Decision Trees as Applied to the Robust Estimation of Diffusion Coefficients in Polyolefins

Olivier Vitrac, Jérôme Lézervant, Alexandre Feigenbaum

AQ: 2 *INRA, UMR FARE 614, Moulin de la Housse, BP 1039, F-51687 Reims Cedex 2, France*

Received 3 November 2004; accepted 1 June 2005
DOI 10.1002/app.23112
Published online in Wiley InterScience (www.interscience.wiley.com).

ABSTRACT: This study dealt with decision trees as used to predict diffusion coefficients (*D*'s) in polyolefins of molecules with molecular weights ranging between 50 and 1200 g/mol at 23 and 40°C. The approach was tested on 657 *D*'s (267 molecules) mainly collected by the European working group SMT-CT98-7513. According to a reptation-like mechanism of transport, three topological molecular descriptors, the Van-der-Waals volume, the gyration radius, and a dimensionless shape parameter, were proposed as both classifiers and regressors. They were calculated from the minimized and oriented structure in the absence of interaction with the polymer matrix. The foreseen ability of regression trees was tested by both cross-validation and bootstrap sampling for a wide number of classes. Optimally pruned trees provided correlation coefficients ranging between 0.74 and 0.96 for each tested polymer. The effects of the volume of diffusing molecules predominated in polyethylene, whereas a combination of the three parameters was required in polypropylene. *D* overestimates, which are particularly useful for checking the compliance of food contact materials, were derived and validated from the upper percentiles of the *D* values observed in each terminal class. The use of decision trees as a tool for data assimilation is discussed. We concluded that the proposed descriptors were *a priori* (without a preliminary fitting) able to gather molecules with similar *D* values without introducing any significant bias.
 J Appl Polym Sci 100: 000–000, 2006

Key words: diffusion; polyolefins; structure-property relations

INTRODUCTION

The identification of relationships between transport properties and chemical structure has major importance in the design of polymer-based products such as packaging materials, membranes, and substrates for biosensors or biotechnological applications. Quantitative structure relationships can be used to either extrapolate results to new substances or materials or infer homology rules for molecules with similar properties. From the conceptual point of view, a multiscale-driven knowledge approach starting from atomistic interactions at the level of picometers and nanometers up to macroscopic scales would be more satisfactory than any other approach because it would explicitly assume a transport mechanism. This strategy is, however, tractable via molecular dynamics simulation or mesoscopic modeling only for small molecules.[1,2] One of the last recorded runs for the simulation of a 100-ns trajectory of limonene in low-density polyethylene (LDPE) was between 77 and 227°C.[3] Because of inherent calculation limitations, more pragmatic approaches may be preferred when predictions are addressed to a large a set of molecules, to medium or large-size molecules, or to relatively low temperatures. Identified correlations between measured properties and chemical structures are a possible alternative. It requires, however, a significant amount of reference values obtained in similar conditions for a representative set of molecules. Recently, a European working group, with whom two of us participated, collected the diffusion coefficients (*D*'s) available in the literature for polyolefins.[4] This study was aimed at the assessment of the feasibility of the use of mathematical modeling to check the compliance of food contact materials against specific migration limits. This possibility was subsequently introduced in an EU regulation (EU-DG SANCO, 2002)[5] as an alternative to costly and time-consuming migration experiments with food simulants. With the logic of the development of robust and simple relationships for *D* estimation in the purpose of compliance checking, an arbitrary Arrhenius-like relationship was proposed. Three predictors were used: a polymer-dependent constant, the molecular weight (*M*) of the considered diffusant, and the temperature. It was demonstrated that this model was almost robust for overestimating the *D*'s of molecules with *M*'s ranging between 100 and 800 g/mol. However, the variable positive bias between the predicted and measured properties up to 3 decades did not make it a possible reliable extension for complex situations (e.g., multilayered materials, recy-

AQ: 18 AQ: 6

Correspondence to: O. Vitrac (olivier.vitrac@reims.inra.fr).

cled materials, reactive systems) or for quantitative risk assessment. Indeed, the combination of bias in the mathematical modeling of the diffusion of substances is responsible for unrealistic migration quantities and uncontrolled uncertainties. A more trustworthy approach is based on the combination of a probabilistic approach and likely D estimates, as described by Vitrac and Hayert[6] (see also Vitrac and coworkers[7,8]).

The purpose of this study was to investigate the use of decision trees [classification and regression tree (CART) algorithms] for the reliable estimation of D's in polyolefins at room temperature or at 40°C. In addition, because CART algorithms generate a hierarchical classification of data, they were used to define comprehensive homologies between the diffusants with regard to their diffusion ability.

CART decision trees are a nonparametric technique for adaptive data-driven modeling; they are quite distinct in form but similar in aim to various forms of nonlinear regression. As an advantage over established parametric techniques, CART algorithm results are nondeterministic (no analytical expression is required), and they can handle both continuous and categorical variables. Because they are clearly constructed, they provide a good compromise between comprehensibility and accuracy and a high computational efficiency and modularity.[9] Also, the recursive partitioning algorithm mimes a very common cognitive approach whereby information is acquired sequentially through a series of questions, with each question depending on the answer to the previous one and each question locally maximizing the expected information about the goal (here the classification of molecules or the prediction of D's).

Because molecular diffusion in amorphous and semicrystalline polymers is related to the size and fractal dimension of the diffusant,[10] predictors related to the three-dimensional (3D) molecular structure of diffusants have been preferred. According to a likely transport mechanism by reptation[11] or in a Rouse regime,[12] topological quantities, including Van der Waals volume (V_{WdV}), the gyration radius (ρ), and a shape parameter ($I_{z/x}$), have been more particularly investigated and tested as classification criteria and D predictors. Similar descriptors (e.g., volume, degree of asymmetry, spherical envelope) were reviewed by Kovarski[13] to explain the different dependences between rotational and translational frequencies observed during the diffusion of low-molecular molecules. The predictive approach was tested on 345 molecules and 628 D data collected by the European project SMT-CT98-7513 (Hinrichs and Piringer)[4] from the literature and industries and on 29 data experimentally obtained by Reynier and coworkers in controlled conditions.[14,15] Because few data were available on the physical properties, such as density and crystallinity, of the tested polymer matrices, different CART models were devised for each polymer type, but they were based on 3D predictors, which were calculated from molecular structures minimized *in vacuo*. This strategy seemed reliable for the prediction of the reptation-like transport of medium-size molecules (i.e., with M's typically ranging between 50 and 1000 g/mol) in noninteracting matrices such as polyolefins as it maximized the steric volume and the unfolding of the diffusant. As a result, the intrinsic topological properties of the diffusants could be stored in the same database and used to predict D's in different polymers.

In this article, we propose a nonparametric and nonlinear 3D quantitative structure relationship that improved the correlations proposed by the SMT-CT98-7513 program,[4] which were fitted on data highly disparate in quality, including both well-documented values and data roughly extrapolated from simple migration tests. Indeed, our growing tree procedure used probabilistic splits that aggregated data (or molecules) that did not present statistically different properties (i.e., that led to a significant overfit bias). For a given set of molecules, the proposed approach could be used to devise either a robust likely D estimate or a D overestimate based on the 50th and 90th percentiles of the available D values.

The article is organized as follows: the Diffusion Coefficients Data section describes the D data available in the EU database[4] and the methodology used to minimize the 3D structures of all of the molecules. General and in-depth analyses of the models are grouped in the following sections. The first subsection presents for each set of molecules a descriptive analysis of calculated topological quantities and discusses how they were related to M, which is the main parameter used as predictor in the technical report.[4] In the second subsection, we analyze the relevance of the chosen predictors *a priori* by checking whether the inferred classification trees relying only on topological properties were able to aggregate molecules with similar D's. The validation of regression trees was studied with computer-intensive approaches, including both cross-validation and bootstrap sampling. This statistical evaluation is presented in a third subsection. The article ends with conclusions and a discussion on the practical use of the proposed models either (1) to perform realistic risk assessment or (2) to check the compliance of plastic materials intended to be in contact with food.

EXPERIMENTAL

D data

The underlying data set consisted of four subsets, as detailed in Table I. Each subset corresponded to D data obtained for a claimed equivalent polymer type and obtained or normalized at the same temperature.

TABLE I
Sets of Data (Polymers, Molecules, Temperature, and *D* Values) Used to Train and Validate the Proposed Decision Trees

Polymer	Temperature (K)	Number of *D* data	Number of molecules	M_{min} (g/mol)	M_{max} (g/mol)	Source
LLDPE–LPDE	296	345	214	16	1177	Ref. 4
MPDE–HDPE	296	142	69	16	1177	Ref. 4
PP	296	141	62	28	1777	Ref. 4
PP	313	29	26	156	807	Refs. 14 and 15

The three first subsets were the 628 data collected by the EU SMT-CT98-7513 program from literature (most data) and from laboratories. Only 302 data (48%) were really determined at the indicated temperature. The other data were either (1) extrapolated from *D* values obtained at a different temperature or (2) derived from migration data (obtained mainly at 40°C). The procedure to standardize all data was documented in the technical report EUR 20604EN.[4] These data were obtained on variable matrices, the density and the crystallinity of which were usually not documented. It was, however, thought that the large number of data including repetitions (up to 10) coupled with such a robust identification technique as CART algorithms made it possible to compensate the inherent discrepancy in each data subset. In the same manner, the large number of different molecular structures (up to 214 for LDPE) might be used to test a very significant number of combinations of molecular parameters to predict *D* values.

By contrast, the last subset comprised data that we already published[14,15] and obtained at 40°C for the same polypropylene (PP) matrix in carefully designed test conditions. The smaller number of data (29 obtained for 26 different molecules) did not, however, allow a general classification of diffusants but was used to test the relevance of the selected predictors on an independent sample and at different temperatures.

3D structure of diffusants

All of the molecular structures were designed within the Materials Studio 3.0 platform (Accelrys). The 3D structure of each molecule was calculated with unconstrained energy minimization within the commercial module of molecular mechanics and dynamics called Discover (Accelrys). The potential energy of the molecules, related to bond and bond angle distortion terms, torsional potentials, intramolecular repulsive and attractive interactions (i.e., Van der Waals), and electrostatic (i.e., Coulombic) interactions, was calculated from the Compass force field (Accelrys). This robust force field, derived from *ab initio* calculations and including both diagonal and off-diagonal cross-coupling terms, is particularly suitable for organic covalently bonded molecules, such as small gas molecules, aromatics, and plastic additives. For the few chemical groups that were not covered by Compass (e.g., phosphoric groups as encountered in secondary antioxidants), a generic covalent force field was used. Minimizations were performed with a conjugate gradient method starting from a random set of initial configurations.

Further calculations and visualizations were performed with a proprietary toolbox called QSPR-MS developed for both the commercial suites Matlab (Mathworks) and the freely available Scilab (INRIA, France). The toolbox was designed to manage native Materials Studio files coded in XML format and to interact with the main modules of molecular mechanics and dynamics.

3D molecular descriptors

Each minimized molecule was oriented along its principal axes. Examples of minimized and subsequently oriented structures are presented in Figure 1 for six typical molecules. The projections along the principal axes (*x*, *y*, and *z*) are also depicted along with the value of the calculated descriptors.

V_{WdV} was determined from the numerical tessellation of the inner volume within its Van der Waals envelope. Our calculations were in good agreement with the simplified summation procedure defined by Zhao et al.[16]

ρ of a molecule including n atoms was calculated from eq. (1):

$$\rho^2 = \langle \|\vec{\mathbf{x}}_i - \vec{\mathbf{x}}_0\|^2 \rangle_{i=1\ldots n} \quad (1)$$

where $\langle\rangle_{i=1\ldots n}$ is the average operator over all atoms of the considered molecule, $\vec{\mathbf{x}}_i$ is the position vector of the atom i and $\vec{\mathbf{x}}_0$ is the position vector of the center of mass. ρ of the diffusant with respect to its centroid is related to the size of an equivalent spherical molecule with a similar moment of inertia.

