

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

4 *DAVID BEAUCHESNE*^{1*}, *PHILIPPE DESJARDINS-PROULX*²,
5 *PHILIPPE ARCHAMBAULT*³, and *DOMINIQUE GRAVEL*²

6 * email: david.beauchesne@uqar.ca

7 ¹ *Université du Québec à Rimouski*

8 ² *Université de Sherbrooke*

9 ³ *Université Laval*

10 September 19, 2016

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. Initial results suggest that pairwise interactions can be predicted with high accuracy. Although results 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 it's 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.

2 Introduction

Large networks of ecological interactions, such as food webs, are complex to characterize (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 evolving approach is to predict interactions using proxies such as functional traits, phylogenies and spatial distributions (e.g. Gravel et al., 2013; 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 to gather the necessary data to apply those methods may still be restrictive, or the data be unavailable altogether, so much so that other methods have been developed to fill the gaps in knowledge (e.g. Schrodtt et al., 2015).

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

3 Methods

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

3.1 Biotic interaction catalogue

We built a biotic interaction catalogue to serve as a set of taxa N_0 for training the algorithm with empirically described interactions. The empirical data used to construct the interaction catalogue was gathered in two successive steps. The first consisted of gathering data from a collection of 94 empirical food webs in marine and coastal ecosystems from which we extracted pairwise taxa interactions (see Brose et al., 2005; Kortsch et al., 2015; GlobalWeb database for more information). We also used a detailed predator-prey interaction database describing trophic relationships between XX predators and their prey (Barnes et al., 2008). From these datasets, only interactions between taxa at the taxonomic scale of the family or higher were selected for inclusion in the catalogue.

As empirical food webs are vastly dominated by non-interactions, 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

94 taxa worldwide. We extracted all interactions available on GloBI for species
 95 belonging to the families of taxa identified through step 1. Interactions were
 96 extracted using the rGloBI package in R (Poelen et al., 2015). As per step 1,
 97 only interactions between taxa at the taxonomic scale of the family or higher
 98 were retained

99 The nomenclature used between datasets and food webs varied substantially.
 100 Taxa names thus had to be verified, modified according to the scientific nomen-
 101 clature and validated. This process was performed using the Taxize package in
 102 R (Chamberlain and Szöcs, 2013; Chamberlain et al., 2014) and manually veri-
 103 fied for errors. The same package was used to extract the taxonomy of all taxa
 104 for which interactions were obtained in previous steps. The complete R code
 105 and data used to build the catalogue is available at https://github.com/david-beauchesne/Interaction_catalog.
 106

107 3.2 Unsupervised machine learning

108 We use the K -nearest neighbor (KNN) algorithm (ref) to predict pairwise in-
 109 teractions for a set of taxa S . The KNN algorithm predicts missing entries
 110 or proposes additional entries by a majority vote based on the K nearest (i.e.
 111 most similar) entries (see Box 1 for an example). In this case, taxa are described
 112 by a set of resources when considered as a consumer, a set of consumers when
 113 considered as a resource and their taxonomy (i.e. kingdom, phylum, class, or-
 114 der, family, genus, species). Similarity between taxa was evaluated using the
 115 Tanimoto similarity measure (ref), which compares two vectors with i elements
 116 based on the number of elements they share and contain:

$$\text{tanimoto}(\mathbf{x}, \mathbf{y}) = \frac{\sum_i x_i \wedge y_i}{\sum_i x_i \vee y_i}, \quad (1)$$

117 where \wedge is bitwise *and*, while \vee is the bitwise *or* operators. Adding a weight-
 118 ing scheme, we can measure the similarity using two different sets of vectors with
 119 i and j elements, respectively.

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

120 where w_t is the weight given to vector i , \mathbf{x}_i , \mathbf{y}_i are the resource or consumer
 121 sets of the two taxa and \mathbf{x}_j and \mathbf{y}_j are the vectors for the taxonomy of two taxa.
 122 When $w_t = 0$ only resource or consumer sets are used to compute similarity,
 123 while $w_t = 1$ solely uses taxonomy.

124 3.3 Predicting interactions, Biotic predictor algorithm, Two- 125 way Tanimoto algorithm, Feng shui name algorithm, 126 Find a name for the algorithm

127 The XXX algorithm is built on a series of logical steps that ultimately predicts
 128 a candidate resources list C_R for each taxon in N_1 (Figure 1). For all consumer

129 taxa T_C in N_1 , the algorithm first verifies whether it has empirical resources
 130 T_R listed in the catalogue (Step S1, Figure 1). When it does, if T_R are also in
 131 N_1 , they are added as predicted resources for T_C (S2, S3). This corresponds
 132 to what we refer to as the catalogue contribution to resource predictions. Two
 133 taxa in N_1 that are known to interact through the catalogue are automatically
 134 assumed to interact in N_1 .

