The Journal of Chemical Physics

Polymer informatics for QSPR prediction of tensile mechanical properties. Case study: Strength at break

Cite as: J. Chem. Phys. (in press) (2022); https://doi.org/10.1063/5.0087392
Submitted: 04 February 2022 • Accepted: 08 May 2022 • Accepted Manuscript Online: 09 May 2022

Fiorella Cravero, Monica Fatima Díaz and Ignacio Ponzoni

AIP Publishing

Accepted to J. Chem. Phys. 10.1063/5.0087392

Polymer informatics for QSPR prediction of tensile mechanical properties. Case study: Strength at break

Fiorella Cravero[1], *Mónica F. Díaz*[2,3*], *and Ignacio Ponzoni*[1,4]

[1]Instituto de Ciencias e Ingeniería de la Computación (ICIC), Universidad Nacional del Sur (UNS) - Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Bahía Blanca, Buenos Aires, 8000, Argentina.

[2]Planta Piloto de Ingeniería Química (PLAPIQUI), Universidad Nacional del Sur (UNS) - Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Bahía Blanca, Buenos Aires, 8000, Argentina.

[3]Departamento de Ingeniería Química (DIQ-UNS), Bahía Blanca, Buenos Aires, 8000, Argentina.

[4]Departamento de Ciencias e Ingeniería de la Computación, (DCIC-UNS), Bahía Blanca, Buenos Aires, 8000, Argentina.

*Author to whom correspondence should be addressed: mdiaz@plapiqui.edu.ar

Keywords: Machine Learning, Visual Analytics, Polymer Informatics, QSPR, Mechanical Properties, Strength at break, Tensile Test

ABSTRACT. The artificial intelligence-based prediction of the mechanical properties derived from the tensile test, plays a key role in assessing the application profile of new polymeric materials, specifically in the design stage, prior to synthesis. This strategy saves time and resources when creating new polymers with improved properties that are increasingly demanded by the market. A quantitative structure-property relationship (QSPR) model for tensile strength at break is presented in this work. The QSPR methodology applied here is based on machine learning tools, visual analytics methods, and expert-in-the-loop strategies. From the whole study, a QSPR model composed of five molecular descriptors that achieved a correlation coefficient of 0.9226 is proposed. We applied visual analytics tools at two levels of analysis: a more general one in which models are discarded for redundant information metrics and a deeper one in which a chemistry expert can make decisions on the composition of the

model in terms of subsets of molecular descriptors, from a physical-chemical point of view. In this way, with the present work, we close a contribution cycle to polymer informatics, providing QSPR models oriented to the prediction of mechanical properties related to the tensile test.

1. INTRODUCTION

Traditionally, for the exploration of structure-property correlations or tailoring of material properties for a specific application, a small series of polymers with chemical variations are synthesized for investigation. The chosen monomers are usually based on an expert's chemical intuition [Guo *et al.*, 2021]. Although this approach may enable polymer chemists to identify materials with suitable properties, it is highly unlikely that this approach will identify the polymer with the optimal possible properties for the intended application. To address this issue, advances have been made in the development of machine learning (ML) methods that can learn the structure-property relationship by computational models [Guo *et al.*, 2021]. Therefore, this interdisciplinary shift in the way of designing new materials represents a significant advance in materials science and it is a long-standing challenge in cheminformatics, specifically in polymer informatics [Kim *et al*., 2019, Chen *et al*., 2021, Kuenneth *et al*., 2021, Lui *et al*., 2021, Sha *et al*., 2021].

These structure-property relationships are used to make predictions of polymer properties, thereby providing a rational way to prioritize among different candidate polymers during the virtual screening step of computational materials design projects [Guo *et al.*, 2021, Wu *et al.*, 2019]. In addition, compound libraries used in virtual screening can also be developed by ML to generate hundreds of virtual candidate molecules with structural variations, without the need to synthesize them [Bilodeau, *et al*., 2022]. Understanding the processes that govern the performance of polymeric materials by exploring modeling results becomes a major challenge in polymer informatics. In this sense, ML has been involved for more than half a century in the

development of techniques that quantitatively relate structure to an activity or a property. These relationships began to be studied sixty years ago in the field of biology [Hansch, and Fujita, 1964], but today they are applied to several others, such as materials science [Meredig, 2019, Chen et al., 2021].

Specifically, in polymer informatics, one of the most studied properties is glass transition temperature (Tg) [Adams, 2010; Chen et al., 2018] beginning with the pioneering works of Katritzky *et al.*[Katritzky *et al.*, 1996], to the present with the work of Kim *et al*, that propose a general ML-based approach to design polymers that meet practically useful, but extreme, property criteria using as a use case: Tg>500 K and bandgap, Eg>6 eV [Kim,*et al.*, 2021]. It is a highly relevant property of polymers because it is related to molecular mobility, material processability, and consequently to macroscopic mechanical properties [Kim, et al., 2019, Karuth et al., 2021, Ward and Sweeney, 2004]. There are other interesting properties [Liu, et al., 2020], such as electrical [Hu, et al., 2020; Tuan-Anh, T. and Zaleśny, R., 2020], or optical [Theodosiou and Kalli, 2020; Erickson, et al., 2020].

In particular, we are interested in mechanical properties derived from the tensile test for high molecular weight linear polymers because all service conditions involve some mechanical loading. Therefore, the choice of a material for a specific application must be based on the knowledge of these properties. They are related to forces outside the polymeric material that are exerted on it and respond in one way or another depending on the type of polymer and the transformation process they have undergone. Generally, plastic materials respond to these forces to which they are subjected by undergoing a deformation that, depending on the case, is greater or lesser and ends in breakage. According to the need and the field of application of each product, whether for packaging, construction, aeronautics, or medicine, for example, some mechanical (tensile, compression, flexural, impact) properties will be determined. The

mechanical properties are very little studied compared to the others, one of the few current references on the prediction of mechanical properties corresponds to the work of Pal and Naskar, [Pal and Naskar, 2021], and another is that of Kuenneth, *et al*. [Kuenneth *et al*., 2021]. On the one hand, Pal and Naskar present a general workflow to predict the properties of ionic thermoplastic elastomers (Ionic-TPEs) through the inference of three different regression ML models. They generated a virtual stress-strain plot (including: tensile strength, elongation at break, hardness (shore A) tear strength), created by the regression model trained for the Ionic-TPEs samples. They prepare only 10 samples (mixtures with different ratios of rubber and stearic acid), measure the stress-strain properties of interest, and then train ML models with these data. On the other hand, Kuenneth *et al*., present a multi-task learning study, where they work with 13,000 polymers and 36 properties, including two mechanical properties. The method presented by them learns "better" as it has more data on the same or other properties, learns one property at a time, and then looks for inherent correlations between the data on the different properties.

Our long-term goal is the development of virtual screening methods for the identification of promising synthetic polymers with a specific industrial profile, based on the quantitative structure-property relationship (QSPR) modeling of mechanical properties. We have already presented a QSPR model for the prediction of elongation at break [Palomba et al., 2014] and two predictive QSPR models for the tensile modulus [Cravero et al., 2019]. The aim of the present work is to close the prediction cycle for the tensile test by presenting a QSPR model for tensile strength at break, thus achieving a complete set of useful prediction tools for materials design practitioners.

2. RESULTS AND DISCUSSIONS

2.1. POLYMER DATABASE: DATA REPRESENTATION, PREDICTED PROPERTY, AND MOLECULAR DESCRIPTORS

It is relevant to understand the inherent complexity of QSPR modeling in the context of polymer informatics. While the chemical structure of a low molecular weight drug candidate determines its biological activity, the structural repeat unit (SRU) of a polymer is only one of many parameters that affects its ultimate utility. Other equally important parameters are polymer molecular weight, molecular weight distribution, and fabrication method. Establishing QSPR correlations using both molecular structure and fabrication characteristics is of utmost significance for prediction purposes. Even though the SRU reduces the structural information on a material (i.e., it does not represent the bulk material), there are vast works in the literature that validate its use. To try to overcome part of this problem, we introduced some descriptors that provide information on average molecular weights, polydispersity, and testing parameters (e.g., crosshead speed).

