The Dual Role of Subaerial Biofilms through the Lens of AI: the case for Causal Networks and Targeted Learning

Gabriele Gianini[1], Francesca Cappitelli[2], Sara Goidanich[3],
Fabrizio Passarini[4] and Letizia Berti[3]

[1] Università degli Studi di Milano-Bicocca, Italy
[2] Università degli Studi di Milano, Italy
[3] Politecnico di Milano, Italy
[4] Università degli Studi di Bologna, Italy

Abstract. Subaerial biofilms (SABs) are microbial communities that form on surfaces exposed to both air and periodic moisture and that can adapt to harsh environmental conditions like UV radiation, and fluctuating temperatures. On the one hand they can protect built surfaces by forming a barrier against environmental stressors, on the other they can also cause deterioration through biological weathering. The balance is complex and depend on a large number of factors. Unfortunately, only a small part of the complex multiscale network of physical, chemical and biological processes is captured by existing mechanistic model; this prompts for the involvement of phenomenological models. In this work we point at the modeling advantages offered by Bayesian Networks (BNs), Causal Networks (or Structural Causal Models, SCMs) and Targeted Learning (TL). The three related frameworks can model the complex interactions between biofilms, substrates, and atmosphere by identifying cause-effect relationships. SCMs can help devise interventions able to keep the biofilm within a desirable range of conditions, furthermore they can help assess the transportability of the discovered causal models across settings; Targeted Learning helps focusing on a specific research question (such as estimating the influence of factors on a target variable; quantifying their interactions; distinguishing between direct and mediated influence) without requiring the full discovery of the joint variable distribution, and using a semiparametric approach, as free as possible from modeling biases. In this work we highlight the considerable potential of the use of those related frameworks to the study of SABs.

Keywords: Subaerial Biofilms, Built Environment, Biofilm-substrate Interactions, Bioprotection, Bayesian Networks, Causal Networks, Structural Causal Models, Targeted Learning.

1 Introduction

Subaerial biofilms (SABs) are microbial communities embedded in an extracellular polymeric substance (EPS) matrix, made up of components like polysaccharides, proteins, and extracellular DNA, that attach to solid surfaces, forming an interface between solid surfaces and air. SABs are highly relevant in many contexts, including built heritage, where their dual role in its protection (bioprotection) and its deterioration is still an open point of discussion.

Unlike biofilms submerged in water, air-exposed biofilms face quickly varying chemical and physical stresses such as drying winds. To combat these stresses, SABs develop irregular structures that conform to the intricate surface topography while at the same time promoting a favorable surface-to-volume ratio, helping to conserve water while contrasting desiccation. Furthermore, the high energy cost of survival promotes symbiotic relationships, for instance between phototrophs and chemoorganotrophs: the former provide fixed carbon, while the latter decompose organic matter.

From the point of view of the built surfaces, SABs can provide protection by forming a barrier against environmental stressors, but they can also cause deterioration through biological weathering, such as acid production and physical deterioration, leading to structural damage over time. Finally, the interaction of SABs with the atmosphere is also interesting, since, at large scale, they can provide several benefits through their roles in carbon sequestration and reduction of atmospheric CO_2, nitrogen cycling, including atmospheric nitrogen fixation, absorption of atmospheric pollutants, and heavy metal uptake in urban or industrial settings.

From this general picture it is clear that the modeling of the interaction among SABs, substrate and atmosphere is complex, spans several magnitude scales and requires the integration of models from different disciplines, such as biology, material science and atmospheric chemistry. This still represents a largely open challenge. An even harder one consists of devising methods to influence the SAB colonies growth to keep them within a range of desirable conditions and achieve a desired level of bioprotection.

One of the main reasons why these objectives are challenging is that *mechanistic models* are available only for a small part of the involved phenomena: by mechanistic models we mean models that seek to model the system's behavior through explicit knowledge of its components and their interactions formalized through mathematical equations or simulations. This lack prompts for the use of *phenomenological models* (or empirical/data-driven models), which focus on describing the observed relationships and patterns in data – e.g. in functional form – without necessarily relying on a deep understanding of the underlying mechanisms. This further raises the need to integrate models of different types and at different scales.

In this work we highlight the considerable potential of the use of Bayesian Networks (BNs), of Causal Networks (Structural Causal Models, SCMs) and of Targeted Learning (TL) as suitable frameworks to build phenomenological models and address a number SAB modeling challenges, and we highlight the innovative character of this application.