135 Otherwise, the algorithm passes to what we refer to as the predictive con-
 136 tribution to resource predictions (S4 to S16), with candidate resources for T_C
 137 identified with the KNN algorithm. If T_R are absent from N_1 , K most similar
 138 resources $T_{R'}$ are identified in N_1 to add to C_R (S4 to S7). Then for all T_C in N_1 ,
 139 the algorithm identifies K most similar consumers $T_{C'}$ in N_0 and extracts their
 140 resource sets (S8). As before, if those resources are found in N_1 (S9) they are
 141 added to C_R (S10 to S12), otherwise K most similar resources $T_{R'}$ are identified
 142 in N_1 (S13) to add to C_R (S14 to S16). A simple working example is presented at
 143 Box 1. Note that other parameters are used in the algorithm, but not presented
 144 here for the sake of message clarity. A more comprehensive mathematical de-
 145 scription of the algorithm and the parameters used is however available through
 146 Figure 1 and the complete R code and data used for the algorithm is available
 147 at https://github.com/david-beauchesne/Predict_interactions.

148 3.4 Algorithm prediction accuracy

149 We used the most extensive and taxonomically detailed datasets included in
 150 the catalogue (**ref**) to assess the prediction accuracy of the algorithm. Testing
 151 accuracy of a particular dataset was done by first removing from the catalogue all
 152 pairwise interacting taxa originating from that dataset. Accuracy was evaluated
 153 using three different statistics:

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

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

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

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

- 156 3. TSS, The True Skilled Statistics (TSS) evaluated prediction success by
 157 considering both true and false predictions, returning a value ranging from
 158 1 (perfect predictions) to -1 (inverted predictions; Allouche et al., 2006):

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

159 where a is the number of links predicted and observed, b is the number
 160 predicted but not observed, c is the number of non-interaction predicted but

interactions observed and d is the number of non-interaction predicted absent and observed. These three statistics give a different perspective on prediction accuracy, focusing in turn on true interactions and non-interactions, and on both true and false predictions.

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

Finally, we evaluated the influence of the comprehensiveness of the catalogue on prediction accuracy. We selected the arctic food web from Kortsch et al. (2015) as a test. This food web was selected as it is highly detailed taxonomically. Furthermore, 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. We iteratively and randomly ($n = 50$ randomizations) removed a percentage of empirical data describing the food web taxa from the catalogue before generating new predictions with the algorithm. We also tested w_t values of 0.5 and 1 to evaluate whether taxonomic similarity could support predictive accuracy in cases when empirical data for species in N_1 in the catalogue is unavailable.

4 Results

4.1 Biotic interaction catalogue

The data compilation process allowed us to build an interaction catalogue composed of 276708 pairwise interactions (interactions = 72110; non-interactions = 204598). A total of 9712 taxa (Superfamily = 15; Family = 591; Subfamily = 29; Tribe = 8; Genus = 1972; Species = 7097) are included in the catalogue, 4159 of which have data as consumers and 4375 as resources.

4.2 Algorithm predictive accuracy

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

Similarity being predominantly measured with resource/consumer sets (w_t closer to 0) yielded better predictions than when measured with taxonomy (w_t

201 closer to 1; Figure 2). Resource/consumer sets therefore appears to serve as
 202 a better predictor of similarity between taxa for interactions predictions. It is
 203 nonetheless interesting to note that although the predictive contribution of the
 204 algorithm decreases as w_t increases, an increased mean and decreased variability
 205 values for the TSS and $Score_y$ statistics is also observed (Figure 2). This
 206 suggests that while using taxonomy for similarity measurements yields lower
 207 predictive accuracy, it may also complement the catalogue contribution by pre-
 208 dicting interactions not captured through empirical data, effectively increasing
 209 the predictive accuracy of the complete algorithm.

210 The partitioning of the catalogue and predictive portions of the algorithm
 211 shows that it is dependent on the comprehensiveness of the catalogue for high
 212 prediction accuracy (Figures 2, 3). As the amount of empirical data available in
 213 the catalogue decreases so does the overall accuracy of the algorithm (Figures 3).
 214 The predictive contribution of the algorithm however slows down the decrease
 215 in the prediction efficiency of the algorithm. Prediction accuracy still remains
 216 around 75% with only 40% of N_1 taxa found in the catalogue (Figures 3).
 217 Furthermore, the use of taxonomy for similarity measurements is more efficient
 218 as empirical data becomes scarcer and no different than resource/consumer sets
 219 for the complete algorithm when ample data is available (Figures 3).

