JCIM
JOURNAL OF
CHEMICAL INFORMATION
AND MODELING

Computational Chemistry

Feature Selection for Polymer Informatics: Evaluating Scalability and Robustness of the FS4RV_{DD} Algorithm using Synthetic Polydisperse Datasets

Fiorella Cravero, Santiago Schustik, María Jimena Martínez,
Gustavo Vazquez, Monica Fatima Diaz, and Ignacio Ponzoni

J. Chem. Inf. Model., **Just Accepted Manuscript** • DOI: 10.1021/acs.jcim.9b00867 • Publication Date (Web): 02 Dec 2019

Downloaded from pubs.acs.org on December 5, 2019

is published by the American Chemical Society. 1155 Sixteenth Street N.W.,
Washington, DC 20036

Feature Selection for Polymer Informatics: Evaluating Scalability and Robustness of the FS4RV$_{DD}$ Algorithm using Synthetic Polydisperse Datasets

*Fiorella Cravero[1], Santiago A. Schustik[1,2], M. Jimena Martínez[3], Gustavo E. Vázquez[4], Mónica F. Díaz[1,5], and Ignacio Ponzoni,[3,6]**

[1] Planta Piloto de Ingeniería Química, Universidad Nacional del Sur - CONICET. Camino La Carrindanga 7000, Bahía Blanca, Argentina.
[2] Comisión de Investigaciones Científicas de la Provincia de Buenos Aires (CIC), Argentina.
[3] Instituto de Ciencias e Ingeniería de la Computación (UNS-CONICET). San Andrés 800, Campus de Palihue, Bahía Blanca, Argentina.
[4] Facultad de Ingeniería y Tecnologías, Universidad Católica del Uruguay. Av. 8 de Octubre 2738, Montevideo, Uruguay.
[5] Departamento de Ingeniería Química (DIQ-UNS), Bahía Blanca, Argentina.
[6] Departamento de Ciencias e Ingeniería de la Computación, (DCIC-UNS), Bahía Blanca, Argentina.

* Author email address: ip@cs.uns.edu.ar

ABSTRACT. The feature selection (FS) process is a key step in the Quantitative Structure-Property Relationship (QSPR) modeling of physicochemical properties in Cheminformatics. In particular, the inference of QSPR models for polymeric material properties constitutes a complex problem because of the uncertainty introduced by the polydispersity of these materials. The main challenge is how to capture the polydispersity information from the molecular weight distribution (MWD) curve to achieve a more effective computational representation of polymeric materials. To date, most of the existing QSPR techniques use only a single molecule to represent each of these materials, but polydispersity is not considered. Consequently, QSPR models obtained by these approaches are being oversimplified. For this reason, we introduced in a previous work a new FS algorithm called Feature

Selection for Random Variables with Discrete Distribution (FS4RV$_{DD}$), which allows dealing with polydisperse data. In the present paper, we evaluate both the scalability and the robustness of the FS4RV$_{DD}$ algorithm. In this sense, we generated synthetic data by varying and combining different parameters: the size of the database, the cardinality of the selected feature subsets, the presence of noise in the data, and the type of correlation (linear and nonlinear). Moreover, the performances obtained by FS4RV$_{DD}$ were contrasted with traditional FS techniques applied to different simplified representations of polymeric materials. The obtained results show that the FS4RV$_{DD}$ algorithm outperformed the traditional FS methods in all proposed scenarios, which suggest the need of an algorithm such as FS4RV$_{DD}$ to deal with the uncertainty that polydispersity introduces in human-made polymers.

KEYWORDS. Feature Selection, Polymer Informatics, Synthetic Database, Polydisperse Data

BRIEF. Evaluating the FS4RV$_{DD}$ algorithm for polymer informatics using synthetic polydisperse datasets.

1. INTRODUCTION

Feature selection (FS) is the task of detecting a set of significant variables (or features) to define a computational model [1]. The main hypothesis to apply an FS method is that data usually contain a large number of irrelevant or redundant variables that can be eliminated without losing significant information. The avoidance of the course of dimensionality, the improvement in the interpretability of the models, and the reduction of overfitting and computational efforts during the training phase of a model are fundamental reasons for applying FS techniques in predictive modeling. Feature selection is a well-known combinatorial optimization problem widely studied for different real-world domains, which typically requires the designing of novel approaches to deal with emerging fields [2] or problems such as imbalanced databases [3].

In Cheminformatics, the area that interfaces between chemistry and computing [4], FS techniques are frequently applied in the initial stage of Quantitative Structure-Activity/Property Relationship (QSAR/QSPR) modeling [5]. QSAR/QSPR models are inferred to predict a biological activity or chemical property in terms of molecular descriptors. A molecular descriptor (MD) represents a portion of information derived of the chemical structures. QSAR/QSPR approaches are applied in order to estimate physicochemical and biological properties of interest, acting as drug prioritization strategy for pharma discovery [6, 7]. In this scenario, the process of choosing the most relevant MDs subgroup for the activity to be modeled is a particular case of the FS problem.

Material discovery is increasingly being impelled by machine learning (ML) methods that extract patterns from preexisting datasets [8]. A singularly challenging issue of QSPR modeling occurs in Polymer Informatics [9, 10], which is an interdisciplinary field that combines tools and knowledge from the computer science and polymer chemistry areas. The key aim behind it is to advance the understanding and design of new custom polymeric materials. This field is focused on the development of algorithms to polymer investigation by methodical computational studies mainly founded on ML algorithms as knowledge recovery techniques [11]. Polymer Informatics, like Cheminformatics, is mainly a design-oriented field; however, for polymer informaticians the modeling of chemical structures is much more complex and demanding in terms of computation than the chemicals studied in drug discovery projects [12]. In this respect, the Polymer Informatics demands a cautious modeling of macromolecules, where each of them is integrated by several large chain-like molecules that consist of many structural repetitive units (SRUs).

Significant efforts are continuously being made in Polymer Informatics as regard the QSPR modeling of polydisperse polymers. A polymeric material consists of various polymer chains of diverse lengths and molecular weights. Unlike a classic drug-like compound, a synthetic polymer is featured by a molecular weight distribution (MWD) rather than by a single molecular weight. In polymeric materials, this phenomenon is termed as polydispersity; it implies that each MD of a human-made polymer has a related distribution of values that could be generated by computing the descriptor value for the diverse polymeric chains and their frequencies [13].

A wide variety of FS methods have been proposed in the area of combinatorial optimization for a variety of application fields [1, 2, 14], including the MDs selection [15-18]. Most published studies use synthetic molecular models; that is, they characterize polymers through MDs calculated either on a single SRU [19-26] or on the central unit of the trimer [27, 28]. Nevertheless, polydispersity has not been considered in most contributions to QSPR modeling in Polymer Informatics [29] in which polymers were characterized by the use of their SRUs or their monomers. Therefore, MDs are only computed for a minimum representation of polymers, which constitutes a simplified view of a polymeric material. In a recently published work [30], MDs corresponding to SRU Molecular Weights (MW) and MDs associated with two representative average molecular weights of polymeric materials: Number Average Molecular Weight (Mn) and Weight Average Molecular Weight (Mw), were integrated into a single database to go through an FS process and train a QSPR model. In this way, the MDs of each polymer were characterized by three values: MW, Mn, and Mw (in Section 2.1, there is a description of this topic). The reported results confirm that predictive models inferred from databases that include several weight instances of polymer representations obtain better performances in terms of generalizability than predictive models generated from typical SRU-based representation databases. This characteristic constitutes a promissory antecedent that supports the experimental efforts of this paper to achieve a better understanding of the impact of polydispersity on the QSPR modeling of polymeric material properties.

In a previous work [13], we addressed the selection of the most relevant features in QSPR modeling in the context of polymeric materials. We presented a FS algorithm to deal with the uncertainty that introduces the polydispersity of polymeric materials. We generated synthetic data - created computationally instead of being taken from real-world events- to evaluate both the scalability and generalizability of the method. However, the experimentation was performed to conduct a proof of concepts. In this sense, the synthetic data generated did not include aspects such as the introduction of noise and the modeling of nonlinear relationships between the synthetic target and the MDs. Therefore, in the present paper, we present the following new contributions in terms of the experimental analysis of the proposed FS method:

1. The use of synthetic targets associated with MDs in both linear and nonlinear ways
2. The analysis of method robustness by introducing noise in the synthetic data.
3. Another approach to evaluate the statistical performance of the novel FS method, contrasting the new approach with the traditional FS algorithms applied to different simplified polymer representations.

In brief, we evaluate the performance of a novel method called Feature Selection Algorithm for Random Variables with Discrete Distribution ($FS4RV_{DD}$) [13], whose accuracy to detect those MDs randomly correlated with the synthetic target was obtained using classification metrics. $FS4RV_{DD}$ use a discrete distribution for represent the value of each MD, and this is the principal difference with traditional representation, which use a single value for each MD. This single value can be calculated or measured on representation of SRU, Mn, Mw or other single instance of weight (simplified representation). On the contrary, $FS4RV_{DD}$ allow to represent the polymer by multiples weights by using a discrete distribution (more details in the following sections).

In other words, we intend to answer the following research question: *given a probabilistic characterization of MDs by using discrete distributions in combination with the $FS4RV_{DD}$ method, is it possible to achieve a more accurate identification of the most relevant MDs than the one obtained from traditional approaches that use a simplified representation of MDs?*

2. MATERIALS AND METHODS

In this section, we present the generation of synthetic data used in the experiments, together with a summary of the $FS4RV_{DD}$ algorithm proposed in Cravero *et al.* [13]. Regarding data, we computed various sizes of synthetic databases. In addition, we explored two scenarios (with and without noise). Figure 1 shows the key steps followed for synthetic data generation, which will be detailed in the next subsections.

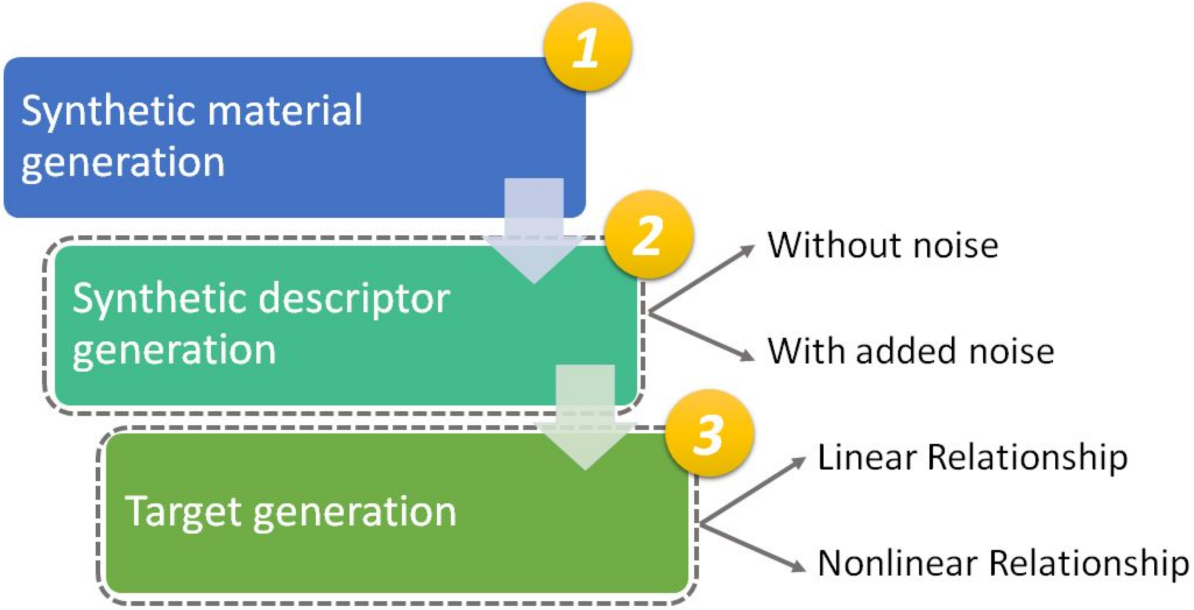


Figure 1. Main steps for the generation of synthetic data.

2.1. INTRODUCING THE POLYDISPERSITY PHENOMENON INTO THE MOLECULAR DESCRIPTOR

A key issue in Polymers Informatics is the lack of benchmark databases [31]. Moreover, the generation of a database with real values of MDs for polymers with high molecular weight is computationally expensive since the main variability factor (i.e., polydispersity) that characterizes a polymeric material database must be modeled. For this reason, the use of synthetic data becomes an advisable approach in this context. Additionally, the random generation of synthetic data is more suitable for scalability and robustness performance tests [32].

