

Towards a system-level causative knowledge of pollinator communities

Serguei Saavedra¹, Ignasi Bartomeus², Oscar Godoy³, Rudolf P. Rohr⁴, Penguan Zu^{5,6}

¹Department of Civil and Environmental Engineering, MIT,
77 Massachusetts Av., 02139 Cambridge, MA, USA. ORCID: 0000-0003-1768-363X

²Estación Biológica de Doñana (EBD-CSIC), Seville, Spain. ORCID: 0000-0001-7893-4389

³Departamento de Biología, Instituto Universitario de Ciencias del Mar (INMAR),
Universidad de Cádiz, E-11510, Royal Port, Spain. ORCID: 0000-0003-4988-6626

⁴Department of Biology - Ecology and Evolution, University of Fribourg
Chemin du Musée 10, CH-1700 Fribourg, Switzerland. ORCID: 0000-0002-6440-2696

⁵Department of Environmental Systems Science, ETH Zurich,
Schmelzbergstrasse 9, CH-8092, Zurich, Switzerland. ORCID: 0000-0002-3222-598X

⁶Department Fish Ecology & Evolution, Swiss Federal Institute of Aquatic Science and Technology (Eawag)
Seestrasse 79, CH-6047, Kastanienbaum, Switzerland.

To whom correspondence should be addressed: *sersaa@mit.edu

Content Type: Opinion Piece

Keywords: causal inference, probability, community ecology, tractability, scalability, nonparametric analysis

2 Abstract

3 Pollination plays a central role both in the maintenance of biodiversity and in crop production.
4 However, habitat loss, pesticides, invasive species, and larger environmental fluctuations are con-
5 tributing to a dramatic decline of numerous pollinators world-wide. This has increased the need
6 for interventions to protect the composition, functioning, and dynamics of pollinator commu-
7 nities. Yet, how to make these interventions successful at the system level remains extremely
8 challenging due to the complex nature of species interactions and the various unknown or un-
9 measured confounding ecological factors. Here, we propose that this knowledge can be derived
10 by following a probabilistic causal analysis of pollinator communities. This analysis implies the
11 inference of interventional expectations from the integration of observational and synthetic data.
12 We propose that such synthetic data can be generated using theoretical models that can enable
13 the tractability and scalability of unseen confounding ecological factors affecting the behavior of
14 pollinator communities. We discuss a road map for how this probabilistic causal analysis can be
15 accomplished to increase our system-level causative knowledge of natural communities.

16 Introduction

17 Pollinators comprise a highly diverse group of species including bees, flies, butterflies, beetles,
18 and some vertebrates [1]. They all have in common a shared interest in visiting flowers to extract
19 resources, collectively and indirectly mediating the reproduction of most of the worldwide plant
20 species [2] and maximizing crop production for 75% of cultivated crops [3]. Hence, pollination is
21 now recognized not only as a key ecosystem function, but also as a key ecosystem service con-
22 tributing to human food security. However, human induced rapid environmental change has been
23 threatening most of these pollinators [4]. On the one hand, habitat destruction and modification
24 is reducing the populations of many pollinator species, often leading to local extirpation. On the
25 other hand, some other species can thrive in human modified ecosystems, but those often face
26 extra pressures such as pesticide exposure, exotic species, or pathogens. In top of that, climate
27 change is altering species' physiological responses, distribution, and activity periods [5]. Overall,
28 we are assisting to a rapid restructuring of pollinator communities world-wide, where their rela-
29 tive abundance, composition, and ecological interactions are being modified with hard to predict
30 consequences for their health.

31 These human pressures on pollinator communities have increased the need for human interven-
32 tions to protect the composition, functioning, and stability of pollinators and their interactions
33 [6]. These interventions include from well established practices such as habitat protection, to more
34 complex actions such as the addition or removal of particular species and their interactions [7].
35 For example, planting field margins [8] or adding managed pollinators [9] have become, respec-
36 tively, popular restoration practices in agricultural systems to increase resources for pollinators or
37 supplement crop pollination. However, these practices often ignore side effects, such as the effects
38 of changes in micro-climate conditions or pathogen prevalence on pollinator health. For instance,
39 a recent study has shown that bumblebees' occupancy patterns in Europe and North America are
40 sensitive to temperature [10]. Similarly, it has been shown how managed pollinator densities not
41 only increases competition among pollinators [11], but also increases parasite loads [12], which
42 can spillover to other species [13]. Yet, as of today, we lack a community-wide framework to guide
43 interventions beyond single species. Indeed, it has been shown that even small local interventions
44 (i.e., at the species level) can have heterogeneous and arbitrary cascading effects across entire
45 communities [14]. This has emphasized the dire need to establishing a system-level causative
46 knowledge of pollinator communities.

47 To address the challenge above, ideally, we need to establish well-defined experiments eliminat-
48 ing all sources of bias (e.g., using randomized controlled trials) and test the effectiveness of a
49 given intervention [15]. However, those sources of bias become extremely difficult to identify
50 and measure in changing natural ecological communities conformed by several co-occurring and
51 interacting species [16]. Moreover, many of these interventions may not be ethical (e.g., species

52 removal) or feasible to perform because pollinators move freely and are difficult to track. This
53 implies that it is instead necessary to obtain interventional knowledge from observational data
54 (e.g., field studies or partially controlled studies) using causal-inference analysis [17]. These ob-
55 servational data (that record for example the observed presence/absence of pollinators) differ
56 from fully controlled studies (that remove or add pollinators) in the sense that observational
57 variables are the result of what is perceived and not of what is intervened by the investigator.
58 Importantly, these observational data are typically confounded by unknown factors (also known
59 as noise, context, or environmental conditions), such as biotic and abiotic variables, making dif-
60 ficult to differentiate between spurious and actual cause-effect relationships. To circumvent this
61 problem, we propose that interventional knowledge can be inferred from the integration of ob-
62 servational and synthetic data. These synthetic data can be generated using theoretical models
63 that can enable the tractability (operationalization and reproducibility) and scalability (gener-
64 alization across dimensions) of unseen confounding factors acting at the community level. This
65 framework can provide a probabilistic knowledge of how likely is a given cause to generate a
66 target effect within a pollinator community (i.e., focusing on the probability of causes instead of
67 effects). In the remainder, we discuss a road map for how this probabilistic causal analysis can be
68 accomplished and illustrate it with a case study.

