Journal of Membrane Science 691 (2024) 122220

Contents lists available at ScienceDirect

Journal of Membrane Science

journal homepage: www.elsevier.com/locate/memsci

Estimating Gas Sorption In Polymeric Membranes From The Molecular Structure: A Machine Learning Based Group Contribution Method For The Non-Equilibrium Lattice Fluid Model (ML-GC-NELF)

Hasan Ismaeel [a], David Gibson [a], Eleonora Ricci [b,1], Maria Grazia De Angelis [a,*]

[a] Institute for Materials and Processes, School of Engineering, University of Edinburgh, Sanderson Building, Robert Stevenson Road, Edinburgh, EH9 3FB, United Kingdom

[b] Department of Civil, Chemical, Environmental and Materials Engineering, University of Bologna, Via Terracini 28, 40131, Bologna, Italy

ARTICLE INFO

Keywords:
NELF model
Group contribution
Equations of state
Gas sorption
Polymer membranes
Machine learning
Lattice fluid theory

ABSTRACT

Since its inception, the non-equilibrium lattice fluid (NELF) model has become a vital tool in correlating and predicting the gas solubility behaviour in glassy polymeric membranes. But like its equilibrium variant, the NELF model is highly constrained by the availability of the pure polymer characteristic parameters, which are not always convenient to obtain as the need arises. In this study, we provide a proof-of-concept for building a machine learning-based group contribution method (ML-GC) for the Sanchez–Lacombe equation of state (EoS) pure polymer parameters. The ML-GC model was built using a modified version of the Marrero and Gani's method, which incorporates machine learning regression into the GC parameterisation process. The final model was capable of reproducing the parameters of a randomly selected test set of polymers, with a diverse range of chemical structures. The resultant average AARD% of the predicted densities in this set is 5.59%, with no polymer exceeding 15%. Moreover, to test the model's capabilities in estimating the parameters of high glass transition temperature polymers, we predicted *a priori* the characteristic parameters of 6 polyimides from the knowledge of their molecular structure. The ML-GC parameters were also incorporated into the NELF model to predict the infinite dilution solubility coefficients (S_0) of some of these polymers and the results were validated against experimental data. Furthermore, the ML-GC-NELF model was also used for the first time to represent effectively the gas solubility isotherms in PIM-PI-SBI and PIM-PI-EA with relatively small magnitudes of the binary interaction parameters (k_{ij}). Despite the small data-set used herein, the model performance was satisfactory, however, as more data are being published in literature, the proposed ML-GC model has the potential of providing even more accurate predictions for a wider range of polymers, ultimately leading to lesser reliance on experimental data for modelling the gas sorption.

1. Introduction

The ubiquity, processability, and affordability of polymeric materials have made them an attractive choice in the development of membranes for gas separation processes [1]. However, the energy efficiency that is associated with polymer-based membrane gas separation processes comes at a cost. Indeed, an inherent permeability–selectivity trade-off, commonly known as the Robeson upper bound, can be observed [2]. A deep understanding of what governs the permeation and selectivity of gases is pertinent for the development of novel polymeric membranes that outperform the current state of the art materials [3].

The transport of a gas penetrant in a dense polymeric membrane can be described using the solution-diffusion (SD) model, which states that the permeability is the product of a kinetic term (i.e the diffusivity) and a thermodynamic factor (i.e. the solubility) [4]. Consequently, being able to systematically analyse the solubility from a modelling prospective is important to characterise both the permeability and selectivity of a gas in a polymeric membrane, particularly for processes targeting CO_2 capture. By and large, CO_2 often exhibits higher solubility in polymers

* Corresponding author at: Institute for Materials and Processes, School of Engineering, University of Edinburgh, Sanderson Building, Robert Stevenson Road, Edinburgh, EH9 3FB, United Kingdom

E-mail address: grazia.deangelis@ed.ac.uk (M.G. De Angelis).

[1] Current address: Institute of Informatics and Telecommunications & Institute of Nanoscience and Nanotechnology, National Centre for Scientific Research "Demokritos", Agia Paraskevi, 15341 Athens, Greece.

https://doi.org/10.1016/j.memsci.2023.122220
Received 21 August 2023; Received in revised form 13 October 2023; Accepted 29 October 2023
Available online 3 November 2023

compared to other light gases, hence, the solubility–selectivity plays a critical role in determining the overall membrane performance [5].

The generality and robustness of equations of state (EoS) have made them a powerful tool in the modelling of the thermodynamics of polymer-containing systems. EoS based on the lattice fluid theory (LF) [6–9], statistical associating fluid theory (SAFT) [10,11], and the perturbed hard-sphere-chain (PHSC) model [12] have been successfully applied in modelling the gas solubility in polymers at equilibrium (i.e. in the rubbery and molten states) [13–16].

Currently, the vast majority of polymers at the Robeson upper bound are glassy [17], which means that they exist in an intrinsic state of non-equilibrium. Hence, the aforementioned EoS approaches fundamentally cannot describe such systems. A popular alternative to EoS that is often capable of describing the gas solubility in glassy polymers is the dual mode sorption (DMS) model [18,19]. The model represents the gas penetrant as two distinct populations of sorbed molecules. On one hand, the first group of molecules are dissolved into the bulk, in a similar manner to rubbery and molten polymers, and is described by Henry's law. On the other hand, the second group is adsorbed onto the microvoids that are present in the glassy polymer phase and is governed by Langmuir adsorption isotherm. The widespread appeal of this model comes from its simplicity and the ease of fit of its three adjustable parameters. While the DMS model can correlate the majority of gas sorption data very well, it may fail in describing complex solubility isotherms such as the sigmoidal shape observed in alcohol sorption in polymers. Moreover, the parameters are dependent on the operating conditions (such as the fitted pressure range), the polymer history, and the penetrant-polymer couple, which impacts the model's predictive capabilities [20–22].

The non-equilibrium thermodynamics for glassy polymers (NET-GP) is another well established framework for modelling the gas sorption in the amorphous glassy phase. In essence, the NET-GP can extend any EoS to account for the non-equilibrium state, and give rise to its non-equilibrium form. The non-equilibrium lattice fluid (NELF) [23–25], based on the Sanchez and Lacombe (SL) EoS [6–8], is the most widely applied case of this approach. Similar to its equilibrium variant, the NELF model requires pure component parameters for both the penetrants and the polymer to carry out the solubility calculations. The gas parameters are often fitted to saturated vapour pressure and liquid density data, while the parameters of the polymer are usually obtained from pressure–volume–temperature (PVT) data in the rubbery region, above the glass transition temperature (T_g). The NELF model was shown to have a remarkable predictive prowess. For instance, Minelli et al. [26] predicted the solubility of mixed gases in glassy polymers from the pure gas parameters and the binary interaction parameters (obtained from the pure gas solubility isotherms) accurately. In contrast, Ricci et al. [27] have conducted a detailed sensitivity analysis on the predictions of mixed-gas sorption from the pure gas-polymer DMS parameters and found that in some cases, the model can only provide a qualitative agreement between the predictions and the sorption data.

Despite that, the applicability of the NELF model is limited by the availability of the pure component parameters, as the required experimental data to be fitted are often not available for polymers and it is not always convenient to obtain them, due to constraints such as time and cost. In addition, the vast majority of the polymers in the vicinity of the Robeson upper bound exhibit high T_g [28]. As a result, some of these polymers would chemically degrade before reaching the rubbery state. Some of these polymers include polyimides, and polymers of intrinsic microporosity (PIMs). To circumvent this issue, the EoS parameters can be acquired from solubility data [29–31], however this method is neither predictive nor general, and would require access to solubility data for multiple gas species.

Another avenue for modelling the gas sorption in polymers is through molecular simulations. In particular, the grand canonical Monte Carlo (GCMC) [32], Gibbs ensemble Monte Carlo (GEMC) [33], staged particle deletion (SPD) Widom [34], test particle insertion [35], and direct particle deletion (DPD) [36,37] methods can be applied for such efforts. However, many computational complexities can arise, especially for the case of a glassy amorphous polymer phase. Firstly, the characteristic relaxation times of the glassy polymer occurs in time scales that are currently inaccessible by atomistic methods, which may lead to problems in the generation of realistic configurations of the glassy polymer. Secondly, at high pressure, or in the presence of sorbing agents, swelling effects must be accounted for. This could be addressed by relying on pre-swollen atomistic packing models [38,39] or by resource-intensive iterative procedures [40]. These complexities, alongside the relatively larger computational cost associated with molecular methods, may restrict the applicability of these simulations to gas sorption processes.

As an intermediate bridge between atomistic methods and EoS approaches, Minelli et al. [41] have proposed a multi-scale approach to model the gas solubility and obtain the EoS parameters by coupling both methods. Here, the PVT behaviour of two high T_g polyimides, Ultem and Kapton, were obtained *in silico* via molecular dynamics (MD) simulations above T_g, which are inaccessible experimentally. From there, the pure polymer parameters of the PC-SAFT EoS were fitted to the PVT data and the pure gas solubilities below (T_g) were modelled using the NET-GP framework (i.e. through the NE-PC-SAFT). Ricci et al. [42] applied the same technique to model the mixed CO_2/CH_4 gas solubility in Matrimid and attained the solubility–selectivity as a function of temperature, pressure and composition. While this approach greatly reduces the computational cost by delegating the solubility predictions to the EoS, this procedure remains intractable when probing the vast chemical space of polymers with unreported EoS parameters.

With the advent of machine learning (ML) as a powerful correlation tool, many data-driven models were built to predict various transport and thermophysical properties of polymeric materials [43–48]. Li et al. [43] have built a backpropagation artificial neural network (ANN) to estimate the solubility of CO_2 and N_2 in polystyrene, and CO_2 in polypropylene. Similarly, Ting and Yuan [44] have trained a radial basis function (RBF) ANN to estimate the CO_2 solubility in 7 different polymers. However, a major limitation of both these works is the lack of chemical structure information inputted into the model. Hence, these applications tend to be limited to a small sub-set of polymers and gas species. ML methods can aid in the development of quantitative structure–property relationships (QSPR), where a functional relationship between the chemical structure of the polymer and the property is determined. ML-QSPR have proven to be indispensable in the field of polymer informatics, and were not only successful in predicting relevant properties [49–51], but also in identifying hypothetical polymers above the Robeson upper bound [45,48].

QSPR methods can also be coupled with EoS to improve their predictive capabilities. The most widely used QSPR for such applications are group contribution methods (GCM). Throughout the past few decades, several GCMs have been applied to estimate the pure and mixture EoS parameters for polymers. High and Danner have developed a group contribution lattice fluid (GCLF) EoS for polymer systems [52,53]. A modified version [9] of this GCLF EoS was later implemented to predict the solubility of CO_2 in polymer melts [14]. Tihic et al. [54] have employed the Constantinou and Gani's GCM [55] to estimate the non-associating pure component parameters for the simplified PC-SAFT. However, in the aforementioned studies, the GC values were regressed to low molecular weight compounds vapour pressure and liquid density experimental data. This may result in a poor extrapolation of the estimations to polymer substances.

Constantinou and Panayiotou [56] have applied Constantinou and Gani's GCM to determine the SL-EoS pure component parameters using a small data set of polymers. Due to the promising results, the model was later extended [57] to include 58 polymers. Moreover, Peters et al. [58,59] have also obtained the non-associating PC-SAFT pure component parameters by the GC parameterisation of a small set of polymers using the Lorentz–Berthelot mixing rule as the objective

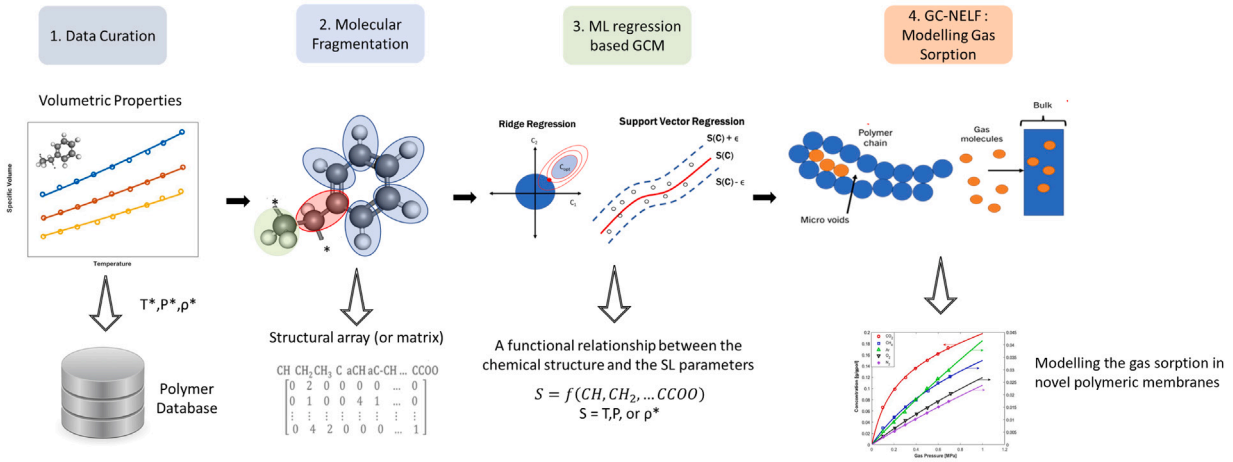


Fig. 1. The workflow for the development of a machine learning based group-contribution non-equilibrium lattice fluid (ML-GC-NELF) model and its application to gas sorption. (**1**) Collection and determination of pure component characteristic parameters from literature and volumetric properties. (**2**) Segmentation of the repeating unit of the polymer database. (**3**) Machine learning-based regression of the group contribution values for the characteristic Sanchez–Lacombe parameters. (**4**) Modelling gas sorption in glassy polymers using the ML-GC-NELF model.

function of the minimisation i.e. by minimising the deviation of the experimentally obtained EoS parameters from the mixing rules values of the groups. While these studies did include polymers in their GC parameterisation scheme, the data sets were relatively small, and a proper test set to verify the ability of these models to generalise to other polymers was lacking.