For the synthetic generation of these databases, it must be considered that polymeric materials have a molecular weight variability called polydispersity; that is, they do not have a single molecular weight (MW), but an MWD represented by an MWD curve. In most current databases and general papers in the literature, only average molecular weights are reported and not the corresponding MWD curves, which is why we had to model the curves of each polymer in our synthetic databases. The MWD curves are characteristic for each material, and they can follow, among others, a normal or a lognormal distribution, and it is crucial to model the polydispersity. In order to solve this problem, we followed a lognormal distribution because it is by far the most commonly observed distribution within asymmetric distributions and the most invoked one in the literature [33, 34]. To represent the polydispersity curve, each chain length under the curve must be correctly modeled; subsequently their MDs must be calculated for each of them. At present, it is not possible to perform this kind of computation with the available software tools for molecular modeling and MDs calculation [35, 36]. In this context, we propose a method for synthetic data generation, which poses a challenge and an opportunity of contribution.

Polymers consist of many SRUs. These SRUs are chemically bonded forming long chains [37]. As previously mentioned, the MW of a human-made polymer is not a single value like occurs with a small molecule. Polymeric materials are polydisperse since they are integrated by a distribution of chain lengths. Accordingly, they must be represented by an MWD curve. The typical MWD curve of a polymeric material is shown in Figure 2. The x-axis represents the different weights of molecular chains (weight), and the y-axis represents the molecular chain number of each weight of a polymer (frequency). As explained before, the lognormal distribution is the one typically used for polymeric materials and the most modeled one. Furthermore, it can describe a broad distribution with two parameters: number average molecular weight (Mn) and weight average molecular weight (Mw) [38], which results adequate for the kind of materials we study.

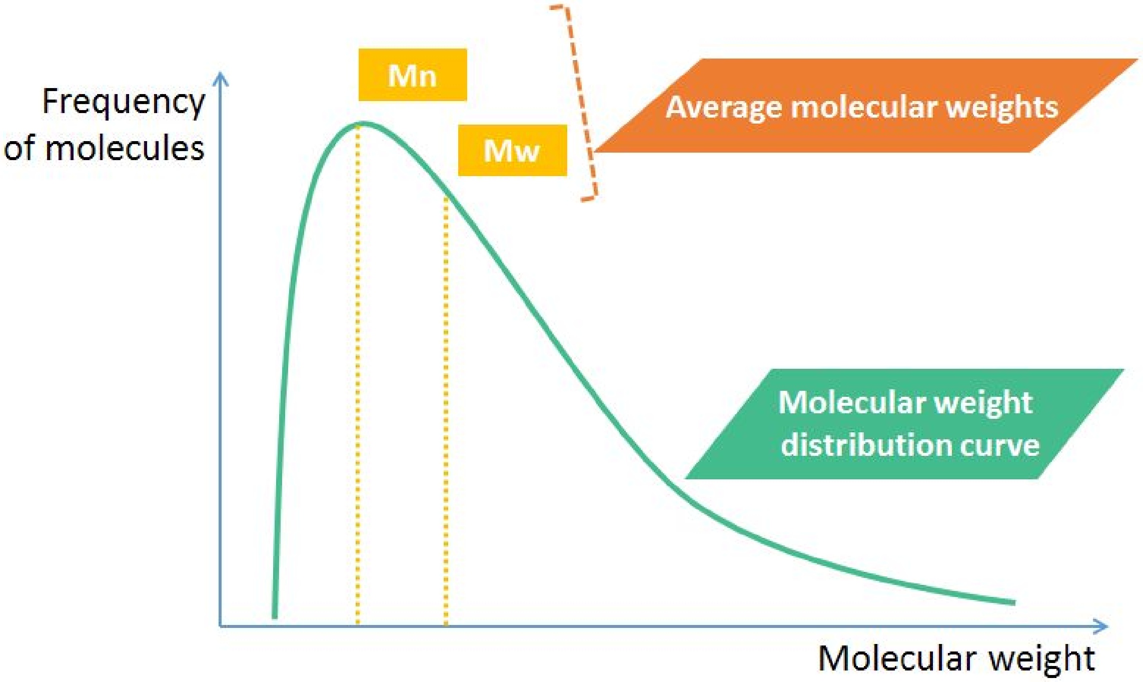

Frequency of molecules
Mn
Mw
Average molecular weights
Molecular weight distribution curve
Molecular weight

Figure 2. Graph of the classic molecular weight distribution (MWD) curve of a human-made polymer, where number average molecular weight (Mn) and weight average molecular weight (Mw) are denoted.

For a better understanding of this topic, it is necessary to know how Mn and Mw are calculated (eq. 1 and eq. 2, respectively).

$$Mn = \frac{\sum_i N_i M_i}{\sum_i N_i} \tag{1}$$

$$Mw = \frac{\sum_i N_i M_i^2}{\sum_i N_i M_i} \tag{2}$$

Where M_i is the molecular weight of one chain length and N_i is the number of chains of that molecular weight. In contrast to Mn, Mw denotes the molecular size instead of just an arithmetic mean [39].

Because polymer properties are dependent on MWD, it becomes crucial to know this feature [40, 41]. At this point, it is intuitive to infer that the descriptor values of each chain length of the MWD curve are needed for QSPR modeling.

Knowledge of the chain length of a polymer is required for the understanding of the polymer physical properties, such as ductility, brittleness and mechanical strength [42]. Because the properties of a polymer are dependent on its molecular weight distribution, the MDs involved in the QSPR model that predict these properties should be characterized considering the polydispersity of the polymer materials (lognormal distribution).

In this context, it was necessary to generate synthetic polymers characterized by this discretized lognormal distribution to obtain the synthetic database (see Fig. 3). This distribution has two parameters μ and σ that correspond to the mean and standard deviation, respectively. In this way, given two randomly-generated values for each polymeric material corresponding to the parameters of molecular weight distribution, it was possible to generate the curve for each polymer in each database (see Fig. 3, part A). Hence, databases of three different sizes were built for this work, namely, with 400, 800, and 1600 materials.

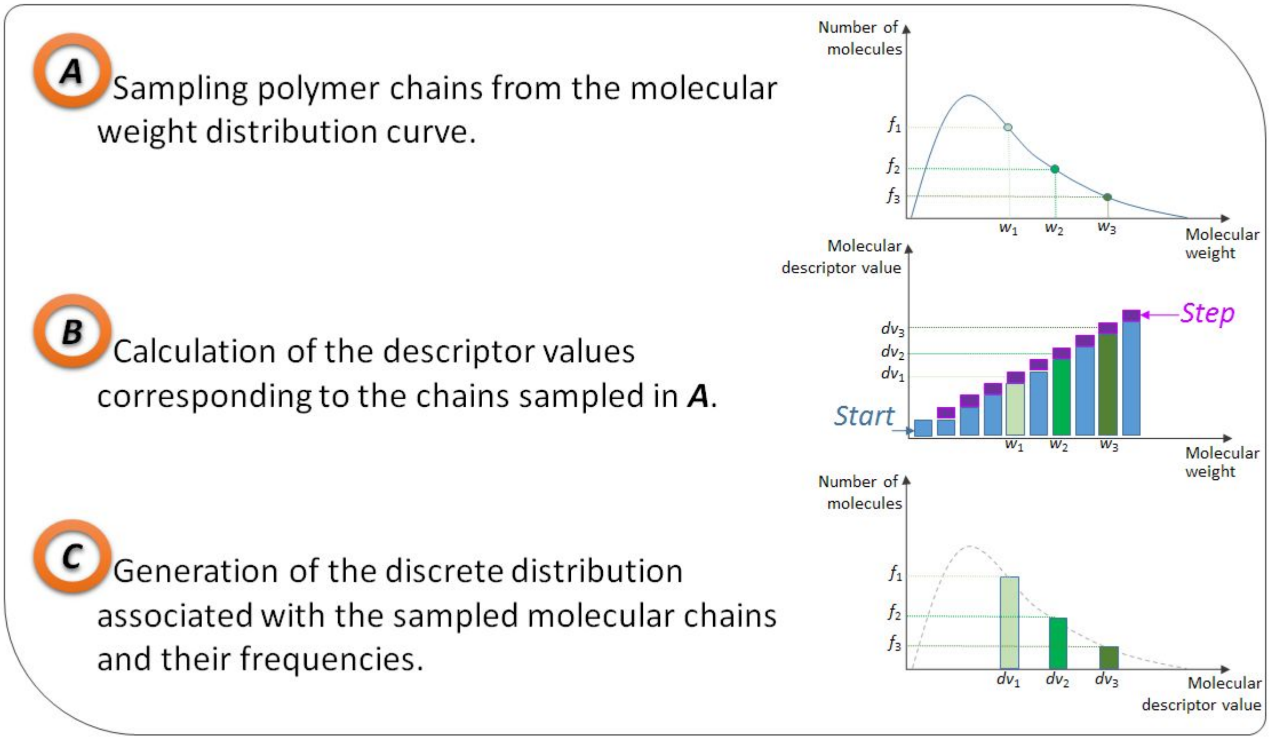

A
Sampling polymer chains from the molecular weight distribution curve.
Number of molecules
f_1
f_2
f_3
w_1
w_2
w_3
Molecular weight
B
Calculation of the descriptor values corresponding to the chains sampled in A.
Molecular descriptor value
Step
dv_3
dv_2
dv_1
Start
w_1
w_2
w_3
Molecular weight
C
Generation of the discrete distribution associated with the sampled molecular chains and their frequencies.
Number of molecules
f_1
f_2
f_3
dv_1
dv_2
dv_3
Molecular descriptor value

Figure 3. Description of steps for introducing the polydispersity phenomenon into the molecular descriptor characterization.

As a second stage, a vector of values for each descriptor was computed by generating $start$ and $step$, two random numbers. Then, the first value of each vector was $start$ and the following values were obtained by adding consecutively the $step$ value until to complete the vector. These observations represent the values of the MD incrementally computed for different molecular chain weights for a polymeric material (see Fig. 3, part B). Accordingly, a vector of coordinates for each material was saved in the material database; the first coordinate symbolizes the different molecular weights (x-axis) and the other one represents the frequencies (y-axis). Figure 3, part C shown this.

2.2. CONSTRUCTION OF THE SYNTHETIC DATABASE

The database was defined as a matrix in which rows are associated with polymeric materials and columns are associated with descriptors (see Fig. 4 part A). For each cell C_{ij}, the dispersion curve corresponding to the *i-th* material M_i was associated with the vector of values associated with the *j-th* descriptor D_j (see Fig. 4 part B). Then, a discrete distribution of values for each descriptorD_j, and for each material M_i could be obtained; the frequency of each descriptor value in the distribution was recovered from the dispersion curve. In particular, in this work, these distributions were obtained by taking 100 samples k from the molecular weight dispersion curves generated in the first stage (see Fig. 4 part C). Each sample corresponded to a molecular weight instance and matched with a descriptor value. Finally, a correlation between each sample k and the target values, expressed by the mean and the variance ($point\_estimator$ and $point\_estimator\_variance$ values), is necessary (see Fig. 4 part D). This is explained in detail in section 2.2.1. Note that different databases were generated by varying the number (400, 800, and 1600) of polymeric materials, but the total number of descriptors included in each database remained fixed at 100.