220 4.3 Southern Gulf of St. Lawrence

221 As an example, we used the XXX algorithm to predict interactions in the south-
 222 ern Gulf of St. Lawrence (SGSL) in eastern Canada. The empirical data and
 223 taxa list come from Savenkoff et al. (2004). They present a list of 29 func-
 224 tional groups for a total of 80 taxa presented at least at taxonomical scale of
 225 the family. Other coarser taxa families were not used for this example (see
 226 Table S1 in Supplementary information (SI) and Savenkoff et al. (2004) for a
 227 complete description of functional groups). As their analysis was performed on
 228 the functional groups rather than the taxa themselves, we used the algorithm
 229 to predict interactions between all 80 taxa selected. We then aggregated them
 230 back to their original functional groups to compare with interactions presented
 231 in Savenkoff et al. (2004). In total, there were empirical data available in the
 232 catalogue for 78% of SGSL taxa (62/80). The algorithm correctly predicted
 233 close to 80% of interactions ($a = 135/170$) and non-interactions ($d = 354/455$)
 234 extracted from Savenkoff et al. (2004). It also predicted an additional 101 inter-
 235 actions (c) that were not noted in Savenkoff et al. (2004) and failed to predict
 236 36 observed interactions that were (c), resulting in a TSS score of 0.57. A vi-
 237 sual comparison of results obtained from the algorithm with interactions noted
 238 in Savenkoff et al. (2004) is available at Figure 4. The network presented is
 239 centered on the observed and predicted interactions of the capelin (*Mallotus*
 240 *villosus*) and piscivorous small pelagic feeders (e.g. *Scomber scombrus* and *Illex*
 241 *illecebrosus*).

242 5 Discussion

243 5.1 Algorithm accuracy

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

252 5.2 Usefulness of taxonomic relatedness

253 While we found that taxonomy could be useful as a complement to predictions
254 made using empirical data, the accuracy of predictions made using the KNN al-
255 gorithm could be improved. Other uses of this machine learning approach have
256 achieved much higher prediction rates (e.g. ?), which suggests that taxonomy
257 may not be the optimal proxy for predicting interactions. While evolutionary
258 history plays a significant role in influencing consumer-resource trait matching
259 and food web structure (Mouquet et al., 2012; Rohr and Bascompte, 2014), phy-
260 logenetic constraints do not account efficiently for certain traits such as body
261 size (Eklöf and Stouffer, 2016). Including traits like body size and metabolism
262 as an additional component of this algorithm could thus help increasing overall
263 prediction accuracy, especially in cases where the catalogue lacks data on taxa
264 for which interactions have to be predicted. Although promising, such an ap-
265 proach would undermine the premise under which this method was built and
266 which constitutes its main strength, *i.e.* predicting interactions in data deficient
267 environments using readily available data.

268 5.3 Interactions classification

269 That $Score_y$ and $Score_{\neg y}$ are inversely proportional means that non-interactions
270 are misclassified as interactions in the process of increasing $Score_y$, consequently
271 decreasing $Score_{\neg y}$. This could either stem from the algorithm poorly predicting
272 non-interactions or from the empirical data itself. Accuracy evaluation assumes
273 that non-interactions from empirical food web are observed data, yet it is usually
274 not the case. Most empirical webs have a strong focus attributed to higher order
275 consumer species and very little attention given to other taxa (?). Furthermore,
276 the methodologies used to obtain consumer-resource data usually relies on gut
277 content analyses, which is efficient at observing interactions, but not so for
278 absence of interactions (?). Misclassified interactions could thus be real, albeit
279 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. 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

- 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>.
- Bartomeus, Ignasi, Dominique Gravel, Jason M. Tylianakis, Marcelo A. Aizen, Ian A. Dickie, and Maud Bernard-Verdier (2016). “A common framework for identifying linkage rules across different types of interactions”. In: *Functional Ecology*, n/a–n/a. ISSN: 02698463. DOI: [10.1111/1365-2435.12666](https://doi.org/10.1111/1365-2435.12666). URL: <http://doi.wiley.com/10.1111/1365-2435.12666>.
- Beauchesne, David, Cindy Grant, Dominique Gravel, and Philippe Archambault (2016). “L’évaluation des impacts cumulés dans l’estuaire et le golfe du Saint-Laurent : vers une planification systémique de l’exploitation des ressources”. In: *Le Naturaliste canadien* 140.2, p. 45. ISSN: 0028-0798. DOI: [10.7202/1036503ar](https://doi.org/10.7202/1036503ar). URL: <http://id.erudit.org/iderudit/1036503ar>.
- Brose, Ulrich, Lara Cushing, Eric L. Berlow, Tomas Jonsson, Carolin Banasek-Richter, Louis-Felix Bersier, Julia L. Blanchard, Thomas Brey, Stephen R. Carpenter, Marie-France Cattin Blandenier, Joel E. Cohen, Hassan Ali Dawah, Tony Dell, Francois Edwards, Sarah Harper-Smith, Ute Jacob, Roland A. Knapp, Mark E. Ledger, Jane Memmott, Katja Mintenbeck, John K. Pinnegar, Björn C. Rall, Tom Rayner, Liliane Ruess, Werner Ulrich, Philip Warren, Rich J. Williams, Guy Woodward, Peter Yodzis, and Neo D. Martinez (2005). “Body sizes of consumers and their resources”. In: *Ecology* 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>.