2.1.1. POLYMERS INCLUDED IN THE DATABASE

The dataset reported by Palomba et al. [Palomba et al., 2014] for high molecular weight linear polymers was used for the present work. This dataset was especially created for the study of mechanical properties derived from tensile tests for amorphous and thermosetting polymers. All the polymer data were taken from PolyInfo [Otsuka, *et al.*, 2011], following homogeneity criteria for testing parameters and standards. Our complete dataset includes values for elongation and tensile strength at break as well as tensile modulus for each polymer. These properties were reported in PolyInfo under similar experimental conditions (testing temperature (20-25 ºC) and standards (ASTM D638, ASTM D882-83, and DIN 53504.53A)), therefore all polymers reported have been tested under their glassy temperature. Nine different chemical

families of compounds are represented in the dataset: Polyoxides/ethers/acetals, Polystyrenes, Polyvinyls, Polyamides/thioamides, Polyesters/thioesters, Polyimides/thioamides, Polysulfones/sulfoxides/sulfonates/sulfoamides, Polyketones/thioketones, and Polysulfides. The polymers can belong to more than one family, in consequence their molecular structures are more diverse and complex, contributing to expanding the chemical space represented in this dataset.

From the original dataset, 76 polymers were included; the dataset was split into two subsets using the same division defined in a previous work [Cravero et al., 2019], where predictive QSPR models for the tensile modulus were presented. One of the subsets was reserved as a testing set (19 polymers), and the other three were used as a training set (57 polymers), preserving the structural diversity from the nine families mentioned. The 76 high molecular weight linear polymers represent nine different families, and most of these polymers belong to more than one of them. This makes it a structurally complex set and therefore expands the chemical space it represents.

2.1.2. TARGET PROPERTY: TENSILE STRENGTH AT BREAK

Tensile strength at break is the property that shows the capability of a polymeric material to resist breaking under tensile stress **(Figure 1)**. It is therefore the force per unit area required to break the material specimen. The crosshead speed at which a sample is pulled apart in the test can affect the results, and this aspect was considered in our database by including these values (ranging from 1 to 100 [mm/min]) as a descriptor in the QSPR technique [Seymour and Carraher, 1998; Ward and Sweeney, 2004]. The values for tensile strength at break in the database range from 7.5 to 103 [MPa].

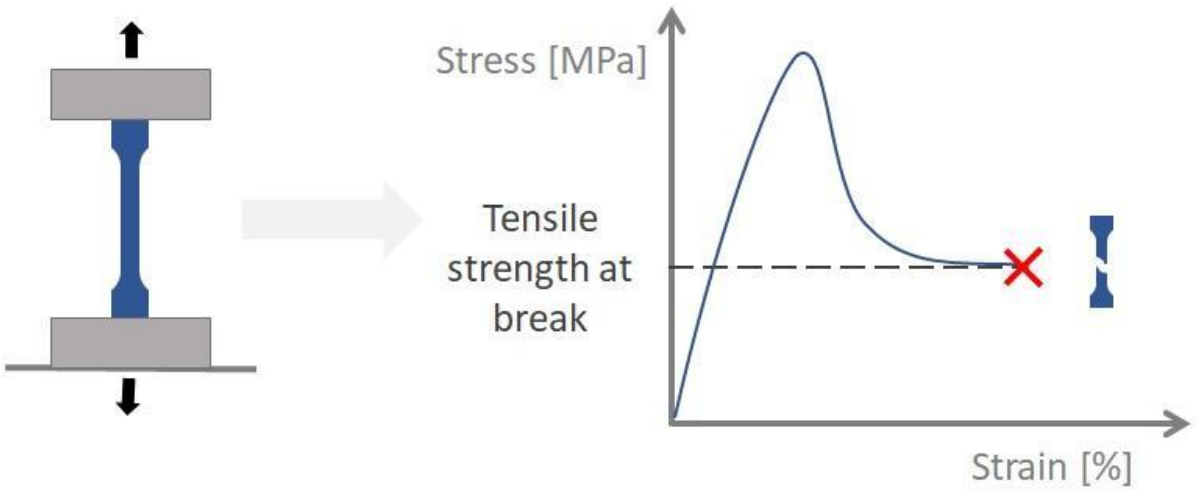


Figure 1. Graphic representation of the tensile test and derived stress-strain curve for a polymeric material. Tensile strength at break [MPa] is plotted on the curve

2.1.3. MOLECULAR DESCRIPTORS

Regarding the characterization of the polymers, two molecular models were used. On the one hand, the hydrogen-end-capped SRU was used, and 998 classical descriptors were calculated for each SRU with the Dragon tool [Mauri et al., 2006]. On the other hand, using the hydrogen-end-capped trimer's central unit as a molecular model, 51 global descriptors were calculated using the Hyperchem software [Froimowitz, 1993]. These descriptors are called macro vision descriptors [Palomba et al., 2012], and they are normalized; that is, they do not vary according to the number of SRUs included in the molecular model used. Other types of features included in this last category are parameters such as the crosshead speed (CHS) at which the tensile test is performed or global data of the polymeric material such as the polydispersity index (PDI), average molecular weights in number (Mn), or average molecular weights in weight (Mw). Finally, a filter was applied to remove features with missing values, highly correlated features, and features with low variance to discard those descriptors that were not useful for inferring the QSPR models describing the molecule. The final pool had a total of 693 molecular descriptors (MDs).

2.2. ML-BASED METHODOLOGY USED FOR QSPR MODELING

Figure 2 shows a flowchart of the methodological steps followed to infer the predictive QSPR models for the tensile strength at break property. In the first step, a percentage of the original dataset is separated to execute the corresponding external validation of the final models. In the second step, the Feature Selection procedure is performed, followed by an analysis of the subset of descriptors obtained. Then, the inference of the QSPR models is executed and the external validation is carried out for each of them. These steps are explained in detail below and the results are discussed.

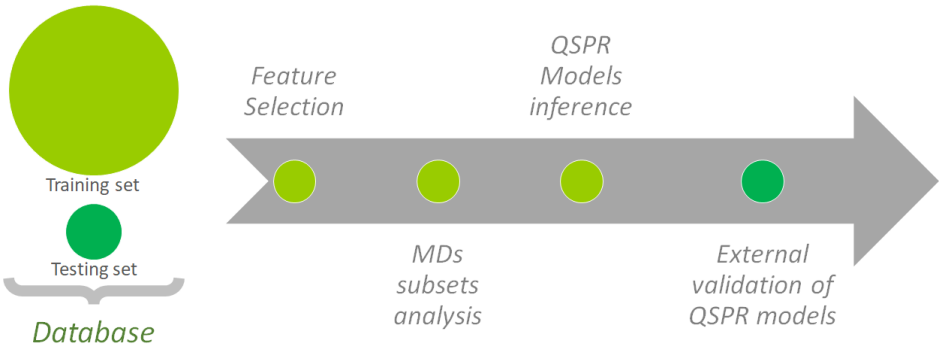


Figure 2. Flowchart of the methodological steps used for generating the QSPR models

2.2.1. *FEATURE SELECTION STEP*

Feature Selection methods are part of the dimensionality reduction methods for pattern recognition systems [Li *et al.*, 2018]. In ML, the aim of these methods is to obtain, from a full set of features, a relevant subset for the inference of a model. In small datasets, where the features (molecular descriptors) are many more than the instances (polymers), using all the features can lead to obtaining models with a high level of overfitting. It is assumed that many of the variables may be irrelevant or redundant for the model. Irrelevant variables are those that do not provide useful information for inference the model and the redundant variables are those that do not provide additional information regarding the remaining selected ones. Besides, a

completely useless variable by itself, it can generate an improvement in the performance of the model if it is taken together with others [Cai *et al.*, 2018.].

Feature selection (FS) plays an important role in the success of an ML model. The features used, together with the training method, influence the final result of a predictive QSPR model. The training methods cannot solve the problem of learning models if the wrong features are used, that is, if they do not provide information. FS can improve results by reducing the noise contributed by some useless features. In this fact lies the importance of doing FS when the dataset is small.

In order to have several different subsets of descriptors, FS techniques were applied to the initial pool of 693 MDs. To perform this task, the Waikato environment for knowledge analysis (Weka) [Hall, et al., 2009] framework was used, experimenting with five different classification functions: one correlation-based feature subset selection (CFS) and four instances of wrapper subset eval (WSE), using different learning techniques: linear regression (WLR), multilayer perceptron (WMLP), random forest (WRF), and random committee (WRC). In addition, the molecular descriptor subset selection (MoDeSuS) [Martínez et al., 2019.] software was used to obtain other five subsets. A summary of the main aspects of the selected subsets is detailed in **Supplementary Information**.

2.2.2. *MOLECULAR DESCRIPTORS SUBSET ANALYSIS*

To begin the analysis of this modeling stage, we have ten subsets of molecular descriptors obtained in the previous stage. These alternative subsets must be judged and selected based not only on the accuracy of their predictions related to the target property, but also on their cardinality (overfitting problem) and information contribution of the descriptors of the same subset (irrelevant or redundant features). For this task, the visual analytics tool ViDeAN

[Martínez et al., 2015] is used to perform a detailed analysis of the subsets and choose the best five in terms of the aforementioned criteria.

