

1 THINKING OUTSIDE THE BOX -  
2 PREDICTING BIOTIC INTERACTIONS IN  
3 DATA-POOR ENVIRONMENTS

4 *DAVID BEAUCHESNE*<sup>1\*</sup>, *PHILIPPE DESJARDINS-PROULX*<sup>2</sup>,  
5 *PHILIPPE ARCHAMBAULT*<sup>3</sup>, and *DOMINIQUE GRAVEL*<sup>2</sup>

6 \* email: [david.beauchesne@uqar.ca](mailto:david.beauchesne@uqar.ca)

7 <sup>1</sup> *Université du Québec à Rimouski*

8 <sup>2</sup> *Université de Sherbrooke*

9 <sup>3</sup> *Université Laval*

10 September 19, 2016

RUNNING TITLE:  
PREDICTING BIOTIC INTERACTIONS IN DATA-POOR ENVIRONMENTS

# 1 Abstract

Large networks of ecological interactions, such as food webs, are complex to characterize, be it empirically or theoretically. The former requires exhaustive observations, while the latter generally requires ample data to be validated. We therefore wondered whether readily available data, namely empirically described interactions in a variety of ecosystems, could be combined to predict species interactions in data deficient ecosystems. To test this, we built a biotic interactions catalogue from a collection of 94 empirical food webs, detailed predator-prey interaction databases and interactions from the Global Biotic Interactions (GloBI) database. We used an unsupervised machine learning method to predict interactions between any given set of taxa, given pairwise taxonomic proximity and known consumer and resource sets found in the interaction catalogue. Results suggest that pairwise interactions can be predicted with high accuracy. Although conclusions are seemingly dependent on the comprehensiveness of the catalogue knowledge of taxonomy was found to complement well the catalogue and improve predictions, especially as empirical information available diminished. Given its high accuracy, this methodology could democratize the use of food webs and network level descriptors in remote location where empirical data is hard to gather. Network characteristics could then be efficiently evaluated and correlated to levels of environmental stressors in order to improve vulnerability assessments of ecosystems to global changes, opening promising avenues for further research and for management initiatives.

**Keywords:** Interactions, machine learning, food webs, K-nearest neighbour, taxonomy, St. Lawrence

## 2 Introduction

Large networks of ecological interactions, such as food webs, are complex to characterize (Polis, 1991; Martinez, 1992; Pascual and Dunne, 2006). Empirical descriptions require exhaustive observations, while theoretical inference generally requires ample data to be validated. For this reason, studies focusing on communities of interacting species remain understudied, even though we acknowledge the importance of considering the reticulated nature of complex networks (Ings et al., 2009; Tylianakis et al., 2008). When time is of the essence, the long term studies required quickly become impractical and the use of network level approaches is relegated to the sideline.

Alternatively, a currently growing approach is to predict interactions using proxies such as functional traits, phylogenies and spatial distributions (e.g. Morales-Castilla et al., 2015; Bartomeus et al., 2016). For example, multiple traits can play a significant role in community dynamics and influence the presence and intensity of biotic interactions, like the influence of body size on predator-prey interactions, a literal take on *big fish eats small fish* (Cohen et al., 2003; Brose et al., 2006; Gravel et al., 2013). However, the time required

53 to gather the necessary data to apply those methods may still be restrictive, or  
 54 the data be unavailable altogether, so much so that other methods have been  
 55 developed to fill the gaps in knowledge (e.g. Schrodtt et al., 2015).

56 We therefore wondered whether more readily available data could be used to  
 57 infer interactions in data deficient ecosystems. There is an increasing amount  
 58 of data describing worldwide species interactions, some freely available through  
 59 the Global Biotic Interactions (GloBI) database (Poelen et al., 2014). Another  
 60 readily available piece of information on species is their taxonomy, through  
 61 initiatives like the World Register of Marine Species (WoRMS; Bailly et al.,  
 62 2016). More than simple nomenclature, evolutionary processes are thought  
 63 to influence consumer-resource relationships (Mouquet et al., 2012; Rohr and  
 64 Bascompte, 2014) so that taxonomically related species would be more likely  
 65 to share similar types of both consumers and resources (Eklöf et al., 2012;  
 66 Morales-Castilla et al., 2015; Gray et al., 2015). Based on that assumption,  
 67 taxonomy might be useful in predicting interactions for species lacking detailed  
 68 information on their biology, but which have a taxonomically related species  
 69 for which such information is available. The objective of this work is thus to  
 70 combine empirical biotic interactions originating from a variety of ecosystems  
 71 with taxonomic relatedness to predict interactions in data deficient ecosystems.  
 72 As an example, we compare the observed interactions in the southern Gulf of  
 73 St. Lawrence (SGSL; Savenkoff et al., 2004) with predictions made using our  
 74 approach.

## 75 3 Methods

76 The objective of our methodology is to predict the interactions between all pairs  
 77 of taxa within an arbitrary set  $N_1$ , using a set of taxa  $N_0$  with empirically de-  
 78 scribed interactions from which we can extract pairs of consumers and resources  
 79 and their taxonomy. We couple the use of empirical data with an unsupervised  
 80 machine learning method to achieve this.

### 81 3.1 Biotic interaction catalogue

82 We built a biotic interaction catalogue to serve as a set of taxa  $N_0$  for with  
 83 empirically described interactions. The empirical data used to construct the  
 84 interaction catalogue was gathered in two successive steps. The first consisted  
 85 of gathering data from a collection of 94 empirical food webs from which we  
 86 extracted pairwise taxa interactions (see Brose et al., 2005; Kortsch et al., 2015;  
 87 University of Canberra, 2016 for more information). We also used a detailed  
 88 predator-prey interaction database describing trophic relationships between ma-  
 89 rine fishes and their prey (Barnes et al., 2008). From these datasets, only in-  
 90 teractions between taxa at the taxonomic scale of the family or higher were  
 91 selected for inclusion in the catalogue. Data used came exclusively from marine  
 92 and coastal ecosystems and encompassed a wide variety of organisms: fungi, al-  
 93 gae, parasites, phytoplankton, zooplankton, benthic and pelagic invertebrates,

demersal and pelagic fishes, marine birds and marine mammals.

As empirical food webs are vastly dominated by non-interactions (96%), these datasets yielded a highly skewed distribution of interactions vs non-interactions. To counterbalance this, the second step of data compilation consisted of extracting observed interactions from the Global Biotic Interaction (GloBI) database (Poelen et al., 2014), which describes binary interactions for a wide range of taxa worldwide. We extracted all trphic interactions available on GloBI for species belonging to the families of taxa identified through step 1. Interactions were extracted using the rGloBI package in R (Poelen et al., 2015). As per step 1, only interactions between taxa at the taxonomic scale of the family or higher were retained.

The nomenclature used between datasets and food webs varied substantially. Taxa names thus had to be verified, modified according to the scientific nomenclature and validated. This process was performed using the Taxize package in R (Chamberlain and Szöcs, 2013; Chamberlain et al., 2014) and manually verified for errors. The same package was used to extract the taxonomy of all taxa for which interactions were obtained in previous steps. The complete R code and data used to build the catalogue is available at [https://github.com/david-beauchesne/Interaction\\_catalog](https://github.com/david-beauchesne/Interaction_catalog).