69 **Observational data: known factors**

70 Given the lack of systematically controlled experiments, observational data from field studies or
71 quasi-controlled experiments (where few factors may be controlled) can provide the raw material
72 to understand the behavior (e.g., composition, dynamics) of a community. This behavior comes
73 in the form of a joint probability distribution $P_{\mathbf{V}}$ over a set of relevant variables \mathbf{V} . For example,
74 studies may record any aspect of community composition as a function of a set of semi-controlled
75 variables such as the presence (or density) of specific pollinators [18], their floral resources in-
76 cluding both the identity of interacting plant species [19] and plant chemical composition [20–22],
77 top down regulators including pathogen [23] and predators [24], as well as several environmental
78 variables such as temperature [25, 26] or pesticide exposure [27, 28]. These observational studies
79 can be either for a specific period of time (across different locations) or measure pollinator com-
80 munities repeatedly over time in order to capture a wider range of temporal conditions affecting
81 pollinators’ population trajectories, which often follow non-linear dynamics [29, 30].

82 While observational data are designed to track potential mechanisms affecting pollinator commu-
83 nities, they cannot establish cause-effect relationships by themselves, only associations [15, 17].
84 That is, following Reichenbach’s principle [31], if two variables (X, Y) are statistically related,
85 then there exists a third variable or set of variables (Z) that causally influenced both (known
86 as confounding effect: $X \leftarrow Z \rightarrow Y$). In some situations, Z coincides with either X or Y (i.e.,

87 $Z = X$ or $Z = Y$), establishing a causal link between X and Y (i.e., $X \rightarrow Y$ or $Y \rightarrow X$).
88 But without knowledge of Z (or when this unknown effect cannot be blocked from the analysis),
89 we cannot safely conclude cause-effect relationships. Thus, conditional distributions (e.g., $P_{Y|X}$)
90 derived from observational data can coincide with causal mechanisms (e.g., $X \rightarrow Y$), but not
91 necessarily. Similarly, two variables (X, Y) may be statistically related if both are the common
92 (confounding) causes of a given effect Z (i.e., $X \rightarrow Z \leftarrow Y$: known as collider in the causal-
93 inference literature [15]) upon which the data is selected (known as selection bias). This problem
94 typically arises when data is filtered or conditioned by Z and $X \not\perp\!\!\!\perp Y|Z$, but $X \perp\!\!\!\perp Y|\{\emptyset\}$ ($\not\perp\!\!\!\perp$ and
95 $\perp\!\!\!\perp$ denote dependence and independence, respectively). Moreover, in a multivariate system, the
96 sources of bias can be originated from direct and indirect common causes and effects. These prop-
97 erties make extremely problematic the interpretation of relationships derived from multivariate
98 regression and meta-analysis that do not have a causal hypothesis [32].

99 For example, let us assume that pollinator abundance is caused by flower abundance, temperature,
100 and some unknown factors. Similarly, let us assume that flower abundance is caused by water
101 availability, temperature, and a subset of the same unknown factors. Then, in a multivariate
102 regression model that includes all factors (except for the unknown) as potential explanations of
103 pollinator abundance, it is likely that water availability will have a strong explanatory effect over
104 pollinator abundance (even though we are conditioning over flower abundance). This happens
105 for the reason that flower abundance introduces a selection bias (collider) between water and the
106 unknown factors, which then gets propagated to pollinator abundance following the cause-effect
107 relationships. Note that flower abundance cannot be eliminated from the regression model ei-
108 ther, because it is needed to partially block the path between water availability and pollinator
109 abundance. This type of examples also illustrates that prediction is different from explanation
110 [33]. Therefore, to infer cause-effect relationships in this example, it is needed to have more
111 information about the underlying causal story and the corresponding unknown confounding fac-
112 tors. In the next sections, we will discuss how to use synthetic data derived from theoretical
113 models to account for confounding unobserved variables, and then how to generate interventional
114 distributions (knowledge) from observational and synthetic data.

115 **Synthetic data: unknown factors**

116 The role of theoretical models has been understood as a formal platform to establish logico-
117 mathematical postulates (formal statements) about how the real-world possibly behaves and to
118 obtain data that can be difficult to generate empirically [34–36]. These postulates are, of course,
119 tautological as they are analytically (or algorithmically) derived from a set of primary principles.
120 It is only possible to falsify these postulates based on their biological interpretation. Thus, the
121 value of theoretical models is to provide hypotheses, predictions, generalization, and systematic

122 links between model parameters (the interpretable factors/context) and the behavior of a system,
123 which can then be revised based on empirical information. The interpretation of theoretical mod-
124 els (model parameters) can range from highly mechanistic to highly phenomenological depending
125 on the level of resolution under investigation [37]. For example, mechanistic interpretations are
126 based on detailed descriptions of ecological processes, such as metabolic rates, nutrients uptake,
127 mobility patterns, predation processes, and behavioral patterns, among others [38, 39]. In turn,
128 phenomenological interpretations are based on summary outcomes that are expressed in terms
129 of model parameters without establishing any specific statement about how exactly these out-
130 comes come to existence (e.g., intrinsic growth rates, species interactions, and death rates, among
131 others). In general, there is no one better model than another (unless there is knowledge about
132 the actual processes and there is capacity to obtain the initial conditions), it all depends on the
133 research question and system under investigation.

134 Regarding pollinator communities (and ecological communities in general), there are two impor-
135 tant properties that need to be considered if one aims to study theoretically and systematically
136 the factors under which several interacting species can coexist [40]: tractability and scalability.
137 We define tractability as the property of a theoretical model to have all its potential solutions
138 fully operationalized, defined, measured, and reproduced over relatively short periods of time (i.e.,
139 polynomial time), enabling a systematic understanding between solutions and parameter values.
140 For example, the Lonsdorf [41] model uses only land use parameters to directly explain pollina-
141 tor densities following a simple equation. Instead, complex models characterized by higher-order
142 polynomials are limited by their intractability (e.g., optimal foraging models [40, 42, 43]). In
143 fact, it has already been proved that it is impossible to write analytically (a closed-form algebraic
144 solution) a polynomial system with degree five or higher with arbitrary coefficients (unknown
145 values) [44]. Note that a simple 3-species system (e.g., two pollinators and one plant) with Type
146 II functional responses (i.e., a non-linear response such as those observed in density-dependent
147 processes arising from competition for floral resources or pathogen spillover) can already form a
148 polynomial of degree eight [45]. This intractability of complex models implies that if the majority
149 of their parameter values are not known a priori (reducing the system to a polynomial of degree
150 four or lower), these models can only be used numerically (simulations). Then, the problem that
151 arises is that it becomes computationally impossible to differentiate the role played by each pa-
152 rameter (e.g., interactions, environmental conditions) in the solutions of the system [40]. While
153 studies have attempted to tackle this complexity by using statistical methods such as Akaike
154 Information Criterion [46], the number of solutions of a polynomial system does not necessarily
155 depend on the number of parameters but on the polynomial degree [45]. Hence, it is not just
156 the lack of data that limits the use of complex models, as it can be perceived [47], it is their
157 intractability, especially in high-dimensional systems [40].