ViDeAN enables visualizations to analyze the correlations between each pair of descriptors in Spearman (S), Kenndal (K), and entropy-based (E) terms. In the ViDeAN plots, in Spearman or Kendall correlation modes, the edges of the graphs represent the Spearman or Kendall rank correlation between the descriptors (nodes). The link color range is: from red to yellow for positive correlation (red = 1) and from blue to light blue for negative correlation (blue = −1). Conditional entropy is a measure of the amount of uncertainty about one random variable when the value of another random variable is known. In entropy-based mode, the links represent the mutual information (MI) between the descriptors, i.e., the reduction of the uncertainty of one of them due to the knowledge of the other, and vice versa. The link color range goes: from light pink when the descriptors are independent (MI = 0) to dark purple when the descriptors are identical (high MI values) [Martínez *et al.*, 2015]. All the screenshots of each of the correlation graphs for each subset can be found in **Supplementary Information Figure SI.1.**

The criteria used to discard subsets were a) high correlation level among the descriptors that conform the subset and b) their cardinality, preferring the lowest ones. That is, first are discarded those subsets formed by MDs with high MI or correlation (in the plots created by ViDeAN this is displayed as dark links of the graphs), and secondly, when the percentage amount of correlation of the MDs is the same, the subset with the highest total number is discarded. All the screenshots of each of the correlation graphs for each subset can be found in **Supplementary Information Figure SI.1. Table 1** shows the percentage of low correlations between each pair of descriptors, that is, the number of links that do not exceed 50% correlation; this means that they have little mutual information, which is desirable.

Table 1. Percentages of pairwise descriptor links in Spearman (S), Kenndal (K), and entropy-based (E) terms of each subset with associated low correlation coefficients

	Cardinality	S (%)	K (%)	E (%)	Selected?
Sub1	10	50	70	60	Yes
Sub2	10	50	70	100	Yes
Sub3	10	30	30	20	No
Sub4	10	50	60	100	Yes
Sub5	10	30	30	50	No
Sub6	33	18	18	78	No
Sub7	12	16	58	83	No
Sub8	8	37	50	62	Yes
Sub9	6	50	60	100	Yes
Sub10	11	36	45	45	No

The percentages in **Table 1** are formed by counting those links that have less than 50% correlation over the total number of links. Therefore, it is desirable to choose those subsets with higher percentages in the table (lower MI). The main steps of the analysis and assessment were the following:

- The first subsets selected for further experimentation were Sub1, Sub2, Sub4, and Sub9 because all of them had at least 50% of the correlations between descriptors belonging to the subset below the midpoint on all correlation scales analyzed. Note that in all of them all the links are light colored, that is, they are below the midpoint of the correlation range.
- Doing the same analysis, subsets Sub3, Sub5, and Sub10 had correlation values above the mean of the range in all the different types of correlations considered (dark links), therefore they were discarded.

- From the remaining subsets, regarding subset Sub6, although it had 78% of entropy-based correlation low values, Spearman and Kendall correlation low values was low (18% in both cases). The subset Sub6 was discarded for this and because of its high cardinality (the highest of the ten subsets).

- Finally, among the subsets Sub7 and Sub8, the latter was selected. Both have low values for both the entropy-based correlation and the Kendall correlation; these were greater than the percentage (50%) threshold defined in this analysis for the three types of correlations considered. However, the cardinality of subset Sub8 was lower, which strongly influenced the decision.

2.2.3. *QSPR MODEL INFERENCE*

QSPR models estimate an objective property of chemical compounds from variables that describe their molecular structure and a mathematical equation that relates them. The predictive QSPR models were inferred using four well-known ML techniques: linear regression (LR), multilayer perceptron (MLP), random forest (RF), and random committee (RC), within the Weka framework [Hall et al., 2009]. More information on how these techniques were used in the present work can be found in **Supplementary Information.**

In total, 20 QSPR models were trained by combining the application of each ML method with each selected subset. Different measures were considered to evaluate the quality of each inferred model to access those with the best statistical performance and physicochemical interpretability. One of these measures was the correlation coefficient, a statistic whose main purpose is to determine the quality of the model to replicate the results and the proportion of variation of the results that can be explained by the model. As it should not be used as the sole guide to select the final models, the values of some error metrics were also considered to perform the evaluation of the inferred models. These errors were mean absolute error (MAE)

and root mean squared error (RMSE). MAE measures the average magnitude of the errors in a set of predictions, without considering their direction. It is the average over the test sample of the absolute differences between prediction and actual observation, where all individual differences have the same weight. RMSE is a quadratic scoring rule that also measures the average magnitude of the error. It is the square root of the average of the squared differences between prediction and actual observation. [Taraji et al., 2017] The criterion for selecting the models in this step was to exceed a correlation coefficient of 0.75; those models whose correlation coefficient was between 0.7 and 0.75 were considered reasonably good models. Additionally, all error measures were expected to be kept low.

Table 2 shows the performances of the QSPR models. All correlation coefficient values are above random results; furthermore, the minimum value is 0.612. The maximum value is 0.936, and 17/20 (85%) models have a correlation coefficient value above 0.75; that is, only 15% of the inferred models have a correlation coefficient below 0.75. If we also consider that 70% of the models are above 0.8, it could be said that the tensile strength at break is a property that can be addressed with this methodology for this type of polymers.

When trained with MLP, subset Sub4 obtained the lowest correlation coefficient; its best performance was achieved when trained using RF. However, the correlation coefficient was barely 0.76, so all four models trained from subset Sub4 were discarded. As well as for subset Sub1, the correlation coefficient values were above 0.8; yet they did not exceed 0.82/0.83, which was low considering that the maximum value achieved exceeded 0.93. It was a stable subset; that is, the models inferred from it had similar correlation coefficient values regardless of the ML method used. However, none of them exceeded the value of 0.85, so its four models were discarded. The same happened with subset Sub2; it was stable, but no model exceeded a

correlation coefficient value of 0.9, so its four models were discarded. From the subsets of molecular descriptors extracted by MoDeSus, subset Sub2 achieved the highest performance.

Table 2. Correlation coefficient and error values for the selected subsets. The highest performances are highlighted in bold

		LR	MLP	RC	RF
Sub1	Correlation coefficient	0.8284	0.8090	0.8213	0.8223
	MAE	12.4241	13.5058	12.4002	12.1383
	RMSE	16.1454	17.4071	16.3708	16.1713
Sub2	Correlation coefficient	0.8597	0.8316	0.8288	0.8672
	MAE	11.2811	13.0056	11.5685	10.5189
	RMSE	14.5203	17.2705	16.0994	14.1359
Sub4	Correlation coefficient	0.7151	0.6117	0.7534	0.7639
	MAE	15.7051	17.0702	13.9597	13.9861
	RMSE	20.2120	27.9884	18.8229	18.3327
Sub8	Correlation coefficient	0.8335	0.7468	**0.9226**	**0.9176**
	MAE	13.0271	15.5269	8.2702	8.3410
	RMSE	15.7964	24.4889	11.0138	11.3948
Sub9	Correlation coefficient	0.7710	**0.9020**	**0.9322**	**0.9361**
	MAE	15.2743	9.7278	8.5314	8.3188
	RMSE	18.1904	12.5023	10.2820	10.1300

Regarding the remaining subsets obtained by Weka, when they were trained by ensemble techniques such as RC or RF, they achieved correlation coefficient values greater than 0.9. Moreover, the model inferred using MLP from subset Sub9 obtained a correlation coefficient value greater than 0.9. Finally, these five QSPR models continued in the next step of external validation. In addition to having the highest correlation coefficient values, the RMSE values of

these final models were lower than those obtained for their counterparts (trained with the same ML method). The MAE values for each of these five models remained below 10, being the lowest values reported in the table. Note that the chosen QSPR models used the subset with the lowest cardinalities among all selected subsets.

2.2.4. *EXTERNAL VALIDATION OF QSPR MODELS*

The goal of the ML algorithm is to obtain a reasonable approximation. The inferred model was the result of the automated generalization procedure. In ML, this generalizability refers to the capability of the model to adapt appropriately to unseen data, within the same applicability domain. For this purpose, the inferred models were applied to the test set (unknown data). The criterion for selecting the models that passed the external validation was that the difference in the correlation coefficient did not exceed 0.05 between the training and the test sets, as a way for avoiding overfitted models. This difference can be seen in the parity plots shown in Figures SI.2 and SI.3, in Supplementary Information. The results in **Table 3** show that the only model that does not meet the condition is the one trained with RF from subset Sub8 (Sub8_RF).

