How can polydispersity information be integrated in the QSPR modeling of mechanical properties?

F. Cravero, S. A. Schustik, M. J. Martínez, M. F. Díaz, and I. Ponzoni

ABSTRACT
Polymer informatics is an emerging discipline that has benefited from the strong development that data science has experienced over the last decade. Machine learning methods are useful to infer QSPR (Quantitative Structure-Property Relationships) models that allow predicting mechanical properties related to the industrial profile of polymeric materials based on their structural repeating units (SRUs). Nonetheless, the chemical structure of the SRU is only one of the many factors that affects the industrial usefulness of a polymer. Other equally relevant factors are polymer molecular weight, molecular weight distribution, and production method, which are related to the inherent polydispersity of this kind of material. For this reason, the computational characterization used for the building of QSPR models for predicting mechanical properties should consider these main factors. The aim of this paper is to highlight recent advances in data science to address the inclusion of polydispersity information of polymeric materials in QSPR modeling. We present two dimensions of discussion: data representation and algorithmic issues. In the first one, we examine how different strategies can be applied to include polydispersity data in the molecular descriptors that characterize the polymers. We explain two data representation approaches designed by our group, named as trivalued and multivalued molecular descriptors. In the second dimension, we discuss algorithms proposed to deal with these new molecular descriptor representations during the construction of the QSPR models. Thus, we present here a comprehensible and integral methodology to address the challenges that polydispersity generates in the QSPR modeling of mechanical properties of polymers.

1. Introduction
There is a growing awareness that plastics or synthetic polymers are part of our lives at any level due to the versatility and low cost of these materials, which have increased their consumption almost 20 times over the last 30 years in multiple industrial applications. A strategy to deal with this high demand is the use of polymer informatics [1–6]. This interdisciplinary field applies the principles and practices of informatics to polymer science to improve the characterization, development, and discovery of new materials [7–12]. It is an emerging field that seeks to achieve efficient and reliable acquisition, management, analysis, and dissemination of data on diverse materials with the goal of greatly reducing the time and risk required to design, produce, and deploy a new material, which generally takes more than 20 years [13,14].

CONTACT I. Ponzoni ipo@cs.uns.edu.ar Instituto de Ciencias e Ingeniería de la Computación (ICIC), Universidad Nacional del Sur (UNS) - Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Bahía Blanca, Argentina

© 2022 The Author(s). Published by National Institute for Materials Science in partnership with Taylor & Francis Group
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
Several authors highlight as a challenge for creating polymeric databases the fact that a synthetic polymer is rarely a single chemical entity [14–16]. In this sense, even the simplest polymers are generally described by distributions. However, the characterization of the most used polymers today usually store only the data from the structural repeating units (SRU) of the polymers, without considering the polydispersity of these materials [14,17]. Hence, it is worth asking if data science can allow us to overcome these challenges.

Data science, especially data mining, develops methods and tools to extract meaningful information and patterns from data. The capability to predict a property value, with a certain degree of accuracy, is one of the most widely used applications of data mining in Chemistry. Emulating chemist in the well-known computer-aided drug design, materials researchers have used data mining for selecting materials with a desired property. Each material is described by features (molecular descriptors) that represent aspects of its chemical structure; and the target property is measured in a real material (after synthesis). Measuring this target requires the synthesis and characterization of materials; consequently, it is the most expensive and time-consuming step. Thus, in silico property prediction techniques have been proposed to use these predictions as a virtual screen to synthesize the minimum possible compounds (see Figure 1).

In particular, the Quantitative Structure–Property Relationship (QSPR) is a methodology that provides a mathematical relationship between the information encoded in the computational representation of the chemical structure of molecules and a given property (e.g. physicochemical, mechanical, etc.). As a result, the prediction of target properties for unknown materials is obtained before synthesis, based on the already known molecular descriptors [18–21]. Currently, QSPR is widely applied to several studies such as the creation of multifunctional databases based on design intent [22] and the prediction of important optical properties in organic [23] and inorganics materials [24]. Specifically, in the case of polymeric materials, the most studied property, and coincidentally the one for which more data is available, is glass transition temperature (T_g) [25,26]; from seminal works with only 22 compounds in their database [27,28] to the current ones with large locatable, accessible, interoperable, and reusable databases [29]. Several challenges are still present in the informatics of materials [3,9,30]; thus, it still constitutes a high-interest research topic nowadays [31–34].

Polymeric materials cannot be treated as in conventional molecular computing of small molecules, since in addition to being several orders of magnitude larger than them, they are usually polydisperse. This means that they are made up of families of molecules of the same structural constitution, but of different sizes; to represent them, it is necessary to refer to the average molecular weights derived from the molecular weight distribution curve. Another important value derived from this curve is the polydispersity index (PDI). It is equal to 1 when exceptionally all the molecules are of the same size; otherwise, it is higher than 1. Therefore, it is intuitive to think that this polydispersity affects the final properties of the polymeric materials. In fact, the mechanical properties are one of the most affected by this intrinsic condition, in

![Figure 1](image-url). Classical process against in silico design process.
which every fraction of molecular size contributes to the final property. This is the reason why a new approach from polymer informatics is needed to successfully address property prediction [35,36].

The aim of this paper is to present the relevance of polydispersity in the QSPR modeling of polymer mechanical properties and to highlight recent developments in data science addressing this issue. Two dimensions of analyses are presented: how to incorporate polydispersity information in the computational representation of polymers and how to address the algorithmic challenges derived from these new representations. In the first one, we explain two data representation approaches designed by our group based on the definition of trivalued and multivalued molecular descriptors. Both representations capture information from the molecular weight distribution of the materials and use an in silico polymerization strategy also developed by our group. In the second analysis, we discuss the algorithms proposed by our group to treat these new representations during the building of the QSPR models.