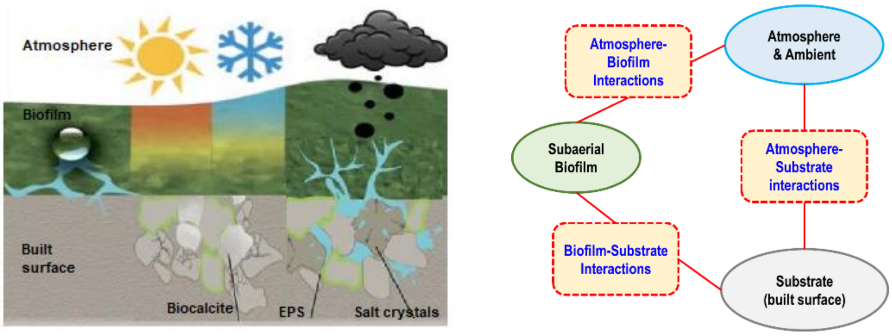

Fig. 1. Left: a stylized view of the overall system encompassing atmospheric agents, substrate layer and SAB layer; Water sequestration outside and within the substrate are also represented, along with the EPS (light green), clay minerals (grey), and salt crystals (white); the role on bioprotection and biodeterioration of these elements is discussed in the text. Right: a schema of the overall system components and their relationships, discussed in the text; the atmosphere-substrate interaction is studied in the absence of biofilm, for comparison.

The paper is structured as follows: first we introduce the SABs bioprotection/biodeterioration problem (Section 2); then we identify the modelling challenges and the requirements the solutions should meet (Section 3) and finally point to the possible methods related to BNs, SCMs and TL that could be used to address the challenges (Section 4).

2 SABs interaction and bioprotection of the built environment

SABs can determine bioprotection or biodeterioration of built materials through different mechanisms, involving both the interactions with the substrate and with the atmosphere, and regulated by the internal biofilm dynamics.

2.1 Internal Biofilm dynamics

SABs have remarkable resilience and self-regenerative properties, especially under stress conditions like drying, physical disturbances, or temperature extremes. This resilience is largely due to the EPS matrix, which acts as a protective barrier and supports the biofilm's ability to recover from damage. The microbial community within the biofilm can repair or regenerate damaged areas by secreting new EPS, re-establishing colony growth, and adapting to the new environmental conditions.

Contamination from other microorganisms can disrupt or benefit the delicate balance of the biofilm's microbial community. These contaminants may outcompete/establish symbiosis with biofilm species, reduce/improve the efficiency of nutrient cycling, or introduce harmful/useful metabolic byproducts. Interestingly, microbiological contamination may sometimes introduce new species that contribute positively to the biofilm's capabilities, such as nitrogen fixation or pollutant degradation. The response to stress of a healthy biofilm is accessible to laboratory experimentation, where stress conditions can be reproduced quite accurately.

To characterize the behavior of the biofilm one can distinguish the features of the EPS matrix from those of the microbial components. The EPS matrix can be described in terms of its composition and structure, for instance in terms of the presence of channels, the polymeric character of the matrix determines a certain level of hydrophobicity; its structures regulate water retention. The microbial component can be characterized by its structure and dynamics. Among the main features are: thickness, cellular biomass/EPS mass ratio; in multicomponent colonies the microbial diversity is important in terms of richness and relative abundances of species, and the nutrient utilization profiles; and self-healing and self-regenerating capacities.

2.2 Biofilm-Substrate Interactions

The presence of biofilm layer over the surface can have a dual role. On the one hand it can: act as a barrier against the deteriorating agents; stabilize the moisture content inside the substrate and decrease the thermal stresses. On the other hand, the biofilm may retain extra water and water vapor inside the substrate and, when the biofilm color is darker than the substrate, may increase the surface temperature during solar irradiation.

The microbial and EPS penetration inside the substrate can on the one hand sta-bilize the crystalline micro-structure of the materials by holding together the grains, while on the other can cause physical damage.

Biomineralization is the process by which microorganisms in biofilms alter their local environment in a way that promotes the precipitation of minerals from the surrounding atmosphere or from substrates they colonize. On the one hand by obtaining calcium from external sources the microorganisms can produce new crystals that can act as consolidation agents, on the other, microorganisms may obtain the calcium from the substrate, thus degrading the material.

