg V$. We take the average of these quantities across the 10 simulations for the plots shown.

When the assumptions of Th. 2.10b are satisfied, the quantity $\max_{V_1,V_2} UI(V \backslash V_1 : V_2)$ is guaranteed to be less than or equal to $CID(\underline{V} \rightarrow \neg V)$ (and all other unique information bounds will be looser than it). These conditions are only satisfied for variable $X$ in the first two models shown. However, variable $Y$ in the first two models satisfies the conditions of Th. 2.10c; as shown, the unique information provides a lower bound in these cases also, implying the constant $C$ in Th. 2.10c does not typically lead to a violation of the bound under the model sampling distribution described above (e.g. uniform sampling of transition kernels). Further, the last model investigates a case in which the DAG assumption of Th. 2.10 is violated, and we have feedback between $X$ and $Y$. The results show that, again, under the model sampling distribution adopted the unique information provides a reliable lower bound here also. We note that in all models, since $Z$ has no children, its causal impact ($CID(\underline{Z} \rightarrow \neg Z)$) is zero; the unique information bound is similarly pushed to 0, hence in the models tested the criterion $UI(V \backslash V_1 : V_2) > 0$ provides a reliable indicator that $V$ has non-zero causal impact. The above implies that

the unique information bounds of Th. 2.10b and c provide a general indicator of causal impact, which are robust to conditions in which the assumptions of the theorem are not strictly met.

## Appendix E Factorizing kernels in Discrete Causal Networks

We provide here further details on the notation we adopt for factorizations of DCN kernels. As specified in Def. 2.1, a DCN requires a kernel function to be specified for each variable $K_i(x_i|x_{Pa(i)})$ representing the conditional distribution of $x_i$ on its parents. We can summarize a DCN model using a 'product of kernels' notation, which we write as either $\dot{\prod}_i K_i(x_i|x_{Pa(i)})$ or $K_1(x_1|x_{Pa(1)}) \cdot K_2(x_2|x_{Pa(2)}) \cdot ....$ We note that, if the $Pa$ relation forms a DAG, this product will directly represent the joint distribution over the DCN variables (subject to no interventions); however, since in general $Pa$ may contain cycles, we adopt the convention that $\dot{\prod}_i K_i(x_i|x_{Pa(i)})$ represents the *set of distributions* which satisfy all the kernel relations. An intervention which sets $x_i$ to value $v$ may be implemented by replacing $K_i(x_i|x_{Pa(i)})$ by $\delta(x_i|v)$, and a solution to the DCN is a choice function which picks a single distribution from the kernel product sets representing each intervention (including the null intervention). For partial products, this notation represents a higher-order conditional kernel, for instance $K(x_i, x_j|x_k) := K(x_i|x_j, x_k) \cdot K(x_j|x_i, x_k)$. Here, $K(x_i, x_j|x_k)$ is a particular conditional distribution which satisfies the kernel product relations between $x_i, x_j$ specified by the lower-order kernels (conditioned on $x_k$). We note that this relation is asymmetric (hence our use of the ':=' symbol, which we note is unrelated to the use of this symbol to denote interventions in [29]); in general, there may be multiple conditional distributions which satisfy the RHS, which the LHS picks from. In a given kernel product, we may thus combine groups of kernels together into higher-order kernels, or split them into multiple lower order kernels, where the former restricts the solution set, and the latter expands it (a particular solution, for instance under the null intervention, is thus represented by the unconditional kernel $K(x_1, ...x_I) = P(x_1, ...x_I)$). A particular DCN selects a 'base-level' factorization, which determines the variable index set $\mathbb{I}$ and thus which interventions may be performed on the model; for instance, if $K(x_i, x_j|x_k)$ is a base-level kernel, then variables $x_i$ and $x_j$ must be treated as a single variable in the DCN, and interventions cannot be applied to $x_i$ and $x_j$ separately. In this sense, once the base level has been set and a particular DCN solution chosen, all higher-order kernels are fully determined, and are used for notational convenience only.

## References

1. Alexandrov, L.B., et. al. Signatures of mutational processes in human cancer. *Nature*, 500(7463), p.415, 2013.
2. Awodey, Steve. *Category theory*. Oxford University Press, 2010.