### 3.2 Unsupervised machine learning

We use the  $K$ -nearest neighbor (KNN) algorithm (Murphy, 2012) to predict pairwise interactions for a set of taxa  $S$ . The KNN algorithm predicts missing entries or proposes additional entries by a majority vote based on the  $K$  nearest (i.e. most similar) entries (see Box 1 for an example). In this case, taxa are described by a set of resources when considered as a consumer, a set of consumers when considered as a resource and their taxonomy (i.e. kingdom, phylum, class, order, family, genus, species). Similarity between taxa was evaluated using the Tanimoto similarity measure, which compares two vectors  $x$  and  $y$  with  $n = |\mathbf{x}| = |\mathbf{y}|$  elements, and is defined as the size of the intersection of two sets divided by their union:

$$\text{tanimoto}(\mathbf{x}, \mathbf{y}) = \frac{|\mathbf{x} \cap \mathbf{y}|}{|\mathbf{x} \cup \mathbf{y}|}, \quad (1)$$

where  $\cap$  is the intersect and  $\cup$  the union of the vectors. Adding a weighting scheme, we can measure the similarity using two different sets of vectors:

$$\text{tanimoto}_t(x, y, w_t) = w_t \text{tanimoto}(\mathbf{T}_C \mathbf{x}, \mathbf{T}_C \mathbf{y}) + (1 - w_t) \text{tanimoto}(\mathbf{T}_T \mathbf{x}, \mathbf{T}_T \mathbf{y}), \quad (2)$$

where  $w_t$  is the weight given to sets of resources or consumers for consumer  $T_C$  and resources  $T_R$  (replace  $T_C$  in equation 2), for taxa  $x$  and  $y$ .  $T_T$  is the taxonomy of taxa  $x$  and  $y$ . When  $w_t = 0$  only resource or consumer sets are used to compute similarity, while  $w_t = 1$  solely uses taxonomy. This approach

130 to consider the relative contribution of two sets of vectors to the Tanimoto  
 131 similarity was developed by Desjardins-Proulx et al. (2016).

### 132 3.3 Predicting interactions

133 The algorithm was built on a series of logical steps that ultimately predicts  
 134 a candidate resources list  $C_R$  for each taxon in  $N_1$  based on empirical data  
 135 available and the similarity among consumers and among resources (Figure 1).  
 136 For all consumer taxa  $T_C$  in  $N_1$ , the algorithm first verifies, for all resources in  
 137 resource set  $T_R$ , if they are found the  $N_0$  (Step S1, Figure 1). When it does,  
 138 all  $T_R$  taxa that are also in  $N_1$  are added as predicted resources for  $T_C$  (Steps  
 139 S2-S3). This corresponds to what we refer to as the catalogue contribution to  
 140 resource predictions. In essence, two taxa in  $N_1$  that are known to interact  
 141 through empirical data in the catalogue are automatically assumed to interact  
 142 in  $N_1$ .

143 Otherwise, the algorithm passes to what we refer to as the predictive con-  
 144 tribution to resource predictions (S4 to S16), with candidate resources for  $T_{Ci}$   
 145 (focal taxa for explanation) identified with the KNN algorithm. For each re-  
 146 source in  $T_R$  that were not in  $N_1$  (Step S2), K most similar resources  $T_{R'}$  are  
 147 identified from  $N_1$  (Steps S4). If similar resources  $T_{R'}$  have a similarity value  
 148 above a minimal similarity threshold, they are added to  $C_R$  as candidate re-  
 149 sources. If not, they are automatically discarded (Steps S5-S7). This minimal  
 150 threshold is an arbitrary parameters (set to 0.3 in our analysis) used to avoid  
 151 predicting resources that have very small and insignificant similarity and hence  
 152 is very unlikely to share consumers and resources with the taxa it is being com-  
 153 pared to.

154 Then for all consumer taxa  $T_C$  in  $N_1$ , K most similar consumers  $T_{C'}$  are  
 155 identified from  $N_0$ . This step aims at extracting sets of potential resources  
 156  $T_R$  from similar types of consumers found in the catalogue (Step S8). Re-  
 157 sources  $T_R$  are added to candidate resources  $C_R$  for  $T_{Ci}$  if they are also found  
 158 in  $N_1$  (Steps S10-S12). Otherwise, Steps S4 to S7 are duplicated to identify  
 159 potential similar resources for  $T_{Ci}$  in  $N_1$  from the set of resources  $T_R$  of sim-  
 160 ilar consumers  $T_{C'}$  (Steps S13-S16). A simple working example is presented  
 161 at Box 1. A comprehensive mathematical description of the algorithm and  
 162 the parameters used is however available through Figure 1 and the complete R  
 163 code and data used for the algorithm is available at [https://github.com/david-](https://github.com/david-beauchesne/Predict_interactions)  
 164 [beauchesne/Predict\\_interactions](https://github.com/david-beauchesne/Predict_interactions).

### 165 3.4 Algorithm prediction accuracy

166 We used datasets including more than 50 taxa (Christian and Luczkovich, 1999;  
 167 Link, 2002; Thompson et al., 2004; Brose et al., 2005; Barnes et al., 2008;  
 168 Kortsch et al., 2015) to assess the prediction accuracy of the algorithm. Testing  
 169 accuracy of a particular dataset was done by first removing from the catalogue all  
 170 pairwise interacting taxa originating from that dataset. Accuracy was evaluated  
 171 using three different statistics:

172 1.  $Score_y$  is the fraction of interactions correctly predicted:

$$Score_y = \frac{a}{a + c} \quad (3)$$

173 2.  $Score_{-y}$  is the fraction of non-interactions correctly predicted:

$$Score_{-y} = \frac{d}{b + d} \quad (4)$$

174 3. TSS, The True Skilled Statistics (TSS) evaluated prediction success by  
 175 considering both true and false predictions, returning a value ranging from  
 176 1 (prefect predictions) to -1 (inverted predictions; Allouche et al., 2006):

$$TSS = \frac{(ad - bc)}{(a + c)(b + d)} \quad (5)$$

177 where  $a$  is the number of links predicted and observed,  $b$  is the number  
 178 predicted but not observed,  $c$  is the number of non-interaction predicted but  
 179 interactions observed and  $d$  is the number of non-interaction predicted absent  
 180 and observed. These three statistics give a different perspective on prediction  
 181 accuracy, focusing in turn on true interactions and non-interactions, and on  
 182 both true and false predictions.

183 We evaluated the three statistics for the complete algorithm and for the cat-  
 184 alogue and the predictions individually to evaluate their respective contribution  
 185 to the algorithm predictive accuracy. Multiple  $w_t$  values were also tested to  
 186 evaluate whether taxa similarity measured as a function of resource/consumer  
 187 sets or taxonomy contributed more significantly towards increased predictive  
 188 accuracy. The same was done with multiple  $K$  values.

189 Finally, we evaluated the influence of the comprehensiveness of the catalogue  
 190 on prediction accuracy. We selected the arctic marine food web from Kortsch  
 191 et al. (2015) as a test. This food web was selected as it is highly detailed  
 192 taxonomically. Furthermore, once removed from the catalogue, almost 100% of  
 193 its taxa still had information available on sets of consumers and resources, which  
 194 necessary for testing the impact of catalogue comprehensiveness on prediction  
 195 accuracy. We iteratively and randomly ( $n = 50$  randomizations) removed a  
 196 percentage of empirical data describing the food web taxa from the catalogue  
 197 before generating new predictions with the algorithm. We also tested  $w_t$  values  
 198 of 0.5 and 1 to evaluate whether taxonomic similarity could support predictive  
 199 accuracy in cases when empirical data for species in  $N_1$  in the catalogue is  
 200 unavailable.

## 201 4 Results

### 202 4.1 Biotic interaction catalogue

203 The data compilation process allowed us to build an interaction catalogue com-  
 204 posed of 276708 pairwise interactions (interactions = 72110; non-interactions =

205 204598). A total of 9712 taxa (Superfamily = 15; Family = 591; Subfamily =  
 206 29; Tribe = 8; Genus = 1972; Species = 7097) are included in the catalogue,  
 207 4159 of which have data as consumers and 4375 as resources.

## 208 4.2 Algorithm predictive accuracy

209 The overall predictive accuracy of the algorithm ranges between 80% to al-  
 210 most 100% in certain cases (Figure 2). Both interactions and non-interactions  
 211 are well predicted by the algorithm. TSS scores are lower than  $Score_y$  and  
 212  $Score_{-y}$  due to misclassified interactions and non-interactions. This can also  
 213 be observed through the effect of varying  $K$  values, which increases the number  
 214 of potential candidate resources for each taxa in the predictive portion of the  
 215 algorithm. Prediction accuracy increases for interactions, while it decreases for  
 216 non-interactions, as  $K$  values increase.