$I_{z/x}$ is defined as the ratio of the moments of inertia along the axes of minimal and maximal inertia with respect to the center of mass, z and x, respectively. For almost spherical molecules, this ratio was close to 1,

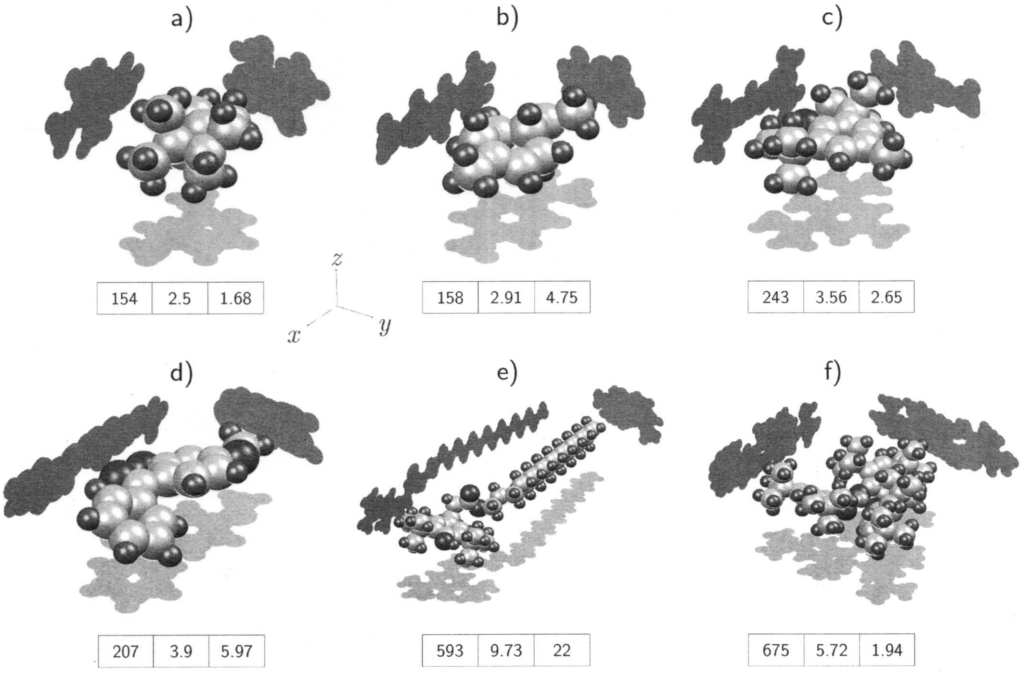


Figure 1 Values of topological descriptors for six typical molecules: (a) β-pinene, (b) limonene, (c) BHT, (d) Chimassorb 90, (e) Irganox 1076, and (f) Irgafos 168. The values are ordered as V_{WdV}, ρ, and $I_{z/x}$. All molecules were oriented along their main axes (x, y, and z); the three main projected surfaces are also depicted.

whereas it was much greater than 1 for linear molecules (Fig. 1). In contrast with ρ, which compares the molecule shape with a sphere, $I_{z/x}$ assesses the similarity of the shape of the molecule with a rod geometry. Because in classical mechanics the moment of inertia with respect to an axis is related to the rotational kinetic energy around the considered axis, this ratio also compares the ability (probability in statistical mechanics) of a molecule to rotate at its centroid around its weakest and strongest axes.

CART procedures

The CART procedure we used was similar to the one described by Breiman et al.[17] it is designed to divide recursively a population of molecules (classification tree) or D's (regression tree) into subpopulations defined by their 3D molecular descriptors in terms of their risk-factor categories (due to either misclassification or error in prediction). The characteristic tree structure was reached by a stepwise division of the population at a node into dichotomous branches in such a way that subpopulations were internally as homogeneous and externally as heterogeneous as possible with respect to some specified criteria. For classification trees, the criterion was based on the Gini index and was calculated as the impurity function $[i_{(t)}]$. For any parent node t, which contained data belonging to $J_t = 2$ number of classes (where J_t is the number of child nodes at node t), $i_{(t)}$ was defined as

$$i_{(t)} = 1 - \sum_{j=1}^{J_t} p^2(j|t) \qquad (2)$$

where $p(j|t)$ is the proportion of each class $j_{j=1\ldots J_t}$ in the node so that $\Sigma_{j=1}^{J} p(j|t)=1$. At each node, the CART procedures used an exhaustive search over all possible variables to identify the split that maximized the decrease in impurity. A branch stopped growing when the impurity could not be further decreased.

For regression purposes, log D was chosen as the dependent variable as a result of the large dispersion of D values. The two-way splitting process was guided by a least-squares error criterion $[\Delta e_{(t,s)}]$. At each parent node t, the best split was the split that maximized a function similar to eq. (2):

$$\Delta e_{(t,s)} = e_{(t)} - \sum_{j=1}^{J_t} e_{(j)} p^2(j|t) \qquad (3)$$

where s is the tested split value. The mean squared error at node j $[e_{(j)}]$, including the N_j data, was calculated as follows: AQ: 8 AQ: 9

$$e_{(j)} = \frac{1}{N_j} \sum_{i=1}^{N_j} [\log_{10}(D_i) - \langle \log_{10}(D_i) \rangle_{i=1\ldots N_j}]^2 \qquad (4)$$

It was verified that the low dimension of the predictor space (3) prevents the unrestrained search pro-

cedure from preferably selecting the variable that generates more splits.[18] The tree construction was stopped when it was determined that there were not enough data to make reliable choices. The stop criterion was based on a variance homogeneity test. This strategy, known as *prepruning,* prevented too much overfitting/overlearning in the initial full tree, as explained later. For a given training sample of molecules, the number of terminal nodes (or classes) in the full tree was, therefore, lower than the number of log D values (or molecules).

Any path from the top node to a leaf (subpartition) could be seen as a conjunction of logical tests on the predictor variables (i.e., decision path). For each terminal leaf of the regression tree, a constant value of the target variable was predicted. Because, for the same training sample of molecules, the achieved full tree structure usually suffered from overfitting (i.e., it was explaining random discrepancies between molecules that were not likely to be features of the larger population of molecules/data), the final structure was inferred from simplifications (i.e., size reduction) of the considered decision tree. As a result of a clear trade-off between the number of partitions (leaves) and the predictive accuracy of the tree, the number of nodes was reduced by the postpruning of the full tree on the basis of cross-validation results. Cross-validations relied on successive validations (10 iterations) of the fitted tree on independent subsamples (blind samples). The initial learning sample was split into 10 subsamples, chosen randomly but with almost equal size. For each subsample, we assessed the prediction error starting from a tree fitted from the 90% remaining data. The optimal tree was a pruned tree whose resusbtitution (simplification) error was of the same magnitude as the cross-validation error. For a given regression tree, a more reliable estimation of the error prediction rate with uncertainty bound was obtained with a bootstrap procedure. The bootstrap technique involved choosing random samples with replacement (i.e., given data can appear multiple times in the same bootstrap sample) and fitting them against available data. The number of observations in each bootstrap sample equaled the number of data in the learning sample. The range of prediction errors between 1000 samples was used to establish the uncertainty on the determination coefficient (r^2).

RESULTS AND DISCUSSION

For a similar set of molecules, the D values are more spread out in solids than in liquids or gases. They may differ by more than a factor of 10^{10} and are difficult to estimate via theoretical models.[19] Diffusion in polymers involves several different mechanisms, depending on the relative size of the diffusant and the mobility of the entangled polymer chains. As a result, diffusion rates should lie between those of liquids and solids. The molecular weight (M) or the spherical representative volume of the diffusant[20] has been proposed in many models to predict or to overestimate D values in polymers on the basis of either semiempirical or mechanistic considerations. The first part of our results reviews possible scaled relationships between collected D values with M. Possible correlations between the proposed diffusant descriptors (V_{WdV}, ρ, and $I_{z/x}$) with M and $\log_{10} D$ are also discussed.

Descriptive statistics of the 3D molecular descriptors

Because, at the scale of the diffusant, diffusion is related to the fluctuation of the position of center of mass, D values are expected to vary with the number of atoms or groups in the molecule. If the movements of individual atoms or groups within the same molecule remains mainly uncorrelated and with nondisparate variances, as a result of constraints or friction with the polymer, the central limit theorem imposes that D values decrease as the reciprocal of M (or equivalently the number of atoms) with a phenomenological exponent connecting D and M (α) close to 2. This transport mechanism is known as *reptation.* The molecules diffuse like snakes because of contour length fluctuations.[21] If a combination of contour length fluctuations and constraint release occurs, Lodge proposed an α value of 2.4.[22] If some self avoidance in the movements of atoms occurs, an α value significantly lower than 2 is expected instead. If the diffusant dynamics is mainly dominated by interactions along its main axes, a Rouse regime may occur; this is identified by an α value of 1.[23] If the transport of the diffusant obeys a hopping mechanism that leads to highly correlated displacements of atoms, the effect of the free volume is predominant and an α value lower than 1 (ca. 0.6) is then expected. On the opposite end, if significant entanglements limit the displacement of the diffusant, α values greater than 3 may occur.[24] Different scaling coefficients in reptation-like models were reviewed by Masaro and Zhu[25] The values, which were reported for the self-diffusion of polymer chains in bulk and in solution, ranging between 0.56 and 3.3. Additional discussions on the connection between the reorientational dynamics of diffusants and segmental dynamics were discussed by Kovarski[26] and Manabe.[27]

To discuss the possible mechanism of the transport of medium-weight molecules in polyolefins, Figure 2 shows the collected D's against M's on a log–log scale. F2
The plotted surface of each marker is proportional to $I_{z/x}$. Straight lines with slopes of $\alpha = 0.6, 1, 2$, and 3 are also plotted. The line $\alpha = 2$ crosses the center of the scatter. The other lines intercept the latter at the lowest M value.

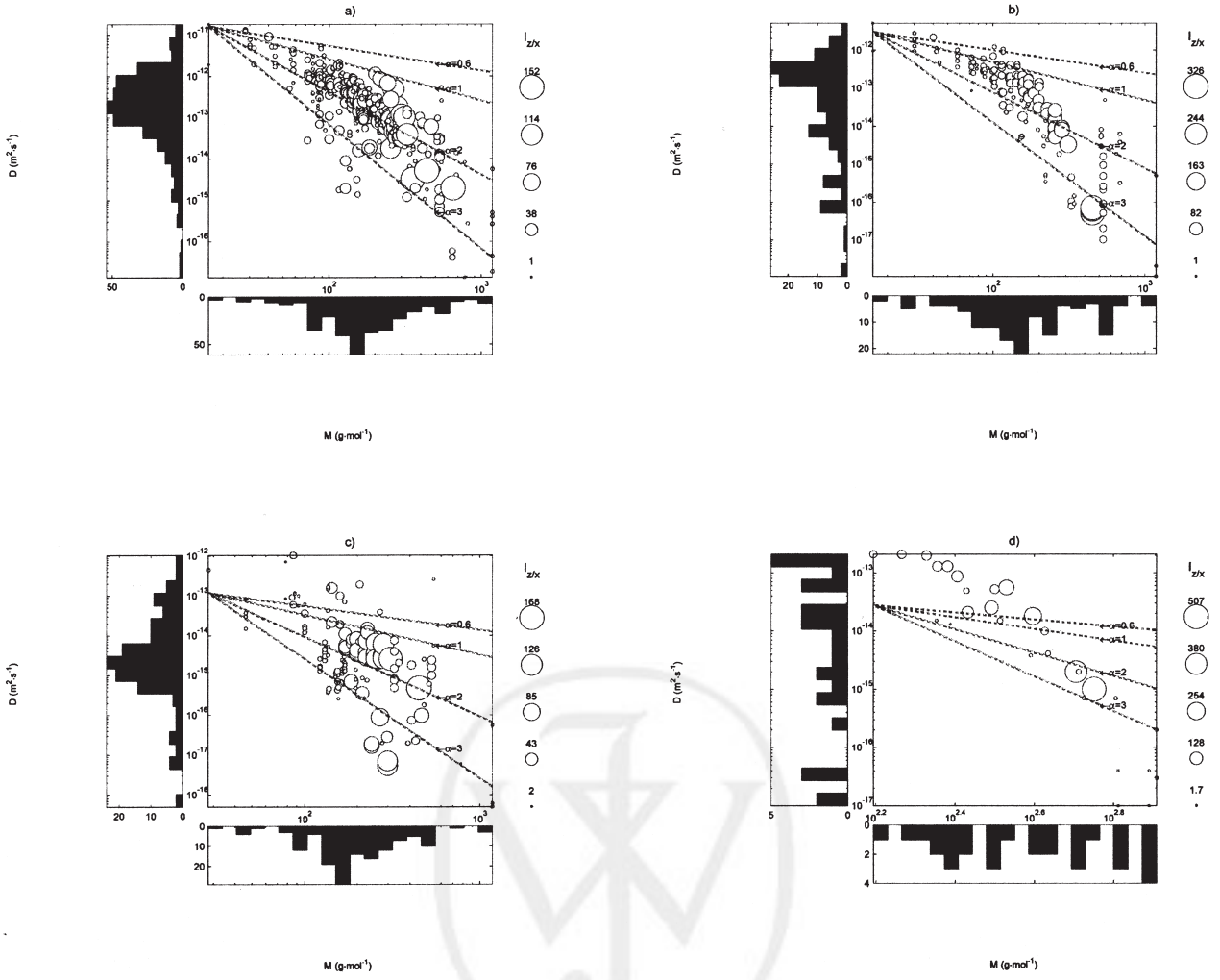


Figure 2 Log–log plot of D versus M in (a) LLDPE–LDPE at 23°C, (b) MDPE°HDPE at 23°C, (c) PP at 23°C, and (d) PP at 40°C. The surface area of each symbol is proportional to $I_{z/x}$. Theoretical variations of D with M are also depicted with the assumption that the transport of the average molecule obeyed a reptation model ($\alpha = 2$).