More recently, ML-QSPR models for the PC-SAFT pure component parameters have surfaced in literature. Matsukawa et al. [60] have built an ANN regression-based GCM to capture the non-linear relationships between a structurally diverse set of low molecular weight compounds and their EoS parameters. Similarly, Habicht et al. [61] used molecular descriptors generated from extended-connectivity fingerprints (ECFPs) [62] to build another ANN model for the EoS parameters of PC-SAFT. However, ML-QSPR are yet to be applied to exclusively predict the pure component EoS parameters of polymers

In this work, a ML regression-based GCM was developed for the estimation of the SL-EoS pure component parameters for polymers and the following protocol was used (see Fig. 1).

(1) A database of 102 polymer SL EoS parameters was built using values from the literature for 37 homopolymers, followed by fitting additional 65 EoS parameters to pressure–volume–temperature (PVT) data for new homopolymers to expand the database.

(2) The Marrero and Gani molecular fragmentation process [63] was employed to determine the structural groups of the monomers (i.e. repeating units) of the polymer data set.

(3) The GC values were obtained for the SL characteristic parameters of the polymers, by using linear regression and ML regression and the results were validated against the SL parameters and density values of a randomly selected test set.

(4) The GC model was used to predict the SL parameters for 6 high T_g polyimides and PIMs, not previously included in the database. For two of these polymers, no SL parameters were available in the literature for comparison.

(5) The ML-GC-NELF model was validated by evaluating the solubility and PVT data in the six polymers and comparing it, where possible, with the NELF model, i.e. the version using the conventionally retrieved values of the polymers SL parameters.

As a result, new SL parameters for 2 high-performance materials (i.e. PIM-PI-EA and PIM-PI-SBI), that could not be previously modelled through an EoS approach, were determined and used to describe the gas solubility via the ML-GC-NELF, to demonstrate a use-case of the proposed modelling strategy.

2. Theory

2.1. The Sanchez and Lacombe Equation of State (SL-EoS)

The Sanchez and Lacombe equation of state (SL-EoS) [6–8] uses the lattice fluid (LF) theory to represent molecules, which are assumed to be ordered in space in a lattice structure. This allows the model to 'coarse-grain' the molecular configurations and interactions, greatly simplifying the process of deriving a close-form expression for the EoS. The SL-EoS can be defined using the reduced forms of the temperature, pressure, and density ($\tilde{T}$, $\tilde{P}$, $\tilde{\rho}$):

$$\tilde{\rho} = 1 - \exp\left[-\frac{\tilde{\rho}^2}{\tilde{T}} - \frac{\tilde{P}}{\tilde{T}} - \tilde{\rho}\left(1 - \sum_i^{N_c} \frac{\varphi_i}{r_i}\right)\right] \quad (1)$$

$$\tilde{T} = \frac{T}{T^*} \qquad \tilde{P} = \frac{P}{P^*} \qquad \tilde{\rho} = \frac{\rho}{\rho^*} \quad (2)$$

where T^*, P^*, and ρ^* are the characteristic SL-EoS parameters, φ_i is the volume fraction of component i at close packing, and r_i is number of lattice sites occupied per unit (or "mer") of component i. The pure polymer characteristic SL parameters are usually obtained by fitting to PVT data. When dealing with mixtures, mixing rules must be applied to determine the characteristic parameters. A detailed description of these rules, alongside other relevant information to the SL-EoS are shown in table S1 (supplementary material).

Since the SL-EoS inception, many pure polymers characteristic parameters have been determined. These parameters have been fitted to various experimental data. The vast majority are from PVT properties, but others have been derived from thermal expansion and pressure coefficients [64], and gas sorption data [29]. In addition, new characteristic parameters were obtained in this work for polymers with available PVT properties. To obtain the best fit, the mean squared error (MSE) was chosen as the cost function:

$$MSE = \frac{1}{n}\sum_i^n (\rho_i^{exp} - \rho_i^{pred})^2 \quad (3)$$

where ρ_i^{exp} and ρ_i^{pred} are the experimental and predicted density values of point i respectively, and n is the total number of experimental points.

2.2. The Non-Equilibrium Lattice Fluid (NELF) model

The non-equilibrium thermodynamics for glassy polymers (NET-GP) [23–25] theory provides the fundamental framework to extend the SL-EoS to its non-equilibrium variant, i.e. the non-equilibrium

lattice fluid (NELF) model. The theory postulates the following: (1) The penetrant-polymer system is homogeneous, isotropic, and amorphous. (2) The system can be described using the macroscopic thermodynamic state variables i.e. temperature (T), pressure (P), and composition (Ω) alongside an additional order parameter, the polymer density (ρ_{pol}), that accounts for the systems departure from equilibrium. In effect, this implies that any two polymer samples at the same temperature, pressure, and composition will display similar thermodynamic behaviour regardless of their histories, provided that their densities are the same. Thus, any thermodynamic potential, for instance, the Helmholtz free energy density (a) can be expressed as:

$$a^{NE} = a^{NE}(T, P, \Omega_i, \rho_{pol}) \tag{4}$$

In this theory, the polymer density can be deemed as an internal state variable for the system, and it can be shown that the non-equilibrium thermodynamic relations that are endowed with such variables are independent of the pressure [65]:

$$\left(\frac{\partial a^{NE}}{\partial P}\right)_{T,\Omega_i,\rho_{pol}} = 0 \tag{5}$$

As a consequence of Eq. (5), a relationship between the non-equilibrium state of the glassy polymer phase and the equilibrium state can be determined:

$$a^{NE}(T, P, \Omega_i, \rho_{pol}) = a^{EQ}(T, \Omega_i, \rho_{pol}) \tag{6}$$

In a similar manner, a relationship between the non-equilibrium chemical potential of penetrant i can be related to the equilibrium value at the same temperature, composition, and polymer density:

$$\mu_i^{NE} = \left(\frac{\partial a^{NE}}{\partial \rho_i}\right)_{T,\Omega_i,\rho_{j\neq i},\rho_{pol}} \Rightarrow \mu_i^{NE}(T, P, \Omega_i, \rho_{pol}) = \mu_i^{EQ}(T, \Omega_i, \rho_{pol}) \tag{7}$$

Hence, the composition of penetrant i in the glassy polymer, in pseudo-equilibrium with the external gas phase, can be calculated using an appropriate EoS by finding an analytical expression for the chemical potential, provided that the experimental value of the polymer density is known. For the NELF model, this expression is given by [66]:

$$\frac{\mu_i}{RT} = \ln(\tilde{\rho}\varphi_i) - \ln(1-\tilde{\rho})\left[r_i^0 + \frac{r_i - r_i^0}{\tilde{\rho}}\right] - r_i - \tilde{\rho}\frac{r_i^0 v_i^*}{RT} \times \left[P_i^* + \sum_j^{N_c} \varphi_j (P_j^* - \Delta P_{ij}^*)\right] + 1 \tag{8}$$

The full description of the variables in Eq. (8) is available in table S1. Embedded in the mixing rules (refer to table S1) is the binary interaction parameter (k_{ij}). This adjustable parameter corrects for the mer–mer energy interaction of the mixture. In the absence of dilation data, especially in the presence of sorbing agents or under high pressures, a simple linear correlation can be used to calculate the swollen polymer's density [67]:

$$\frac{1}{\rho_{pol}} = \frac{1}{\rho_{pol}^0}\left(1 + \sum_i^{N_c-1} k_{sw,i} P_i\right) \tag{9}$$

where ρ_{pol}^0 is the dry polymer density, P_i is the partial pressure of penetrant i, and $k_{sw,i}$ is the swelling coefficient of penetrant i. In this work, only pure gas sorption will be considered, thus, a single pair of a binary interaction parameter and a swelling coefficient is sufficient. The binary interaction parameter is obtained through best fit in the low pressure region of the gas solubility isotherm. In contrast, the swelling coefficient is obtained by fitting to the higher pressure region of the curve. Usually, as a first order approximation, both of these parameters are set to zero.

2.3. The Marrero and Gani group contribution method

In this work, the Marrero and Gani molecular fragmentation method is used to determine the structural groups present in the polymer database. This method is found to be well suited for this database since it is capable of distinguishing between the finer details of many aromatic and non-aromatic rings present in the chemical structure of the polymers investigated here. For the sake of brevity, the reader is referred to the original work for the fragmentation rules [63]. To summarise, the molecular structure of the repeating unit can be segmented into three types of groups: first, second, and third order groups (FOGs, SOGs, and TOGs). FOGs contain a large set of simple groups and must be used to describe all the polymers repeating units. In contrast, not all polymers have higher order groups, but for those that do, they could give a better representation of the proximity effects, fine structural differences, and isomeric differences. The groups found in the materials comprising our database are listed in Tables 3, 4, and 5 below. It should be noted that new groups that are not found in the list provided by Marrero and Gani have been created specifically to describe this polymer data set. For example, groups like Si-O, $Si(CH_3)_3$ etc. are used to describe silicon containing polymers. To determine the group contribution (GC) values, Marrero and Gani [63] have proposed the following equation:

$$X - X_0 = \sum_i N_i C_i + w \sum_j M_j D_j + z \sum_k O_k E_k \tag{10}$$

In the present work, X is the characteristic SL-EoS parameter (T^*, P^*, or ρ^*), X_0 is a universal SL-EoS parameter constant, i, j, and k are the indexes assigned to the various first, second, and third order groups respectively. N_i, M_j, and O_k represent the occurrences of each FOG, SOG and TOG in the database. Finally, C_i, D_j, and E_k corresponds to the GC values of the groups. The machine learning implementation of this method will be discussed in Section 3.3.

3. Results and discussion

3.1. Populating the polymer database

To build the database for the SL parameters of homopolymers we firstly collected values published in the literature on 37 polymers (refer to table S2 in the supplementary information). Most of the parameters in this list were obtained by best fitting PVT data above the glass transition to the SL EoS model predictions. In some cases, e.g. Matrimid or PIM-1, such data are not obtainable, as the polymer chemically degrades before reaching the glass transition point. In such cases, one can use indirect routes, i.e. performing molecular simulations of the PVT behaviour above the hypothetical glass transition temperature or, as in the case of some the parameters reported in table S2, best-fitting gas solubility data onto the NELF predicted values [29,67,68]. To extend the database, we fitted 65 new SL polymers parameters. We chose polymers for which PVT data were available in the literature above the glass transition temperature. The fitting procedure is based on the non-linear square fitting of the model's parameters, with Eq. (3) as the objective function. The numerical computations were carried out using MATLAB's *fmincon* solver. The algorithm chosen for this task is the interior-point method. The list of new parameters is reported in Table 1 below.

3.2. Fragmenting the monomers and splitting the database

The database of 102 polymers were fragmented into FOGs, SOGs, and TOGs as reported in Tables 2, 3, and 4 respectively. For further machine learning processing, the database was split into a training set and test set. 83 polymers were randomly assigned to the training set and 19 to the test set with the only condition that all 90 groups were present in the training set. Fig. 2 and Table 5 showcase the composition of the groups and the different classes of polymers found in both sets respectively.

Table 1
List of polymer SL characteristic parameters fitted in this work from PVT data.