217 Similarity being predominantly measured with resource/consumer sets ( $w_t$   
 218 closer to 0) yielded better predictions than when measured with taxonomy ( $w_t$   
 219 closer to 1; Figure 2). Resource/consumer sets therefore appears to serve as  
 220 a better measure of similarity between taxa for interactions predictions. It is  
 221 nonetheless interesting to note that although the predictive contribution of the  
 222 algorithm decreases as  $w_t$  increases, an increased mean and decreased variabil-  
 223 ity values for the TSS and  $Score_y$  statistics is also observed (Figure 2). This  
 224 suggests that while using taxonomy for similarity measurements yields lower  
 225 predictive accuracy, it may also complement the catalogue contribution by pre-  
 226 dicting interactions not captured through empirical data, effectively increasing  
 227 the predictive accuracy of the complete algorithm.

228 The partitioning of the catalogue and predictive portions of the algorithm  
 229 shows that it is dependent on the comprehensiveness of the catalogue for high  
 230 prediction accuracy (Figures 2, 3). As the amount of empirical data available in  
 231 the catalogue decreases so does the overall accuracy of the algorithm (Figures 3).  
 232 The predictive contribution of the algorithm however slows down the decrease  
 233 in the prediction efficiency of the algorithm. Prediction accuracy still remains  
 234 around 75% with only 40% of  $N_1$  taxa found in the catalogue (Figures 3).  
 235 Furthermore, the use of taxonomy for similarity measurements is more efficient  
 236 as empirical data becomes scarcer and no different than resource/consumer sets  
 237 for the complete algorithm when ample data is available (Figures 3).

## 238 4.3 Southern Gulf of St. Lawrence

239 As an example, we predict interactions in the southern Gulf of St. Lawrence  
 240 (SGSL) in eastern Canada. The empirical data and taxa list come from Savenkoff  
 241 et al. (2004). They present a list of 29 functional groups for a total of 80 taxa  
 242 presented at least at taxonomical scale of the family. Other coarser functional  
 243 groups were not used for this example (see Table S1 in Supplementary informa-  
 244 tion (SI) and Savenkoff et al. (2004) for a complete description of documented  
 245 groups). As their analysis was performed on the functional groups rather than  
 246 the taxa themselves, we used the algorithm to predict interactions between all 80

247 taxa selected. We then aggregated them back to their original functional groups  
248 to compare with interactions presented in Savenkoff et al. (2004). In total, there  
249 were empirical data available in the catalogue for 78% of SGSL taxa (62/80).  
250 The algorithm correctly predicted close to 80% of interactions ( $a = 135/170$ )  
251 and non-interactions ( $d = 354/455$ ) extracted from Savenkoff et al. (2004). It  
252 also predicted an additional 101 interactions that were not noted in Savenkoff  
253 et al. (2004) and failed to predict 36 observed interactions that were, resulting in  
254 a TSS score of 0.57. A visual comparison of results obtained from the algorithm  
255 with interactions noted in Savenkoff et al. (2004) is available at Figure 4. The  
256 network presented is centered on the observed and predicted interactions of the  
257 capelin (*Mallotus villosus*) and piscivorous small pelagic feeders (e.g. *Scomber*  
258 *scombrus* and *Illex illecebrosus*).

## 259 5 Discussion

### 260 5.1 Algorithm accuracy

261 We show that out of the box interaction inference for a set of taxa with incom-  
262 plete or unavailable preexisting information can be achieved with high accuracy  
263 using a combination of empirical data describing biotic interactions and tax-  
264 onomic relatedness. Although the efficiency of the algorithm is dependent on  
265 the comprehensiveness of the interactions catalogue, taxonomic proximity acts  
266 as a complement to increase the number of observed interactions correctly pre-  
267 dicted. Taxonomic proximity also supports the efficiency of the algorithm when  
268 catalogue comprehensiveness decreases.

### 269 5.2 Usefulness of taxonomic relatedness

270 While we found that taxonomy could be useful as a complement to predictions  
271 made using empirical data, the accuracy of predictions made using the KNN  
272 algorithm could be improved. While evolutionary history plays a significant role  
273 in influencing consumer-resource trait matching and food web structure (Mou-  
274 quet et al., 2012; Rohr and Bascompte, 2014), phylogenetic constraints do not  
275 account efficiently for certain traits such as body size (Eklöf and Stouffer, 2016).  
276 Including traits like body size and metabolism as an additional component of  
277 this algorithm could thus help increasing overall prediction accuracy, especially  
278 in cases where the catalogue lacks data on taxa for which interactions have to be  
279 predicted. Although promising, such an approach would undermine the premise  
280 under which this method was built and which constitutes its main strength, *i.e.*  
281 predicting interactions in data deficient environments using readily available  
282 data.

### 283 5.3 Interactions classification

284 That  $Score_y$  and  $Score_{\neg y}$  are inversely proportional means that non-interactions  
285 are misclassified as interactions in the process of increasing  $Score_y$ , consequently



decreasing  $Score_{-y}$ . This could either stem from the algorithm poorly predicting non-interactions or from the empirical data itself. Accuracy evaluation assumes that non-interactions from empirical food web are observed data, yet it is usually not the case. Most empirical webs have a strong focus attributed to higher order consumer species and often uneven effort made to thoroughly detail species interactions (Dunne, 2006). Furthermore, the methodologies used to obtain consumer-resource data, often relying on gut content analyses, which is efficient at observing interactions, may be inefficient to detect absence of interactions in natural systems (Dunne, 2006). This is especially true with our methodology, where we predict interactions between species whose co-occurrence may have been observed in the other ecosystems we are using to predict interactions. Misclassified interactions could thus be real, albeit unobserved through empirical data available.

## 5.4 Southern Gulf of St. Lawrence

The St Lawrence example (Figure 4 and SI) provides great material to discuss predictions in greater detail. The algorithm fails to predict 20% of interactions presented in Savenkoff et al. (2004). Interactions that failed to be predicted were mainly centered on invertebrate species (e.g. polychaetes and mollusks) and large functional groups described by coarse taxonomic categories (e.g. diatoms) alongside few species in Savenkoff et al. (2004) (e.g. piscivorous small pelagic feeders; Table S3). As we focused on the taxa at least at the scale of family, it is likely that their functional groups had a broader range of possible interactions included than what the algorithm could predict using only a few taxa. Furthermore, the efficiency of the algorithm greatly depends on the underlying empirical data that defines the catalogue. If the empirical data used to build the catalogue focuses on higher order consumers, it should come as no surprise that the algorithm would be afflicted by the same limitations.

The algorithm also predicts substantially more interactions than those presented in Savenkoff et al. (2004) (Figure 4; Table S2). The catalogue is not currently built to take into account life stages of species. Considering life stages and the fact that they are not explicitly considered in the catalogue could explain additional interactions that seem suspicious at first, like the surprise amount of additional interactions predicted for small piscivorous pelagic feeders as consumers (Figure 4). Due to the aggregated nature of the SGSL web, we believe the TSS score to be an underestimate of the efficiency of the algorithm.

## 5.5 Perspectives

Overall, we believe our method performs well and offers promising avenues for further applied research and management initiatives. Interaction strength and species co-occurrence are major attributes affecting the probability of observing interactions. Interaction strength is instrumental to understanding community dynamics, stability and robustness (Laska and Wootton, 1998; Morales-Castilla et al., 2015), while the co-occurrence of species encloses valuable information

on interactions and is a pre-requisite for them to exist (Cazelles et al., 2016). Considering them in our methodology would be highly valuable to correctly assess interactions in a given ecosystem and predict the spatial distribution of interaction networks.

Given its high efficiency and simplicity, our methodology could broaden the use and the accessibility of food webs and network level descriptors for integrative management initiatives such as cumulative impacts assessments and systematic planning (Giakoumi et al., 2015; Beauchesne et al., 2016), especially for remote locations where empirical data is hard to gather. Network characteristics could be efficiently evaluated and correlated to levels of multiple environmental stressors to assess the vulnerability of ecosystems to global changes (Albouy et al., 2014). We believe that the development of such predictive approaches could represent the first much needed steps towards the use of ecological networks in systematic impacts assessments.

## 6 Acknowledgements

We thank the Fond de Recherche Québécois Nature et Technologie (FRQNT) and the Natural Science and Engineering Council of Canada (CRSNG) for financial support. This project is also supported by Québec Océan, the Quebec Centre for Biodiversity Science (QCBS), and the Notre Golfe and CHONeII networks. We also wish to thank K. Cazelles for the help, constructive comments and suggestions.

