Formalising contextual expert knowledge for causal discovery in linked knowledge graphs about transformation processes: application to processing of bio-composites for food packaging

Mélanie Munch, Patrice Buche, Helene Angellier-Coussy, Cristina Manfredotti, Pierre-Henri Wuillemin

Formalizing Contextual Expert Knowledge for Causal Discovery in linked Knowledge Graphs about Transformation Processes: Application to processing of bio-composites for food packaging

Melanie Munch

I2M, U. Bordeaux, INRAE, Talence, France
Mélanie Munch holds a doctorate in computer science from the Paris-Saclay university. She currently has a post-doctoral position in Bordeaux University.
E-mail: melanie.munch@u-bordeaux.fr **Orcid:** 0000-0001-6704-1446

Patrice Buche and Hélène Angellier-Coussy

IATE, U. Montpellier, INRAE, CIRAD, Montpellier SupAgro, Montpellier, France
Patrice Buche holds a doctorate in computer science from Renne 1 university. He is a research engineer and qualified Senior Researcher at the head of the Knowledge Engineering group of JRU IATE.
E-mail: patrice.buche@inrae.fr **Orcid:** 0000-0002-9134-5404
Hélène Angellier-Coussy holds a doctorate in material science from Grenoble 1 University. She is a lecturer in Montpellier University.
Orcid: 0000-0001-5482-7095

Cristina Manfredotti

UMR MIA-Paris, AgroParisTech, INRAE, U. Paris-Saclay, Paris, France
Cristina Manfredotti holds a doctorate in computer science from Milano-Bicocca University. She is a lecturer in the engineering AgroParisTech.
E-mail: cristina.manfredotti@agroparitech.fr **Orcid:** 0000-0003-4217-2591

Pierre-Henri Wuillemin

Sorbonne Universites, UPMC, U. Paris 06, CNRS UMR, LIP6, Paris, France
Pierre-Henri Wuillemin has a doctorate in computer science from Paris 6 University, where he holds a position of lecturer.
E-mail: pierre-henri.wuillemin@lip6.fr **Orcid:** 0000-0003-3691-4886

Abstract: With numerous parameters and criteria to take into account, transformation processes are a challenge to model and reason about. This work can be eased thanks to knowledge graphs, which are a widespread practice for formalizing knowledge associated with structured and specialized vocabulary about a given domain. They allow to draw semantic relations between concepts, and thus offer numerous tools for reasoning over complex queries. Yet, some of these queries in transformation processes might rely on an additional layer hard to transcribe: uncertainty. In this article, we present how knowledge graphs and probabilistic models can benefit each other for reasoning over transformation processes and address the necessity of formalizing contextual expert knowledge for this combination. We then show how this can be used for (1) reverse engineering approaches and (2) linking knowledge bases, through a detailed example on the process of bio-composites for food packaging.

Keywords: Knowledge graph; Probabilistic model; Expert knowledge; Causality; Linked Open Data.

1 Introduction

Knowledge bases allow the formalization of complex domains thanks to their two-folds structures: the ontology proposes a structured vocabulary, described by classes and relations among these; while the knowledge graph, which represents a set of facts, is used to populate this ontology with facts about the represented

domain. This dichotomy between the structure and the data allows a great freedom for knowledge engineers: a same ontology can thus be used for multiple purposes, depending on its genericity and the knowledge base used. This is especially useful when dealing with broad domains such as transformation processes; although their structure can be usually summarized by the same broad concepts (*steps* with *observations* linked together through temporal precedence relations), the diversity of their applications is a challenge to model. Moreover, the facts used to describe them (usually measures resulting from the observations) are subject to variability: some measuring instruments might be more or less accurate, protocols can be changed, ... In this case, reasoning on these different values can quickly become tricky, and complex queries might require to comply with uncertainty. Moreover, in the case of transformation processes, causal knowledge cannot be overlooked, as it gives tools for their better formulation. Yet, causal discovery from data alone is a complicated challenge. To deal with this issue, previous works have presented the combination of probabilistic models with knowledge graphs [1], in order to take into account this uncertainty and to provide tools for causal probabilistic inferences: "If I know that A has an influence over B, then if A takes this value, what is the probability of B taking that value?". However, both knowledge bases and probabilistic models work on different assumptions, making their union sometimes complicated. Indeed, knowledge bases suppose the Open-World Assumption (OWA), meaning that *what is not represented might still exists*. On the contrary, probabilistic models assume the Close-World Assumption: *what is not represented cannot exist*. Thus, learning a probabilistic model presuppose to be able to close the OWA, i.e. being able to distinguish impossible facts ("There is no particle going faster than the light") to missing ones ("The cake's temperature has not been measured due to a default in the thermometer; but the cake does have a temperature"). To do so, in this article we propose a way to introduce and formalize contextual expert knowledge in the case of transformation processes using the Process and Observation Ontology (PO^2), in order to answer complex queries about the represented domain. In this case, we consider contextual expert knowledge as information about the domain not represented in the knowledge base because of its dependence to the context of the query we wish to answer. Both queries and expert knowledge are provided by human expert of the domain: the challenge of our approach is to provide simple tools for integrating this information into the learning of a probabilistic model. We demonstrate that this new source of information allows for complex approaches such as reverse engineering, whose results can be used to enrich knowledge graphs through linked open data (LOD). More generally, we denote this approach as PO^2 ONtology Discovery (POND).

As an illustration, we consider the processing of bio-composites for food packaging. As the massive amount of plastics used each year results in a constant accumulation of wastes in our environment, with harmful effects on our eco-systems and human health [2], innovative technologies are developed for the production of bio-sourced, biodegradable and recyclable materials in order to increase the circularity of plastics. Among bio-polymers, poly(3-hydroxybutyrate-co-3-hydroxyvalerate), called PHBV, is a promising bacterial bio-polymer that is biodegradable in soil and ocean and that can be synthesized from all kinds of carbon residues. One main limitation for its large use at the industrial level is its high cost that still exceeds 5€/kg [3]. To prevent this issue, the development of PHBV bio-composites loaded with lignocellulosic fillers obtained by dry fractionation of organic solid waste and residues is studied: in addition of the decrease in PHBV's cost, it is also motivated by an improvement of the carbon footprint and a reduction of the global warming [4]. This modulates the overall technical performance while giving value to organic residues, thus favoring a cradle-to-cradle concept and promoting the circular economy. The maximum filler content is targeted to decrease the overall cost and the environmental impact of PHBV-based materials. However, the augmentation of added lignocellulosic fibers has a negative impact over the bio-composite's brittleness and its processability [5][6]. It was shown that the stress and strain at break were all the more preserved and the highest possible filler level all the more high that the fiber size was little, due to reduced film heterogeneity and improved wetting of fibers by the polymer [7][5]. The positive effect of reduced filler size could be negatively counterbalanced by the higher energy required to reduce the size of lignocellulosic particles. To reason with this problem, we thus propose to represent the domain thanks to a knowledge base built on an ontology dedicated to transformation processes; and learn a probabilistic model guided by formalized contextual expert knowledge.

Original contributions of this article are (1) the complete integration of the Process and Observation Ontology (PO^2) in a pipeline dedicated to answer causal expert questions; (2) a proposition of conceptual expert knowledge formalization allowing reverse-engineering and LOD approaches; (3) a meta-analysis of different projects connected to the domain of processing bio-composites for food packaging.

Section 2 presents the state of the art necessary for understanding POND: the ontology used, probabilistic models, and the combination of both in a causal discovery context. Section 3 introduces POND and highlights our contributions to the state of the art on the combination of knowledge bases and probabilistic models in the context of expert knowledge integration. Section 4 illustrates POND through the example of biocomposite packaging. We base our study on an innovative knowledge base, composed from different projects.

2 Background

2.1 Knowledge Bases

2.1.1 Definition

While there are many interpretations for the term knowledge base (KB), we adopt here the definition of [8]: a KB is defined as the combination of an ontology structure and a large population expressed in RDF format[1]. More formally, we denote a KB as $KB = (O, F)$ with O an ontology represented in OWL[2] and F a knowledge graph represented in RDF[3].

- O is defined by the set of its classes $\mathcal{C}$ and properties $\mathcal{P}$=$DP \cup OP$, with DP the set of its datatype properties and OP of its object properties. We denote an instantiation i of $\mathcal{I} = \mathcal{C} \cup \mathcal{P}$ as $i \in \mathcal{I}$.
- F is defined by the set of its triples (s, p, o), with $s \in \mathcal{C}$, $p \in \mathcal{P}$. Depending of p, o is either an instantiation of $\mathcal{C}$ (if $p \in OP$) or a literal (if $p \in DP$).

