polymers

MDPI

Article

Atomistic Descriptors for Machine Learning Models of Solubility Parameters for Small Molecules and Polymers

Mingzhe Chi [1], Rihab Gargouri [2], Tim Schrader [1], Kamel Damak [2], Ramzi Maâlej [2] and Marek Sierka [1,*]

[1] Otto Schott Institute of Materials Research, Friedrich Schiller University Jena, 07743 Jena, Germany; mingzhe.chi@uni-jena.de (M.C.); tim.schrader@uni-jena.de (T.S.)
[2] Georesources Materials Environment and Global Changes Laboratory (GEOGLOB), Faculty of Sciences of Sfax, Sfax University, Sfax 3018, Tunisia; rihab.gargouri.etud@fss.usf.tn (R.G.); kamel.damak@fss.usf.tn (K.D.); ramzi.maalej@fss.usf.tn (R.M.)
* Correspondence: marek.sierka@uni-jena.de

Abstract: Descriptors derived from atomic structure and quantum chemical calculations for small molecules representing polymer repeat elements were evaluated for machine learning models to predict the Hildebrand solubility parameters of the corresponding polymers. Since reliable cohesive energy density data and solubility parameters for polymers are difficult to obtain, the experimental heat of vaporization ΔH_{vap} of a set of small molecules was used as a proxy property to evaluate the descriptors. Using the atomistic descriptors, the multilinear regression model showed good accuracy in predicting ΔH_{vap} of the small-molecule set, with a mean absolute error of 2.63 kJ/mol for training and 3.61 kJ/mol for cross-validation. Kernel ridge regression showed similar performance for the small-molecule training set but slightly worse accuracy for the prediction of ΔH_{vap} of molecules representing repeating polymer elements. The Hildebrand solubility parameters of the polymers derived from the atomistic descriptors of the repeating polymer elements showed good correlation with values from the CROW polymer database.

Keywords: machine learning; polymer; properties prediction

Citation: Chi, M.; Gargouri, R.; Schrader, T.; Damak, K.; Maâlej, R.; Sierka, M. Atomistic Descriptors for Machine Learning Models of Solubility Parameters for Small Molecules and Polymers. *Polymers* **2022**, *14*, 26. https://doi.org/10.3390/polym14010026

Academic Editor: Ming-Jay Deng

Received: 29 October 2021
Accepted: 15 December 2021
Published: 22 December 2021

1. Introduction

Computer-aided predictions of polymer solubility and miscibility with small molecules and drugs are of fundamental importance in a number of industrial applications, including the use of polymers as drug carriers in the growing field of nanomedicine [1]. Among various approaches, solubility and miscibility predictions based on Hildebrand solubility parameters are often used for polymer blends, polymer solutions and polymer–drug mixtures [2]. The Hildebrand model uses a solubility parameter (SP), δ, defined as the square root of the cohesive energy density:

$$\delta = \sqrt{\frac{E_{coh}}{V_m}} \quad (1)$$

where E_{coh} is the cohesive energy, and V_m is the molar volume. The miscibility of two substances can be estimated by comparing the absolute value of the difference in their SPs. If it is more than 2 $MPa^{1/2}$, the two substances are deemed immiscible, and with a difference of less than 2 $MPa^{1/2}$, they are considered miscible [3]. The factor 2 $MPa^{1/2}$ was determined on the basis of empirical considerations [3]. The Hildebrand SP can also be used to roughly estimate the Flory–Huggins interaction parameter [4], which is another useful tool for predicting the miscibility of polymer blends [5].

For low-molecular-weight compounds, E_{coh} and δ can be estimated from the heat of vaporization:

$$E_{coh} \approx \Delta H_{vap} - RT \quad (2)$$

where ΔH_{vap} is the heat of vaporization [2]. However, for polymers, SP is difficult to obtain from experiments [6]. Various experimental methods can be used to indirectly derive SP, such as hot-stage microscopy, differential scanning calorimetry (DSC) and ultraviolet spectroscopy [7,8], but these methods provide limited accuracy and can only be used for a small range of polymer species.