158 In turn, we define scalability as the property of a model to establish clear and invariant rules
159 across dimensions, enabling extensions from simple to complex natural communities. For ex-
160 ample, the Lonsdorf model [41] is designed to track central place foragers (e.g., bees), where a
161 key piece of the model is the foraging range from a central point in the landscape; but it is not
162 scalable to wanderers (e.g., flies and butterflies), which move freely over the landscape tracking
163 resources. Similarly, it has been demonstrated that insights derived from classic work on coexis-
164 tence using 2-species Lotka-Volterra models cannot be directly extrapolated to higher dimensions
165 [48]. Therefore, simple phenomenological or simple mechanistic models can be understood as the
166 simplification (reduction of polynomial degree and free parameters) of complex models to enhance
167 the tractability and scalability of a system. However, it is central to fully understand how they
168 should be used.

169 For instance, generic phenomenological models can be written in the form $\dot{\mathbf{N}} = \mathbf{N}f(\mathbf{N}, \mathbf{U})$, where
170 $\dot{\mathbf{N}}$ represents the time derivative of species density, and f is a given function describing the
171 relationship among endogenous \mathbf{N} variables and contextual parameters \mathbf{U} [36]. Note that having
172 the vector \mathbf{N} in front of the function f guarantees the impossibility of negative densities (or species
173 revival without immigration). A classic phenomenological model that follows this formalism is
174 the linear Lotka-Volterra (LV) model [49, 50]: $\dot{\mathbf{N}} = \mathbf{N}(\mathbf{r} + \mathbf{A}\mathbf{N})$, where \mathbf{r} typically represents
175 species intrinsic growth rates and \mathbf{A} is the so-called interaction matrix (summarizing the positive
176 or negative per capita effect of one species upon individuals of another species). While the linear
177 LV model can be derived from first principles, such as energy conservation or thermodynamic
178 limits, it can be phenomenological interpreted as the first-order approximation (derived from
179 the Taylor expansion) of the unknown function f [35]. This can then make the elements of the
180 linear LV model to be interpreted as endogenous variables \mathbf{N} , a set of time-invariant interaction
181 parameters summarized in \mathbf{A} , and contextual parameters \mathbf{r} . This interpretation allows both
182 the tractability and scalability of a multispecies community. That is, the analytical solution
183 is $\mathbf{N}^* = -\mathbf{A}^{-1}\mathbf{r}$ (setting $\dot{\mathbf{N}} = 0$), making possible the one-to-one mapping between \mathbf{N}^* and \mathbf{r}
184 [51]. This means that the constraints imposed by \mathbf{A} on contextual factors \mathbf{r} to generate a given
185 endogenous behavior \mathbf{N}^* can be systematically analyzed regardless of the number of species in
186 the system.

187 Importantly, tractable and scalable models become good candidates towards increasing our system-
188 level causative understanding of pollinator communities. Indeed, by conceptualizing the function
189 f above as an approximation to a structural causal model [15, 17] (i.e., $X = f_X(\mathbf{V}_X, \mathbf{U}_X)$, where
190 f_X is a time-invariant function defining the cause-effect relationships of X , \mathbf{V}_X is the set of causes
191 of X , and \mathbf{U}_X is the random noise/context affecting X defined by $P_{\mathbf{U}_X}$), it is possible to obtain
192 theoretical probability distributions of unknown factors \mathbf{U} (e.g., \mathbf{r} in the LV model) compatible
193 with a given behavior of \mathbf{N}^* as dictated by a set of invariant rules (e.g., \mathbf{A} in the LV model).

194 For example, in the linear LV model, by assuming that $\mathbf{r} \in \mathbb{R}^S$ (where S is the dimension of the
195 system) is a priori randomly and uniformly distributed (conforming with ergodicity and inde-
196 pendence from initial conditions [52]), it is possible to calculate analytically the range of feasible
197 unknown conditions (i.e., $\mathbf{U} \subseteq \mathbf{r}$ and $P_{\mathbf{U}}$) leading to a given set of species (i.e., $I \subseteq R$, where R
198 is the set of species within a community) with positive densities at equilibrium ($\mathbf{N}_I^* > 0$) [53, 54].
199 Moreover, we can calculate the expected number of species with positive densities at equilibrium
200 $E[\mathbf{N}^* > 0]$ (or the probability of persistence of each single species within a community) [52]. Note
201 that if \mathbf{A} is also derived from a probability distribution (i.e., $P_{\mathbf{A}}$), the range of feasible unknown
202 conditions remains characterized by $P_{\mathbf{U}}$. Importantly, extracting these conditions requires the
203 inference (empirical parameterization) of invariant rules (e.g., \mathbf{A}). While challenging, it has been
204 shown that this properties can be approximated with commonly available data, such as species
205 abundances or presence/absence data [14, 55–58]. We provide a case study in the last section.

206 Probability of causes

207 While observational data per se are not enough to obtain a causative knowledge about pollina-
208 tor communities, they can be translated into interventional distributions using causal-inference
209 techniques [15, 17]. Recently, promising causal-inference methods have been developed, such as
210 inverse modelling approaches [59, 60] or empirical dynamical modeling [61], but these methods
211 require large amounts of data which for several reasons can be difficult to obtain. To partially
212 circumvent this problem, we propose that probabilistic causal-inference approaches [15] used in
213 economics, social science, and medicine can be good candidates for inferring interventional dis-
214 tributions (i.e., how likely is a given cause to generate a target effect) in pollinator communities.

215 First and foremost, probabilistic causal inference requires a causal graph involving the set of
216 relevant variables (nodes) \mathbf{V} (e.g., $\mathbf{V} = \{X, Y\}$, $X \rightarrow Y$) upon which to test causal relationships
217 (edges) [15]. These graphs serve as a guideline (testable hypothesis) to understand the potential
218 paths linking causes and effects, which are necessary to study in order to eliminate spurious
219 associations (due to confounding and selection bias). In general, causal graphs should be drawn
220 based on expert knowledge or intuition about how the world works, and should not be drawn
221 based on the observed correlations on data (otherwise, it will be circular). These graphs act as a
222 hypothetical causal story, which can be followed after identifying and corroborating its testable
223 implications expressed as unconditional and conditional independencies between variables (in
224 causal-inference analysis, this is called d-separation of variables [15]). For instance, a lack of
225 correlation between two variables in any context does not immediately invalidate a potential
226 direct causal link (since we cannot be sure of having sampled all potential values within the sample
227 space); however, a lack of correlation in all contexts after conditioning by a potential confounder
228 (i.e., $X \not\perp\!\!\!\perp Y|\{\emptyset\}$, but $X \perp\!\!\!\perp Y|Z$) does support the hypothesized causal graph $X \leftarrow Z \rightarrow Y$ (i.e.,

229 no direct causal effect between X and Y). Remember that a correlation between two variables
230 is not enough evidence to support a potential causal link. Thus, causal graphs inform about
231 both the likely dependencies and established independencies between variables. If the data do
232 not corroborate the causal graph, then a new causal story must be drawn and tested.