Name	T* [K]	P* [MPa]	ρ^* [g/cm^3]	Data Ref.[a]
poly(l-lactic acid)	584.30	827.30	1.388	[69]
poly(4-chloro styrene)	801.80	418.80	1.277	[69]
nylon 6	817.70	682.30	1.119	[69]
poly(ethylene adipate)	657.30	643.80	1.267	[69]
poly(ethylene naphthenate)	784.10	854.90	1.409	[69]
nylon 9	808.80	507.60	1.031	[69]
nylon 11	792.70	495.70	1.009	[69]
nylon 12	778.40	476.20	0.996	[69]
nylon 66	788.70	756.60	1.138	[69]
nylon 67	842.60	580.10	1.091	[69]
nylon 68	833.90	584.10	1.073	[69]
nylon 69	824.40	556.20	1.061	[69]
nylon 610	793.90	600.90	1.065	[69]
nylon 612	809.50	474.40	1.028	[69]
nylon 13/13	764.20	484.60	1.007	[69]
poly(ethylene isophthalate)	739.90	687.70	1.404	[69]
poly(ethylene succinate)	675.20	702.00	1.370	[69]
poly(1-octene)	672.50	351.30	0.914	[69]
poly(hexamethylene terephthalamide)	848.20	626.50	1.224	[69]
polyethersulfone	920.70	609.70	1.431	[69]
poly(isobutyl methacrylate)	581.47	667.07	1.175	[69]
poly(n-propyl acrylate)	642.80	440.80	1.135	[69]
poly(n-propyl methacrylate)	656.86	417.35	1.143	[69]
poly(n-butyl acrylate)	650.50	417.80	1.101	[69]
poly(n-hexyl methacrylate)	619.90	448.30	1.078	[69]
poly(lauryl methacrylate)	635.90	437.25	0.999	[69]
nylon 4/6	867.35	631.97	1.154	[69]
nylon 7	820.89	603.57	1.083	[69]
polyvinylidene fluoride	615.00	680.00	1.810	[69]
poly(vinyl formal)	703.70	595.40	1.282	[69]
poly(vinyl butyral)	669.40	509.10	1.155	[69]
polyvinyl carbazole	832.79	517.07	1.271	[69]
poly(acrylic acid)	839.20	749.17	1.489	[69]
poly(methacrylic acid)	674.10	1096.20	1.443	[69]
poly(vinyl fluoride)	759.00	487.90	1.349	[69]
poly(butylene terephalate)	732.89	762.28	1.325	[69]
poly bisphenol-A Isophthalate	821.19	550.30	1.277	[69]
polyarylate	827.50	610.10	1.268	[69]
poly ether ether ketone	930.80	462.40	1.285	[69]
polyetherimide (Ultem 1000)	940.00	473.50	1.330	[69]
poly(trimethylene terephthalate)	811.80	527.80	1.287	[70]
phenoxy resin (polyhydroxyl ether)	769.84	618.17	1.230	[69]
poly(azomethine ether) (n = 4)[b]	826.37	681.48	1.214	[69]
poly(azomethine ether) (n = 7)[b]	761.70	663.30	1.204	[69]
poly(azomethine ether) (n = 8)[b]	754.70	708.07	1.209	[69]
poly(azomethine ether) (n = 9)[b]	753.45	632.75	1.174	[69]
poly(azomethine ether) (n = 10)[b]	892.47	445.86	1.157	[69]
poly(azomethine ether) (n = 11)[b]	732.70	626.90	1.168	[69]
polyisoprene	672.38	446.86	0.955	[69]
poly(1-hexene)	603.30	600.20	0.908	[71]
poly(1-heptene)	664.00	300.17	0.881	[71]
poly(1-nonene)	605.00	300.00	0.902	[71]
poly(1-undecene)	633.00	300.66	0.903	[71]
poly(1-tridecene)	639.48	300.13	0.905	[71]
poly(1-octadecene)	640.95	300.00	0.896	[71]
poly(tert-butyl acrylate)	627.34	368.93	1.058	[72]
Poly(dimethylsilylene methylene)	686.20	371.60	0.926	[73]
poly(methylphenylsilylene methylene)	734.00	430.80	1.113	[74]
polymethylphenylsiloxane	657.69	403.85	1.180	[75]
poly(methyl-p-tolyl siloxane)	582.56	385.88	1.337	[75]
poly hexafluoropropylene oxide	554.70	294.50	2.084	[69]
DP1,1[c]	780.43	558.69	1.253	[76]
DP1,2[d]	797.70	534.70	1.178	[76]
DP1,3[e]	777.10	534.90	1.202	[76]
poly(butylene succinate)	684.89	567.55	1.253	[77]

[a] Reference for the experimental data used for the fitting.
[b] n represents the number of spacer $-CH_2-$ groups in the main chain.
[c] Poly[oxy(2,2-dimethyl propane-1,3-diyl) carboxybisphenyl4,4′-dicarbonyl].
[d] Poly[oxy(2-methyl, 2-ethyl propane-1,3-diyl) carboxybisphenyl4,4′-dicarbonyl].
[e] Poly[oxy(2-methyl, 2-n-propyl propane-1,3-diyl) carboxybisphenyl4,4′-dicarbonyl].

Table 2
List of first-order groups (FOGs). For the definition of the groups refer to [63].

No.	Group	No.	Group
1	CH	27	CH_2Cl
2	CH_2	28	SiO[a]
3	CH_3	29	CF_2
4	C	30	aC–SO_2
5	CH=CH	31	Si[a]
6	aCH	32	C=C
7	aC–CH	33	aC–Cl
8	aC–CH_3	34	$CONHCH_2$
9	CH_3COO	35	aC fused w/arom. ring
10	CCOO	36	aC–CONH
11	CH (cyclic)	37	O (cyclic)
12	CH_2 (cyclic)	38	N (cyclic)
13	CH_2O	39	aC fused w/non-arom. subring
14	aC	40	OH
15	aC–O	41	COOH
16	CH_3O	42	CHF
17	aC–OOC	43	aC–CO
18	aC–C	44	CO (cyclic)
19	CF_3	45	C (cyclic)
20	aC–COO	46	aC–N=CH[a]
21	CHCOO	47	aC–C=C
22	CH_2COO	48	aC–CN
23	CHCN	49	aC–OH
24	CHCl	50	O
25	CH=C	51	CF
26	Cl	52	aC–Si[a]

[a] These are new groups that are not listed in Ref. [63].

Table 3
List of second-order groups (SOGs). For the definition of the groups refer to [63].

No.	Group	No.	Group
53	CH_2–CH_m=CH_n (m,n in 0..2)	63	CH_m=CH_n–Cl (m,n in 0..2)
54	$(CH_3)_2CH$	64	CH_3–CH_m=CH_n (m,n in 0..2)
55	AROMRINGs$^1s^2$	65	CH_{cyc}–CH_2
56	CH_3COOCH or CH_3COOC	66	CHOH
57	CH_{cyc}–OOC	67	CHCOOH or CCOOH
58	AROMRINGs$^1s^2s^3s^5$	68	AROMRINGs$^1s^2s^4$
59	AROMRINGs$^1s^4$	69	AROMRINGs$^1s^2s^4s^5$
60	COO–CH_n–CH_m–OOC (n,m in 1..2)	70	CH_{cyc}–CH_3
61	AROMRINGs$^1s^3$	71	$(CH_3)_3C$
62	OOC–CH_n–CH_m–COO (n,m in 1..2)	72	$Si(CH_3)_3$[a]

[a] These are new groups that are not listed in Ref. [63].

3.3. GC regression

In this work, a three-level regression is carried out. Firstly, the first order group contributions (C_i) are determined alongside the universal constant X_0, as explained in detail in the following paragraph. This is done while assigning 0 to the SOG and TOG coefficients w and z. Then, the values of C_i and X_0 are kept constant during the second level regression. Here, w is 'switched on' by assigning to it a value of unity, and the SOG contributions (D_j) are determined while keeping z equal to zero. Similarly, the third level regression is carried out by keeping C_i, D_j, and X_0 constant and switching on z, which leads to the determination of the third order group contribution (E_k) values. When dealing with small data sets, the presence of many groups can become problematic: the larger the number of groups, the more sparse the feature set becomes [78]. As a consequence, even a simple linear model can over-fit, when enough groups are present, and the model would not be able to generalise well to new polymers [79]. To ameliorate this, a minor adjustment to the Marrero and Gani regression procedure is introduced herein. The first order group term of Eq. (10) will be replaced with a surrogate model produced from machine learning (ML) algorithms, while the higher order groups regression will remain the same since their input arrays are not dense enough for ML.

3.3.1. Determination of first order groups

We relied on two ML algorithms (1) ridge regression (RR) and (2) support vector regression (SVR), that are considered well suited for small data sets, and compared their performance in the regression of T^*, P^*, and ρ^*. The RR method is a linear model modified with a regularisation term (α) introduced into the cost function to reduce the chances of over-fitting [80]:

$$\min_{\mathbf{C},X_0} \sum_{l}^{n} (X_l - \mathbf{C}^T\mathbf{N}_l - X_0)^2 + \alpha\|\mathbf{C}\|_2^2 \tag{11}$$

where $\mathbf{C}$ is a vector that holds the GC values of all FOGs for all the n polymers considered in the summation (i.e. C_i), and $\mathbf{N}_l$ is the vector containing the occurrences of each FOG in polymer l. The larger the magnitude of the positive scalar hyperparameter α, the lower the variance of the final model becomes. In other words, the sensitivity of the model towards the data points diminishes as the model gets more regularised. In the case of SVR, the formal representation of the optimisation problem of the FOGs becomes [79]:

$$\begin{aligned} \min_{\mathbf{C},X_0,\zeta_l,\hat{\zeta}_l} \quad & \frac{1}{2}\|\mathbf{C}\|_2^2 + c\sum_{l}^{n}(\zeta_l + \hat{\zeta}_l) \\ \text{s.t.} \quad & \mathbf{C}^T\mathbf{N}_l + X_0 - X_l \le \epsilon + \zeta_l \\ & X_l - \mathbf{C}^T\mathbf{N}_l - X_0 \le \epsilon + \hat{\zeta}_l \end{aligned} \tag{12}$$

where ζ_l, $\hat{\zeta}_l$ are slack variables, and c, ϵ are hyperparameters. The value ϵ specifies a region of space (or 'tube') in which no penalty is incurred when an instance (i.e. data point) falls under it, while c is a regularisation term which has the opposite effect of α, thus the lower the magnitude of c, the higher the regularisation effect and vice versa [81]. Another important feature of SVR is its support of kernel tricks. From expression 12, it can be seen that SVR is a linear model, but with the aid of kernel tricks, the model can also solve non-linear problems in a computationally efficient manner. A more rigorous explanation of the mathematics of kernel methods can be found elsewhere [79,81]. Thus, the tuning of these kernel tricks hyperparameters must also be taken into consideration when building the ML models. For the problem investigated herein, we found that the Radial Basis Function (RBF) kernel had produced the best results. The RBF kernel introduces a new regularisation hyperparameter, γ, which defines the influence of a single data point on the model i.e. the smaller the value of γ is, the larger the influence of the instances on the model. It is also important to note that both of these algorithms are sensitive to feature scaling (i.e. the algorithm performs better if the variable scales are the same), hence, the FOG structural array was standardised:

$$N'_{i,l} = \frac{N_{i,l} - \bar{N}_i}{\sigma_i} \tag{13}$$

where $N'_{i,l}$ is the new standardised occurrence number of group i in polymer l, $N_{i,l}$ is the original occurrence number of group i in polymer l, $\bar{N}_i$ is the average occurrence number of group i in the training set, and σ_i is the standard deviation of the occurrence number of group i.

3.3.2. Tuning the hyperparameters

In this work, grid search was used to tune the hyperparameters [82]. Here, the training set was split into five validation sets or 'folds', in addition to a 'primary' set that includes all FOG groups (see Fig. 3). Then, the model was trained on the primary set and four of these folds and tested on the last remaining one. This was repeated five times until all of the folds had been used as a scoring set (i.e. the set chosen for testing). The selection of the hyperparameter combination is based on the best average performance over the five evaluations. This technique is sometimes called the '5-fold cross validation grid search'. Fig. 3 below provides an illustration of the procedure.

The performance metric selected as the scoring function is the coefficient of determination (R^2):

$$R^2 = 1 - \frac{\sum(X_l - \hat{X}_l)^2}{\sum(X_l - \bar{X})^2} \tag{14}$$

Table 4
List of third-order groups (SOGs) For the definition of the groups refer to [63].

No.	Group	No.	Group
73	aC–CO–aC (diff. rings)	82	aC–O–aC (diff. rings)
74	aC–CH_m–aC (diff. rings)(m in 0..2)	83	aC–CO_{cyc} (fused ring)
75	aC–SO_n–aC (diff. rings)(n in 1..4)	84	aC–$NH_{n,cyc}$ (diff. rings) (n in 0..1)[a]
76	AROMFUSED [2]s[2]	85	aC–$CH_{n,cyc}$ (diff. rings) (n in 0..1)
77	CH_{cyc}–$(CH_m)_n$–CH_{cyc} (m > 0; n in 0..2)	86	aC–$CH_{n,cyc}$ (fused rings) (n in 0..1)
78	aC–aC(diff. rings)	87	aC–N=CH–aC (diff. rings)[a]
79	aC–$NH_{n,cyc}$(fused rings) (n in 0..1)	88	aC–O_{cyc} (fused rings)
80	COO–$(CH_n)_m$–OOC (m > 2; n in 0..2)	89	CH (multi-ring)
81	aC–O–CH_n–aC (diff. rings) (n in 0..2)	90	CH_{cyc}–CH_{cyc} (diff. rings)

[a] These are new groups that are not listed in Ref. [63].