In the rest of this article, we will consider a KB with the Process and Observation Ontology as O, which will allows us to reason about all transformation processes.

2.1.2 The Process and Observation Ontology PO²

PO[2] is a generic ontology dedicated to the representation of transformation processes. Initially thought for food science [9], it has been developed through NeON methodology's scenario 6 [10], by reworking a pre-existant ontology dedicated to the eco-conception of transformation processes [11]. It has been recently used for bio-composite products such as food packaging. Figure 1 presents an overview of its main different parts, described by 67 concepts and 79 relations. A **transformation process** is represented as a succession of **steps**, linked to a **temporal entity** which allows them to be situated in relation to each other. Every step is attached to experimental **results** that can be measured at different scales and units on **components** (which represents factor of interest). PO[2] 2.0 version is implemented in OWL 2 [12], and published on AgroPortal [13] with the public licence Creative Commons Attribution International (CC BY 4.0) [14].

2.2 Probabilistic Models

In this section, we will detail two particular probabilistic graphical models: Bayesian Networks (BNs), and their oriented-object extension, Probabilistic Relational Models (PRMs). We will also give an overview of the Essential graph (EG), common to both, that will allow us to implement causal reasoning.

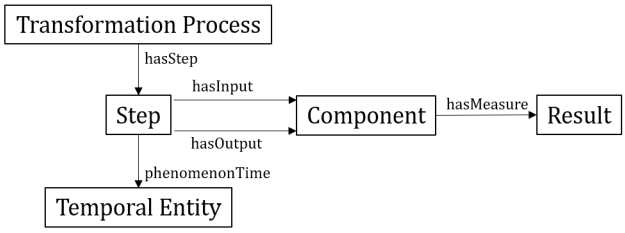

Figure 1 Simplified representation of the main parts of PO[2] ontology. For the sake of simplicity, we will keep it for the rest of our article.

2.2.1 Bayesian Networks

A **Bayesian network** (BN) is the representation of joint probabilities over a set of probabilistic variables encoded as a directed acyclic graph. As such, a BN captures two pieces of information: each arrow represents a probabilistic dependency between two variables, and each variable is defined by its conditional probability table, which describe its different possible values and the corresponding probabilities. Learning a BN is done in two times: first the structure, then the probabilistic dependencies. In our case, it is done using the classical Greedy Hill climbing Algorithm [15] with a BIC score [16]. In this article, we deal with a particular case of BN: the **causal BN** (cBN), which is a BN each relation of whom transcribe a causal relation. An example of cBN is given in Figure 5, while Table 3 shows an example of probabilistic dependencies (in this case, the evolution of the probabilities of the **strain at break** variable in accordance with the **filler content** variable).

2.2.2 Essential Graphs

For each BN, it is possible to deduce its essential graph (EG), a semi-oriented acyclic graph encoding its Markov equivalence class. While BN and EG share the same structure (same nodes and dependencies), some relations are not oriented in the EG. This (un)orientation transcribes the necessity to keep the orientation for preserving the independences encoded in the graph. If a relation is not oriented in the EG, then it can be reversed in the BN without modifying the underlying independences within the variables; on the contrary, if it is oriented, then reversing it would require to learn again the BN's structure. By definition, multiple BNs can have the same EG. Figure 2 illustrates two examples of BNs from the same Markov equivalence class and their associated EG. As we will see further later, the EG is an information source that can be used for causal discovery in a certain context (that we will better define in the following).

2.2.3 Probabilistic Relational Models

PRMs extend BNs with object-oriented notions, allowing to define classes and relations between these. While BNs only require one layer of information to be described, PRMs need two: (1) the relational schema (RS), which

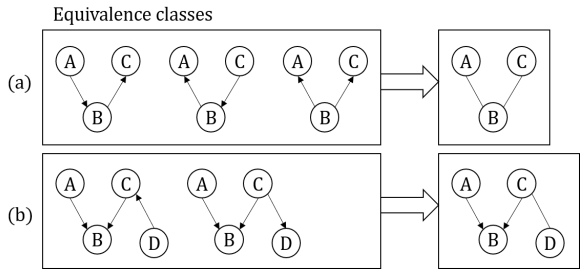

Figure 2 Example of two Markov equivalence classes and their associated essential graph. (a) BN's relations can be reversed in all directions without modifying the independence relations between the variables: the EG is thus completely unoriented. (b) In this example, variables A, B and C share a particular structure that cannot be modified: it is, thus, indicated in the EG with oriented relations.

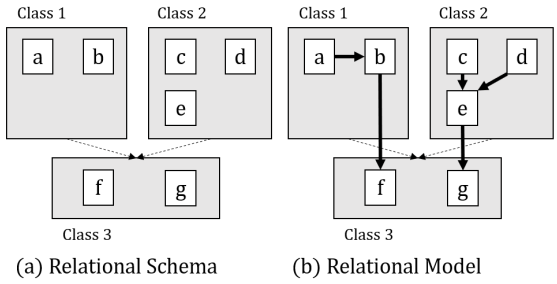

Figure 3 Example of PRM represented with its (a) upper and (b) lower layers of information.

proposes a qualitative description of its classes and their associated variables (denoted here as attributes); and (2) the relational model (RM), which relates all quantitative information about the attributes' probabilistic relations. Two classes in the RS can be linked with relational links which orientate the direction of probabilistic relations between attributes: for instance, the link from **Class 1** toward **Class 3** in the RS (Fig.3 (a)) forces the orientation between the b and f attributes in the RM (Fig.3 (b)). On another hand, relations are not oriented within a class. Once both the RS and RM have been defined, a PRM can be instantiated in model similar to a BN. Learning a PRM from scratch can be a hard task, as it requires two learning: a first for the RS, then one for the RM. In our case, we choose to manually construct the RS, in order to reduce the learning complexity to that of a BN [17]. As a consequence, the learning of the relations between the variables is oriented by the constraints we put in the RS: this allows us to learn under constraints to introduce external knowledge.

2.2.4 Learning under constraints

Learning under constraints in the case of BNs allows to greatly enhance the precision of the learned model, both for structure [18] or parameters learning [19]. This is all the more verified in the case of small databases[4] [20]. To this purpose, previous work have proposed to force a complete [21] or partial [22] node ordering in order to introduce directional constraints during the learning. In our case, the RS manual definition allows to replicate this partial node ordering while also enabling the introduction of contextual expert knowledge given by both the KB and the expert, which creates a favorable context for causal knowledge discovery [23]. It is important to note that, in our case, we consider this new knowledge as causal, allowing to learn our model under *causal constraints*.

2.3 Knowledge Discovery

2.3.1 Causal Discovery

Correlation is not causation: when aiming for causal discovery, a model must answer some specific criteria in order to avoid inferring false facts. Among them, **causal sufficiency** is a key factor, as it guarantees the absence of the influence from factors not represented in the model, which could hinder its predictions [24]. As an illustration, one might consider the fact that a model representing both a person's *shoes size* and *reading ability* will find a strong correlation between the two, despite the lack of a real causal relation between these two variables. This is due to the fact that this model is missing a third variable, which has an influence over the first two: the person's *age*. Another important criteria is the dataset quality: in the case of missing or biased data, or in presence of deterministic cases, causal discovery is not possible [25]. To continue our last example, if we consider a biased dataset collected among child geniuses, we would find links between *age* and *reading abilities* which would not be representative for the general population. In a more general setting, this representativity is a key point of causal discovery. Discovering causal knowledge from data can only be done with independence tests between the different variables [24, 26]. Other works have proposed the use of GE for learning causal models: [27] presents two optimal strategies for suggesting interventions in order to learn causal models; [28] and [29] use the GE to suggest a limited number of interventions in order to build a cBN. However, all of these works do not integrate the addition of external contextual knowledge. In order to achieve our goal, we have chosen to look into the combination of such models with knowledge bases.