363 Brose, Ulrich, Tomas Jonsson, Eric L. Berlow, Philip Warren, Carolin Banasek-
364 Richter, Louis-Félix Bersier, Julia L. Blanchard, Thomas Brey, Stephen
365 R. Carpenter, Marie-France Cattin Blandenier, Lara Cushing, Hassan Ali
366 Dawah, Tony Dell, Francois Edwards, Sarah Harper-Smith, Ute Jacob, Mark
367 E. Ledger, Neo D. Martinez, Jane Memmott, Katja Mintenbeck, John K.
368 Pinnegar, Björn C. Rall, Thomas S. Rayner, Daniel C. Reuman, Liliane
369 Ruess, Werner Ulrich, Richard J. Williams, Guy Woodward, and Joel E.
370 Cohen (2006). “Consumer-resource body-size relationships in natural food
371 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)
372 [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)
373 [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).

374 Cazelles, Kévin, Miguel B. Araújo, Nicolas Mouquet, and Dominique Gravel
375 (2016). “A theory for species co-occurrence in interaction networks”. In:
376 *Theoretical Ecology* 9.1, pp. 39–48. ISSN: 1874-1738. DOI: [10.1007/s12080-](https://doi.org/10.1007/s12080-015-0281-9)
377 [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)
378 [0281-9](http://link.springer.com/10.1007/s12080-015-0281-9).

379 Chamberlain, Scott A. and Eduard Szöcs (2013). “taxize: taxonomic search
380 and retrieval in R”. In: *F1000Research* 2. ISSN: 2046-1402. DOI: [10.12688/](https://doi.org/10.12688/f1000research.2-191.v1)
381 [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)
382 [2-191/v1](http://f1000research.com/articles/2-191/v1).

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

386 Cohen, Joel E, Tomas Jonsson, and Stephen R Carpenter (2003). “Ecological
387 community description using the food web, species abundance, and body
388 size.” In: *Proceedings of the National Academy of Sciences of the United*
389 *States of America* 100.4, pp. 1781–6. ISSN: 0027-8424. DOI: [10.1073/pnas.](https://doi.org/10.1073/pnas.232715699)
390 [232715699](https://doi.org/10.1073/pnas.232715699). URL: [http://www.ncbi.nlm.nih.gov/pubmed/12547915%](http://www.ncbi.nlm.nih.gov/pubmed/12547915)
391 [20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=](http://www.ncbi.nlm.nih.gov/pubmed/12547915)
392 [PMC149910](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC149910).

393 Eklöf, Anna, Matthew R. Helmus, M. Moore, and Stefano Allesina (2012). “Rel-
394 evance of evolutionary history for food web structure”. In: *Proceedings of the*
395 *Royal Society of London B: Biological Sciences* 279.1733.