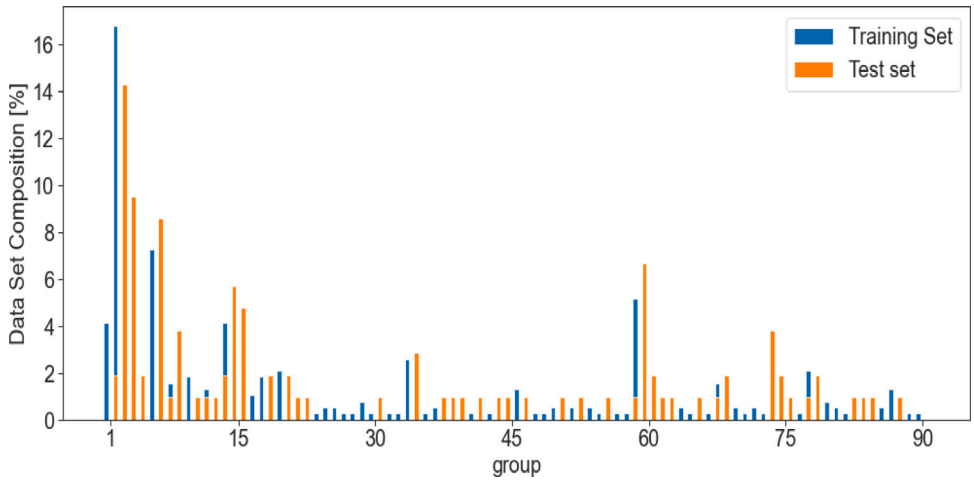


Fig. 2. The composition of the groups in the data set.

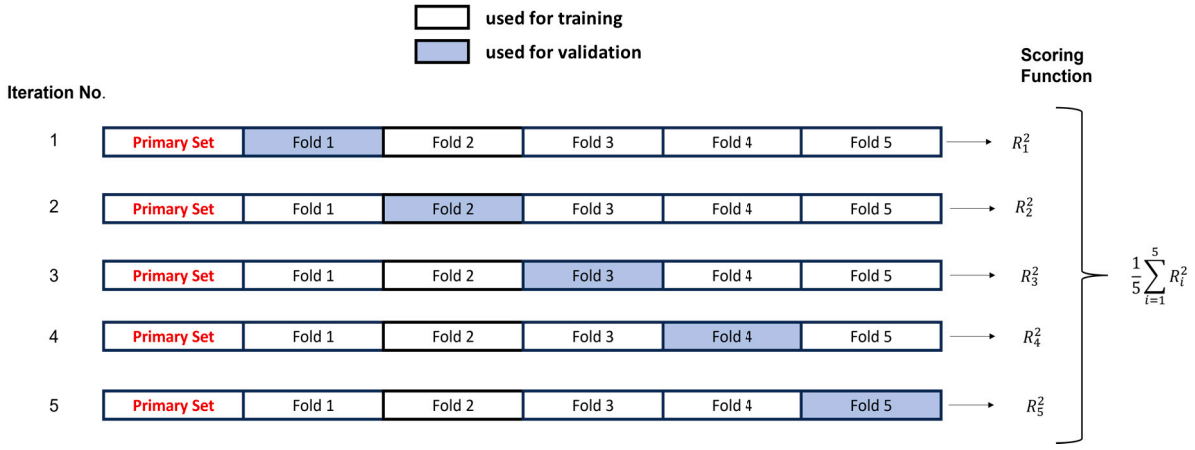


Fig. 3. A graphical representation of the grid search algorithm. For every combination of the hyperparameters values that are specified by the user, the model evaluates the scoring function five times. The primary set is present in each evaluation since it holds all groups. In each iteration, a different validation set is used (highlighted in blue). Then, the scoring function is averaged over the five iterations. The combination of hyperparameters that resulted in the best performance is selected as the optimum.

where X_l is the real value of the SL-parameter of polymer l, $\hat{X}_l$ is the predicted value of polymer l, and $\bar{X}$ is the mean value of the real SL parameters. All of the algorithms used for building the models were taken from the open source python library scikit-learn [82].

For each EoS parameter, the accuracy obtained using RR and SVR were compared in order to select the best model. The models hyperparameters were tuned by performing the 5-fold cross validation grid search described earlier on the training set (i.e. the 83 polymers set). Once the hyperparameters were tuned, the performance of the models was assessed using the test set. For T^* and ρ^*, RR had performed best, while for P^*, SVR predictions were better. A detailed discussion of the models performance can be found in Section 3.4. The final surrogate models' optimised hyperparameters are showcased in Table 6.

3.3.3. Adding higher order groups

Higher order groups were then added to the picture, which were linearly regressed as described earlier. The aim of such groups is to provide, to some extent, a better representation of the proximity effects and structural differences between the polymers. Some polymers in this database, for example, poly(ethylene isophthlate) and poly(ethylene terephthlate), are constitutional isomers of each other. Hence, FOGs alone would not be able to distinguish them apart. Higher order groups also 'absorb' some of the errors made by the previous groups. In

Table 5
The classification of the polymer samples in the data set.

Polymer class	Training set	Test set
Polyesters	12	3
Polyamides	11	3
Polyalkenes	8	1
Polymethacrylates	7	2
Polyacrylates	5	1
Si-Containing polymers	6	1
Polyazomethines	5	1
Polyglycols	3	1
Polycarbonates	4	0
Polystyrenes	4	1
Polyvinylethers	2	1
Polyimides	1	1
Polydienes/halodienes	3	0
Polysulfones	1	1
PIMs	1	0
Polyketones	0	1
Polyhydroxyethers	1	0
Polynitriles	1	0
Polyphenylethers	0	1
Polyvinylesters	1	0
Polyethers	1	0
Other Polyolefin/haloolefins	5	0
Polyvinylcarbazoles	1	0
Total data points	83	19

Table 6
The optimised hyperparameters for the surrogate models of FOGs.

Parameter	Machine learning algorithm	Hyperparameters
T^*	Ridge regression	$\alpha = 15$
P^*	Support vector regression	c = 321, $\gamma = 0.01$[a], $\epsilon = 0.1$
ρ^*	Ridge regression	$\alpha = 1.3$

[a] γ is a hyperparameter associated with the radial basis function (RBF) kernel.

practice, however, the error reduction observed in the training set is not universal, i.e. some of the polymers may perform worse. Moreover, the presence of higher order groups may worsen the predictions accuracy in the test set [83]. As such, the effects of higher order groups must be examined.

3.4. Model performance

3.4.1. Prediction of SL parameters

Table 7 displays the effects of the GC method based on FOG only, FOG and SOG, or FOG, SOG, and TOG on the estimation of the SL parameters. We considered two error indicators, namely the average absolute relative deviation (AARD%) and the root mean squared error (RMSE). As it can be observed, the effect of higher order groups is minuscule. However, in order to be able to distinguish between isomers, the higher order groups were kept in the final model. In the case of ρ^*, the test set AARD% (the primary metric considered for this work) slightly increases with the addition of TOGs, hence only FOGs and SOGs were included.

The parity charts of each of the parameter models are illustrated in Fig. 4. As it can be seen, the performance of all models, both on the training and test sets, are satisfactory. However, it is clear from Fig. 4(b) that P^* is the worse performing parameter, exhibiting an AARD of 5.99% and 10.82% in the training and test sets respectively. Initially, a RR model was built to predict P^*, however, the regularised linear model was not capable of capturing the structural–property trend very accurately, hence, a SVR model was built to introduce non-linearities into the function instead. The results of the SVR model, despite not being as accurate as the T^* and ρ^* models, is still acceptable.

3.4.2. Prediction of polymer density

The fitting of the SL characteristic parameters on PVT data is achieved through non-linear least-squares regression. The objective function of these optimisation problems is infested with local minima, and, as a consequence, global minima are not guaranteed [84]. Moreover, depending on the initial guesses provided to the algorithm, different local optima may be returned as a solution. As a consequence, no universal set of characteristic parameters can be obtained based on a particular set of experimental data. This is further exacerbated by the fact that, in some instances, the temperature and pressure ranges at which the fitting is carried over is different for the same polymer, resulting in entirely different sets of parameters that are reported in literature [85]. To truly test the model's performance, comparing the predicted SL parameters AARD alone would not suffice, thus, the ML-GC-SL EoS density predictions must be compared to the experimental values for a more comprehensive analysis of the model efficacy. Tables 8 and 9 show the predicted SL parameters and the performance of the models for the test set respectively. The average AARD% of the density predictions of the ML-GC-SL model is around 5.59%, with all density AARDs% being less than 15%. These results are reasonably good, considering that the estimations are entirely predictive, the data set is limited in size, and the test set is diverse in terms of chemical structure. At face value, over-represented polymer classes, like polyamides and polyazomethine ethers, have their parameters predicted with satisfactory accuracies. There are of course, exceptions to the rule. Poly(methacrylic acid) parameters, for example, are poorly predicted. This may be caused by the presence of the –COOH group, which is absent from the rest of the polymethacrylates in the data set. In the SL-EoS, T^* represents the average mer–mer interactions of the polymer, P^* is a proxy for the cohesive energy density at closed packing, and ρ^* is the density of the polymer at closed packing [6,31]. Therefore, as a polar substance, poly(methacrylic acid) will exhibit higher values of these parameters compared to the other polymethacrylates. This trend is also observed in the training set, where the parameter values of poly(acrylic acid) is higher than the rest of the members of its class. In the training set, the group –COOH only occurs once, and its group contribution value comes from poly(acrylic acid). This may have contributed to the poor predictions of poly(methacrylic acid), since their experimentally fitted values for T^* and P^* are vastly different. On the other hand, the error in the ρ^* prediction is much less profound due to the similar experimentally fitted values of this parameter for both polymers. In contrast, poly(n-propyl methacrylate) predictions were more accurate, owing to its structural similarity to the rest of the polymethacrylates training instances. We note that in the presence of associating groups, models like the non-random hydrogen bonding (NRHB) model [86–88] and the PC-SAFT EoS [10,89–91] are better suited for describing such systems because they explicitly consider these interactions in their formulation.

Despite having 8 other polymers of the same class in the training set, poly(1-heptene) density error is still higher than average. The reason for this may be that polyalkenes only contain alkyl groups (i.e. CH_2, CH_3 etc.) and no other unique groups that distinguish them from the rest of the other instances. The alkyl groups are present in many of the polymers in the data set, hence, their group contribution values also incorporate the effects of their presence in other polymer classes, like polyesters and polyglycols, which may lead to lower accuracies when it comes to predicting the parameters of polyalkenes. This is also apparent in the training set, where some of the highest errors belong to polyalkenes.

Polymers like poly(o-methylstyrene), poly(ethylene terephtalate), DP1,2 and polysulfone had their PVT properties predicted with relatively high accuracies. This may be due to the presence of analogous polymers with similar chemical structures and SL parameters values in the training set. For example, the SL parameters of polysulfone were extrapolated from the same structural groups as polyethersulfone, which is found in the training set. The only difference however, is that

Table 7
The effect of the group order on the average absolute relative deviation (AARD%) and the root mean square error (RMSE).

Group	T^* [K]		P^* [MPa]		ρ^* [g/cm^3]	
	Training set	Test set	Training set	Test set	Training set	Test set
AARD [%]						
FOG	3.74	5.45	5.86	11.53	1.97	4.92
SOG	3.34	5.28	5.96	10.83	1.88	4.89
TOG	3.26	5.19	5.99	10.82	1.80	4.98
RMSE						
FOG	34.82	56.92	75.17	121.64	3.42E−2	7.96E−2
SOG	32.95	58.38	70.98	113.03	3.30E−2	7.96E−2
TOG	32.81	57.89	70.59	112.90	3.20E−2	8.04E−2

The AARD The RMSE is defined as $\sqrt{\frac{1}{n}\sum_i^n (X_i - \hat{X}_i)^2}$.

Table 8
The ML-GC-SL characteristic parameters predictions for the test set.