2. Polydisperse synthetic polymers and mechanical properties

As explained above, the size of the polymeric materials is represented in the molecular weight distribution curve. Two widely used average weights are derived from there: \( Mn \) (number-average molar weight) and \( Mw \) (weight-average molar weight), which are graficated in Figure 2 and their mathematical formulas are described in Equations (1) and (2), respectively. In addition, the PDI is obtained from the ratio of both \((Mw/Mn)\). At this point, it is relevant to understand that the molecular weight distribution has a great influence on mechanical properties. For example, for polypropylene (PP), the higher their average molecular weights, the higher the modulus of elasticity and the tensile strength (properties derived from the tensile test) [36]. For this reason, companies produce relatively new PPs in which the extremes of a remarkably high molecular weight have been eliminated without involving a considerable increase in the extremes of a low molecular weight, thus decreasing the PDI (narrow molecular weight distribution curve). Consequently, the resulting PP has a lower level of elasticity [37].

\[
Mn = \frac{\sum N_i M_i}{\sum N_i} \quad (1)
\]

\[
Mw = \frac{\sum (N_i M_i) M_i}{\sum N_i M_i} \quad (2)
\]

where \( N_i \) is the number of molecules of molecular weight \( M_i \).

![Figure 2. Typical molecular weight distribution curve for polymeric materials.](image)

The mechanical properties of a material are those that involve a reaction to an applied force. Although mechanical tests are standardized, the derived properties are not constant and often change as a function of temperature, loading rate, and other conditions (testing parameters). In particular, the tensile test consists of subjecting the specimen of a material to a variable tensile force at a constant crosshead speed until it fractures (Figure 3(a)). The stress–strain curve derived from the test provides information on the stiffness, strength, toughness, and ductility of the polymer, that is, how the material behaves under the testing conditions (Figure 3(b)) [38,39].

3. Data science and QSPR modeling in polymer informatics

QSPR is one of the most widely used methodologies for in silico molecular modeling. Once validated, these models are applied to predict the properties of new compounds [40–42]. The main advantages of QSPR models are the reduction of time and cost because of the computational design prior to synthesis. In this context, data science has a key role, since it allows inferring QSPR models using machine learning techniques applied to discover patterns and relationships among data [43].

In the context of polymeric materials, QSPR modeling presents major challenges due to the uncertainty introduced by the polydispersity of these materials and the few reliable data available to build databases. In this sense, both the calculation and the selection of molecular descriptors become more complex and computationally demanding tasks because of the large size of the polymers and their characterization, which is given by the molecular weight distribution curve and not by a single value,
as in the case of drug molecules. Most of the existing QSPR models for polymeric materials use a single molecule to describe each material without considering polydispersity; these approaches tend to simplify the computational representation of the material using only its SRU [14,17,44,45]. Then, the calculation of the molecular descriptors is done using the shortest chain of polymers that characterizes the material. In this context, it becomes necessary to add the information from the molecular weight distribution curve to achieve more reliable predictive models, as explained in the following sections.

Furthermore, the limited availability of data poses a problem in itself. Compared to other molecules, existing studies on the properties of polymers, particularly those related to mechanical properties, are very scarce. Especially, if we are looking for cured datasets, most of them have little data, and molecular weight distribution curves or other structural characteristics of polymers useful to predict the target property are rarely reported. Likewise, different experimental factors that determine the attainment of the final value of the target property must be considered, such as those related to the processing of the material, the thermal history, and the property measurement test (e.g. the tensile test parameters), since these experimental parameters influence the value of the measured property. Consequently, there arises the need to homogenize the data and gather those obtained under the same conditions (temperature, crosshead speed, specimen shape, etc.). At the same time, the task of homogenizing the data is difficult, since there is not a large amount of experimental data generated under similar conditions. Therefore, developing a polymer database for mechanical properties is arduous work [14,15,44].

4. Trivalued representation of polymers

The first proposal incorporates relevant information from the weight distribution curve through a trivalued representation of the molecular descriptors of the polymers [46]. We work with an ad hoc database that has 77 homopolymers, linear and amorphous, curated according to criteria for the processing of the material, the thermal history, and the property measurement test norm. For this, trivalued representation, in addition to the SRU, it is necessary to compute the two polymeric chains whose lengths correspond to the $M_n$ and $M_w$ values of the material by means of in silico polymerization. In particular, in our database the Mn varied within a range from 4700 to 765,000 [g/mol] and the Mw varied from 19,500 to 2,200,000 [g/mol]. Thus, each material will be represented by the Simplified Molecular Input Line Entry Specification (SMILES) corresponding to three chemical structures relevant for its characterization. Hereafter, we will refer to these three SMILES encodings as instances of different weight of the same material mat, generically denoting them as mat$_{SRU}$, mat$_{Mn}$, and mat$_{Mw}$. Therefore, the $i$-th polymer in a materials database will have the following chemical structures associated with it: mat$_{SRU}$, mat$_{Mn}$, and mat$_{Mw}$. This process is outlined in Figure 4.

Regarding the in silico polymerization, we developed a computational tool called PolyMaS [47], which characterizes an instance of weight in 2D using the SMILES code that allows a more compact representation than the connection tables. This computational polymerization is performed from the SMILES code (character string) of the SRU that contains the indicated head and tail. Then, through operations with these character strings, a head-tail polymerization is generated, resulting in a single polymer chain of the
required length for the instance to be represented. Note, the computational polymerization expression refers to the macromolecule generation procedure.