233 Causal graphs are nonparametric by construction since they do not depend on the specific form
234 of causal relationships, they only specify the (lack of) existence of a causal relationship between
235 variables. While most of the standard work on probabilistic causal inference has been developed
236 for directed acyclic graphs (no mutual causality or feedback processes), cyclic graphs can also
237 be analyzed, especially under equilibrium conditions [62]. Importantly, these causal graphs need
238 to take into account both observed and unknown common factors (typically, these unknown
239 factors can be and are excluded from the graph if they are all mutually exclusive [15]). In
240 some situations, the potential confounding effects of unknown factors (context) can be eliminated
241 using standard causal-inference techniques (e.g., using the so-called front-door and back-door
242 criteria, or using latent variables [15]). Note that latent variables are typically used in structural
243 equation modeling assuming linearity for all variables [17, 63]. However, when these unknown
244 common factors cannot be eliminated or linearity cannot be assumed or validated, we propose to
245 approximate these factors by deriving them from theoretical models (as explained in the previous
246 section). Specifically, these unknown factors can be characterized by $P_{\mathbf{U}}$, an expected value, or
247 can be transformed into binary variables using heuristic rules [52, 54, 57]. We provide a case
248 study in the following section.

249 The translation from observational distributions to interventional distributions is rooted on *do*-
250 calculus [15], which are the rules for moving from observations to interventions using the causal
251 graph. That is, causal inference moves (whenever identifiable) from the probabilistic association
252 $P(y|x)$ to the probabilistic causal association $P(y|\text{do}(x))$, where y is the value of the potential
253 effect Y and x is the value taken after the intervention on the inferred cause X . The nomenclature
254 $\text{do}(x)$ implies that we are not just merely observing X to take the value of x , but we need to
255 make it have it (e.g., removing a species from a community). This action is then represented in a
256 modified causal graph by eliminating all the incoming edges (causes) from an intervened variable
257 (since its value is no longer dependent on mechanisms, but on a given action). It is typically
258 assumed that mechanisms $P(y|\text{do}(x))$ are independent from each other, invariant, and follow the
259 arrow of time (i.e., causes before effects), allowing to apply probabilistic Markov properties (i.e.,
260 each variable is independent from its non-causal variables—known as ancestors—given its causes—
261 known as parents [15]).

262 Given a directed acyclic causal graph G and disjoint variables X, Y, Z and W (these variables
263 can also be empty sets), *do*-calculus involves three rules to move from observational to interven-
264 tional distributions (see Figure 1) [15]: (1) Insertion/deletion of observations: $P(y|\text{do}(x), z, w) =$

265 $P(y|do(x), w)$ if $(Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}}}$, where $G_{\overline{X}}$ is graph G after the removal of all the incoming edges
 266 to X . This rule establishes the conditions under which it is possible to remove conditional vari-
 267 ables from the analysis. (2) Action/observation exchange: $P(y|do(x), do(z), w) = P(y|do(x), z, w)$
 268 if $(Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{XZ}}}$, where $G_{\overline{XZ}}$ is graph $G_{\overline{X}}$ after the removal of all the outgoing edges from Z .
 269 This rule establishes the conditions under which it is possible to replace additional actions (acting
 270 as confounders) with observational data. (3) Insertion/deletion of actions: $P(y|do(x), do(z), w) =$
 271 $P(y|do(x), w)$ if $(Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{XZ(W)}}}$, where $Z(W)$ is the set of Z -variables that are not ancestors
 272 of any W -variable in $G_{\overline{X}}$. This rule establishes the conditions under which it is possible to remove
 273 additional actions (acting as confounders) from the analysis. Note that while path analysis [17]
 274 can be used instead of do-calculus, only the latter is a nonparametric framework that can be used
 275 with any sort of data without making any assumptions.

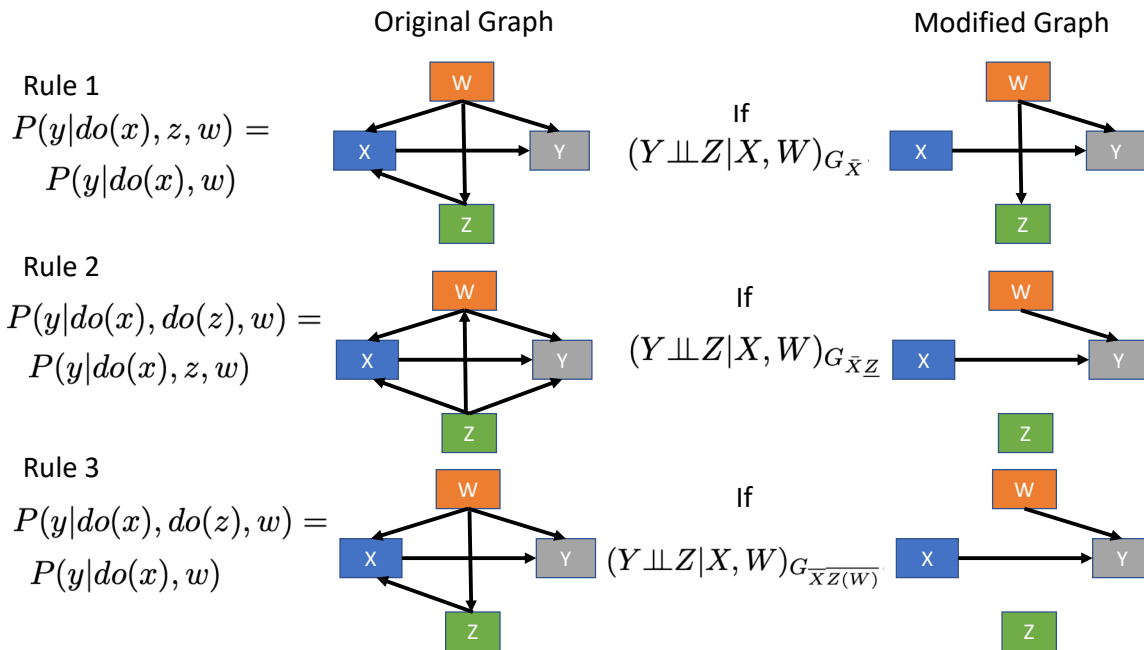


Figure 1: **do-calculus**. The translation from interventional $P(do(x))$ to observational $P(x)$ distributions can be achieved following the rules of do-calculus [15]. The figure depicts the three do-calculus rules on a graph G with disjoint variables X, Y, Z and W (see main text). Rule 1 is used for insertion/deletion of observations. Rule 2 is used for action/observation exchange. Rule 3 is used for insertion/deletion of actions. Here, $G_{\overline{X}}$ is graph G after the removal of all the incoming edges to X , $G_{\overline{XZ}}$ is graph $G_{\overline{X}}$ after the removal of all the outgoing edges from Z , and $Z(W)$ is the set of Z -variables that are not ancestors of any W -variable in $G_{\overline{X}}$. Note that $\perp\!\!\!\perp$ and $|$ denote independence and conditional on, respectively. The graphs in the left column vary for illustration purposes of each rule.