## References

- Albouy, Camille, Laure Velez, Marta Coll, Francesco Colloca, François Le Loc’h, David Mouillot, and Dominique Gravel (2014). “From projected species distribution to food-web structure under climate change”. In: *Global Change Biology* 20.3, pp. 730–741. ISSN: 13541013. DOI: [10.1111/gcb.12467](https://doi.org/10.1111/gcb.12467). URL: <http://doi.wiley.com/10.1111/gcb.12467>.
- Allouche, Omri, Asaf Tsoar, and Ronen Kadmon (2006). “Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS)”. In: *Journal of Applied Ecology* 43.6, pp. 1223–1232. ISSN: 00218901. DOI: [10.1111/j.1365-2664.2006.01214.x](https://doi.org/10.1111/j.1365-2664.2006.01214.x). URL: <http://doi.wiley.com/10.1111/j.1365-2664.2006.01214.x>.
- Bailly, N et al. (2016). *World Register of Marine Species (WoRMS)*. \url=http://www.marinespecies.org. URL: <http://www.marinespecies.org>.
- Barnes, C., D. M. Bethea, R. D. Brodeur, J. Spitz, V. Ridoux, C. Pusineri, B. C. Chase, M. E. Hunsicker, F. Juanes, A. Kellermann, J. Lancaster, F. Ménard, F.-X. Bard, P. Munk, J. K. Pinnegar, F. S. Scharf, R. A. Rountree, K. I. Stergiou, C. Sassa, A. Sabates, and S. Jennings (2008). “Predator and prey body sizes in marine food webs”. In: *Ecology* 89.3, pp. 881–881. DOI: [10.1890/07-1551.1](https://doi.org/10.1890/07-1551.1). URL: <http://doi.wiley.com/10.1890/07-1551.1>.

368 Bartomeus, Ignasi, Dominique Gravel, Jason M. Tylianakis, Marcelo A. Aizen,  
 369 Ian A. Dickie, and Maud Bernard-Verdier (2016). “A common framework for  
 370 identifying linkage rules across different types of interactions”. In: *Functional*  
 371 *Ecology*, n/a–n/a. ISSN: 02698463. DOI: [10.1111/1365-2435.12666](https://doi.org/10.1111/1365-2435.12666). URL:  
 372 <http://doi.wiley.com/10.1111/1365-2435.12666>.

373 Beauchesne, David, Cindy Grant, Dominique Gravel, and Philippe Archambault  
 374 (2016). “L’évaluation des impacts cumulés dans l’estuaire et le golfe du Saint-  
 375 Laurent : vers une planification systémique de l’exploitation des ressources”.  
 376 In: *Le Naturaliste canadien* 140.2, p. 45. ISSN: 0028-0798. DOI: [10.7202/](https://doi.org/10.7202/1036503ar)  
 377 [1036503ar](https://doi.org/10.7202/1036503ar). URL: <http://id.erudit.org/iderudit/1036503ar>.

378 Brose, Ulrich, Lara Cushing, Eric L. Berlow, Tomas Jonsson, Carolin Banasek-  
 379 Richter, Louis-Félix Bersier, Julia L. Blanchard, Thomas Brey, Stephen  
 380 R. Carpenter, Marie-France Cattin Blandenier, Joel E. Cohen, Hassan Ali  
 381 Dawah, Tony Dell, Francois Edwards, Sarah Harper-Smith, Ute Jacob, Roland  
 382 A. Knapp, Mark E. Ledger, Jane Memmott, Katja Mintenbeck, John K.  
 383 Pinnegar, Björn C. Rall, Tom Rayner, Liliane Ruess, Werner Ulrich, Philip  
 384 Warren, Rich J. Williams, Guy Woodward, Peter Yodzis, and Neo D. Mar-  
 385 tinez (2005). “Body sizes of consumers and their resources”. In: *Ecology*  
 386 86.9, pp. 2545–2545. ISSN: 0012-9658. DOI: [10.1890/05-0379](https://doi.org/10.1890/05-0379). URL: <http://doi.wiley.com/10.1890/05-0379>.

388 Brose, Ulrich, Tomas Jonsson, Eric L. Berlow, Philip Warren, Carolin Banasek-  
 389 Richter, Louis-Félix Bersier, Julia L. Blanchard, Thomas Brey, Stephen  
 390 R. Carpenter, Marie-France Cattin Blandenier, Lara Cushing, Hassan Ali  
 391 Dawah, Tony Dell, Francois Edwards, Sarah Harper-Smith, Ute Jacob, Mark  
 392 E. Ledger, Neo D. Martinez, Jane Memmott, Katja Mintenbeck, John K.  
 393 Pinnegar, Björn C. Rall, Thomas S. Rayner, Daniel C. Reuman, Liliane  
 394 Ruess, Werner Ulrich, Richard J. Williams, Guy Woodward, and Joel E.  
 395 Cohen (2006). “Consumer-resource body-size relationships in natural food  
 396 webs”. In: *Ecology* 87.10, pp. 2411–2417. DOI: [10.1890/0012-9658\(2006\)](https://doi.org/10.1890/0012-9658(2006)87[2411:CBRINF]2.0.CO;2)  
 397 [87\[2411:CBRINF\]2.0.CO;2](https://doi.org/10.1890/0012-9658(2006)87[2411:CBRINF]2.0.CO;2). URL: [http://doi.wiley.com/10.1890/0012-](http://doi.wiley.com/10.1890/0012-9658(2006)87[2411:CBRINF]2.0.CO;2)  
 398 [9658\(2006\)87\[2411:CBRINF\]2.0.CO;2](http://doi.wiley.com/10.1890/0012-9658(2006)87[2411:CBRINF]2.0.CO;2).

399 Cazelles, Kévin, Miguel B. Araújo, Nicolas Mouquet, and Dominique Gravel  
 400 (2016). “A theory for species co-occurrence in interaction networks”. In:  
 401 *Theoretical Ecology* 9.1, pp. 39–48. ISSN: 1874-1738. DOI: [10.1007/s12080-](https://doi.org/10.1007/s12080-015-0281-9)  
 402 [015-0281-9](https://doi.org/10.1007/s12080-015-0281-9). URL: [http://link.springer.com/10.1007/s12080-015-](http://link.springer.com/10.1007/s12080-015-0281-9)  
 403 [0281-9](http://link.springer.com/10.1007/s12080-015-0281-9).

404 Chamberlain, Scott A. and Eduard Szöcs (2013). “taxize: taxonomic search  
 405 and retrieval in R”. In: *F1000Research* 2. ISSN: 2046-1402. DOI: [10.12688/](https://doi.org/10.12688/f1000research.2-191.v1)  
 406 [f1000research.2-191.v1](https://doi.org/10.12688/f1000research.2-191.v1). URL: [http://f1000research.com/articles/](http://f1000research.com/articles/2-191/v1)  
 407 [2-191/v1](http://f1000research.com/articles/2-191/v1).

408 Chamberlain, Scott A., Eduard Szöcs, Carl Boettiger, Karthik Ram, Ignasi Bar-  
 409 tomeus, and John Baumgartner (2014). *taxize: Taxonomic information from*  
 410 *around the web*. URL: <https://github.com/ropensci/taxize>.

411 Christian, Robert R. and Joseph J. Luczkovich (1999). “Organizing and un-  
 412 derstanding a winter’s seagrass foodweb network through effective trophic

levels". In: *Ecological Modelling* 117.1, pp. 99–124. ISSN: 03043800. DOI: [10.1016/S0304-3800\(99\)00022-8](https://doi.org/10.1016/S0304-3800(99)00022-8).

Cohen, Joel E, Tomas Jonsson, and Stephen R Carpenter (2003). "Ecological community description using the food web, species abundance, and body size." In: *Proceedings of the National Academy of Sciences of the United States of America* 100.4, pp. 1781–6. ISSN: 0027-8424. DOI: [10.1073/pnas.232715699](https://doi.org/10.1073/pnas.232715699). URL: <http://www.ncbi.nlm.nih.gov/pubmed/12547915> <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC149910>.