Polymer	T^* [K]	P^* [MPa]	ρ^*[g/cm^3]
poly (o-methylstyrene)	749.76	357.23	1.0961
poly(2,6-dimethyl-1,4-phenylene oxide)	724.11	515.00	1.1454
poly(tetrahydrofuran)	634.79	482.27	1.1588
poly (ethylene terephtalate)	756.41	679.67	1.3623
nylon 6	768.28	521.69	1.1089
nylon 612	808.98	556.38	1.0262
nylon 13/13	796.13	519.38	0.9260
poly(ethylene succinate)	678.76	649.53	1.3011
poly(n-propyl acrylate)	642.38	454.15	1.1941
poly(n-propyl methacrylate)	652.85	459.42	1.1720
poly(vinyl butyral)	645.12	591.61	1.3439
poly(metharcylic acid)	840.46	678.81	1.4741
poly ether ether ketone	819.15	556.54	1.2506
polyetherimide (Ultem 1000)	939.70	473.91	1.4487
poly(azomethine ether (n = 9)	791.98	619.75	1.1777
poly(1-heptene)	680.21	342.57	0.9463
poly(methyl-p-tolyl siloxane)	666.23	407.48	1.1599
DP1,2	778.77	546.85	1.2026
polysulfone	895.60	591.29	1.2634

polysulfone substitutes one of the aC-SO_2 with an aC-C and two CH_3 groups in polyethersulfone. Since the difference between the experimentally fitted T^* of these two polymers is the largest, its error was also anticipated to be higher than the other two parameters. Similarly, the experimentally fitted P^* values were very close, making it the most accurate parameter predicted for this polymer.

The model is also capable of predicting the SL parameters for polymers of families that did not participate in the training process reasonably well. For example, poly(2,6-dimethyl-1,4-phenylene oxide), or PPO, had its parameters predicted from groups found in polycarbonates, polyesters etc., and the results were successful. Figs. 5 to 8 showcase some of the PVT predictions made by the ML-GC-SL in comparison to the experimental values for some of the polymers in the test set.

3.5. Extending the model predictions beyond the database: estimating the SL parameters of polyimides

3.5.1. Predicting the SL parameters for new high T_g polyimides

The validation of the ML-GC models built in the previous sections was found satisfactory, however, in order to obtain the ML-GC-NELF model version to use for production, the surrogate models were re-trained (using the same hyperparameters) with the entirety of the data set (i.e. training and test sets), to improve the models accuracy. Table 10 showcases the new AARD and RMSE values of the data set. The new values of the SL parameters for the entire data-set can be found in the supplementary information section (Table S3). Then, the SL parameters of six new high T_g were determined. This set of new polymers was not part of the 102 polymer data set that was used to train and test the models in the previous sections. This was done to ensure that we do not introduce any bias in the models during the optimisation process of the hyperparameters. In other words, we wanted to build ML models that were fully capable of predicting the SL parameters of high T_g polymers, but to preserve the integrity of the GCM, the models were first tuned and tested on a randomly selected polymer set and then used to evaluate the SL parameters of materials that had never been exposed to the models before. The list of the new high T_g polymers, along side their predicted SL parameter values, parameter and density AARD can be found in Table 11, and the experimentally fitted SL parameters are reported in Table S2.

For two of these polymers, PIM-PI-EA and PIM-PI-SBI, no SL characteristic parameters are available in literature, perhaps due to the lack of sufficient gas sorption data, and the values reported here are *a priori* predictions of the ML-GC-NELF approach based on their monomer structure. For the other polyimides with experimentally fitted SL parameters, the absolute relative errors of P^* and ρ^* are reasonably good, especially when compared to the data set's AARD. The AARD of T^* on the other hand, is relatively higher. However, as mentioned earlier, comparing the AARD of the parameters alone can be misleading, thus, the results must also be validated against experimental data. Since we are dealing with high T_g polymers, not all of the polyimides listed below have experimental PVT data. Hence, in the upcoming sections, the results will be also validated against solubility data.

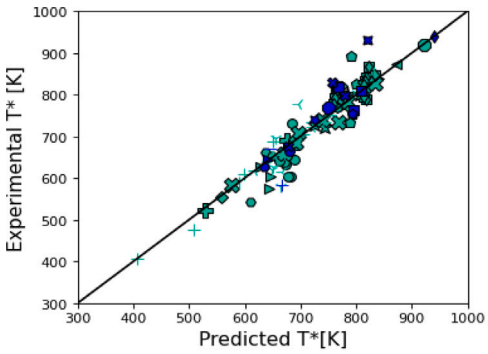
3.5.2. Verifying the ML-GC-NELF predictions against PVT data: 6FDA-6FpDA and 6FDA-ODA

Scherillo et al. [88] have reported the SL-parameters (found in Table S2) of 6FDA-6FpDA and 6FDA-ODA by fitting them to the experimental PVT properties. Since the authors did not publish the experimental data, the ML-GC-SL predictions were compared against the predictions made by the SL-EoS, using the experimentally fitted parameters above T_g. The resultant density AARD's for 6FDA-6FpDA and 6FDA-ODA are both satisfactory (5.79% and 1.00% respectively) as shown in Table 11. It is also important to note that, despite having comparable parameter errors, the ML-GC-SL has produced a more accurate prediction of the PVT data in the case of 6FDA-ODA than 6FDA-6FpDA. Hence, it is always advisable to validate the SL parameter predictions against experimental values. Fig. 8 showcases the PVT predictions made by both models for 6FDA-ODA.

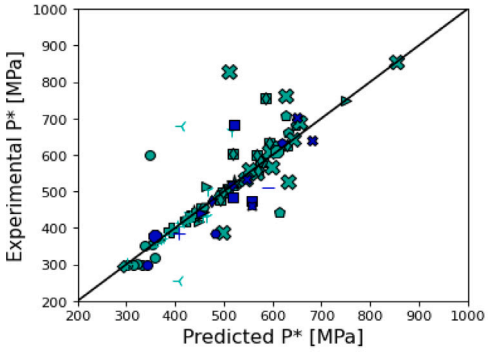
3.5.3. Verifying the ML-GC-NELF predictions against solubility data: the infinite dilution solubility coefficient

As mentioned earlier, comparing the error to the experimentally fitted SL parameters alone is not sufficient in assessing the model's efficacy. However, most of the high T_g polyimides have no density data in the rubbery region available in literature. To this end, the infinite dilution solubility coefficient S_0 of pure gases in the polymer was used to validate the model performance. At the limit of zero pressure, the expression for the solubility coefficient using the NELF model is given by [29,96]:

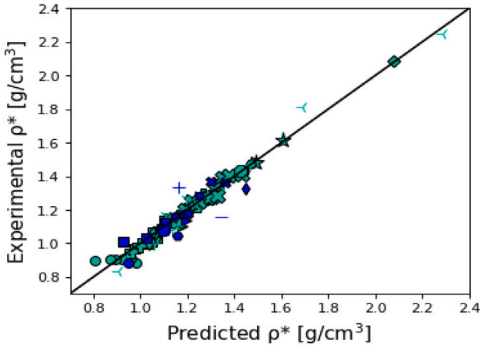
$$\ln(S_0) = \ln\left(\frac{T_{stp}}{P_{stp}T}\right) + r_1^0\left\{\left[1 + \left(\frac{v_1^*}{v_2^*} - 1\right)\frac{\rho_2^*}{\rho_2^0}\right]\right.$$



(a) The performance of the T* model using all structural groups.



(b) The performance of the *P** model using all structural groups.



(c) The performance of the *ρ** model using groups up to second-order.

Fig. 4. The parity charts of the final models performances of the parameters T^{*}*, P^{*}, and ρ^{*}. The experimental values refer to the parameters fitted to experimental data, and the predicted values are the values generated by the GC model.

$$\times \ln\left(1-\frac{\rho_2^0}{\rho_2^*}\right)+\left(\frac{v_1^*}{v_2^*}-1\right)+\frac{\rho_2^0}{\rho_2^*}\frac{T_1^*}{T}\frac{2}{P_1^*}(1-k_{12})\sqrt{P_1^*P_2^*}\Bigg\} \tag{15}$$

where the subscripts STP, 1, and 2 represent standard temperature and pressure, the penetrant, and the polymer respectively, for more details about the variables please refer to Table S1 (supplementary information). In addition to the penetrant and the polymer SL parameters, Eq. (15) also requires the dry polymer density (ρ_2^0) as an input to carry out the calculations. Tables S4 and Table 11 list the penetrant SL parameters and the density values of the polymers respectively. The experimentally fitted SL parameters of the listed polyimides with no PVT data were obtained from gas sorption data using Eq. (15), usually by setting the binary interaction parameter to $k_{12} = 0$ as a first order approximation [29,31]. At diminishing pressure, this assumption

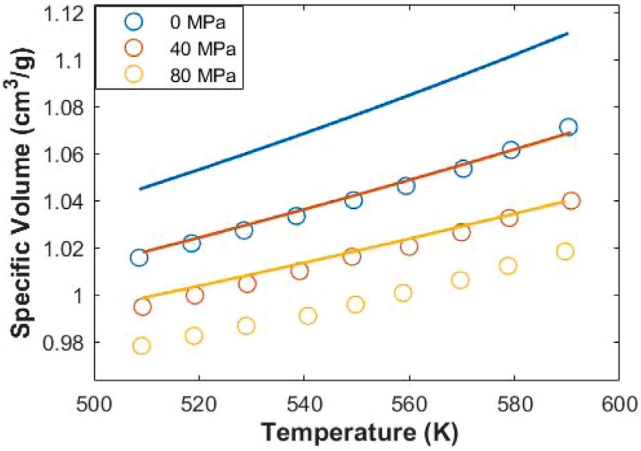


Fig. 5. PVT data of Nylon 6. The continuous curves are the ML-GC-SL predictions and the discrete points are the experimental data. *Source:* The experimental data are taken from Ref. [69].

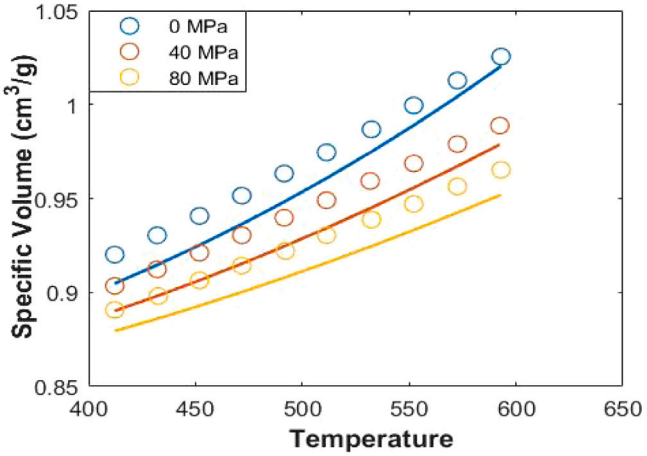


Fig. 6. PVT data of DP1,2. The continuous curves are the ML-GC-SL predictions and the discrete points are the experimental data. *Source:* The experimental data are taken from Ref. [76].

Table 9
AARD% of the ML-GC-SL characteristic parameters and the predicted PVT data of the test set.

Polymer	T* [%]	P* [%]	ρ^* [%]	Density AARD	*Ref.*[a]
poly(o-methylstyrene)	2.38	5.49	1.58	0.83	[92]
poly(2,6-dimethyl-1,4-phenylene oxide)	2.02	0.39	1.22	3.12	[69]
poly(tetrahydrofuran)	1.40	24.94	11.10	11.06	[56]
poly(ethylene terephtalate)	8.87	6.03	0.10	2.99	[69]
Nylon 6	6.04	23.54	0.93	2.33	[69]
Nylon 6/12	0.06	17.28	0.17	0.55	[69]
Nylon 13/13	4.18	7.18	8.02	6.97	[69]
poly(ethylene succinate)	0.53	7.47	5.01	4.75	[69]
poly(n-propyl acrylate)	0.07	3.03	5.21	5.15	[69]
poly(n-propyl methacrylate)	0.61	10.08	2.56	2.27	[69]
poly(vinyl butyral)	3.63	16.21	16.38	14.78	[69]
poly(methacrylic acid)	24.68	38.08	2.17	10.39	[69]
poly ether ether ketone	12.00	20.36	2.44	7.51	[69]
polyetherimide (Ultem 1000)	0.03	0.09	9.37	8.85	[69]
poly(azomethine ether (n = 9)	5.11	2.05	0.31	1.70	[93]
poly(1-heptene)	2.44	14.13	7.46	7.95	[71]
poly(methyl-p-tolyl siloxane)	14.36	5.60	13.26	11.24	[75]
DP1,2	2.37	2.27	4.25	1.43	[76]
polysulfone	7.90	1.45	2.99	2.27	[69]
Average	5.19	10.82	4.98	5.59	

[a] References for the experimental PVT data.

is reasonable to make when dealing with light gases and n-alkanes vapours. Another reason why this approximation is made is to avoid having to fit k_{12} for every gas simultaneously with the pure parameters, and since every single data-point represents a single gas species, doing so is inadvisable if one would like obtain a reliable and generalisable set of the pure polymer parameters.