Table 3. External validation results: correlation coefficient values obtained for the test set, the difference obtained between test and training sets, the mean value of the correlation coefficient using y-Randomization, and the performance classification criteria according to Roy [Roy et al., 2016]

Models	Test	Difference	Y-Rand	Roy's Class
Sub 8_RC	0.8851	**0.0380**	0.1083	Moderate
Sub 8_RF	0.8660	0.0520	0.1088	Moderate
Sub 9_MLP	0.9163	**-0.0140**	0.1068	Moderate
Sub 9_RC	0.9029	**0.0290**	0.1088	Moderate
Sub 9_RF	0.9203	**0,0160**	0.1074	Moderate

QSPR modeling has the risk of finding correlations by chance, since models consider features selected from large descriptor pools. Therefore, it is desirable to perform tests showing the performance of models built from pseudo-descriptors, in which the models derived from random data have lower predictive potential (the quality of QSPR models) than the original models [Lipiński and Szurmak, 2017]. To verify whether the obtained models could be explained by such "chance correlations", we used the Y-scrambling method [Khan and Roy, 2018], in which property data are randomly shuffled to change their true order altering any meaningful structure-property relationships [Muller et al., 2015]. All models passed this test, as the mean, obtained from 1000 random trials, was as far below the correlation coefficient obtained by each of the original models. Roy's test was also run. This allowed classifying the inferred model predictions according to their quality level in Good, Bad, or Moderate, after discarding the possibility of systematic errors. In the present study, all QSPR models passed the test with a Moderate quality level [Roy et al., 2016]. More information in **Supplementary Information.**

Generalizability, non-randomness, and statistical parameter analysis form part of the good practice in QSAR. However, considering the possible physicochemical interpretations of a model is desirable. Out of the final five models, four of them (discarding the RF-trained model from subset Sub8) passed all tests. Consequently, all of them are valid models for the prediction of tensile strength at break. Nevertheless, we selected the MLP-trained model from subset Sub9 to perform physicochemical analysis because it obtained a higher correlation coefficient value for unseen data, thus ruling out the possibility of overfitting. In addition, it had only six features and the lowest cardinality. The features or molecular descriptors for the model Sub9_MLP are: ETA_dEpsilon_D (measure of the contribution of hydrogen bond donor atoms, that is, the presence of groups such as –OH, –NH2, –SH, etc. T -*ETA indices*-), MSD (mean square distance index [Balaban]; MSD decreases with increasing molecular branching in an isomeric

series -*Topological indices*-), nROH (number of hydroxyl groups -*Functional group counts*-), maxssssC (Maximum ssssC -*Atom-type E-state indices*-), nHdsCH, and maxHdsCH (Count of and Maximum atom-type H E-State: = CH-, respectively -*Atom-type E-state indices*-).

To be able to perform a physicochemical interpretation of the best models and, consequently, to intervene on them to increase their reliability, a chemist's knowledge was introduced at this step. This technique is called expert-in-the-loop and allows quickly capturing an expert's experience and knowledge in the domain to obtain better models in less time than it would take to test all combinatorics [Ristoski, et al., 2020; Schustik, et al., 2021].

Note that the two features nHdsCH and maxHdsCH are about the same information; this fact could be confirmed by using ViDeAN. This software tool can generate different types of plots and graphs containing several pieces of information. In the present study, it allowed us to make a more detailed analysis of the contribution of each descriptor to the subset. **Figure 3** shows a scatterplot for the relationship of nHdsCH (on the left) and maxHdsCH (on the right) with the target property. Both features cover similar information spaces, where most of the points lie on the y-axis, that is, they have a value of zero and a few have medium and high property values for low descriptor values.

Another detailed analysis that could be performed using ViDeAN is that of the mutual information between these descriptors by examining the correlation graphs (**Figure 4**). In this case, the Kendal correlation is presented in Figure 4a and the Spearman correlation in Figure 4b. In both graphs, the link between nHdsCH and maxHdsCH is red. This color represents the maximum possible intensity, which means that they have mutual information; in other words, they do not provide different information.

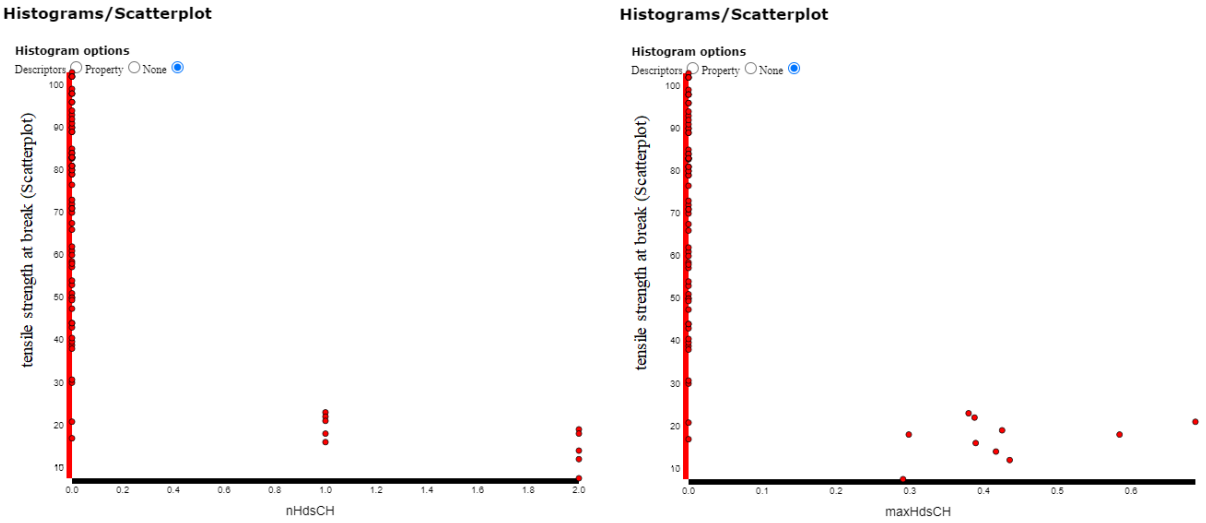


Figure 3. Screenshot of a scatterplot ViDeAN's visualization for the relationship of nHdsCH (on the left) and maxHdsCH (on the right) with the target property

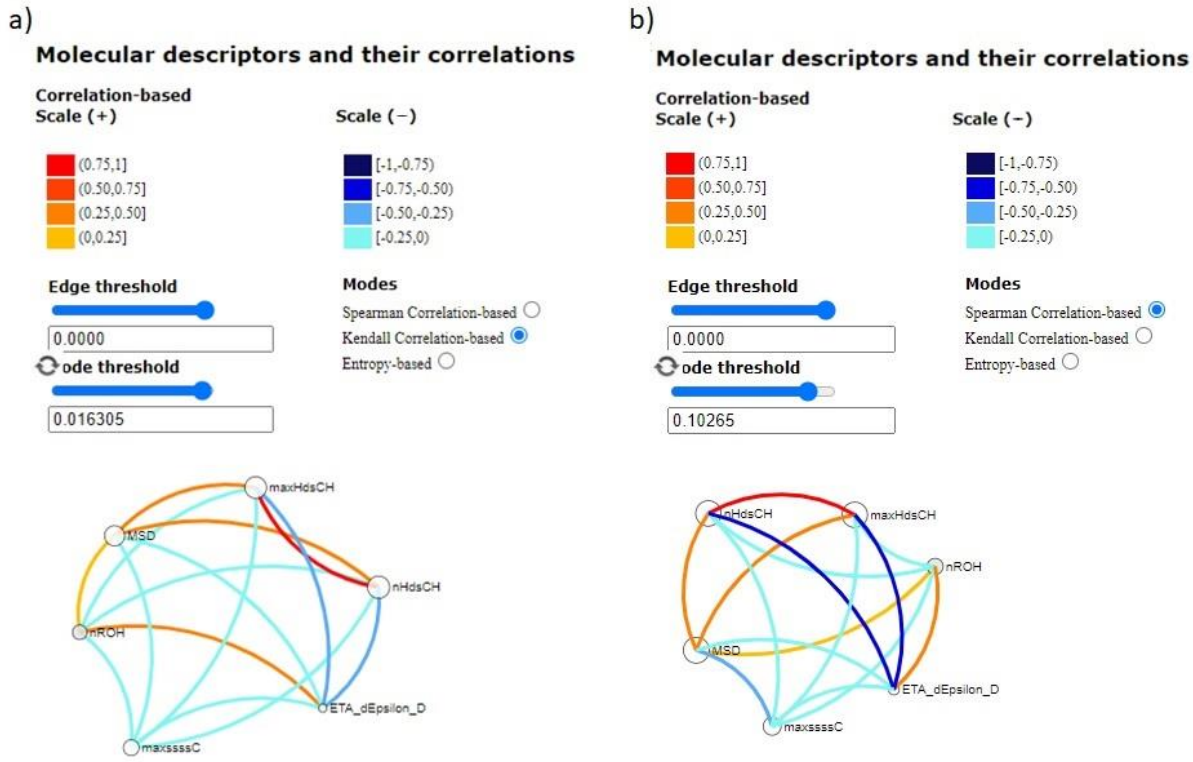


Figure 4. Molecular descriptors and their (a) Kendall and (b) Spearman correlation