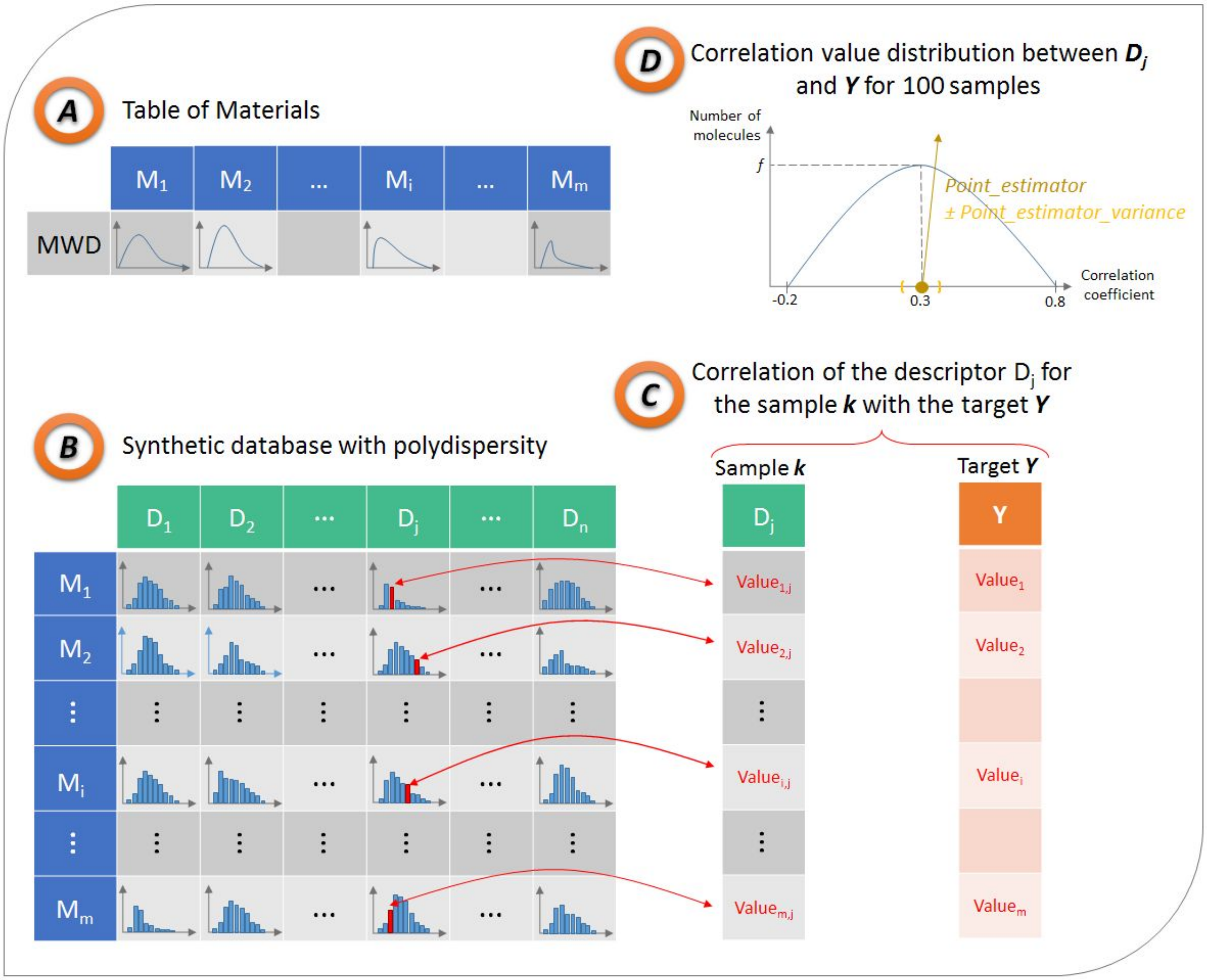


Figure 4. Graphical scheme of the conceptual construction of the database. Note that it contains polydisperse data and not a unique value.

2.2.1. BUILDING THE TARGET

The third stage is the target value calculation. To build the targets, we used the dispersity curves and the descriptor value vector generated in the previous steps. Firstly, MDs were randomly selected, generating three different scenarios by picking 5, 10, and 20 descriptors. Then, simple mathematical operations were used to correlate these descriptors with the target variable. Two scenarios were created: linear target and nonlinear target. In the linear scenario, the linear target (Y_L) was generated by adding the s values of the selected descriptors (eq. 3), where s is the total number of descriptors selected.

$$Y_L = \sum_{i=1}^{s} D_{p_i} \qquad (3)$$

On the other hand, the nonlinear target (Y_{nL}) was calculated by adding the product of selected descriptors taken in pairs (eq. 4). The sub-index in p are the positions of the chosen descriptor and *s* is the total number of descriptors involved in building the target. For example, when the descriptors randomly selected are five, the nonlinear target is $Y_{nL} = D_{p_1} * D_{p_2} + D_{p_3} * D_{p_4} + D_{p_5}$. The formula used for target building was inspired in the *label target functions* called *sum* and *nonlinear* available in RapidMiner [43] for synthetic data generation. The methodology for generating these two scenarios is shown in Figure 5.

$$Y_{nL} = \sum_{i=1}^{s} (D_{p_i} * D_{p_{i+1}}) + \ldots + (D_{p_{s-1}} * D_{p_s}) \qquad (4)$$

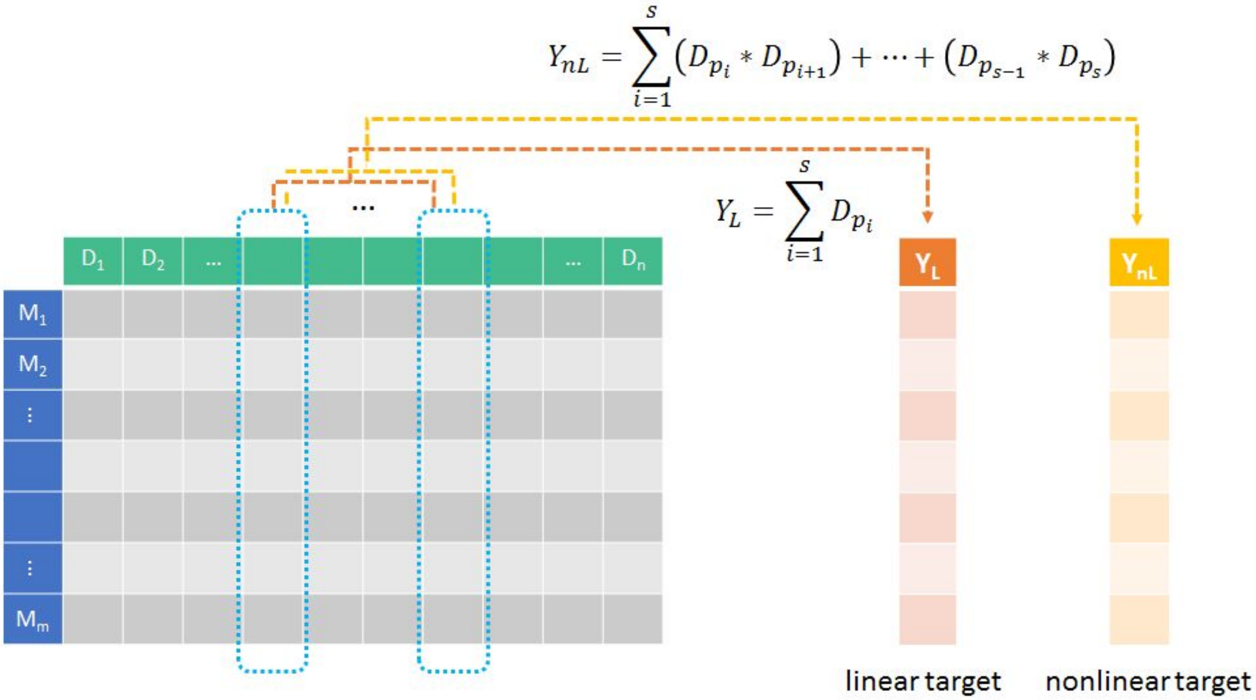

$Y_{nL} = \sum_{i=1}^{s} (D_{p_i} * D_{p_{i+1}}) + \cdots + (D_{p_{s-1}} * D_{p_s})$
$Y_L = \sum_{i=1}^{s} D_{p_i}$
...
D_1
D_2
...
...
D_n
M_1
M_2
⋮
⋮
M_m
Y_L
Y_{nL}
linear target
nonlinear target

Figure 5. Target generation for both linear and nonlinear targets.

At this point, we had databases with 400, 800, and 1600 polymeric materials, and we generated for each of them a target for each quantity of selected descriptors (5, 10, and 20). These combinations provided nine possible scenarios (400-5, 400-10, 400-20, 800-5, 800-10, 800-20, 1600-5, 1600-10, and 1600-20).

2.2.2. SCENARIOS WITH AND WITHOUT NOISE

Experiments were conducted for two scenarios: data with noise and data without noise (see Fig. 1). The noise scenario was generated using the jitter library of R [44]. For this scenario, we considered a random percentage of noise ranging ± 5%, simulating the error level within the instrument accuracy of ± 5% corresponding to a Waters Scientific Chromatograph model 150-CV (Size Exclusion Chromatography). This technique is applied to measure the MWD of polymers [45]. We added this noise percentage to the MD values. This process was performed for all polymeric materials in the database. Therefore, including these two new alternatives, we defined a combination of eighteen possible scenarios; that is, the nine scenarios mentioned in the previous section, but considering the options with noise and without noise.

2.3. FEATURE SELECTION FOR POLYDISPERSE DATA USING FS4RV$_{DD}$

The selection of the most relevant subset of MDs for a specific property is a crucial step in the inference of QSPR models. In the case of polymeric materials, this issue is a special and particular case of FS problems in which the variables are polydisperse data (i.e., they present uncertainty). In this sense, the FS4RV$_{DD}$ algorithm was implemented to deal with discrete distributions such as input variables. The method is summarized in this subsection (see Fig. 6), but a detailed presentation of the FS4RV$_{DD}$ can be found in Cravero *et al.* (2018) [13]. The algorithm consists of two key phases: the MDs ranking and the MDs elimination based on the correlation among them. The first phase is generated based on the linear correlations between MDs and target property (see Fig. 4 part C and Fig. 5). In the second one, MDs that are highly correlated with each other are eliminated. These two phases are detailed as follows.

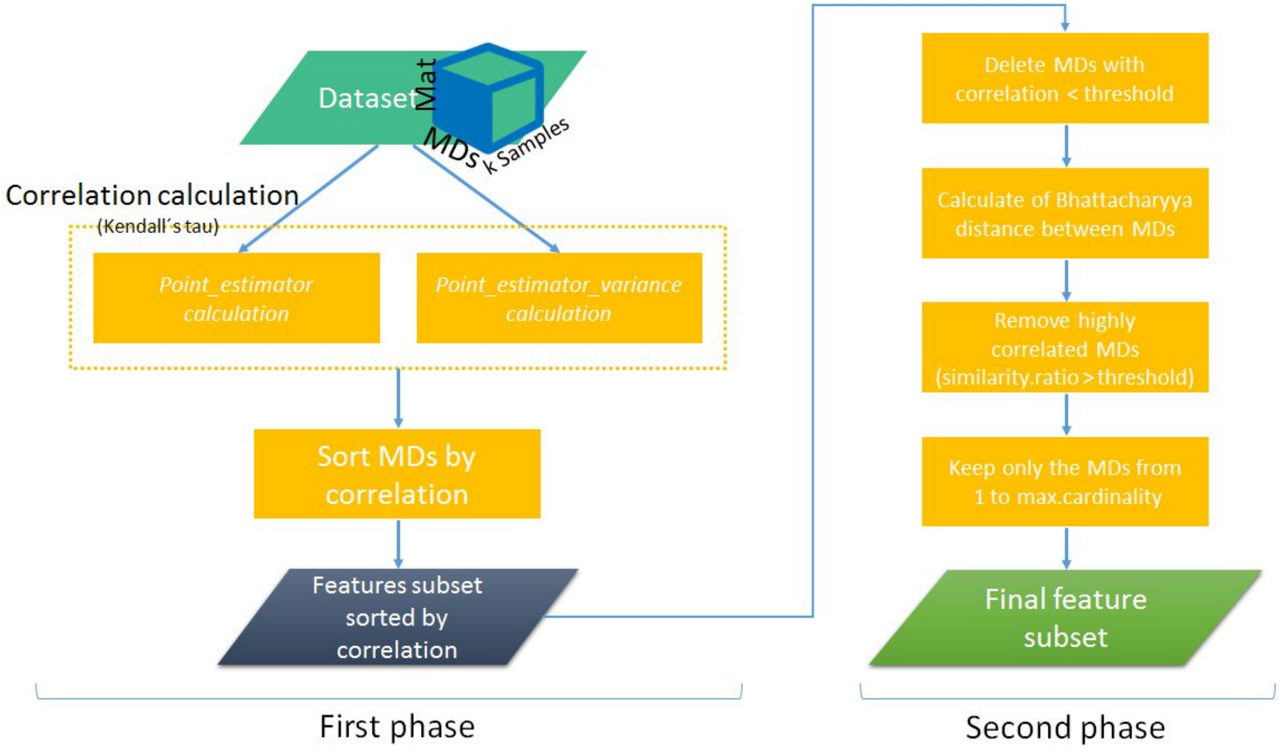

Dataset
Mat
MDs
k Samples
Correlation calculation
(Kendall´s tau)
Point_estimator calculation
Point_estimator_variance calculation
Sort MDs by correlation
Features subset sorted by correlation
Delete MDs with correlation < threshold
Calculate of Bhattacharyya distance between MDs
Remove highly correlated MDs (similarity.ratio > threshold)
Keep only the MDs from 1 to max.cardinality
Final feature subset
First phase
Second phase

Figure 6. Global methodology for our Feature Selection Method.