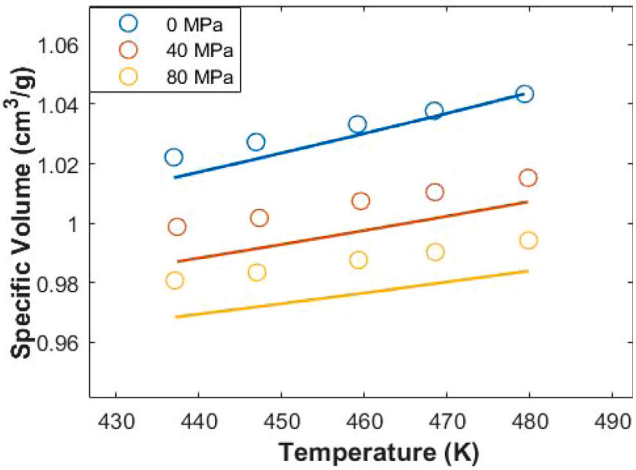


Fig. 7. PVT data of poly(o-methylstyrene). The continuous curves are the ML-GC-SL predictions and the discrete points are the experimental data. *Source:* The experimental data are taken from Ref. [92].

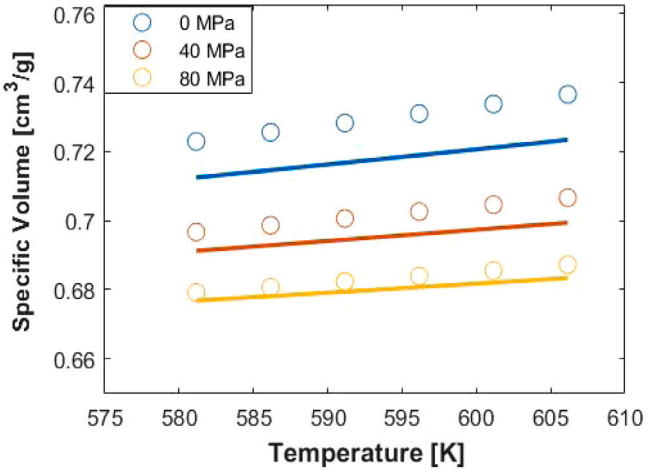


Fig. 8. A comparison of the ML-GC-SL density predictions (continuous curves) against the experimentally-fitted SL predictions (discrete points) for 6FDA-ODA. *Source:* The experimentally fitted SL parameters were taken from Ref. [88].

Table 10
The re-trained models error indicators. These models were used to predict the SL parameters of the polyimides that are not present in the data-set.

Parameter	AARD [%]	RMSE[a]
T^*	3.07	31.75
P^*	6.22	71.65
ρ^*	2.27	3.8E−2

[a] The RMSE units of T*, P* and ρ^* are in K, MPa and g/cm^3 respectively.

In this work, the predicted infinite dilution solubility coefficients were calculated using this first order approximation through the ML-GC-NELF model and the results were compared against the experimental value, which is obtained by the extrapolation to zero pressure of the experimental data fitting given by the DMS relationship [97]:

$$S_0 = k_d + C'_H b \tag{16}$$

where k_d is Henry's constant, C'_H is Langmuir capacity, and b is the hole affinity. These parameters are polymer-penetrant specific and are fitted to gas sorption data. The DMS parameters values can be found in Tables S5 to S8. Finally, for benchmark purposes, the original NELF solubility coefficient at infinite dilution calculated with the first-order approximation assumption, will also be compared to both the ML-GC-NELF predictions and the experimental values. The difference between the two models lies in the way the SL parameters were estimated for the polymer, namely using experimental data in the case of NELF, and the molecular monomeric structure in the case of the ML-GC-NELF model.

The ML-GC-NELF and the NELF infinite dilution solubility coefficient predictions for various gases are displayed in Fig. 9. It has to be noticed, as explained above, that since the SL parameters are available only for two of the four polymers examined in this section, the NELF model calculations can be run only on this subset and the comparison between the two versions of the model shall be confined to these materials. Starting with the PIMs, the accuracy of the S_0 predictions ranges from excellent to satisfactory, as it can be seen from the RMSE values reported in Table 12. Fig. 9 indicates that the S_0 values for PIM-PI-EA are good for all the gases tested. In the case of PIM-PI-SBI, the error seems to be more prominent in N_2, with the other gases predictions being acceptable. In the upcoming Section 3.5.4, it will be shown that for those two PIMs, only a minor adjustment for the gases k_{ij}s would be required to fit the solubility isotherms, which is indicative of the good accuracy of the estimated SL parameters.

It can also be observed from Table 12 that the RMSE of 6FDA-DAM with regards to S_0 is higher than the rest of the other polymers, due to the poor estimation of the solubility coefficients of the alkane vapours by both models. However, for the results of the other gases tested herein, the accuracy of the models are comparable.

Table 11
The predicted SL parameters, parameter AARD, density AARD, and the dry polymer densities of the polyimides selected. The density AARD are based on the experimentally-fitted SL parameters.

Polymer	T* [K]		P* [MPa]		ρ^* [g/cm^3]		Density AARD[a]	ρ^0_{pol}
	Prediction	Error	Prediction	Error	Prediction	Error		
PIM-PI-SBI	651.71	–	494.70	–	1.4661	–	–	1.12 [94]
6FDA-6FpDA	817.31	8.96%	509.07	6.84%	1.8521	2.55%	5.79%[b]	1.58 [88]
HAB-6FDA	900.77	25.11%	499.01	3.72%	1.6390	1.87%	–	1.41 [31]
6FDA-ODA	887.23	10.32%	493.05	6.41%	1.6202	2.28%	1.00%[c]	1.49 [88]
6FDA-DAM	824.52	7.78%	496.65	3.47%	1.5135	8.83%	–	1.33 [95]
PIM-PI-EA	902.34	–	521.89	–	1.4669	–	–	1.12 [94]
Average	–	**13.04%**	–	**5.11%**	–	**3.88%**	**3.40%**	–

[a] Data taken from Ref. [88]. The AARD was calculated based on the predictions of the experimentally fitted SL parameters.
[b] Calculation range : 593–626 K and 0–80 MPa.
[c] Calculation range : 581–606 K and 0–80 MPa.

Table 12
RMSE[a] of the S_0 calculations based on the first-order approximation ($k_{12} = 0$) for the ML-GC-NELF and the NELF models.

Polymer	NELF	ML-GC-NELF
6FDA-DAM	305.85	161.3
PIM-PI-EA	–	2.73
HAB-6FDA	2.40	7.94
PIM-PI-SBI	–	13.17

[a] cm^3(STP)/cm^3(polymer)atm.

In addition, it is also worth noting that the quality of fit of the alkane vapour values for 6FDA-DAM is superior in the case of the ML-GC-NELF model, which explains the lower RMSE value in comparison to the NELF model. Finally, for HAB-6FDA, the ML-GC-NELF model seems to over predict S_0 in all of the three gases investigated. This may be attributed to the relatively higher error made in estimating T* in comparison to P* and ρ^* (see Table 11). Regardless, the ML-GC-NELF S_0 predictions all seem to be very satisfactory for the polyimides tested, especially considering that there were no chemically analogous molecules in the terms of chemical structure found in the data-set (i.e. no 6FDA-based polymers exist in the data-set, and PIM-1 differs in chemical structure from PIM-PI-EA and PIM-PI-SBI).

3.5.4. Modelling of the pure gas solubility isotherms in high T_g polymers using the ML-GC-NELF model

As mentioned above, the ML-GC model allowed us to obtain the SL parameters for two polymers that were previously unknown in literature, PIM-PI-SBI and PIM-PI-EA. The predicted SL parameters were used to model the light gas sorption for these polymers. Figs. 10 and 11 show the resultant experimental gas solubility isotherms and the ML-GC-NELF calculations. For these calculations we could refine the k_{ij} as customary. In addition, the swelling coefficient k_{sw} was also fitted to the sorption data at the high pressure region of the isotherms. From Table 13, it can be seen that the magnitudes of the adjusted $k_{ij}s$ are within 0.1 for all the gas species and the polymers involved. This minor adjustment to the mixing rule is another indication of the accuracy of the predicted SL parameters. As it can be observed in Fig. 9, the S_0 test conducted on PIM-PI-SBI SL parameters shows that the quality of fit with respect to CH_4 and N_2 is worse in comparison to the other gases, which explains the relatively higher values of k_{ij} required to fit their isotherms.

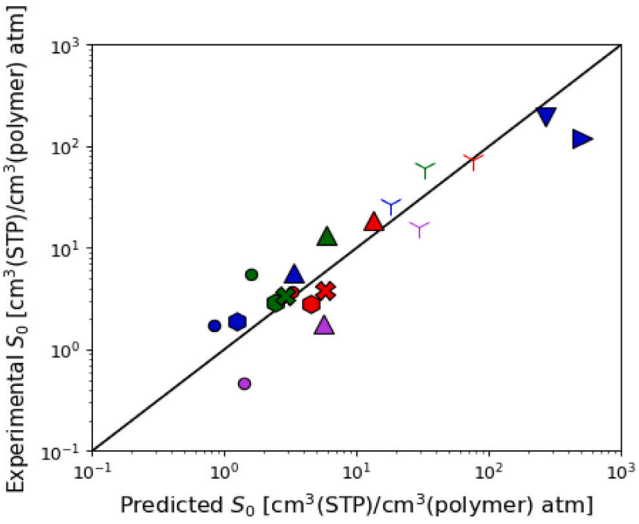
This also explains the relatively higher k_{ij} for Ar and O_2 in PIM-PI-EA. As one would expect, the swelling coefficient is the highest in CO_2 for all polymers, as it is the strongest sorbing agent of the listed gases. This is followed by CH_4, which seems to exhibit the second highest swelling coefficient. This behaviour can be confirmed visually from the experimental data in Figs. 10 and 11, where a downward concavity is observed in both these gas species (and is most prominent in CO_2). For the other light gases (N_2, O_2, Ar), the solubility isotherms are, in effect, linear, with minimal to no swelling that can be observed based on the k_{sw} values.

4. Conclusion

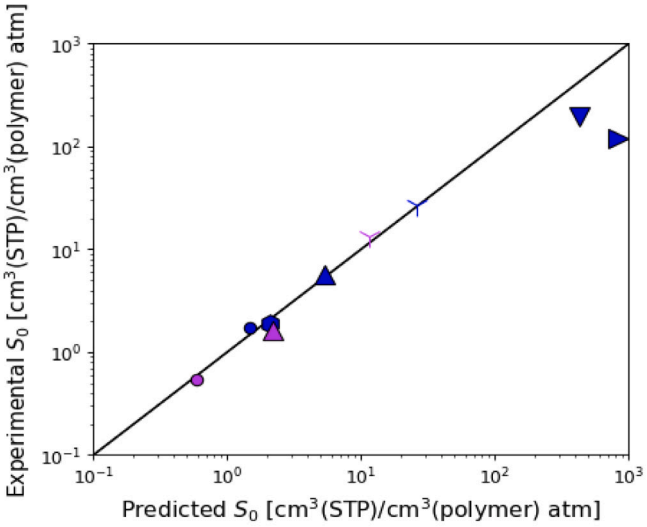
The NELF model, developed by Doghieri and Sarti [23], is a valuable tool for the systematic characterisation of gas sorption in glassy polymeric membranes, and remains to be one of the most powerful tools to date for such purposes. But like its equilibrium counterpart, the NELF model predictive capabilities are constrained by the availability of pure and mixture parameters, specifically for polymers. In the past, the issue of predicting pure polymer Sanchez–Lacombe parameters was addressed through group contribution methods. However, the efficacy of these models was not tested rigorously due to the limited size of the data sets. In this work, a data set of 102 polymers was curated to build a group contribution model for the SL parameters. Roughly 80% of these polymers were used for training, while the remaining 20% were used for evaluating the model's performance. The group contribution method was based on the Marrero and Gani's approach [63], with a minor adjustment to the multi-step regression procedure, which involves replacing the first-order term with a surrogate machine learning model.

The resultant regularised machine learning models were able to reproduce the SL parameters with satisfactory agreement to the experimentally fitted values. Coupled with the Sanchez–Lacombe equation of state, the group contribution model was also validated against experimental pressure–volume–temperature properties, leading to an average density AARD of 5.59% for the test set. Using the same hyperparameters, the models were retrained on the entire data set to determine the SL parameters of high T_g polymeric membranes, consisting of 6FDA-based polyimides and PIMs. Many of these non-conventional membranes lack experimental PVT data, hence another experimental metric, known as the infinite dilution solubility coefficient S_0 was also used to validate the models. Through the first-order approximation (i.e. $k_{ij} = 0$), the predicted ML-GC-NELF S_0 values were found to be comparable to the original NELF predictions (which is reliant on the experimentally fitted polymer SL parameters) and the experimental S_0. From this list of polymers, PIM-PI-SBI and PIM-PI-EA had their light gas sorption isotherms calculated through the ML-GC-NELF model. For these PIM polymers, the ML-GC-NELF was able to capture the trend of the experimental gas solubility isotherms successfully with minor adjustment to k_{ij}, which is indicative of the accuracy of the predicted SL parameters.