2.3.2 Discovery with knowledge bases

As introduced in the previous section, causal knowledge from data alone is a tedious task. For this reason, several works have undertaken the task of combining ontologies with probabilistic models in order to discover new relations. For instance, multiple ontologies extensions allow to directly integrate probabilistic reasoning (such as BayesOWL [30, 31], or HyProb-Ontology [32]). If these extensions allow to reason with probabilistic relations, they do not allow to learn these relations. Other works use the ontology's structure to build a BN, by translating the object properties as probabilistic

[33] or causal [34] relations; this preconception cannot be applied to PO^2, where numerous relations do not encode causal dependencies. Other methods, finally, have been developed to answer precise cases, which cannot be applied to our approach: for instance, [35] uses predefined models to undertake medical diagnostics, which cannot be extended to other medical applications than the one they are presenting. On this note, it is important to note that while POND is based on a single ontology, its complexity allows to address many different applications cases. As long as a domain can be represented by a transformation process within the ontology's structure, then we can apply our approach.

3 POND: PO² Ontology Discovery

In this section we present POND, whose goal is the formalization and integration of contextual expert knowledge for the learning of a probabilistic model in order to reason on transformation processes. In this section, we will present the different sources of knowledge and how they can be used to design and answer complex probabilistic and causal queries. Since next section will give an application with a reverse engineering approach, we will give, in this part, a particular focus on causal discovery.

3.1 Knowledge Integration

3.1.1 Expression

Expert knowledge can come from: (1) experimental data, gathered from different sources (such as publications, books, or productions from different projects); (2) direct interviews with experts of the considered domain. Most of these information can be directly structured by PO^2: this concerns factual and descriptive facts (which are the different steps, which observations are done, what are the measured values). This integration of knowledge within the KB serves two goals:

- it builds a complete and coherent thesaurus, which will be used as a common ground when exchanging with the expert during the contextual formalization part;
- it helps define the potential future variables that would be integrated in the model for its learning.

This last point is a key point: since probabilistic models' learning is based on statistical tools, in order to learn one we need to be able to build a learning database composed of concrete values (such as numbers, strings, ...). In an ontology, those are usually given by datatype properties which, in our case, are themselves fed by the data gathered at the beginning in POND. As a consequence, this part helps us elicit the future variables on which we will be able to reason.

From this newly defined KB, the expert can express **Expert Questions** that we categorize in two parts. Some stays at a descriptive level and can be answered by directly querying the ontology (through SPARQL queries, for instance). We denote them as **Competency Questions** (CQ). Others are based on probabilistic reasoning and require the learning of a model (in our case, a BN). These are denoted as **Knowledge Questions** (KQ). In this article, we will focus on how POND is equipped to answer these and, more especially, how it can be used to answer causal KQs (cKQs). Given X_i and X_j two groups of variables in the domain, we formalize a cKQ as:

cKQ_1 , does X_i has an influence over X_j? (*i.e.*, does changing the values of X_i will impact the values of X_j).

cKQ_2 , what is the impact of X_i over X_j? (*i.e.*, how the values of X_j evolve along those of X_i).

These two questions illustrate the double reading given by the cBNs: while cKQ_1 focuses on the descriptive aspect of the learned relations (which can be deduced directly from the graph), cKQ_2 rather questions their nature (which can be analysed with the conditional probability tables).

3.1.2 Integration

Once the cKQ defined, the model can be learned. As described in Section 2.2, our goal here is to transcribe the contextual expert knowledge that couldn't be directly described by PO^2 (such as causal relations "A has an influence over B") into a RS for guiding the learning of a PRM. The originality of POND is how this integration is dealt with:

1. Through the mapping of the variables from the ontology toward the model.
2. With the definition of the precedence constraints.

Mapping of the variables. As mentioned in the introduction, the trickiest part of combining ontologies with probabilistic models is that it requires to close the OWA, *i.e.* defining which pieces of information are relevant, which are missing and which can be compared. The first point requires the expert to define, within the domain, which variables they want to see represented in the model. For instance, if one has a KQ about a cooking recipe, the weather outside would probably not be relevant (unless the atmospheric humidity has to be taken into account!). More generally, not all variables represented in the KB have to be represented in the model and only expert knowledge can help us distinguish important from irrelevant ones. On the opposite, a missing variable might compromise the learning of the model: having no information about the quantity of ingredients would definitively be a hindrance to the modeling of cooking recipe. Once again, only an expert can evaluate if a KQ on their domain can be answered with the available variables. Once the necessary variables

Process	Step	Component	Result	**Variable**
$recipe_1$	*cooking*	*oven*	*temperature*	**CT**
$recipe_2$	*cooking*	*oven*	*temperature*	**CT**

Table 1 Template defined within POND for describing the variables to be integrated into the learning database. In this table, we consider our cooking example: we have two different recipes 1 and 2, for which we have measured the oven's temperature during the cooking. We represent all of their attributes' values as a single variable, denoted **CT** (Cooking Temperature).

```
Select ?value
Where {
<Process> po2:hasStep <Step>.
<Step> po2:hasInput/po2:hasOutput <Component>.
<Component> po2:hasMeasure <Result>.
<Result> po2:hasValue ?value.
}
```

Figure 4 Template used for the automatic querying of the variables values. <Process>, <Step>, <Component> and <Result> are given by Tab.1.

have been defined, the expert has to define how to extract their value in order to constitute the learning database. To do so, we define the template presented in Tab.1, for which the expert indicates for each variable they wish to represent, which values they want to associate with them. From there, a template (Fig.4) is automatically filled for generating queries which are used to build a learning dataset. In this way, the expert can easily define what should be used and which values can be compared in order to define the different variables. The strength of this approach is that every interaction is done using specific vocabulary defined beforehand in accordance with the expert: they do not have to write the SPARQL queries themselves.

Defining Precedence Constraints. Once the variables are all defined, we can start building the RS. To do so, we first define a common class in which we put all the variables. Then, the expert can express contextual causal knowledge about the different variables in order to create more classes. Given X_i and X_j two groups of variables, these pieces of information can be formalized as:

- "X_i can have an influence over X_j, but not the contrary". In this case, two classes are created in the RS, such that (1) X_i and X_j are separated in a class each and (2) a relational slot is traced between the class that owns X_i toward the class that owns X_j.
- "X_i and X_j are independent". In this case, two classes are created, such that (1) X_i and X_j are separated in a class each, and (2) no relational slot is traced between the two classes.
- "The relation between X_i and X_j is unknown". In this case, X_i and X_j are put in the same class.

To be noted, some of this causal knowledge can be automatically deduced from the KB: for instance, in the case of temporality, we can easily infer that variables that are measured at time t might have an influence over measures made at time $t + n$, but not the contrary.

Our contribution in this part is the formalization of the expert knowledge and its integration within the learning: thanks to the definition of PO^2, every represented transformation process can be easily integrated into a SR through the definition of a common vocabulary with the expert. From there, a learning database can be automatically extracted and then used for learning the probabilistic relations of the RM.

3.1.3 Causal Validation

Once the RS built with the expert, the RM can be learned; it then becomes possible to instantiate the PRM (defined with the RS and the RM) in a BN. In our case, we consider that the integrated expert knowledge allows to learn the model under causal constraints, thus favoring a context for causal discovery [1]. This is due to the fact that we consider this model as learned at the intersection of two sets of models: (1) every model whose relations' orientation reflects the database's inner constraints (expressed, as we have seen in Sect.2, by the EG); and (2) every model that respects the expert's causal constraints (defined in the previous section by the RS). Once again, it is important to recall that we consider that all the following deductions are made under the assumption that we are indeed in a context favorable to causal discovery (as detailed in Sect.2.3.1). If so, we can now proceed to the causal validation of the model with the expert, which consists in looking at each relation:

- If a relation is learned between two variables between which an expert precedence constraint has been drawn (*i.e.*, the two variables are in different classes with a relational slot between them), then the causality of the relation is validated by the expert knowledge.
- If a relation is learned and is also oriented in the EG, then causality is validated by the data. This deduction is based on the following reasoning: if the learning database can be trusted, then this relation could not be oriented in another way without disrespecting the inner constraints of this database.

To be noted, these points are independent: a relation is causally validated as long as at least one of these points is verified. In the case where a relation does not fall into one of these categories, it is impossible to deduce its causality. An application of this validation protocol is given as an example in Sect. 4.2.2.