2.3.1 FIRST PHASE OF FS4RV$_{DD}$

The correlation between each MD and the target variable is measured. This correlation is computed using the Kendall's Tau distance [46]. Each descriptor D_j is represented by several samples k of the discrete distributions this descriptor has in association with each polymeric material M_i. Then, the method calculates the existing correlation between each sample k and the target values, obtaining a distribution of correlation values. Next, the method obtains a correlation measure for each descriptor expressed by the mean and the variance ($point\_estimator$ and $point\_estimator\_variance$ values) of this distribution of correlation between descriptor D_j and target Y calculated for $k = 100$ values (see Fig. 4 Part D). Finally, a descriptor ranking is generated, sorting them from the biggest to the smallest $point\_estimator$ values. This ranking is the output of the first phase.

2.3.2 SECOND PHASE OF FS4RV$_{DD}$

Those DMs that have a correlation value ($point\_estimator$) lower than a predefined threshold are eliminated and they are not part of the ranking (see Fig. 6). The objective here is to obtain a subgroup of MDs with a small level of redundancy. For this procedure, a pairwise contrast is performed with the remaining MDs in the ranking to identify MDs that have similar discrete distributions. The Bhattacharyya Distance [47] is applied to compare the distributions of each pair of MDs for each material. If two MDs present a higher degree of similarity than the threshold, the DM located further down the ranking is eliminated. Finally, the final subgroup of selected descriptors is defined considering the maximum cardinality that user had previously settled.

2.3.3 PERFORMANCE EVALUATION

In a previous work [13], the FS4RV$_{DD}$ method was evaluated by generating QSPR models from the selected variables and discussing the accuracy of these models as an undirected way for assessing the quality of the FS output. In the present paper, we sought to evaluate the FS method using a more straightforward approach instead of computing QSPR models. The idea was to determine the capability of the FS method for detecting the MDs that had been formerly correlated with the target variables in the generation process of the artificial datasets. In this way, this new approach for the performance evaluation of the FS4RV$_{DD}$ method can be analyzed as a classification problem in which the FS selection algorithm classifies the MDs into two classes: correlated and not correlated with the target variable. Therefore, the metrics used for assessing the method performance can be the ones typically used in binary classification: Percentage of Correctly Classified (%CC), also known as "Accuracy" in Machine Learning community, or the Non-Error Rate (NER). The %CC represents the number of correct predictions obtained on the total number of samples. On the other hand, the value of NER is the arithmetic mean of class sensitivity, that is, the average of the percentages of samples that were correctly classified for each class.

An additional challenge is to define a fair experimental framework for performance comparisons with other approaches, because there are no similar methods that work with descriptors characterized by discrete distributions. For this reason, we decided to compare our method with state of the art FS methods using representations based on a single value. In polymer modeling, representations based on SRU [18-29] and representations based on average weight values (i.e., Mn and Mw) [30, 35, 48] are the most reasonable and sane single value representations. Therefore, we decided to use the smallest value of the MWD of a polymer as an analogue for the SRU-based representation of this polymer. Similarly, we decided to use the average value of the MWD of a polymeric material as an analogue for the Mn-based representation of this material. In this way, two additional databases with single value representations of MDs could be defined: one of them considering only the first value of the discrete distribution as the descriptor value and another one considering only the average value of discrete distribution as the descriptor value.

Regarding the FS methods used for these single value representations, we worked with the Weka tool [49]. We used the *AttributesSelection* method included in Weka, where the *attributes* correspond to the MD of our databases. It provides a fair and direct comparison with respect to our FS method in terms of classification metrics. In particular, we used the following parameters: *CorrelationAttributeEval* as Attribute Evaluator and Ranker as Search Method. *CorrelationAttributeEval* evaluates the worth of an attribute by measuring its Pearson's correlation with the target and Ranker builds a rank among the attributes considering their individual evaluations.

3. RESULTS AND DISCUSSIONS

In this section, the experiment results are discussed in terms of classification metrics. The variables (MDs) selected by the FS method that matched those that had been initially selected to create the target property were labeled as True Positive (TP), whereas those MDs that were incorrectly selected by the method (i.e., those that were not part of the group considered to build the target) were labeled as False

Positive (FP). At this point, it is important to remember that three synthetic databases integrated for 400, 800, and 1600 polymeric materials had been created. Each database contained one hundred MDs and different scenarios had been defined by changing the conditions for creating the target variables. Targets had been generated by varying the number of MDs randomly selected (5, 10, or 20), changing the type of correlation (linear and nonlinear), and adding (or not) noise to the data. Therefore, the combinations of these experimental conditions provided 36 different scenarios. Additionally, we computed different metrics that will be explained in other subsections.

3.1 PERCENTAGE OF DESCRIPTORS CORRECTLY CLASSIFIED BY THE FS4RV$_{DD}$ METHOD

Figure 7 shows the percentage of MDs correctly classified (%CC) for the noiseless scenario. Regarding the obtained results, the algorithm accuracy increased as the database size (number of polymers) increased, but it decreased when the number of MDs used for creating the synthetic targets increased. These results correspond to the expected performance behavior. When comparing the results obtained for databases with equal sizes, we can conclude that misclassifications increased when more MDs had to be retrieved. On the other hand, the performances improved when the databases were bigger, which is a logical consequence of the increase in the number of polymers available for detecting the MDs correlated with the target. In other words, when more data were available for the MDs and the targets, it was easier to detect a pattern among them. Regarding the type of correlations, the performance behaviors were quite similar.

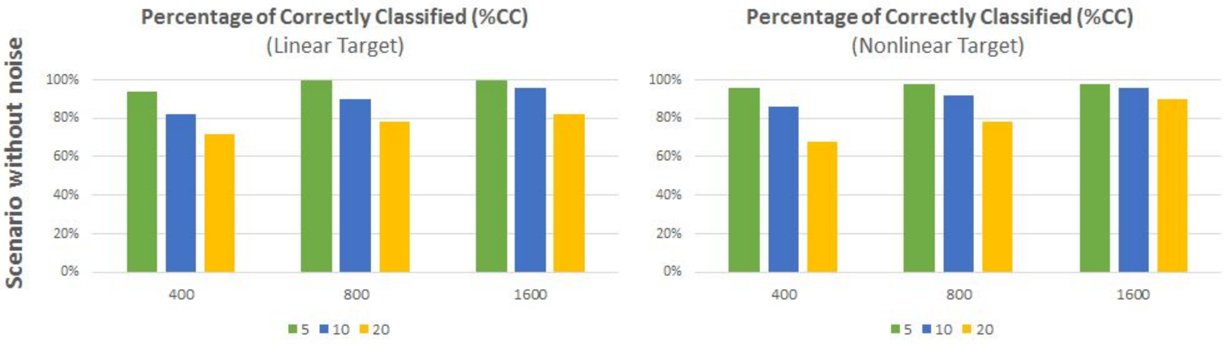


Figure 7. Percentage of MDs correctly classified by $FS4RV_{DD}$ for noiseless experiments.

Incorrect experimental data hinders improvement of QSAR models; counterintuitively, adding noise can improve their performance in some adaptive algorithms [50]. Figure 8 shows that the behavior in terms of %CC had similar tendencies to the ones reported for the noiseless scenario. The only difference was a slight performance improvement in general terms. The phenomenon in which a signal that is normally too weak to be detected by a sensor can be boosted by adding white noise to the signal is known as stochastic resonance. A similar effect could happen in this case; as data were sufficiently similar, by adding noise the method could better detect the different features (MDs).

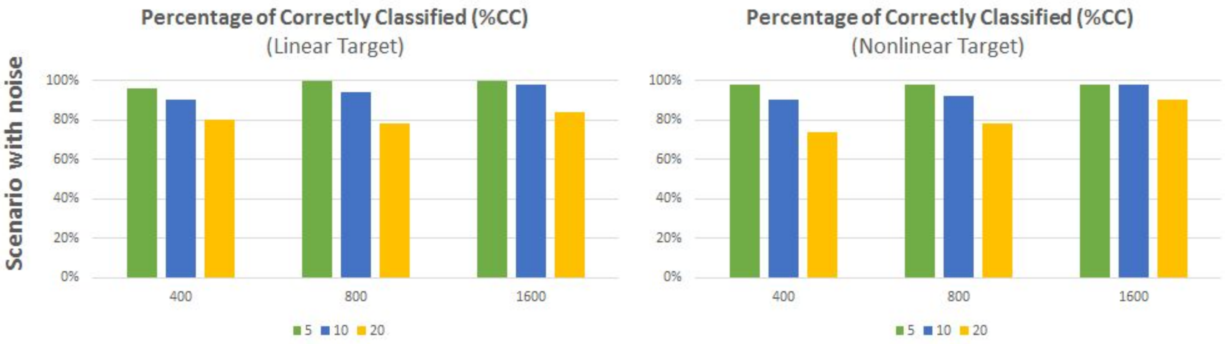


Figure 8. Percentage of MDs correctly classified by $FS4RV_{DD}$ for experiments with noise.

3.2 NON-ERROR RATE VALUES OBTAINED BY THE $FS4RV_{DD}$ METHOD

Percentage classification accuracy is not always enough to evaluate a model when some imbalances occur between classes because its value will be biased to the most numerous class. In our experiments, in the scenarios in which the number of MDs to be selected was low (5 or 10), the unbalance was relatively high. A recommended measure to correct these situations is the Non-Error Rate (NER) [51].

The following graphs show the results obtained for the Non-Error Rate for both scenarios: without noise (Fig. 9) and adding noise (Fig. 10). For the smallest databases, in all scenarios, the performance decreased. In the remaining cases, the performance measures also had a lower decrement. This general reduction in terms of the percentages reported by NER values in comparison with the %CC values is reasonable because NER metric makes corrections in the presence of unbalanced samples and %CC does not. Note that, as it had happened with the %CC results, the addition of noise improved the NER values.

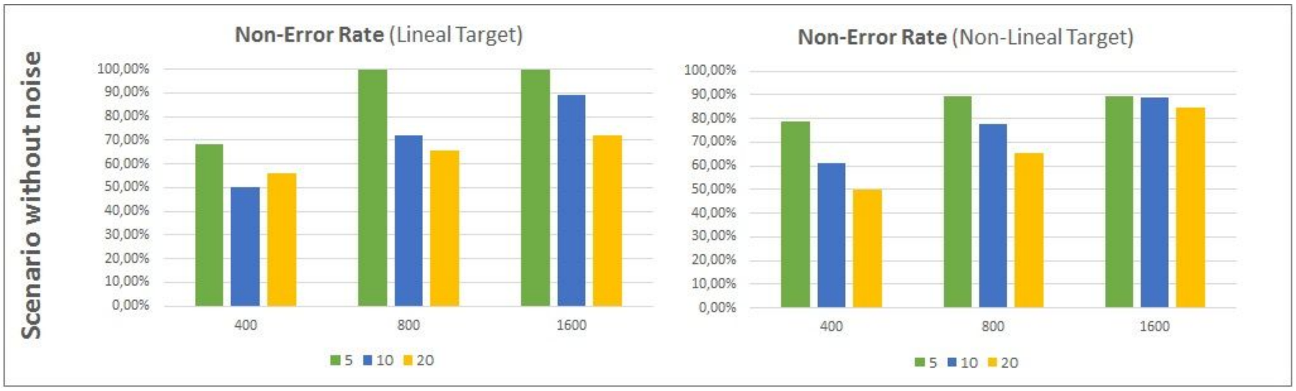


Figure 9. Non-Error Rate measures obtained by FS4RV$_{DD}$ for noiseless experiments.