We decided to remove each of these two molecular descriptors and analyze the performance of the resulting models. When removing the nHdsCH descriptor, the performance in terms of

correlation coefficient increased to 0.905; when removing the maxHdsCH descriptor, the correlation coefficient increased to 0.9054. Therefore, we concluded that apparently neither descriptor is more relevant than the other. When both descriptors were excluded, the performance of the model dropped to 0.8419, as expected. This means that both descriptors provide relevant information, but it is the same, confirming what was observed in the ViDeAN visualizations. Then, we chose a model to continue our analysis. We selected the model "Sub9_MLP - masHdsCH" because of the minimal difference in performance.

Note in Fig. 4b that the dark blue links join each of these descriptors with ETA_dEpsilon_D. Therefore, an attempt was made to remove it from the model with the best performance, that is, the model with five descriptors "Sub9_MLP - masHdsCH". The resulting model with four descriptors dropped to a correlation coefficient value of 0.748, so it was decided to undo this last elimination. **Figure 5** shows a scatterplot for ETA_dEpsilon_D, in which the coverage of the information space is quite different from that of the other two descriptors; this could explain the drop in the performance of the model when it was removed.

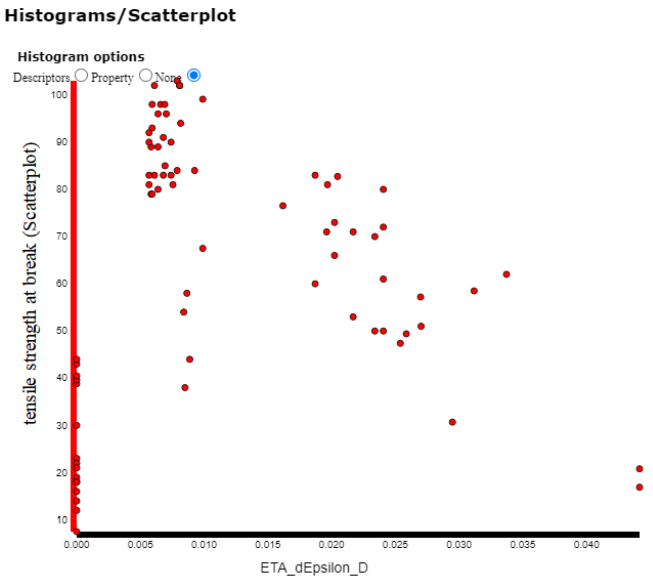


Figure 5. Screenshot of a scatterplot ViDeAN's visualization for the relation of ETA_dEpsilon_D with the target property

Therefore, the final model is integrated by the following five molecular descriptors: ETA_dEpsilon_D, MSD, nROH, maxssssC, and nHdsCH. This model has a performance in terms of correlation coefficient of 0.9054 for training and 0.9226 for external validation. In addition, the y-randomization value is 0.1057 and Roy's class is Moderate. Therefore, these results show the high performance and confidence achieved by the QSPR model. In this way, with the incorporation of the domain expert in the modeling cycle, it was possible to reduce the cardinality of the model by 16% without decreasing its performance. The ready-to-use Weka model is available in **Supplementary Information.**

3. CONCLUSION

Following the proper steps of a QSPR methodology based on ML and visual analytics tools, the tensile strength at break property derived from the tensile test was modeled in this work. The study comprised six consecutive central steps for the analysis: construction of a database integrated by suitable molecular descriptors, selection of subsets of features relevant to the studied property, learning of QSPR models derived from the subsets of selected features, assessment of the QSPR models with higher performances, intervention of the model through the incorporation of an expert-in-the-loop, and a statistical validation procedure. During this process, 20 candidate QSPR models were generated by combining different subsets of selected molecular descriptors and different supervised learning methods.

In a clearly data-driven step, we selected a QSPR model composed of six molecular descriptors that was generated using the MLP method and achieved a correlation coefficient of 0.9163. At this step, visual analytics was used to filter subsets for a global analysis of subsets. In a subsequent step, in which an expert’s knowledge and experience were incorporated, the model selected in the previous step was intervened. This model was reduced in cardinality, eliminating

a molecular descriptor that did not provide additional information. This expert-guided analysis was performed through visual analytics, making use of ViDeAN for a more detailed analysis of each of the descriptors that constituted the subset from which the QSPR model was trained. From the whole analysis and ML process, we recommend a QSPR model composed of five molecular descriptors that was generated using the MLP method and achieved a correlation coefficient for external validation of 0.9226.

In this way, through an exhaustive experimental design, it was possible to infer a QSPR model for the tensile strength at break property, thus completing the cycle of QSPR models oriented to the prediction of mechanical properties related to the tensile test: elongation at break, tensile modulus, and tensile strength at break. We consider that these QSPR models will be useful for the design of new polymeric materials aided by artificial intelligence techniques, saving time and resources against to classical techniques.

SUPPLEMENTARY MATERIAL

See **Supplementary Information** for the complete dataset and the main aspects of the selected subsets, their screenshots of each correlation plots as well as the techniques to inference and validate the predictive QSPR models.

ACKNOWLEDGEMENTS

This work was partially supported by the Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina, [Grant N° PIP 112-2017-0100829], the Universidad Nacional del Sur (UNS), Bahía Blanca, Argentina, [Grants N° PGI 24/N052 and PGI 24/ZM17], and the Agencia Nacional de Promoción Científica y Tecnológica [Grants PICT 2018-04533 and PICT-2019-03350]. We thank Dr. María Jimena Martínez for running the feature selection with the MoDeSuS tool.

AUTHOR DECLARATIONS

Conflict of Interest

The authors declare no conflict of interest.

REFERENCES

Adams, N. (2010). Polymer informatics. In Polymer Libraries (pp. 107-149). Springer, Berlin, Heidelberg. https://doi.org/10.1007/12_2009_18

Bilodeau, C., Jin, W., Jaakkola, T., Barzilay, R., & Jensen, K. F. (2022). Generative models for molecular discovery: Recent advances and challenges. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, e1608. https://doi.org/10.1002/wcms.1608

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. Journal of Polymer Science Part B: Polymer Physics, 56(11), 877-885. https://doi.org/10.1002/polb.24602

Chen, L., Pilania, G., Batra, R., Huan, T. D., Kim, C., Kuenneth, C., & Ramprasad, R. (2021). Polymer informatics: Current status and critical next steps. *Materials Science and Engineering: R: Reports*, *144*, 100595. https://doi.org/10.1016/j.mser.2020.100595

Cravero, F., Martínez, M. J., Ponzoni, I., & Diaz, M. F. (2019). Computational modelling of mechanical properties for new polymeric materials with high molecular weight. *Chemometrics and Intelligent Laboratory Systems*, *193*, 103851. https://doi.org/10.1016/j.chemolab.2019.103851

Erickson, M. E., Ngongang, M., & Rasulev, B. (2020). A refractive index study of a diverse set of polymeric materials by QSPR with quantum-chemical and additive descriptors. *Molecules*, *25*(17), 3772. https://doi.org/10.3390/molecules25173772

Froimowitz, M. (1993). HyperChem: a software package for computational chemistry and molecular modeling. *Biotechniques*, 14(6), 1010-1013. Molecular Modeling System, Release 8.0.7 for Windows Hypercube, Inc., Gainesville, USA (2009)

Guo, K., Yang, Yang, Z., Yu, C. H., & Buehler, M. J. (2021). Artificial intelligence and machine learning in design of mechanical materials. *Materials Horizons*, *8*(4), 1153-1172. DOI: 10.1039/D0MH01451F

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*, *11*(1), 10-18.https://doi.org/10.1145/1656274.1656278

Hansch, C., & Fujita, T. (1964). p-σ-π Analysis. A method for the correlation of biological activity and chemical structure. *Journal of the American Chemical Society*, *86*(8), 1616-1626.https://doi.org/10.1021/ja01062a035