For the tested polymers, the distribution of D values was not directly correlated to the distribution of M values. For linear low-density polyethylene (LLDPE)–LDPE, the mechanism of reptation ($\alpha = 2$) was very likely for a wide range of molecules. An upper bound of D's was established with the assumption of a Rouse regime and $\alpha = 1$. Also, a hopping mechanism based only on free volume considerations was very unlikely. A lower bound of D's would exist in a restrained mechanism such that $\alpha > 3$. As the D values varied up to 3 decades for the same M, we inferred that complementary parameters were required to predict D. We calculated that up to 40% of the observed variance within the available data could not be explained by M. Different mechanisms of transport could occur for diffusants with similar M values. Figure 2(a) shows in particular that the deviation from a reptation mechanism ($\alpha = 2$) toward an entangled mechanism ($\alpha > 2$) was more probable for asymmetric molecules with a $I_{z/x}$ value greater than 30. However, we emphasize that the observed discrepancy in D values should have been partially related to the experimental error or to the procedure to extrapolate data from different temperatures and/or from migration data. So to prevent misleading conclusions, suitable models should not generate significant bias between values available within this database and the predicted ones. In addition, an optimal predictive model should lead to an error similar in magnitude to the error inherent to the experimental and normalization methodologies. In our approach, a quantification of the uncontrolled error, which was assumed to be random, was provided either by the repetition of D measurements with different methodologies (when available) or by the comparison of D values obtained with homologous molecules (almost always achievable).

Very similar conclusions were drawn from the D values obtained with medium-density polyethylene (MDPE)–high-density polyethylene (HDPE) normalized at 23°C. The set of studied molecules was differ-

AQ: 3-4
AQ: 18
AQ: 5

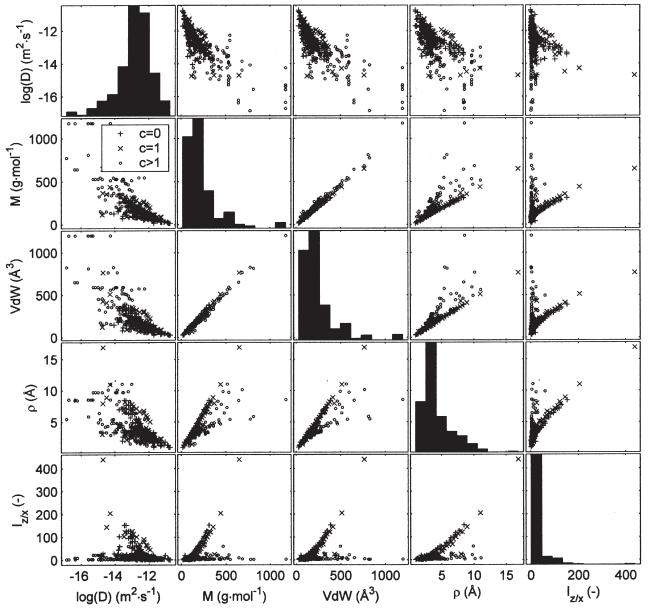


Figure 3 Scatter plots of the predictors and the dependent variable grouped according to *c*. Only the results corresponding to LLDPE–LDPE at 23°C are depicted.

ent in number and type among the four studied polymers. The heterogeneity of α values was particularly high between the 62 molecules tested in PP at 40°C, with a likely mechanism such that α was greater than or equal to 2. In the set of molecules tested at 40°C in PP, all 29*D* values were derived from the same plastic material without any extrapolation. This shows that the α values were lower than 1 for molecules with *M* values below 200 g/mol, whereas α values greater than 3 were more probable for diffusants with higher *M* values. The low number of molecules and the absence of homologous series made impossible to accurately locate the transition between a transport mechanism controlled by the free volume of the diffusant and a mechanism controlled by constraints between the diffusant and the polymer.

In this study, we examined three properties (V_{WdV}, ρ, and $I_{z/x}$) that could be related to the previous wide range of identified mechanisms of transport at the molecular scale. Because these quantities were extensive, they were expected to be partially correlated with the number of atoms and, hence, with *M*. The correlation between all of these quantities and $\log_{10} D$ were analyzed and are shown as scatter plots in Figure 3 for the LDPE data at 23°C. A topological index, the third-order Kier and Hall cluster count index[28] (*c*), was used to group data according to the size and the degree of the branching of the molecules. A *cluster* was defined by a pattern with a connectivity similar to a isobutane. As a result, the parameter *c* measured the number of isobutane patterns in a molecule. The distributions of *M*, V_{WdV}, ρ, and $I_{z/x}$ were asymmetric to the right, whereas the $\log_{10} D$ distribution was asymmetric to the left. The high linear correlation between the *M* and V_{WdV} quantities was responsible for the large similarity between their distributions, which presented both two modes and similar ranges when expressed in grams per mole and in Å^3. ρ and $I_{z/x}$ were only partially correlated with *M*. The relationship between ρ and *M* was almost linear for $c = 0$ and $c = 1$ but with slightly different slopes. When *c* was greater than 1, ρ and *M* were poorly correlated. Similarly, $I_{z/x}$ was nonlinearly correlated with *M* when *c* was less than or equal to 1 and was independent of *M* when *c* was greater than 1. ρ and $I_{z/x}$ were both correlated with $\log_{10} D$ except for low $I_{z/x}$ values (i.e., for almost spherical molecules). Compared to *M*, these predictors generated similar or even lower scatter with collected $\log_{10} D$ values when *c* was less than or equal to 1. $I_{z/x}$ seemed, in particular, a pertinent predictor of $\log_{10} D$ values for molecules characterized by $c = 0$ (i.e., without any ring and poorly ramified).

Similar observations were made with the three other sets of molecules. The main results are summarized in Figure 4. $\log_{10} D$ was significantly correlated with either *M* or V_{WdV}. A correlation existed with ρ, which varied significantly according to the value of *c*. This effect was clearly observable for the set of molecules tested in MDPE–HDPE and was more tenuous for both sets of molecules tested in PP. Finally, only molecules with *c* values lower or equal than 1 exhibited a decreasing correlation of $\log_{10} D - I_{z/x}$. This last result confirmed that $I_{z/x}$ was a pertinent parameter mainly for molecules that did not present a globular or highly clustered structure. *D* values spreading over more than 2 decades independent of the used descriptors was observed for molecules with *M* above 550 g/mol. This discrepancy was related to a higher experimental error because the corresponding very low *D* values were the more difficult to assess. Besides, the diffusion behavior of such molecules may have been more sensitive to the detailed microstructure of the polymer.

Prediction of $\log_{10} D$ from an *a priori* tree classification of molecules

Intuitively, a pertinent predictor based on the molecular structure of diffusants should be also a good classifier because it is expected to regroup molecules with similar $\log_{10} D$ values whatever the tested polymer. In other terms, a decision tree, based on the proposed properties (V_{WdV}, ρ, and $I_{z/x}$), which is not trained to predict *D*'s, should have been able to classify *a priori* a set of molecules according to their $\log_{10} D$ values. The ability of the classification tree to gather *a priori* molecules with homogeneous $\log_{10} D$ values was analyzed and is shown in Figure 5 for the four sets of molecules. For each set, a maximal classification

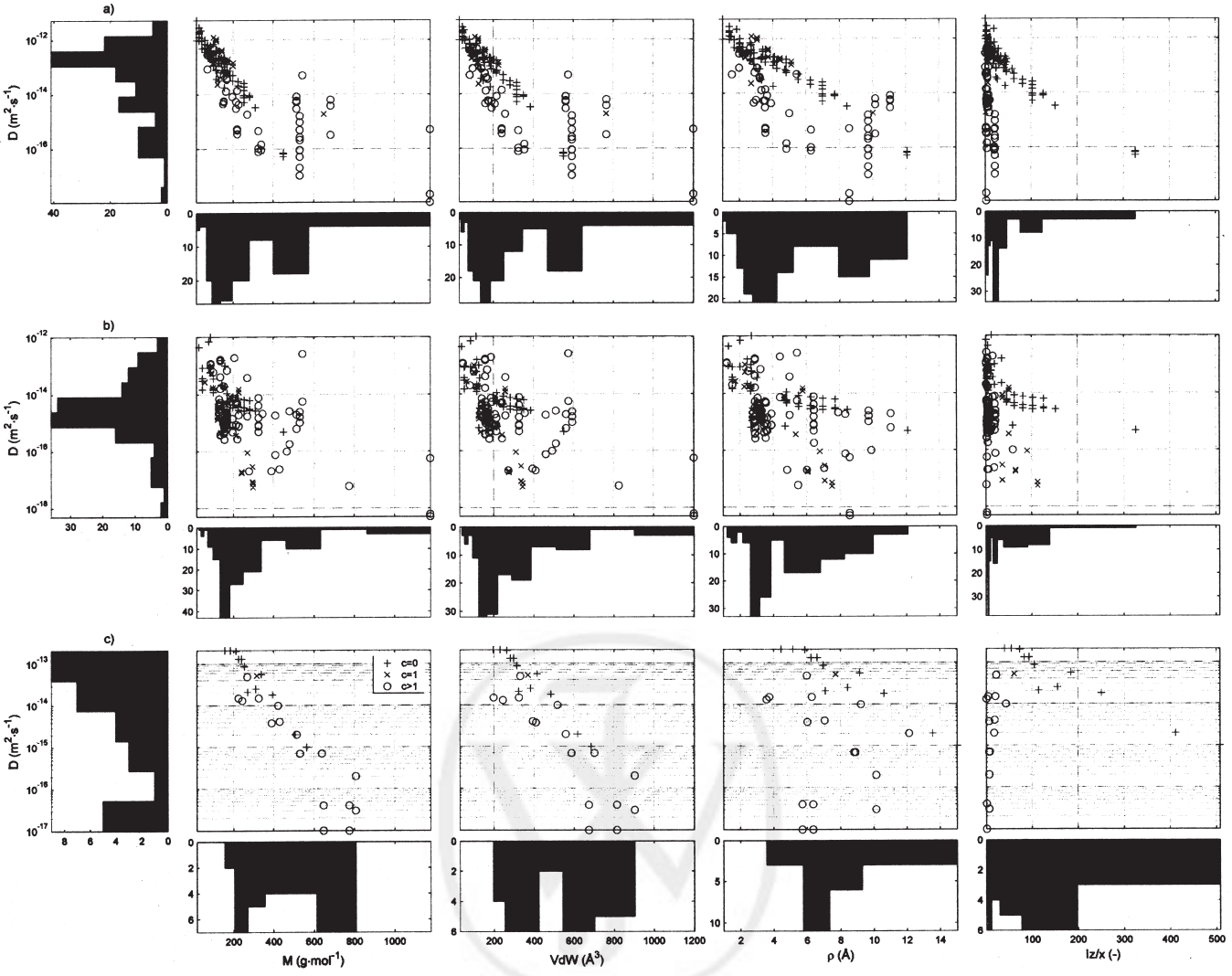


Figure 4 Scatter plots of the dependent variable, $\log_{10} D$, versus each tested predictor in (a) MDPE–HDPE at 23°C, (b) PP at 23°C, and (c) PP at 40°C. The data are grouped according to *c*.

tree was first optimized over the whole sample of molecules to their V_{WdV}, ρ, and $I_{z/x}$ values. The tree was afterwards successively pruned until the number of terminal nodes (i.e., classes) was reduced to a singleton. For each set of molecules, Figure 5 plots the value of the average of $\log_{10} D$ in each class according to the number of classes (here, the square root of number of classes) and the distribution of $\log_{10} D$ values. The obtained figure is a two-dimensional topological representation of the initial full tree that describes how $\log_{10} D$ values were shifted toward the average of the whole sample when the information on molecular structure of diffusants was reduced. To illustrate the drastic effect of the data reduction achieved with the iterative classification process, Figure 5 depicts also the links (in gray lines) between the raw D values (i.e., as collected) and typical D values obtained with the full classification tree. We emphasize that the full classification tree minimized the number of molecules per class and gathered molecules whose topological parameters were the most homogeneous. Along the pruning level or, equivalently, along the number of terminal nodes, the iterative classification was interpreted differently. Starting from the full classification, disparate D values for similar molecular structures appeared as fingered leaves when the complexity was decreased. The amplitude of digitations provided a possible measure of the inhomogeneity in the initial D values mainly due to the uncertainty in the initial data set. Also, the shape of the so-plotted tree revealed whether $\log_{10} D$ data were close after an *a priori* classification. A bush shape was generated by a significant number of crossover branches due to significant initial misclassifications of data. By contrast, a regular shape/homothetic shape was interpreted as a convenient classification without, however, providing any information on its optimality to predict D values.