Ideally, this causal discovery has for goal to causally validate all relations from the probabilistic model, thus allowing to define a cBN. However, even if not all relations are validated, it offers a solid ground for:

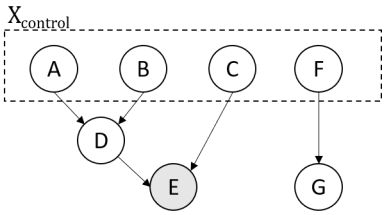

Figure 5 Example of a short cBN. The $X_{control}$ set represents all control variables, *i.e.* variables on which the expert can intervene. E represents the target variable.

- **Help the expert criticize the model.** Since we aim to model real domains, direct evaluation to validate every relation are often hard (even impossible) to carry out directly (too expensive, time consuming, ...). By presenting the learned potentially causal relations to the expert, we give them a tool for questioning them from their own knowledge, and even formulate new hypotheses that could be experimentally verified (as presented in the next section).

- **Answering cKQs.** The resolution of a cKQ requires a causal reasoning: cKQ_1 will look at the presence (or absence) of a relation between the concerned variables; while cKQ_2 will use these relations to reason on their conditional probability tables. If the needed relations are not causally validated, it is impossible to answer these cKQs.

It is important to note that in case of non-validation, several solutions are offered to the expert: they can provide more expert knowledge (as new data to enrich the database, or causal precedence constraints to refine the RS); new experiments can be suggested, in order to fill identified lack of knowledge; of the KQ can be redesigned. If, however, none of these solutions are possible, the KB is judged inadequate to answer the given KQ.

3.2 Causal Inferences

Until now, we have seen how to integrate expert knowledge in order to learn a model and how to validate it. If this would be generally adequate for answering questions such as cKQ_1 (by checking the presence/absence of causally validated relations), cKQ_2 requires a more in-depth analysis. As an illustration, we consider the BN presented in Fig.5 as causally validated and the following cKQ_{ex}: "Which intervention should I do to maximize the value of E?", which is a combination of a cKQ_1 ("Which variables have an impact over E?") and a cKQ_2 ("What is the influence of these variables over E?").

In order to answer cKQ_{ex}, we first have to identify the variables on which it is possible to intervene. Denoted as **control variables**, they are characterized by the fact that they can be controlled. For instance, the quantity of ingredients in a recipe are control variables, as we can directly and easily change their values; on another hand, the texture of a mixture cannot be changed: we may have to intervene on the quantity of the ingredients that compose it, or the way it is blended, but cannot modify it directly. Control variables can depend of the context of the question and thus have to be discussed with the expert to see what it is possible. In Fig.5, we consider that A, B, C and F are control variables (denoted as $X_{control}$), while D is not: even if it has an influence over E, it cannot be used in our reasoning. Moreover, by looking at the graph, we can see that while F is a control variable, it has no influence whatsoever over E. Thus, we define $X_{inter} = \{A, B, C\}$ the set of variables which can answer cKQ_{ex}: since we are considering a cBN, modifying this set will have an influence over E. In practice, this means that for each combination of the variables' values, E's values and probability's distribution are affected: each part constitutes a potential scenario that we have to evaluate in order to determine which one will answer cKQ_{ex} the better. In order to do so, the expert is asked to define their own *criteria of acceptability*, *i.e.* the goal they wish to attain, such as "Which values are best for the target variable?", or "Which conditions should apply on X_{inter}?". More formally, these criteria are distinguished in two sorts:

- **Hard Criteria.** Some values or combinations of values are impossible to obtain: corresponding scenarios have, then, to be removed. For instance, the expert might wish that the sum of all variables' values does not reach a certain point; or they might want to exclude some values of E. In our case, since we wish to maximize E, we exclude its lower values.

- **Soft Criteria.** In some cases, the expert has to sort their preferences depending on the context. It could be that a high value of E is not that interesting if A is also high; or that a lower value of E with a higher probability of realisation is a more interesting scenario than a high value that has no chance to occur.

Defining these criteria allows to better define the expert's needs and, thus, helps to find the best answer to QCc_{ex}. Sec.4.3 shows how this approach can be concretely applied to a reverse-engineering approach.

3.3 Enrichment through LOD

As introduced previously, the semantization of the data allows the use of LOD approaches to enrich our approach. While LOD under uncertainty has already been tackled in other works, it mostly focuses on dealing with uncertain graphs [36] or data [37]. In our case, the uncertainty comes from our query: the alignments we need to do depend on both the KQs and the learned model. In particular, once the model learned, we can consider that all attributes defined within POND using

Tab.1 can also be extended using other KBs, thus providing more data to reason with. An example of this approach is given in Sect.4.3.3, where a KB different from PO² is used to query for new prospective data.

Given a CQ defined by the expert, we formally extend POND with a set of triples $< KB^i$, CQ^i, mapping$^i >$, with

- KB^i a SPARQL endpoint;
- CQ^i the SPARQL query associated to CQ corresponding to the scheme of KB^i;
- mapping^i a set of *exact-match* alignments between the concepts of CQ and CQ^i;

This definition is based on the assumption that it is possible to get for each KB^i a template similar to the one presented in Fig.4, with which a query can be defined to gather all relevant values. In the case of POND, this query is automatically deduced from Tab.1; here, it must be formulated manually. As such, while this approach could in theory be used to enrich the learning dataset, its scalability when considering numerous attributes can be discussed. On another hand, once the model learned, the **control variables** defined in Sect.3.2 becomes potential candidates: in this case, querying new KBs allows to predict new results from them, which are not present in the original KB. In short, this approach should be preferred when enriching the original dataset with incomplete new data, in order to predict the missing information or, as we will illustrate in the next section, provide new data for the expert to reason with.

4 Application to Food Composite Packaging

As presented in the introduction, our application aims to define the formulation of lignocellulosic-based biocomposites based on a reverse engineering approach, i.e. based on targeted specifications including functional properties, but also environmental impact and economic cost. The solution we present is based on the combination of ligno-cellulosic biomass and a polymer matrix (PHBV). Lignocellulosic fillers (LFs) are produced by dry fractionation of raw lignocellulosic biomasses. They are available in the form of powders (also called particles) whose granulometry and biochemical composition depend on both the nature of the raw biomass and the fractionation itinerary. The formulation means the choice of the filler biochemical composition, size and weight content, all these three parameters being demonstrated to have a strong impact on the performance of the final product. In this section, we will present how POND can help to study this problem.

4.1 KB presentation

Data was collected from five different projects dedicated to the development of biocomposites constituted of PHBV, as the polymer matrix, and ligno-cellulosic fillers stemming from organic solid wastes. Nine different lignocellulosic biomass are considered : vine shoots, i.e. an agricultural residue of viticulture, corresponding to the pruning wood (*H2020 NoAW*); olive pomace, i.e. the residue of olive oil manufacturing, corresponding to a mixture of residual skin, pulp and fragments of the crushed stone and pine bark (*Chercheur d'avenir région Languedoc-Roussillon MALICE*); wheat straw, i.e. the by-product of wheat culture (*FP7 EcoBioCAP*); five fractions of lignocellulosic fillers derived from urban parks and gardens green waste: a branches-rich fraction, a grasses-rich fraction, a leaves-rich fraction, and two fractions constituted of a mixture of constituents (*H2020 Resurbis*). In order to define a ground for comparison, pure cellulose fibers as well as wood fibers have been considered in the frame of the H2020 Usable, H2020 NoAW and H2020 RESURBIS projects. This constitutes in the end a database of 88 formulations described by 15 different attributes [38].

4.2 Expert Knowledge Integration

Expert knowledge integration is done in two times: (1) the mapping of interesting attributes from the KB to the RS, and (2) the definition of the potential precedence constraints. In this section, we present the principal results used for learning the final result, as well as an example of an expert critic.

4.2.1 RS Definition

Attributes selection. [5] The expert describes a LF with two main categories of attributes: biochemical composition with the plant's main organic constituents (**cellulose**, **hemicellulose**, **lignin**) and inorganic content (**ash**); and apparent median diameter (**d50**), which is the main parameter describing the granulometry of a powder. It is worth noting that this list of attributes is not exhaustive. Additional attributes regarding the granulometry and the shape of particles could have been added. However, since such characteristics are not systematically assessed in the projects, neither in literature, they were not considered in the present case study. A third attribute is associated to LFs and describes the formulation : the **weight filler content**, which indicates the quantity of filler (*i.e.*, the CL) that was added to the compounds. Final composite materials are described by four groups of attributes: tensile parameters (**stress at break**, **strain at break** and **Young's modulus**); **permeability** (to water vapour); thermal properties (**crystallization** and **melting** temperatures); thermal degradation (**onset** and **peak** temperatures). The neat polymer matrix (i.e. without the addition of LFs) is characterized by the same attributes. For each composite, data are normalized to those of the corresponding matrix, processed exactly in the same conditions as the composites. Fig.6 presents a summary of this categories and how a formulation is represented in our model.