3. Ay, N. and Polani, D. Information flows in causal networks. In *Advances in complex systems*, 11(01), pp.17-41, 2008.
4. Bertschinger, N., Rauh, J., Olbrich, E., Jost, J. and Ay, N. Quantifying unique information. *Entropy*, 16(4), pp.2161-2183, 2014.
5. Bongers, S., Peters, J., Schölkopf, B. and Mooij, J.M. Theoretical Aspects of Cyclic Structural Causal Models. *arXiv preprint* arXiv:1611.06221, 2018.
6. Calcott, B. and Sterelny, K. *The major transitions in evolution revisited*. The MIT Press, 2011.
7. Chalupka, K., Eberhardt, F. and Perona, P. Multi-level cause-effect systems. In *Artificial Intelligence and Statistics*, (pp. 361-369), 2016.
8. Dentro, S.C., Leshchiner, I., Haase, K., Tarabichi, M., Wintersinger, J., Deshwar, A.G., Yu, K., Rubanova, Y., Mcintyre, G., Vazquez-Garcia, I. and Kleinheinz, K. Portraits of genetic intra-tumour heterogeneity and subclonal selection across cancer types. *bioRxiv*, p.312041, 2018.
9. Felsenstein, J. *Theoretical Evolutionary Genetics*, 2016. Online book at: evolution.genetics.washington.edu/pgbook/pgbook.html
10. Frank, Steven A. Natural selection maximizes Fisher information. In *Journal of Evolutionary Biology* 22, no. 2 (2009): 231-244.
11. Geiger, P., Janzing, D. and Schölkopf, B. Estimating Causal Effects by Bounding Confounding. In *Proceedings of the Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, 2014.
12. Griffith, V. and Koch, C. Quantifying synergistic mutual information. In *Guided Self-Organization: Inception* (pp. 159-190). Springer, Berlin, Heidelberg, 2014.
13. Griffiths, P.E., Pocheville, A., Calcott, B., Stotz, K., Kim, H. and Knight, R. Measuring causal specificity. *Philosophy of science*, 82(4), pp.529-555, 2015.
14. Hoel, E.P. When the map is better than the territory. *Entropy*, 19(5), p.188, 2017.
15. Houle, D., Govindaraju, D.R. and Omholt, S., 2010. Phenomics: the next challenge. *Nature reviews genetics*, 11(12), p.855, 2010.
16. Itani, S., Ohannessian, M., Sachs, K., Nolan, G. P., and Dahleh, M. A. Structure learning in causal cyclic networks. In *JMLR Workshop and Conference Proceedings*, volume 6, page 165176, 2010.
17. Janzing, D. and Scholkopf, B. Causal inference using the algorithmic Markov condition. *IEEE Transactions on Information Theory*, 56(10), pp.5168-5194, 2010.
18. Janzing, D., Balduzzi, D., Grosse-Wentrup, M. and Schölkopf, B. Quantifying causal influences. In *The Annals of Statistics*, 41(5), pp.2324-2358, 2013.
19. Koller, D. and Friedman, N. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.
20. Krakauer, D.C., Page, K.M. and Erwin, D.H. Diversity, dilemmas, and monopolies of niche construction. *The American Naturalist*, 173(1), pp.26-40, 2008.
21. Lawlor, D.A., Harbord, R.M., Sterne, J.A., Timpson, N. and Davey Smith, G. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statistics in medicine*, 27(8), pp.1133-1163, 2008.
22. Krakauer, D., Bertschinger, N., Olbrich, E., Ay, N. and Flack, J.C. The information theory of individuality. *arXiv preprint arXiv:1412.2447*, 2014.
23. J. M. Mooij, D. Janzing, and B. Schölkopf. From Ordinary Differential Equations to Structural Causal Models: the deterministic case. In *Proceedings of the Twenty-Ninth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 440–448, 2013.
24. Nowak, M.A., Tarnita, C.E. and Wilson, E.O. The evolution of eusociality. *Nature*, 466(7310), p.1057, 2010.
25. Okasha, S. *Evolution and the levels of selection*. Oxford University Press, 2006.
26. Okasha, S. The relation between kin and multilevel selection: an approach using causal graphs. *The British Journal for the Philosophy of Science*, 67(2), pp.435-470, 2015.
27. Paulsson, J. Multileveled selection on plasmid replication. *Genetics*, 161(4), pp.1373-1384, 2002.
28. Pearl, J. *Causality*. Cambridge university press, 2009.
29. Peters, J., Janzing, D. and Schölkopf, B. *Elements of causal inference: foundations and learning algorithms*. MIT press, 2017.

30. Rauh, J., Bertschinger, N., Olbrich, E. and Jost, J., June. Reconsidering unique information: Towards a multivariate information decomposition. In *IEEE International Symposium on Information Theory (ISIT)*, (pp. 2232-2236), 2014.
31. Rice, S.H. *Evolutionary theory: mathematical and conceptual foundations*. Sunderland, MA: Sinauer Associates, 2004.
32. Rubenstein, P.K., Weichwald, S., Bongers, S., Mooij, J.M., Janzing, D., Grosse-Wentrup, M. and Schölkopf, B. Causal consistency of structural equation models. In *Proceedings of the Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, 2017.
33. Temko, D., Tomlinson, I.P., Severini, S., Schuster-Böckler, B. and Graham, T.A. The effects of mutational processes and selection on driver mutations across cancer types. *Nature communications*, 9(1), p. 1857, 2018.
34. Tononi, G. and Sporns, O., Measuring information integration. *BMC neuroscience*, 4(1), p.31, 2003.
35. Traulsen, A. and Nowak, M.A. Evolution of cooperation by multilevel selection. *Proceedings of the National Academy of Sciences*, 103(29), pp.10952-10955, 2006.
36. Wagner, A. Causal drift, robust signaling, and complex disease. *PloS one*, 10(3), p.e0118413, 2015.
37. Williams, P.L., Beer, R.D. Nonnegative decomposition of multivariate information. *CoRR*, abs/1004.2515, 2010.
38. Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M.R., Powell, J.E., Montgomery, G.W., Goddard, M.E., Wray, N.R., Visscher, P.M. and Yang, J. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nature genetics*, 48(5), p.481, 2016