AQ: 18

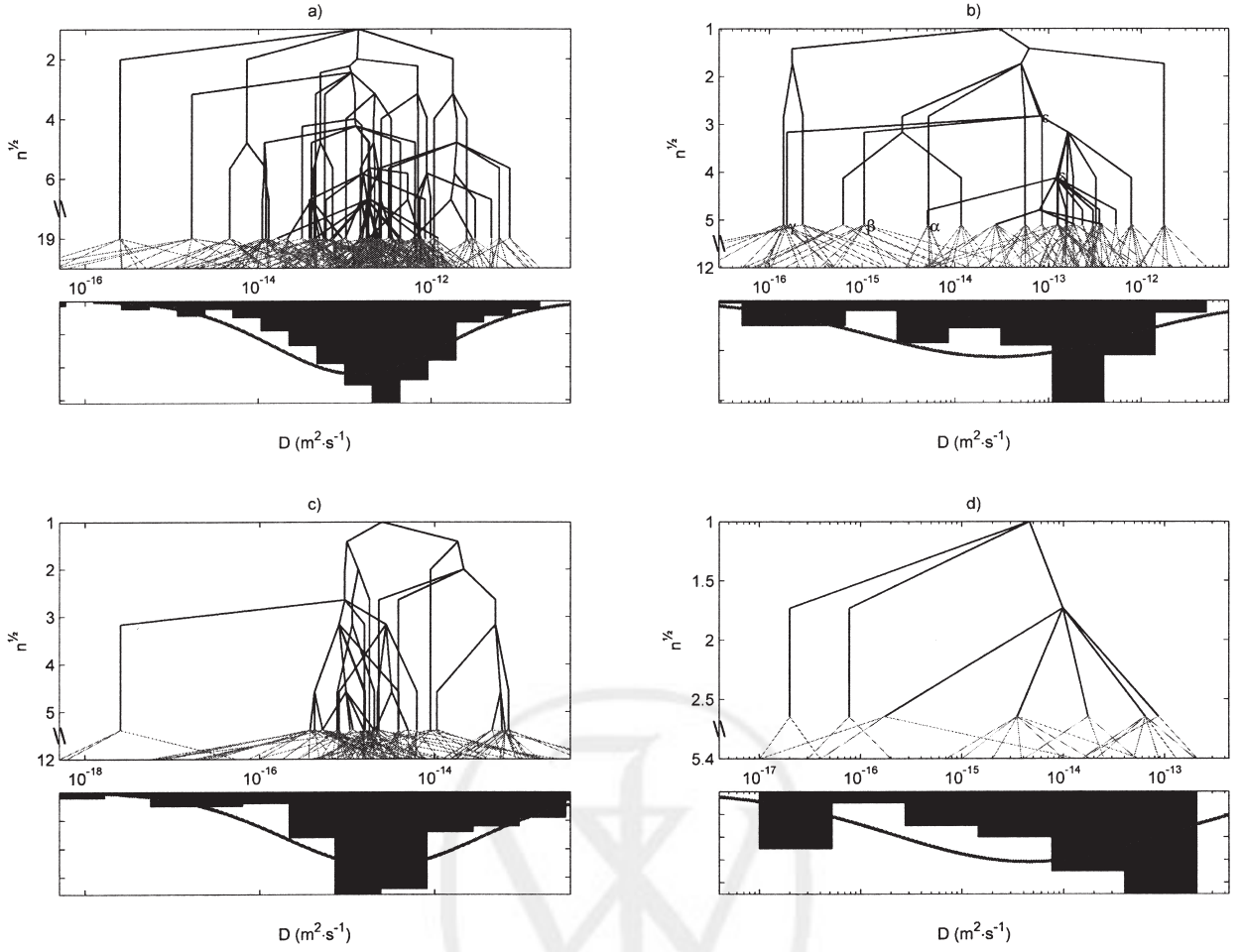


Figure 5 *A priori* clustering of *D* values according to the pruning level of the full classification tree for (a) LLDPE–LDPE at 23°C, (b) MDPE°HDPE at 23°C, (c) PP at 23°C, and (d) PP at 40°C. *n* is the remaining number of terminal nodes after tree pruning. For each class of molecules, the typical *D* value was assumed to be equal to the average *D* value. The distribution of *D* values (bins) and the corresponding fitted log-normal distribution (continuous line) are also depicted. The gray lines depict how individual *D* data were grouped into classes. α, β, γ, δ, and ε are the typical classes that are discussed in the section Prediction of $\log_{10} D$ from Regression Trees.

A variance analysis on the $\log_{10} D$ values demonstrated that an *a priori* classification of molecules according to the full classification tree resulted in an intraclass variability of a similar magnitude as the estimated uncertainty in the collected *D* values (Table II). Both variances had statistically similar magnitudes in samples tested in LLDPE–LDPE at 23°C and in PP at 40°C. After classification, the intraclass variance was only about 72 and 140% higher after the proposed *a priori* classification of samples tested in MDPE–

TABLE II
Comparison of the Intraclass Variance of $\log_{10} D$ Estimates with the Uncertainty in Collected Data

Polymer	Temperature (K)	*r* (%)	*F*	$p > F$	Significance
LLDPE–LPDE	296	26	0.91	0.76	Not significant
MPDE–HDPE	296	54	1.7	0.006	Very significant
PP	296	66	2.4	<0.001	Extremely significant
PP	313	12	2.3	0.26	Not significant

F = ratio between the intraclass variance assessed after the classification and the typical variance observed between repetitions (uncertainty in collected data). Because the uncertainty in the collected data was only based on molecules including repeated *D* values, *F* values lower than 1 could occur: *r* = the fraction of molecules including repeated *D* values. The classification was based on the optimal classification tree. The uncertainty was assessed from the variance in *D* values collected for the same molecules.

HDPE at 23°C and in PP at 40°C, respectively. Because the initial uncertainty and the achieved intraclass variability were almost homogeneous, we inferred that the three tested parameters were able to classify *a priori* and in a satisfactory manner the collected $\log_{10} D$ values. This conclusion was strengthened by the low disordered shape of the classification trees when the number of terminal nodes was reduced. In particular, we emphasize that molecules with extreme diffusivities were appropriately identified even with very simplified classification trees. The best results were obtained with the sets of molecules including the largest diversity (LLDPE–LDPE, MDPE–HDPE). For all sets of molecules, the risk of misclassification seemed higher for molecules with $\log_{10} D$ values close to the average of the whole sample. Because this risk was drastically reduced when the complexity of the tree was reduced, this effect could have been related to the high variability of initial data. Indeed, the average diffusivities in the polyolefins (for the three sets at 23°C) were estimated from methods that were highly disparate in quality and that included both interpolated and noninterpolated data. The uncertainty was estimated up to 1.5 decades (ranges between digitations). On the contrary, extreme diffusivities were thought to be more reliable as they were mainly inferred from well-documented migration experiments (case of low diffusivities) or from well- controlled gas permeation experiments (case of high diffusivities).

Prediction of $\log_{10} D$ from regression trees

The regression procedure acted as a learning stage dedicated to minimizing the risk of misclassification (as previously defined) of molecules. A regression tree, therefore, took advantage of intrinsic classification properties of the tested predictors and also of significant improvements due to learning. The results derived from regression trees can finally be interpreted in the same way as classification results depicted in Figure 5.

To finely analyze the effect of learning/fitting, the full regression tree based on the whole sample of data were first plotted (Fig. 6) and compared with the results derived from the corresponding full classification tree (Fig. 5). The average error of misprediction was analyzed for different simplification levels of the full regression tree starting from both cross-validation testing and bootstrap sampling. The so-considered optimal regression tree is plotted in Figures 10 and 11 (shown later) with the conditional tests that made it possible a practical use for D prediction.

Regressions tress based on the whole sample

Figure 6 plots the full regression tree corresponding to Figure 5. All four regression trees exhibited a more regular structure without significant crossovers. They revealed, in particular, how extreme values could be combined with other values along the simplification process, which was aimed at reducing the number of terminal nodes. Starting from the full tree (including 75, 30, 28, and 5 terminal nodes, respectively, for the sets tested in LDPE–MDPE at 23°C, MDPE–HDPE at 23°C, PP at 23°C and PP at 40°C), we first combined the median values before combining the intermediate and extreme values. The trees, whose number of terminal nodes was reduced by a factor up to four, kept, therefore, almost similar ranges of $\log_{10} D$ values. The classification procedure combined with a learning stage acted to preserve the main features of the initial distribution of the $\log_{10} D$ values after the classification. As a result, the number of recombinations was higher where the number of D data was higher, that is, mainly around the mode of the distribution of $\log_{10} D$ values, whereas it was significantly lower for the molecules with extreme D values.

The effect of learning is discussed for three small subsets of molecules tested in MDPE–HDPE at 23°C [classes α, β, and γ as noted in Fig. 5(a)] that were nonoptimally classified for D prediction when the complexity was reduced. The subsets α, β, and γ are depicted in Figure 7; they included, respectively, 2, 4, and 1 molecules with Ms higher than the average. The first subset, as generated by *a priori* classification, was merged into a class of 29 molecules [δ; 42% of tested molecules, depicted in Fig. 8(b)] when the number of terminal nodes was reduced from 26 down to 17. This poorly selective grouping led to an overestimate of the average D value of the class a by a factor up to 75. The mode of grouping of the two last subsets was more dramatic for D prediction, as they were merged into the same huge class of 53 molecules (77% of the total) located on the opposite side of the D scale. After additional grouping, the average D values of subsets β and γ were overestimated by a factor of up to 3 decades when the number of terminal nodes was reduced down to 8.

When a regression tree was used instead of a classification tree, the molecules included in subsets α, β, and γ were distributed into three distinct classes, δ, ε_1, and ε_2, including, respectively, 12, 5, and 5 molecules. The list of properties of the corresponding molecules is summarized in Table III. Staring with a full regression tree consisting of 30 terminal nodes, we overestimated the average D values of the initial classes only by a factor lower than 5 for molecules belonging to subsets α and β, whereas it was underestimated only by a factor of 4 for molecules belonging to subset γ. This significant improvement in the classification due to regression (learning) is illustrated in Figure 8 with a comparison of the molecules belonging to the equivalent class δ when a classification tree was substituted by a regression tree. The new class was smaller and

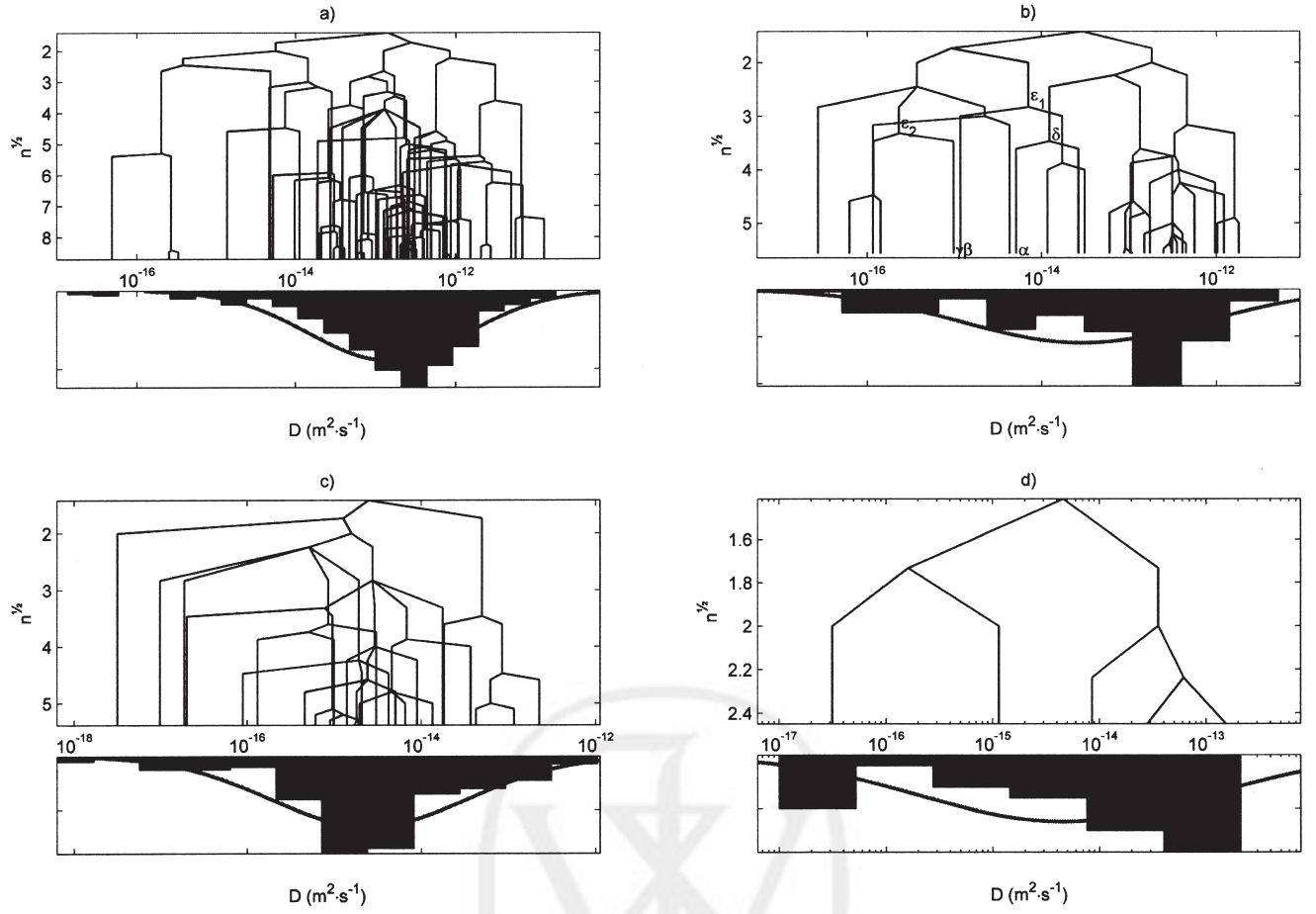
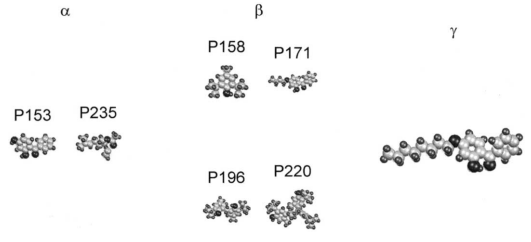