276 Case study

277 We illustrate some of the concepts above using the following example. Figure 2 depicts a hypo-
 278 theoretical, directed, acyclic, causal graph to study the within-season pollinator abundance dynamics

279 of a pollinator community [30, 64]. Specifically, in the example, we study how the relative abun-
280 dance of flowering plants at a given time t (noted as A and measured as the ratio between the
281 number of plant species and pollinator species at time t) affects the rate of change of the pollina-
282 tor community at time $t + 1$ (noted as B and measured as the absolute difference in the pollinator
283 community between time $t + 1$ and t , and divided by the observation at time t , providing a
284 detrended measure). In addition, the causal graph (Fig. 2) assumes that temperature affects
285 both A and B (written as C and measured as the mean temperature at time t). Note that C
286 also works as a trend factor. Finally, we also assume that unknown factors D (the context) act
287 as confounding effects of A and B . Following the concepts expressed in the previous section, we
288 propose (see below for details) to quantify the unknown factors D using synthetic data derived
289 from the linear LV model (i.e., $P_{\mathbf{U} \subseteq \mathbf{r}}$) leading to the presence of the observed pollinator com-
290 munity at time t (i.e., $N_I^* > 0$). Integrating observational and synthetic data, the graph in Fig.
291 2 is complete and informs us about the variables that need to be blocked (controlled for) using
292 do-calculus in order to infer the cause-effect relationships between observed variables. Note that
293 it is assumed that each of these variables is random in the sense that they are all affected by
294 mutually exclusive independent noise, allowing us to omit this other type of variables from the
295 causal graph [15].

296 To put numbers to this example, we use publicly available data recording species interactions
297 between pollinators and flowering plants on a daily basis (whenever weather allowed) in a high-
298 arctic site during the springs of 1996 and 1997 [30, 64]. These data allow us to directly measure
299 variables A , B , and C above for a given observed day t . To measure the theoretical context (D)
300 for each day t , we first inferred the daily interaction matrices \mathbf{A}_t and then measure the fraction
301 of conditions compatible with the persistence of all observed pollinators $\omega(\mathbf{A}_t)$. To infer \mathbf{A}_t , we
302 use a niche-based inference [58, 65], which is one of the simplest methods yet well ecologically
303 motivated. Specifically, we use the monopartite projection $\mathbf{M}_t = \mathbf{B}_t^T \mathbf{B}_t$, where \mathbf{B}_t is the binary
304 matrix for day t formed by the observed pollinators as columns and observed plants as rows. This
305 binary matrix has entries $B_{ki} = 1$ if the pollinator i is observed interacting with plant k , otherwise
306 $B_{ki} = 0$. In turn, the off-diagonal entries of \mathbf{M}_t correspond to the number of plant resources
307 shared between two pollinator species. The higher the resource overlap between pollinators i and
308 j (i.e., the value of M_{ij}), the higher their level of competition. By normalizing the entries of \mathbf{M}_t
309 by the sum of their column ($A_{ij} = \frac{M_{ij}}{\sum M_{ij}}$), we infer a pollinator competition matrix \mathbf{A}_t for each
310 time t .

311 To infer $\omega(\mathbf{A}_t)$ [30], we calculate the fraction of intrinsic growth rates ($\mathbf{U} \subseteq \mathbf{r}$) leading to the daily
312 set of competing pollinators according to a (tractable and scalable) linear LV model. Specifically,
313 we calculate this as:

$$\omega(\mathbf{A}_t) = \left(\frac{2^{S_t} \text{vol}(D_F(\mathbf{A}_t) \cap \mathbb{B}^{S_t})}{\text{vol}(\mathbb{B}^{S_t})} \right)^{\frac{1}{S_t}},$$

314 where $\text{vol}(\mathbb{B}^S)$ is the volume of the normalized S_t -dimensional parameter space of intrinsic growth
 315 rates (\mathbf{r}) at day t , 2^{S_t} normalizes the parameter space to the positive orthant (because for sim-
 316 plification we are summarizing the pollinator community as a competition system, all intrinsic
 317 growth rates are restricted to positive values), and $\text{vol}(D_F(\mathbf{A}_t) \cap \mathbb{B}^S)$ corresponds to the
 318 volume of the intersection of the the parameter space with the feasibility domain: $D_F(\mathbf{A}_t) =$
 319 $\left\{ \mathbf{U} = N_1^* \mathbf{v}_1 + \dots + N_S^* \mathbf{v}_S, \text{ with } N_1^*, \dots, N_S^* > 0 \right\}$, where \mathbf{v}_i is the i th column vector of the in-
 320 teraction matrix \mathbf{A}_t [54]. Thus, $\omega(\mathbf{A}_t) \in [0, 1]$ is a probabilistic measure, which can be efficiently
 321 computed and compared across dimensions [30, 54].