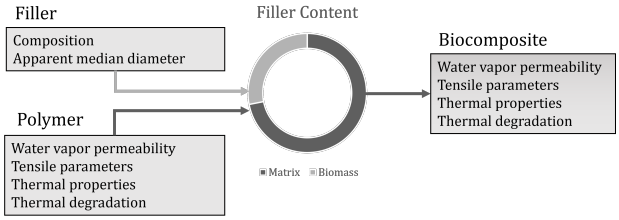

Figure 6 Summary of the main groups of attributes for each formulation.

Knowledge Question Definition. Considering now our application, we would like to know how a given filler (characterized by its composition and its apparent median diameter) and its weight content within the material influence the tensile parameters of the final biocomposites. It is worth to remind that our objective is to achieve the best balance between the highest filler content (allowing to decrease the cost and the environmental impact of the PHBV-based materials) and the smallest decrease in mechanical properties. Be V the group of all attributes and V_i the group of all attributes from the group i, we formally define cKQ_{bio}: "Which characteristics allow to optimize the tensile parameters of the biocomposite?" as the combination of two cKQ:

cKQ_{1b} "Does x has an influence over y?", with $x \in V$ $V_{tensile}$ and $y \in V_{tensile}$.

cKQ_{2b} "What is the causal impact of x over y?"

Answering cKQ_{bio} will, thus, be done in two times: first answering cKQ_1b to find the set of variables x that have an influence over the tensile parameters in the model; and then cKQ_2b to quantify the impact of these variables and find the most optimal combination.

What we learn will then allow us to answer the question "Which characteristics for the LFs and which filler content for achieving the smallest reduction of tensile parameters and the highest filler content?". Among all attributes represented in our KB, only those concerning the filler and the tensile parameters would be relevant to answer this question. However, in a context of causal discovery and in order to ease the expert critics, all attributes are included for the learning.

Attributes Discretization. Most of the considered attributes are continuous (*i.e.* they cannot be automatically sorted into distinct discrete categories). However, by definition, classical BNs cannot be learned from this kind of data: thus it is necessary to go through the discretization of the variables. This step is important as it can influence the different relations learning and thus change the model's interpretation. Two ways for discretizing a variable were considered in the present study:

- Defining n quantiles (*i.e.*, the sets of values were separated such that each part had the same number of individuals;
- Asking the expert for a discretization.

Lignin content (%)	**FC** (%)	**SB** (no unit)
[0;19] *(32)*	[2;4] *(10)*	[0.2;0.5] *(19)*
]19.4;26.4] *(30)*	]4;11] *(34)*	]0.5;0.8] *(44)*
]26.4;49] *(23)*	]11;21] *(22)*	]0.8;0.9] *(9)*
	]21;50] *(19)*	]0.9;1.07] *(9)*

Table 2 Example of the discretization used in our application for the **Lignin content**, the **Filler Content** (FC) and the ratio between the strain at break value of the composite and the strain at break value of the corresponding PHBV matrix (SB). *(Number of data)*

While the first method is easier and less prone to errors (since categories are not over-represented compared to each other), it can sometimes have no sense in regard to the model we wish to learn. In our case, we aim to learn whether the final characteristics of the biocomposite are impacted by the filler or not: if the value of the composite attribute is < 1, this means that this attribute has been degraded by the addition of LFs, as compared to the neat matrix; if it is > 1, this means that the attribute has been improved. Considering for example an interval of [0.6;1.2] would not be a useful discretization, as we could not assess the impact of the filler over the attribute. That is why asking experts for inputs is important in order to determine the thresholds and patterns we would like to observe. In our case, the two methods are applied. While control variables (i.e. variables concerning the fillers: composition and apparent median diameter) are discretized in 3 quantiles, target variables (i.e. tensile parameters, water vapour permeability, thermal properties and thermal degradation temperatures, as well as weight filler content) are discretized with the expert's input, which reflects the increase or decrease of the biocomposite's attributes values. An example of this discretization is detailed in Tab.2.

Precedence Constraints Definition. The expert first defines two global precedence constraints that will be precised in the following:

- Between the LF's attributes and the biocomposite's ones. This is due to the fact that the first are considered as control variables (*i.e.*, the expert has an influence over them), while the second are considered as a consequence of the first. We thus define two classes in the RS, with a relational link from the class containing all LF's attributes toward the class containing all the biocomposite's ones.
- Between the different categories of attributes. This distinction allows to take into consideration each category as a sub-group independent from the others (*e.g.*, tensile properties have no impact over the thermal properties). To model this, the class that owns all of the LF's attributes is subdivided in new sub-classes with no relational link between them.

4.2.2 Expert Critic

Once the attributes and their precedence constraints defined, a first model is learned (Fig.7). Expert critic can then be gathered in two times by using the causal validation method described in Sect.3.1.3:

- Since the sum of the CLs' constituents is 100, it is coherent to learn relations between them (*e.g.* between **Cellulose** and **Ash**). Yet, those bear no sense from a causal point of view: they should be independent, as we cannot say that "The value of **Cellulose** causes the value of the **Ash**". Thus, we create four distinct sub-classes, one for each constituent, and compartmentalize them. This modeling choice will lead to the definition of an hard constraint (that will be further explained in Sect.4.3) that guarantees a CL is technically possible (*i.e* the sum of its constituents is not higher that 100).
- The **crystallization** temperature cannot be explained by the **melting point**: the relation learned between the two reflects a correlation, not a causation. In order to address this, the two are put in distinct sub-classes.
- The **peak** can explain the **onset** temperature, but not the contrary. The two attributes are separated and attributed to a class each between which a relational link is set.
- Against all expectations, the **strain at break** is not explained by any attribute. A new discretization is tested in order to better represent the attribute: we change the one represented in Tab.2 so that we now distinguish the cases]0.8;1] (15 examples) and]1;1.07] (3 examples).
- It is also unexpected that the filler content has no impact on neither the mechanical properties (strain at break, stress at break, Young's modulus) nor the water vapor permeability.
- Finally, the learned model shows no relation between the size of the lignocellulosic fillers (characterized by the apparent median diameter d50) and the mechanical properties, which is very unexpected.

Moreover, expert critic allowed to highlight knowledge holes (*i.e.*, cases not represented in the KB that could lead to incomplete models). Indeed, when looking at the conditional probability tables in more details, the learned model shows that when the **filler content** $\in$]21;50], then the **melting** temperature $\notin$]1;1.02]. This could be due for two reasons: (1) it is indeed a gap in knowledge and the KB needs to be completed; (2) it is an expected result that does not require supplementary data. As explained in our introduction, this is a typical case where, in order to close the OWA, we need an expert input to determine which case we are considering.

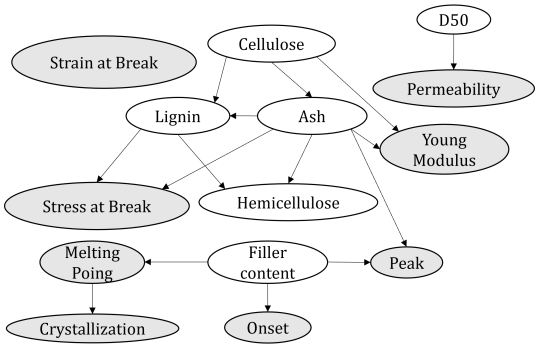

Figure 7 Model learned after one iteration and criticized by the expert as explained in Sect.4.2.2. Control and target variables are respectively depicted in white and grey.

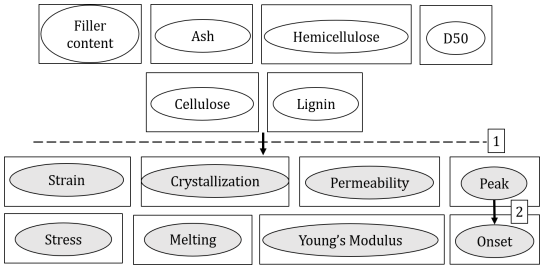

Figure 8 Relational Schema defined after expert critic. Each rectangle represents a class in the RS, while each ellipse represents an attribute. For the sake of readability, relational links are indicated in two ways: (1) indicates that all classes above the dashed line have a relational link toward those under the line; (2) indicates that a link is established from the **peak** toward the **onset** temperature.