Also the EPS matrix filling the space between grains and acting as link between the substrate and the biofilm (see Figure 1) can play a double role: it can stabilize the free water content of the support thus reducing swelling phenomena of clay minerals and improving resistance to salt decay; however, the swelling of the EPS itself in presence of water may cause mechanical stress. Furthermore, the bio-films may speed up the process of salt crystallization by sequestrating the water from the space between grains.

Among the most relevant properties characterizing the substrate there are: chemical composition, porosity, surface topography (roughness/smoothness), hydrophobicity/hydrophilicity, water vapor permeability, presence of nutrients, residuals of preexisting biofilms, presence of antimicrobial agents; also variables as temperature and level of moisture (which change within daily and seasonal cycles) are important.

2.3 Biofilm-atmosphere (and -ambient) interactions

Functional to the understanding of the overall biofilm good health and resilience is also the modeling of the interaction of the biofilm with the atmosphere with respect to carbon fixation and depollution processes.

The ability of biofilms to fix CO_2, particularly when phototrophic organisms are present, plays a crucial role in maintaining the metabolic health of the biofilm. By converting atmospheric CO_2 into organic compounds through photosynthesis, these microorganisms supply essential nutrients that can support the growth and sustainability of the entire biofilm community.

Biofilms can act as biological filters, capturing and metabolizing pollutants such as particulate matter (PM10, ultrafine particles) and harmful chemicals like sulfur and nitrogen oxides. These are absorbed into the EPS matrix, where they can be degraded or neutralized by specific microbial species. This bioremoval process not only contributes to the environmental clean-up but also helps mitigate the damage pollutants can cause to the substrate, and in some cases provides the microorganisms with a potential source of nutrients or energy.

The atmosphere may be obviously characterized by its composition, especially in terms of CO_2 and Nitrogen concentration, presence of gaseous pollutants, or substances harmful to the EPS or substrate; other relevant characteristics the ambient can be described in terms of the power spectrum and photoperiod of radiations, distinguishing infrared, from optical and UV frequency bands; also the temperature and moisture cycles are important in the regulation of biofilm functions.

Overall, the factors determining bioprotection and biodeterioration are manifold and, as mentioned above, can display contrasting effects.

3 Modelling Challenges

Central to the pursuit of the SABs modeling objectives in the context of the rate biodeterioration/bioprotection of the built environment is the task of developing models of the relevant interconnected subsystems (support, atmosphere, SABs and subsystems, such as the EPS matrix and different microbial taxa), portraying *correlations* and *cause-effect links* among the different factors of interest, within and across subsystems.

Those models are essential not only for forecasting the free development and behavior of the system, but also for predicting its behavior under the *effect of interventions* (this is functional to the purpose of keeping the SAB within a range of desirable conditions). This task entails the collection of both *observational data* and *interventional data* (the latter ideally from randomized experiments) and their integration with available domain knowledge, to build, adapt or calibrate mechanistic and phenomenological models of the subsystems and combine them in a comprehensive model.

This multifaceted task is confronted to serious challenging problems, whose solutions should meet a number of requirements, discussed below.

3.1 Complex coupled multi-scale phenomena

From the description of the interactions in previous section it becomes apparent that the overall substrate-biofilm-atmosphere system is highly heterogeneous, and that each component is by itself a complex system, characterized by a large number of structural and dynamic variables. Typically, the subsystems interactions are insufficiently understood (Cappitelli 2014)(Zavrel 2019), and detailed models exist only for some mechanisms in certain subsystems (Li 2022) (Picioreanu 2007) (Lesnik 2017a and 2017b), furthermore quantitative modeling is generally lacking (Zavrel 2019). In addition, the overall system features complex coupled chemical, physical and biological phenomena, interacting over <u>different but interconnected spatial scales</u> (e.g. metabolic pathways, cell-to-cell signaling, ecology of bacterial colony, interactions with substrate, urban environment). The currently available mechanistic models – first-principle-based biochemistry, environmental chemistry and physics models, represented in the analytic form of equations or as simulative models – are characterized by a relatively targeted focus and cannot individually span the system as a whole.

> Requirement A. These aspects prompt for the adoption and development of alternative phenomenological (empirical/data-driven) models, and for the construction of novel integrated hybrid (mechanistic and phenomenological) *multiscale models*.