Figure 6 *A posteriori* clustering of *D*'s according to the pruning level of the regression tree for (a) LLDPE–LDPE at 23°C, (b) MDPE–HDPE at 23°C, (c) PP at 23°C, and (d) PP at 40°C. *n* is the remaining number of terminal nodes after tree pruning. For each class of molecules, the typical *D* value was assumed to be equal to the average *D* value. The distribution of *D* values (bins) and the corresponding fitted log-normal distribution (continuous line) are also depicted. α, β, γ, δ, and ε are classes that correspond to the classes noted in Figure 5 (b).

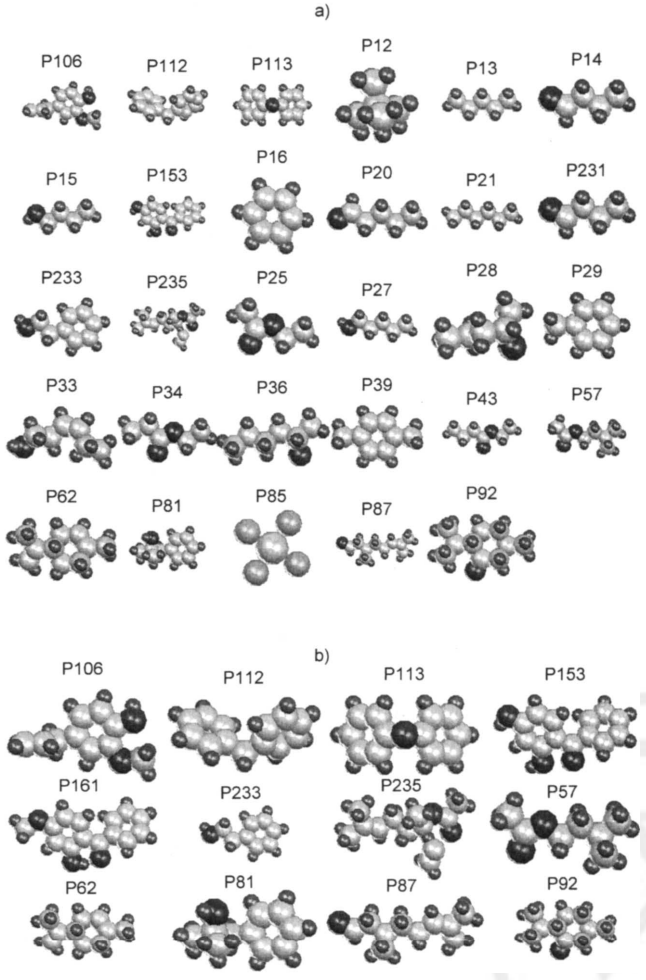
appeared more homogeneous, as it contained only molecules with similar sizes including one or two rings or equivalently ramified shapes so that they had similar inertia. This last class exhibited a 95% confidence range of *D* values lower than 1 decade, which was in very good agreement with our initial intuition: molecules with similar 3D structures had similar *D* values, as they obeyed the same transport mechanisms. However, this analysis does not provide any information on the robustness of the classification. AQ: 18



AQ: 13 **Figure 7** CPK molecular structure of molecules corresponding to subsets α, β, and γ [defined in Fig. 5(b)]. All molecules were minimized *in vacuo* and projected along their two principal axes.

Postpruning optimization

The previous qualitative description of regression trees was not sufficient to determine which pruning level was optimal for each set of molecules. The cross-validation and bootstrap results are summarized in Figure 9(a,b) for the set of molecules tested in LDPE–MDPE at 23°C. Similar results were obtained for the three other sets of molecules. A too-high reduction in the tree complexity was responsible for a dramatic increase of the average quadratic error of prediction [χ^2; Fig. 9(a,b)]. By contrast, increasing the tree complexity beyond a certain threshold led to an increasing prediction error [Fig. 9(a)]. r^2's exhibited similar behavior, as their values were significantly lower in the bootstrap samples than in the whole sample when the F9



AQ: 13 **Figure 8** CPK molecular structures of molecules belonging to the class δ as defined in (a) Figure 5 (b) (classification) and (b) Figure 6 (b) (regression). All molecules were minimized *in vacuo* and projected along their two principal axes.

number of terminal nodes increased [Fig. 9(b)]. These effects are known as *underfitting* and *overfitting*. Underfitting produced excessive bias in the predicted $\log_{10} D$ values, whereas overfitting generated excessive variance between the predictions. Both measures of performance on validation and bootstrap samples provided different optima, ranging between 8 and 22 classes. Cross-validations usually led to a lower number of classes than that estimated by bootstrap sampling and led to values that varied significantly (ca. 30%) between repetitions. Within the expected range of the optimal pruning range, the choice to grow a given branch or to remove nodes was based on our expertise and on the capacity of the generated tree to gather similar chemical structures and to preserve some specific features related to some outstanding chemical structures initially present within the set of molecules. AQ: 18

The risk of overfitting was all the higher when the sample used for fitting/learning was small (e.g., set tested in PP at 40°C). In our approach, the risk of overfitting was counterbalanced by the use of a low number of molecular attributes that seemed both pertinent classifiers and predictors even without prior fitting/learning. For robustness purposes, the number of terminal nodes was chosen to be almost proportional to the square root of the number of different molecules in each tested sample. AQ: 18

Optimized (postpruned) regression trees

The regression with optimum postpruning levels are plotted in Figures 10 and 11 for the four tested samples, and their main characteristics are reported in Table IV. The $\log_{10} D$ value at each node (either terminal or not) corresponded, respectively, to the average and to the 90th percentile (Figs. 10 and 11) of the gathered $\log_{10} D$ values. Thus, Figures 10(a) and 11(a–c) plot realistic $\log_{10} D$ estimates, whereas Figures 10(b) and 11(d–f) plot $\log_{10} D$ overestimates with a risk of 10%. The corresponding splitting conditions appeared at each node so that the left branch had to be chosen when the test value was true and the right branch had to be chosen otherwise. F10/F11/T4

The plotted regression trees included 25, 9, 10, and 4 terminal nodes for the molecules tested in LDPE–MDPE at 23°C, MDPE–HDPE at 23°C, PP at 23°C and PP at 40°C (Table IV) , respectively. The order of appearance of each variable along each regression tree revealed their respective weights for the prediction of $\log_{10} D$. The effects of volume (V_{WdV}) predominated in all of the samples tested in polyethylene [Figs. 10(a) and 11(a)], whereas both ρ, and V_{WdV} controlled the dispersion of the $\log_{10} D$ values of the sample tested in PP at 23°C. In the sample tested in PP at 40°C, a cutting volume of 538 $Å^3$ combined with different $I_{z/x}$ values defined how the values of the D's were spread over the $\log_{10} D$ scale. Because of the risk of overfitting, the combination of the three predictors was mainly used to separate the classes of the most documented molecules (i.e., including a significant number of repetitions). As a result of the overrepresentation of molecules with M's ranging between 120 and 450 g/mol in the database (see Figs. 3 and 4), each pruned tree exhibited a dense structure close to the median of the $\log_{10} D$ values and was sparser outside. The entanglement level was low and was at a maximum where the density was at a maximum. As a result, the regression trees based on the average $\log_{10} D$ values of each class were well-balanced and almost symmetric [Figs. 10(a) and 11(a–c)]. On the contrary, regression trees based on the 90th percentile values deviated toward high $\log_{10} D$ values [Figs. 10(b) and 11(d–f)]. The shift was higher, up to 2 decades, for classes corresponding to the highest $\log_{10} D$ values, and they were on the opposite lower for classes with the lowest $\log_{10} D$ values. For each leaf/terminal class including

TABLE III
Molecular Descriptors and $\log_{10} D$ Data for All Diffusants Presented in Figures 8 and 9

AQ: 12

Code	Chemical name	CAS no.	V_{WdV}	ρ	$I_{z/x}$	$\log_{10} D$	$\log_{10} (D)_{min}$	$\log_{10} (D)_{max}$	n
P106	2-Methoxy-4-(2-propenyl)phenol (eugenol)	97-53-0	162.0	3.32	4.98	−13.9	−13.9	−13.9	1
P112	Diphenylmethane	101-81-5	172.0	3.23	5.41	−13.5	−13.5	−13.5	1
P113	Diphenyl oxide (phenoxy benzene)	101-84-8	164.0	3.21	5.07	−13.4	−13.4	−13.4	1
P12	Neopentane	463-82-1	95.7	1.98	1.00	−12.8	−13.1	−12.6	2
P13	*n*-Pentaue	109-66-0	96.1	2.35	9.15	−12.6	−12.6	−12.6	1
P14	Butanal (*n*-Butylaldehyde)	123-72-8	81.8	2.00	11.39	−12.4	−12.4	−12.4	1
P15	Butyl alcohol	71-36-3	87.4	2.20	9.85	−12.5	−12.6	−12.3	2
P153	2,4-Dihydroxybenzophenone (DHB)	131-56-6	191.0	3.53	4.93	−14.4	−14.5	−14.3	2
P158	2,6-Di-*tert*-butyl-4-methylphenol (ionol or BHT)	128-37-0	243.0	3.56	2.65	−15.0	−15.5	−13.9	4
P16	Benzene	71-43-2	84.1	2.02	2.00	−12.4	−12.5	−12.4	2
P161	2-Hydroxy-4-methoxy benzophenone (Chimassorb 90)	131-57-7	207.0	3.90	5.97	−14.4	−14.4	−14.4	1
P171	2-Hydroxy-4-butoxybenzophenone	15131-43-8	260.0	5.00	9.94	−14.7	−14.7	−14.7	1
P196	2-2 -Methylenebis(6-*tert*-butyl-4-methylphenol (Antioxydant 2246)	119-47-1	356.0	4.82	4.88	−15.9	−16.0	−15.8	2
P20	Pentanal	110-62-3	98.6	2.32	15.50	−12.6	−12.6	−12.6	1
P21	*n*-Hexane	110-54-3	113.0	2.70	13.35	−12.4	−12.5	−12.2	3
P220	1,1,3-Tris(2-methyl-4-hydroxy-5-*tert*-butyl-phenyl)butane (Topanol CA)	1843-03-4	575.0	5.44	2.61	−13.3	−13.3	−13.3	1
P231	*n*-Butylaldehyde	123-72-8	81.8	2.00	11.39	−12.3	−12.4	−12.2	2
P233	2-Phenylethylalcohol	60-12-8	126.0	2.70	6.65	−13.6	−13.6	−13.6	1
P235	3,7-Dimethyl-1,6-octadien-3-yl acetate (linalyl acetate)	115-95-7	216.0	3.66	4.52	−14.1	−14.1	−14.1	1
P25	Ethylacetate	141-78-6	90.1	2.35	4.73	−12.6	−12.7	−12.5	2
P27	1-Pentanol	71-41-0	105.0	2.53	14.50	−12.5	−12.5	−12.5	1
P28	2-Pentanol	6032-29-7	105.0	2.27	3.74	−12.6	−12.6	−12.6	1
P29	Toluene	108-88-3	101.1	2.32	3.17	−12.5	−12.5	−12.5	2
P33	*cis*-3-Hexen-1-ol	928-96-1	116.0	2.66	6.49	−12.8	−12.8	−12.8	1
P34	Ethylpropionate	105-37-3	107.0	2.72	6.70	−11.9	−11.9	−11.9	1
P36	2-Hexanol	626-93-7	121.0	2.72	7.21	−12.8	−12.8	−12.8	1
P39	*p*-Xylene	106-42-3	117.0	2.65	4.73	−12.3	−12.3	−12.3	1
P43	Ethylbutyrate	105-54-4	124.0	3.05	8.01	−12.0	−12.1	−12.0	2
P57	Amylaceticester (isoamylacetate)	123-92-2	141.0	3.06	7.75	−13.1	−13.1	−13.1	1
P62	4-Isopropenyl-1-methyl-1-cyclohexen (limonene)	138-86-3	158.0	2.91	4.75	−13.2	−13.2	−13.2	1
P81	Dimethylbenzylcarbinol	100-86-7	159.0	2.87	4.07	−14.3	−14.3	−14.3	1
P85	Carbon tetrachloride	56-23-5	81.5	1.55	2.00	−13.3	−13.3	−13.3	1
P87	3,7-Dimethyl-6-octen-1-al (citronellal)	106-23-0	178.0	3.46	10.84	−14.3	−14.3	−14.3	1
P92	2-Isopropyl-5-methylhexanone (menthone)	89-80-5	173.0	3.00	3.68	−14.2	−14.2	−14.2	1

n = the number of D values per molecule.