Hu, H., Zhang, F., Luo, S., Chang, W., Yue, J., & Wang, C. H. (2020). Recent advances in rational design of polymer nanocomposite dielectrics for energy storage. *Nano Energy*, *74*, 104844. https://doi.org/10.1016/j.nanoen.2020.104844

Katritzky, A. R., Rachwal, P., Law, K. W., Karelson, M., & Lobanov, V. S. (1996). Prediction of polymer glass transition temperatures using a general quantitative structure− property relationship treatment. Journal of chemical information and computer sciences, 36(4), 879-884. https://doi.org/10.1021/ci950156w

Karuth, A., Alesadi, A., Xia, W., & Rasulev, B. (2021). Predicting glass transition of amorphous polymers by application of cheminformatics and molecular dynamics simulations. *Polymer*, *218*, 123495. https://doi.org/10.1016/j.polymer.2021.123495

Khan, P. M., & Roy, K. (2018). QSPR modelling for prediction of glass transition temperature of diverse polymers. *SAR and QSAR in Environmental Research*, *29*(12), 935-956. https://doi.org/10.1080/1062936X.2018.1536078

Kim, C., Chandrasekaran, A., Jha, A., & Ramprasad, R. (2019). Active-learning and materials design: The example of high glass transition temperature polymers. *MRS Communications*, *9*(3), 860-866.https://doi.org/10.1557/mrc.2019.78

Kim, C., Batra, R., Chen, L., Tran, H., & Ramprasad, R. (2021). Polymer design using genetic algorithm and machine learning. *Computational Materials Science*, *186*, 110067. https://doi.org/10.1016/j.commatsci.2020.110067

Kuenneth, C., et al. (2021). Polymer informatics with multi-task learning. Patterns, 2(4), 100238. https://doi.org/10.1016/j.patter.2021.100238

Kuenneth, C., Rajan, A. C., Tran, H., Chen, L., Kim, C., & Ramprasad, R. (2021). Polymer informatics with multi-task learning. Patterns, 2(4), 100238. https://doi.org/10.1016/j.patter.2021.100238

Liu, Y., Esan, O. C., Pan, Z., & An, L. (2021). Machine learning for advanced energy materials. *Energy and AI*, *3*, 100049. https://doi.org/10.1016/j.egyai.2021.100049

Lipiński, P. F., & Szurmak, P. (2017). SCRAMBLE'N'GAMBLE: a tool for fast and facile generation of random data for statistical evaluation of QSAR models. *Chemical Papers*, *71*(11), 2217-2232. 10.1007/s11696-017-0215-7

Liu, Y., Niu, C., Wang, Z., Gan, Y., Zhu, Y., Sun, S., & Shen, T. (2020). Machine learning in materials genome initiative: A review. *Journal of Materials Science & Technology*, *57*, 113-122. https://doi.org/10.1016/j.jmst.2020.01.067

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. *Journal of cheminformatics*, *7*(1), 1-17.https://doi.org/10.1186/s13321-015-0092-4

Martínez, M. J., Razuc, M., & Ponzoni, I. (2019). MoDeSuS: A machine learning tool for selection of molecular descriptors in QSAR studies applied to molecular informatics. *BioMed research international*, *2019*. https://doi.org/10.1155/2019/2905203

Mauri, A., Consonni, V., Pavan, M., & Todeschini, R. (2006). Dragon software: An easy approach to molecular descriptor calculations. *Match*, *56*(2), 237-248. - DRAGON for Windows (Software for Molecular Descriptor Calculations), Version 5.5; Talete srl: Milan, Italy, 2007.

Meredig, B. (2019). Five high-impact research areas in machine learning for materials science. https://doi.org/10.1021/acs.chemmater.9b04078

Muller, C., Pekthong, D., Alexandre, E., Marcou, G., Horvath, D., Richert, L., & Varnek, A. (2015). Prediction of drug induced liver injury using molecular and biological descriptors. Combinatorial Chemistry & High Throughput Screening, 18(3), 315-322.

Otsuka, S., Kuwajima, I., Hosoya, J., Xu, Y., & Yamazaki, M. (2011, September). PoLyInfo: Polymer database for polymeric materials design. In *2011 International Conference on Emerging Intelligent Data and Web Technologies* (pp. 22-29). IEEE. doi: 10.1109/EIDWT.2011.13.

Pal, S., & Naskar, K. (2021). Machine learning model predict stress-strain plot for Marlow hyperelastic material design. Materials Today Communications, 27, 102213. https://doi.org/10.1016/j.mtcomm.2021.102213

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. *Journal of Molecular Graphics and Modelling*, *38*, 137-147.https://doi.org/10.1016/j.jmgm.2012.04.006

Palomba, D., Vazquez, G. E., & Díaz, M. F. (2014). Prediction of elongation at break for linear polymers. *Chemometrics and Intelligent Laboratory Systems*, *139*, 121-131.https://doi.org/10.1016/j.chemolab.2014.09.009

Ristoski, P., Zubarev, D. Y., Gentile, A. L., Park, N., Sanders, D., Gruhl, D., Kato, L. & Welch, S. (2020). Expert-in-the-loop AI for Polymer Discovery. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (pp. 2701-2708). https://doi.org/10.1145/3340531.3416020

Roy, K., Das, R. N., Ambure, P., & Aher, R. B. (2016). Be aware of error measures. Further studies on validation of predictive QSAR models. *Chemometrics and Intelligent Laboratory Systems*, *152*, 18-33. https://doi.org/10.1016/j.chemolab.2016.01.008

Schustik, S. A., Cravero, F., Ponzoni, I., Díaz, M. F. (2021). Polymer informatics: Expert-in-the-loop in QSPR modeling of refractive index. *Computational Materials Science*, *194*, 110460. https://doi.org/10.1016/j.commatsci.2021.110460

Seymour, R.B. and Carraher, C.E. (1998). *Introducción a la química de los polímeros*. Editorial Reverté, 3ra ed. Barcelona, España.

Sha, W., Li, Y., Tang, S., Tian, J., Zhao, Y., Guo, Y., ... & Cheng, S. (2021). Machine learning in polymer informatics. *InfoMat*, *3*(4), 353-361.https://doi.org/10.1002/inf2.12167

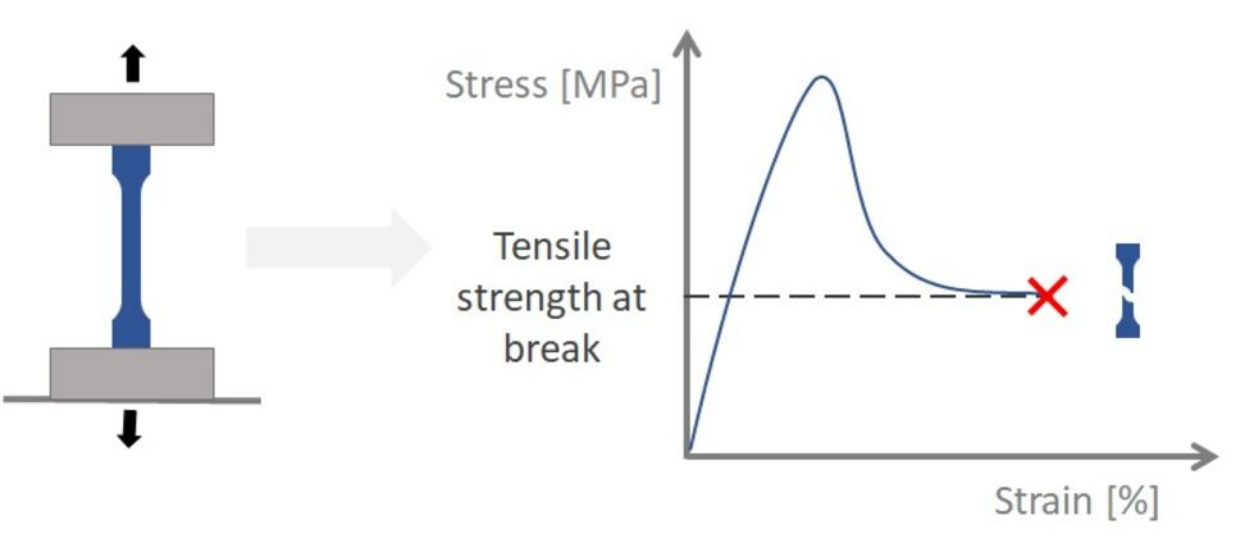
Taraji, M., Haddad, P. R., Amos, R. I., Talebi, M., Szucs, R., Dolan, J. W., and Pohl, C. A. (2017). Error measures in quantitative structure-retention relationships studies. *Journal of Chromatography A*, *1524*, 298-302. https://doi.org/10.1016/j.chroma.2017.09.050