In this example, the expert had indeed confirmed that the lack of increase of the **melting** temperature was coherent with a high **filler content**. This is explained by the fact that the **melting** temperature should only decrease due to the thermic degradation linked to the LFs introduction and thus should not increase with the **filler content**. Fig.8 presents the final RS that we used to learn the model that we will consider in the rest of this article.

4.3 Answering the Knowledge Question

We now consider a validated cBN learned from the KB and the RS presented in Fig.8. From this model, we first answer cKQ_{1b} by looking at the attributes that have a direct relation toward the tensile parameters. Those are represented in Fig.9: the **filler content** and three compositional parameters, the **ash**, **cellulose** and **lignin**. In order to answer cKQ_{2b} and then cKQ_{bio}, we need to look at the conditional property tables. However, we can see that not all tensile parameters are explained by the same attributes: depending on the parameter, answering the cKQ requires one of the two following interventions: (1) the **filler content** and (2) the CL's **composition**.

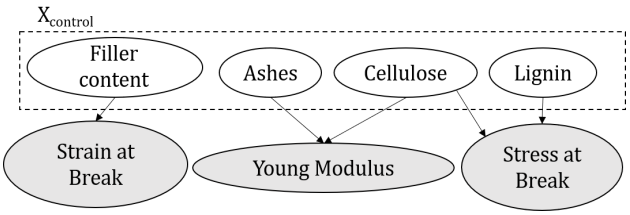

Figure 9 Extract of the final validated BN used for answering cKQ_{bio}. Since all relations are influenced by the expert precedence constraints we consider this model as *causal*.

	Strain at Break			
Filler	]0.24;0.5]	]0.5;0.8]	]0.8;1]	]1;1.07]
]2;4]	0.0076	0.4924	***0.4924***	0.0076
]4;11]	0.002	***0.770***	0.1620	***0.0660***
]11;21]	0.3624	0.4522	0.1826	0.0028
]21:50]	***0.5747***	0.2630	0.1071	0.0552

Table 3 Conditional probability table for the **strain at break**. (***Maximum Likelihood***)

4.3.1 Optimal Filler Content

According to Fig.9, the **filler content** has an impact over the mechanical properties through the **strain at break**. This is coherent with the problem described in the introduction: since the PHBV main limitation is its high cost, introducing low-cost LFs could help tackle this issue by filling the polymer as much as possible. However, this introduction has consequences on the polymer's brittleness, as illustrated in our model. From the conditional property table of the variable (Tab.3), multiple readings and answers are possible depending on the expert's criteria of acceptability (as described in Sect.3.2):

- If we are aiming for the highest possible value of the **strain at break** (]1;1.07]), probabilities are all almost zero. This criteria is, thus, not possibly considerable.

- If we aim to obtain the second best value for the **strain at break** (between 0.8 and 1), a **filler content** $\in$]2;4] can be considered: it guarantees a probability of success of 0.49, meaning that on average half of the final products would reach the desired value (and would have a value between 0.5 and 0.8 on the other cases).

- When considering an industrial process, stability in the result would be something the expert would want to pay attention to. In this case, a criteria of acceptability can be placed not upon the value, but on the probability of success, in order to guarantee this stability. In this case, it should be better to consider a **filler content** $\in$]4;11], which guarantees a **strain at break** $\in$]0.5;0.8] with a probability of 0.77.

Biomass	P(HC_1)
Wheat straw	0.0015
Olive pomace	0.1
Pine bark	0.12
Vine Shoot	0.4
Cellulose	0.47
Wood flour	0.51
Urban parks and green residues	0.83

Table 4 Biomasses present in our KB and their probability of respecting HC_1: "The **Young's modulus** and **stress at break** have values over 0.8".

4.3.2 Optimal LF composition

According to the cBN presented in Fig.9, the **Young's modulus** and **stress at break** depend on the composition of the considered LF (more specifically, its **ash**, **cellulose** and **lignin** content). In this section, we will use this information to find the most optimal biomass among the one tested during the different projects. To do so, we first need to define the different criteria of acceptability.

Criteria of acceptability definition. Similarly to the previous section, we first define one hard criteria of acceptability in accordance with the expert's expectation:

HC_1 We want our target variables to have optimal values: we fix **Young's modulus** $> 0.8 \cap$ **Stress at break** > 0.8.

Biomass Evaluation. From the learned model, we simulate experiments using the biomasses present in our KB. To do so, we define a Competency Question (CQ) CQ_{b1}: "What are the biomasses represented in the KB and what is their composition in **Lignin**, **Ash** and **Cellulose**?", which can be answered by a SPARQL query as defined in Sect.3.1.1. Then, for each profile of composition, we evaluate the probability of reaching HC_1. Results are presented in Tab.4. As we can see, the best ones are obtained when using urban parks and green residues, while other LFs (such as the pine bark) have a clear negative impact.

Discussion. These first results have to be considered within the framework of our learned model. As we have seen, relations between attributes are dependant of the values represented in the KBs. Thus, classifications similar to the one presented in Tab.4 allows to see trends, but can also be erroneous. In particular, expert evaluation highlighted the fact that the **filler content** and the **apparent median diameter** should have an impact over the characteristics, which is not illustrated in these results.

4.3.3 Discovering new LFs

In the previous section, we have evaluated LFs that had already been tested in order to find the most suitable

one. In this section, we will extend these results by suggesting new LFs that have not been tested and see if the previous results could be improved. To do so, we define CQ_{b2}: "Which LFs have for characteristics the ones returned by cKQ_{bio}?". On the contrary to CQ_{b1}, the SPARQL query will not be run in PO^2, but in another KB in order to test potential LFs. As such, the first issue to tackle is the alignment of the two different KBs. Similarly to the expert critic section, we use here knowledge engineering tools to offer new perspectives on the represented domain.

Ontology Alignment. For the following, we consider a KB structured by the @Web ontology [39], which is dedicated to the representation of n-ary relations. The associated knowledge graph contains information about LFs; however, since the structure is not the same as PO^2, we first need to align its vocabulary on the one used in our project, using the mapping defined in Sec.3.3. In our case, we need to map the control variables (**Ashes**, **Cellulose** and **Lignin**): for each, we define a SPARQL query to retrieve the information using the @Web's structure. This allows to draw correspondences between the different concepts through the matching of (1) the PO^2 query generated from a template and (2) the @Web query. An example of the queries for the retrieval of the **Ash** value is given in Fig.10.

Defining new Acceptability Criteria. In order to query for novel LFs, we first need to define new criteria of acceptability to filter the potentially interesting ones:

HC_2 In order to propose realistic LFs, it is important that the sum of its constituents does not exceed 100 (in other terms, the simulated biomass must be physically possible). By fixing $x \in \{$**Ash**, **Cellulose**, **Lignin**$\}$ and their associated interval $[x_{min}; x_{max}]$ determined by the discretization; we fix HC_1 such that $\sum_x x_{min} < 100$.

HC_3 The probability of realization must be higher than 0.25. This value was chosen arbitrarily in order to elect LFs whose probability of success are not too low.

SC_1 In the case where no LF is found, we extend our research range to similar potential candidates, *i.e.* candidates whose composition is very close to the recommended composition. In order to evaluate such substitutes, we define a quality's score S_m. Be a potential LF m, its composition x_m (with $x \in \{$**Ash**, **Cellulose**, **Lignin**$\}$) and the target interval $[x_{min}; x_{max}]$ (*i.e.*, the target interval recommended by the model), we have $S_m = \sum_x \sigma(m, x)$ with $\sigma(m, x) = min(abs(x_m - x_{min}), abs(x_m - x_{max}))$. To be noted, $\sigma(m, x) = 0$ if $x_m \in [x_{min}; x_{max}]$.