Molecular descriptors can be calculated for each of these instances, defining three databases for the same group of materials: DB_{SRU}, DB_{Min} and DB_{Mw}. Note that each database refers to the same universe of materials and molecular descriptors; the only thing that varies is the instance of weight represented in each case. For this reason, the value mdA_i^{Mw} corresponds to the calculation of the descriptor mdA made on the instance of weight Mw of the i-th material. In this way, it is possible to define the concatenation of the DB_{SRU}, DB_{Min} and DB_{Mw} databases as the trivalued representation of the molecular descriptors of a materials database. Figure 5 shows the definition of the database with trivalued representation, DB_{TRI}.

4.1. Algorithmic construction of QSPR using the trivalued representation

Once the DB_{TRI} is obtained, the next step to build the QSPR model is to identify the most relevant descriptor values for the property to be predicted. This task can be performed through the feature selection algorithms typically used in data science. This representation allows identifying situations in which a molecular descriptor is relevant for a particular instance of weight but not for the rest, as well as detecting those that are relevant for all instances. For example, Cravero et al. [46] identified that many selected descriptors were computed for several instances of weight, which allows promoting the design of a multivalued representation.

Descriptor selection can be carried out in a completely automatic way, directly taking the output of the feature selection procedure (whose search algorithm can be, for example, a neural network, random committees, or a random forest), or also this output can be intervened by a domain expert (expert-in-the-loop), using visual analytical methods, such as VIDEAN [48]. This second option can help to obtain more interpretable models in physicochemical terms, since the expert’s knowledge allows prioritizing descriptors according to their semantic meanings. The last step to obtain the predictive model consists of conducting a supervised machine learning process using the information from the selected molecular descriptors and the target property.

Figure 4. Use of the SRU and weight distribution curve of a material for in silico polymerization of instances of different weight.

Figure 5. Generation of a database with trivalued molecular descriptors.
4.2. Performance analysis of the impact of using the trivalued representation

In Cravero et al. [46], an exhaustive performance analysis of using trivalued representations were presented. In particular, the main research question explored in that work was: ‘is it advisable to integrate, in a single database, molecular descriptors corresponding to polymeric chains of different characteristic weights related to the molecular weight distribution curves of the materials?’ In the mentioned work, an in-house database was developed and applied by our research lab [49]. The polymers included in this database are homopolymers, linear and amorphous. The 77 initial polymers were characterized by their SRU in SMILES code. To obtain the DB\textsubscript{SRU}, DB\textsubscript{Mn} and DB\textsubscript{Mw} databases in that work, it was necessary to generate the macromolecules that reached the average weights of the polymers in the database by using the PolymaS software tool [47]. After that, once all molecular descriptors for the three databases were computed, the trivalued database was defined as the DB\textsubscript{TRI} joining all the molecular descriptors (see Figure 5). Note that this integrated DB was denoted as DB\textsubscript{Global} in Cravero et al. [46], but it is conceptually the same DB. Therefore, this fourth database contains all the information associated to the three different instances of molecular weight of polymeric materials, characterizing polymeric materials by capturing part of their polydispersity. Regarding the studied targets, three mechanical properties derived from the tensile test were included in the performance analysis: tensile modulus, elongation at break and tensile strength at break [50].

Several experiments were executed in Cravero et al. [46] to establish performance comparisons among the three representations based only on single-polymeric chains, DB\textsubscript{SRU}, DB\textsubscript{Mn} and DB\textsubscript{Mw}, also known as unique weight instance representations, and the trivalued representation, DB\textsubscript{TRI}. In these experiments, different machine learning strategies were explored both for selecting the most useful molecular descriptors and for learning the most accurate QSPR models. From all these experiments, we will only focus here on the comparison between the best QSPR models obtained from the unique weight instance representations and the trivalued representation because these results bring answer to the early mentioned research question, but the detailed results of the remaining experiments can be found in Cravero et al. [46].

The generalizability of the QSPR\textsubscript{TRI} models can be compared with the unified generalizability quantification of QSPR models generated from unique weight instance representations (DB\textsubscript{SRD}, DB\textsubscript{Mn} and DB\textsubscript{Mw}), which were denoted as QSPR\textsubscript{SRU}, QSPR\textsubscript{Mn}, and QSPR\textsubscript{Mw} models respectively. This methodology consists of computing the $R^2$ metric of the alternative models generated from the unique weight instance representations (SRU, Mn, and Mw) in an aggregated manner called QSPR\textsubscript{AWI} (all weight instances; AWI). For example, the $R^2$ value of the QSPR\textsubscript{AWI} aggregated model for the tensile modulus property is 0.9609. This value was obtained by computing the $R^2$ value that corresponds to the complete set of prediction results obtained by QSPR\textsubscript{SRU}, QSPR\textsubscript{Mn}, and QSPR\textsubscript{Mw} models for the same property when these models are applied in their external validation datasets respectively (adding the three testing outputs in a unique set of results). Therefore, the validation outputs obtained by these three QSPR models are interpreted as the $R^2$ of a unified QSPR model.

