Prediction of the Glass-Transition Temperatures of Linear Homo/Heteropolymers and Cross-Linked Epoxy Resins

Chisa Higuchi, Dragos Horvath, Gilles Marcou, Kazunari Yoshizawa, Alexandre Varnek

Prediction of the Glass Transition Temperatures of Linear Homo/heteropolymers and Cross-linked Epoxy Resins

Chisa Higuchi,[a,b] Dragos Horvath,[a] Gilles Marcou,[a] Kazunari Yoshizawa,[b] and Alexandre Varnek[a*]

[a] Laboratoiry of Chemoinformatics, UMR 7140 CNRS-Univ. Strasbourg, 4 rue Blaise Pascal, Strasbourg 67000, France
[b] Institute for Materials Chemistry and Engineering, Kyushu University, Fukuoka 819-0395, Japan

***** *varnek@unistra.fr*

Abstract

This work proposes a unified approach to predict glass transition temperatures (T_gs) of linear homo/hetero-polymers and cross-linked epoxy resins by machine-learning approaches based on descriptors of reagents undergoing polymerization, represented in a formal way such as to encompass all the three scenarios: linear homo- and heteropolymers, plus network heteropolymers. The "formal" representation of reagents is a problem-specific, herein designed standardization protocol of compounds, sometimes differing from typical structure curation rules in chemoinformatics. For example, heteropolymers are represented by the two partner reagents, while homopolymers are depicted as formal "heteropolymers" with identical partners. The key rule proposed here is to choose "formal" monomers such as to minimize the number of marked atoms, involved in bonds being formed or changing bond order. Accordingly, carbonyl compounds are rendered as the less stable vinyl alcohol tautomer, following the same formalism as in olefin polymerization, in order to minimize the total number of formal polymerization mechanisms and herewith provide the most general framework encompassing a maximum of polymerization processes. ISIDA (In Silico design and Data Analysis) fragment counts with special status given to the "marked atoms" participating in the polymerization process were combined using "mixture" strategies to generate the final polymer descriptors. Three predictive models based on SVR (Support Vector Regression) are discussed here. After reproducing results of Katritzky et al. with a local model applicable only to linear homo/hetero-polymers, an epoxy resin-specific model applicable to both linear and network forms was built. Eventually, the general model applicable to all these families was constructed. In 12 × repeated 3-fold cross-validation challenges, it displayed the highest accuracy of Q^2 = 0.920, RMSE = 34.3 K over the training set of 270 polymers, and R^2 = 0.779, RMSE 35.9 K for an external test set of 119 polymers. GTM (Generative Topographic Mapping) analysis produced a 2D map of "polymer chemical space", highlighting the various classes of polymers included in the study and their relationship with respect to T_g values. The epoxy-specific and general models are publicly available on our web server: http://infochim.u-strasbg.fr/webserv/VSEngine.html.

1. Introduction

When a liquid is cooled sufficiently fast (supercooled), crystallization can be avoided and a glass is formed with an amorphous structure.[1] This occurs at the glass transition temperature (T_g), where the increasing viscosity reaches 10^{13} poise that the material can be regarded as solid.[1,2] Under constant pressure, temperature-dependent changes of liquid volume or enthalpy largely differ below and above T_g. It should be noted that the glass transition is only a kinetic phenomenon and does not mark a phase transition, so that glassy polymer is not in thermodynamic equilibrium below its T_g.[1,3] T_g is an important indicator of key processing and performance properties such as heat resistance, durability, and adhesion of the polymer since the heat capacity, the coefficient of thermal expansion, and viscosity are affected by glass transition. T_g is generally measured using Differential Scanning Calorimetry (DSC) or Dynamic Mechanical Thermal Analysis (DMTA).[4–5,7] Glass transition occurs over a relatively wide temperature range and depends on conditions such as measurement method, experimental period, and pressure under measurement.[4,8] T_g is also highly dependent on the structure of the polymer (crosslinking, chain stiffness), constitutive (additives, fillers, impurities), and conformation (stereo regularity). Therefore, it is difficult to uniquely determine by experiment, and the differences between the reported values of T_g in literature can be very large.

Numerous T_g prediction models have been already developed.[5–6, 9–24] Krevelen and co-workers developed the basis of group additive property (GAP) method.[9–10] The GAP method predicts polymer T_gs as a sum of calibrated contributions associated to typical substructures present in the monomers. Starting from here, many studies have been attempted to improve prediction accuracy and applicability domain by calibrating the contributions for additional substructures. Bicerano used a data set of 320 polymers[5] to build a model that combined a weighted sum of structural parameters along with the solubility parameter of each polymer. A linear regression procedure was used to produce a model with a standard deviation of 24.65 K and a correlation coefficient of 0.9749. However, no external data set compounds were withheld to validate this model. Most of these approaches gave relatively good predictive correlations, but they are only applicable for polymers containing chemical structural groups previously investigated.

At the end of the 1980's, more general QSPR (Quantitative Structure–Property Relationships)-based T_g predictions were developed.[14–17] Hopfinger and co-workers used molecular modeling to generate polymer descriptors (conformational entropy, mass moments, and intermolecular interactions) used to complement the group-specific terms in GAP models.[14–15] Waegell and co-workers approached modeling by using an Energy, Volume, Mass (EVM) QSPR model.[16–17] For linear and branched aliphatic acrylate and methacrylate polymers, the standard deviation from linear regression was 12 K with an R^2 value of 0.96. This model allowed the prediction of polymer T_g values not used for training of the original multiple linear regression, with an average absolute error of 10%. In the 1990's, T_g prediction focused on models without explicit knowledge of polymer 3D structure and without falling back to a predefined set of substructures of known contributions.[6,18–21] Katritzky and co-workers generated

over 400 constitutional, topological, geometrical, and quantum chemical descriptors directly from the molecular structure of the unit block in the polymer with the Comprehensive Descriptors for Structural and Statistical Analysis (CODESSA) program.[18–19] They showed that T_g divided by the molar weight of the repeating unit (M) improved squared correlation coefficient, resulting in an R^2 value of 0.946. In cross-validation of their training set, the T_g values for the 88 linear homopolymers, from the results of T_g/M prediction, with a standard error of 0.33 K mol g^{-1}.

All the discussed approaches for predicting T_g values were developed on the basis of homopolymers, despite the importance of epoxy-amine copolymers in commercial applications. There are, however, studies concerning a small number on amine-cured epoxy resins.[22–24] Bellenger et al. predicted about 40 T_gs of epoxy-amine copolymers based on the additivity law for copolymers and the contribution of cross-linked structures.[23] They have compared several physical and empirical approaches of the effect of cross-linking on T_g. Morrill et al. have predicted T_gs for epoxy-amine copolymers with the CODESSA program.[24] They succeeded to predict the T_g changes depending on the epoxy-amine molar ratio. However, their data set was rather limited.

So far, thus, most T_g prediction studies use molecular descriptors of the repeat unit in the polymer, which implies that they are only applicable to homopolymers –$(A)_n$– or linear 1:1 heteropolymers –$(AB)_n$– which can formally be regarded as "homopolymers" of unit AB. By contrast, repeating units cannot be always found in epoxy resins forming a network structure. As a result, T_g prediction of linear homopolymer and epoxy–amine copolymers were so far treated separately, because of assumed incompatibility of input descriptors: classical molecular descriptors of the repeat unit, on one hand, *versus* a combination of descriptors of copolymer reagents (and information about molar ratios) for epoxy-amine resins. Since epoxy–amine copolymers form a complicated 3D network structure, we herein advocate focusing on copolymer reagent structures (the polyamines and polyepoxides) and their molar fraction to generate "implicit" descriptors for the resulting polymer.

The herein proposed strategy is possible because the basis of QSPR is the neighborhood behavior principle: similar molecules are likely to have similar properties. Molecules are represented as points in a descriptor space, with the vector of descriptors defining their coordinates. In this space, similarity metrics are defined, allowing to quantitative assess the distances (dissimilarity) between molecules. The actual machine learning method – here, Support Vector Regression SVR – basically impacts on the concrete way in which the concept of "neighborhood" is defined in that space, ranging from straightforward Euclidean distance to sophisticated non-linear "kernels", but has no alter the fundamental principle of QSPR. In this work, "molecules" are actually monomer pairs, described by the concatenated ISIDA descriptor vectors. The QSAR principle is – for the best or the worst - completely oblivious of mechanistic aspects. Let us exemplify this on hand of the case of two heteropolymers (A,B) with glass transition temperature T_g, versus (A',B) at T_g'. Assume A and A' to be rather analogous monomers of the same type – for example, A and A' could be epoxides differing by the presence

of an additional ethyl group in A'. How does the model achieve prediction of T_g" for the polymer (A",B) – where A" features an methyl group as new substituent? It will infer that the point (A",B) is situated in descriptor space roughly halfway between reference points (A,B) and (A',B) – and therefore T_g"=(T_g+T_g')/2. Why does T_g change to T_g" upon addition of an ethyl substituent? It may be because its presence triggers a change of monomer reactivity and impacts on chain length and/or on the degree of branching of the polymer – or it may simply strengthen hydrophobic contacts between polymer chains, or it may trigger a combination of all the above effects. This question is – fortunately – not requiring an answer in QSPR, which simply assumes that, whatever impact an ethyl group has on T_g shift, the impact of the methyl group will be roughly half as strong. This enables prediction of the property of the methyl analogue without needing to synthesize it and characterize its chain length and branching. As far as the similarity relationship between the "formal" representations of polymers in descriptor space correctly renders the inter-species "distances" (in the sense that higher distances statistically correlate with increased T_g differences) the QSPR model will be predictive. By this approach, the propyl derivative will experience a T_g shift of 150% with respect to T_g-T_g'|, butyl – 200%, etc. This simplistic extrapolation is obviously prone to fail at a certain point, and the final model cannot be better than the data it was based on. If the butyl derivative is however part of the training set, providing experimental evidence that the local trend in T_g is actually flattening off with respect to alkyl chain length, then the QSPR model will gain in accuracy for both the propyl and pentyl derivatives, direct neighbors of the butyl reference point. Of course, physico-chemical descriptors like mean chain length and branching could be used in QSPR, with many potential benefits, but at very high cost: these would need to be either measured (limiting application to existing polymers only) or predicted, by an approach that remains to be defined. However, such descriptors would not fundamentally change the empirical nature of the approach, but simply redefine the "formal" descriptor space. Perhaps the latter would support some metric leading to statistically more robust models – but the gain in predictive power is unlikely to compensate for the loss of applicability to not yet synthetized species. The formal representation employed here was chosen such as to allow "virtual screening" of polymer candidates before their synthesis, by simple interpolation in a formal space based on monomer structures. It was assumed that the final state of the polymer is implicitly "encoded" in the structures of its monomers – which is a limiting hypothesis: chain length and degree of branching might also depend on polymerization reaction conditions. However, the explicit impact of conditions could not be considered here, because of lacking data. Otherwise, reaction conditions could be entered as novel descriptors[1, 2], in completion of the vector derived from monomer structures. This would still render the approach independent of knowledge of structural details, e.g. able to perform virtual screening.