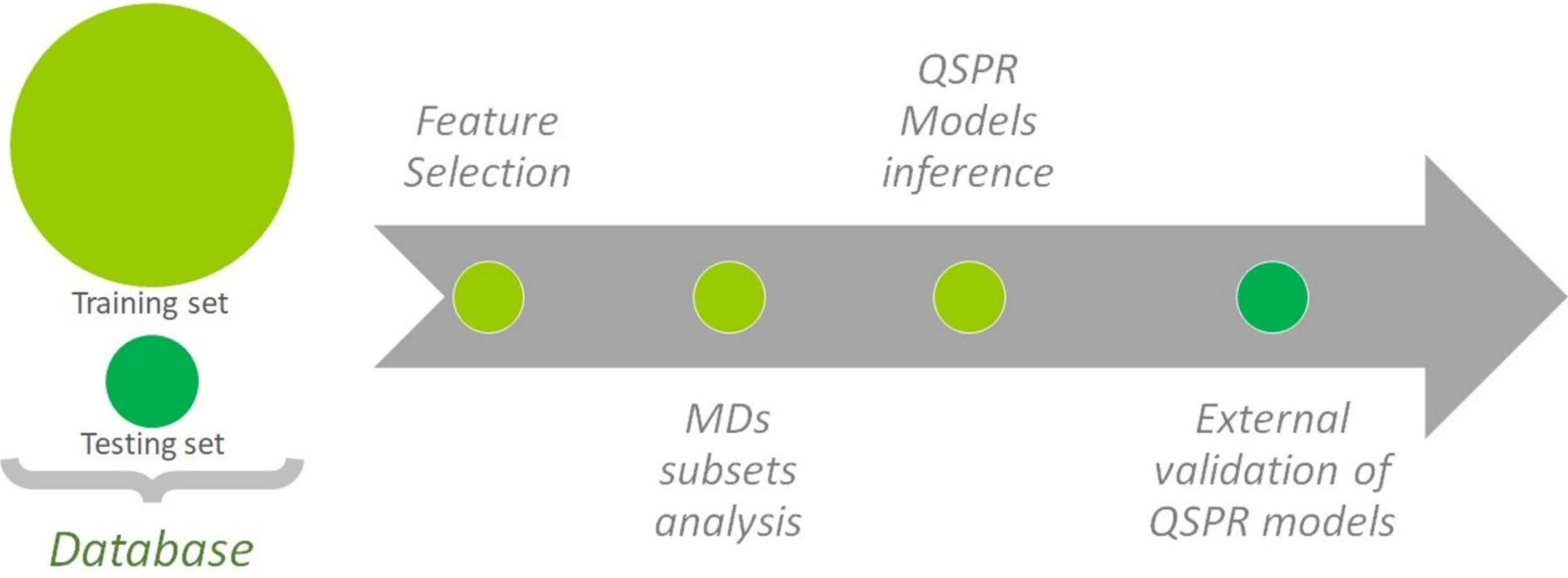
Theodosiou, A., & Kalli, K. (2020). Recent trends and advances of fibre Bragg grating sensors in CYTOP polymer optical fibres. *Optical Fiber Technology*, *54*, 102079.https://doi.org/10.1016/j.yofte.2019.102079

Tuan-Anh, T., & Zaleśny, R. (2020). Predictions of high-order electric properties of molecules: Can we benefit from machine learning?. *ACS omega*, *5*(10), 5318-5325. https://doi.org/10.1021/acsomega.9b04339

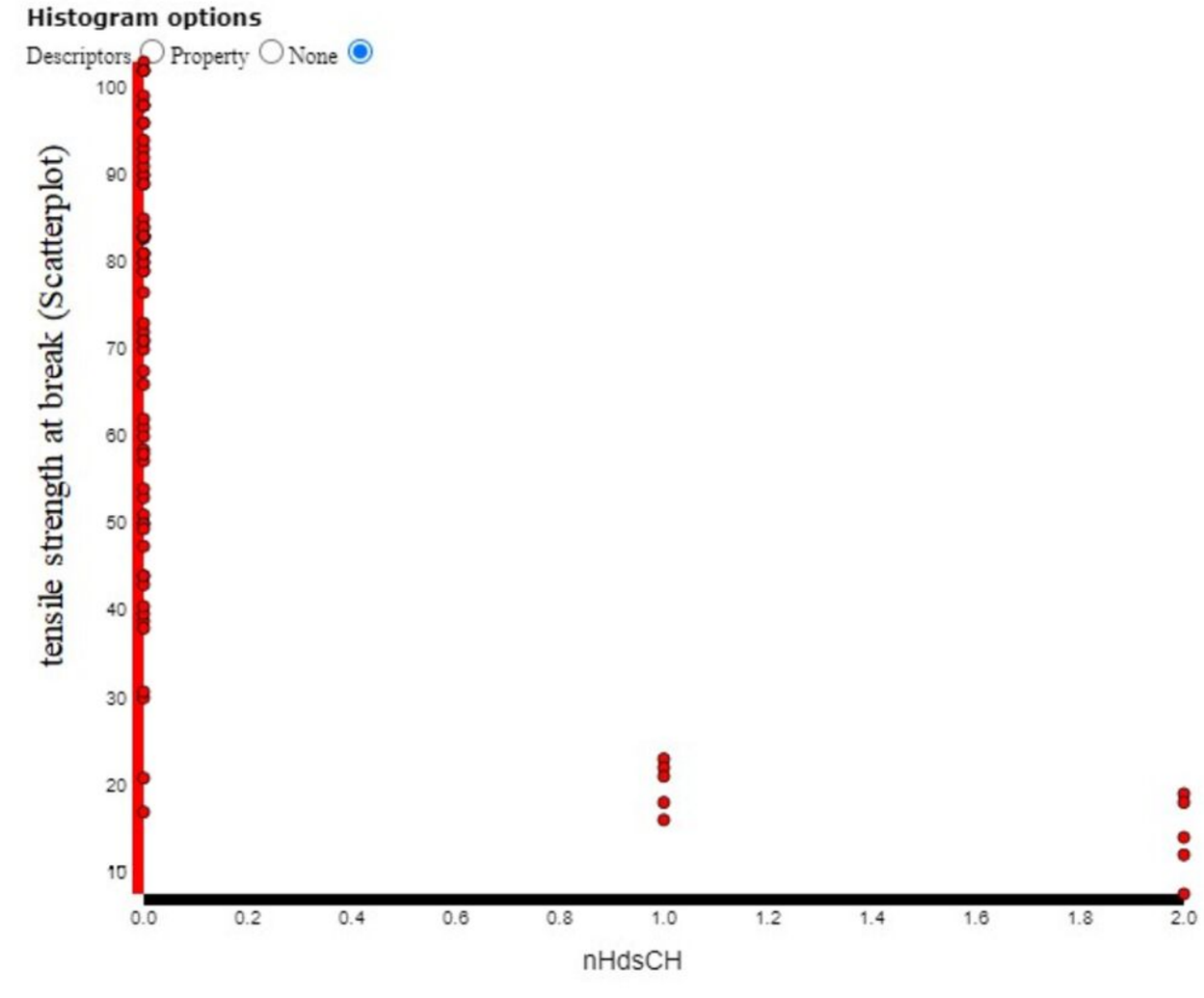
Ward, I. M., and Sweeney, J. (2004) An introduction to the mechanical properties of solid polymers, *England: John Wiley & Sons, Ltd*.