Table 1 shows the $R^2$ performances of the QSPR\textsubscript{AWI} models reported by Cravero et al. [46] for the three target properties in contrast with the $R^2$ metrics achieved by the QSPR\textsubscript{TRI} models for the same property validation datasets. In all cases, the trivialized representation achieves the highest $R^2$ values, supporting with evidence an affirmative response to our research question.

| Target Property          | QSPR model | $R^2$ | MAE  | RMSE |
|--------------------------|------------|-------|------|------|
| Tensile Modulus          | AWI        | 0.9679| 6.8495| 8.3651|
| Elongation at Break      | TRI        | 0.9725| 8.1382| 9.4827|
| Tensile strength at Break| AWI        | 0.9002| 1.1571| 1.4524|

5. Multivalued representation of polymers

In the second representation, we proposed to define the molecular descriptors of the polymers from data sampled from the weight distribution curve [51,52]. For this representation, in addition to the SRU, it is necessary to sample a set of $s$ instances of different weight $w_i$ from the weight distribution curve of the material together with their corresponding frequencies $f_i$. Ideally, these instances should be equally distributed along the x-axis of the weight distribution curve, covering as wide as possible the weight range of the material. Once the sampling is completed, the in silico polymerization of the polymer chains corresponding to the $s$ sampled weights must be carried out. Thus, for the same material mat, we will have $s$ instances of weight that characterize it, which are encoded as SMILES and which we will call consecutively as $mat_1, mat_2, \ldots, mat_s$. Thus, the $i$-th polymer in a material database will have the following $s$ associated chemical structures: $mat_1, mat_2, \ldots, mat_s$. Furthermore, as we also know the frequencies $f_i$ of these chains in the bulk material, we can say that the
$i$-th polymer will be characterized by the following $s$ pairs: $(m_{i,1}, f_1), (m_{i,2}, f_2), \ldots, (m_{i,s}, f_s)$. This process is outlined in Figure 6.

Each of these instances can be calculated with its molecular descriptors, thus being able to define databases for the same group of materials: $DB_1, DB_2, \ldots, DB_s$. Note that each database refers to the same universe of materials and molecular descriptors; the only thing that varies is the instance of weight represented in each case. For this reason, the value $mdA_{i}^{i}$ corresponds to the calculation of the descriptor $mdA$ carried out on the instance of weight $mat_{i}^{i}$ of the $i$-th material. In this way, it is possible to define a multivalued representation of the molecular descriptors of a material as a discrete distribution that combines the values of the descriptors computed for the different instances of weight with their corresponding frequencies. Following our notation, the discrete distribution corresponding to the descriptor $mdA$ for the $i$-th polymer, denoted as $mdA_{i}^{i}$, would be defined by the following sequence of pairs: $(mdA_{i,1}^{i}, f_1), (mdA_{i,2}^{i}, f_2), \ldots, (mdA_{i,s}^{i}, f_s)$. Thus, a $DB_{MUL}$ database corresponding to this multivalued representation will be integrated by molecular descriptors modeled using discrete distributions closely related to the weight distribution curves of the materials. The definition of the database with multivalued representation, $DB_{MUL}$, is outlined in Figure 7.

5.1. Algorithmic construction of QSPR using the multivalued representation

Unlike the classic representations based on the SRU and the trivalued representation, where the molecular descriptors are characterized as real or integer numbers, the multivalued representation defines the descriptors as discrete distributions. For this reason, it is not possible to apply in the multivalued representation the traditional feature selection algorithms to identify the most relevant descriptors or the traditional supervised learning methods to learn the QSPR model. In other words, the multivalued representation requires specific algorithms to infer the predictive models.

To drive the selection of molecular descriptors, our group proposed an algorithm called FS4RV$_{DD}$, as an acronym for Feature Selection for Random Variables with Discrete Distribution. This method was proposed by Cravero et al. [52] and a more in-depth evaluation of its performance, in terms of scalability and noise level of the data, was later presented in Cravero et al. [51]. Below is a self-contained overview of this method.

FS4RV$_{DD}$ is encoded in two sequential phases, as illustrated in Figure 8. In the first phase, a pairwise correlation analysis is performed to rank the molecular descriptors from the highest to the lowest correlated with the target property to be predicted. In terms of our notation, we will express, for example, that the
descriptor \( mDA \) of a database with \( q \) polymers will be represented by a vector with the form: \( \{mDA^1, mDA^2, \ldots, mDA^i, \ldots, mDA^q \} \), where for each \( i \)-th polymer \( mDA^i \) will contain the discrete distribution defined by the following pairs: \( (mDA^i_j, f^i_j) \), where \( f^i_j \) and \( s \) will represent the number of samples that were taken from the weight distribution curve. On the other hand, the values of the target property to be predicted can be represented by a real-valued vector that we can denote as \( tp = [tp^1, tp^2, \ldots, tp^i, \ldots, tp^q] \), where \( tp^i \) corresponds to the target property value for the \( i \)-th polymer of \( DB_{MUL} \).

Since we cannot directly calculate a correlation between the vector that represents a molecular descriptor, such as \( mDA \), and a vector of real numbers, such as \( tp \), in \( FS4RV_{DD} \) we sample \( k \) times the discrete distributions contained in a descriptor, thus generating \( k \) vectors of real numbers. Continuing with our example, in the case of the descriptor \( mDA \), each sample \( mdA_k \) would be a vector formed by \( q \) real values, where the \( i \)-th input of the vector \( mDA_k \) would come from sampling the discrete distribution \( mDA^i \). Then, it is possible to calculate the Pearson correlation between each of these \( k \) vectors sampled from \( mDA \) and the vector \( tp \), and thus define a vector \( p_{mDA}^{kp} \) that will contain \( k \) correlation values computed between both variables. Finally, the point estimator \( (\mu) \) of \( p_{mDA}^{kp} \) and its corresponding variance \( (f) \) are computed to rank the descriptors. The higher the value of the point estimator \( p_{mDA}^{kp} \), the higher the position of \( mDA \) in the ranking. Between descriptors with same point estimator values, the variance will be used to break the tie, being the winner the one with the lowest variance.