3.2 Heterogeneous kinds of knowledge and uncertainty

Furthermore, the knowledge available to build those models is not always represented by *quantitative information*, but often by information of *qualitative, non-numerical* character. Another important point is that experimentation only measures a part of what is relevant: only some of the variables and their relationships can be measured (the 'known knowns'). For some variables, their existence is known, but they are not measured (the 'known unknowns'). However, a further portion of the relevant variables may not even be known, though their existence is hypothesized (the 'unknown unknowns', or 'unrecognized sources of uncertainty'), this issue is seldom addressed in the scientific modeling practice (Capote 2020) (Schnabel 2021) (Kornilov 2021).

> Requirement B. These aspects prompt for the adoption of hybrid models set to cope with *both quantitative and qualitative knowledge,* and able to combine them with the new experimental data. It also requires the models to *account for the different kinds of uncertainty* of the structure and of the parameter of the different representations.

3.3 Need for interventional data

What is more, one has to consider that not only the collection of *observational data* from in-vitro and in-vivo experiments implies a substantial cost, but also that the acquisition of new cause-effect information, i.e. *interventional data*, from randomized experiments (where a group of control variables is set to a collection of representative values) entails a considerable outlay of resources.

> Requirement C. In *experimental resource optimization* one should benefit from techniques to extract interventional data from observational data (e.g. through instrumental variable techniques, mediation analysis or causal calculus as detailed below), thus replacing costly randomized experiments with observational ones wherever applicable.

3.4 Data sparsity, weak models and transportability

Almost always the data collected have a high cost of collection and require long periods of observation. As a consequence, the analysis has to be performed on *sparse data* (panel/longitudinal data typically of a small number of use cases), characterized by statistical and systematic uncertainties. This issue is entangled with the lack of well- established models for the relationships of interest. Often one cannot assume specific functional or statistical relationships among variabilities: parametric models unsupported by evidence can lead to misspecification and biases. Weak conclusions can hinder the possibility to generalize the findings and hinder the transportability of the causal models and inferences to different settings.

> Requirement D. Addressing these issues requires flexible and robust statistical methods, particularly based on non-parametric or semi-parametric statistical techniques, for example from the AI sub-field of Machine Learning. The methods should allow to assess the transportability of the causal models and causal reasoning findings to different scenarios.

4 Possible modeling approaches: Bayesian Nets, SCMs and TL

Framework have arisen within relatively recent developments in Artificial Intelligence and Computational Intelligence, able to address all the requirements entailed by the above mentioned challenges – A. the capability of building hybrid (both mechanistic and phenomenological) integrated multiscale models, B. the ability of incorporating quantitative and qualitative expert knowledge accounting for different kinds of model and data uncertainty, C. the capability of predicting in notable cases the result of interventions without costly randomized experiments, D. the capability of coping with sparse data and the lack of parametric models.

A first mature methodological framework addressing many of those requirements is the one centered on *Bayesian Networks* (BNs) (Pearl 1994) (Smith 2010) (Scanagatta 2019) and their extensions. BNs are probabilistic models based on directed acyclic graphs (DAGs), whose nodes represent random variables, and whose directed arcs capture probabilistic dependence (conditional probabilities, qualitative relations).

A second mature framework which is rapidly expanding thanks to a lively research community is the one of *Structural Causal Models (SCMs)*, whose graph-based representation core is represented by Causal Networks (Pearl 2009): in these networks directed arcs have a very specific semantics, they signify cause-effect dependence.

Involving the SCMs as a key component, and encompassing various supervised Machine Learning algorithms as devices to learn the form of non-parametric relationship, is the Targeted Learning framework (Van der Laan 2011) (Van der Laan 2018).

4.1 Requirement A: hybrid mechanistic and phenomenological models

The BN framework allows for representations at different granularities (with part-of and is-a relationships), allows variable grouping and multi-scale representations (Kim 2022) (Becker 2021) and more general object-oriented inspired constructs (Molina 2010) (Benjamin-Fink 2017). Additionally, the graph form grants easy interpretability of these models. They have been used in different use cases in Biology use cases, mostly in Ecology (Rigosi 2015), Molecular Biology (e.g. gene expression, metabolic paths) (Friedman 2000), Neuroscience (Bielza 2014), Complex Biological System modelling (Smith 2010), but have had a limited use in the study of biofilm related phenomena (Hodges 2010) (Kannan 2020), and to the best of our knowledge, have never been applied to SABs.