In Silico design and Data Analysis (ISIDA) descriptors[25–26] monitor the occurrence of user-defined fragments in compounds, furthermore supporting "marked-atom"[27] strategies (where specified atoms are "marked" in the input structure and herewith acquire special status: the molecular fragments containing marked atoms will hence be counted separately from fragments occurring in the non-marked molecular "bulk"). This approach is perfectly suited to capture

structural information about the copolymer reagents, all while marking the atoms involved in the formation of new bonds during the polymerization process. Marked-atom descriptors of the copolymerization reagents can then be combined (with optionally weighing by their molar ratio) into a final descriptor vector of the copolymer. This approach is however not limited to epoxy-amine copolymers, but also applies for linear 1:1 heteropolymers –(AB)$_n$– of other chemical classes, thus opening the perspective of a more general T_g prediction model. For example, polyamides will be described by the combined descriptors of the dicarboxylic acid with marked carboxyl carbon and diamine with marked amino N atoms. Eventually, the present study undertook one more step towards generalization: it was assumed that homopolymers –(A)$_n$– can be formally described as *1:1 copolymers of A with itself, and herewith amenable to the same description protocol as implemented for genuine copolymers.* This original strategy enabled the first-time development of a unified, homo- and heteropolymer-competent T_g prediction model. Implicitly, the now possible fusion of the various local datasets exploited in previous studies lead to an unprecedented wealth of training information and external validation opportunities. Support Vector Regression (SVR)[28] was used for modeling, driving the selection of best-suited polymer description schemes (which result from the several considered marked-atom monomer fragmentation schemes and monomer descriptor combination strategies). Generative Topographic Mapping (GTM)[29–31] was employed to create a 2D map of "polymer chemical space", highlighting the various classes of (co)polymers included in the study, and being used to analyze the features and problems of the predictive model. Proceeding in three steps, T_g models of increasing generality are realized: (1) T_g prediction is performed on Katritzky's data set of homo- and linear –(AB)$_n$– copolymers, for which published modeling results serve as a benchmark to assess the pertinence of the herein proposed method. Next, (2) an epoxy-resin-specific model was developed, based on T_g data set of epoxy resin gathered from published literatures. Finally, (3) a global data set including both linear and network homo- and heteropolymer was compiled and used to fit the general, final model. Models obtained at steps (2) and (3) are publicly available on the QSAR prediction web server http://infochim.u-strasbg.fr/webserv/VSEngine.html.

2. Methods

2.1. Data Sources

To support the modeling workflow shown in Figure 1, T_g values of 389 polymers were collected from the literature. 270 of these constituted the "global" set for the general model. They contain

(i) 88 compounds from Katritzky's set.[19] These also served to build an alternative model to the published one, with the herein proposed technology for benchmarking purposes (by cross-validation),

(ii) 50 epoxy resins,[24,32–43] which also served for calibration of an epoxy resin-specific model, and

(iii) 132 homo- and heteropolymers from Bicerano et al.,[5] which only contributed to the global set (no "local" model was fitted for these). These were selected because they included completely novel chemotypes shown to fall outside the Applicability Domain (AD) of the local model (i).

The remaining 119 polymers, composed of 102 linear and 17 epoxy-amine copolymers were kept apart, as test sets for external validation. The 17 epoxy resins represent the novel compounds listed in ref 20 but not already present amongst the 50 compounds mentioned above (ii). They served both in the test set of the global model as well as for the epoxy-resin-specific model (ii). The 102 linear homo/heteropolymers stem from Bicerano's article and were kept as external test after exclusion of the 132 items which were mandatorily part of training set as chemically complementary to Katritzky's polymers.

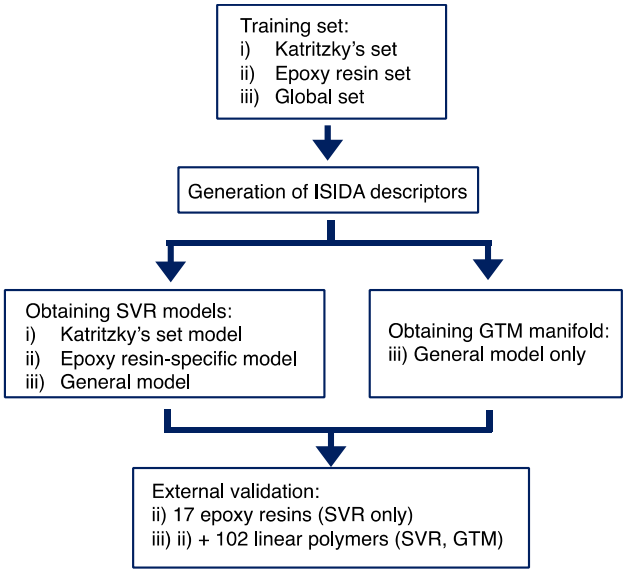


Figure 1. Workflow for the modeling of the T_g for homopolymers and copolymers.

2.2. Data preparation

The training data used in the present work was compiled from the various precursory articles mentioned above. However, given the herein targeted goal of a maximal generality model, structural data had to be significantly reorganized and standardized to fit our purposes. As already mentioned, the most general case is represented by heteropolymers – thus, the input required for modeling must contain the structures of the two monomers involved in copolymerization (dot-concatenated SMILES of the two species must be prepared for input). Implicitly, in homopolymers the structure of the only monomer "copolymerizing" with itself must also be reported twice. As the studied polymers result from diverse chemistries, there is no clear rule to define which of the two monomers must be reported first in the SMILES pair –

therefore, the order in which the two SMILES are concatenated is irrelevant. However, another key request is to report (in the second column of SMILES file), the molar ratio of the first monomer in the SMILES pair. For example, in a network epoxy-amine resin incorporating two moles of triamine A for three moles of diepoxy compound E, the input line can be either "E.A 0.6" or "A.E 0.4", where 0.6 is the molar fraction of E (3/5) and 0.4 is the one of A (2/5). For a homopolymer of monomer M, the input line will invariably be "M.M 0.5"

Before employment in model building, monomer structures must undergo structure standardization. Since the ultimate goal of this work was to achieve publicly available models operating on our multipurpose QSAR prediction server, submitted structures will necessarily undergo the thereon implemented "classical" standardization protocol (removal of counterions and mixtures – this specific option can and must be toggled off to allow processing of above-mentioned input files, e.g. standardize every mixture component –, conversion to "basic" aromatic form, split-charge nitro groups, etc.). For this reason, it is required to enter the "formal structures" of monomer reagents rather than structures of unit fragments in the polymer chain (with unsatisfied valences). These "formal structures" of the reagents are those atom-marked representations of the reagents which are easiest to convert to the polymeric form (with a minimal rearrangement of bonds). Polyethylene $–[CH_2\text{-}CH_2]_n-$ can be obviously derived from the structure of it monomer, ethylene – which coincides with the "formal structure" [CH2:1]=[CH2:1] to be used (note ":1" represents the mapping labels associated to the atoms connecting to other monomers – the same map label "1" can be used for all atoms involved in polymerization). However, as shown in Figure 2, $–[CH_2\text{-}CH(OH)]_n-$ is the polymerization product of acetaldehyde,

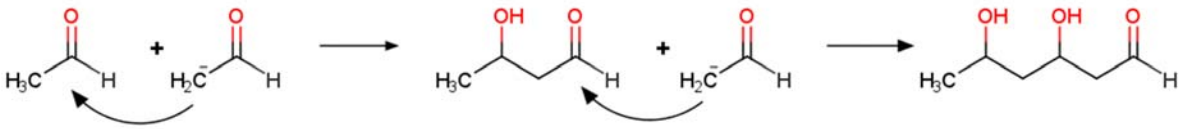


Figure 2. Polymerization of acetaldehyde.

a reaction proceeding by the addition of the carbanion resulting from α-proton extraction by a base to the carbonyl group. Formally it is nevertheless easier to describe this polymer as "polyvinylalcohol" stemming from [CH2:1]=[CH2:1]O. The key rule adopted in this work is to minimize the number of marked atoms, involved in bonds being formed or changing bond order. With vinyl alcohol, only the two carbons need to be marked. By contrast, taking acetaldehyde as such for monomer would require flagging of both carbons *and* the carbonyl O. Using this concept of "formal structures" for monomers, even copolymers in which the chain unit stems from three molecules may be described by a pair of formal structures. For example, above-mentioned polyvinylalcohols may react with another aldehyde, forming 1,3-dioxane rings as stable acetals. The product (Figure 3) may nevertheless be described in a way that is compatible with modeling constraints, by assuming the two "formal" copolymer structures to be (i) the vinyl alcohol and (ii) the vinylalcohol hemiacetal of the ring-closing aldehyde. Both all the "ethylene" carbons (as responsible for the C-C concatenation) as well as the vinyl alcohol O and the

hemiacetal carbon C(O)(X)OH need to be marked. The latter couple is responsible for dioxane ring closure – with elimination of water, which corresponds to the unmarked -OH of the hemiacetal.