As a proof-of-concept, the results of this ML-GC model seem promising, and just like any data-centric model, the proposed model has the potential to be improved as more data on polymers are collected. Important classes of polymers, such as thermally rearranged (TR) polymers and bio-polymers, to name a few, were missing in this data set. These polymers, and many more, will be added to build better GC models in the foreseeable future. Moreover, the same principles used in this GC method, can also be exploited to estimate the parameters of other EoS (i.e. PC-SAFT), and that will be the focus of our future work.



(a) The performance of the ML-GC-NELF model in predicting the solubility coefficient at infinite dilution.



(b) The performance of the NELF model in predicting the solubility coefficient at infinite dilution.

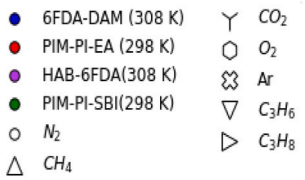


Fig. 9. The parity charts of the solubility coefficients at infinite dilution. The experimental values were obtained by the dual mode sorption model and the predicted values were obtained through the first-order approximations made by the ML-GC-NELF and the NELF models.

Table 13
The binary interaction parameters and the swelling coefficients for each of the polymer-penetrant pairs.

Polymer	CO_2		CH_4		N_2		O_2		Ar	
	k_{ij}	k_{sw}	k_{ij}	k_{sw}	k_{ij}	k_{sw}	k_{ij}	k_{sw}	k_{ij}	k_{sw}
PIM-PI-EA	0.023	0.032	0.020	0.021	−0.032	0	0.065	0	0.080	0
PIM-PI-SBI	−0.055	0.047	−0.0844	0.020	−0.100	0	−0.02	0	−0.055	0

k_{sw} units in MPa^{-1}.

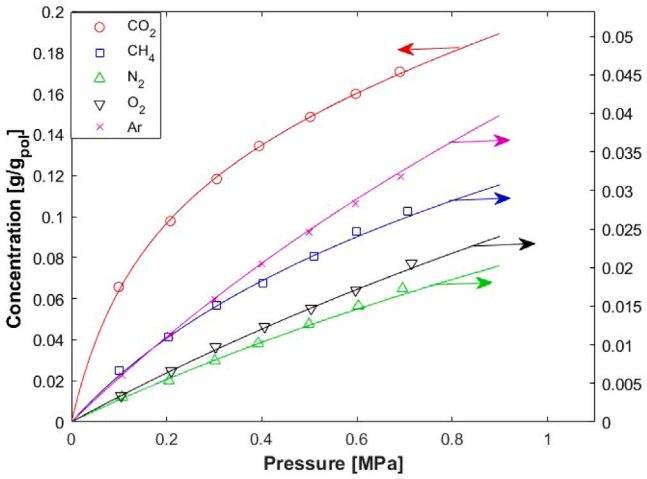


Fig. 10. The gas solubility isotherms in PIM-PI-SBI (25 °C). The discrete points are the experimental values, and the continuous curves are the ML-GC-NELF calculations. *Source:* Experimental data are taken from Ref. [94].

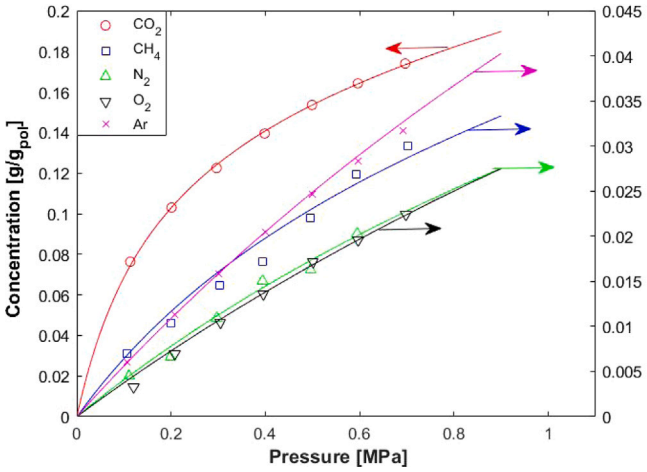


Fig. 11. The gas solubility isotherms in PIM-PI-EA (25 °C). The discrete points are the experimental values, and the continuous curves are the ML-GC-NELF calculations. *Source:* Experimental data are taken from Ref. [94].

CRediT authorship contribution statement

Hasan Ismaeel: Conceptualization, Methodology, Software, Data curation, Writing – original draft. **David Gibson:** Software, Data curation, Data analysis. **Eleonora Ricci:** Conceptualization, Methodology, Data curation, Writing, Editing. **Maria Grazia De Angelis:** Conceptualization of this study, Methodology, Writing, Editing.

Declaration of competing interest

The authors declare no conflict of interest.

Data availability

Data will be made available on request.

Acknowledgement

We are grateful for the Royal Society of Edinburgh (RSE) for financially supporting this work under the "A Machine Learning-Aided Modelling Platform for the design of Hydrogen-Ready materials" grant award 2915.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.memsci.2023.122220. These include tables that provide the details of the SL-EoS, the predicted and fitted parameters used in this work.

References

[1] H. Lin, Y. Ding, Polymeric membranes: chemistry, physics, and applications, J. Polym. Sci. 58 (18) (2020) 2433–2434.
[2] B.D. Freeman, Basis of permeability/selectivity tradeoff relations in polymeric gas separation membranes, Macromolecules 32 (2) (1999) 375–380.
[3] L.M. Robeson, The upper bound revisited, J. Membr. Sci. 320 (1–2) (2008) 390–400.
[4] J. Wijmans, R. Baker, The solution-diffusion model: a review, J. Membr. Sci. 107 (1–2) (1995) 1–21.
[5] H. Lin, B.D. Freeman, Materials selection guidelines for membranes that remove CO2 from gas mixtures, J. Mol. Struct. 739 (1–3) (2005) 57–74.
[6] I.C. Sanchez, R.H. Lacombe, An elementary molecular theory of classical fluids. Pure fluids, J. Phys. Chem. 80 (21) (1976) 2352–2362.
[7] R.H. Lacombe, I.C. Sanchez, Statistical thermodynamics of fluid mixtures, J. Phys. Chem. 80 (23) (1976) 2568–2580.
[8] I.C. Sanchez, R.H. Lacombe, Statistical thermodynamics of polymer solutions, Macromolecules 11 (6) (1978) 1145–1156.
[9] B.C. Lee, R.P. Danner, Prediction of polymer-solvent phase equilibria by a modified group-contribution EOS, AIChE J. 42 (3) (1996) 837–849.

[10] J. Gross, G. Sadowski, Perturbed-chain SAFT: An equation of state based on a perturbation theory for chain molecules, Ind. Eng. Chem. Res. 40 (4) (2001) 1244–1260.

[11] N. Von Solms, M.L. Michelsen, G.M. Kontogeorgis, Computational and physical performance of a modified PC-SAFT equation of state for highly asymmetric and associating mixtures, Ind. Eng. Chem. Res. 42 (5) (2003) 1098–1105.

[12] Y. Song, S.M. Lambert, J.M. Prausnitz, A perturbed hard-sphere-chain equation of state for normal fluids and polymers, Ind. Eng. Chem. Res. 33 (4) (1994) 1047–1057.

[13] I.C. Sanchez, P.A. Rodgers, Solubility of gases in polymers, Pure Appl. Chem. 62 (11) (1990) 2107–2114.

[14] M. Hamedi, V. Muralidharan, B. Lee, R. Danner, Prediction of carbon dioxide solubility in polymers based on a group-contribution equation of state, Fluid Phase Equilibria 204 (1) (2003) 41–53.

[15] F. Sabzi, M.R. Talaghat, A. Hosseini, Prediction of water vapor sorption in the polymeric membranes using PHSC equation of state, J. Nat. Gas Sci. Eng. 21 (2014) 757–763.

[16] N. Von Solms, M.L. Michelsen, G.M. Kontogeorgis, Prediction and correlation of high-pressure gas solubility in polymers with simplified PC-SAFT, Ind. Eng. Chem. Res. 44 (9) (2005) 3330–3335.

[17] L.M. Robeson, Q. Liu, B.D. Freeman, D.R. Paul, Comparison of transport properties of rubbery and glassy polymers and the relevance to the upper bound relationship, J. Membr. Sci. 476 (2015) 421–431.

[18] R.M. Barrer, J.A. Barrie, J. Slater, Sorption and diffusion in ethyl cellulose. Part III. Comparison between ethyl cellulose and rubber, J. Polym. Sci. 27 (115) (1958) 177–197.

[19] A.S. Michaels, W.R. Vieth, J.A. Barrie, Solution of gases in polyethylene terephthalate, J. Appl. Phys. 34 (1) (1963) 1–12.

[20] V.I. Bondar, Y. Kamiya, Y.P. Yampol'skii, On pressure dependence of the parameters of the dual-mode sorption model, J. Polym. Sci. B 34 (2) (1996) 369–378.

[21] M.G. De Angelis, G.C. Sarti, Solubility of gases and liquids in glassy polymers, Annu. Rev. Chem. Biomol. Eng. 2 (1) (2011) 97–120.

[22] M. Minelli, G.C. Sarti, 110Th anniversary: Gas and vapor sorption in glassy polymeric membranes—Critical review of different physical and mathematical models, Ind. Eng. Chem. Res. 59 (1) (2020) 341–365.

[23] F. Doghieri, G.C. Sarti, Nonequilibrium lattice fluids: A predictive model for the solubility in glassy polymers, Macromolecules 29 (24) (1996) 7885–7896.

[24] G.C. Sarti, F. Doghieri, Predictions of the solubility of gases in glassy polymers based on the NELF model, Chem. Eng. Sci. 53 (19) (1998) 3435–3447.

[25] F. Doghieri, G.C. Sarti, Predicting the low pressure solubility of gases and vapors in glassy polymers by the NELF model, J. Membr. Sci. 147 (1) (1998) 73–86.

[26] M. Minelli, S. Campagnoli, M.G. De Angelis, F. Doghieri, G.C. Sarti, Predictive model for the solubility of fluid mixtures in glassy polymers, Macromolecules 44 (12) (2011) 4852–4862.

[27] E. Ricci, M. De Angelis, Modelling mixed-gas sorption in glassy polymers for CO2 removal: A sensitivity analysis of the dual mode sorption model, Membranes 9 (1) (2019) 8.

[28] L.M. Robeson, Z.P. Smith, B.D. Freeman, D.R. Paul, Contributions of diffusion and solubility selectivity to the upper bound analysis for glassy gas separation membranes, J. Membr. Sci. 453 (2014) 71–83.

[29] M. Galizia, M.G. De Angelis, G.C. Sarti, Sorption of hydrocarbons and alcohols in addition-type poly(trimethyl silyl norbornene) and other high free volume glassy polymers. II: NELF model predictions, J. Membr. Sci. 405–406 (2012) 201–211.

[30] M. Minelli, G.C. Sarti, Gas permeability in glassy polymers: A thermodynamic approach, Fluid Phase Equilib. 424 (2016) 44–51.

[31] M. Galizia, K.A. Stevens, Z.P. Smith, D.R. Paul, B.D. Freeman, Nonequilibrium lattice fluid modeling of gas solubility in HAB-6FDA polyimide and its thermally rearranged analogues, Macromolecules 49 (22) (2016) 8768–8779.

[32] M.P. Allen, D.J. Tildesley, Computer Simulation of Liquids, Vol. 1, Oxford University Press, 2017.

[33] A.Z. Panagiotopoulos, U.W. Suter, R.C. Reid, Phase diagrams of nonideal fluid mixtures from Monte Carlo simulation, Ind. Eng. Chem. Fundam. 25 (4) (1986) 525–535.

[34] G.C. Boulougouris, I.G. Economou, D.N. Theodorou, On the calculation of the chemical potential using the particle deletion scheme, Mol. Phys. 96 (6) (1999) 905–913.

[35] B. Widom, Some topics in the theory of fluids, J. Chem. Phys. 39 (11) (1963) 2808–2812.

[36] M.R. Siegert, M. Heuchel, D. Hofmann, A generalized direct-particle-deletion scheme for the calculation of chemical potential and solubilities of small- and medium-sized molecules in amorphous polymers, J. Comput. Chem. 28 (5) (2007) 877–889.

[37] M.G. De Angelis, G.C. Boulougouris, D.N. Theodorou, Prediction of infinite dilution benzene solubility in linear polyethylene melts via the direct particle deletion method, J. Phys. Chem. B 114 (19) (2010) 6233–6246.