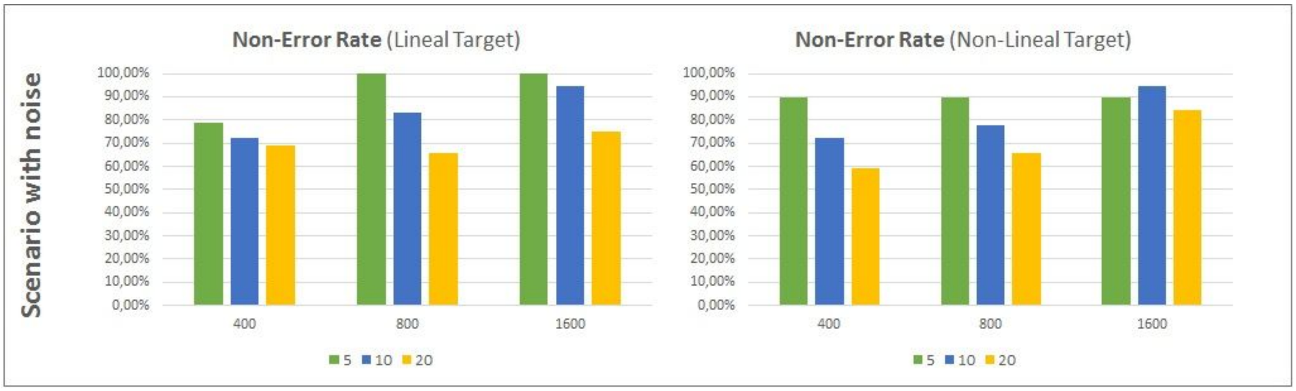


Figure 10. Non-Error Rate measures obtained by FS4RV$_{DD}$ for experiments with noise.

3.3 COMPARING THE FS4RV$_{DD}$ METHOD WITH OTHER APPROACHES

Once evaluated the performance of our method (FS4RV$_{DD)}$, it is time to compare it with the state of the art methods. In polymer informatics, the SRU-based representation (or monomers) is typically used to make QSPR models. As previously explained, the lowest value of the reported weight distribution curve is used as an analogue for SRU. Figure 11 shows the results for the Non-Error Rate for the lowest value (analogous to SRU) for noiseless scenarios and Figure 12 shows the results for scenarios with noise. The analysis of these results becomes a chance to answer our research question from the synthetic data generated for this work. It is evident that the performance dropped abruptly in contrast with the results achieved by FS4RV$_{DD}$. In addition, it did not show a notorious improvement when noise was added, as it happened for the representation that considered the whole polydispersity curve. For this reason, we can conclude that the SRU-based representation did not contain enough information to capture the structural complexity of polydispersity, making it difficult to identify the most relevant MDs for the target variable under study.

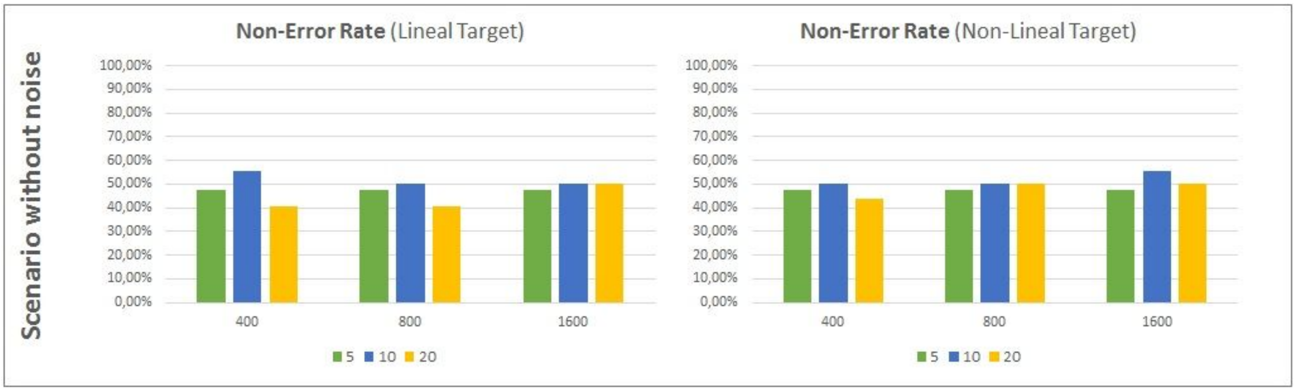


Figure 11. Non-Error Rate obtained by a traditional FS method using Lowest Value based-representations (analogous to the SRU-based representation) for noiseless experiments.

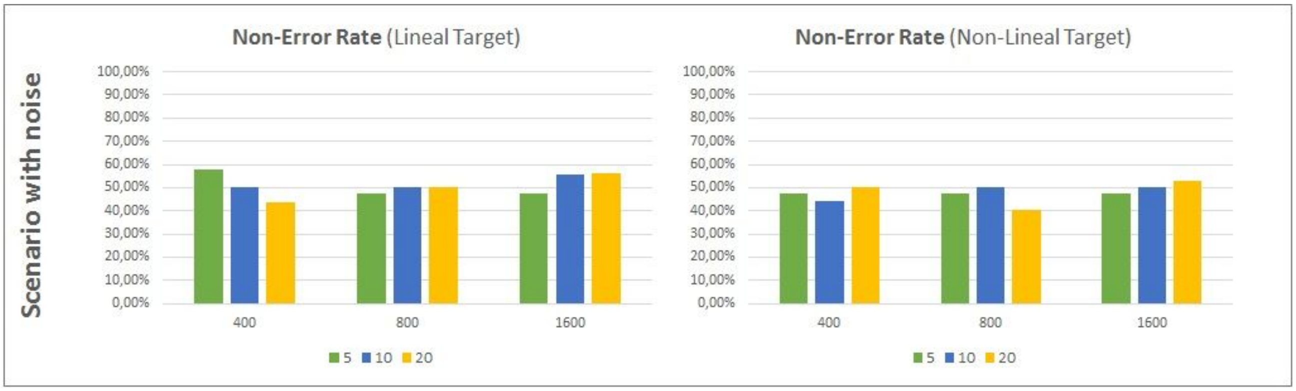


Figure 12. Non-Error Rate obtained by a traditional FS method using Lowest Value based-representations (analogous to the SRU-based representation) for experiments with noise.

Finally, our group considered the average molecular weights as the representative instance of polydisperse materials [30]. To represent this instance, the mean value of the polydispersity curve was taken into account. In general, the NER values for noiseless scenarios shown in Figure 13 and the ones for scenarios without noise shown in Figure 14 were larger than the ones reported in Figure 11 and 12. In fact, they are more similar to the results obtained by FS4RV$_{DD}$. As the average descriptor values are more informative of polydispersity than the lower ones, it was expected that the performance showed improvements. Therefore, there was no disagreement between the obtained results and the expected ones. Neverthless, in the noiseless scenarios for nonlinear targets, the NER values were low, comparable to those obtained for Lowest Value based-representations. An explanation of these results could be that a simplified representation that summarizes the MDs distributions in average values, as it occurred in this case, might require more effort to find more complex targets (nonlinear scenario).

From these experiments, we can conclude that the representation based on average molecular weights could be more suitable than representations based on SRU in terms of the results obtained for these analogue synthetic representations. Nevertheless, if we compare the performances achieved by both single value based-representations with the results obtained by the FS4RV$_{DD}$ method, it is clear that the characterization of MDs by means of discrete distributions is the best representation to capture polydispersity.

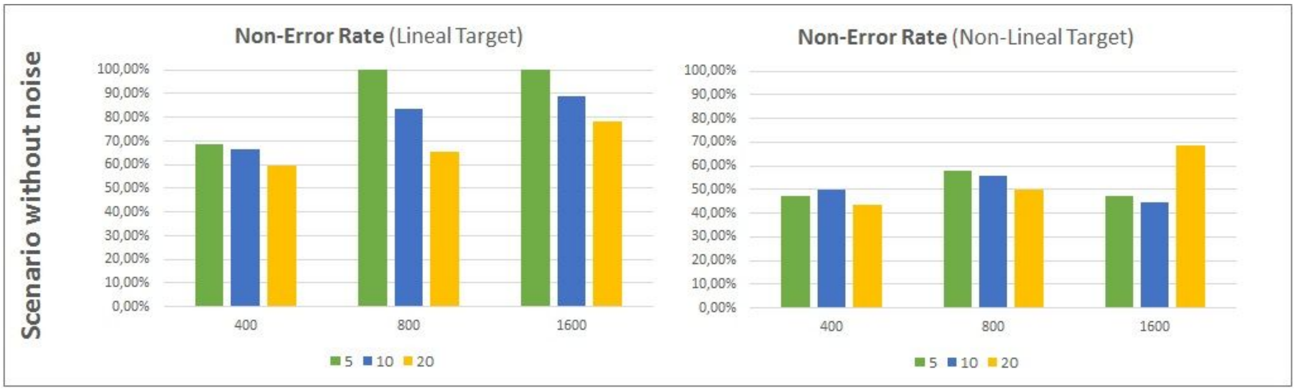


Figure 13. Non-Error Rate obtained by a traditional FS method using Mean Value based-representations (analogous to the average weight-based representation) for noiseless experiments.

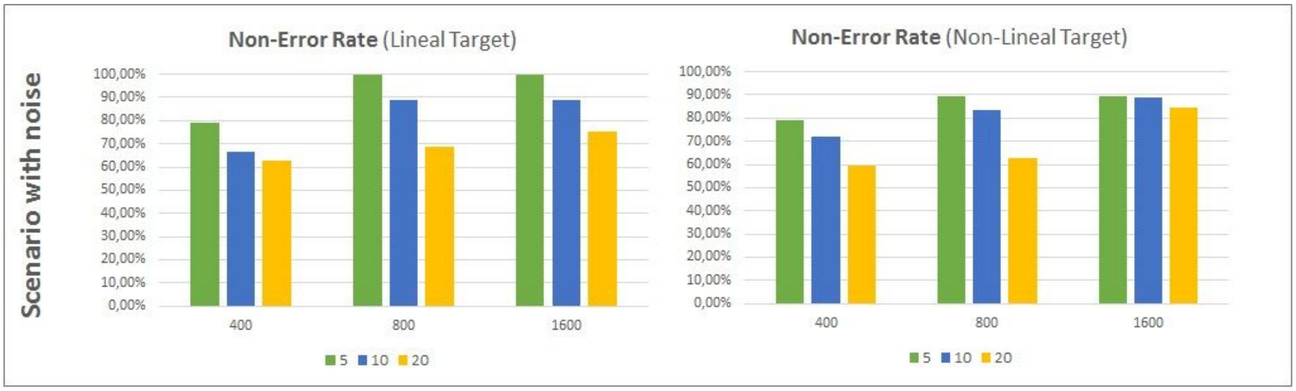


Figure 14. Non-Error Rate obtained by a traditional FS method using Mean Value based-representations (analogous to the average weight-based representation) for experiments with noise.

Another correlation study was relevant to exclude the possibility that the MDs wrongly selected by the method (false positives) were highly correlated with those that were not selected by mistake (false negatives). If false positives and false negatives correspond to well correlated MDs, this means that the method did not make a serious mistake during the selection process; it was only replacing the correct selections with good alternatives, which invalidates our previous analyses and discussions. This correlation study allowed us to discard this concern, because in all cases the correlations between false positives and false negatives were low. The complete report of this study is included as Supporting Information.

As a summary, we can conclude that the representation based on the whole polydispersity curve always achieved competitive results for all scenarios (note that these models included frequency information). The other single value based-representations (lowest and mean) had poor performances in most cases. Therefore, to answer the formulated research question, it can be said that the characterization of polydispersity by the probabilistic distributions derived from the MDs from the curve and their frequencies allow a better selection of the variables involved in the target construction than other instances of representation with less associated information.

4. CONCLUSIONS

In this work, we present a stringent performance evaluation of an FS algorithm for discrete variables with uncertainty called $\mathrm{FS4RV_{DD}}$, designed for addressing a challenging application in the area of QSPR for Polymer Informatics. This new experimental study required the generation of synthetic data to emulate the polydispersity phenomenon, which plays a central role in the design of new polymeric materials. In this regard, we proposed several scenarios considering different numbers of materials (400, 800, and 1600) and selected MDs (5, 10, and 20), the kind of correlations with the target (linear and nonlinear), and the presence (or not) of noise in the data.