4.2 Requirement B: quantitative and qualitative expert knowledge

BNs provide a rational method for the integration of the best attainable data from a variety of sources, including expert opinion (Schreiber 2017), (Hu 2023), prior knowledge (Pollino 2005), simulation results and empirical data (Wooldridge and Done 2004). Thus they allow information potentially of different quality, to be merged and easily updated. BNs account by construction for probabilistic uncertainty. Interval probability variants of BNs have been formalized (Mauà 2020) in the form of credal networks. Qualitative networks and hybrid quantitative-qualitative models have also been developed, (Zhang 2020) (Sazal 2018), and methods for a stepwise construction of hybrid networks have been proposed (Mitra 2018).

4.3 Requirement C: predicting the result of interventions

BN framework has been the basis for the subsequent development of the modeling framework of *Causal Networks* (Pearl 2009). This represents a mature research area of Artificial Intelligence; the area is characterized by an active community, whose established results have still to be fully exploited in Biology (Shipley 2012), and whose recent developments open the way for innovative solutions in several application areas. Causality research can generally be divided into two main branches: Causal Discovery and Causal Reasoning (or Causal Calculus).

Causal Discovery allows to *induce*, from observational data, *phenomenological causal models*. Those models can complement other well established *mechanistic models* from physics, chemistry and biology, and can act as *surrogate models* when high complexity or lack of information preclude the development of models based on known elementary mechanisms. This is precisely the case of SABs: developing models of the microbiome from first principles is difficult due to the high number of structural and dynamic variables and to the unknown or unmeasured interactions as well as a limited understanding of the mechanisms that underlie these interactions.

With *Causal Reasoning*, which mainly consisting of Pearl's Causal Calculus (the do-Calculus) and its extensions (Pearl 2009, 2021), once the cause-effect relationships that rule a system are available, one can predict the system behavior, should one of its parts be actively manipulated from outside: Causal Reasoning can quantify the actual influence of a factor variable over a target variable by filtering out those contributions due to contextual confounding factors, affecting both the factors and the target.

Thus, Causal Reasoning in several cases, allows to *predict the effects of the interventions using* the much less expensive *observational data*: this information can, in those cases, replace the collection of data through costly randomized experiments. Using Causal Reasoning and, when available, interventional data, one can quantify and rank the main drivers of complex system behavior. Developing suitable syntheses of this information can lead to the definition of *indicators summarizing the condition of the complex system*.

Causal Networks, furthermore, allow for handling missing data in a principled way (Mohan 2021), and can simplify knowledge update because they can break the problem in independent parts. Recent developments allow to deal with feed backs and cycles, common situations in biological systems.

Although applied to individual biology problems – ecology, molecular biology, neuroscience, see (Smith 2010), (Hamilton 2007), (Arhonditsis 2007), (Tsamardinos 2009) (Triantafillou 2017) (Tsourtis 2020) – *those tools have seldom been used to discovery and reasoning over an integrated set of models unfolding over different spatial and time scales, and never used in the SAB bio-protection or the air quality improvement scenarios of interest.*

Recent and potentially impactful developments of Causal Inference worth mentioning include the following methods:

1) *Joint discovery and reasoning*. the ABCI framework by (Toth 2022) integrates Causal Discovery and Causal Reasoning into a single process, and enables to jointly infer a posterior over causal models and queries of interest, allowing to save considerable resources.
2) *Multiple interventions*. Further efficiency can be gained by simultaneous controlled interventions on multiple variables (Multi-Perturbation Experimental Design), that reduces the number of needed experiments (Triantafillou 2015) (Sussex 2021) to elicit causal knowledge.
3) *Crude interventions*. A related problem is that classical atomic interventions are often not available (e.g., changing DNA base-pairs), instead, we only have access to indirect or crude interventions (e.g., one intervenes on several hub genes in full awareness that this may spur unintended interactions with other biological processes). Gultchin (Gultchin 2021) proposes (a) a two-step *mediator variable method* for predicting the causal responses of crude interventions (b) a testing procedure to identify mediators of crude interventions. Alier (Ailer 2023) proposes a sequential procedure for the selection of instrumental variables (variables that only affect the outcome indirectly, via the treatment variable).
4) *Partially unknown interventions*. Hagele (Hagele 2023) addresses the problem of imprecision intrinsic to many interventional setting and proposes a framework for causal discovery with partially known interventions, by introducing priors about the interventions in the formalism.

4.4 Requirement D: Non-parametric methods and transportability assessment

Data sparsity and the need for robust and non-parametric statistical methods can be addressed with the help of Targeted Learning (TL) (Van der Laan 2011) (Van der Laan 2018). TL integrates causal modeling and statistical estimation into a roadmap (Petersen 2014) to address key questions (such as estimating the influence of factors on a target variable; quantifying their interactions; distinguishing between direct and mediated influence), without requiring the full discovery of the joint variable distribution. TL encompasses SCMs as a fundamental component and uses Machine Learning algorithms to discover the functional relationships between nodes.