Desjardins-Proulx, Philippe, Timothée Poisot, and Dominique Gravel (2016). "Ecological interactions and the Netflix problem". In:

Dunne, JA (2006). "The network structure of food webs". In: *networks: linking structure to dynamics in food webs*. URL: <https://books.google.ca/books?hl=en&%7B%5C%7Dlr=%7B%5C%7Ddid=bF3JoZgoo24C%7B%5C%7Ddoi=fnd%7B%5C%7Dpg=PA27%7B%5C%7Ddq=The+Network+Structure+of+Food+Webs%7B%5C%7Dots=00mi%7B%5C%7DKSWEi%7B%5C%7Dsig=kVzdEtE5toSzP0Kq7bXnBwHghaY>.

Eklöf, Anna, Matthew R. Helmus, M. Moore, and Stefano Allesina (2012). "Relevance of evolutionary history for food web structure". In: *Proceedings of the Royal Society of London B: Biological Sciences* 279.1733.

Eklöf, Anna and Daniel B. Stouffer (2016). "The phylogenetic component of food web structure and intervality". In: *Theoretical Ecology* 9.1, pp. 107–115. ISSN: 1874-1738. DOI: [10.1007/s12080-015-0273-9](https://doi.org/10.1007/s12080-015-0273-9). URL: <http://link.springer.com/10.1007/s12080-015-0273-9>.

Giakoumi, Sylvaine, Benjamin S. Halpern, Loïc N. Michel, Sylvie Gobert, Maria Sini, Charles-François Boudouresque, Maria-Cristina Gambi, Stelios Katsanevakis, Pierre Lejeune, Monica Montefalcone, Gerard Pergent, Christine Pergent-Martini, Pablo Sanchez-Jerez, Branko Velimirov, Salvatrice Vizzini, Arnaud Abadie, Marta Coll, Paolo Guidetti, Fiorenza Micheli, and Hugh P. Possingham (2015). "Towards a framework for assessment and management of cumulative human impacts on marine food webs". In: *Conservation Biology* 29.4, pp. 1228–1234. ISSN: 08888892. DOI: [10.1111/cobi.12468](https://doi.org/10.1111/cobi.12468). URL: <http://doi.wiley.com/10.1111/cobi.12468>.

Gravel, Dominique, Timothée Poisot, Camille Albouy, Laure Velez, and David Mouillot (2013). "Inferring food web structure from predator-prey body size relationships". In: *Methods in Ecology and Evolution* 4.11. Ed. by Robert Freckleton, pp. 1083–1090. ISSN: 2041210X. DOI: [10.1111/2041-210X.12103](https://doi.org/10.1111/2041-210X.12103). URL: <http://doi.wiley.com/10.1111/2041-210X.12103>.

Gray, Clare, David H. Figueroa, Lawrence N. Hudson, Athen Ma, Dan Perkins, and Guy Woodward (2015). "Joining the dots: An automated method for constructing food webs from compendia of published interactions". In: *Food Webs* 5, pp. 11–20. ISSN: 23522496. DOI: [10.1016/j.fooweb.2015.09.001](https://doi.org/10.1016/j.fooweb.2015.09.001).

Ings, Thomas C., José M. Montoya, Jordi Bascompte, Nico Blüthgen, Lee Brown, Carsten F. Dormann, François Edwards, David Figueroa, Ute Jacob, J. Iwan Jones, Rasmus B. Lauridsen, Mark E. Ledger, Hannah M. Lewis, Jens M. Olesen, F.J. Frank van Veen, Phil H. Warren, and Guy Woodward (2009).

459 “Review: Ecological networks - beyond food webs”. In: *Journal of Ani-*  
460 *mal Ecology* 78.1, pp. 253–269. ISSN: 00218790. DOI: [10.1111/j.1365-](https://doi.org/10.1111/j.1365-2656.2008.01460.x)  
461 [2656.2008.01460.x](https://doi.org/10.1111/j.1365-2656.2008.01460.x). URL: [http://doi.wiley.com/10.1111/j.1365-](http://doi.wiley.com/10.1111/j.1365-2656.2008.01460.x)  
462 [2656.2008.01460.x](http://doi.wiley.com/10.1111/j.1365-2656.2008.01460.x).

463 Kortsch, Susanne, Raul Primicerio, Maria Fossheim, Andrey V. Dolgov, and  
464 Michaela Aschan (2015). “Climate change alters the structure of arctic ma-  
465 rine food webs due to poleward shifts of boreal generalists”. In: *Proceedings*  
466 *of the Royal Society of London B: Biological Sciences* 282.1814.

467 Laska, Mark S. and J. Timothy Wootton (1998). “Theoretical concepts and  
468 empirical approaches to measuring interaction strength”. In: *Ecology* 79.2,  
469 pp. 461–476. DOI: [10.1890/0012-9658\(1998\)079\[0461:TCAEAT\]2.0.CO;2](https://doi.org/10.1890/0012-9658(1998)079[0461:TCAEAT]2.0.CO;2);  
470 URL: [http://doi.wiley.com/10.1890/0012-9658\(1998\)079\[0461:](http://doi.wiley.com/10.1890/0012-9658(1998)079[0461:TCAEAT]2.0.CO;2)  
471 [TCAEAT\]2.0.CO;2](http://doi.wiley.com/10.1890/0012-9658(1998)079[0461:TCAEAT]2.0.CO;2).

472 Link, J (2002). “Does food web theory work for marine ecosystems?” In: *Ma-*  
473 *rine Ecology Progress Series* 230, pp. 1–9. ISSN: 0171-8630. DOI: [10.3354/](https://doi.org/10.3354/meps230001)  
474 [meps230001](https://doi.org/10.3354/meps230001). URL: [http://www.int-res.com/abstracts/meps/v230/p1-](http://www.int-res.com/abstracts/meps/v230/p1-9/)  
475 [9/](http://www.int-res.com/abstracts/meps/v230/p1-9/).

476 Martinez, Neo D. (1992). “Constant connectance in community food webs”. In:  
477 *American Naturalist* 139.6, pp. 1208–1218. URL: [http://www.jstor.org/](http://www.jstor.org/stable/2462337)  
478 [stable/2462337](http://www.jstor.org/stable/2462337).

479 Morales-Castilla, Ignacio, Miguel G. Matias, Dominique Gravel, and Miguel B.  
480 Araújo (2015). “Inferring biotic interactions from proxies”. In: *Trends in*  
481 *Ecology & Evolution* 30.6, pp. 347–356. ISSN: 01695347. DOI: [10.1016/j.](https://doi.org/10.1016/j.tree.2015.03.014)  
482 [tree.2015.03.014](https://doi.org/10.1016/j.tree.2015.03.014).

483 Mouquet, Nicolas, Vincent Devictor, Christine N. Meynard, Francois Munoz,  
484 Louis-Félix Bersier, Jérôme Chave, Pierre Couteron, Ambroise Dalecky, Colin  
485 Fontaine, Dominique Gravel, Olivier J. Hardy, Franck Jabot, Sébastien Lavergne,  
486 Mathew Leibold, David Mouillot, Tamara Münkemüller, Sandrine Pavoine,  
487 Andreas Prinzing, Ana S.L. Rodrigues, Rudolf P. Rohr, Elisa Thébaud, and  
488 Wilfried Thuiller (2012). “Ecophylogenetics: advances and perspectives”. In:  
489 *Biological Reviews* 87.4, pp. 769–785. ISSN: 14647931. DOI: [10.1111/j.1469-](https://doi.org/10.1111/j.1469-185X.2012.00224.x)  
490 [185X.2012.00224.x](https://doi.org/10.1111/j.1469-185X.2012.00224.x). URL: [http://doi.wiley.com/10.1111/j.1469-](http://doi.wiley.com/10.1111/j.1469-185X.2012.00224.x)  
491 [185X.2012.00224.x](http://doi.wiley.com/10.1111/j.1469-185X.2012.00224.x).

492 Murphy, Kevin P. (2012). *Machine learning : a probabilistic perspective*. MIT  
493 Press, p. 1067. ISBN: 9780262018029.