Wu, S., Kondo, Y., Kakimoto, Ma. *et al*. Machine-learning-assisted discovery of polymers with high thermal conductivity using a molecular design algorithm. *npj Comput Mater* **5,** 66 (2019). https://doi.org/10.1038/s41524-019-0203-2

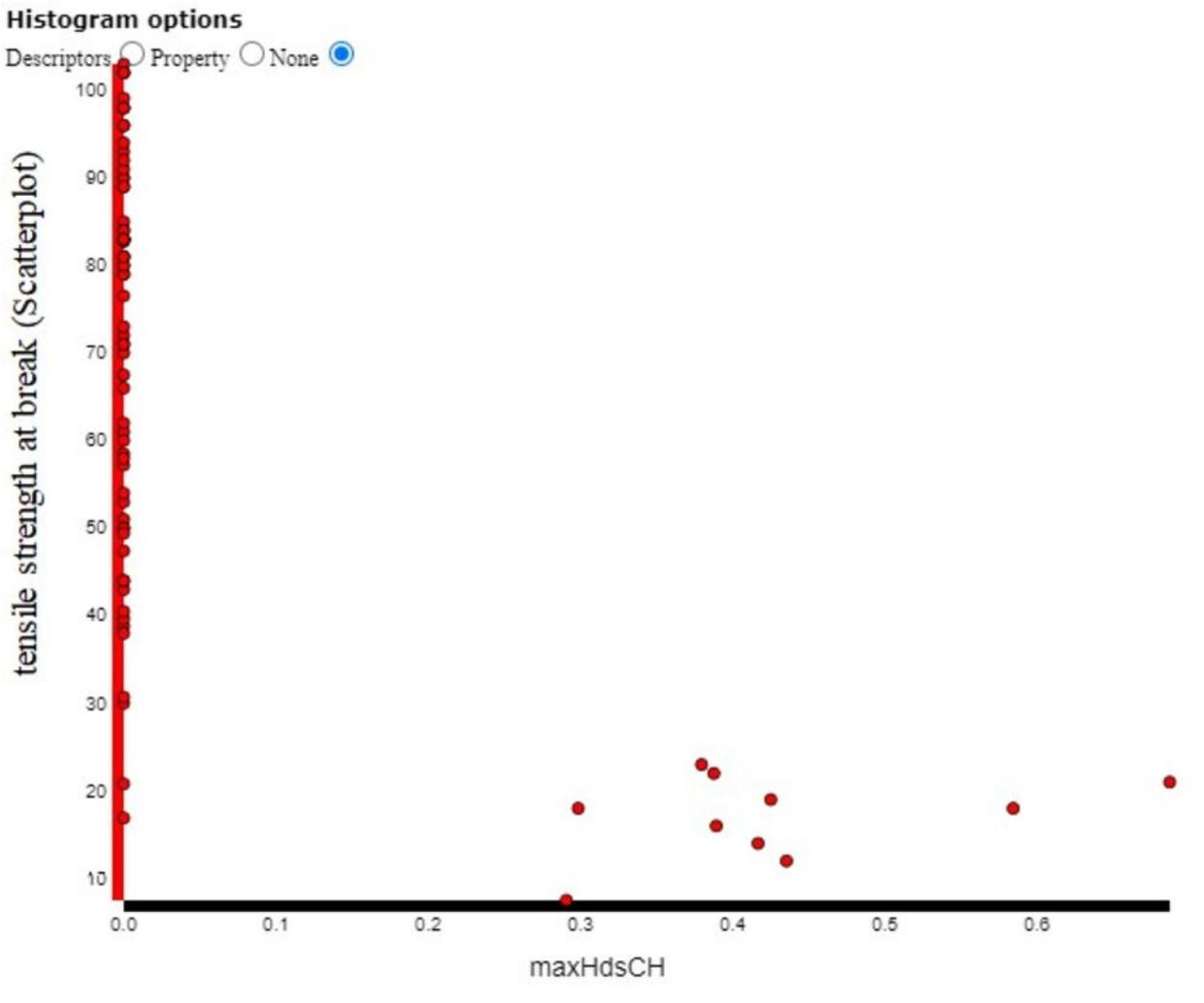

Stress [MPa]
Tensile strength at break
Strain [%]


Training set
Testing set
Database
Feature Selection
QSPR Models inference
MDs subsets analysis
External validation of QSPR models

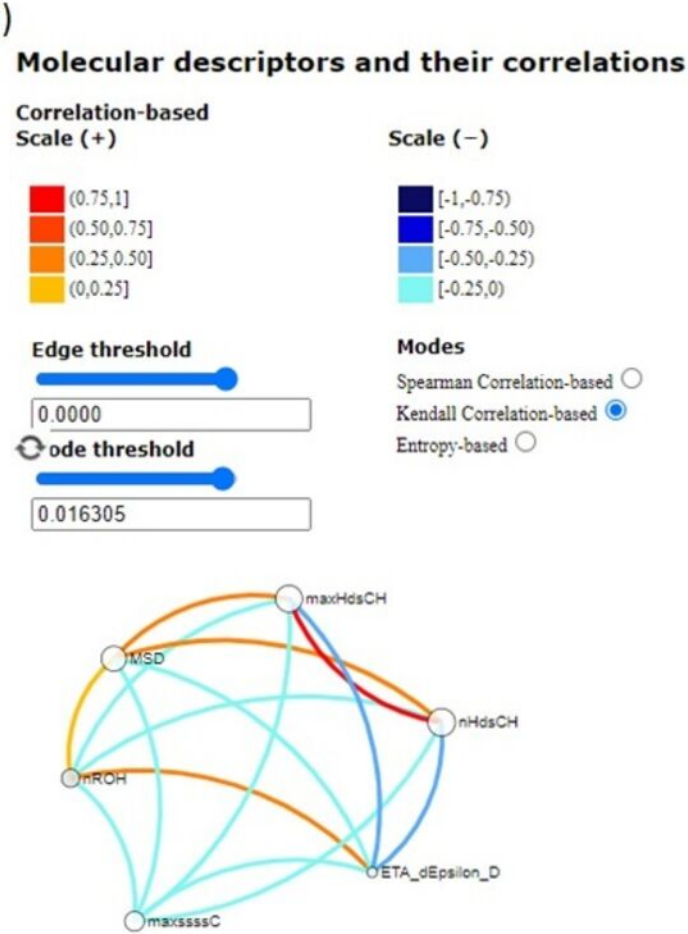
Histograms/Scatterplot



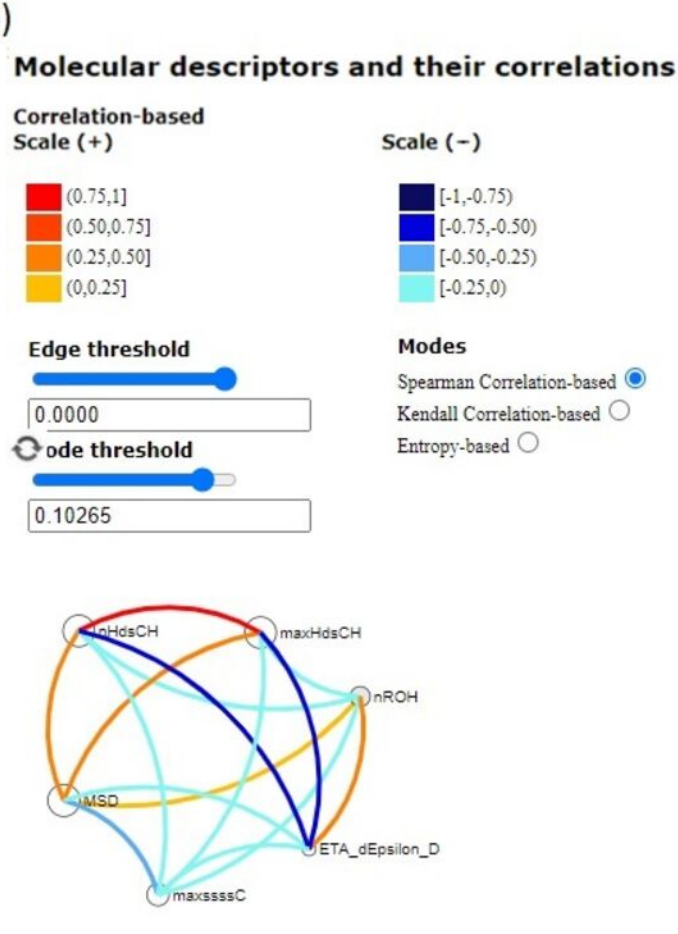
Histograms/Scatterplot



a)


Molecular descriptors and their correlations
Correlation-based
Scale (+)
Scale (−)
(0.75,1]
(0.50,0.75]
(0.25,0.50]
(0,0.25]
[-1,-0.75)
[-0.75,-0.50)
[-0.50,-0.25)
[-0.25,0)
Edge threshold
0.0000
ode threshold
0.016305
Modes
Spearman Correlation-based
Kendall Correlation-based
Entropy-based
maxHdsCH
MSD
nHdsCH
nROH
ETA_dEpsilon_D
maxssssC

b)


Molecular descriptors and their correlations
Correlation-based
Scale (+)
Scale (−)
(0.75,1]
(0.50,0.75]
(0.25,0.50]
(0,0.25]
[-1,-0.75)
[-0.75,-0.50)
[-0.50,-0.25)
[-0.25,0)
Edge threshold
0.0000
ode threshold
0.10265
Modes
Spearman Correlation-based
Kendall Correlation-based
Entropy-based
nHdsCH
maxHdsCH
nROH
MSD
ETA_dEpsilon_D
maxssssC

Histograms/Scatterplot