The use of TL and SCMs (Pearl 2009, 2021) fosters reproducibility of research procedures and results and supports the quantification of the dependence on the assumptions (sensitivity analysis). Thanks to TL and SCMs one can assess the transportability of the research findings across scenarios (simply based on statistical correlations in a domain one cannot tell whether a model can be transported to another): a set of techniques developed within the SCM framework allow to assess which relations are transportable and what the transport formula should be (Degtiar 2023) (Pearl 2011, 2014, 2015).

5 Conclusions

The study of the dual role (bioprotection and biodeterioration) of SubAerial Biofilms and the set-up of measures able to keep the biofilm within a desirable range of conditions require the capability of building integrated phenomenological models. One should organize heterogeneous qualitative-quantitative knowledge about the multi-scale systems involved through an organic framework handling different kinds of uncertainty, and able to guide the incremental planning towards the most informative experimental steps. In this work we argued that these requirements can be fulfilled by Bayesian Networks and by an integrated framework encompassing the Structural Causal Models (SCMs) framework (Pearl 2009), and the non-parametric inference framework of Targeted Learning (TL) (which relies also on Machine Learning techniques) (Van der Laan 2011) (Van der Laan 2018).

Acknowledgments

The work was partially supported by the MUSA-Multilayered Urban Sustainability Action project, funded by the European Union-NextGenerationEU, under the Mission 4 Component 2 Investment Line of the National Recovery and Resilience Plan (NRRP) Strengthening of research structures and creation of R&D "innovation ecosystems" (CUP G43C22001370007, Code ECS00000037); Program "piano sostegno alla ricerca" PSR and the PSR-GSA-Linea 6; Project ReGAInS (code 2023-NAZ-0207/DIP-ECC-DISCO23), funded by the Italian University and Research Ministry, within the Excellence Departments program 2023-2027 (law 232/2016); and - FAIR - Future Artificial Intelligence Research - Spoke 4 - PE00000013 - D53C22002380006, funded by the European Union - Next Generation EU within the project NRPP M4C2, Investment 1.,3 DD. 341 - 15 March 2022. The views and opinions expressed are those of the authors alone and do not necessarily reflect those of the European Union or the Italian MUR. Neither the EU nor the Italian MUR can be held responsible for them.

References

(Ailer 2023) Ailer, E., Hartford, J., & Kilbertus, N. (2023). Sequential Underspecified Instrument Selection for Cause-Effect Estimation. ArXiv. /abs/2302.05684

(Arhonditsis 2007) Arhonditsis GB, Qian SS, Stow CA, Lamon EC, Reckhow KH (2007) Eutrophication risk assessment using Bayesian calibration of process-based models: application to a mesotropic lake. Ecol Model 208:215-229 - https://doi.org/10.1016/j.ecolmodel.2007.05.020

(Becker 2021) Becker, A. K., et al. (2021). From heterogeneous healthcare data to disease-specific biomarker networks: A hierarchical Bayesian network approach. PLoS Computational Biology, 17(2), e1008735.

(Benjamin-Fink 2017) Nicole Benjamin-Fink, Brian K. Reilly, A road map for developing and applying object-oriented bayesian networks to WICKED problems, Ecological Modelling, Volume 360, 2017, Pages 27-44

(Bielza 2014) Bielza, Concha, and Pedro Larrañaga. "Bayesian networks in neuroscience: a survey." Frontiers in computational neuroscience 8 (2014): 131.

(Capote 2020) Capote, Roberto, et al. "Unrecognized sources of uncertainties (USU) in experimental nuclear data." Nuclear Data Sheets 163 (2020): 191-227.

(Cappitelli 2014) Cappitelli, F., Andrea Polo, and F. Villa. "Biofilm formation in food processing environments is still poorly understood and controlled." Food Engineering Reviews 6 (2014): 29-42.

(Degtiar 2023) Degtiar and S. Rose, A Review of Generalizability and Transportability. Annual Review of Statistics and Its Application, 2023. 10(1): p. 501-524. DOI: 10.1146/annurev-statistics-042522-103837.

(Friedman 2000) Friedman, Nir, et al. "Using Bayesian networks to analyze expression data." Proceedings of the fourth annual international conference on Computational molecular biology. 2000.