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

400 Giakoumi, Sylvaine, Benjamin S. Halpern, Loïc N. Michel, Sylvie Gobert, Maria
401 Sini, Charles-François Boudouresque, Maria-Cristina Gambi, Stelios Kat-
402 sanevakis, Pierre Lejeune, Monica Montefalcone, Gerard Pergent, Christine
403 Pergent-Martini, Pablo Sanchez-Jerez, Branko Velimirov, Salvatrice Vizzini,
404 Arnaud Abadie, Marta Coll, Paolo Guidetti, Fiorenza Micheli, and Hugh P.
405 Possingham (2015). “Towards a framework for assessment and management
406 of cumulative human impacts on marine food webs”. In: *Conservation Biol-*
407 *ogy* 29.4, pp. 1228–1234. ISSN: 08888892. DOI: [10.1111/cobi.12468](https://doi.org/10.1111/cobi.12468). URL:
408 <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). "Review: Ecological networks - beyond food webs". In: *Journal of Animal Ecology* 78.1, pp. 253–269. ISSN: 00218790. DOI: [10.1111/j.1365-2656.2008.01460.x](https://doi.org/10.1111/j.1365-2656.2008.01460.x). URL: <http://doi.wiley.com/10.1111/j.1365-2656.2008.01460.x>.
- Kortsch, Susanne, Raul Primicerio, Maria Fossheim, Andrey V. Dolgov, and Michaela Aschan (2015). "Climate change alters the structure of arctic marine food webs due to poleward shifts of boreal generalists". In: *Proceedings of the Royal Society of London B: Biological Sciences* 282.1814.
- Laska, Mark S. and J. Timothy Wootton (1998). "Theoretical concepts and empirical approaches to measuring interaction strength". In: *Ecology* 79.2, 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). URL: [http://doi.wiley.com/10.1890/0012-9658\(1998\)079\[0461:TCAEAT\]2.0.CO;2](http://doi.wiley.com/10.1890/0012-9658(1998)079[0461:TCAEAT]2.0.CO;2).
- Martinez, Neo D. (1992). "Constant connectance in community food webs". In: *American Naturalist* 139.6, pp. 1208–1218. URL: <http://www.jstor.org/stable/2462337>.
- Morales-Castilla, Ignacio, Miguel G. Matias, Dominique Gravel, and Miguel B. Araújo (2015). "Inferring biotic interactions from proxies". In: *Trends in Ecology & Evolution* 30.6, pp. 347–356. ISSN: 01695347. DOI: [10.1016/j.tree.2015.03.014](https://doi.org/10.1016/j.tree.2015.03.014).
- Mouquet, Nicolas, Vincent Devictor, Christine N. Meynard, Francois Munoz, Louis-Félix Bersier, Jérôme Chave, Pierre Couteron, Ambroise Dalecky, Colin Fontaine, Dominique Gravel, Olivier J. Hardy, Franck Jabot, Sébastien Lavergne, Mathew Leibold, David Mouillot, Tamara Münkemüller, Sandrine Pavoine, Andreas Prinzing, Ana S.L. Rodrigues, Rudolf P. Rohr, Elisa Thébaud, and Wilfried Thuiller (2012). "Ecophylogenetics: advances and perspectives". In: *Biological Reviews* 87.4, pp. 769–785. ISSN: 14647931. DOI: [10.1111/j.1469-185X.2012.00224.x](https://doi.org/10.1111/j.1469-185X.2012.00224.x). URL: <http://doi.wiley.com/10.1111/j.1469-185X.2012.00224.x>.
- Pascual, M and JA Dunne (2006). *Ecological networks: linking structure to dynamics in food webs*. URL: <https://books.google.ca/books?hl=en%7B%5C%7Dlr=%7B%5C%7Ddid=YpQRDAAQBAJ%7B%5C%7Ddoi=fnd%7B%5C%7D>

454 %7Dpg=PP1%7B%5C%7Ddq=Pascual+and+Dunne+2006+interactions%7B%
 455 5C%7Dots=K4a5d62r9X%7B%5C%7Dsig=01fs%7B%5C%7DfXV1pgP6IeP1jBIb3B61rU.
 456 Poelen, Jorrit H., Stephen Gosnell, and Sergey Slyusarev (2015). *rglobi: R In-*
 457 *terface to Global Biotic Interactions*. URL: [https://cran.r-project.org/](https://cran.r-project.org/package=rglobi)
 458 [package=rglobi](https://cran.r-project.org/package=rglobi).
 459 Poelen, Jorrit H., James D. Simons, and Chris J. Mungall (2014). “Global biotic
 460 interactions: An open infrastructure to share and analyze species-interaction
 461 datasets”. In: *Ecological Informatics* 24, pp. 148–159. ISSN: 15749541. DOI:
 462 [10.1016/j.ecoinf.2014.08.005](https://doi.org/10.1016/j.ecoinf.2014.08.005).
 463 Rohr, Rudolf P. and Jordi Bascompte (2014). “Components of Phylogenetic Sig-
 464 nal in Antagonistic and Mutualistic Networks”. In: *The American Naturalist*
 465 184.5, pp. 556–564. DOI: [10.1086/678234](https://doi.org/10.1086/678234). URL: [http://www.journals.](http://www.journals.uchicago.edu/doi/10.1086/678234)
 466 [uchicago.edu/doi/10.1086/678234](http://www.journals.uchicago.edu/doi/10.1086/678234).
 467 Savenkoff, Claude, Hugo Bourdages, Douglas P. Swain, Simon-Pierre Despatie,
 468 J. Mark Hanson, Red Méthot, Lyne Morissette, and Mike O. Hammil (2004).
 469 *Input data and parameter estimates for ecosystem models of the southern*
 470 *Gulf of St. Lawrence (mid-1980s and mid-1990s)*. Tech. rep. Mont-Joli, Québec,
 471 Canada: Canadian Technical Report of Fisheries, Aquatic Sciences 2529, De-
 472 partment of Fisheries, and Oceans, p. 105.
 473 Schrod, Franziska, Jens Kattge, Hanhuai Shan, Farideh Fazayeli, Julia Joswig,
 474 Arindam Banerjee, Markus Reichstein, Gerhard Bönsch, Sandra Díaz, John
 475 Dickie, Andy Gillison, Anuj Karpatne, Sandra Lavorel, Paul Leadley, Chris-
 476 tian B. Wirth, Ian J. Wright, S. Joseph Wright, and Peter B. Reich (2015).
 477 “BHPMF - a hierarchical Bayesian approach to gap-filling and trait predic-
 478 tion for macroecology and functional biogeography”. In: *Global Ecology and*
 479 *Biogeography* 24.12, pp. 1510–1521. ISSN: 1466822X. DOI: [10.1111/geb.](https://doi.org/10.1111/geb.12335)
 480 [12335](https://doi.org/10.1111/geb.12335). URL: <http://doi.wiley.com/10.1111/geb.12335>.
 481 Tylianakis, Jason M., Raphael K. Didham, Jordi Bascompte, and David A.
 482 Wardle (2008). “Global change and species interactions in terrestrial ecosys-
 483 tems”. In: *Ecology Letters* 11.12, pp. 1351–1363. ISSN: 1461023X. DOI: [10.](https://doi.org/10.1111/j.1461-0248.2008.01250.x)
 484 [1111/j.1461-0248.2008.01250.x](https://doi.org/10.1111/j.1461-0248.2008.01250.x). URL: [http://doi.wiley.com/10.](http://doi.wiley.com/10.1111/j.1461-0248.2008.01250.x)
 485 [1111/j.1461-0248.2008.01250.x](http://doi.wiley.com/10.1111/j.1461-0248.2008.01250.x).