The performance achieved by $\mathrm{FS4RV_{DD}}$ was contrasted with traditional FS approaches used in Polymer Informatics, in which polymers are typically characterized by simplified representations. In particular, the main goal of this study was to respond to the following question: *given a probabilistic characterization of MDs by using discrete distributions in combination with the $FS4RV_{DD}$ method, is it possible to achieve a more accurate identification of the most relevant MDs than the one obtained from traditional approaches using a simplified representation of MDs?*

The analysis of the experiment results provides evidence to answer this question affirmatively, because in all scenarios the $\mathrm{FS4RV_{DD}}$ algorithm outperformed, or at least achieved similar performance, in comparison with the other FS alternatives. In this sense, the motivating idea behind this paper is to reinforce the arguments about the need of a special FS algorithm designed for Polymer Informatics, since the uncertainty in the polymerization process requires specialized algorithms for representing and dealing with this phenomenon. Since the market demand for custom polymeric materials is growing, we hope that the results of this work show an open field of research and encourage the possibility of new design alternatives for the production of novel industrial polymeric materials.

ACKNOWLEDGMENTS. This work was partially supported by the Argentinean National Council of Scientific and Technological Research (CONICET for its acronym in Spanish) [grant PIP 112-2012-0100471] and by the Universidad Nacional del Sur (UNS); Bahía Blanca, Argentina, [grants PGI 24/N042 and PGI 24/ZM17].

REFERENCES

1 Li, Y., Li, T., & Liu, H. (2017). Recent Advances in Feature Selection and its Applications. *Knowledge and Information Systems*.

2 Tommasel, A., & Godoy, D. (2018). A Social-aware online short-text Feature Selection Technique for Social Media. *Information Fusion*.

3 Klimenko, K., Rosenberg, S. A., Dybdahl, M., Wedebye, E. B., Nikolov, N. G. (2019). QSAR Modelling of a Large Imbalanced Aryl Hydrocarbon Activation Dataset by Rational and Random Sampling and Screening of 80,086 REACH Pre-registered and/or Registered Substances. *PloS one*.

4 de Souza, J. C. S., Claudino, S. G., da Silva Simões, R., Oliveira, P. R., Honório, K. M. (2016, July). Recent Advances for Handling Imbalancement and Uncertainty in Labelling in Medicinal Chemistry Data Analysis. In *2016 SAI Computing Conference (SAI)* IEEE.

5 Eklund, M., Norinder, U., Boyer, S., Carlsson, L. (2014). Choosing Feature Selection and Learning Algorithms in QSAR. *J. Chem. Inf. Model*.

6 Li, J., Fong, S., Siu, S., Mohammed, S., Fiaidhi, J., Wong, K. K. (2016). WITHDRAWN: Improving Classification of Protein Binders for Virtual Drug Screening by Novel Swarm-based Feature Selection Techniques. *Computerized Medical Imaging and Graphics*.

7 Ponzoni I., Sebastián-Pérez V., Requena-Triguero C., Roca C., Martínez M.J., Cravero F., Díaz M.F., Páez J.A., Gómez Arrayás R., Adrio J., Campillo N.E. (2017). Hybridizing Feature Selection and Feature Learning Approaches in QSAR Modeling for Drug Discovery. *Sci. Rep.*

8 Jennings, P. C., Lysgaard, S., Hummelshøj, J. S., Vegge, T., Bligaard, T. (2019). Genetic Algorithms for Computational Materials Discovery Accelerated by Machine Learning. *npj Comput. Mater.*

9 Adams, N. (2010). Polymer informatics. In *Polymer Libraries*. Springer, Berlin, Heidelberg.

10 Audus, D. J., & de Pablo, J. J. (2017). Polymer informatics: Opportunities and Challenges. *ACS Macro Lett.*

11 Liu, Y., Zhao, T., Ju, W., Shi, S. (2017). Materials Discovery and Design using Machine Learning. *J. Materiomics*.

12 Huan, T. D., Mannodi-Kanakkithodi, A., Kim, C., Sharma, V., Pilania, G., Ramprasad, R. (2016). A Polymer Dataset for Accelerated Property Prediction and Design. *Sci. Data*.

13 Cravero F., Schustik S., Martínez M.J., Díaz M.F., Ponzoni I. (2018). FS4RV$_{DD}$: A Feature Selection Algorithm for Random Variables with Discrete Distribution. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*. Springer, Cham.

14 Singh, R. K., & Sivabalakrishnan, M. (2015). Feature Selection of Gene Expression Data for Cancer Classification: a Review. *Procedia Computer Science*.

15 Soto A.J., Cecchini R.L., Vazquez G.E., Ponzoni I. (2008). A Wrapper-Based Feature Selection Method for ADMET Prediction Using Evolutionary Computing. In *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*. Lecture Notes in Computer Science.

16 Soto A.J., Cecchini R.L., Vazquez G.E., Ponzoni I. (2009). Multi-Objective Feature Selection in QSAR Using a Machine Learning Approach. *Molecular Informatics*.

17 Martínez, M. J., Ponzoni, I., Díaz, M. F., Vazquez, G. E., Soto, A. J. (2015). Visual Analytics in Cheminformatics: user-supervised Descriptor Selection for QSAR Methods. *J. Cheminf*.

18 Cravero, F., Martínez, M.J., Vazquez, G.E., Díaz, M.F., Ponzoni, I. (2016). Feature learning Applied to the Estimation of Tensile Strength at Break in Polymeric Material Design. *J. Iintegrative Bioinformatics*.

19 Cao, C., & Lin, Y. (2003). Correlation between the Glass Transition Temperatures and Repeating Unit Structure for High Molecular Weight Polymers. *J. Chem. Inf. Comput. Sci.*

20 Duce, C., Micheli, A., Starita, A., Tiné, M. R., Solaro, R. (2006). Prediction of Polymer Properties from their Structure by Recursive Neural Networks. *Macromol. Rapid Commun*.

21 Yu, X. L., Yi, B., & Wang, X. Y. (2008). Prediction of the Glass Transition Temperatures for Polymers with Artificial Neural Network. *J. Theor. Comput. Chem*.

22 Liu, W., & Cao, C. (2009). Artificial Neural Network Prediction of Glass Transition Temperature of Polymers. *Colloid Polym. Sci*.

23 Toropova, A. P., Toropov, A. A., Kudyshkin, V. O., Leszczynska, D., Leszczynski, J. (2014). Optimal Descriptors as a Tool to predict the Thermal Decomposition of Polymers. *J. Math. Chem*.

24 Jabeen, F., Chen, M., Rasulev, B., Ossowski, M., Boudjouk, P. (2017). Refractive Indices of Diverse Data Set of Polymers: A Computational QSPR based Study. *Comput. Mater. Sci.*

25 Chen, M., Jabeen, F., Rasulev, B., Ossowski, M., Boudjouk, P. (2018). A Computational Structure–Property Relationship Study of Glass Transition Temperatures for a Diverse Set of Polymers. *J. Polym. Sci., Part B: Polym. Phys.*

26 Cravero, F., Martínez, M. J., Ponzoni, I., Díaz, M. F. (2019). Computational Modelling of Mechanical Properties for New Polymeric Materials with High Molecular Weight. *Chemom. Intell. Lab. Syst.*

27 Katritzky, A. R., Lan, X., Yang, J. Z., Denisko, O. V. (1998). Properties and Synthetic Utility of N-substituted Benzotriazoles. *Chemical reviews*, *98*(2), 409-548.

28 Palomba, D., Vazquez, G. E., & Díaz, M. F. (2012). Novel Descriptors from Main and Side Chains of High-Molecular-Weight Polymers Applied to Prediction of Glass Transition Temperatures. *J. Mol. Graphics Modell.*

29 Wu, K., Sukumar, N., Lanzillo, N. A., Wang, C., Ramprasad, R., Ma, R., Baldwin, A. F., Sotzing, G., Breneman, C. (2016). Prediction of Polymer Properties using Infinite Chain Descriptors (ICD) and Machine Learning: Toward Optimized Dielectric Polymeric Materials. *J. Polym. Sci., Part B: Polym. Phys*.

30 Cravero, F., Schustik, S. A., Martínez, M. J., Barranco, C. D., Díaz, M. F., Ponzoni, I. (2019). Computer-Aided Design of Polymeric Materials: Computational Study for Characterization of Databases for Prediction of Mechanical Properties under Polydispersity. *Chemom. Intell. Lab. Syst*.

31 Ma, R., Liu, Z., Zhang, Q., Liu, Z., Luo, T. (2019). Evaluating Polymer Representations via Quantifying Structure-Property Relationships. *J. Chem. Inf. Model.*

32 Brinkhoff T. (2009) *Real and Synthetic Test Datasets*. In: LIU L., ÖZSU M.T. (eds) Encyclopedia of Database Systems. Springer, Boston, MA.

33 Lorenzini, P., Pons, M., & Villermaux, J. (1992). Free-Radical Polymerization Engineering—III. Modelling Homogeneous Polymerization of Ethylene: Mathematical Model and New Method for Obtaining Molecular-Weight Distribution. *Chem. Eng. Sci.*

34 Brandolin, A., Lacunza, M. H., Ugrin, P. E., Capiati, N. J. (1996). High-Pressure Polymerization of Ethylene. An Improved Mathematical Model for Industrial Tubular Reactors. *Polym. React. Eng.*

35 Cravero, F., Schustik, S., Martínez, M.J., Ponzoni, I., Díaz, M.F. (2017). Macro Approach to Molecular Modelling of Linear Polymers Applied to Estimation of Tensile Modulus for New Materials Development. In *VIII International Symposium on Materials*.

36 Cravero F., Martínez M.J., Vazquez G.E., Ponzoni I., Díaz M.F. (2016). Representación de la Estructura Molecular de Polímeros Sintéticos de Alto Peso. In *XXXI Congreso Argentino de Química*.

37 McCrum N.G., Buckley C.P., & Bucknall C.B. (1997). *Principles of polymer engineering*. Oxford; New York: Oxford University Press.

38 Monteiro, M. J. (2015). Fitting Molecular Weight Distributions Using a Log-Normal Distribution Model. *Eur. Polym. J.*

39 Chanda M. (2013). *Introduction to Polymer Science and Chemistry: A Problem-Solving Approach*, CRC Press (Eds.).

40 Martin, J. R., Johnson, J. F., & Cooper, A. R. (1972). Mechanical Properties of Polymers: the Influence of Molecular Weight and Molecular Weight Distribution. *J. Macromol. Sci., Rev. Macromol. Chem.*

41 Dobkowski, Z. (1981). General Approach to Polymer Properties Dependent on Molecular Characteristics. *Eur. Polym. J.*

42 Sheu, W. S. (2001). Molecular Weight Averages and Polydispersity of Polymers. *J. Chem. Educ.*

43 Mierswa, I., & Klinkenberg, R. (2018). RapidMiner Studio (9.1) [Data science, machine learning, predictive analytics]. Retrieved from [https://rapidminer.com/]

44 R Core Team (2017). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from [https://www.R-project.org]

45 Pantano, I. A. G., Díaz, M. F., Brandolin, A., Sarmoria, C. (2009). Mathematical Modeling of the Catalytic Degradation of Polystyrene in the Presence of Aluminum Chloride. *Polym. Degrad. Stab.*

46 McLeod, A.I. (2015). Package 'Kendall'. Retrieved from [https://cran.r-project.org/web/packages/Kendall/Kendall.pdf]

47 Bhattacharyya A., (1943). On a Measure of Divergence Between two Statistical Populations Defined by Probability Distributions. *Bull. Calcutta Math. Soc*.