The lower S_m is, the closer the proposed LF is to the recommendation.

```
Select ?ash_min ?ash_max ?ash_u
Where {
# Get the biomass
?it rdf:type ?process.
?it core:hasForStep ?step.
?step core:hasOutput ?lsf.
?lsf core:isComposedOf ?compo.

# Get the ash value
?obs_compo sosa:hasFeatureOfInterest ?lsf.
?obs_compo core:observationResult ?ash.
?ash ssn:hasProperty ?ash_ppt.
?ash_ppt rdf:type domain_att:Ash_rate.
?ash schema:minValue ?ash_min.
?ash schema:maxValue ?ash_max.
?ash schema:unitText ?ash_unit.
}
```

(a) PO^2 Query

```
Select ?ash_min ?ash_max ?ash_u
Where {
# Get the biomass
?biomass rdf:type atweb:biomass_carac_relation.
?biomass atweb-core:hasResultConcept ?Result.

# Get the ash value
?biomass atweb-core:hasAccessConcept ?Ash.
?Ash rdf:type atweb:ash_rate .
?Ash atweb-data:hasForFS ?fuzzySet.
?fuzzySet atweb-data:hasForFuzzyElement ?fuzzyEl.
?fuzzyEl atweb-data:hasForMinKernel ?ash_min .
?fuzzyEl atweb-data:hasForMaxKernel ?ash_max.
?fuzzySet atweb-data:hasForUnit ?ash_u.
}
```

(b) @Web Query

Figure 10 Comparison of two queries within (a) PO^2 and (b) @Web. For the sake of simplicity, we only consider a simple query for retrieving the **Ash** value. In the actual experiment, queries where defined to retrieve all relevant compositional parameters in one iteration.

Electing new LFs. From the criteria of acceptability defined, we can now generate scenarios using the learned model. This is done is 4 steps:

1. Defining a BN that integrates HC_1 and HC_2.
2. Testing all combinations of biomass and assessing the probability p of respecting HC_1 and HC_2.
3. If $p > 0.25$ (as defined in HC_3), then a query is made over @Web to find the potential CLs that match the scenario.
4. If no CL is found, then we compute for each new potential CL the S_m score.

Step 1. allows to directly integrate our constraints into the BN by defining two new deterministic variables, as illustrated in Fig.11. They take two values: *True* if their condition is respected and *False* otherwise. An example of HC_1 conditional probability table is given in Tab.5: when both the values of the **Young's modulus** and the **Stress at Break** are above 0.8, then the variable takes the value True. From there, we can simulate all possible scenarios and easily check whether these two constraints are respected by computing $P(HC_1|HC_2)$ (*i.e.*, the probability of verifying HC_1 knowing that HC_2 is met). If so, we apply HC_3 to

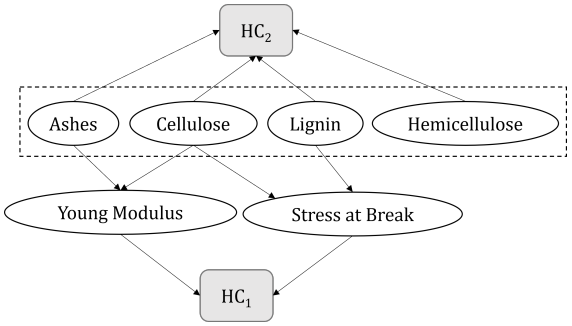

Figure 11 In order to integrate the hard constraints HC_1 and HC_2 into our reasoning, we translate them as new variables in our cBN.

Young's modulus	**Stress at break**	HC_1 True	False
]0.2;0.8]	]0;0.8]	0	1
]0.2;0.8]	]0.8;1]	0	1
]0.2;0.8]	]1;1.4]	0	1
]0.8;1]	]0;0.8]	0	1
]0.8;1]	**]0.8;1]**	**1**	**0**
]0.8;1]	**]1;1.4]**	**1**	**0**
]1;1.5]	]0;0.8]	0	1
]1;1.5]	**]0.8;1]**	**1**	**0**
]1;1.5]	**]1;1.4]**	**1**	**0**

Table 5 Definition of the deterministic variable HC_1

determine the potential CL's composition we should be looking in @Web, be it an exact (step 3) or a close (step 4) match.

Queries result. The multiple queries performed returned fifteen results, partially presented in Tab.6. Each of these scenarios evaluates the probability of success of HC_1 (knowing that we already respect HC_2 and HC_3). Among the two scenarios presented in this article, the most probable ($p = 0.99$) is not an exact match: the closest CL is the pine bark, with an S-score of 5.26 (which is due to its too low **ash** content compared to the recommendation). The second presented scenario, while lower in term of probability ($p = 0.82$), is a perfect match with the rice husk. Despite this difference in probabilities, which would indicate that experiments with pine bark would always give perfect results, the rice husk should be favored for tests. Indeed, it is important to recall that one of the BN's limits is that it is a discrete model: behaviors around thresholds could be hard to predict. In the case of the pine bark, we have seen that its **ash** content is too low (1.44 on average), which already question its potential results; but in addition, its composition in **lignin** (27.33 in average) places it just above the recommended **lignin** content, making its results even more uncertain. Rice husk, on the contrary, presents compositions rather far from the discretization limits. It would, thus, seems safer to test this LF the first time. Compositions of the two LFs are presented in Tab.7.

p	0.99	p	0.82
Ash	[6.7;24.7]	**Ash**	[6.7;24.7]
Cellulose	[10.9; 25.6]	**Cellulose**	[25.6;33]
Lignin	[26.4; 49]	**Lignin**	[19.4; 26.4]
Exact	∅	**Exact**	Rice Husk
Similar	Pine Bark	**Similaire**	∅
S_{Pin}	5.26	S_{Riz}	0

Table 6 Example of results corresponding to the defined acceptability criteria and their probability p of achievement. When no exact match was found, an S-score was calculated to find the LF closest to the target.

	Ash	**Cellulose**	**Lignin**
Pine Bark	1.44	20.6	27.33
Rice Husk	14.5	31.9	25.7

Table 7 Composition of two potential LFs.

Discussion. In conclusion, if the choice of discretizations bears a meaning in the considered domain, it also always introduces biases: classifying a value in a given category can sometimes be tricky and some categories can, thus, be artificially augmented compared to others that are not enough represented in the KB. This highlights once more the importance of representativity in a learning database for machine learning algorithms: more diversity and examples would allow to limit these thresholds effects. However, this is not always easy when considering domains where obtaining data is costly, such as in biology.

5 Conclusion

In this article, we presented POND, a complete workflow dedicated to answering expert questions on knowledge bases representing transformation processes modeled by the PO^2 ontology. We focused on causality and on the tools offered by causal discovery (such as reverse engineering), by presenting the introduction of expert knowledge at different steps of the modeling. This is based on two points: the establishment of a common standardized vocabulary through the PO^2 ontology and the formalization of expert knowledge that cannot be directly expressed in a KB because it depends on the context.

We, then, illustrated this approach through a concrete application on bio-composite packaging. Thanks to the ontology, this workflow allows the expert to easily handle the expert knowledge to be integrated on one hand, and to add and modify it on the fly. Finally, we have defined a formalization of different expert constraints to guide the reading of the learned BN in order to elicit the most interesting answers from the user's perspective. Thus, through our illustration, we have presented several possible answers and identified potential new materials to be tested, still untested in the original database. This last part was done through the alignment with another KB, which opens new ways

for enriching the knowledge graph. This highlights the three possible uses of POND:

- Once a model is learned, it can be used for explanation, prediction and control of the different variables.
- Through expert critic, it can pinpoint potential knowledge holes and suggest new experiments to strengthen the learned model.
- Because it is based on semantized data, it can be extended to other KB through LOD tools in order to gather new information not represented in the original KB.

As in all causal analysis, it is very important to consider the context in which the learning was done (which was detailed in Sect.2.3). In this work, we have considered that the consulted expert has a reliable knowledge of the domain and no contradiction was integrated (which would not always be the case when confronted with a group of experts whose opinions can divert). The integration of these possible dissensions and their modeling in order to establish the learning of an optimal model is an avenue of research that we wish to explore.

Similarly, the recommendations established by the BN and generated automatically allow for a list of rules establishing more or less credible scenarios. For example, if we look at the Table 3, it seems highly unlikely that a high load (between 21 and 50) will result in an improved stress at failure (with a near zero probability of 0.06). The use of these rules to assess the credibility of new information or the suitability of the KB is another line to be explored in further work.

Finally, we have presented in this article an original approach to LOD, which proposes in our case to integrate new data not present in the original KB in order to enrich the knowledge discovery part. Future works should seek how to develop this part by suggesting new potential relevant data from other KB, using the learned model.