494 Pascual, M and JA Dunne (2006). *Ecological networks: linking structure to dy-*  
495 *namics in food webs*. URL: [https://books.google.ca/books?hl=en%](https://books.google.ca/books?hl=en%7B%5C%7Dlr=%7B%5C%7Ddid=YpQRDAAQBAJ%7B%5C%7Ddoi=fnd%7B%5C%7Dpg=PP1%7B%5C%7Ddq=Pascual+and+Dunne+2006+interactions%7B%5C%7Dots=K4a5d62r9X%7B%5C%7Dsig=01fs%7B%5C%7DfXV1pgP6IeP1jBIb3B61rU)  
496 [7B%5C%7Dlr=%7B%5C%7Ddid=YpQRDAAQBAJ%7B%5C%7Ddoi=fnd%7B%5C%](https://books.google.ca/books?hl=en%7B%5C%7Dlr=%7B%5C%7Ddid=YpQRDAAQBAJ%7B%5C%7Ddoi=fnd%7B%5C%7Dpg=PP1%7B%5C%7Ddq=Pascual+and+Dunne+2006+interactions%7B%5C%7Dots=K4a5d62r9X%7B%5C%7Dsig=01fs%7B%5C%7DfXV1pgP6IeP1jBIb3B61rU)  
497 [7B%5C%7Dpg=PP1%7B%5C%7Ddq=Pascual+and+Dunne+2006+interactions%7B%](https://books.google.ca/books?hl=en%7B%5C%7Dlr=%7B%5C%7Ddid=YpQRDAAQBAJ%7B%5C%7Ddoi=fnd%7B%5C%7Dpg=PP1%7B%5C%7Ddq=Pascual+and+Dunne+2006+interactions%7B%5C%7Dots=K4a5d62r9X%7B%5C%7Dsig=01fs%7B%5C%7DfXV1pgP6IeP1jBIb3B61rU)  
498 [5C%7Dots=K4a5d62r9X%7B%5C%7Dsig=01fs%7B%5C\\_%7DfXV1pgP6IeP1jBIb3B61rU](https://books.google.ca/books?hl=en%7B%5C%7Dlr=%7B%5C%7Ddid=YpQRDAAQBAJ%7B%5C%7Ddoi=fnd%7B%5C%7Dpg=PP1%7B%5C%7Ddq=Pascual+and+Dunne+2006+interactions%7B%5C%7Dots=K4a5d62r9X%7B%5C%7Dsig=01fs%7B%5C%7DfXV1pgP6IeP1jBIb3B61rU).

499 Poelen, Jorrit H., Stephen Gosnell, and Sergey Slyusarev (2015). *rglobi: R In-*  
500 *terface to Global Biotic Interactions*. URL: [https://cran.r-project.org/](https://cran.r-project.org/package=rglobi)  
501 [package=rglobi](https://cran.r-project.org/package=rglobi).

502 Poelen, Jorrit H., James D. Simons, and Chris J. Mungall (2014). “Global biotic  
503 interactions: An open infrastructure to share and analyze species-interaction

504 datasets". In: *Ecological Informatics* 24, pp. 148–159. ISSN: 15749541. DOI:  
505 [10.1016/j.ecoinf.2014.08.005](https://doi.org/10.1016/j.ecoinf.2014.08.005).

506 Polis, GA (1991). "Complex trophic interactions in deserts: an empirical critique  
507 of food-web theory". In: *American naturalist* 138.1, pp. 123–155. URL: [http:  
508 //www.jstor.org/stable/2462536](http://www.jstor.org/stable/2462536).

509 Rohr, Rudolf P. and Jordi Bascompte (2014). "Components of Phylogenetic Sig-  
510 nal in Antagonistic and Mutualistic Networks". In: *The American Naturalist*  
511 184.5, pp. 556–564. DOI: [10.1086/678234](https://doi.org/10.1086/678234). URL: [http://www.journals.  
512 uchicago.edu/doi/10.1086/678234](http://www.journals.uchicago.edu/doi/10.1086/678234).

513 Savenkoff, Claude, Hugo Bourdages, Douglas P. Swain, Simon-Pierre Despatie,  
514 J. Mark Hanson, Red Méthot, Lyne Morissette, and Mike O. Hammil (2004).  
515 *Input data and parameter estimates for ecosystem models of the southern*  
516 *Gulf of St. Lawrence (mid-1980s and mid-1990s)*. Tech. rep. Mont-Joli, Québec,  
517 Canada: Canadian Technical Report of Fisheries, Aquatic Sciences 2529, De-  
518 partment of Fisheries, and Oceans, p. 105.

519 Schrodtt, Franziska, Jens Kattge, Hanhuai Shan, Farideh Fazayeli, Julia Joswig,  
520 Arindam Banerjee, Markus Reichstein, Gerhard Bönisch, Sandra Díaz, John  
521 Dickie, Andy Gillison, Anuj Karpatne, Sandra Lavorel, Paul Leadley, Chris-  
522 tian B. Wirth, Ian J. Wright, S. Joseph Wright, and Peter B. Reich (2015).  
523 "BHPMF - a hierarchical Bayesian approach to gap-filling and trait predic-  
524 tion for macroecology and functional biogeography". In: *Global Ecology and*  
525 *Biogeography* 24.12, pp. 1510–1521. ISSN: 1466822X. DOI: [10.1111/geb.  
526 12335](https://doi.org/10.1111/geb.12335). URL: <http://doi.wiley.com/10.1111/geb.12335>.

527 Thompson, Ross M., Kim N. Mouritsen, and Robert Poulin (2004). "Importance  
528 of parasites and their life cycle characteristics in determining the structure  
529 of a large marine food web". In: *Journal of Animal Ecology* 74.1, pp. 77–85.  
530 DOI: [10.1111/j.1365-2656.2004.00899.x](https://doi.org/10.1111/j.1365-2656.2004.00899.x). URL: [http://doi.wiley.com/  
531 10.1111/j.1365-2656.2004.00899.x](http://doi.wiley.com/10.1111/j.1365-2656.2004.00899.x).

532 Tylianakis, Jason M., Raphael K. Didham, Jordi Bascompte, and David A.  
533 Wardle (2008). "Global change and species interactions in terrestrial ecosys-  
534 tems". In: *Ecology Letters* 11.12, pp. 1351–1363. ISSN: 1461023X. DOI: [10.  
535 1111/j.1461-0248.2008.01250.x](https://doi.org/10.1111/j.1461-0248.2008.01250.x). URL: [http://doi.wiley.com/10.  
536 1111/j.1461-0248.2008.01250.x](http://doi.wiley.com/10.1111/j.1461-0248.2008.01250.x).

537 University of Canberra (2016). *Food Web Database - University of CANBERRA*.  
538 URL: <http://globalwebdb.com/>.

539 **6.1 Box 1**

540 The algorithm follows a series of logical steps to predict resources for all taxa  
541 in an arbitrary set of taxa  $N_1$  using a set of taxa  $N_0$  with empirically described  
542 interactions from which we can extract sets of consumers and resources and their  
543 taxonomy. In this example, we are predicting interactions for a fictitious  $N_1 =$   
544  $\{T_1, T_9, T_{10}, T_{11}, T_{12}\}$  using  $N_0$  with information on 12 taxa. This catalogue  
545 holds information on consumer or resource for 10 taxa and the taxonomy for all  
546 12 taxa in the list.

$N_0$ taxa ID	taxonomy	resource	consumer
$T_1$	$\{a, b, c\}$	$\{T_2, T_3, T_{12}\}$	$\{T_4\}$
$T_2$	$\{e, f, g\}$		$\{T_1, T_5\}$
$T_3$	$\{i, j, k\}$		$\{T_5\}$
$T_4$	$\{m, n, o\}$	$\{T_1, T_5\}$	
$T_5$	$\{a, b, d\}$	$\{T_8, T_9\}$	$\{T_4\}$
$T_6$	$\{i, q, r\}$	$\{T_2, T_8\}$	$\{T_4\}$
$T_7$	$\{e, f, h\}$		$\{T_1, T_6\}$
$T_8$	$\{s, t, u\}$		$\{T_5, T_6\}$
$T_9$	$\{s, t, v\}$		$\{T_5\}$
$T_{10}$	$\{i, j, l\}$		
$T_{11}$	$\{m, n, p\}$		
$T_{12}$	$\{q, r, s\}$		$\{T_1\}$

547 Similarity between all pairs of taxa in  $N_0$  is measured for consumer, resource  
548 and taxonomic proximity using equation 1. The upper triangular matrix repre-  
549 sents similarity measured with taxa sets of resources/consumers, while the lower  
550 triangular represents taxonomic similarities. For consumer/resource set similar-  
551 ities, values of 0 mean that similarity equals 0 for both similarity measurements.