more than one $\log_{10} D$ value and possibly more than one molecule, the 90th percentile provided a pessimistic D value (i.e., a reliable overestimate) that could be used to assess with enough safety the migration of any known or unknown molecule. If more safety is required, the number of terminal nodes could be reduced to increase the number of data corresponding to each terminal class and, therefore, to increase the robustness of the estimation of the upper percentile. We emphasize that the proposed approach does not require any assumption on the distribution of D values. One can decrease the risk of underestimating the true D coefficient as low as possible by choosing *a priori* an appropriate upper percentile.

Decision trees as a tool for data assimilation

Regression trees could replace advantageously raw collected data, which are intrinsically disparate in number and quality. This process is known as *data assimilation* and is aimed at the control and standardization of the quality of available data. Depending on the goal, this process may be performed continuously when new data are available or in batch mode when a significant amount of new data are collected. Furthermore, it may be used to devise either realistic D values or convenient overestimates. The sampling of molecules or methodologies and the choice of predictors may significantly affect how data are assimilated into

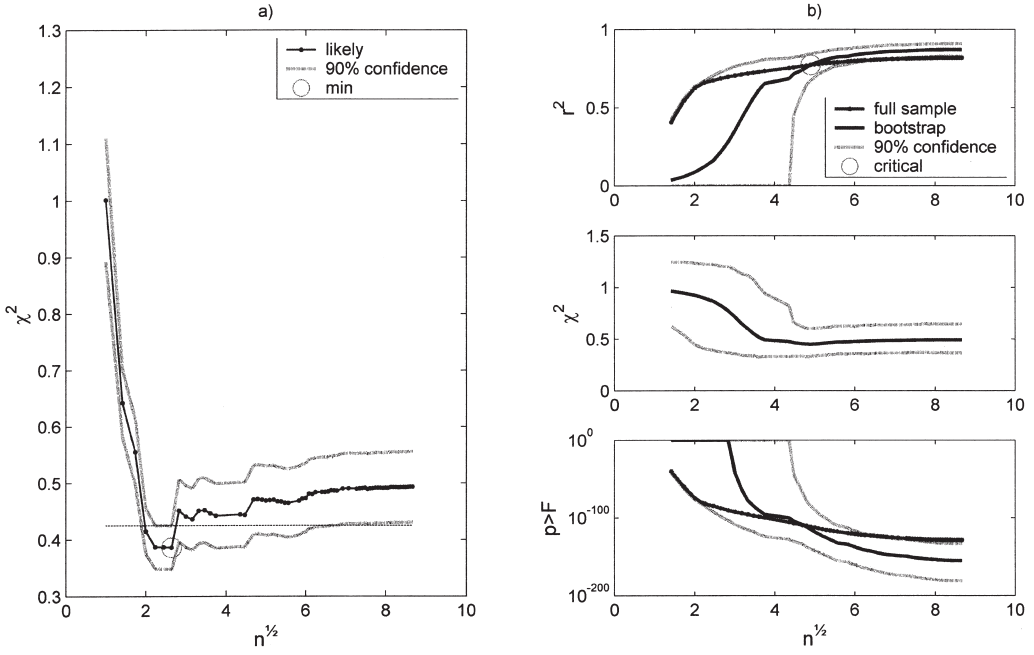


Figure 9 Effect of tree pruning on the error on the prediction error (for LLDPE–LDPE at 23°C). The results were obtained with (a) 10-fold cross-validation and (b) 1000 bootstrap samples. The results are expressed as r^2 values, sums of χ^2 values, and $p > F$ values.

new cognitive structures such as decision trees. To illustrate these possible effects, mismatches between the data and the model, based on either the average and the 95th percentile of each terminal/leaf class, are plotted in Figures 12 and 13, respectively, for each set of molecules. Predicted values obtained with the full regression tree and with the optimally pruned tree are plotted. Figures 12 and 13 do not depict the validation results, as the regression trees were, in this case, fitted over the whole sample, but illustrate the closeness (Fig. 12) or the overestimation (Fig. 13) of the predicted results when they replaced the initial data. Because, the tested predictors were also pertinent *a priori* classifiers, the intrinsic ability of the predictors to assimilate new (unknown) data is illustrated in Figure 14.

Figure 12 shows that regression coefficients (*r*'s) were higher than 0.89 even for the pruned trees. In other words, 89% of the whole variance was at least explained by the proposed models. The discrepancy between the predicted values and the data is summarized in Table V. The deviation between *D* values ranging between 0.51 and 1.15 decades for the full tree and did not increase dramatically when the pruning level was increased because it ranged between 0.79 and 1.31 decades for the optimally pruned tree. As a result of few *D* values for large molecules (>700 g/mol), the deviation was higher when the expected *D* value was lower.

Figure 13 presents how the *D* values based on the 95th percentile of each class overestimated the *D* value available in the database. The number of values, which were lower than the reference values, is also depicted. A quantification of the possible *D* underestimation is given in Table VI. As expected, the full tree overestimated or exactly predicted the true *D*. An equality condition corresponded mainly to a terminal class/leaf that included a single *D* value. As a pruned tree gathers more data and molecular structures, the size of each class was expected to increase and its 95th percentile could either underestimate or overestimate the data-training sample. A satisfactory pruning level should have generated an amount of underestimations that was similar to an *a priori* risk level (i.e., 5% for a regression tree based on 95th percentile values). The optimally pruned regression trees verified underestimation rates of 4.9, 3.5, 5.0, and 0%, respectively for the sets of molecules tested in LDPE–MDPE at 23°C, MDPE–HDPE at 23°C, PP at 23°C, and PP at 40°C (Fig. 13). The maximum underestimation was lower than 0.36 decades and occurred mainly for the set tested in MDPE–HDPE at 23°C. Compared to the noise present in the data and the generally recognized accuracy of conventional methods to assess *D* values, these differences were considered of poor significance. We emphasize that the use of well-controlled *D* data (the set of molecules tested in PP at 40°C) did not lead to any estimation of *D* values.

In the previous discussion, we assumed that a complete training of the regression tree was performed on a significant amount of data. In practice, it would be useful to estimate an unknown *D* or its

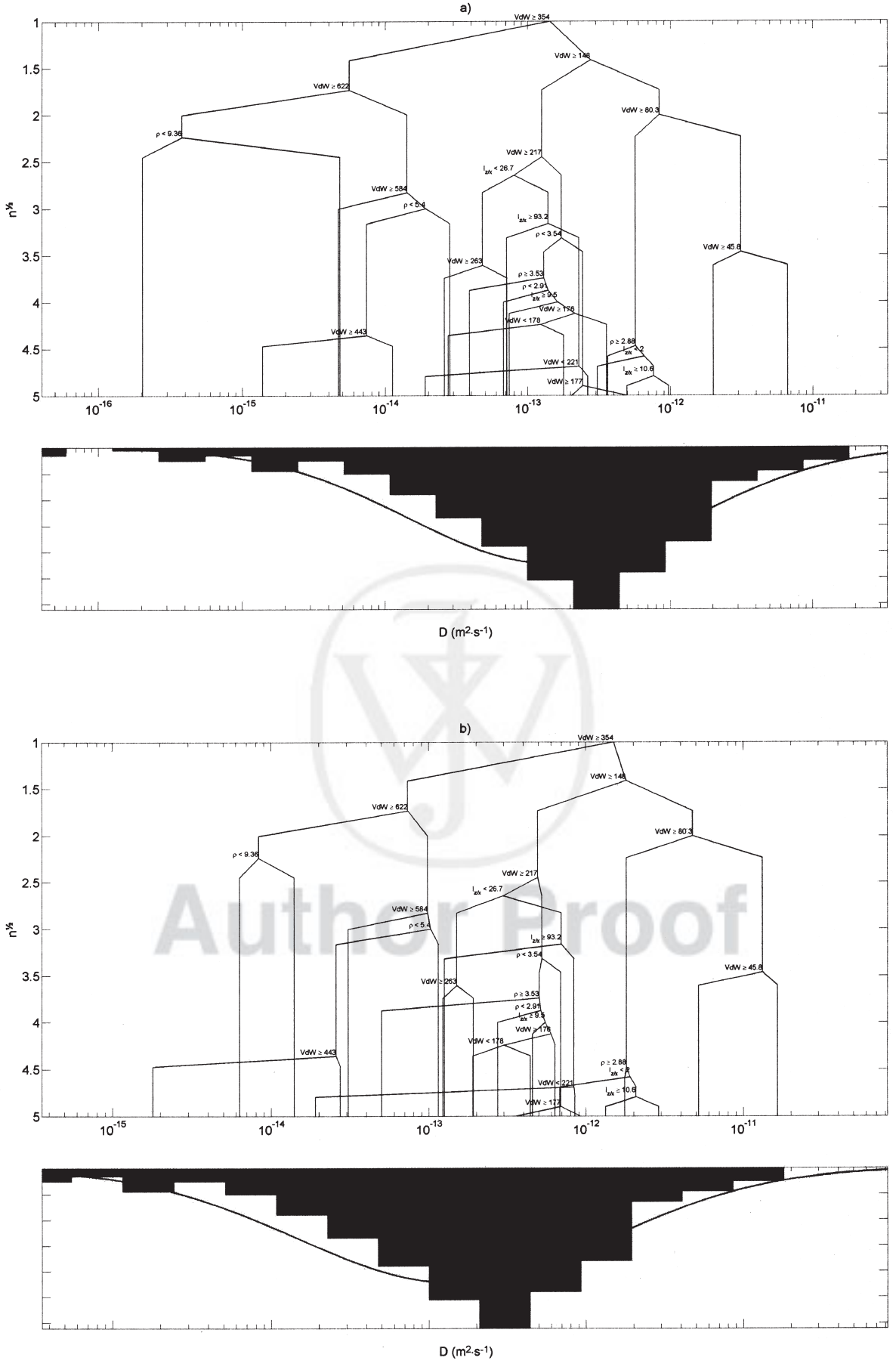


Figure 10 Optimized (postpruned) regression trees for LLDPE–LDPE at 23°C based on the (a) average $\log_{10} D$ values of each class and (b) 90th percentile $\log_{10} D$ values of each class. The corresponding splitting conditions are indicated at each node.