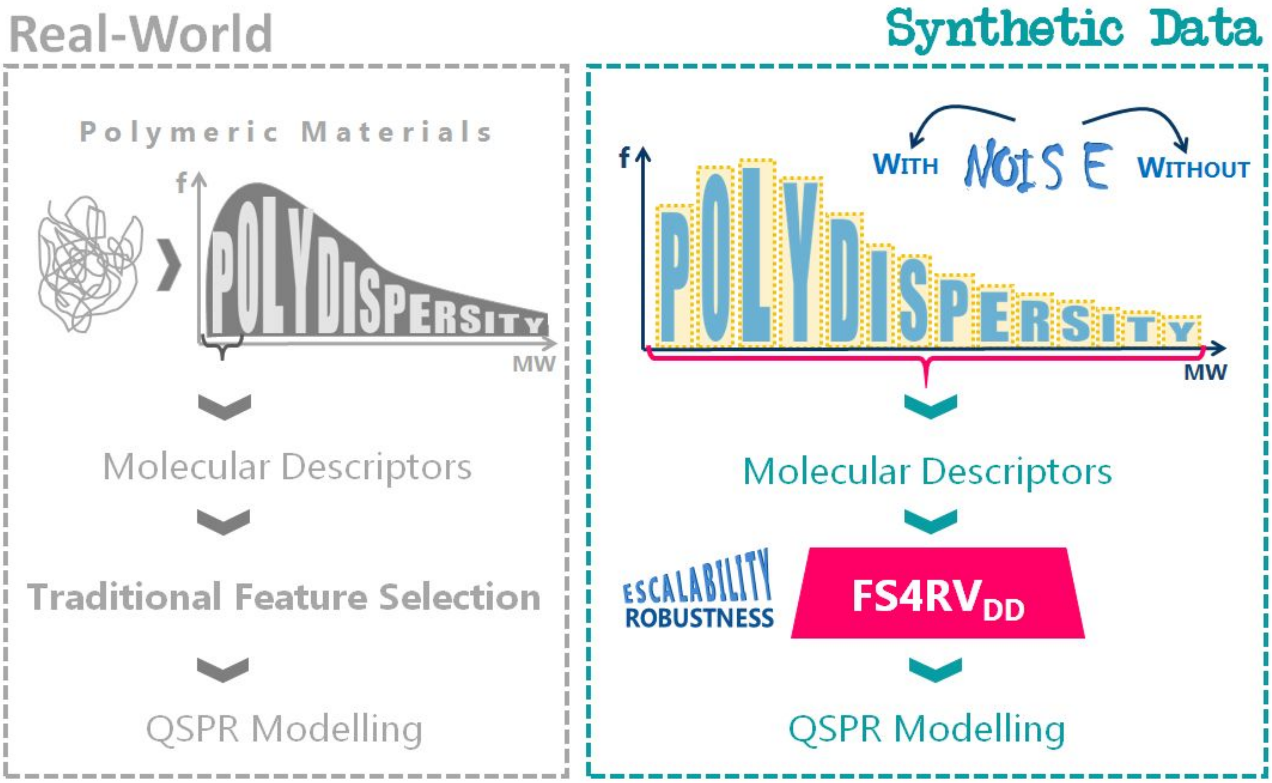
48 Cravero F., Schustik S., Martínez M.J., Barranco C.D., Díaz M.F., Ponzoni I. (2019) Feature Selection and Polydispersity Characterization for QSPR Modelling: Predicting a Tensile Property. In *International Conference on Practical Applications of Computational Biology & Bioinformatics*. Springer, Cham.

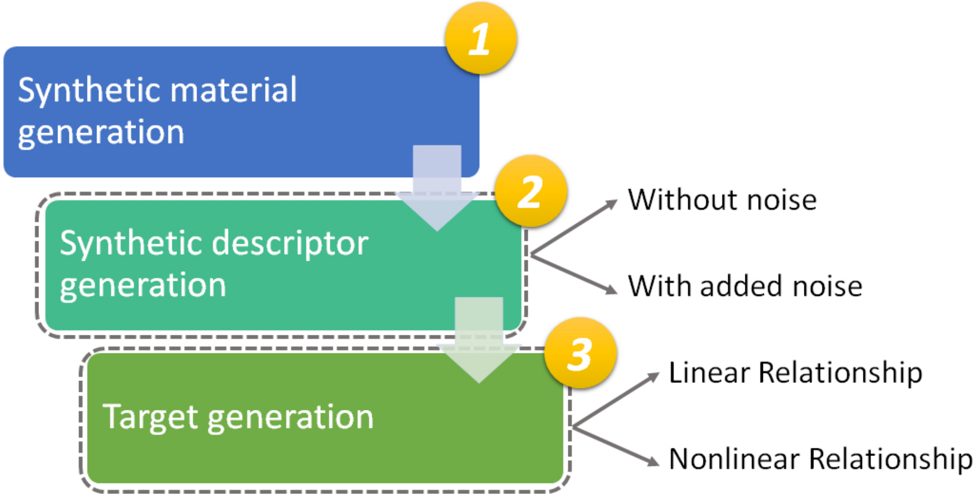
49 Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I. H. (2009). The WEKA Data Mining Software: an Update. *ACM SIGKDD Explorations Newsletter*.

50 Kay, S. (2000). Can detectability be improved by adding noise?. *IEEE signal processing letters*.

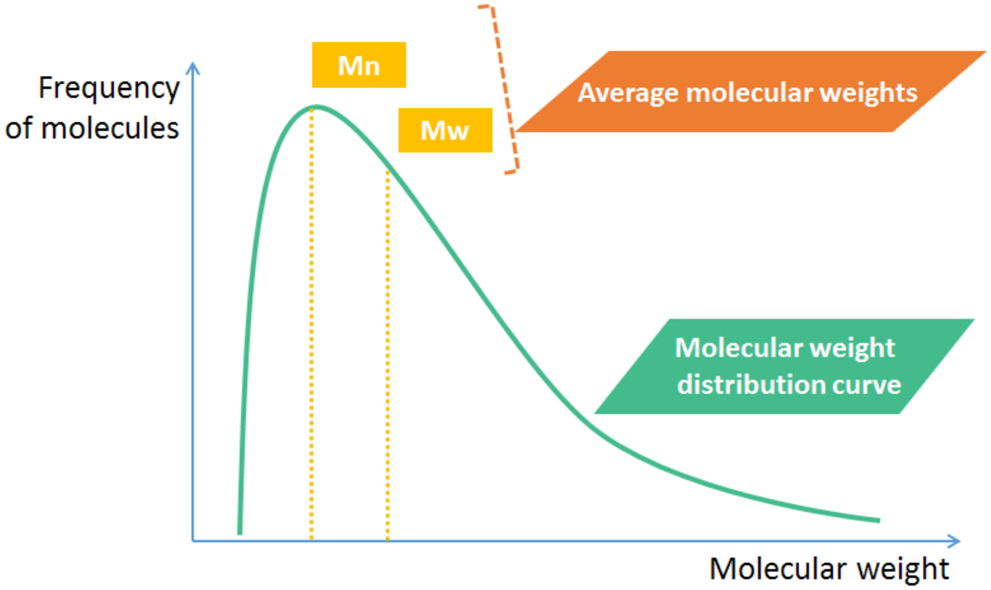
51 Ballabio, D., Grisoni, F., & Todeschini, R. (2018). Multivariate Comparison of Classification Performance Measures. *Chemom. Intell. Lab. Syst*.

For Table of Contents Only

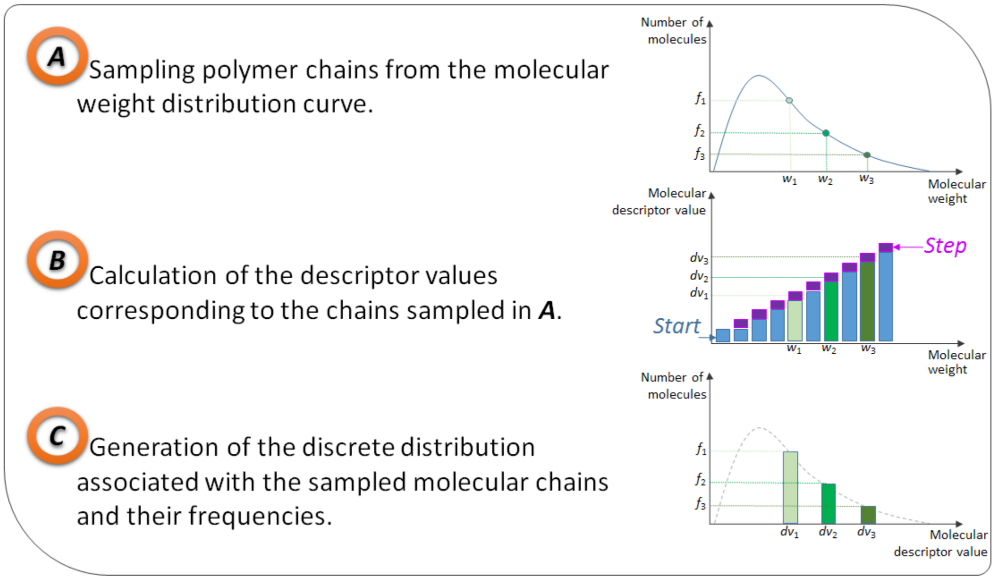




Main steps for the generation of synthetic data.


Frequency of molecules
Mn
Mw
Average molecular weights
Molecular weight distribution curve
Molecular weight

Graph of the classic molecular weight distribution (MWD) curve of a human-made polymer, where number average molecular weight (Mn) and weight average molecular weight (Mw) are denoted.

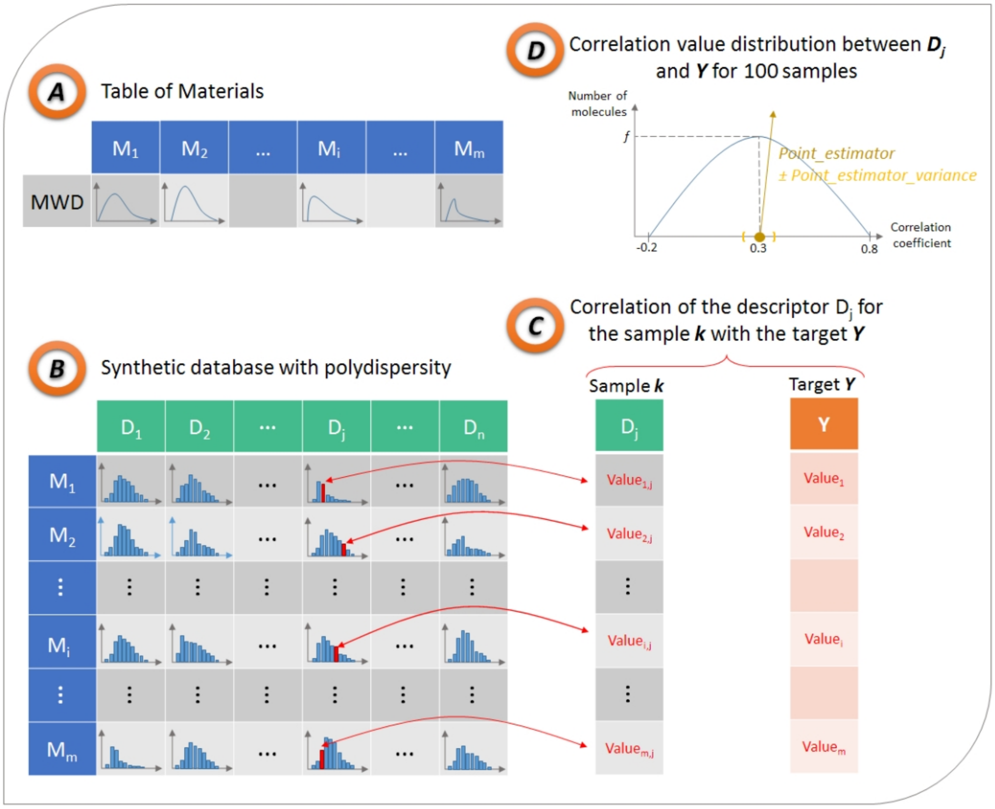


A Sampling polymer chains from the molecular weight distribution curve.

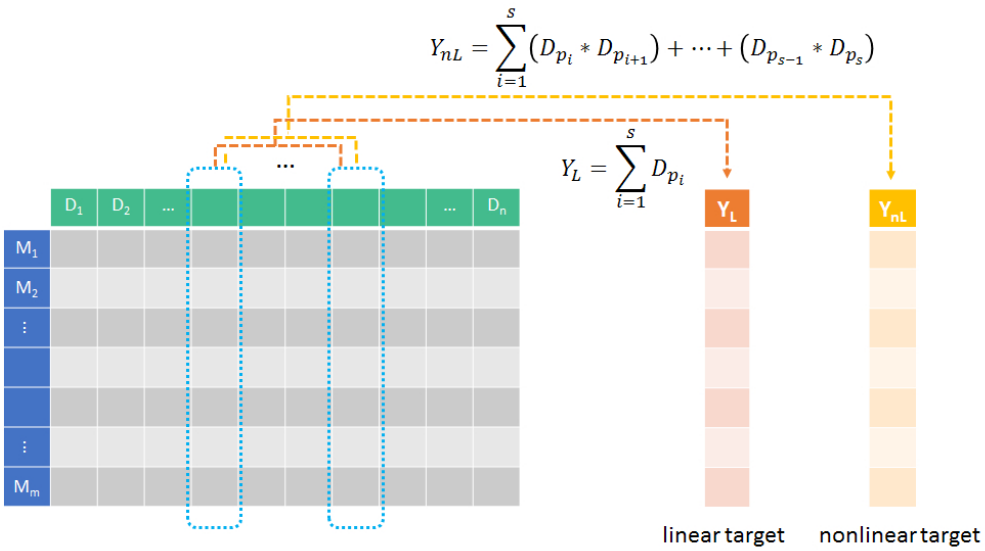
B Calculation of the descriptor values corresponding to the chains sampled in ***A***.