552

$$\text{tanimoto}(T_Cx, T_Cy) / \text{tanimoto}(T_Rx, T_Ry)$$

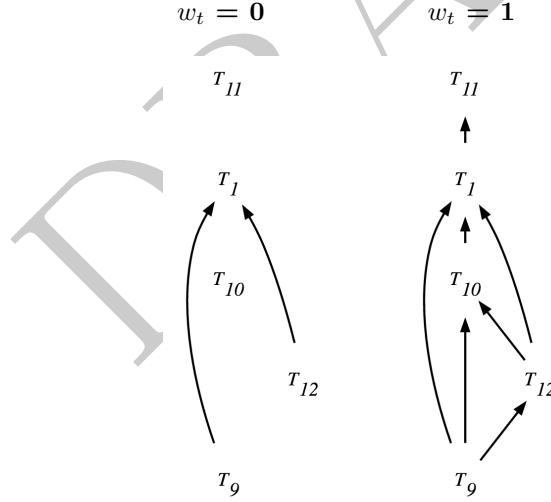
	$T_1$	$T_2$	$T_3$	$T_4$	$T_5$	$T_6$	$T_7$	$T_8$	$T_9$	$T_{10}$	$T_{11}$	$T_{12}$
$T_1$	-	0	0	0	0/1	0.3/1	0	0	0	0	0	0
$T_2$	0	-	0/0.5	0	0	0	0/0.3	0/0.3	0/0.5	0	0	0/0.5
$T_3$	0	0	-	0	0	0	0	0/0.5	0/1	0	0	0
$T_4$	0	0	0	-	0	0	0	0	0	0	0	0
$T_5$	0.5	0	0	0	-	0.3/1	0	0	0	0	0	0
$T_6$	0	0	0.2	0	0	-	0	0	0	0	0	0
$T_7$	0	0.5	0	0	0	0	-	0/0.3	0	0	0	0/0.5
$T_8$	0	0	0	0	0	0	0	-	0	0	0	0
$T_9$	0	0	0	0	0	0	0	0.5	-	0	0	0
$T_{10}$	0	0	0.5	0	0	0.2	0	0	0	-	0	0
$T_{11}$	0	0	0	0.5	0	0	0	0	0	0	-	0
$T_{12}$	0	0	0	0	0	0.5	0	0.2	0.2	0	0	-

$$\text{tanimoto}(T_Tx, T_Ty)$$

553 From these, the algorithm goes through logical steps (Figure 1) to identify  
554 a candidate resource list  $C_R$  for each taxon in  $N_1$  using either empirical data  
555 directly or  $K$  most similar taxa with equation 2. Going through the process for  
556  $T_1$ , using  $K = 1$  and  $w_t = 1$ :

Steps		Catalogue	Prediction
1	$I(T_1, T_R)$ in $N_0$ ?		
2	$T_R$ in $N_1$ ?		
4-7	$T_2 = \text{no} \rightarrow t(T_2, T_{R'}, w_t) = \text{NA}$	$\{\}$	$\{\}$
4-7	$T_3 = \text{no} \rightarrow t(T_2, T_{R'}, w_t) = T_{10} = 0.5$	$\{\}$	$\{T_{10}\}$
3	$T_{12} = \text{yes}$	$\{T_{12}\}$	$\{T_{10}\}$
8	$t(T_1, T_{C'}, w_t) = T_5 = 0.5$		
9	$I(T_5, T_R)$ in $N_1$ ?		
13-16	$T_8 = \text{no} \rightarrow t(T_8, T_{R'}, w_t) = T_9 = 0.5$	$\{T_{12}\}$	$\{T_9, T_{10}\}$
10-12	$T_9 = \text{yes}$	$\{T_9, T_{12}\}$	$\{T_9, T_{10}\}$

557 The logical steps allow us to predict a set of resources for  $T_1 = \{T_9, T_{10},$   
558  $T_{12}\}$ . Doing it for all taxa in  $N_1$  with  $w_t = 0$  and 1 predicts the following  
559 networks:





## 6.2 Figures

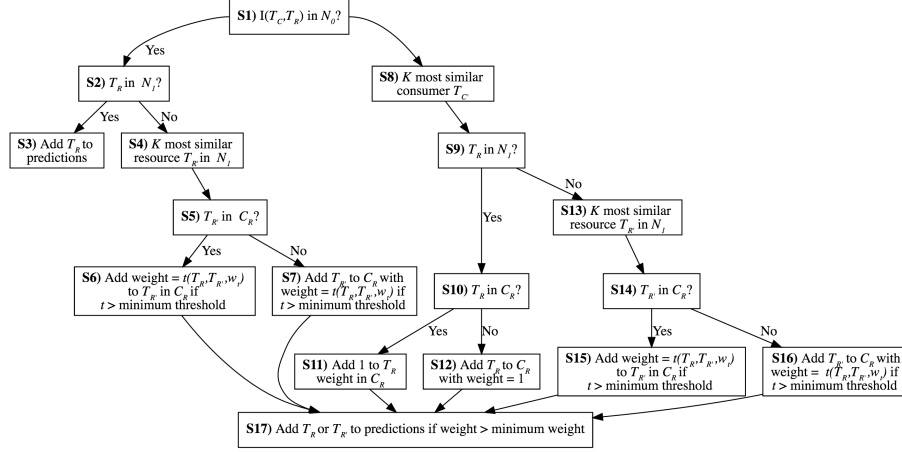


Figure 1: Description of logical steps used by the algorithm to suggest a list of candidate resources ( $C_R$ ) for each consumer tax ( $T_C$ ) in an arbitrary set of  $N_1$  for which interactions are predicted, using a set of taxa  $N_0$  with empirically described interactions. Interactions between consumer and resource taxa are denoted as  $I(T_C, T_R)$ .  $K$  is the number of most similar neighbours selected for the KNN algorithm;  $t$  stands for tanimoto in equation 1;  $w_t$  is the weight given to sets of resources and consumers in equation 2; the minimum threshold is an arbitrary value setting the minimal similarity value accepted for taxa to be considered as close neighbours in the KNN algorithm; the weight is the value added to a candidate resource each time it is added to  $C_R$ ; the minimum weight is the minimal weight value accepted for candidate resources to be selected as predicted sources in the algorithm.