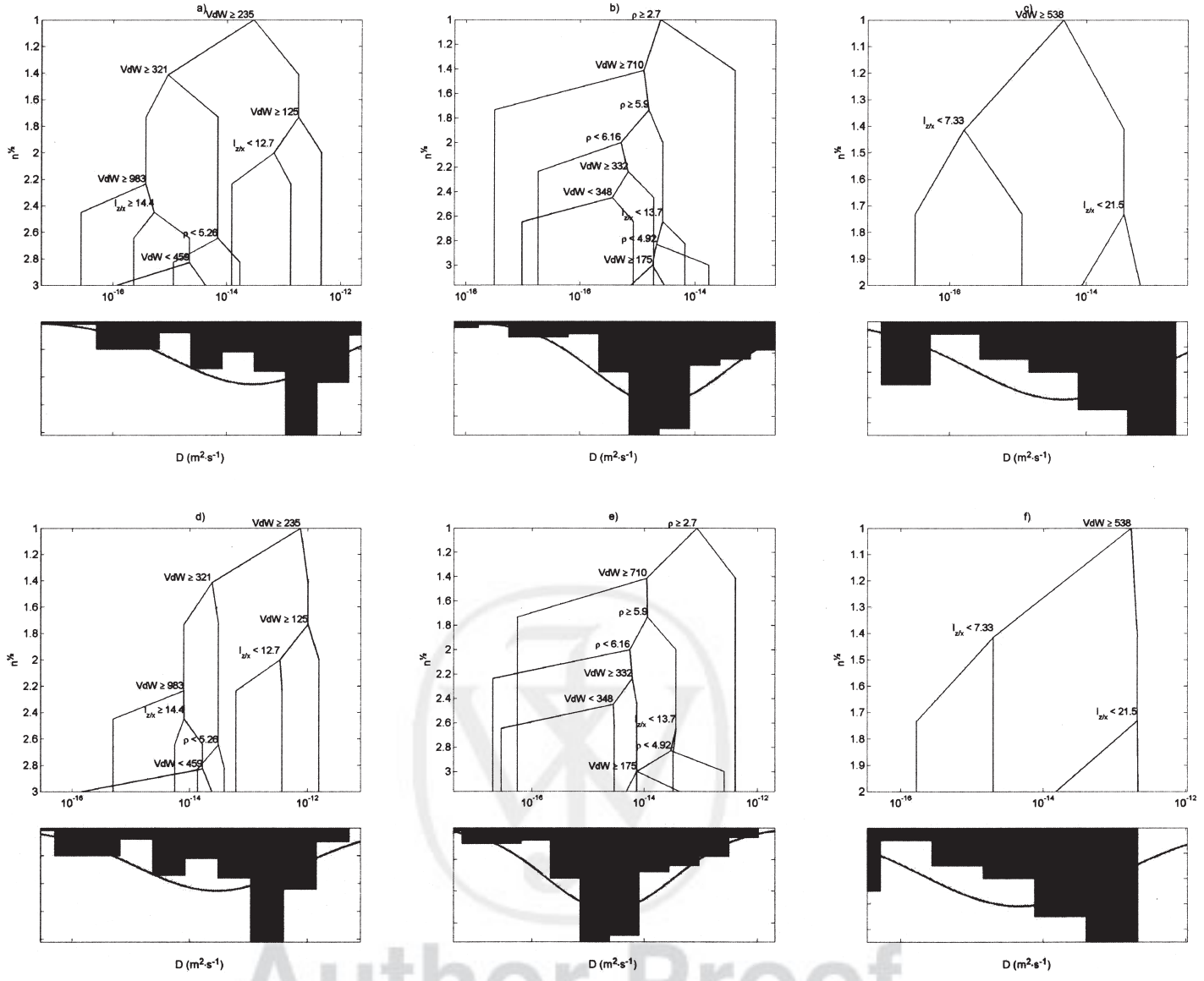


Figure 11 Optimized (postpruned) regression trees based on the average $\log_{10} D$ values of each class for (a) MDPE–HDPE at 23°C, (b) PP at 23°C, and (c) PP at 40°C and based on the 90th percentile $\log_{10} D$ values of each class for (d) MDPE–HDPE at 23°C, (e) PP at 23°C, and (f) PP at 40°C. The corresponding splitting conditions are indicated at each node.

TABLE IV
Main Properties of the Optimal Regression Tree for Each Set of Molecules

Polymer	Temperature (K)	Number of terminal nodes	Number of molecules per class	
			Minimum	Maximum
LLDPE–LPDE	296	25	1	23
MPDE–HDPE	296	9	1	27
PP	296	10	2	14
PP	313	4	3	13

overestimate from more general conditions, for instance, from its molecular structure similarities with a small set of molecules with well-known properties. Thus, we would intuitively expect that the new molecule would have a D similar to the molecules that resemble it. This homologous approach based on an *a priori* classification trees is compared in Figure 14 to the approach requiring an *a posteriori* adjustment. We confirmed that both approaches were highly correlated and presented a significant bias. The variability in D estimates, depicted as interquartile range, which was not minimal when no training was performed, was mainly responsible for the observed differences. Unexpected behaviors appeared for very few molecules and were mainly

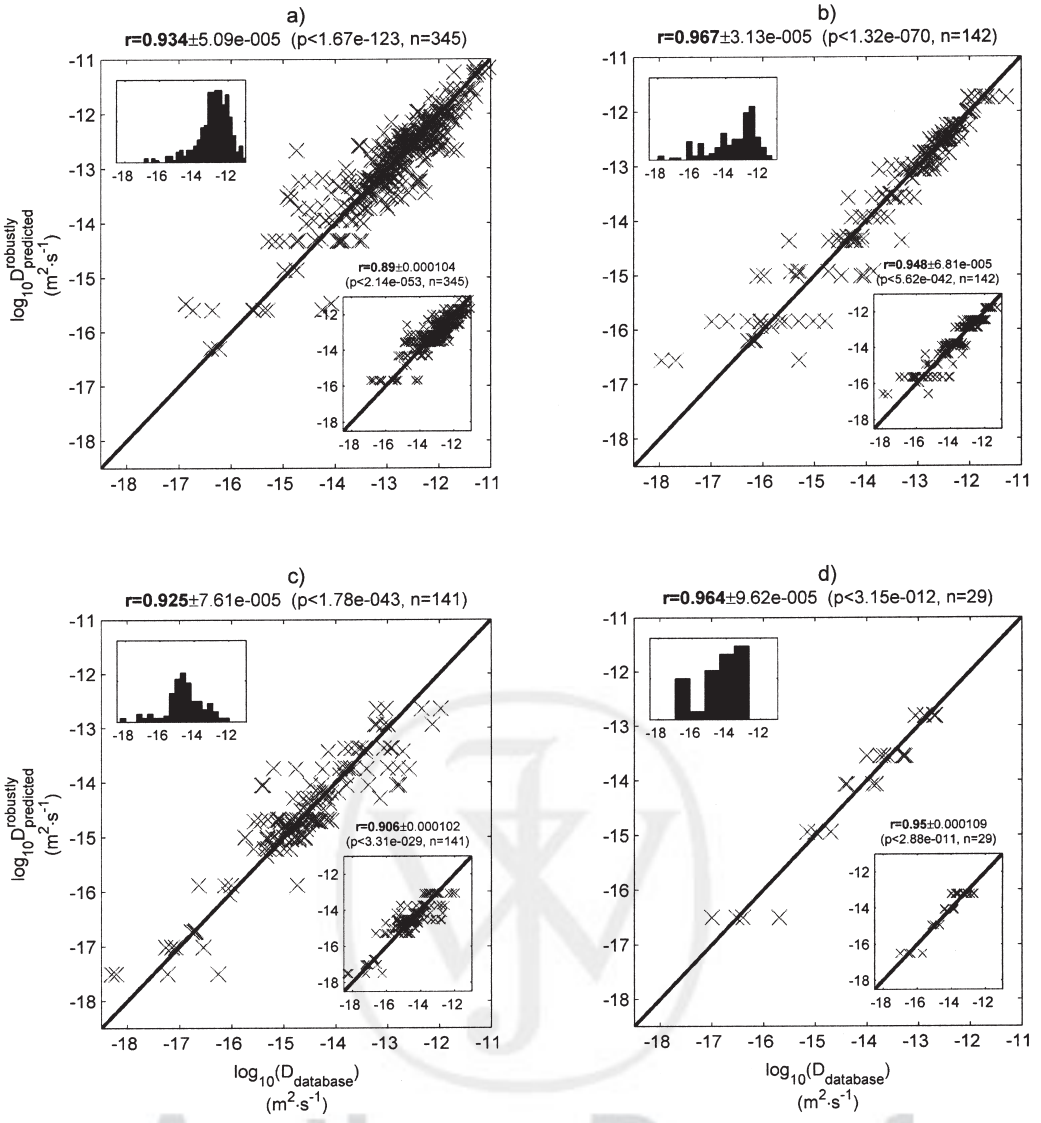


Figure 12 Predicted average $\log_{10} D$ values versus values available in the database for (a) LLDPE–LDPE at 23°C, (b) MDPE–HDPE at 23°C, (c) PP at 23°C, and (d) PP at 40°C. Each main subplot represents the values predicted with the full regression tree, whereas the lower inserted subplots present the results inferred with the optimally pruned regression tree. The upper inserted subplots depict the distribution of $\log_{10} D$ values as available in the database.

related to experimental errors. Deviations typically ranged between 1 and 2 decades.

CONCLUSIONS

This article presents new predictive estimators of D's based on decision trees for molecules with M values ranging between 50 and 1200 g/mol in polyolefin materials. The main features of decision trees for the prediction of D's are that they are nondeterministic (non-theory-dependent) and nonlinear (able to handle complex behaviors or interactions). The final models were highly comprehensible and should be adaptable for a wide range of practical uses: the design of materials with controlled transport properties, the verification of compliance of food contact materials, and food sanitary surveys based on a risk assessment of the migration of substances from packaging materials into food articles.

AQ: 18

According to a likely transport mechanism involving reptation-like motions, three topological descriptors of the molecular structure in the absence of interaction with the polymer matrix (V_{WdV}, ρ, and $I_{z/x}$) were used as both classifiers and predictors. The whole approach was tested on 628 D's (242 molecules) collected by the European SMT-CT98-7513 project and normalized at 23°C for three typical polyolefin materials (LLDPE–LDPE, MDPE–HDPE, PP) and on 29 D's

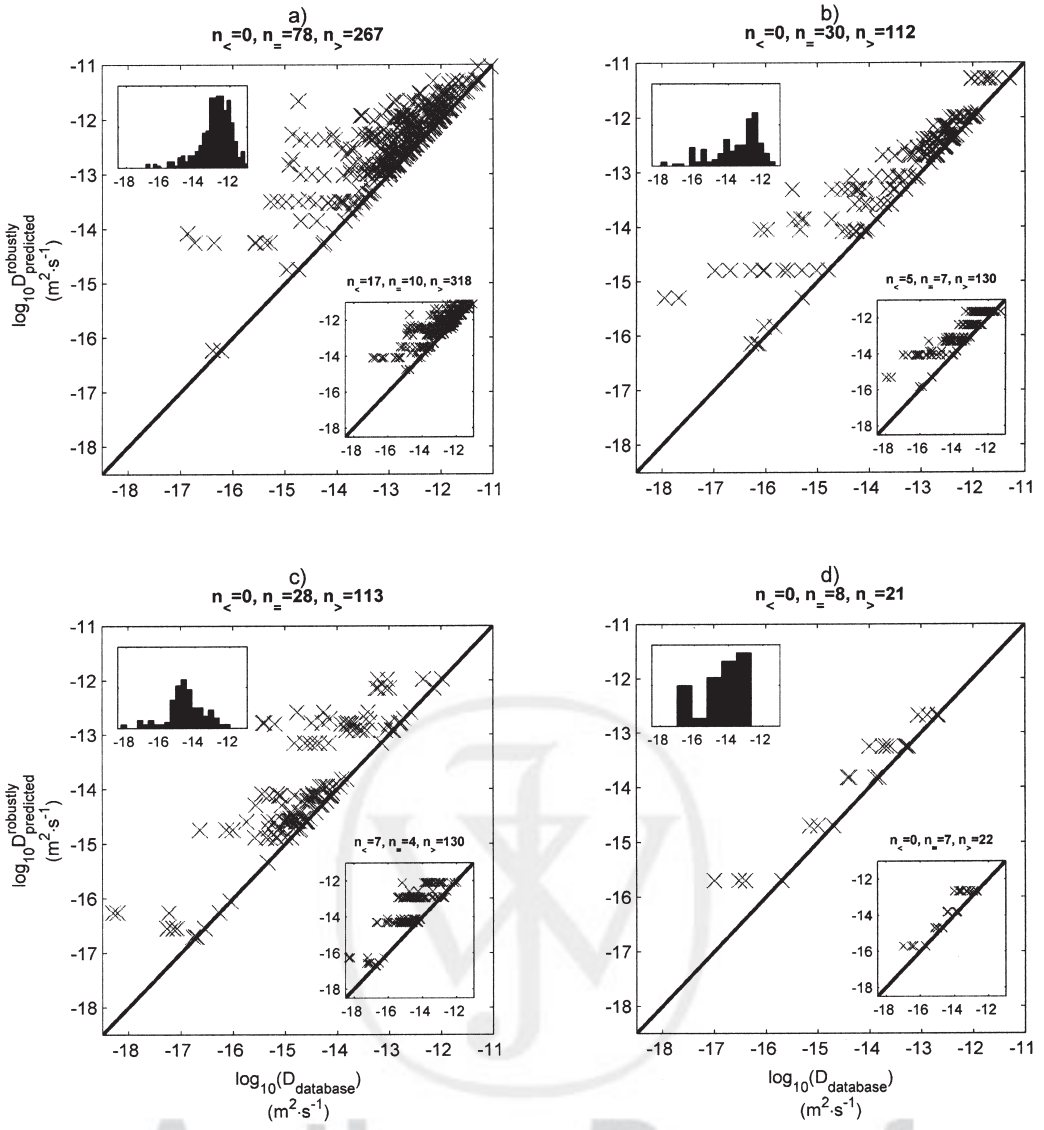


Figure 13 Predicted 95th percentile of the $\log_{10} D$ values versus values available in the database for (a) LLDPE–LDPE at 23°C, (b) MDPE–HDPE at 23°C, (c) PP at 23°C, and (d) PP at 40°C. Each main subplot represents the values predicted with the full regression tree, whereas the lower inserted subplots present the results inferred with the optimally pruned regression tree. The upper inserted subplots depict the distribution of values as available in the database.