In the second phase of \( FS4RV_{DD} \), the best \( c \) descriptors will be selected from an iterative exploration of the ranking elaborated in the previous phase, where the cardinality \( c \) of the selected subset is a parameter defined by the user of the algorithm. This process will prioritize the choice of the highest correlated with the target variable descriptors (i.e. the first ones in the ranking), but they must not be similar to each other. In other words, it is sought that the \( c \) selected molecular descriptors contribute with non-redundant information.

To measure the similarity between two descriptors \( mDA = \{mDA^1, mDA^2, \ldots, mDA^i, \ldots, mDA^q \} \) and \( mDB = \{mDB^1, mDB^2, \ldots, mDB^i, \ldots, mDB^q \} \), it is necessary to compare the similarities that both present vectors have for the same entries; that is, is necessary to compare the similarity between each pair of discrete distributions \( (mDA^1, mDB^i) \), for \( i \) varying from \( 1 \) to \( q \). This can be achieved by calculating the Bhattacharyya Distance [53], a function denoted as \( B_{dist} \) that generates values between 0 and 1, where zero indicates the highest degree of similarity between the contrasted distributions. In this way, we can define a distance vector between the descriptors \( mDA \) and \( mDB \) denoted as \( dist_{mDA}^{mDB} = [B_{dist}(mDA^1, mDB^1), B_{dist}(mDA^2, mDB^2), \ldots, B_{dist}(mDA^q, mDB^q)] \). Finally, we will establish that \( mDA \) and \( mDB \) contain similar information if the ratio between the number of entries of \( dist_{mDA}^{mDB} \) that have a similarity value lower than \( U_{BD} \) and the total number of entries, \( q \), is greater or equal to \( U_{SR} \), where \( U_{BD} \) and \( U_{SR} \) are two thresholds defined by the user as parameters of the \( FS4RV_{DD} \) algorithm.

The last step to obtain the QSPR model consists of conducting a supervised machine learning process using the information from the selected molecular descriptors and the target property. To accomplish this, it is necessary to define supervised learning methods designed to solve regression problems in which the descriptive variables correspond to discrete distributions. Figure 9 outlines a possible algorithmic methodology to carry out this task proposed by Cravero et al. [52]. The strategy is very simple and consists of generating \( k \) samples of all the discrete distributions associated to the different molecular descriptors selected to build the QSPR model. Thus, our original supervised learning problem with descriptors represented as discrete distributions is now transformed into a new
5.2. Performance analysis of the impact of using the multivalued representation

In Cravero et al. [51], an exhaustive performance analysis of using multivalued representations were presented. In particular, the main research question explored in that work was: given a probabilistic characterization of molecular descriptors by using discrete distributions in combination with the FS4RVDD method, is it possible to achieve a more accurate identification of the most relevant molecular descriptors than the one obtained from traditional approaches that use a simplified representation of the molecular descriptors.

Developing a database with multivalued characterization of real polymers is a complex task. First, it is required to collect the weight distribution curves of the materials in order to sample various instances of different weights and their frequencies in the bulk material. These curves are not usually reported in publications, unlike what usually happens with the availability of \( M_n \) and \( M_w \) values required for trivalued representation. Additionally, the calculation of the molecular descriptor values for the largest instances sampled from these curves is computationally expensive since the huge size of these macromolecules. For all these reasons, in Cravero et al. [51] and Cravero et al. [52] synthetic databases were generated for evaluating a proof of concept of the multivalued representation and studying the soundness and scalability of FS4RVDD algorithm, where each material, \( \text{mat}_i \), is represented by a discrete distribution. Then, this discrete distribution is a template for obtaining the discrete distributions of the molecular descriptors associated to the \( i \)-th simulated material. A detailed explanation of this procedure is provided in Cravero et al. [51].

Three synthetic databases integrated for 400, 800, and 1600 polymeric materials were built. Each database included 100 molecular descriptors and diverse scenarios were defined by varying the conditions for building the synthetic target variables. Targets had been created by changing the number of molecular descriptors randomly selected (5, 10, or 20), modifying the kind of correlation (linear and nonlinear), and incorporating (or not) noise to the data. In summary, the combinations of these conditions generated 36 different experimental scenarios.

An extra challenge was to set a fair experimental framework for performance comparisons with other approaches, because there are no similar methods that work with molecular descriptors characterized by discrete distributions. For this reason, Cravero et al. [51] decided to compare FS4RVDD method with state-of-the-art feature selection methods using representations based on a single value. In particular, Cravero et al. [51] decided to use the smallest value of the discrete distribution of \( \text{mat}_i \) simulated polymer as an analogue for the SRU-based representation of this material. In this way, an additional synthetic database with single value representations of molecular descriptors could be defined considering only the first value of the discrete distribution as the single descriptor value.
Regarding the performance metrics, the idea was to determine the capability of a feature selection method for detecting the molecular descriptors that had been previously correlated with the target variables during the construction of the simulated data sets. Then, the performance can be assessed as in a classification problem, in which the feature selection algorithm classifies the molecular descriptors into two classes: correlated and not correlated with the simulated target variable. In this work, we will focus on the Percentage of Correctly Classified (%CC) molecular descriptors, also known as Accuracy in the Machine Learning research, but other additional metrics and experiments were also reported in detail by Cravero et al. [51].

For the single value representation, the AttributesSelection method included in Weka [54] was used in that work, where the attributes correspond to the molecular descriptors of the simulated database. It provides a fair and direct comparison with respect to the FS4RVDD method in terms of classification metrics. In particular, the following parameters: CorrelationAttributeEval as attribute evaluator and Ranker as search method were used.