(Gultchin 2021) Gultchin, L., Watson, D., Kusner, M., & Silva, R. (2021, July). Operationalizing complex causes: A pragmatic view of mediation. In International Conference on Machine Learning (pp. 3875-3885). PMLR.

(Hagele 2023) Hägele, A., Rothfuss, J., Lorch, L., Somnath, V. R., Schölkopf, B., & Krause, A. (2023, April). Bacadi: Bayesian causal discovery with unknown interventions. In International Conference on Artificial Intelligence and Statistics (pp. 1411-1436). PMLR.

(Hamilton 2007) Hamilton GS et al. (2007) Investigating the use of a Bayesian Network to model the risk of Lyngbya majuscula bloom initiation in Deception Bay, Queensland, Australia. Hum Ecol Risk Assess 13:1271-1287

(Hodges 2010) Hodges, Andrew P., et al. "Bayesian network expansion identifies new ROS and biofilm regulators." PloS one 5.3 (2010): e9513.

(Hu 2023) Hu, H., & Kerschberg, L. (2020). Improving Causal Bayesian Networks using Expertise in Authoritative Medical Ontologies. ACM Transactions on Computing for Healthcare.

(Husmeier 2005) Husmeier, D. (2005). Introduction to learning Bayesian networks from data. In Probabilistic modeling in bioinformatics and medical informatics (pp. 17-57). London: Springer London.

(Kannan 2020) Kannan, Pranav, et al. "A systems-based approach for modeling of microbiologically influenced corrosion implemented using static and dynamic Bayesian networks." Journal of Loss Prevention in the Process Industries 65 (2020): 104108.

(Kim 2022) Kim, T., Lee, D., Shin, J., Kim, Y., & Cha, Y. (2022). Learning hierarchical Bayesian networks to assess the interaction effects of controlling factors on spatiotemporal patterns of fecal pollution in streams. Science of The Total Environment, 812, 152520.

(Kornilov 2021) Kornilov, Nikolay V., Vladimir G. Pronyaev, and Steven M. Grimes. "Systematic distortion factor and unrecognized source of uncertainties in nuclear data measurements and evaluations." Metrology 2.1 (2021): 1-18.

(Lesnik 2017a) Lesnik, Keaton Larson. "Consider the community: developing predictive linkages between community structure and performance in microbial fuel cells." (2017).

(Lesnik 2017b) Lesnik, Keaton Larson, and Hong Liu. "Predicting microbial fuel cell biofilm communities and bioreactor performance using artificial neural networks." Environmental science & technology 51.18 (2017): 10881-10892.

(Li 2022) Li, Zhuo, et al. "Model development of bioelectrochemical systems: A critical review from the perspective of physiochemical principles and mathematical methods." Water Research 226 (2022): 119311.

(Mauà 2020) Mauá, D. D., & Cozman, F. G. (2020). Thirty years of credal networks: Specification, algorithms and complexity. Internat. Journal of Approximate Reasoning, 126, 133-157.

(Mitra 2018) Mitra, E. D., Dias, R., Posner, R. G., & Hlavacek, W. S. (2018). Using both qualitative and quantitative data in parameter identification for systems biology models. Nature communications, 9(1), 3901.

(Mohan 2021) Mohan, K., & Pearl, J. (2021). Graphical models for processing missing data. Journal of the American Statistical Association, 116(534), 1023-1037.

(Pearl 1994) Pearl, Judea. "Belief networks revisited." (1994).

(Pearl 2009) J. Pearl, Causality: Models, Reasoning, and Inference, 2nd Edition. 2009: Cambridge University Press.

(Pearl 2011) Pearl. J and Barenboim. E. Transportability of causal and statistical relations: a formal approach. in 2011 IEEE 11th Int. Conf. on Data Mining Workshops. 2011.

(Pearl 2014) J. Pearl and E. Bareinboim, External Validity: From Do-Calculus to Transportability Across Populations. Statistical Science, 2014. 29(4).

(Pearl 2015) J. Pearl, Generalizing Experimental Findings. Journal of Causal Inference, 2015. 3(2): p. 259-266.

(Pearl 2021) J. Pearl, Probabilistic and Causal Inference: The Works of Judea Pearl, ed. H. Geffner, R. Dechter, and J.Y. Halpern. 2022: ACM Association for Computing Machinery.