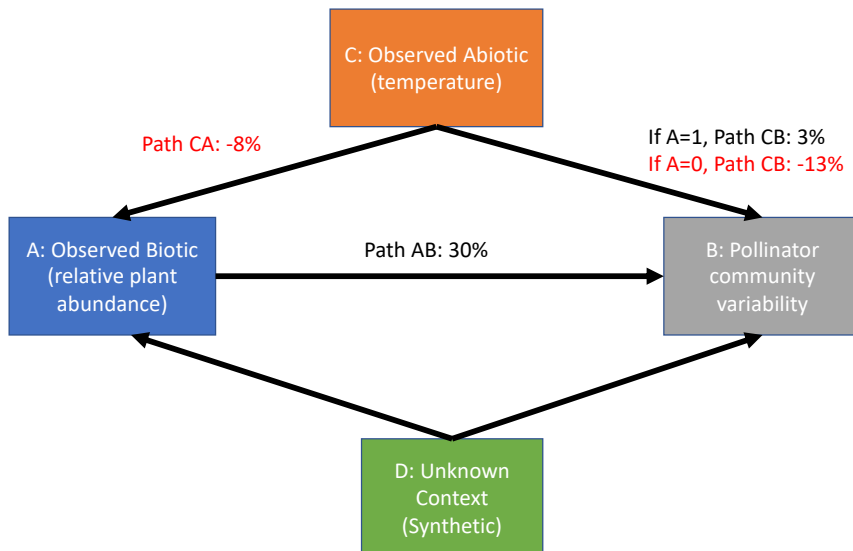


Figure 2: Illustrative example of cause-effect relationships of a phenological process in a pollinator community. However, this effect needs to be separated from potential confounders. The figure depicts a directed acyclic causal graph, where each box (node) corresponds to a random variable, and each edge corresponds to a direct causal effect. We consider that each causal relationship is autonomous and independent from the others. Each node is a random variable since it is also assumed that mutually exclusive random noise affects each node. Following do-calculus rules (see text and Fig. 1), for three paths, we show the estimated change in probability of observing a high value (above the median of the population) given a high value of its direct cause (see text). The variables in this graph should not be always equated to the variables in Figure 1. For example, variable C can be equivalent to variable X or Z in Figure 1 depending on the rule applied.

322 Similar to path analysis in structural equation modeling [17, 63], to apply probabilistic causal
 323 inference with continuous data, it can be possible to use linear regressions (or Pearson correlations)
 324 if it is assumed that the effects are linear, monotonic, and noise is Gaussian. Spearman rank
 325 correlations can be used if at least monotonicity is achieved. Instead, nonparametric tools can

326 be used whether or not these assumptions above are fulfilled. While nonparametric tools provide
327 generality and should be preferred, their application to continuous data can be rather challenging.
328 Thus, whenever possible, the data can be discretized [15]. Here, for illustration purposes, we
329 transform all our variables into binary values, using the median of each variable (per year) as
330 the cut-off value: values higher than the median are considered one, otherwise zero. While this
331 may be perceived as a disadvantageous simplification, it actually allows us to efficiently work on
332 a general nonparametric framework (i.e., using probability distributions).

333 We test the causal graph shown in Figure 2. Here, the only testable d-separation (conditional
334 or unconditional) is between temperature (C) and context (D). That is, there is no direct
335 path between these two variables, and their path gets naturally blocked (no need to condition on
336 anything) by A and B , which act as colliders. This d-separation can be tested by the unconditional
337 independence as $P(d|c) = P(d)$. Using a G^2 -test (χ^2 -test can also be used for binary data or
338 permutation tests [15, 17]), we found no statistical relationship between C and D ($p = 0.39$,
339 lower values indicate dependence). Note that if the hypothesis would not have been supported
340 by d-separation, a new causal graph must be drawn and tested. Below, we compute the effects of
341 temperature on the relative abundance of flowering plants (Path CA), the effect of temperature
342 on community variability (Path CB), and the effect of relative abundance of flowering plants on
343 community variability (Path AB).

344 The interventional distribution (probability of cause) of Path CA is written as $P(a|do(c))$. This
345 causal relationship can be inferred using observational distributions following rule 2 of do-calculus.
346 That is, we can write $P(a|do(c)) = P(a|c)$ by setting $Y = A$, $Z = C$, and $W = X = \emptyset$ in Figure
347 1. Because we are using binary variables, the average causal effect [15] of c on a (i.e., ACE_{CA})
348 is given by $\frac{\partial}{\partial w} E[A|do(c)]$ and can be written as $ACE_{CA} = P(a = 1|c = 1) - P(a = 1|c = 0)$. We
349 found that $ACE_{CA} = -0.08$, meaning that if temperature is high (i.e., above the population me-
350 dian) there is a decrease in probability of 8% that the relative plant abundance will be high (i.e.,
351 above its population median). However, using a G^2 test, we found that this effect is not largely
352 different ($p = 0.56$) from what would be expected by chance alone given the data. In turn, the in-
353 terventional distribution of Path CB can be calculated as $P(b|do(c), do(a))$. Note that Path CB is
354 mediated by A , which needs to be controlled for. However, conditioning (i.e., $P(b|do(c), x)$) opens
355 the collider between C and D , creating a spurious association between C and B . To eliminate this
356 noise, it is then necessary to intervene on A (i.e., $do(a)$). Using marginalization and the Markov
357 property, we can write $P(b|do(c), do(a)) = \sum_d P(b, d|do(c), do(a)) = \sum_d P(b|do(c), do(a), d)P(d)$.
358 Following rule 2 twice (setting first $Z = A$, $Y = B$, $X = C$, and $W = D$; and second $Z = C$,
359 $Y = B$, $X = \emptyset$, and $W = \{A, D\}$ in Fig. 1), we can write $\sum_d P(b|c, a, d)P(d)$. In this case,
360 we can perform two separated analyses: one for $a = 1$ and the other for $a = 0$. We found that
361 for $a = 1$, $ACE_{CB} = 0.03$ (G^2 test: $p = 0.43$). While for $a = 0$, $ACE_{CB} = -0.13$ (G^2 test:

362 $p = 0.008$). This implies that under high flower abundance, temperature has almost no effect on
363 pollinator variability. Instead, under low flower abundance, if temperature is high (i.e., above the
364 population median), there is a decrease in probability of 13% that the variability of the pollinator
365 community will be also high (i.e., above its population median).

366 Finally, following the methodologies above, we calculate the effect of relative plant abundance
367 on community variability (Path AB) as $P(b|do(a)) = \sum_{cd} P(b|a, c, d)P(c, d)$. We found that
368 $ACE_{AB} = 0.30$ (G^2 test: $p = 0.06$), meaning that if relative plant abundance is high (i.e.,
369 above the population median) there is an increase in the probability of 30% that the community
370 variability will be high (i.e., above its population median). It is worth mentioning that if we do
371 not take into account the context (D), the causal effect of A (relative flower abundance) on B
372 (pollinator community variability) can be overestimated $ACE_{AB} = 0.86$ (G^2 test: $p = 0.003$),
373 leading to potential prediction errors of interventions. It is also important to mention that a
374 linear multivariate regression of B on all the other three variables (using normalized data instead
375 of binary) produce qualitatively similar results as the ones reported above. While this equivalence
376 between nonparametric and parametric methods is not expected to be always true [15], working
377 under a causal hypothesis (as we have done here) can establish a more informative regression
378 analysis that can then be translated into causal analysis under the assumption of linearity.