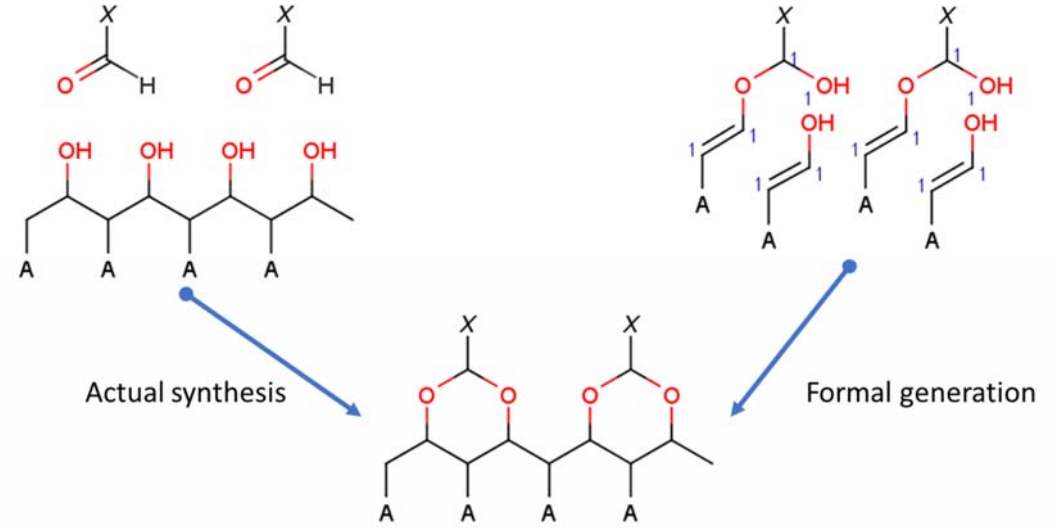


Figure 3. Actual synthesis and formal representation of a polymer containing a 1,3-dioxane-based unit resulting from ring closure by acetal formation in a reaction following the formation of the polyvinylalcohol chain. Formally, this can be described as the "copolymerization" of a vinyl alcohol molecule with the hemiacetal formed by another vinyl alcohol molecule and the ring-closing aldehyde.

As could be seen in the above-mentioned example, some polymerization processes (the archetypical ones being polyamide or polyester synthesis) involve elimination of some leaving groups (typically -OH, e.g. formation of water). This leaving group is kept in the reacting monomers, even though it will be absent in the actual polymer for which prediction of T_g is attempted. Owing to the fact that a leaving -OH group is by definition connected to a marked carbon participating in reaction, this signature can be allegedly exploited by the machine learning algorithm to differentially treat leaving -OH by contrast to regular hydroxy groups in the polymer. However, leaving groups other than -OH (it is chemically possible to obtain polyamides by reacting diamines with diacyl chlorides, with HCl as coproduct, for example) should be rendered as -OH in "formal" monomer structures, for coherence (not following this rule will place the input structure outside the model's applicability domain). Note that Bicerano's set also feature two "atypical" polytriazine imide-based structures: these were ignored in the present study. Albeit they could be formally represented according to a scheme similar to Figure 3, they were excluded from the study because they are radically new chemotypes, and two of them are clearly not enough to allow any meaningful learning of specific features of this polymer class.

The above-mentioned simple rules are not meant to define an exhaustive protocol for the rendering of arbitrary polymer structures – a still open problem in chemoinformatics. Note that in QSPR the representation of modeled items need not be "correct" from a physical point of view but must be neighborhood-behavior compliant and, to this purpose, coherent to the ad-hoc

representation rules established as a constitutive part of the model. The herein advocated rules are not necessarily the best, nor are they generalizable to any polymers – they simply have the merit to allow the unification of so-far distinct modeling problems into a common framework. Any further evolution of the model may imply an evolution of the so-far established rendering rules, in expectation of the development of a general-purpose polymer rendering system and the advent of large-scale polymer structure-property databases. So far, the rules are not general, but merely coherent and self-sufficient (in the sense that classes of polymers not representable by these rules are not part of this training set, so would be by definition out of its applicability domain). The user wishing to submit polymer structures for T_g prediction on our web server is therefore encouraged, if in doubt about how to render an input structure, to seek for similar entries in the training set provided in Supplementary Material as an example.

2.3. ISIDA (In Silico design and Data Analysis) descriptors

To generate fragment descriptors, ISIDA Fragmentor 2017 (see Supplementary Material) was applied to each reacting monomer, as rendered after the standardization step. ISIDA descriptors are topological fragments descriptors based on 2D chemical structures.[25–27] Each element of the vector corresponds to the number of occurrences of an associated substructure, where the considered substructures are defined by the user, in specifying key parameters. The most important key parameter is the type of considered fragments (I – linear sequences, II – atom-centered fragments, III – topological triplets). Upper and lower bounds for the considered fragment sizes are also mandatory: considered options ranged from 2 to 15 for sequences, from 1 to 5 for atom-centered fragments, and from 3 to 7 for triplets. The following options were also used at choice: charges on atoms (FormalCharge), accounting for the terminal atoms of a fragment exclusively (AtomPairs), exploring all possible paths instead of shortest paths (AllPaths) or restricted paths (Restricted). All employed fragmentation schemes generated both default fragments and specific fragments containing marked atoms (marked-atom strategy #3). For example, in sequence counts, enabling the marked-atom strategy implies that the number of propyl fragments CCC which do not contain any marked atoms will be counted separately (assigned to a distinct vector element) from propyl moieties with a terminal ([C:1]CC) or respectively central (C[C:1]C) marked C, which have each a distinct, dedicated vector element of their own. Refer to the above-cited Fragmentor manual for technical details about these options. A total of 42 different fragmentation schemes were considered, in order to select the best suited one for modeling (their list, following standard Fragmentor nomenclature, is provided in Supplementary Material).

The actual polymer $–(A_xB_{1-x})_n-$ is eventually described by combined descriptors of their monomers in eq (1), with the first elements of the descriptor vector stemming from the summing, and the last ones from the absolute differences of (molar ratio-weighed or not) monomer descriptors.

$$\overrightarrow{D^x} = \left[xD_{A,i} + (1-x)D_{Bi}, i = 1 \ldots N; |xD_{Ai} - (1-x)D_{Bi}|, i = 1 \ldots N\right] \quad (1)$$
$$\overrightarrow{D^0} = \left[D_{A,i} + D_{Bi}, i = 1 \ldots N; |D_{Ai} - D_{Bi}|, i = 1 \ldots N\right]$$

where D_A and D_B are the descriptor values of the individual monomers A and B, respectively, while x is the molar fraction of the first listed monomer, A and N the dimension of descriptor D in the chemical space containing all monomers, irrespective of their reactive class. Polymer descriptors will thus have a maximal dimensionality of 2N – and typically much less, noting that in the case of homopolymers the absolute difference contributions will systematically be zero. Since, for each of the 42 different fragmentation schemes applied to monomers, the two distinct combination strategies – with and without accounting of molar ratio – are applied, a total of 84 distinct descriptor spaces competed in the evolutionary strategy to be selected as the best support for optimally cross-validating Support Vector Machine models, *vide infra.*

2.4. Building and validation of the models

SVR models were built and validated using the ε-SVR algorithm implemented in the libSVM package.[45] Optimally parameterized SVR models, including descriptor choice as a key degree of freedom, were built according to the evolutionary procedure,[46] which provides both descriptor space selection and optimization of the operational parameters (epsilon, kernel type, cost, gamma) of the SVR method. The SVR models have been built for homopolymer, copolymer, and general data sets.

The predictive performance of the SVR models has been estimated by squared determination coefficient calculated in three-fold cross-validation (Q^2) repeated 12 times after the data reshuffling (12 × 3-CV) and eventually on the external test set (R^2) which are shown in eq 2, and Root-Mean-Squared Error (RMSE) which is shown in eq 3.

$$Q^2(R^2) = 1 - \frac{\sum_{i=1}^{n}(Y_{exp,i} - Y_{pred,i})^2}{\sum_{i=1}^{n}(Y_{exp,i} - < Y >_{exp})^2} \quad (2)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(Y_{exp,i} - Y_{pred,i})^2}{n}} \quad (3)$$

Here Y_{exp} and Y_{pred} are experimental and predicted values of T_g respectively, n is the number of data points, while $<Y>_{exp}$ is the mean of experimental values.

2.5. Generative Topographic Mapping

Generative Topographic Mapping (GTM) is a nonlinear mapping method used for data visualization originally described by Bishop.[29] The approach is basically a fuzzy-logics-driven generalization of Self-Organizing (Kohonen) maps[3]. A six-page brief introduction to this relatively new technology which is steadily gaining visibility in recent chemoinformatics publications is provided as Supplementary Material for the interested reader. In GTM, a 2D

latent space (called manifold) is embedded into the descriptor space. The manifold represents a grid of k × k nodes; each node is mapped in the initial descriptor space using the mapping function y(x, W). An "item" (here – a polymer depicted by its ISIDA descriptor vector) will be considered to "reside" on one or several of the k × k nodes that are closest to its descriptor space position. The fuzzy (real-number) truth value of the assumption "item n is a resident on node k" is termed "responsibility" R_{nk}, meaning that – unlike in Kohonen maps – an item may be a "time-sharing" resident of several nodes, such that $\sum_k R_{nk} = 1$. The mapping function is given as a grid of m × m radial basis functions (RBFs). To build a GTM-based QSAR model, the weighted average of properties of all molecules associated with any particular node is used to "color" the manifold according to that property, achieving a meaningful separation of items with different properties, or assignable to different classes. Here, the map parameters were tuned in order to achieve maximization of the separations of the different polymer classes as given in the literature[5,9] (Table 1, listing 17 distinct polymer types assigned to both training and test items). Map tuning followed the evolutionary procedure already described, using only the five descriptor spaces employed in SVM models as potential candidates for the GTM descriptor space and addressing the classical tunable GTM parameters (the number of RBF kernels, the number of grid points, the width factor of radial basis functions, and the regularization coefficient). The global model training set served both as frame set (items to guide the fitting of the manifold in descriptor space) and selection sets (providing items to be optimally separated on the map – here, by chemical class). The optimality criterion was the mean ability to separate (balanced accuracy; BA) members from non-members of the 10 most well-represented polymer classes (with at least 10 examples in the global model training set) following the cross-validated projection of the global training set on the current manifold. Once an optimal manifold in the above-mentioned sense was selected, it was also "colored" by T_g values, leading to a coherent landscape with "red" and "blue" zones populated by high and low T_g polymers, respectively. White zones represent unpopulated areas.

Table 1. 17 classes of polymer and numbers of polymers in train/test set for each class.