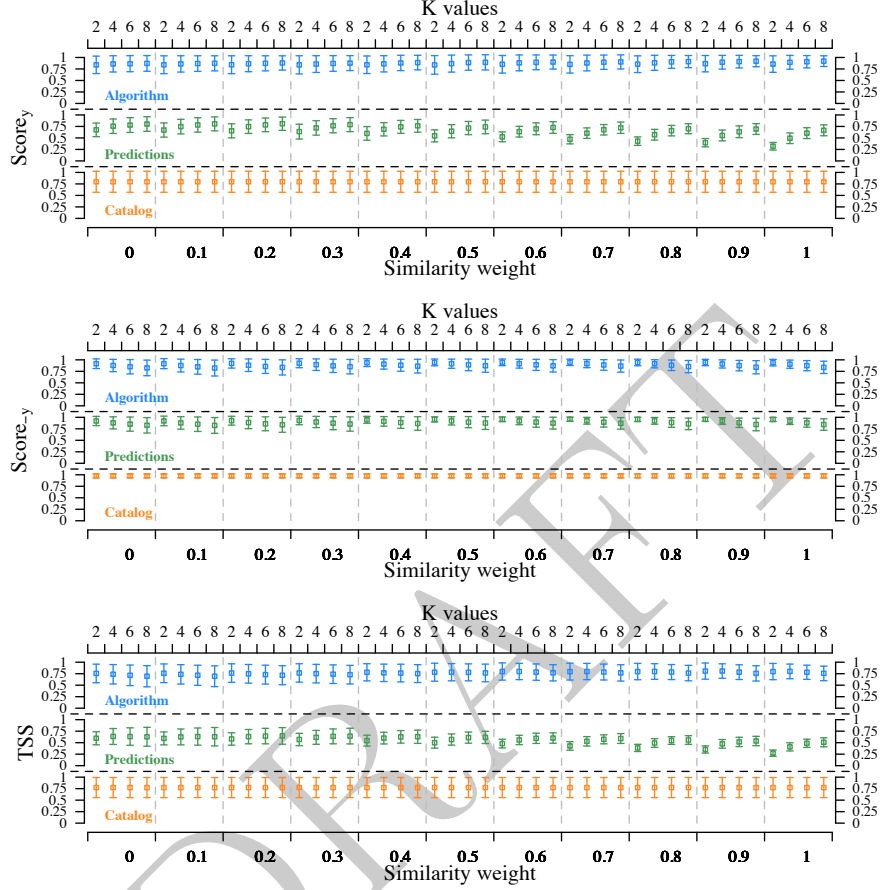


Figure 2: Representation of the three statistics (*i.e.*  $Score_y$ ,  $Score_{-y}$  and TSS) used to evaluate the accuracy of the algorithm as a function of  $K$  values tested (*i.e.* 2, 4, 6 and 8 most similar neighbours, top  $x$ -axis) and weight for taxonomy (bottom  $x$ -axis), which varies between 0 and 1. A weight of 0 means that similarity is measured only using set of resources/consumers for each taxa, while a weight of 1 means that similarity is based solely on taxonomy. For each statistic, the topmost panel presents prediction accuracy for the complete algorithm, the middle panel corresponds to predictions made through the predictive portion of the algorithm (Steps S4-S16; Figure 1) and the bottom panel presents the catalogue contribution for the algorithm (Steps S1-S3; Figure 1). Note that the sum of the predictive and catalogue contributions can be over 100% as there is overlap between predictions made through both. The 7 datasets used for this analysis contained over 50 taxa (Thompson2005; Christian and Luczkovich, 1999; Link, 2002; Brose et al., 2005; Barnes et al., 2008; Kortsch et al., 2015)

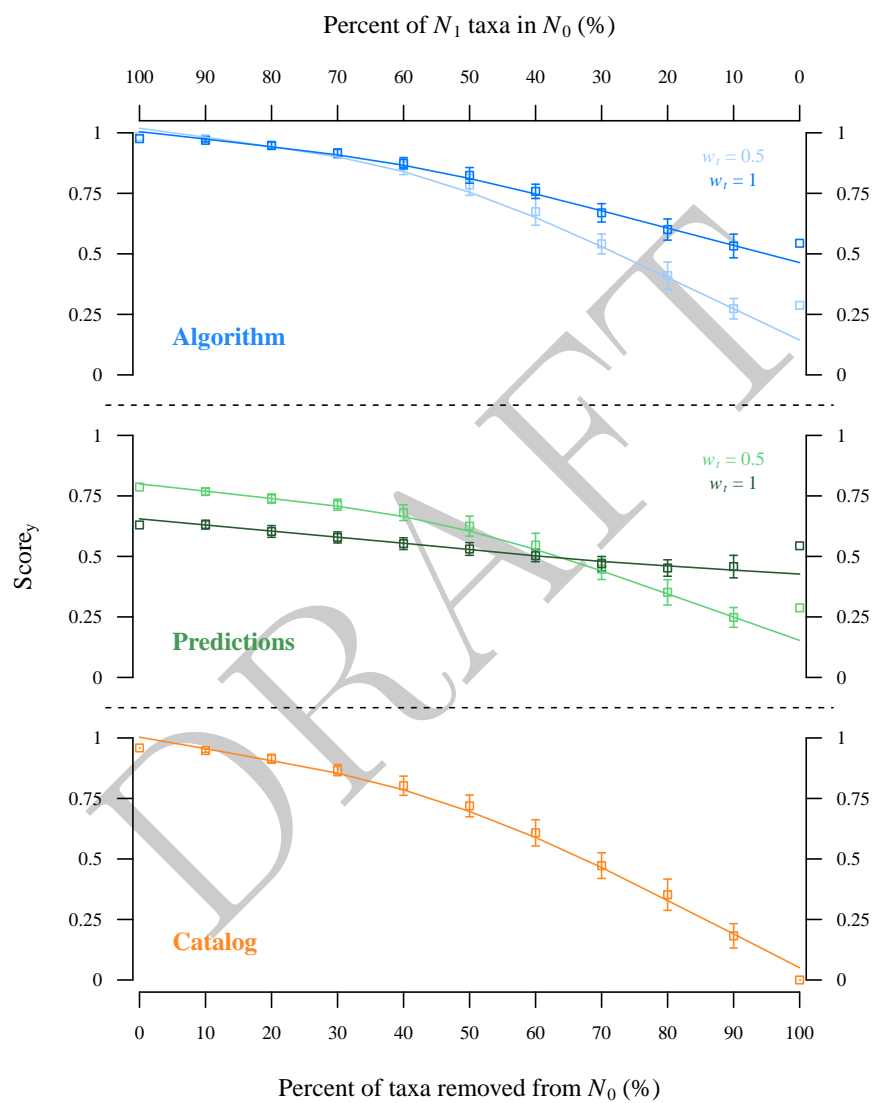


Figure 3: Caption on next page.

Figure 3: Representation of  $Score_y$  as a function of catalogue comprehensiveness, *i.e.* the amount of information on sets of consumer and resources available in the catalogue. The sensitivity of the algorithm to data accuracy was evaluated with the arctic food web from Kortsch et al. (2015). This food web was highly detailed taxonomically. Once removed from the catalogue, almost 100% of its taxa still had information available on sets of consumers and resources, which necessary for testing the impact of catalogue comprehensiveness on prediction accuracy. A random percentage of data available in the catalogue for taxa in the food web (*i.e.* 0 to 100%) was iteratively removed ( $n = 50$  randomizations) before generating new predictions with the algorithm.  $w_t$  values of 0.5 and 1 were evaluated to verify the usefulness of taxonomy in supporting predictive accuracy. The topmost panel presents prediction accuracy for the complete algorithm, the middle panel corresponds to predictions made through the predictive portion of the algorithm (Steps S4-S16; Figure 1) and the bottom panel presents the catalogue contribution for the algorithm (Steps S1-S3; Figure 1). Note that the sum of the predictive and catalogue contributions can be over 100% as there is overlap between predictions made through both.

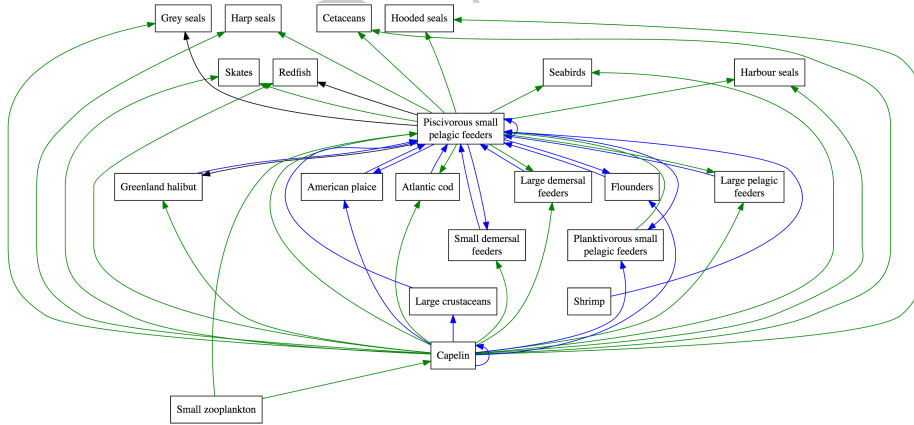


Figure 4: Example of predicted interactions with the network of the southern Gulf of St. Lawrence (Savenkoff et al., 2004), centered around the interactions of the capelin (*Mallotus villosus*) and piscivorous small pelagic feeders (*e.g.* *Scomber scombrus* and *Illex illecebrosus*). Edge with colors green were both predicted and observed (26), black were observed only (3) and blue were predicted only (19). Arrows are pointed towards consumers.