C Generation of the discrete distribution associated with the sampled molecular chains and their frequencies.

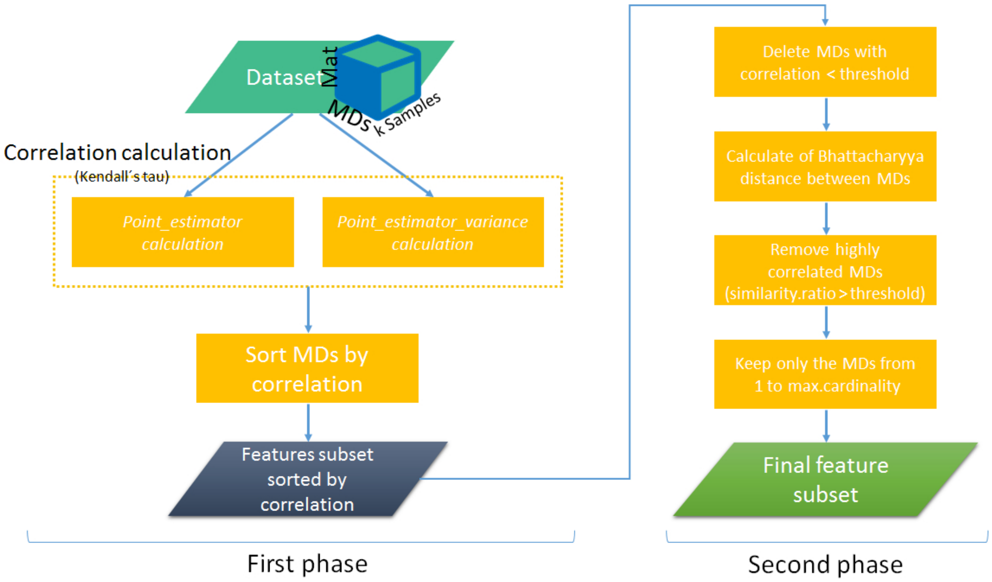
Description of steps for introducing the polydispersity phenomenon into the molecular descriptor characterization.



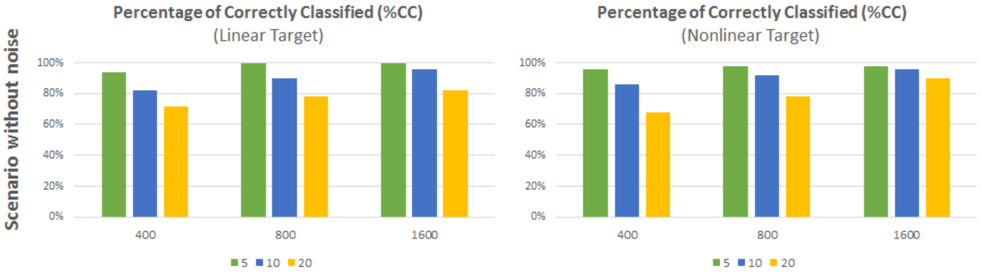
Graphical scheme of the conceptual construction of the database. Note that it contains polydisperse data and not a unique value.


$Y_{nL} = \sum_{i=1}^{s} (D_{p_i} * D_{p_{i+1}}) + \cdots + (D_{p_{s-1}} * D_{p_s})$
...
$Y_L = \sum_{i=1}^{s} D_{p_i}$
D_1
D_2
...
...
D_n
M_1
M_2
⋮
⋮
M_m
Y_L
Y_{nL}
linear target
nonlinear target

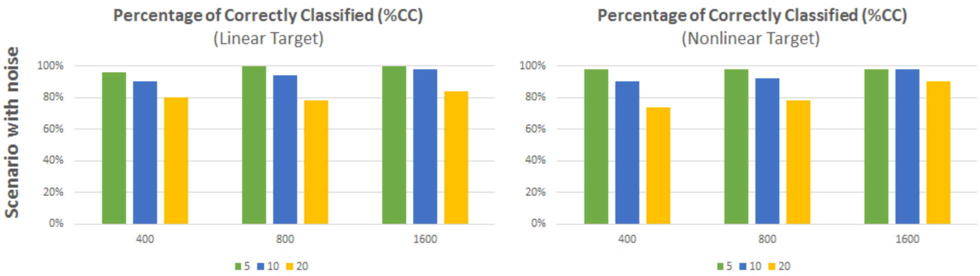
Target generation for both linear and nonlinear targets.


Dataset
Mat
MDs
k Samples
Correlation calculation
(Kendall´s tau)
Point_estimator calculation
Point_estimator_variance calculation
Sort MDs by correlation
Features subset sorted by correlation
First phase
Delete MDs with correlation < threshold
Calculate of Bhattacharyya distance between MDs
Remove highly correlated MDs (similarity.ratio > threshold)
Keep only the MDs from 1 to max.cardinality
Final feature subset
Second phase

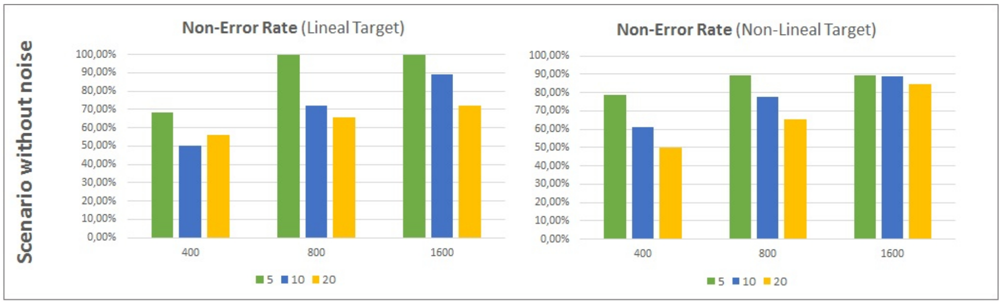
Global methodology for our Feature Selection Method.



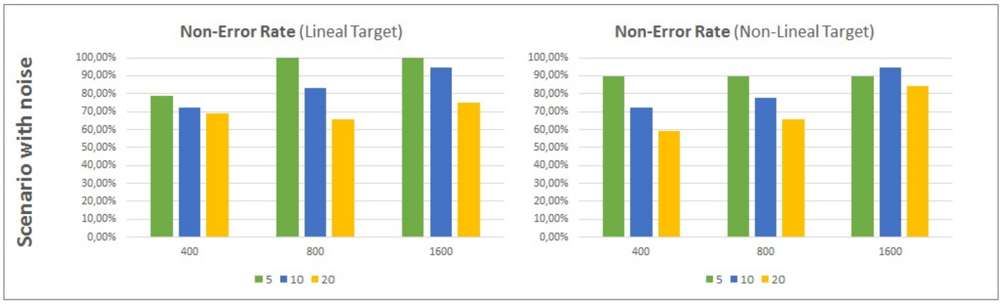
Percentage of MDs correctly classified by FS4RVDD for noiseless experiments.



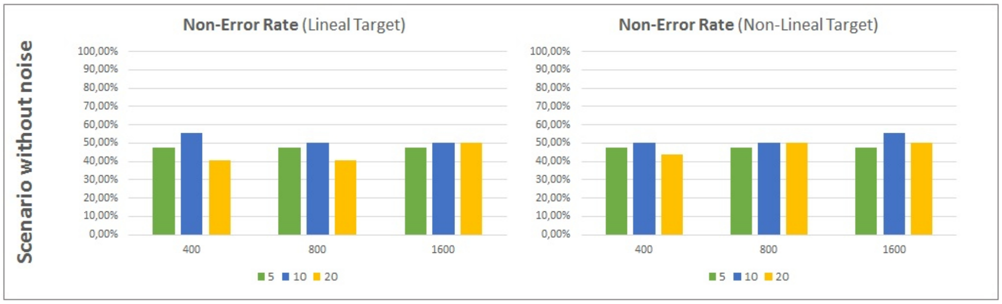
Percentage of MDs correctly classified by FS4RVDD for experiments with noise.



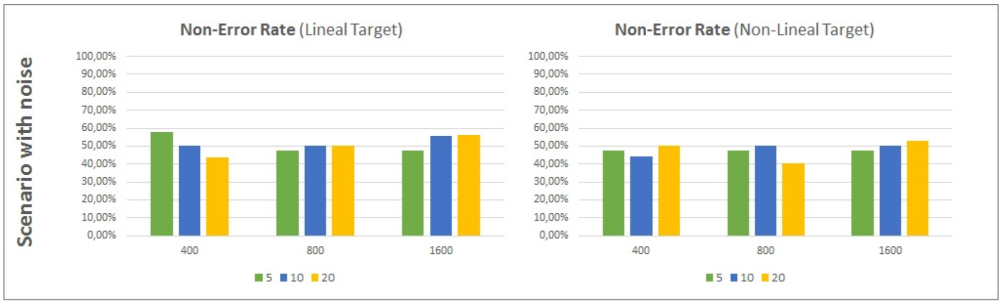
Non-Error Rate measures obtained by FS4RVDD for noiseless experiments.



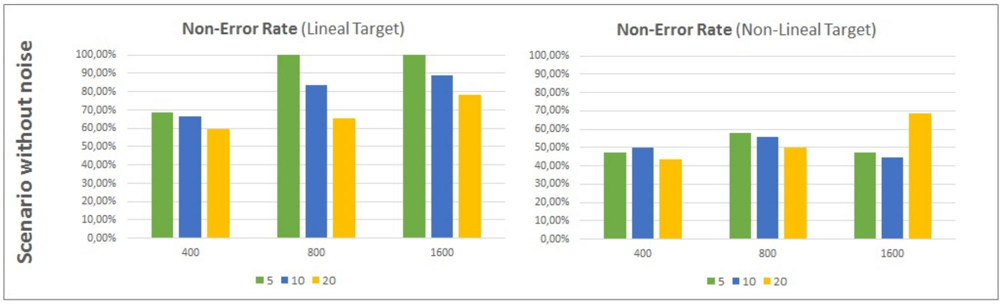
Non-Error Rate measures obtained by FS4RVDD for experiments with noise.



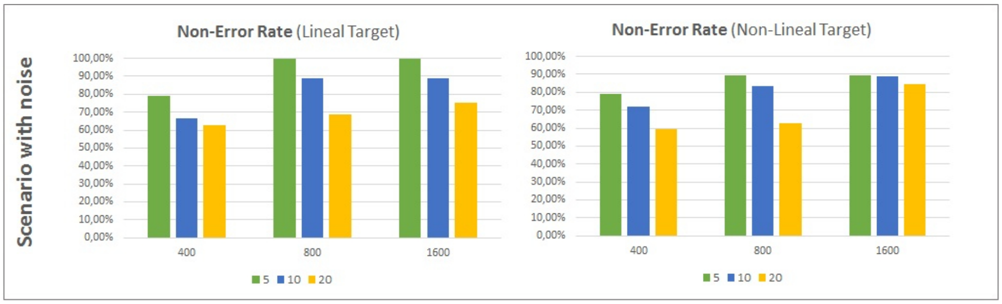
Non-Error Rate obtained by a traditional FS method using Lowest Value based-representations (analogous to the SRU-based representation) for noiseless experiments.



Non-Error Rate obtained by a traditional FS method using Lowest Value based-representations (analogous to the SRU-based representation) for experiments with noise.



Non-Error Rate obtained by a traditional FS method using Mean Value based-representations (analogous to the average weight-based representation) for noiseless experiments.



Non-Error Rate obtained by a traditional FS method using Mean Value based-representations (analogous to the average weight-based representation) for experiments with noise.