6.1 Box 1

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

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

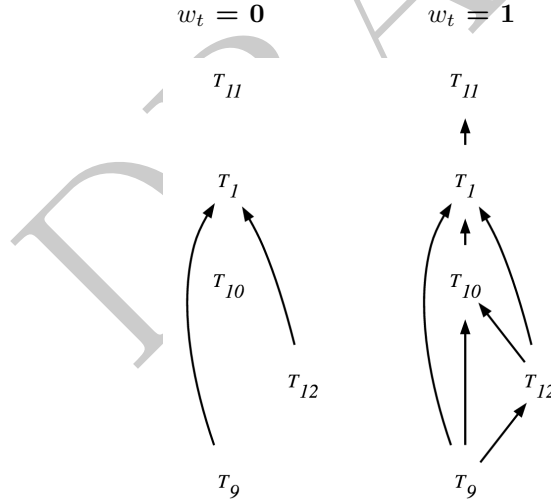
tanimoto(T_Cx, T_Cy) / tanimoto(T_Rx, T_Ry)												
	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8	T_9	T_{10}	T_{11}	T_{12}
T_1	-	0	0	0	0/1	0.3/1	0	0	0	0	0	0
T_2	0	-	0/0.5	0	0	0	0/0.3	0/0.3	0/0.5	0	0	0/0.5
T_3	0	0	-	0	0	0	0	0/0.5	0/1	0	0	0
T_4	0	0	0	-	0	0	0	0	0	0	0	0
T_5	0.5	0	0	0	-	0.3/1	0	0	0	0	0	0
T_6	0	0	0.2	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	-

tanimoto(T_Tx, T_Ty)

500 From these, the algorithm goes through logical steps (Figure 1) to identify
 501 a candidate resource list C_R for each taxon in N_1 using either empirical data
 502 directly or K most similar taxa with equation 2. Going through the process for
 503 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	$T_9 = \text{yes}$	$\{T_9, T_{12}\}$	$\{T_9, T_{10}\}$

504 The logical steps allow us to predict a set of resources for $T_1 = \{T_9, T_{10},$
 505 $T_{12}\}$. Doing it for all taxa in N_1 with $w_t = 0$ and 1 predicts the following
 506 networks:



6.2 Figures

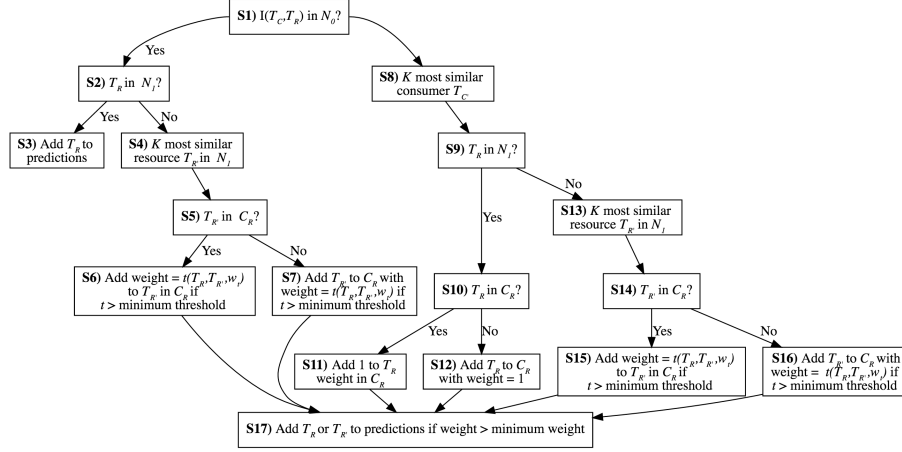


Figure 1: Description of logical steps used by the algorithm to suggest a list of candidate resources (C_R) for each consumer taxa (T_C) in an arbitrary set of N_1 for which interactions are predicted, using a set of taxa N_0 with empirically described interactions. Interactions between consumer and 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 and the minimum weight is the minimal weight value accepted for candidate resources to be selected as predicted sources in the algorithm.

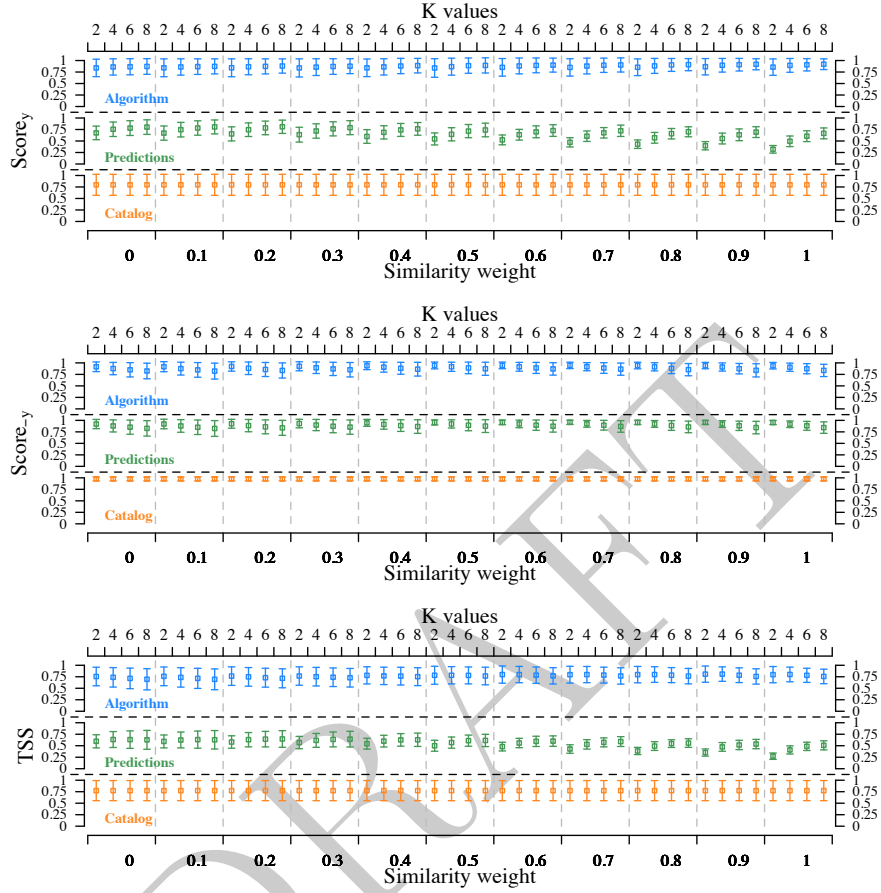


Figure 2: The graph presents the three statistics (*i.e.* $Score_y$, $Score_{-y}$ and TSS) used to evaluate the accuracy of the algorithm as a function of as a function of K values tested (*i.e.* 2, 4, 6 and 8 most similar seighbours, top x -axis) and trait weight (bottom x -axis), which varies between 0 and 1, and . 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 statistics, the topmost graph presents prediction accuracy for the complete algorithm, the middle graph corresponds to predictions made through the predictive portion of the algorithm (Steps S4-S16; Figure 1) and the bottom graph 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 (**Christian1999**; **Link2002**; **Thompson2005**; Brose et al., 2005; Barnes et al., 2008; Kortsch et al., 2015)