	Classes	Numbers of data	
		Train	Test
1	Epoxy resin	50	17
2	Polyolefin	15	3
3	Polystyrene	13	36
4	Polyvinyl	24	10
5	Polyacrylic	30	13
6	Polyhalo-olefin	9	2
7	Polydiene	5	5
8	Polyoether	17	6
9	Polysulphide	2	0

10	Polyester	12	11
11	Polyamide	4	15
12	Polyimide	20	0
13	Polyamide-imide	1	0
14	Polycarbonate	26	1
15	Polyimine	20	0
16	Silicon-containing polymer	14	0
17	Polyxylene	8	0

3. Results and Discussion

3.1. Reproducibility of Katritzky's results by the proposed modeling strategy

The results of the Katritzky set model are shown in Table 2. The descriptor set producing the SVR model of maximal robustness (estimated by 3-CV Q^2 value) was based on atom-centered fragments of the length 1–3 (See Table S1). This model returned a Q^2 value of 0.727, RMSE of 34.3 K, with the worst misprediction error of 110.4 K. In previous works by Katritzky et al., T_gs for homopolymers calculated from predicted T_g/M values reported a Q^2 value of 0.754 based on 3-CV model, and a worst error of 111 K. Our results are consistent with Katritzky's, which means that SVR modeling with purely topological ISIDA fragment counts for homopolymers works as well as the more sophisticated model employing constitutional, geometrical, and quantum chemical descriptors. Although T_g/M was the property modeled by Katritzky, we can directly predict the T_g values. Most important, note that in Katritzky's work a "homopolymer"-specific strategy was used, focusing on the repeating unit in the polymer – which means that copolymers may only be predicted if they are 1:1 linear concatenation products of the two monomers. A polyamide is modeled as a "homopolymer" of the amide unit -C(=O)-A-C(=O)N-B-N- in Katritzky's approach, while the same species is rendered as a genuine heteropolymer in this work. The proposed descriptor scheme is thus robust in supporting simultaneous processing of genuine homopolymers and 1:1 linear copolymers, without the need to explicitly generate the repeating unit.

Table 2. T_g predictive accuracy from 12 × 3-CV models for Katritzky's, epoxy resin, and global training sets.

	Katritzky's set model	Epoxy resin-specific model	General model
Number of data points	88	50	270
T_g range /K	190–409	280–531	130–685
Q^2	0.727	0.864	0.920
RMSE /K	34.3	21.5	34.3
Max error /K	110.4	44.0	137.2

3.2. Epoxy resin-specific model

The results of the epoxy resin specific model are shown in Table 2. The descriptor set producing the SVR model of maximal robustness (estimated by 3-CV Q^2 value) was based on atom-centered fragments of the length 1–4 with AtomPairs option (See Table S1). This model performed well with a Q^2 value of 0.864, RMSE of 21.5 K, with the worst misprediction error of 44.0 K. In the previous work by Morrill,[21] the leave-one-out cross-validated coefficient of determination was 0.995, which is higher than our Q^2 result. The difference between the number of data set and the diversity of epoxy resin can be parts of the reasons why our Q^2 value were lower than the Morrill's result. 50 epoxy resins are contained in our data set, on the other hand, Morrill's data set has only 13 data points. Additionally, there was small diversity of epoxy resin in Morrill's data base because Morrill et al. applied only DGEBA for their prediction as a representative epoxy resin, while we have 6 kinds of epoxy resin in training set.

Comparing with Katritzky's set model (section 3.1), epoxy resin-specific model returned higher accuracy. There are two possibilities for the reason. (1) Because the number of data points decreased, regression model fitted on data points more exactly. Since if the model fits too exactly to a particular data set, the model fails to fit additional external test set, we have checked the reliability of our model by using scrambled T_g data set to avoid this risk. T_gs were randomly mixed to create no correlated epoxy resin–T_gs data set. The same procedure as epoxy-specific model of SVR modeling was applied to this data set, at least it is proved that epoxy-specific model was not overfitting model although the number of the data set was relatively small. (2) Basically epoxy–amine copolymers have similar structures; all of epoxies have epoxy group, most of them also contain benzene rings, and all of amines contain amine groups. Therefore, the prediction should be more accurate than Katritzky's data set which has big diversity and different chemical groups. The similarity of epoxy resin structures will be assured by GTM maps discussed in section 2.4.

A consensus model was generated from 5 models which showed high robustness to predict 17 epoxy resins of external test set. The 5 different models have several types of descriptor set; two of them contain the sequence of the length within 2–6 with FormalCharge and AtomPairs options, other two contain atom-centered fragments of the length within 1–3 with FormalCharge, and another model is based on the sequence of the length 2–7. Descriptors in former 4 models were not multiplied by molar ratio, while descriptors in the last model were multiplied by molar ratio (See Table S2). Prediction of the external test set have returned an R^2 value of 0.687, RMSE of 22.3 K, with the worst misprediction error 50.2 K. In the previous work by Bellenger et al.,[20] "The results, which are given in Table V, are generally in good agreement with the experimental data; the average error of the prediction being less than 3%." To compare with this, our results showed the average error of the prediction less than 4% which is relatively acceptable.

3.3. General model

To create the general model, 132 linear homo/heteropolymers set have been added to combined Katritzky's data set and epoxy resin set as training data. The results of the general model are shown in Table 2. This model performed with a Q^2 value of 0.920, RMSE of 34.3 K, with the worst misprediction error of 137.2 K. The plot of predicted T_g values versus experimental T_g values is shown in Figure 4. The descriptor set producing the SVR model of maximal robustness (estimated by 3-CV Q^2 value) was based on atom-centered fragments of the length 1–3 with FormalCharge option (See Table S1). This model returned higher accuracy than epoxy resin-specific model. The increasing number of data points must have affected Q^2 high accuracy, since Q^2 was improved even though the diversity of data points much wider than Katritzky's set model and epoxy resin-specific model.

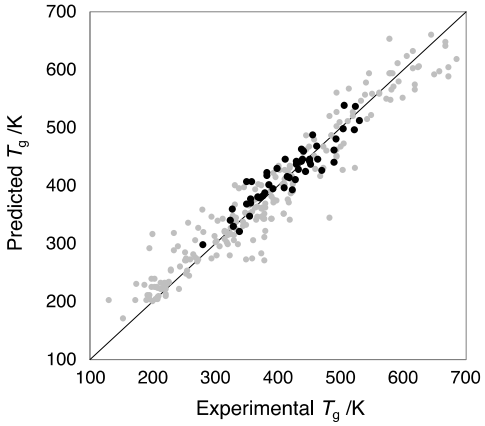


Figure 4. Plot of predicted T_g values versus experimental T_g values for 270 training data set consisted of 220 linear homo/heteropolymers (gray) and 50 epoxy resins (black) from 12 × 3-CV prediction.

The external test set consisted of 102 linear homo/heteropolymers from Bicerano's data set and 17 epoxy–amine copolymers from Bellenger's data set were predicted from a consensus model generated from 5 models which showed high robustness. The 5 models which is the base of the consensus model have atom-centered fragments of the length within 1–4, additionally one of them contained FormalCharge option (See Table S2). This external validation showed results of an R^2 value of 0.779, RMSE of 35.9 K, with the worst misprediction error of 127.1 K. The plot of predicted T_g values versus experimental T_g values is shown in Figure 5. We have divided the test set into linear polymer part and epoxy resin part to check each accuracy. The breakdown is shown in Table 3 together with external validation results of epoxy resin-specific model (section 3.2). As shown in Table 3, general model succeeded to improve the prediction accuracy of epoxy resin part of test set comparing with results of epoxy resin-specific model. It is interesting that the model could get better results for the epoxy resin prediction when homo/heteropolymers have been added into training set. According to both results of 3-CV for

training data set and external validation for test set, the greater diversity of polymer structures can be considered to lead to the better T_g predictions. This notwithstanding, the imprecision of prediction (which is in part an echo of the intrinsic experimental noise affecting training data) is certainly not good enough to consider this model as a reliable replacement of experimental T_g assessment. But QSPR models in general, and in all their domains of application – from drug design to material science – are rarely accurate enough to substitute themselves to experimental measure. Their goal is rather to act as filtering/prioritizing tools, selecting a small set of experimentally verifiable compounds with a maximal probability to contain molecules of desired properties, out of the very large pool of possible molecular structures. This approach is mainly intended to serve for the design of new materials – like, for example, considering a combinatorial matrix of E epoxides × A polyamines, running the *in Silico* prediction for all the $E \times A$ putative compounds, and selecting a pool of few dozen combinations amongst the ones predicted to have T_g within the desired range. The statistics of the model ensures that selected polymers will have a significantly enhanced probability to actually meet the desired T_g constraints, comparatively to randomly picked members of the $E \times A$ product matrix. The intended use makes it primordial to rely on a formal representation of polymers which can be applied to not yet synthesized, virtual species of unknown mean chain lengths, degree of reticulation, etc.

Generation of model based on such a diverse database consisted of homo/heteropolymer and cross-linked epoxy resins has never been attempted before, and we found this general model for global set can predict T_gs with better accuracy especially for epoxy resins.

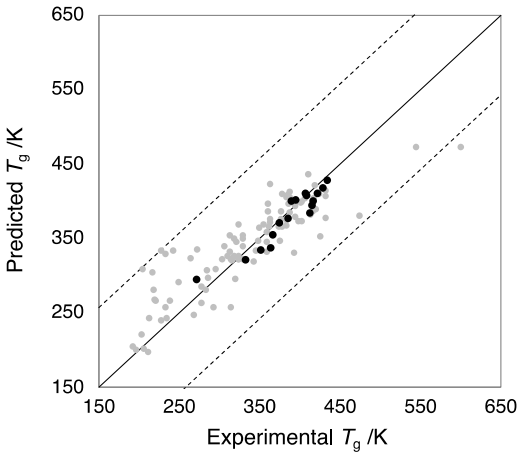


Figure 5. Plot of predicted T_g values versus experimental T_g values for 119 external data set consisted of 102 linear homo/hetero polymers (gray) and 17 epoxy resins (black).

Table 3. Results of external test set based predicted from consensus models of epoxy resin set and global set. For general model, results for the entire set and breakdown consisted of linear polymer and epoxy resin parts are shown.