Table 2 shows %CC values achieved by the multi-valued and single-valued representations using FS4RVDD and AttributesSelection as feature selection algorithms, respectively, for different experimental conditions. Only in one experiment, that correspond with the database that have 400 simulated materials, 20 molecular descriptors selected, linear target calculated and without noise in the data, the high performance is obtained by the single-valued representation. In the remaining 35 scenarios the multivalued representation achieves the highest %CC values, which supports an affirmative response to our research question.

### 6. Concluding remarks

The demand for polymer design is experiencing strong growth, with changing scenarios that are shaping new and complex challenges, which favor an increase in the volume of data available for the discovery of new materials, making more feasible the application of data science techniques. However, achieving an adequate use of these techniques requires rigorous and realistic computational representation models of polymers. In this regard, a shortcoming identified by many authors is the lack of characterization of polydispersity in the most used representations for QSRR modeling of this type of materials, such as those based on SRU. For this reason, in this paper, two alternative computational representations of polymers, previously developed for our research group, were described to show how an expert can incorporate information from the weight distribution curve of a material by using trivalued and multivalued representations.

The trivalued representation constitutes a more rigorous characterization of a polymer than the characterizations based on a single polymer chain, such as those supported exclusively on SRU. In this sense, the experiments reported by Cravero et al. [46] provided clear evidence that the combination of data from the molecular descriptors obtained for the three instances of weight considered (SRU, Mn, and Mw) allows generating more precise QSRR models than those obtained using a single instance of weight, whatever the instance of weight considered. Another favorable aspect of this representation is that, in general, the Mn and Mw values of polymeric materials are usually provided in publications; therefore, it is not difficult to access this information. Beyond these advantages, it is worthy to note that the trivalued representation does not take advantage of the information about the frequency that each instance of weight has in the weight distribution curve of the material. These frequencies have a clear impact on the properties of a polymer and, for this reason, they were considered to define the multivalued representation.

In our opinion, the multivalued representation is conceptually the most realistic characterization for polymers in the context of QSRR modeling, because it captures the information about the molecular weights and frequencies of several polymeric chains that define the bulk material. The results reported by Cravero et al. [51, 54] showed that this alternative is clearly superior to the single-value representations, like it is the SRU-based representation,
for detecting the most relevant molecular descriptors and capture more information from the weight distribution curve of the materials than the trivalued representation presented in the previous section.

Nevertheless, these approach presents several important challenges. Firstly, to the already well-known problem of the lack of availability of databases of polymeric materials for QSRR modeling, the need of weight distribution curves for each material is added here. This limitation is so strong that in the publications in which this representation was presented it was necessary to evaluate its feasibility and performance using simulated data. For this reason, it should be acknowledged that the trivalued representation is supported by data that is currently more accessible.

Secondly, the new algorithms that the multivalued representation demands pose an important challenge. In the present work, we describe the $FS4RV_{DD}$ feature selection algorithm and a consensus QSRR modeling method, both developed by our group, which provide a first alternative solution. However, there is certainly much work to be done in this regard. An alternative we are currently evaluating is to define a new multivalued representation based on molecular descriptors represented as fuzzy numbers. In this way, the information from the weight distribution curve would continue to be sampled to generate the molecular descriptors, but it would also take advantage of all the methodologies that fuzzy logic provides.

Finally, beyond the challenges mentioned above, we consider that multivalued representations will be an affordable alternative in the medium term, thanks to the strong advances that data science is making in the field of Materials Sciences. This should undoubtedly lead to an ever-increasing availability of the type of databases required to support these rigorous approaches.

Acknowledgements

This work was partially supported by the Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina, [Grant Nº PIP 112-2017-0100829], the Universidad Nacional del Sur (UNS), Bahía Blanca, Argentina, [Grants Nº PGI 24/N042 and PGI 24/ZM17], and the Agencia Nacional de Promoción Científica y Tecnológica [Grants Nº PICT 2018-04533 and PICT-2019-03350].

Author contributions

All the authors helped to design, implement, and validate the trivalued and multivalued methodologies discussed in this article. In particular, in this manuscript, F.C. and M.J.M. focused on the presentation of data science approaches to QSRR modeling and their antecedents in polymer informatics. S.A. S. centered on the presentation of the in silico polymerization method. M.F.D. described all aspects of the growing relevance of the polymer industry, the issues related to polydispersity, and the challenges and opportunities in the prediction of polymer mechanical properties. I.P. conceived the mathematical formulation used to present the trivalued and multivalued representations and the description of their associated algorithms. Finally, all the authors reviewed and approved the submitted manuscript.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was partially supported by the Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina [Grant Nº PIP 112-2017-0100829], the Universidad Nacional del Sur (UNS), Bahía Blanca, Argentina, [Grants Nº PGI 24/N042 and PGI 24/ZM17], and the Agencia Nacional de Promoción Científica y Tecnológica [Grants Nº PICT 2018-04533 and PICT-2019-03350].

Notes on contributors

Fiorella Cravero holds her BSc. degree in Bioinformatics from the Universidad Nacional de Entre Ríos. She received her PhD in Computer Science from the Universidad Nacional del Sur in 2020. She is working as a postdoctoral fellow in the BioChemTIc group at ICIC (UNS-CONICET). Her research on the interdisciplinary field of polymer informatics focuses on the computational representation of high molecular weight polymers.

Santiago A. Schustik received his degree in Computer Systems Engineering in February 2017 and began his PhD in Computer Science in the field of molecular computing in April 2017, both at the Universidad Nacional del Sur (UNS). His research focuses on the design of QSRR models for polymer informatics working at PLAPIQUI (UNS-CONICET).