(26 molecules) measured in PP at 40°C. The three descriptors, which were chosen independently of the polymer matrix, were derived from the molecular structure of each molecule. All 267 molecules were minimized *in vacuo* and were subsequently oriented along their main axes.

The three parameters were partially correlated with M with a correlation rate that varied significantly with the 3D shape of the considered molecules. Thus, ρ and $I_{z/x}$ of nonspherical molecules were poorly correlated with M. We also verified that none of the distributions of each parameter could alone fit the distribution of the collected D's. Also, with an *a priori* tree classification, we were able to gather pertinent molecules with similar D's. Tree regressions increased the previously mentioned predictability and significantly reduced the risk of misclassification. Both cross-validations and bootstrap sampling demonstrated that the proposed approach was predictive and robust when the regression tree was conveniently simplified by prepruning and postpruning. In particular, we demonstrated that correlation coefficients ranging between 0.74 and 0.96 (90% confidence interval assessed by bootstrap sampling) were achievable for all of the polymers. In detail, the optimal regression trees involved all of the tree descriptors with almost the same importance. For a rough classification of diffusing behaviors over a scale of D's ranging between 10^{-17} and 10^{-13} m^2/s, V_{WdV} predominated for the subset of molecules tested in polyethylene, whereas a combination of ρ and of V_{WdV} was required in PP at 23°C.

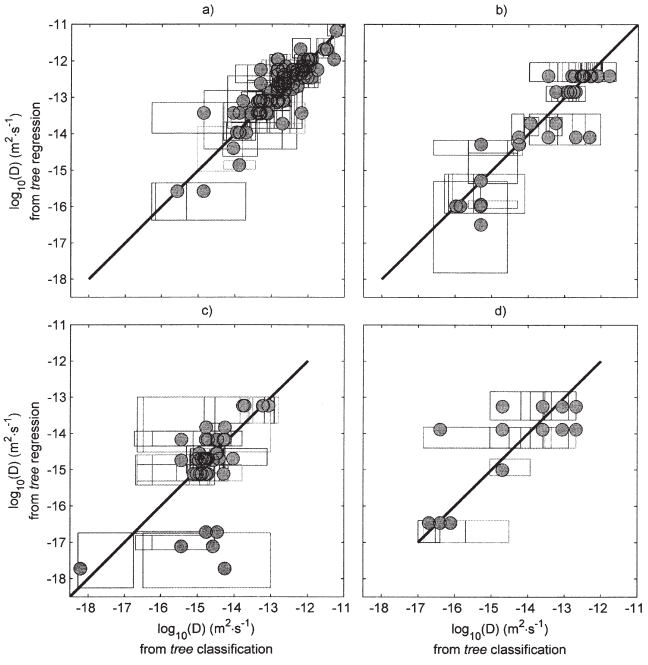


Figure 14 Comparison of $\log_{10} D$ values obtained from CARTs after optimal pruning for (a) LLDPE–LDPE at 23°C, (b) MDPE–HDPE at 23°C, (c) PP at 23°C, and (d) PP at 40°C.

For the small subset of molecules tested at 40°C, a similar range of D, V_{WdV}, and $I_{z/x}$ values were predominant.

AQ: 18

As regression trees aim at gathering molecules with similar D's, a methodology was devised to robustly calculate overestimates of D's from the upper percentiles of values collected within the same class. By assuming a risk of 10% to underestimate the D's collected in the current database, we tested the 90th percentile values of each terminal class/nodes on an optimally pruned tree. For all polymers, the risk level was never exceeded and the maximum underestimation was slightly higher (by a factor 2) than the conventional uncertainty in the measurements of D's. In addition to their foreseen ability and their iterative construction process, decision trees were also proposed for the standardization, (re)evaluation, and/or adjustment of new data or of previously collected data. This operation, known as data assimilation, aims at transforming an initial set of disparate data into well-validated data for the purposes defined previously. This approach was strengthened by a demonstration that the three proposed descriptors used as either as classifiers or as regressors led to similar conclusions. Because of the absence of any fitting/learning when they were used as classifiers, the risk of misclassification was higher but did not exhibit a significant bias. As a result, the three molecular descriptors that did not account for the interactions with the polymer matrix seemed remarkable descriptors to define the analogies between diffusants.

Further work is desirable to extend the proposed decision trees to new polymers and to account for temperature effects. Many properties that control the experimental values of D, such as the temperature, polymer density, and crystallinity, could be intuitively handled through a discrete formulation as decision trees do. Indeed, for practical uses, D's need to be known for a given range of a given parameter (e.g., temperature). An estimation of the uncertainty due to this simplification could be straightforwardly and robustly assessed from the lower and upper percentiles of each terminal leaf/node.

TABLE V
Maximum Deviations Between the Predicted D Values and the Values Available in the Database

Polymer	Temperature (K)	95th percentile of $\lvert\log_{10}(D_{predicted}/D_{database})\rvert$ Full tree	Optimally pruned tree
LLDPE–LPDE	296	1.04	1.08
MPDE–HDPE	296	1.14	1.15
PP	296	1.15	1.31
PP	313	0.51	0.79

The predicted D values were based on the average of all data available at each node/class.

TABLE VI
Maximum Underestimations of the D Values and the Values Available in the Database

Polymer	Temperature (K)	5th percentile of $\log_{10}(D_{predicted}/D_{database})$ for all values that verify $D_{predicted} < D_{database}$ Full tree	Optimally pruned tree
LLDPE–LPDE	296	None	0.23
MPDE–HDPE	296	None	0.36
PP	296	None	0.22
PP	313	None	None

The predicted D values were based on the average of all data available at each node/class.

NOMENCLATURE

D	diffusion coefficient (m^2/s)
V_{WdV}	Van der Waals volume [$Å^3$]
ρ	gyration radius [Å]
$\vec{\mathbf{x}}_\mathbf{i}$	position vector of the atom i
$\vec{\mathbf{x}}_\mathbf{0}$	position vector of the center of mass
$\langle\ \rangle$	average operator
$I_{z/x}$	shape parameter
$i_{(t)}$	impurity function at node t
t	parent node index
J_t	total number of child nodes at node t

$\Delta e_{(t,s)}$	least-squares error criterion that splits the node t with the split value
s	split value
$e_{(j)}$	mean squared error at node j
M	molecular weight (g/mol)
$p(j\|t)$	proportion of each class $j_{j=1\ldots J_t}$ at node t
α	phenomenological exponent connecting D and M
c	third-order Kier and Hall cluster count index
r^2	determination coefficient
χ^2	average quadratic error of prediction
$p > F$	significance levels
r	regression coefficient

References

1. Goddard, W. A., III; Cagin, T.; Blanco, M.; Vaidehi, N.; Dasgupta, S.; Floriano, W.; Belmares, M.; Kua, J.; Zamanakos, G.; Kashihara, S.; Iotov, M.; Gao, G. Comput Theor Polym Sci 2001, 11(5), 329.
2. Theodorou, D. N. SIMU Programme Newsletter 2000, No. 1, 19.
3. Karlsson, G. E.; Johansson, T. S.; Gedde, U. W.; Hedenqvist, M. S. Macromolecules 2002, 35, 7453.
4. Hinrichs, K.; Piringer, O. Evaluation of Migration Models to Be Used Under Directive 90/128/EEC; Technical Report EUR 20604EN; European Commission, Directorate-General for Research: Brussels, Belgium, 2002.
5. Official J 2002, L220, 18.
6. Vitrac, O.; Hayert, M. AIChE J 2005, 51, 1080.
7. Vitrac, O.; Challe, B.; Leblanc, J.-C.; Feigenbaum, A. Food Additives Contam, to appear.
8. Vitrac, O.; Leblanc, J.-C.; Feigenbaum, A. Food Additives Contam, submitted.
9. Loh, W. Y.; Vanichsetakul, N. J Am Stat Assoc 1988, 83(403), 715.
10. Cheng, Y.; Prud'homme, R. K.; Thomas, J. L. Macromolecules 2002, 35, 8111.
11. Favre, E.; Leonard, M.; Laurent, A.; Dellacherie, E. Colloids Surf A 2001, 194, 197.
12. Harmandaris, V. A.; Mavrantzas, V. G.; Theodorou, D. N.; Kröger, M.; Ramírez, J.; Öttinger, H. C.; Vlassopoulos, D. Macromolecules 2003, 36, 1376.
13. Kovarski, A. L. In Molecular Dynamics of Additives in Polymers; VSP: Utrecht, The Netherlands, 1997; Chapter 4.
14. Reynier, A.; Dole, P.; Humbel, S.; Feigenbaum, A. J Appl Polym Sci 2001, 82, 2422.
15. Reynier, A.; Dole, P.; Feigenbaum, A. J Appl Polym Sci 2001, 82, 2434.
16. Zhao, Y. H.; Abraham, M. H.; Zissimos, A. M. J Org Chem 2003, 68, 7368.
17. Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. Classification and Regression Trees; Wadsworth: Pacific Grove, CA, 1984.
18. Loh, W. Y.; Shih, Y. S. Stat Sinica 1997, 7, 815.
19. Cussler, E. L. Diffusion, Mass Transfer in Fluids Systems, 2nd ed.; Cambridge University Press: Cambridge, England, 1997.
20. Brandt, W. W. J. Phys Chem 1959, 63, 1080.
21. Willmann, R. D. J Chem Phys 2002, 116, 2688.
22. Lodge, T. P. Phys Rev Lett 1999, 83, 3218.
23. Bandyopadhyay, T.; Ghosh, S. K. J Chem Phys 2003, 119, 572.
24. Bird, R. B.; Stewart, W. E.; Lightfoot, E. N. Transport Phenomena, 2nd ed.; Wiley: New York, 2002.; Chapter 17.
25. Masaro, L.; Zhu, X. X. Prog Polym Sci 1999, 24, 731.
26. Kovarski, A. L. In Molecular Dynamics of Additives in Polymers; VSP: Utrecht, The Netherlands, 1997; Chapter 7.
27. Manabe, S.-I. In Diffusion in Polymers; Neogi, P., Ed.; Marcel Dekker: New York, 1996; Chapter 5.
28. Kier, L. B.; Hall, L. H. Molecular Connectivity in Chemistry and Drug Research; Academic: New York, 1976.

AQ1: Please confirm the last column of the table as edited per journal style, with reference number citations inserted in place of the table footnotes.

AQ2: Please spell out all abbreviations in the affiliation, including INRA, UMR, FARE, and BP.

AQ3: Please confirm that LDPE is defined correctly here.

AQ4: Please confirm that LLDPE is defined correctly here.

AQ5: Please confirm that MDPE and HDPE are defined correctly here.

AQ6: Please define EU.

AQ7: Please confirm that CART is defined correctly here.

AQ8: Please define $e_{(t)}$ after eq. (3).

AQ9: Please spell out N_j.

AQ10: Please confirm that PP is defined correctly here.

AQ11: Please define $M_{\min}$ and $M_{\max}$ in the table footnote or spell them out in the table.

AQ12: Please define CAS, $\log_{10}(D)_{\min}$, and $\log_{10}(D)_{\max}$ in the table footnote or spell them out in the table.

AQ13: Please spell out CPK.

AQ14: Unless this journal is paginated by issue (i.e., each issue begins on page 1), please delete the issue number.

AQ15: If the publication status has changed, please update the reference.

AQ16: Please provide the names of all the authors.

AQ17: Please confirm the location of the publisher as added.

AQ18: Please confirm the sentence as edited.

AQ19: Please consider changing phrasing in this paragraph such as [font"Arial MT"]"[/font"Arial MT"]the first subsection[font"Arial MT"]"[/font"Arial MT"] to the actual title of the section of the article. It is not clear to which sections you are referring throughout this paragraph.

AQ20: Please provide the location of the manufacturer (city and state in the United States and city and country elsewhere).