	General model			Epoxy resin-specific model
	All	Linear polymer	Epoxy resin	
Number of data points	119	102	17	17
R^2	0.779	0.761	0.848	0.687
RMSE /K	35.9	38.3	15.6	22.3
Max error /K	127.1	127.1	29.6	50.2

The dramatic improvement of epoxy resin predictions by the global model compared to the dedicated model is an interesting example of inductive transfer of knowledge. The problem with the epoxy resin-specific model is its already very small training set. Data fusion allowed information conveyed by the linear polymers to complement the epoxy resin data, leading to overall better predictions. This is not trivial – if a training set containing two completely unrelated chemical classes (e.g. learning structure-activity relationships for the first chemical class is not helpful to understand the behavior of the second class), then building separate "local" models for each class (splitting the training set in a "divide and conquer" strategy) is the more rational approach. There are reasons to believe that a "divide and conquer" approach for T_g prediction may help some of the unaddressed challenges of the present work – conceiving, for example, specific models based on polymer subsets with T_g values obtained from a same strictly controlled experimental protocol may for example answer the problem of the noise affecting our training data. Unfortunately, "divide and conquer" only makes sense for initial training sets large enough to accommodate specific subsets still capable of supporting the fit of a robust local model. The alternative to "divide and conquer" is fitting of a unified model including experimental conditions – even experimental conditions describing the polymerization process, as these are likely to affect key parameters of resulting chains or networks – as additional fitable parameters in machine learning[1, 2]. Unfortunately, tracing back the reported T_g values to their experimental protocol in absence of centralized public polymer databases requires human resources beyond our capacities. At the given amount and quality of training data, the unifying approach adopted here was in our opinion the best strategy to adopt – as the above-illustrated results show. The key source for further improvement of the model will foremost be conditioned by an improvement of quantity and content of experimental information available for training.

3.4. Polymer space analysis by GTM visualization

16 GTMs highlighting the (fuzzy) separation of members from non-members of all classes except class 9 are shown in Figure 6. Class 9 has been removed since there are only two polysulfides in this data set, as shown in Table 1. The map supporting these fuzzy classification landscapes is based on atom-centered fragments of restricted atoms and bonds with a length of

1–3, with multiplication by molar ratio. All compounds (training set and test set of global set confounded) have been projected. Most classes are indeed well separated from each other (BAs for 10 most well-represented polymer classes in the cross-validated projection of the global training set were more than 0.88, such as 1.00 for class 1, 0.98 for class 5 and 12), which means that the selection of descriptors was well chosen by GA in SVR modeling. In particular, members of Class 1 and 5 are nearly perfectly separated from any other classes. This is not surprising for epoxy resins, which indeed stand out as the only (potentially) network polymers for which T_g data were available.

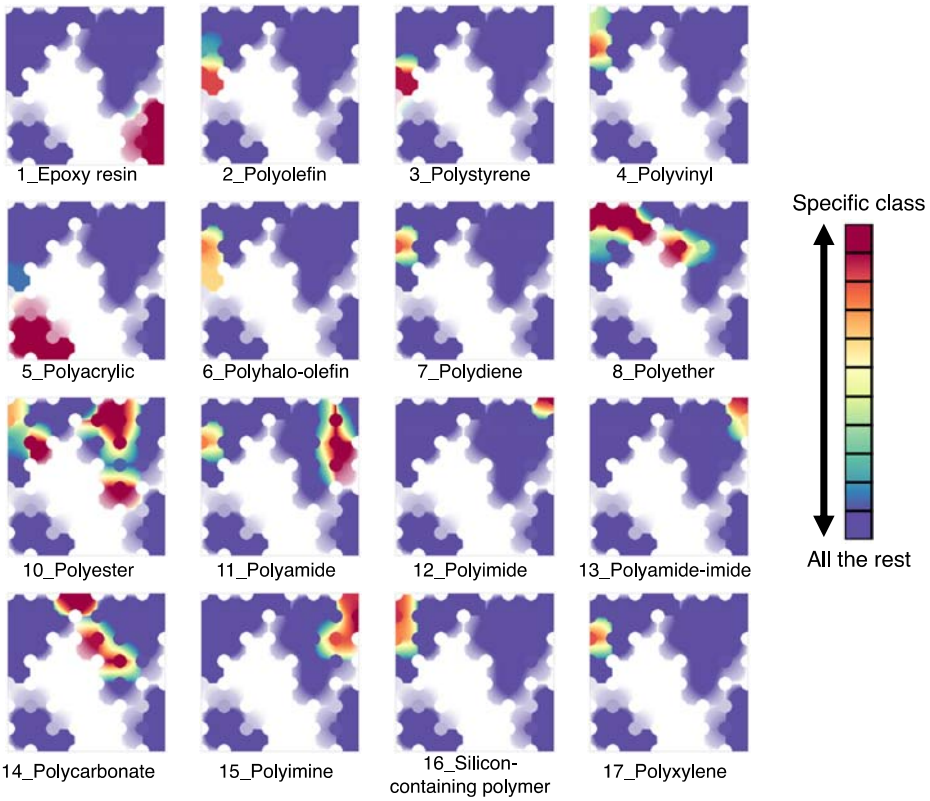


Figure 6. Polymer class landscapes highlighting the map positions of 16 relevant chemical classes, built on hand of on the global data set. The dark red area represents zones exclusively featuring resident compounds belonging to the specified class, while dark blue areas are populated by polymers of any other class *except* the specified one. At intermediate colors, class members "mingle" with representatives of other classes, and their relative occurrences are linearly related to the color scale to the right. Color transparency is modulated by the total number of residents (cumulated responsibility or "density").

Some polymer classes do however overlap to significant extents, but this can be perfectly well explained on behalf of the chemical similarity of structures, which transcends the rather rigid labelling by chemical class. Classes 2 to 7 are homochain polymers which are classified based on the type of side chains, while 8 to 17 are heterochain polymers which contains some

elements or chemical groups in main chain such as oxygen, sulfur, carbonyl group amide group, benzene rings. They are classified based on the type of main chain. Overlap on the map may arise because of the similarity of either main chains or side chains. Unsurprisingly, "polycarbonates" and "polyesters" are overlapping – carbonates are technically esters of the carbonic acid, after all. Also, the distinction between "Polyvinyl" compounds and "Polyhalo-olefins" is not clear – neither is the separation of these two classes. More interesting is the case of overlapping classes 7 (polydiene) and 17 (polyxylene). Clearly, one would expect aromatic polyxylenes to be distinct from polydienes – however, the formal monomers (Figure 7) used to describe polyxylene formation are, too, nothing but polyenes. Aromaticity is an unpredicted consequence of the reaction – therefore, the reagent-based similarity of the two classes – the underlying reason of the observed overlap – is not reflected in the final product. This is a limitation on (formal) monomer-based representation advocated here in order to unify modeling of T_g for both linear and network polymers. Regardless of this, the T_g landscape shown in Figure 8 indicate that compounds on node 3 and surrounding nodes have relatively low values compared to the global T_g range. (Strictly speaking, it should be noted that the T_g ranges are 171–293 K for class 7, 298–373 K for class 17, they are not overlapping.) On the opposite, significant diversity may occur within a chemical class. For example, compounds of class 10 spread on some nodes of the map because some of them have normal carbon chain, others have benzene rings with ester groups in the main chain (Figure 9).

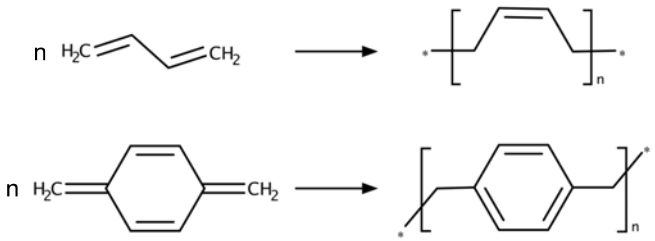


Figure 7. Polymerization of class 7 (polydiene) and class 17(polyxylene).

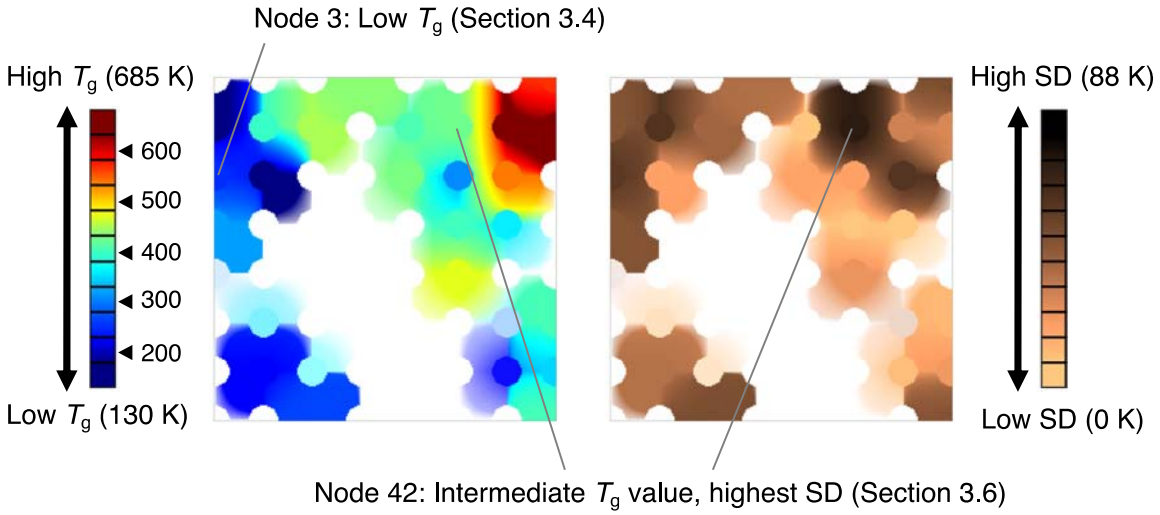


Figure 8. T_g landscape (left) and the landscape of the T_g standard deviation (SD) at each node (right).