[38] M. Heuchel, M. Böhning, O. Hölck, M.R. Siegert, D. Hofmann, Atomistic packing models for experimentally investigated swelling states induced by CO2 in glassy polysulfone and poly(ether sulfone), J. Polym. Sci. B 44 (13) (2006) 1874–1897.

[39] N.F.A. van der Vegt, W.J. Briels, M. Wessling, H. Strathmann, The sorption induced glass transition in amorphous glassy polymers, J. Chem. Phys. 110 (22) (1999) 11061–11069.

[40] T. Spyriouni, G.C. Boulougouris, D.N. Theodorou, Prediction of sorption of CO2 in glassy atactic polystyrene at elevated pressures through a new computational scheme, Macromolecules 42 (5) (2009) 1759–1769.

[41] M. Minelli, M.G. De Angelis, D. Hofmann, A novel multiscale method for the prediction of the volumetric and gas solubility behavior of high-T g polyimides, Fluid Phase Equilib. 333 (2012) 87–96.

[42] E. Ricci, M. Minelli, M.G. De Angelis, A multiscale approach to predict the mixed gas separation performance of glassy polymeric membranes for CO2 capture: the case of CO2/CH4 mixture in Matrimid®, J. Membr. Sci. 539 (2017) 88–100.

[43] M. Li, X. Huang, H. Liu, B. Liu, Y. Wu, A. Xiong, T. Dong, Prediction of gas solubility in polymers by back propagation artificial neural network based on self-adaptive particle swarm optimization algorithm and chaos theory, Fluid Phase Equilib. 356 (2013) 11–17.

[44] X. Ru-Ting, H. Xing-Yuan, Predictive calculation of carbon dioxide solubility in polymers, RSC Adv. 5 (94) (2015) 76979–76986.

[45] J.W. Barnett, C.R. Bilchak, Y. Wang, B.C. Benicewicz, L.A. Murdock, T. Bereau, S.K. Kumar, Designing exceptional gas-separation polymer membranes using machine learning, Science Advances 6 (20) (2020) eaaz4301.

[46] Q. Yuan, M. Longo, A.W. Thornton, N.B. McKeown, B. Comesaña-Gándara, J.C. Jansen, K.E. Jelfs, Imputation of missing gas permeability data for polymer membranes using machine learning, J. Membr. Sci. 627 (2021).

[47] C.L. Ritt, M. Liu, T.A. Pham, R. Epsztein, H.J. Kulik, M. Elimelech, Machine learning reveals key ion selectivity mechanisms in polymeric membranes with subnanometer pores, Science Advances 8 (2) (2022) eabl5771.

[48] J. Yang, L. Tao, J. He, J.R. McCutcheon, Y. Li, Machine learning enables interpretable discovery of innovative polymers for gas separation membranes, Science Advances 8 (29) (2022) eabn9545.

[49] L. Tao, G. Chen, Y. Li, Machine learning discovery of high-temperature polymers, Patterns 2 (4) (2021).

[50] G. Chen, L. Tao, Y. Li, Predicting polymers' glass transition temperature by a chemical language processing model, Polymers 13 (11) (2021).

[51] S. Wu, Y. Kondo, M.a. Kakimoto, B. Yang, H. Yamada, I. Kuwajima, G. Lambard, K. Hongo, Y. Xu, J. Shiomi, C. Schick, J. Morikawa, R. Yoshida, Machine-learning-assisted discovery of polymers with high thermal conductivity using a molecular design algorithm, npj Comput. Mater. 5 (1) (2019).

[52] M. High, R. Danner, A group contribution equation of state for polymer solutions, Fluid Phase Equilib. 53 (1989) 323–330.

[53] M.S. High, R.P. Danner, Application of the group contribution lattice-fluid EOS to polymer solutions, AIChE J. 36 (11) (1990) 1625–1632.

[54] A. Tihic, G.M. Kontogeorgis, N. Von Solms, M.L. Michelsen, A predictive group-contribution simplified PC-SAFT equation of state: Application to polymer systems, Ind. Eng. Chem. Res. 47 (15) (2008) 5092–5101.

[55] L. Constantinou, R. Gani, New group contribution method for estimating properties of pure compounds, AIChE J. 40 (10) (1994) 1697–1710.

[56] D. Boudouris, L. Constantinou, C. Panayiotou, A group contribution estimation of the thermodynamic properties of polymers, Ind. Eng. Chem. Res. 36 (9) (1997) 3968–3973.

[57] D. Boudouris, L. Constantinou, C. Panayiotou, Prediction of volumetric behavior and glass transition temperature of polymers: a group contribution approach, Fluid Phase Equilib. 167 (1) (2000) 1–19.

[58] F.T. Peters, F.S. Laube, G. Sadowski, Development of a group contribution method for polymers within the PC-SAFT model, Fluid Phase Equilib. 324 (2012) 70–79.

[59] F.T. Peters, M. Herhut, G. Sadowski, Extension of the PC-SAFT based group contribution method for polymers to aromatic, oxygen- and silicon-based polymers, Fluid Phase Equilib. 339 (2013) 89–104.

[60] H. Matsukawa, M. Kitahara, K. Otake, Estimation of pure component parameters of PC-SAFT EoS by an artificial neural network based on a group contribution method, Fluid Phase Equilib. 548 (2021).

[61] J. Habicht, C. Brandenbusch, G. Sadowski, Predicting PC-SAFT pure-component parameters by machine learning using a molecular fingerprint as key input, Fluid Phase Equilib. 565 (2023).

[62] D. Rogers, M. Hahn, Extended-connectivity fingerprints, J. Chem. Inf. Model. 50 (5) (2010) 742–754.

[63] J. Marrero, R. Gani, Group-contribution based estimation of pure component properties, Fluid Phase Equilib. 183–184 (2001) 183–208.

[64] I.C. Sanchez, A.C. Balazs, Generalization of the lattice-fluid model for specific interactions, Macromolecules 22 (5) (1989) 2325–2331.

[65] B.D. Coleman, M.E. Gurtin, Thermodynamics with internal state variables, J. Chem. Phys. 47 (2) (1967) 597–613.

[66] M.G. Baschetti, F. Doghieri, G.C. Sarti, Solubility in glassy polymers: Correlations through the nonequilibrium lattice fluid model, Ind. Eng. Chem. Res. 40 (14) (2001) 3027–3037.

[67] M. Minelli, K. Friess, O. Vopička, M.G. De Angelis, Modeling gas and vapor sorption in a polymer of intrinsic microporosity (PIM-1), Fluid Phase Equilib. 347 (2013) 35–44.

[68] M. Minelli, G. Cocchi, L. Ansaloni, M.G. Baschetti, M. De Angelis, F. Doghieri, Vapor and liquid sorption in matrimid polyimide: Experimental characterization and modeling, Ind. Eng. Chem. Res. 52 (26) (2013) 8936–8945.
[69] P. Zoller, D.J. Walsh, Standard Pressure Volume Temperature Data for Polymers, Technomic, Lancaster, 1995.
[70] I. Aravind, J. Pionteck, S. Thomas, Transreactions in poly trimethylene terephthalate/bisphenol-A polycarbonate (PC) blends analysed by pressure-volume-temperature measurements, Polym. Test. 31 (1) (2012) 16–24.
[71] J.-s. Wang, R.S. Porter, J.R. Knox, Physical properties of the poly(1-olefin)s. Thermal behavior and dilute solution properties, Polym. J. 10 (6) (1978) 619–628.
[72] A. Gitsas, G. Floudas, H.-J. Butt, T. Pakula, K. Matyjaszewski, Effects of nanoscale confinement and pressure on the dynamics of pODMA-b-ptBA-b-pODMA triblock copolymers, Macromolecules 43 (5) (2010) 2453–2462.
[73] R.-D. Maier, M. Kopf, D. Mäder, F. Koopmann, H. Frey, J. Kressler, Thermodynamics of polymer blends of poly(isobutylene) and poly(dimethylsilylenemethylene), Acta Polym. 49 (7) (1998) 356–362.
[74] R.-D. Maier, r. Kressler, B. Rudolf, P. Reichert, F. Koopmann, H. Frey, R. Mu, Macromolecules 29 (5) (1996) 1490–1497.
[75] M. Paluch, R. Casalini, A. Patkowski, T. Pakula, C.M. Roland, Effect of volume changes on segmental relaxation in siloxane polymers, Phys. Rev. E 68 (3) (2003) 031802.
[76] M. Hess, J. Pionteck, Thermodynamic properties of a series of semi-rigid polyesters, Mater. Res. Innov. 6 (2) (2002) 51–54.
[77] Y. Sato, K. Inohara, S. Takishima, H. Masuoka, M. Imaizumi, H. Yamamoto, M. Takasugi, Pressure-volume-temperature behavior of polylactide, poly(butylene succinate), and poly(butylene succinate-co-adipate), Polym. Eng. Sci. 40 (12) (2000) 2602–2609.
[78] R.E. Bellman, Dynamic Programming, Princeton University Press, 2010.
[79] Z.-H. Zhou, Machine Learning, Springer Singapore, Singapore, 2021.
[80] A.N. Tikhonov, V.Y. Arsenin, Solutions of ill-posed problems, SIAM Rev. 21 (2) (1979) 266–267.
[81] A. Gron, Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, first ed., O'Reilly Media, Inc., 2017.
[82] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: machine learning in python, Journal of Machine Learning Research 12 (85) (2011) 2825–2830.
[83] J. Frutiger, C. Marcarie, J. Abildskov, G. Sin, A comprehensive methodology for development, parameter estimation, and uncertainty analysis of group contribution based property models—An application to the heat of combustion, J. Chem. Eng. Data 61 (1) (2016) 602–613.
[84] F. Vogt, A self-guided search for good local minima of the sum-of-squared-error in nonlinear least squares regression, J. Chemom. 29 (2) (2015) 71–79.
[85] K. Von Konigslow, C.B. Park, R.B. Thompson, Evaluating characteristic parameters for carbon dioxide in the sanchez-lacombe equation of state, J. Chem. Eng. Data 62 (2) (2017) 585–595.
[86] C. Panayiotou, M. Pantoula, E. Stefanis, I. Tsivintzelis, I.G. Economou, Nonrandom hydrogen-bonding model of fluids and their mixtures. 1. Pure fluids, Ind. Eng. Chem. Res. 43 (20) (2004) 6592–6606.
[87] C. Panayiotou, I. Tsivintzelis, I.G. Economou, Nonrandom hydrogen-bonding model of fluids and their mixtures. 2. Multicomponent mixtures, Ind. Eng. Chem. Res. 46 (8) (2007) 2628–2636.
[88] G. Scherillo, L. Sanguigno, M. Galizia, M. Lavorgna, P. Musto, G. Mensitieri, Non-equilibrium compressible lattice theories accounting for hydrogen bonding interactions: Modelling water sorption thermodynamics in fluorinated polyimides, Fluid Phase Equilib. 334 (2012) 166–188.
[89] J. Gross, G. Sadowski, Application of the perturbed-chain SAFT equation of state to associating systems, Ind. Eng. Chem. Res. 41 (22) (2002) 5510–5515.
[90] M. Kleiner, G. Sadowski, Modeling of polar systems using PCP-SAFT: An approach to account for induced-association interactions, J. Phys. Chem. C 111 (43) (2007) 15544–15553.
[91] L. Liu, S.E. Kentish, Modeling of carbon dioxide and water sorption in glassy polymers through PC-SAFT and NET PC-SAFT, Polymer 104 (2016) 149–155, Rheology.
[92] A. Quach, R. Simha, Pressure-volume-temperature properties and transitions of amorphous polymers; polystyrene and poly (orthomethylstyrene), J. Appl. Phys. 42 (12) (1971) 4592–4606.
[93] P.W. Wojtkowski, Aromatic-aliphatic azomethine ether polymers and fibers, Macromolecules 20 (4) (1987) 740–748.
[94] M. Lanč, K. Pilnáček, C.R. Mason, P.M. Budd, Y. Rogan, R. Malpass-Evans, M. Carta, B.C. Gándara, N.B. McKeown, J.C. Jansen, O. Vopička, K. Friess, Gas sorption in polymers of intrinsic microporosity: The difference between solubility coefficients determined via time-lag and direct sorption experiments, J. Membr. Sci. 570–571 (2019) 522–536.
[95] H. Sejour, Investigation of Dithiolenes For Propylene/Propane Membrane Separations (Ph.D. thesis), Georgia Institute of Technology, Atlanta, 2007.
[96] M.G. Baschetti, F. Doghieri, G.C. Sarti, Solubility in glassy polymers: Correlations through the nonequilibrium lattice fluid model, Ind. Eng. Chem. Res. 40 (14) (2001) 3027–3037.
[97] W. Vieth, J. Howell, J. Hsieh, Dual sorption theory, J. Membr. Sci. 1 (1976) 177–220.