379 This example is not intended to demonstrate a general effect and serves only for illustration
380 purposes. For example, we try to explain a fairly simple community metric such as changes
381 in overall relative abundance. Furthermore, many more variables can be explicitly taken into
382 account (instead of being summarized in the unknown confounding factors), such as abundance
383 of pathogens, herbivores, chemical compounds, humidity, etc, and it is important to identify the
384 main players in line with the hypothesized causal graphs. Moreover, it is important to note
385 that the theoretical model has also sensible assumptions, such as that resource overlap among
386 pollinators is a good proxy of competition. We hope future work can build on this to establish
387 causal knowledge at the pollinator community-level.

388 Conclusions

389 It has long been recognized that causation does not always coincides with correlation. This
390 premise has been extensively applied when studying the behavior (i.e., variables) of complex
391 natural systems, where multiple factors can be responsible for the patterns observed in nature.
392 This has not been an exception when investigating pollinator communities. As a consequence,
393 the majority of work has carefully stated correlations, which respond to what do we see in nature.
394 However, in the face of rapid environmental change, we need to take bolder research programs and
395 answer the questions of why and when the behavior of pollinator communities is affected. These
396 goals can be achieved by conducting experimental studies. Nevertheless, manipulating all factors

397 related to the behavior of entire pollination communities can be unrealistic. Instead, these goals
398 can be achieved by using causal-inference techniques. Yet, very often these techniques cannot
399 be applied due to the nature of the causal story and the unknown/unmeasured factors acting
400 as confounders. While not exhaustive, here we have provided a brief overview of how to apply
401 probabilistic causal inference from the integration of observational and synthetic data. We propose
402 that synthetic data can be used as a proxy for unknown confounding factors by deriving them
403 from theoretical models that attain the desired properties of tractability (provide a systematic
404 link between model parameters and solutions) and scalability (can be applied across dimensions).
405 At the very least, we hope this overview can illustrate that a causal probabilistic analysis can
406 allow us to speak the causal language in pollination studies that for long has been prevented by
407 the dominance of multivariate regressions and meta-analyses without causal hypotheses [32].

408 **Acknowledgments** Funding to SS was provided by NSF grant No. DEB-2024349. IB and OG
409 acknowledges funding from the Simplex project (PRPCGL2017-92436-EXP). OG acknowledges
410 financial support provided by the Spanish Ministry of Economy and Competitiveness (MINECO)
411 and by the European Social Fund through the Ramón y Cajal Program (RYC-2017-23666). RPR
412 acknowledges funding from the Swiss National Science Foundation, grant no. 31003A_182386.
413 PZ acknowledges the Swiss National Science Foundation Spark scheme, under grant number
414 CRSK-3_196506.

415 **Competing financial interests** The authors declare no competing financial interests.

416 **Author contributions** SS designed and performed the study. All authors contributed with
417 ideas and wrote the manuscript.

418 **Data accessibility** The data and R code supporting the results can be found at [https://](https://github.com/MITEcology/Saavedra_etal_causal_example)
419 github.com/MITEcology/Saavedra_etal_causal_example.

References

- [1] Winfree R, Bartomeus I, others (2011) Native pollinators in anthropogenic habitats. *Annual Review of Ecology*.
- [2] Ollerton J, Winfree R, Tarrant S (2011) How many flowering plants are pollinated by animals? *Oikos* 120:321–326.
- [3] Klein AM, et al. (2007) Importance of pollinators in changing landscapes for world crops. *Proc. of the Royal Society B* 274:303–313.
- [4] Potts SG, et al. (2016) Safeguarding pollinators and their values to human well-being. *Nature* 540:220–229.
- [5] Goulson D, Nicholls E, Botías C, Rotheray EL (2015) Bee declines driven by combined stress from parasites, pesticides, and lack of flowers. *Science* 347.
- [6] Garibaldi LA, et al. (2017) Farming Approaches for Greater Biodiversity, Livelihoods, and Food Security. *Trends in Ecology & Evolution* 32:68–80.
- [7] Menz MHM, et al. (2011) Reconnecting plants and pollinators: challenges in the restoration of pollination mutualisms. *Trends in Plant Science* 16:4–12.
- [8] Scheper J, et al. (2015) Local and landscape-level floral resources explain effects of wildflower strips on wild bees across four European countries. *J. Appl. Ecol.*
- [9] Klein AM, Boreux V, Fornoff F, Mupepele AC, Pufal G (2018) Relevance of wild and managed bees for human well-being. *Current Opinion in Insect Science* 26:82–88.
- [10] Soroye P, Newbold T, Kerr J (2020) Climate change contributes to widespread declines among bumble bees across continents. *Science* 367:685–688.
- [11] Henry M, Rodet G (2018) Controlling the impact of the managed honeybee on wild bees in protected areas. *Scientific Reports* 8:9308.
- [12] Dynes TL, Berry JA, Delaplane KS, Brosi BJ, de Roode JC (2019) Reduced density and visually complex apiaries reduce parasite load and promote honey production and overwintering survival in honey bees. *PLoS One* 14:e0216286.
- [13] Graystock P, Goulson D, Hughes WOH (2014) The relationship between managed bees and the prevalence of parasites in bumblebees. *PeerJ* 2:e522 Publisher: PeerJ Inc.
- [14] Bartomeus I, Saavedra S, Rohr RP, Godoy O (2021) Experimental evidence of the importance of multitrophic structure for species persistence. *Proceedings of the National Academy of Sciences* 118:e2023872118.
- [15] Pearl J (2009) *Causality* (Cambridge Univ. Press, Cambridge).
- [16] Kimmel K, Dee LE, Avolio ML, Ferraro PJ (2021) Causal assumptions and causal inference in ecological experiments. *Trends in Ecol. Evol.* doi.org/10.1016/j.tree.2021.08.008.
- [17] Shipley B (2016) *Cause and Correlation in Biology* (Cambridge University Press).
- [18] Brosi BJ, Briggs HM (2013) Single pollinator species losses reduce floral fidelity and plant reproductive function. *Proceedings of the National Academy of Sciences* 110:13044.
- [19] Biella P, et al. (2018) Experimental loss of generalist plants reveals alterations in plant-pollinator interactions and a constrained flexibility of foraging. *bioRxiv* p 279430.
- [20] Zu P, et al. (2020) Information arms race explains plant-herbivore chemical communication in ecological communities. *Science* 368:1377–1381.