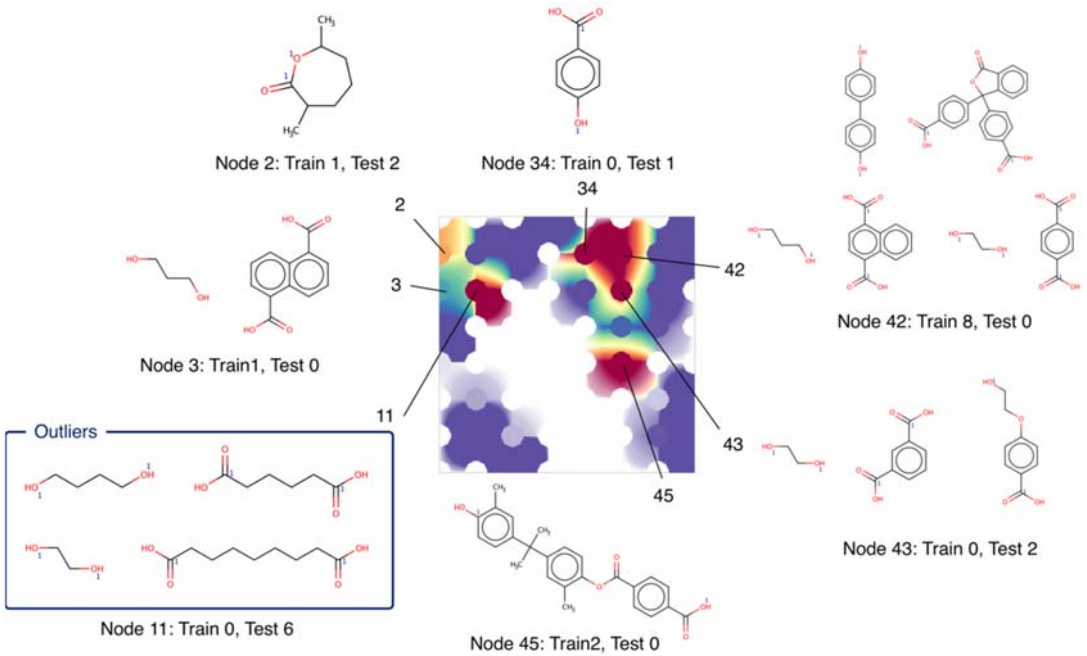


Figure 9. Representative structures, the number of training/test on each node of the map for class 10.

3.5. Understanding outliers within their chemical space context

According to the external validation of the SVR model in section 2.3, three outliers were evidenced: two polyesters (class 10) and one polyamide (class 11). Figure 9 and Figure 10 show landscapes of classes 10 and 11, representative structures, and the number of training/test set compounds residing in each node. Or, the nodes in which these outliers reside do not harbor any

training set compounds. In other words, if the GTM model would have been used as Applicability Domain delimiter, these outliers would have counted as excluded from the AD. Note that they did pass the less stringent test of "fragment control" used by the SVR model, but they do not stand out in terms of rare fragments – they rather stand out in terms of how these fragments are interconnected. Aliphatic diacids and diols are well represented in the training set – yet, their combinations are not. Table 4 shows the accuracy of T_g predictions for test set compounds of class 10 and 11 based on SVR general model in section 2.3. Although the worst misprediction error of class 11 is higher than that of class 10, RMSE and R^2 values of class 11 were much better than for class 10, regardless of the fact that class 10 has relatively more training set data than class 11. As discussed above, the imbalance coverage of polymer chemical space by train/test data is the reason for this. According to Figure 9, most of test data of class 10 is on node 11, which does not have any training item residing here. On the other hand, in Figure 10, most of test data resides in node 3, which is the residence node of one training item with a T_g value matching rather closely the ones of external compounds. As a consequence, their T_g predictions were quite accurate.

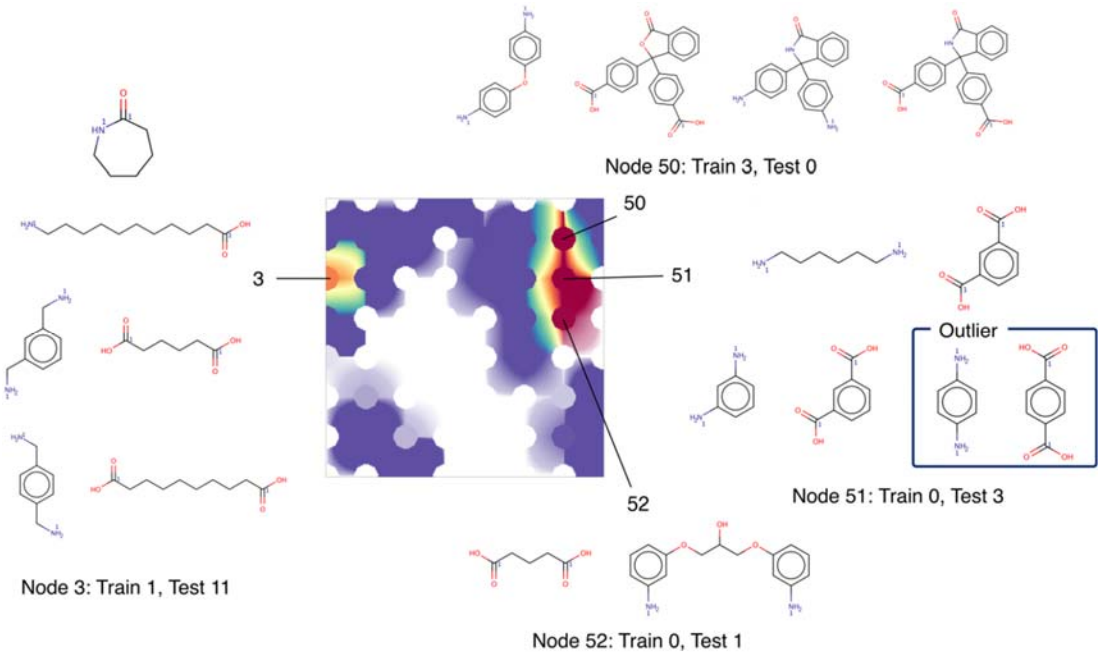


Figure 10. Representative structures, the number of training/test on each node of the map for class 11.

Table 4. External validation results from SVR general model in section 2.3. for each class 10 and 11.

Classes	Numbers of data	Accuracy of T_g prediction for test data on general model

		Train	Test	R^2	RMSE	Max Error
10	Polyester	12	11	–0.18	72.6	105.7
11	Polyamide	4	15	0.73	44.5	127.1

3.6. The T_g Landscape of Polymer Space

A key advantage of GTM is that a convenient manifold may be used to support any arbitrary property landscape. Even though the map was chosen for its propensity to separate chemical classes of polymers, it is nevertheless able to display a T_g landscape, with clearly separated high- and low-temperature areas. It is important to note upfront that, since the color code reflects averages of T_g values projected on each node, the interpretation of both high and low temperature nodes is straightforward: these chemical space zones are predominantly populated by polymer with extremal (high, respectively low) T_g values. By contrast, zones with "medium" T_g corresponding to intermediate spectral colors might arise either due to a local concentration of polymers with intermediate T_g values, or due to the cohabitation of low- and high-temperature polymers. To lift this uncertainly, the T_g landscape can be associated to the landscape of the T_g standard deviation at each node (Figure 8). The node with the highest divergence of T_g values of residing polymers can be clearly located in the Figure. The eight residents therein have a mean T_g of 435.0 ± 88 K. In all other nodes, resident polymers have better focused T_g ranges.

In Figure 11, low T_g areas (in blue) in the North-West (top left) mainly accommodate simple carbon polymers. Moving North, mean T_g values correspond to compounds contain oxygen, carbonyl groups, or carboxyl group. Eventually, the North-East is a high T_g-area, populated with polyamides and -imines. It is thus apparent that structure similarity as captured by the map implies similarity of T_g values. While the polymer class is *per* se a partial indicator of expectable T_g values, the map (and, of course, the predictive models) provide additional accuracy. In particular, epoxy resins form a well-separated class of polymers with strongly varying T_g values. This variability is well reproduced by the map, which provides a fine split of class 1 epoxy resins into subfamilies of higher and lower T_g, depending on their degree of reticulation. On the other hand, in the case of class 5, there can be special reason why they were well separated even though they have similar T_g ranges from top left compounds on the map, which can be interesting topic to investigate in detail for the future work.

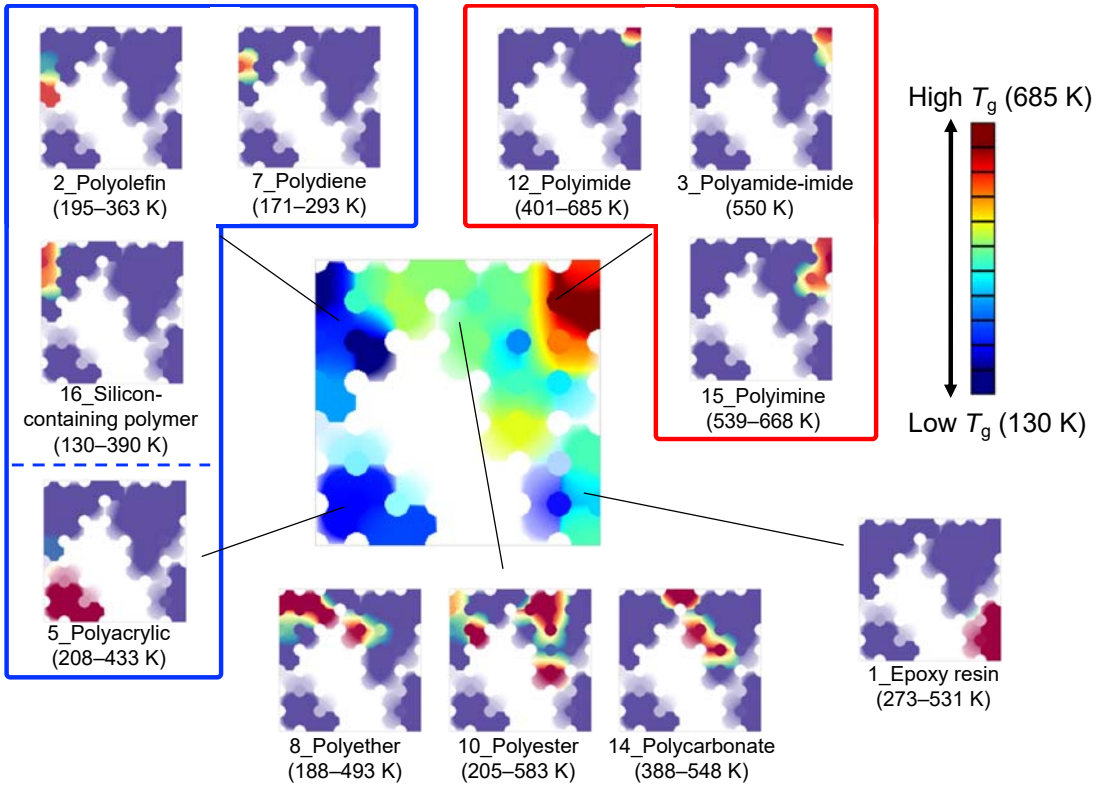


Figure 11. T_g landscape and representative maps on some areas depending on the T_g values.