María Jimena Martínez has held a PhD in Computer Science from the Universidad Nacional del Sur since 2017. She is currently an Assistant Researcher at ISISTAN (UNICEN-CONICET). Her interest focuses on machine learning methods in cheminformatics and polymer informatics.

Mónica Fátima Díaz is a Professor in the Department of Chemical Engineering at the Universidad Nacional del Sur. She earned her PhD in Chemistry at UNS, with expertise in the field of polymeric materials. She is an Independent Researcher at PLAPIQUI (UNS-CONICET). She is active in cheminformatics, polymer informatics, and environmental chemistry.
Ignacio Ponzoni holds a PhD in Computer Science. He is an Principal Researcher at the ICIC (UNS-CONICET), where he leads the BioChemTICs group. His research interests focus on the development of data science and methods applied to bioinformatics and cheminformatics.

He was the president of the Argentinean Association for Computational Biology and Bioinformatics (A2B2C) from May 2013 to May 2015. Nowadays, he is the vice-head of the Institute of Computer Science and Engineering.

**ORCID**

F. Cravero http://orcid.org/0000-0002-9816-4147  
S. A. Schusterski http://orcid.org/0000-0003-1402-7055  
M. J. Martinez http://orcid.org/0000-0002-0443-5795  
M. F. Diaz http://orcid.org/0000-0002-6680-8067  
I. Ponzoni http://orcid.org/0000-0002-6923-9592

**References**

[1] Venkatraman V, Alshberg BK. Designing high-refractive index polymers using materials informatics. Polymers. 2018;10(1):103.

[2] Rasulev B, Casanola-Martin G. Qsar/QspR in polymers: recent developments in property modeling. J Quant Struct -Prop Relat. 2020;5(1):80–88.

[3] Chandrasekaran A, Kim C, Ramprasad R. Polymer genome: a polymer informatics platform to accelerate polymer discovery. In: Schütt K, Chmiela S, Von Lilienfeld O, et al, editors. Machine learning meets quantum physics. lecture notes in physics. Vol. 968. Cham: Springer; 2020. p. 397–412.

[4] Kuenneth C, Rajan AC, Tran H, et al. Polymer informatics with multi-task learning. Patterns. 2021;2(4):100238.

[5] Sha W, Li Y, Tang S, et al. Machine learning in polymer informatics. Infomat. 2021;3(4):353–361.

[6] Chen L, Piliana G, Batra R, et al. Polymer informatics: current status and critical next steps. Mat Sci Eng R. 2021;144:100595.

[7] Wu S, Kondo Y, Kakkimoto MA, et al. Machine-learning-assisted discovery of polymers with high thermal conductivity using a molecular design algorithm. Comp Mater. 2019;5(1):1–11.

[8] Himanen L, Geurts A, Foster AS, et al. Data-driven materials science: status, challenges, and perspectives. Adv Sci. 2019;6(21):1900808.

[9] Suh C, Fare C, Warren JA, et al. Evolving the materials genome: how machine learning is fueling the next generation of materials discovery. Annu Rev Mater Res. 2020;50:1–25.

[10] Wang AYT, Murdock RJ, Kauwe SK, et al. Machine learning for materials scientists: an introductory guide toward best practices. Chem Mater. 2020;32(2):4954–4965.

[11] Liu Y, Esan OC, Pan Z, et al. Machine learning for advanced energy materials. Energy Artif Intell. 2021;3:100049.

[12] Lin TS, Rebello NJ, Beech HK, et al. PolyDAT: a generic data schema for polymer characterization. J Chem Inf Model. 2021;61(3):1150–1163.

[13] Mulholland GJ, Paradiso SP. Perspective: materials informatics across the product lifecycle: selection, manufacturing, and certification. APL Mater. 2016;4(5):053207.

[14] Audus DJ, de Pablo JJ. Polymer informatics: opportunities and challenges. ACS Macro Lett. 2017;6(10):1078–1082.

[15] Otsuka S, Kuwajima I, Hosoya J, et al. PolYlnfo: polymer database for polymeric materials design. In 2011 International Conference on Emerging Intelligent Data and Web Technologies; Tirana, Albania. IEEE, 2011, 22–29.

[16] Hill J, Mulholland G, Persson K, et al. Materials science with large-scale data and informatics: unlocking new opportunities. MRS Bull. 2016;41(5):399–409.

[17] Holmes PF, Bohrer M, Kohn J. Exploration of polymethacrylate structure-property correlations: advances towards combinatorial and high-throughput methods for biomaterials discovery. Prog Polym Sci. 2008;33(8):787–796.

[18] Gantzer P, Creton B, Nieto-Draghi C. Inverse-QSPR for de novo design: a review. Mol Inform. 2020;39(4):1900087.

[19] Sukumar N, Krein M, Luo Q, et al. MQSPR modeling in materials informatics: a way to shorten design cycles? J Mater Sci. 2012;47(21):7703–7715.

[20] Wu K, Natarajan B, Morkowchuk L, et al. From drug discovery QSAR to predictive materials QSpar: the evolution of descriptors, methods, and models. In: Krishna Rajan, editor. Informatics for materials science and engineering. Oxford (UK): Butterworth-Heinemann. 2013. 385–422.

[21] Stumpfe D, Bajorath J. Current trends, overlooked issues, and unmet challenges in virtual screening. J Chem Inf Model. 2020;60:4112–4115.

[22] Broderick S, Rajan K. Informatics derived materials databases for multifunctional properties. Sci Technol Adv Mater. 2015;16:013501.

[23] Erickson ME, Ngogang M, Rasulev B. A refractive index study of a diverse set of polymeric materials by QSPR with quantum-chemical and additive descriptors. Molecules. 2020;25(17):3772.

