# Bayesian Networks as a Novel Tool to Enhance Interpretability and Predictive Power of Ecological Models

**Edwin Hui**[1] **Richard Stafford**[2] **Iain M. Matthews**[1] **V. Anne Smith**[1]

[1]Centre for Biological Diversity , School of Biology, University of St Andrews, St Andrews, Fife KY16 9TH, United Kingdom

[2]Department of Life and Environmental Sciences, Faculty of Science and Technology, Bournemouth University, Poole, BH12 5BB, United Kingdom

## 1 EXTENDED ABSTRACT

In today's world, it is becoming increasingly important to have the tools to understand, and ultimately to predict the response of ecosystems to disturbance [Steffen et al., 2018]. However, understanding such dynamics is not simple. Ecosystems are a complex network of species interactions, and therefore any change to a population of one species will have some degree of community level effect [Stafford et al., 2013]. Computational inference of complex networks presents an efficient route to reveal complex interactions such as those within an ecosystem and have been demonstrated to work on some natural complex systems. For example, the use of Bayesian networks (BN) has seen successful applications in molecular biology and ecology, where it was able to recover known links in the respective systems it was applied to [Chen and Mar, 2018, Friedman et al., 2000, Hecker et al., 2009, Milns et al., 2010, Mitchell et al., 2021].

A key strength of BNs is the ability to at least semi-quantify the strength of interactions between variables (or species). This can be done by utilizing the influence score [Yu et al., 2004] and the mutual information (MI). Here, the influence score represents the direction and magnitude of influence, where scores range from -1 to 1, with a score of exactly 0.0 representing non-monotonic (e.g. hump– or U–shaped) influence. This can be complemented with the MI, which measures the degree of mutual dependence between two variables, where it quantifies the amount of information obtained about one random variable through the observation of another random variable. Using this information, it follows that relative variable importance with respect to a certain variable of interest can be inferred from the revealed network. This has two key implications. Firstly, knowing the relative interaction strength of certain species in relation to others is invaluable in the field of ecology. For example, knowing the strength of interactions between predator, prey and competitors would give a clear understanding of how ecological communities are structured and regulated. Secondly, using the structure of the revealed network via the Markov blanket to infer the relevant variables in relation to a target variable could potentially serve as a novel variable selection tool in the field of machine learning. To this end, we evaluate the potential usefulness of BNs in two aspects. Firstly, we apply BN inference on species abundance data from a rocky shore ecosystem in Scotland, a system with well documented links, to test the ecological validity of the revealed network. Secondly, we evaluate BNs as a novel variable selection method to train an artificial neural network (ANN) for each component species of the network. To evaluate the effectiveness of BN as a variable selection method, we compare the performance of the ANN with and without the BN-based variable selection. Finally, to benchmark this approach against previous methods, we compare the performance of the ANN with variable selection against a generalised linear model (GLM) with variable selection, where variable selection has been performed by the BN.

The application of Bayesian networks to rocky shore ecosystems predicted relationships between species well and provided relative weights to indicate the importance of the interactions. Although these results are not those expected to be obtained immediately after experimental manipulations, these results match what one would expect from a 'static' rocky shore system (i.e. those which are in a 'stable state' rather than those adapting post experimental disturbance). For example, the grazing relationships revealed from the network was consistent with prior knowledge as macroalgae has been documented to provide food (via spores) and shelter for grazing snails, whilst simultaneously acting as a buffer from wave action [Norton et al., 1990]. Here, a positive link was found between grazers and macroalgae– however, in a dynamic system with the consideration of time, we would also expect grazers to reduce the level of establishing macroalgae [Hidalgo et al., 2008]. Therefore, it could also be argued that there should be a negative link found between grazers and macroalgae. Given that sampling occurred at a time where macroalgae had already established, these patches therefore acted as a refuge for littorinids. This

*Accepted for 15th Bayesian Modeling Applications Workshop at UAI 2021* (BMAW2021).

led to a positive link being recovered, which was consistent with expectations for areas with established macroalgae [Norton et al., 1990].