4. Conclusions

The successfully addressed challenge of this work was to propose a unified framework for chemoinformatics modeling of the glass transition temperature T_g of both network and linear homo- and heteropolymers, which were traditionally addressed by distinct approaches. The key to solve this problem was the unified description paradigm of these polymers, by means of molar-ratio-sensitive "mixing" of atom-marked ISIDA fragment counts of the "formal" monomers – following typical mixture modeling strategies, where in homopolymers the single monomer is considered in 1:1 "mixture" with itself. This "formal monomer"-based strategy accommodates both linear and network polymers, while classical approaches based on descriptors of the repeating unit only work for linear homo- and 1:1 heteropolymers $–(AB)_n–$, i.e. "homopolymers" of repeating unit AB. "Formal" monomers are rendered according to herein defined, specific standardization rules – following not the actual chemical mechanism of polymerization, but aiming to minimize the number of simplest schemes that could be used to formally describe the polymerization process. For example, aldehyde polymerization is easiest rendered as the formal polymerization of vinyl alcohol, the unstable tautomer of the aldehyde. This may be mechanistically untrue but has the merit to describe these polymers in a similar way to "other" polyolefins. This unified rendering of all polymers as pairs of monomers (identical, for

homo- or different, for heteropolymers) was shown to apply even to some ternary polymers, if a judicious choice of formal monomers is made. The rule here is to minimize the number of marked atoms, involved in bonds being formed or changing bond order. While this rule is clearly established, it must be nevertheless pointed out that the herein performed standardization is semi-automatic and required human reflection and decision-making for specific cases. A fully automated implementation of rendering polymers by their "formal" monomers would require an additional technical development, which is not a priority knowing that, unlike drug-like molecules having structures that can be directly accessed from electronic databases, a universal standard for polymer databases is not yet established, making chemical name to structure conversion an unavoidably human intervention requiring step, anyway.

Starting from the file of pairs of monomers and molar ratio information, descriptor calculation, model building and prediction are fully automated, following standard QSAR procedures. In a first step, a focused approach excluding network epoxy resins was challenged to reproduce Katritzky's previous study and achieved comparable results. Specific modeling of epoxy resins (network or not) also proved to be robust. Eventually, the general model covering both linear homo/heteropolymers and cross-linked epoxy resins showed the highest accuracy (Q^2 = 0.920, RMSE = 34.3 K for training set of 270 polymers, and R^2 = 0.779, RMSE 35.9 K for external test set of 119 polymers) of three models. Especially, this model performed better predicting epoxy resins T_g than the dedicated, epoxy resin-specific model. The greater diversity of polymer structures has thus a significant impact in improving T_g predictions *across* polymer classes. Generation of models based on such a diverse database has never been attempted before, and it was shown to be helpful for improving predictions for small polymer families, where paucity of training data automatically limits the applicability domain of local, dedicated models.

Eventually, polymer space analysis using GTM landscapes highlighted several interesting insights. Outliers mispredicted during the external validation of the model, were shown to reside in chemical space zones with insufficient training data. GTM landscapes allow a clear separation of chemically distinct polymer families, but also highlighted an interesting case (polydienes *versus* polyxylenes) where monomer-based rendering was pushed to its limits – aromatization occurring during the polymerization process cannot be captured by monomer-based descriptors. The GTM may also harbor the landscape of the property of interest T_g herewith permitting the intuitive oversight of the association of polymer classes to glass temperature ranges.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website.

The formal SMILES representation of the compound set is provided in a multi-column text file (DataSets.txt) featuring smiles, molar ratio, experimental T_gs, polymer class, participation in either of the following subsets: training set of Katritzky's model, training set of Epoxy model, global training set, global external test set – see header line.

One document of the provided archive reports model-related information: the list of considered fragmentation schemes, out of which descriptor sets supporting SVR model of maximal robustness were selected, and the descriptor sets involved in the SVR consensus models.

In addition, the ISIDA Fragmentor manual is provided for the reader interested in in-depth understanding of ISIDA fragmentation schemes, whilst a short introductory document about Generative Topographic Mapping revisits the basics of this rather new but potent dimensionality reduction and mapping technique.

AUTHOR INFORMATION

Corresponding Author

* E-mail: varnek@unistra.fr

ORCID

Dragos Horvath: 0000-0003-0173-5714

Kazunari Yoshizawa: 0000-0002-6279-9722

Alexandre Varnek: 0000-0003-1886-925X

Notes

The authors declare no competing financial interest.

ISIDA GTM software is developed by the Laboratoire de Chemoinformatique Strasbourg and can be obtained upon request (visit http://infochim.u-strasbg.fr/spip.php?rubrique41).

ACKNOWLEDGEMENT

We thank Dr Fanny Bonachera for the help with GTM calculations. CH thanks Kyushu University for supporting her stay at the University of Strasbourg. KY acknowledges KAKENHI Grant numbers JP15K13710, and JP17H03117 from Japan Society for the Promotion of Science (JSPS) and the Ministry of Education, Culture, Sports, Science and Technology of Japan (MEXT), the MEXT Projects of “Integrated Research Consortium on Chemical Sciences”, “Cooperative Research Program of Network Joint Research Center for Materials and Devices”, “Elements Strategy Initiative to Form Core Research Center”, and JST-CREST “Innovative Catalysts”.

Reference

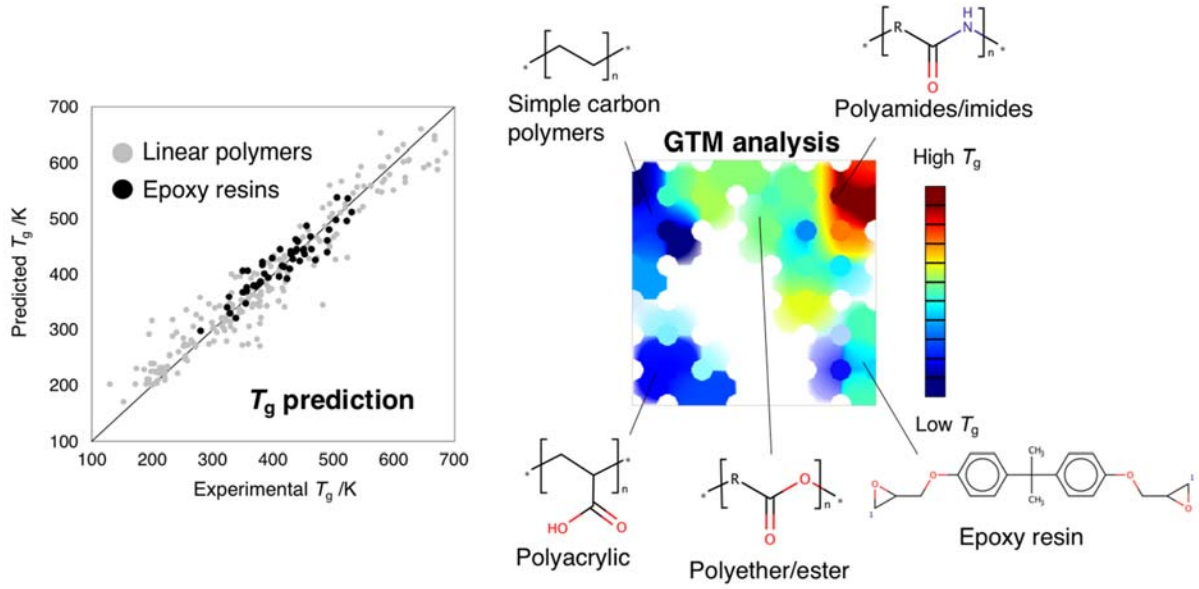
1. Geirhos, K.; Lunkenheimer, P.; Loidl, A. Johari-Goldstein Relaxation Far below Tg: Experimental Evidence for the Gardner Transition in Structural Glasses? *Phys. Rev. Lett.* **2018**, *120* (8), 85705.
2. Debenedetti, P. G.; Stillinger, F. H. Review Article Supercooled Liquids and the Glass Transition. *Nature* **2001**, *410*, 259–267.
3. Bhattacharya, S.; Suryanarayanan, R. A. J. Local Mobility in Amorphous Pharmaceuticals — Characterization and Implications on Stability. *J. Am. Pharm. Assoc.* **2009**, *98* (9), 2935–2953.
4. Sperling, L. H. *Introduction to Physical Polymer Science*, 4th ed; John Wiley & Sons, Inc. New Jersey, 2006.
5. Bicerano, J. *Prediction of Polymer Properties*, 3nd ed.; Marcel Dekker, Inc. New York, 2002.
6. Mattioni, B. E.; Jurs, P. C. Prediction of Glass Transition Temperatures from Monomer and Repeat Unit Structure Using Computational Neural Networks. *J. Chem. Inf. Comput. Sci.* **2002**, *42* (2), 232–240.
7. Cheremisinoff, N. P. *Polymer characterization: laboratory techniques and analysis*; Noyes Publications, New Jersey, 1996.
8. Krause, S.; Gormley, J. J.; Roman, N.; Shetter, J. A.; Watanabe, W. H. Glass Temperatures of Some Acrylic Polymers. *J. Polym. Sci. Part A* **1965**, *3* (10), 3573–3586.
9. Van Krevelen, D. W. *Properties of polymers*, 4th completely revised ed.; Elsevier, Oxford, 2009.
10. Weyland, H. G.; Hoftyzer, P. J.; Van Krevelen, D. W. Prediction of the Glass Transition Temperature of Polymers. *Polymer* **1970**, *11* (2), 79–87.
11. Barton, J. M. Relation of Glass Transition Temperature to Molecular Structure of Addition Copolymers. *J. Polym. Sci. Part C* **1970**, *30* (1), 573–597.
12. Lee, W. A.; Establishment, R. A. Calculation of the Glass Transition Temperatures of Polymers. Part I. Homopolymers and Copolymers with Alkyl Side Chains. *Polymer* **1970**, *8*, 555–570.