- [21] Zu P, et al. (2021) Pollen sterols are associated with phylogeny and environment but not with pollinator guilds. *New Phytologist* 230:1169–1184.
- [22] Kantsa A, et al. (2017) Community-wide integration of floral colour and scent in a mediterranean scrubland. *Nature Ecology & Evolution* 1:1502.
- [23] Adler LS, Barber NA, Biller OM, Irwin RE (2020) Flowering plant composition shapes pathogen infection intensity and reproduction in bumble bee colonies. *Proceedings of the National Academy of Sciences* 117:11559–11565.
- [24] Dukas R, Morse DH (2003) Crab spiders affect flower visitation by bees. *Oikos* 101:157–163.
- [25] Frund J, Zieger SL, Tschardt T (2013) Response diversity of wild bees to overwintering temperatures. *Oecologia* 173:1639–1648.
- [26] Zaragoza-Trello C, Vilá M, Botías C, Bartomeus I (2020) Interactions among global change pressures act in a non-additive way on bumblebee individuals and colonies. *Functional Ecology* 35:420–434.
- [27] Rundlof M, et al. (2015) Seed coating with a neonicotinoid insecticide negatively affects wild bees. *Nature* 521:77–80.
- [28] Zaragoza-Trello C, Vilá M, Bartomeus I (2021) Interaction between warming and landscape foraging resource availability on solitary bee reproduction. *Journal of Animal Ecology* doi.org/10.1111/1365-2656.13559.
- [29] Clark T, Luis AD (2020) Nonlinear population dynamics are ubiquitous in animals. *Nature ecology & evolution* 4:75–81.
- [30] Song C, Saavedra S (2018) Structural stability as a consistent predictor of phenological events. *Proc. R. Soc. B* 285:20180767.
- [31] Reichenbach H (1956) *The direction of time* (The University of California Press).
- [32] Bareinboim E, Pearl J (2016) Causal inference and the data-fusion problem. *PNAS* 113:7345–7352.
- [33] Shmueli G (2010) To explain or to predict? *Statistical Science* 25:289–310.
- [34] Strogatz SH (2014) *Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering* (Westview press).
- [35] Svirzhev YM, Logofet DO (1983) *Stability of Biological Communities* (Mir Publishers).
- [36] Case TJ (2000) *An Illustrated Guide to Theoretical Ecology* (Oxford Univ. Press, Oxford).
- [37] Valdovinos FS (2019) Mutualistic networks: moving closer to a predictive theory. *Ecology letters* 22:1517–1534.
- [38] Banks HT, et al. (2017) Modeling bumble bee population dynamics with delay differential equations. *Ecological Modelling* 351:14–23.
- [39] Haussler J, Sahlin U, Baey C, Smith HG, Clough Y (2017) Pollinator population size and pollination ecosystem service responses to enhancing floral and nesting resources. *Ecol. Evol.*
- [40] AlAdwani M, Saavedra S (2020) Ecological models: higher complexity in, higher feasibility out. *J. of the Roy. Soc. Interface* 17:20200607.
- [41] Lonsdorf E, et al. (2009) Modelling pollination services across agricultural landscapes. *Annals of Botany* 103:1589–1600.

- [42] Valdovinos FS, et al. (2016) Niche partitioning due to adaptive foraging reverses effects of nestedness and connectance on pollination network stability. *Ecol. Lett.* 19:1277–1286.
- [43] Peralta G, Stouffer DB, Bringa EM, Vázquez DP (2020) No such thing as a free lunch: interaction costs and the structure and stability of mutualistic networks. *Oikos* 129:503–511.
- [44] Abel NH (1826) Démonstration de l'impossibilité de la résolution algébrique des équations générales qui passent le quatrième degré. *Journal für die reine und angewandte Mathematik* 1:65–96.
- [45] AlAdwani M, Saavedra S (2019) Is the addition of higher-order interactions in ecological models increasing the understanding of ecological dynamics? *Mathematical Biosciences* 315:108222.
- [46] Mayfield MM, Stouffer DB (2017) Higher-order interactions capture unexplained complexity in diverse communities. *Nature Ecology & Evolution* 1:0062.
- [47] Martyn JTE, et al. (2021) Identifying 'useful' fitness models: balancing the benefits of added complexity with realistic data requirements in models of individual plant fitness. *The American Naturalist* 197:415–433.
- [48] Song C, Barabás G, Saavedra S (2019) On the consequences of the interdependence of stabilizing and equalizing mechanisms. *The American Naturalist* 194:627–639.
- [49] Lotka AJ (1920) Analytical note on certain rhythmic relations in organic systems. *Proceedings of the National Academy of Sciences* 6:410–415.
- [50] Volterra V (1926) Fluctuations in the abundance of a species considered mathematically. *Proceedings of the National Academy of Sciences* 118:558–560.
- [51] Rohr RP, et al. (2016) Persist or produce: a community trade-off tuned by species evenness. *Am. Nat.* 188:411–422.
- [52] Saavedra S, Medeiros LP, AlAdwani M (2020) Structural forecasting of species persistence under changing environments. *Ecology Letters* 23:1511–1521.
- [53] Saavedra S, Rohr RP, Olesen JM, Bascompte J (2016) Nested species interactions promote feasibility over stability during the assembly of a pollinator community. *Ecology and Evolution* 6:997–1007.
- [54] Song C, Rohr RP, Saavedra S (2018) A guideline to study the feasibility domain of multi-trophic and changing ecological communities. *J. of Theoretical Biology* 450:30–36.
- [55] Xiao Y, et al. (2017) Mapping the ecological networks of microbial communities. *Nature Communications* 8:1–12.
- [56] Maynard DS, Miller ZR, Allesina S (2020) Predicting coexistence in experimental ecological communities. *Nature Ecology & Evolution* 4:91–100.
- [57] Deng J, Angulo MT, Saavedra S (2021) Generalizing game-changing species across microbial communities. *ISME Communications* 1:1–8.
- [58] Cenci S, Montero-Castaño A, Saavedra S (2018) Estimating the effect of the reorganization of interactions on the adaptability of species to changing environments. *J. of Theor. Bio.* 437:115–125.
- [59] Ives AR, Dennis B, Cottingham K, Carpenter S (2003) Estimating community stability and ecological interactions from time-series data. *Ecological monographs* 73:301–330.
- [60] Almaraz P, Oro D (2011) Size-mediated non-trophic interactions and stochastic predation drive assembly and dynamics in a seabird community. *Ecology* 92:1948–1958.

- [61] Sugihara G, et al. (2012) Detecting causality in complex ecosystems. *science* 338:496–500.
- [62] Mooij JM, Janzing D, Scholkopf B (2013) From ordinary differential equations to structural causal models: the deterministic case. *UAI'13: Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence* p 440–448.
- [63] Wei N, et al. (2021) Pollinators contribute to the maintenance of flowering plant diversity. *Nature* doi.org/10.1038/s41586-021-03890-9.
- [64] Olesen JM, Stefanescu C, Traveset A (2011) Strong, long-term temporal dynamics of an ecological network. *PLoS One* 6:e26455.
- [65] Song C, Altermatt F, Pearse I, Saavedra S (2018) Structural changes within trophic levels are constrained by within-family assembly rules at lower trophic levels. *Ecology Letters* 21:1221–1228.