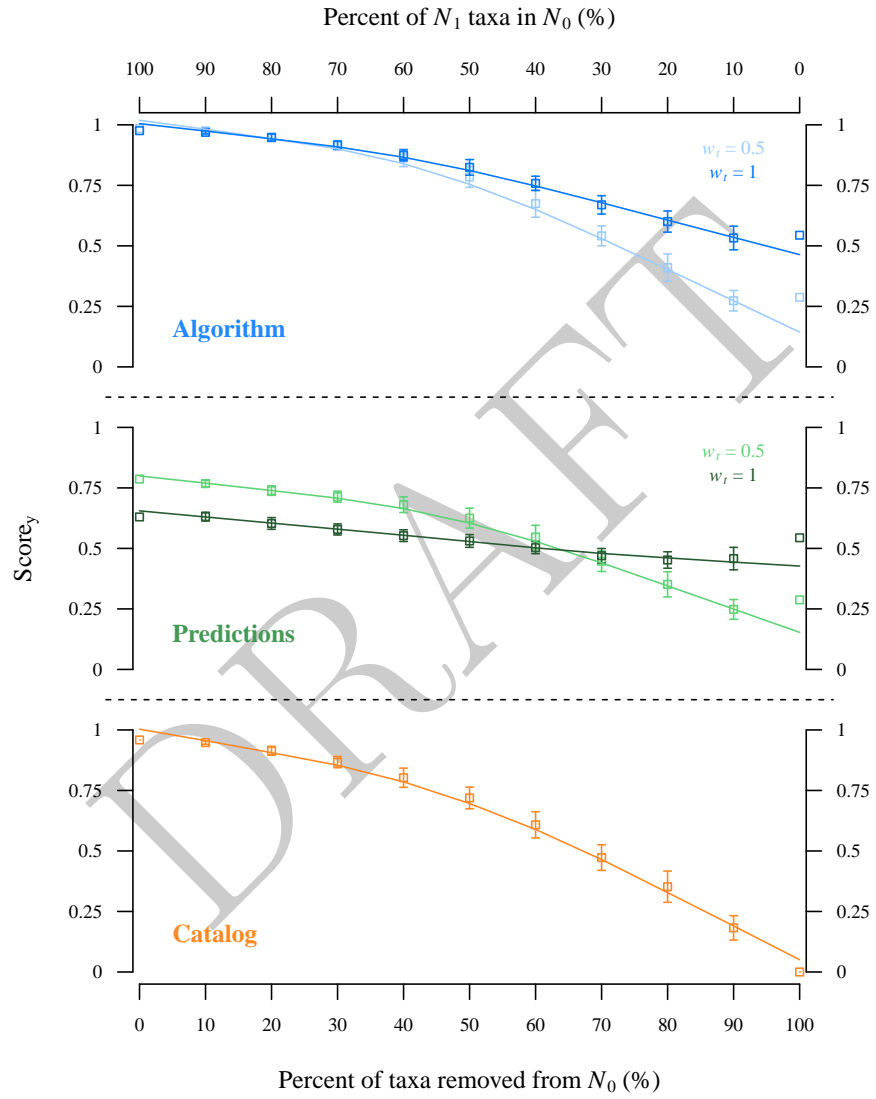


Figure 3: Caption on next page.

Figure 3: Graph presenting $Score_y$ as a function of catalogue comprehensiveness, *i.e.* the amount of information on sets of consumer and resources available in the catalogue. We tested this on 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 graph presents prediction accuracy for the complete algorithm, the middle graph corresponds to predictions made through the predictive portion of the algorithm (Steps S4-S16; Figure 1) and the bottom graph 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.

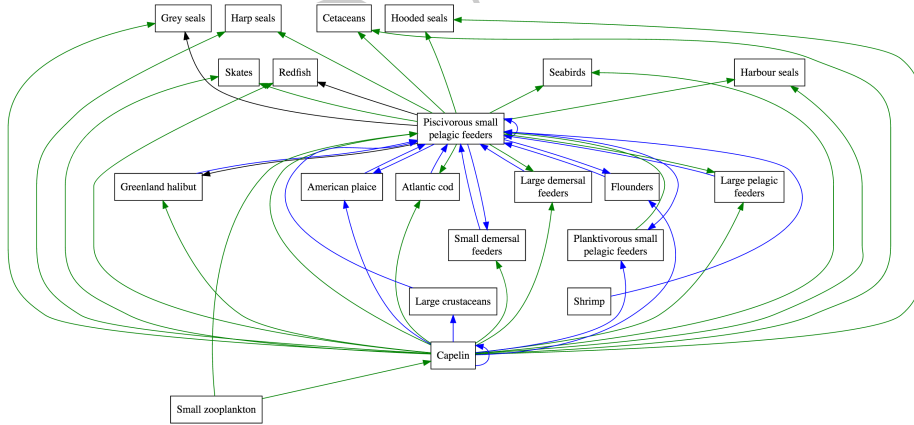


Figure 4: Example of results from the algorithm with the Network of the southern Gulf of St. Lawrence (Savenkoff et al., 2004) centered on interactions of the capelin (*Mallotus villosus*) and piscivorous small pelagic feeders (*e.g.* *Scomber scombrus* and *Illex illecebrosus*). Edge with colors green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