(Petersen 2014) Petersen, Maya L., and Mark J. van der Laan. "Causal models and learning from data: integrating causal modeling and statistical estimation." Epidemiology 25.3 (2014): 418-426.

(Picioreanu 2007) Picioreanu, Cristian, et al. "A computational model for biofilm-based microbial fuel cells." Water research 41.13 (2007): 2921-2940.

(Pollino et al. 2005) Pollino CA, Woodberry O, Nicholson AE, et al. 2005. Parameterising Bayesian networks: A case study in ecological risk assessment In: Proceedings of the 2005 International Conference on Simulation and Modelling, Bangkok, Thailand.

(Rigosi 2015) Rigosi, Anna, et al. "Determining the probability of cyanobacterial blooms: the application of Bayesian networks in multiple lake systems." Ecological applications 25.1 (2015): 186-199.

(Ruiz-Perez 2021) Ruiz-Perez, Daniel, et al. "Dynamic bayesian networks for integrating multi-omics time series microbiome data." Msystems 6.2 (2021): 10-1128.

(Sazal 2018) Sazal, Musfiqur, et al. "Inferring directional relationships in microbial communities using signed Bayesian networks." BMC genomics 21 (2020): 1-11.

(Scanagatta 2019) Scanagatta, M., Salmerón, A., & Stella, F. (2019). A survey on Bayesian network structure learning from data. Progress in Artificial Intelligence, 8, 425-439.

(Schnabel 2021) Schnabel, Georg, et al. "Nuclear data evaluation with Bayesian networks." arXiv preprint arXiv:2110.10322 (2021).

(Schreiber 2017) Schreiber, J. M., & Noble, W. S. (2017). Finding the optimal Bayesian network given a constraint graph. PeerJ Computer Science, 3, e122.

(Smith 2010) Smith, V. A. (2010). Revealing structure of complex biological systems using Bayesian networks. Network Science: Complexity in Nature and Technology, 185-204.

(Sussex 2021) Sussex, S., Uhler, C., & Krause, A. (2021). Near-Optimal Multi-Perturbation Experimental Design for Causal Structure Learning. In Advances in Neural Information Processing Systems (pp. 777–788). Curran Associates, Inc..

(Toth 2022) Toth, C., Lorch, L., Knoll, C., Krause, A., Pernkopf, F., Peharz, R., & Kügelgen, J. (2022). Active Bayesian Causal Inference. In Advances in Neural Information Processing Systems (pp. 16261–16275). Curran Associates, Inc..

(Triantafillou 2017) Triantafillou, Sofia, et al. "Predicting causal relationships from biological data: Applying automated causal discovery on mass cytometry data of human immune cells." Scientific reports 7.1 (2017): 12724.

(Triantafillou 2015) Triantafillou, Sofia, and Ioannis Tsamardinos. "Constraint-based causal discovery from multiple interventions over overlapping variable sets." The Journal of Machine Learning Research 16.1 (2015): 2147-2205.

(Tsamardinos 2009) Tsamardinos, Ioannis, and Sofia Triantafillou. "The possibility of integrative causal analysis: Learning from different datasets and studies." Engineering Intelligent Systems 17.2 (2009): 163.

(Tsourtis 2020) Tsourtis, Anastasios, Yannis Pantazis, and Ioannis Tsamardinos. "Inference of Stochastic Dynamical Systems from Cross-Sectional Population Data." arXiv preprint arXiv:2012.05055 (2020).

(Van der Laan 2011) M.J. Van der Laan and S. Rose, Targeted learning: causal inference for observational and experimental data. Vol. 4. 2011: Springer.

(Van der Laan 2018) M.J. Van der Laan and S. Rose, Targeted learning in data science. 2018: Springer.

(Wooldridge and Done 2004) Wooldridge S and Done T. 2004. Learning to predict large-scale coral bleaching from past events: A Bayesian approach using remotely sensed data, in-situ data, and environmental proxies. Coral Reefs 23:96-108

(Zavrel 2019) Zavřel, Tomáš, Marjan Faizi, Cristina Loureiro, Gereon Poschmann, Kai Stühler, Maria Sinetova, Anna Zorina, Ralf Steuer, and Jan Červený. "Quantitative insights into the cyanobacterial cell economy." Elife 8 (2019): e42508.

(Zhang 2020) Zhang, Y., Apley, D. W., & Chen, W. (2020). Bayesian optimization for materials design with mixed quantitative and qualitative variables. Scientific reports, 10 (1), 4924.