References

[1] Mélanie MUNCH, Juliette Dibie-Barthelemy, Pierre-Henri Wuillemin, and cristina manfredotti. Interactive Causal Discovery in Knowledge Graphs. In *PROFILES/SEMEX@ISWC 2019*, volume 2465 of *CEUR Workshop Proceedings*, pages 78–93, Auckland, New Zealand, October 2019. CEUR-WS.org.

[2] Roland Geyer, Jenna Jambeck, and Kara Law. Production, use, and fate of all plastics ever made. *Science Advances*, 3, 2017.

[3] E. Bugnicourt, P. Cinelli, A. Lazzeri, and V. Alvarez. Polyhydroxyalkanoate (PHA): Review of synthesis, characteristics, processing and potential applications in packaging. *Express Polymer Letters*, 8(11):791–808, 2014.

[4] Grégoire David, Giovanna Croxatto Vega, Joshua Sohn, Anna Ekman Nilsson, Arnaud Hélias, Nathalie Gontard, and Helene Angellier-Coussy. Using life cycle assessment to quantify the environmental benefit of upcycling vine shoots as fillers in biocomposite packaging materials. *International Journal of Life Cycle Assessment*, 2020.

[5] M.-A. Berthet, H. Angellier-Coussy, V. Chea, V. Guillard, E. Gastaldi, and N. Gontard. Sustainable food packaging: Valorising wheat straw fibres for tuning phbv-based composites properties. *Composites Part A*, 72(Complete):139–147, 2015.

[6] Jordi Girones, Thi To Loan Vo, Erika Di Giuseppe, and Patrick Navard. Natural filler-reinforced composites: Comparison of the reinforcing potential among technical fibers, stem fragments and industrial by-products. *Cellulose Chemistry and Technology*, 2017.

[7] Beatriz Montano Leyva, Gabriela Silva, Emmanuelle Gastaldi, Patricia Isabel Torres Chavez, Nathalie Gontard, and Hélène Angellier-Coussy. Biocomposites from wheat proteins and fibers: Structure/mechanical properties relationships. *Industrial Crops and Products*, 43:545–555, 2013.

[8] Lisa Ehrlinger and Wolfram Wöß. Towards a definition of knowledge graphs. In *SEMANTiCS (Posters, Demos, SuCCESS)*, 2016.

[9] Liliana Ibanescu, Juliette Dibie, Stéphane Dervaux, Elisabeth Guichard, and Joe Raad. Po2- a process and observation ontology in food science. application to dairy gels. *Metadata and Semantics Research*, pages 155–165, 2016.

[10] Mari Carmen Suárez-Figueroa, Asunción Gómez-Pérez, and Mariano Fernández-López. The neon methodology for ontology engineering. In Mari Carmen Suárez-Figueroa, Asunción Gómez-Pérez, Enrico Motta, and Aldo Gangemi, editors, *Ontology Engineering in a Networked World*, pages 9–34. Springer, 2012.

[11] Juliette Dibie, Stéphane Dervaux, Estelle Doriot, Liliana Ibanescu, and Caroline Pénicaud. $[MS]^2O$ - A multi-scale and multi-step ontology for transformation processes: Application to micro-organisms. In *ICSS*, pages 163–176, 2016.

[12] W3C OWL Working Group. Owl 2 web ontology language document overview (second edition). https://www.w3.org/TR/owl2-overview/.

[13] Juliette Dibie, Stéphane Dervaux, Liliana Ibanescu, and Joe Raad. Agroportal - process and observation ontology. http://agroportal.lirmm.fr/ontologies/PO2.

[14] Creative Commons. Attribution 4.0 international (cc by 4.0). https://creativecommons.org/licenses/by/4.0/.

[15] David Maxwell Chickering. Optimal structure identification with greedy search. *J. Mach. Learn. Res.*, 3(null):507–554, mar 2003.

[16] Gideon Schwarz. Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461 – 464, 1978.

[17] Nir Friedman, Lise Getoor, Daphne Koller, and Avi Pfeffer. Learning probabilistic relational models. In *IJCAI*, page 1300–1307. Morgan Kaufmann Publishers Inc., 1999.

[18] C.P. De Campos, Z. Zhi, and Q. Ji. Structure learning of bayesian networks using constraints. In *ICML*, pages 113–120, 2009.

[19] C. P. De Campos and Q. Ji. Improving bayesian network parameter learning using constraints. In *ICPR*, pages 1–4, 2008.

[20] Melanie Munch, Pierre-Henri Wuillemin, Cristina Manfredotti, Juliette Dibie, and Stephane Dervaux. Learning probabilistic relational models using an ontology of transformation processes. In *OTM 2017 Conferences*, pages 198–215, 2017.

[21] Gregory F. Cooper and Edward Herskovits. A bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9(4):309–347, 1992.

[22] Pekka Parviainen and Mikko Koivisto. Finding optimal bayesian networks using precedence constraints. *Journal of Machine Learning Research*, 14:1387–1415, 2013.

[23] Melanie Munch, Juliette Dibie, Pierre-Henri Wuillemin, and Cristina E. Manfredotti. Towards interactive causal relation discovery driven by an ontology. In *FLAIRS*, pages 504–508, 2019.

[24] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT press, 2nd edition, 2000.

[25] Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 10:524, 2019.

[26] Louis Verny, Nadir Sella, Séverine Affeldt, Param Singh, and Hervé Isambert. Learning causal networks with latent variables from multivariate information in genomic data. *PLOS Computational Biology*, 13, 2017.

[27] Alain Hauser and Peter Bühlmann. Two optimal strategies for active learning of causal models from interventional data. *International Journal of Approximate Reasoning*, pages 926—-939, 2014.

[28] Karthikeyan Shanmugam, Murat Kocaoglu, Alexandros G Dimakis, and Sriram Vishwanath. Learning causal graphs with small interventions. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[29] Federico Castelletti and Guido Consonni. Discovering causal structures in bayesian gaussian directed acyclic graph models. *Journal of the Royal Statistical Society Series A, Royal Statistical Society*, 183:1727–1745, 2020.

[30] Zhongli Ding, Yun Peng, and Rong Pan. *BayesOWL: Uncertainty Modeling in Semantic Web Ontologies*, pages 3–29. Springer Berlin Heidelberg, 2006.

[31] Shenyong Zhang, Yi Sun, Yun Peng, and Xiaopu Wang. Bayesowl: A prototype system for uncertainty in semantic web. *ICAI*, 2:678–684, 2009.

[32] Abdul-Wahid Mohammed. Knowledge-oriented semantics modelling towards uncertainty reasoning. *SpringerPlus*, 5, 2016.

[33] Stefan Fenz. Exploiting experts' knowledge for structure learning of bayesian networks. *Data And Knowledge Engineering*, 73:73 – 88, 2012.

[34] Montassar Ben Messaoud, Philippe Leray, and Nahla Ben Amor. Semcado: A serendipitous strategy for learning causal bayesian networks using ontologies. *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, pages 182–193, 2011.

[35] Giacomo Bucci, Valeriano Sandrucci, and Enrico Vicario. Ontologies and bayesian networks in medical diagnosis. *HICSS*, pages 1–8, 2011.

[36] Ahmed El Amine Djebri, Andrea G. B. Tettamanzi, and Fabien Gandon. Publishing uncertainty on the semantic web: Blurring the lod bubbles. In Dominik Endres, Mehwish Alam, and Diana Şotropa, editors, *Graph-Based Representation and Reasoning*, pages 42–56, Cham, 2019. Springer International Publishing.

[37] Denis Krompaß, Maximilian Nickel, and Volker Tresp. Querying factorized probabilistic triple databases. In *The Semantic Web – ISWC 2014*, page 114–129, Berlin, Heidelberg. Springer-Verlag.

[38] Mélanie Munch, Patrice Buche, Stéphane Dervaux, Amélie Breysse, Marie-Alix Berthet, Grégoire David, Sarah Lammi, Fleur Rol, Amandine Viretto,

and Hélène Angellier-Coussy. Biocomposites from poly(3-hydroxybutyrate-co-3-hydroxyvalerate) and lignocellulosic fillers: Processes stored in data warehouse structured by an ontology. *Data in Brief*, page 108191, 2022.

[39] Patrice Buche, Juliette Dibie-Barthelemy, Liliana L. Ibanescu, and Lydie Soler. Fuzzy Web Data Tables Integration Guided by a Termino-Ontological Resource. *IEEE Transactions on Knowledge and Data Engineering*, 25(4):805–819, 2013.