[24] Kuz’mın VE, Ognichenko LN, Zinchenko VF, et al. QSpar models for predicting of the melting points and refractive indices for inorganic substances: components of the optical film-forming materials. Int J Quant Struct -Prop Relat. 2020;5(1):1–21.

[25] Adams N. Polymer Informatics. In: Meier M, Webster D, editors. Polymer libraries. Adv polym sci. Vol. 225. Berlin Heidelberg: Springer; 2010. p. 107–149.

[26] Chen M, Jabeen F, Rasulev B, et al. A computational structure–property relationship study of glass transition temperatures for a diverse set of polymers. J Polym Sci Pol Phys. 2018;56(11):877–885.

[27] Katritzky AR, Rachwal P, Law KW, et al. Prediction of polymer glass transition temperatures using a general quantitative structure–property relationship treatment. J Chem Inform Comp Sci. 1996;36(4):879–884.

[28] Ma R, Luo T. P11M: a benchmark database for polymer informatics. J Chem Inf Model. 2020;60(10):4684–4690.
[29] Tchoua R, Hong Z, Audus D, et al. Developing databases for polymer informatics. B Am Phys Soc. 2020;65(1).

[30] Ramprasad R, Batra R, Pilania G, et al. Machine learning in materials informatics: recent applications and prospects. Comp Mater. 2017;7(1):1–13.

[31] Peerless JS, Milliken NJ, Oweida TJ, et al. Soft matter informatics: current progress and challenges. Adv Theor Simul. 2019;2(1):1800129.

[32] Kim C, Chandrasekar An, Huan TD, et al. Polymer genome: a data-powered polymer informatics platform for property predictions. J Phys Chem C. 2018;122(31):17575–17585.

[33] Rickman JM, Lookman T, Kalinin SV. Materials informatics: from the atomic-level to the continuum. Acta Mater. 2019;168:473–510.

[34] Chen L, Kim C, Batra R, et al. Frequency-dependent dielectric constant prediction of polymers using machine learning. Comp Mater. 2020;6(1):1–9.

[35] Cravero F, Martinez MJ, Ponroni I, et al. Computational modelling of mechanical properties for new polymeric materials with high molecular weight. Chemometr Intell Lab. 2019;193:103851.

[36] van Krevelen DW, Te Nijenhuis K. Properties of polymers: their correlation with chemical structure, their numerical estimation and prediction from additive group contributions. 4th ed. Amsterdam: Elsevier; 2009.

[37] Petrocuyo SRL. Polypropylene: distribution of molecular weights. Buenos Aires Argentina. [cited 2021 Oct 20]. Available from: http://www.petrocuyo.com/en/technology/polypropylene/molecular-weights

[38] Seymour RB, Carraher CE. Structure-property relationships in polymers. New York NY: Plenum Publishing Corporation; 1984.

[39] Im W, Sweeney J. thee. In: Ward M, Sweeney J, editors. An introduction to the mechanical properties of solid polymers. 2nd ed. England: John Wiley & Sons Ltd; 2004. p. 241–272.

[40] Tropsha A. Best practices for QSAR model development, validation, and exploitation. Mol Inform. 2010;29(6-7):476–488.

[41] Cherkasov A, Muratov EN, Fourches D, et al. QSAR modeling: where have you been? Where are you going to? J Med Chem. 2014;57(12):4977–5010.

[42] Muratov EN, Bajorath J, Sheridan RP, et al QSAR without borders. Chem Soc Rev. 2020;49(11):3525–3564.

[43] Keyvanpour MR, Shirzad MB. An analysis of QSAR research based on machine learning concepts. Curr Drug Disc Technol. 2020;18(1):17–30.

[44] Palomba D, Vazquez GE, Diaz MF. Prediction of elongation at break for linear polymers. Chemometr Intell Lab. 2014;139:121–131.

[45] Cravero F, Schustik S, Martinez MJ, et al. Feature selection and polydisperisty characterization for QSRR modelling: predicting a tensile property. In International Conference on Practical Applications of Computational Biology & Bioinformatics, Springer, Cham. 2018, 43–51.

[46] Cravero F, Schustik S, Martinez MJ, et al. Computer-aided design of polymeric materials: characterization of databases for prediction of mechanical properties under polydisperisty. Chemometr Intell Lab. 2019;191:65–72.

[47] Schustik S, Cravero F, Martinez MJ, et al. PolyMaS: a new software to generate high molecular weight polymer macromolecules from repeating structural units. Polimery. 2021;6(5):293–297.

[48] Cravero F, Martinez MJ, Vazquez GE, et al. Feature learning applied to the estimation of tensile strength at break in polymeric material design. J Integr Bioinf. 2016;13(2):15–29.

[49] Ward IM, Sweeney J. Mechanical properties of solid polymers. Bradford UK: John Wiley & Sons; 2012.

[50] Cravero F, Schustik S, Martinez MJ, et al. Feature selection for polymer informatics: evaluating scalability and robustness of the FS4RVDD algorithm using synthetic polydisperse datasets. J Chem Inf Model. 2020;60(2):592–603.

[51] Cravero F, Schustik S, Martinez MJ, et al. FS4RD: a feature selection algorithm for random variables with discrete distribution. Medina J, Ojeda-Aciego M, Verdegay J, et al editors In: 2018 International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems. Springer: Cham; 2018. Vol. 855, pp. 211–222.

[52] Bhattacharyya A. On a measure of divergence between two statistical populations defined by probability distributions. Bull Calcutta Math Soc. 1943;35:99.

[53] Hall M, Frank E, Holmes G, et al. The WEKA data mining software: an update. ACM SIGKDD Explor Newsl. 2009;11(1):10–18.