13. Wiff, D. R.; Altieri, M. S.; Goldfarb, I. J. Predicting Glass Transition Temperatures of Linear Polymers, Random Copolymers, and Cured Reactive Oligomers from Chemical Structure. *J. Polym. Sci., Part B, Polym. Phys.* **1985**, *23*, 1165–1176.
14. Hopfinger, A. J.; Koehler, M. G.; Pearlstein, R. A.; Tripathy, S. K. Molecular Modeling of Polymers. IV. Estimation of Glass Transition Temperatures. *J. Polym. Sci. Part B Polym. Phys.* **1988**, *26* (10), 2007–2028.
15. Koehler, M. G.; Hopfinger, A. J. Molecular Modelling of Polymers: 5. Inclusion of Intermolecular Energetics in Estimating Glass and Crystal-Melt Transition Temperatures. *Polymer* **1989**, *30* (1), 116–126.
16. Cypcar, C. C.; Camelio, P.; Lazzeri, V.; Mathias, L. J.; Waegell, B. Prediction of the Glass Transition Temperature of Multicyclic and Bulky Substituted Acrylate and Methacrylate Polymers Using the Energy, Volume, Mass (EVM) QSPR Model. *Macromolecules* **1996**, *29* (27), 8954–8959.
17. Camelio, P.; Cypcar, C. C.; Lazzeri, V.; Waegell, B. A Novel Approach toward the Prediction of the Glass Transition Temperature: Application of the EVM Model, a Designer QSPR Equation for the Prediction of Acrylate and Methacrylate Polymers. *J. Polym. Sci. Part A Polym. Chem.* **1997**, *35* (13), 2579–2590.
18. Katritzky, A. R.; Rachwal, P.; Law, K. W.; Karelson, M.; Lobanov, V. S. Prediction of Polymer Glass Transition Temperatures Using a General Quantitative Structure-Property Relationship Treatment. *J. Chem. Inf. Comput. Sci.* **1996**, *36* (4), 879–884.
19. Katritzky, A. R.; Sild, S.; Lobanov, V.; Karelson, M. Quantitative Structure - Property Relationship (QSPR) Correlation of Glass Transition Temperatures of High Molecular Weight Polymers. *J. Chem. Inf. Comput. Sci.* **1998**, *38* (2), 300–304.
20. Cao, C.; Lin, Y. Correlation between the Glass Transition Temperatures and Repeating Unit Structure for High Molecular Weight Polymers. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (2), 643–650.
21. Liu, W.; Cao, C. Artificial Neural Network Prediction of Glass Transition Temperature of Polymers. *Colloid Polym. Sci.* **2009**, *287* (7), 811–818.
22. Lee, G.; Hartmann, B. Glass Transition Temperature Predictions in Some Epoxy Polymers. *J. Appl. Polym. Sci.* **1983**, *28* (2), 823–830.

23. Bellenger, V.; Verdu, J.; Morel, E. Effect of Structure on Glass Transition Temperature of Amine Crosslinked Epoxies. *J. Polym. Sci. Part B Polym. Phys.* **1987**, *25* (6), 1219–1234.
24. Morrill, J. A.; Jensen, R. E.; Madison, P. H.; Chabalowski, C. F. Prediction of the Formulation Dependence of the Glass Transition Temperatures of Amine-Epoxy Copolymers Using a QSPR Based on the AM1 Method. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (3), 912–920.
25. Varnek, A.; Fourches, D.; Horvath, D.; Klimchuk, O.; Gaudin, C.; Vayer, P.; Solov'ev, V.; Hoonakker, F.; Tetko, I.; Marcou, G. ISIDA - Platform for Virtual Screening Based on Fragment and Pharmacophoric Descriptors. *Curr. Comput. Aided-Drug Des.* **2008**, *4* (3), 191–198.
26. Ruggiu, F.; Marcou, G.; Varnek, A.; Horvath, D. ISIDA Property-Labelled Fragment Descriptors. *Mol. Inform.* **2010**, *29* (12), 855–868.
27. Ruggiu, F.; Solov'ev V.; Marcou, G.; Horvath, D.; Graton, J.; Le Questel, J. Y.; Varnek, A. Individual Hydrogen-Bond Strength QSPR Modelling with ISIDA Local Descriptors: a Step Towards Polyfunctional Molecules. *Mol.Inform.* **2014**, *33* (6–7), 477–487.
28. Drucker, H.; Burges, C. J.; Kaufman, L.; Smola, A.; Vapnik, V. Support Vector Regression Machines. *Advances in neural information processing systems* **1997**, *9*, 155–161.
29. Bishop, C. M.; Svenseń, M.; Williams, C. K. I. GTM: The Generative Topographic Mapping. *Neural Comput.* **1998**, *10*, 215− 234.
30. Kireeva, N.; Baskin, I. I.; Gaspar, H. A.; Horvath, D.; Marcou, G.; Varnek, A. Generative Topographic Mapping (GTM): Universal Tool for Data Visualization, Structure-Activity Modeling and Dataset Comparison. *Mol. Inf.* **2012**, *31*, 301−312.
31. Gaspar, H. A.; Baskin, I. I.; Marcou, G.; Horvath, D.; Varnek, A. GTM-Based QSAR Models and Their Applicability Domains. *Mol. Inf.* **2015**, *34*, 348−356.
32. Cook, W. D.; Mayr, A. E.; Edward, G. H. Yielding behaviour in model epoxy thermosets – II. Temperature dependence. *Polymer* **1998**, *39* (16), 3725–3733.
33. Chen, M. C.; Hourston, D. J.; Sun, W. B. Miscibility and Fracture Behaviour of an Epoxy Resin-Bisphenol-A Polycarbonate Blend. *Eur. Polym. J.* **1992**, *28* (12), 1471–1475.
34. Lee, J. Y.; Shimb, M. J.; Kim, S. W.; Characteristics of the DGEBA/MDA system modified with glutaronitrile. *Mater. Chem. Phys.* **1996**, *44*, 74–78.

35. Knorr, D. B.; Yu, J. H.; Richardson, A. D.; Hindenlang, M. D.; McAninch, I. M.; La Scala, J. J.; Lenhart, J. L. Glass Transition Dependence of Ultrahigh Strain Rate Response in Amine Cured Epoxy Resins. *Polymer.* **2012**, *53*, 5917–5923.
36. Bellenger, V.; Dhaoui, W.; Morel, E.; Verdu, J. Packing Density of the Amine-Crosslinked Stoichiometric Epoxy Networks. *J. Appl. Polym. Sci.* **1988**, *35*, 563–571.
37. Carfagna, C.; Apicella, A.; Nicolais, L. The Effect of the Prepolymer Composition of Amino-Hardened Epoxy Resins on the Water Sorption Behavior and Plasticization. *J. Appl. Polym. Sci.* **1982**, *27*, 105–112.
38. Chang, T. D.; Carr, S. H.; Brittain, J. O. Studies of Epoxy Resin Systems: Part A: A Study of the Origins of the Secondary Relaxations of Epoxy Resins by Thermally Stimulated Depolarization. *Polym. Eng. Sci.* **1982**, *22* (18), 1205–1212.
39. Garcia, F. G.; Soares, B. G.; Pita, V. J. R. R.; Sánchez, R.; Rieumont, J. Mechanical Properties of Epoxy Networks Based on DGEBA and Aliphatic Amines. *J. Appl. Polym. Sci.* **2007**, *106*, 2047–2055.
40. Perret, B.; Schartel, B.; Stöß, K.; Ciesielski, M.; Diederichs, J.; Döring, M.; Krämer, J.; Altstädt, V. A New Halogen-Free Flame Retardant Based on 9,10-Dihydro-9-oxa-10-phosphaphenanthrene-10-oxide for Epoxy Resins and their Carbon Fiber Composites for the Automotive and Aviation Industries. *Macromol. Mater. Eng.* **2011**, *296*, 14–30.
41. Gumen, V. R.; Jones, F. R.; Attwood, D. Prediction of the Glass Transition Temperatures for Epoxy Resins and Blends Using Group Interaction Modelling. *Polymer.* **2001**, *42*, 5717–5725.
42. Zhou, J.; Lucas, J. P. Hygrothermal Effects of Epoxy Resin. Part II: Variations of Glass Transition Temperature. *Polymer.* **1999**, *40*, 5513–5522.
43. Pineda, Á. F. E.; Garcia, F. G.; Soares, B. G.; Simões, A. Z.; Silva, E. L. Comparative Study of Glycerol Diglycidyl Ether/Aliphatic Amines Networks. *EIJST*. **2017**, *6* (7), 48–65.
44. Available from: http://infochim.u-strasbg.fr/
45. Chang, C.; Lin, C.; Tieleman, T. LIBSVM : A Library for Support Vector Machines. *ACM Trans. Intell. Syst. Technol.* **2008**, *307*, 1–39.
46. Horvath, D.; Brown, J.; Marcou, G.; Varnek, A. An Evolutionary Optimizer of Libsvm Models. *Challenges* **2014**, *5* (2), 450–472.

TOC graphics



1. Glavatskikh, M.; Madzhidov, T.; Baskin, I. I.; Horvath, D.; Nugmanov, R.; Gimadiev, T.; Marcou, G.; Varnek, A., Visualization and Analysis of Complex Reaction Data: The Case of Tautomeric Equilibria. *Molecular Informatics* **2018,** *37* (9-10).

2. Glavatskikh, M.; Madzhidov, T.; Horvath, D.; Nugmanov, R.; Gimadiev, T.; Malakhova, D.; Marcou, G.; Varnek, A., Predictive Models for Kinetic Parameters of Cycloaddition Reactions. *Molecular informatics* **2018**.

3. Kohonen, T., *Self-Organizing Maps*. Springer: Heidelberg, Belin, Germany, 2001.