When utilizing the results from the BN as a variable selection tool to train ANNs, there were general improvements to model performance in ANN models with variable selection compared to models with no variable selection. This difference was significant for barnacles and macroalgae. From our models, there was a clear distinction of predictive performance of sessile species (barnacles and macroalgae) and mobile species (limpets and littorinids), where models of sessile species performed better than models of mobile species. From a causal perspective, this would make sense. Given the barnacles and macroalgae confer alternative assemblages on the rocky shore, the presence of one sessile species has strongly affects the presence of another. However, it should be noted that physical abiotic factors such as wave exposure and site angle also have a strong influence on the distribution of these sessile species. Here, while species count data was sufficient in training an accurate model for sessile species, the lack of physical abiotic factors limited ability to fully capture the variation in grazers. Factors such as exposure to wave action, desiccation, and shore angle, along with biotic factors such as predation and competition all have a dynamic role in structuring intertidal communities [Raffaelli and Hawkins, 1999].

Here, we demonstrate that coupling the results from the BN can complement the strong predictive abilities of ANNs. BNs show strong potential in revealing ecological networks, however due to the discretization of data, it is sub-optimal as a predictive tool as we would be limited to a classification type problem. On the other hand, ANN show very strong predictive capabilities, where the predictive power was significantly superior compared to GLM models. However, due to its black box nature, the results are generally difficult if not impossible to interpret. Given that explaining ecological relationships is integral in the study of ecology, ANNs may not be optimal as a standalone model. Thus, combining the results from BNs and ANNs can in turn can overcome each models' respective shortcomings, where BNs can be used to explain the underlying relationships between variables, which can inform the training of ANNs to ultimately develop strong predictive models.

To conclude, this novel Bayesian-neural network approach shows promise in the field of ecology as it can achieve two important features: 1) it has potential to reveal ecological network structures in different ecosystems, where existing relationships between species and other functional components are not known; and 2) it can guide the training of powerful predictive models by serving as a robust variable selection tool.

## References


S. Chen and J.C. Mar. Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. *BMC Bioinformatics*, 19(1):1–21, 2018.

N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using Bayesian Networks to Analyze Expression Data. *Journal of Computational Biology*, 7(3–4):601–620, 2000.

M. Hecker, S. Lambeck, S. Toepfer, E. van Someren, and R. Guthke. Gene regulatory network inference: Data integration in dynamic models—A review. *Biosystems*, 96(1):86–103, 2009.

F.J. Hidalgo, F.N. Firstater, E. Fanjul, M.C. Bazterrica, B.J. Lomovasky, J. Tarazona, and O.O. Iribarne. Grazing effects of the periwinkle *Echinolittorina peruviana* at a central Peruvian high rocky intertidal. *Helgoland Marine Research*, 62(1):73–83, 2008.

I. Milns, C.M. Beale, and V. Anne Smith. Revealing ecological networks using Bayesian network inference algorithms. *Ecology*, 91(7):1892–1899, 2010.

E.G. Mitchell, M.I. Wallace, V.A. Smith, A.A. Wiesenthal, and A.S. Brierley. Bayesian Network Analysis reveals resilience of the jellyfish *Aurelia aurita* to an Irish Sea regime shift. *Scientific Reports*, 11(1):1–14, 2021.

T.A. Norton, S.J. Hawkins, N.L. Manley, G.A. Williams, and D.C. Watson. Scraping a living: a review of littorinid grazing. *Hydrobiologia*, 193(1):117–138, 1990.

D. Raffaelli and S. Hawkins. *Intertidal Ecology*. Springer Netherlands, 1999.

R. Stafford, V.A. Smith, D. Husmeier, T. Grima, and B. Guinn. Predicting ecological regime shift under climate change : New modelling techniques and potential of molecular-based approaches. *Current Zoology*, 59(3): 403–417, 2013.

W. Steffen, J. Rockström, K. Richardson, T.M. Lenton, C. Folke, D. Liverman, C.P. Summerhayes, A.D. Barnosky, S.E. Cornell, M. Crucifix, J.F. Donges, I. Fetzer, S.J. Lade, M. Scheffer, R. Winkelmann, and H.J. Schellnhuber. Trajectories of the Earth System in the Anthropocene. *Proceedings of the National Academy of Sciences*, 115(33):8252–8259, 2018.

J. Yu, V.A. Smith, P.P. Wang, A.J. Hartemink, and E.D. Jarvis. Advances to Bayesian network inference for generating causal networks from observational biological data. *Bioinformatics*, 20(18):3594–3603, 2004.