In addition to experimental methods, SP can also be calculated by empirical approaches and computer simulations. A group-contribution method (GC) is an empirical approach, which uses the sum of the contributions of structural and functional groups to estimate polymer properties [9]. GC is easy to apply but has limited accuracy due to the use of empirical assumptions. Although new GC approaches are being developed, a general model that can cover a wide range of polymer species and polymer properties is not available [10]. Atomistic simulations employing force fields and interatomic potential functions are another tool for predicting polymer properties [5,11]. However, accurate SP predictions using atomistic simulations are computationally demanding, especially for polymers and compounds with complex structures [12].

In this regard, data-driven approaches based on machine learning (ML) models have become an appealing alternative to simple empirical approaches and atomistic simulations. ML models have been developed to predict the physical or chemical properties of materials with good accuracy, including solubility parameters [11]. However, the predictive power of ML models depends heavily on the availability of accurate and consistent target data covering a wide structural and compositional range, as well as unbiased descriptors, i.e., readily available observables that can be linked to the target property [13]. Such observables can be derived, e.g., from experimental data or from quantum chemical or atomistic calculations [13,14]. It is difficult to obtain relevant experimental data on target properties and descriptors for predicting the SP of polymers. In the case of descriptors, one approach is to use the features derived for monomer molecules [13]. However, polymer properties may be fundamentally different from those of the monomer. Therefore, small organic molecules that are structurally similar to the repeating element (RE) of the polymer may be a better choice. For a given polymer, different molecules can be identified that represent REs, as shown in Figure 1 for polyethylene glycol (PEG). Both ethylene glycol and ethanol, as polar molecules forming strong hydrogen bonds, are poor choices for deriving molecular descriptors for ML models for PEG.

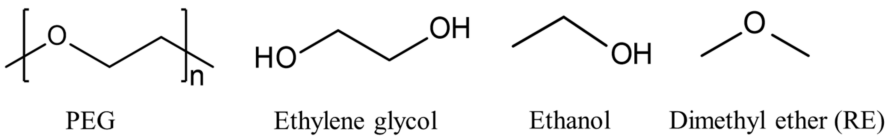


Figure 1. Polyethylene glycol (PEG) and different choices of small molecules with a structural motif of the repeating element of PEG.

In this work, descriptors derived from the atomic structure and quantum chemical calculations of small molecules as potential polymer REs are evaluated for ML models of the polymer SP. Since reliable cohesive energy density and SP data for polymers are difficult to obtain in experiments and simulations, a surrogate target property is used to evaluate the descriptors, namely, the experimental heat of vaporization ΔH_{vap} of the small molecules. For low-molecular-weight compounds, ΔH_{vap} can usually be determined with good accuracy [15]. Subsequently, the relationship between ΔH_{vap} of the polymer RE and the available SP of the polymers is investigated.

2. Method

2.1. Molecular Datasets

ML models for predicting ΔH_{vap} were trained and tested on a dataset of small organic molecules including hydrocarbons, alcohols, acids, amines, ketones, aldehydes, nitriles, organic chlorides and benzene derivatives. A summary of the dataset is shown in Table 1. Figure 2 shows examples of the largest molecules used. The ML models were then applied to

another dataset of organic molecules with structural similarity to REs of popular polymers to predict ΔH_{vap} and correlate it with the polymer SP. This dataset is summarized in Table 2.

Table 1. Summary of the molecules included in the training dataset with experimental ΔH_{vap} (see Supplementary Materials, Tables S1 and S6).

Type	Formula	Size *n*	Number of Molecules
hydrocarbons	C_nH_{2n+2}	1–10	10
acids	$C_nH_{2n+1}COOH$	0–8	9
alcohols	$C_nH_{2n+1}OH$	1–9	9
ketones	$C_nH_{2n}O/C_6H_5COCH_3$	3–7	6
amines	$C_nH_{2n+1}NH_2$	1–6	5
aldehydes	$C_{n-1}H_{2n-1}CHO/C_6H_5CHO$	3–6	5
nitriles	$C_nH_{2n+1}CN$	1, 3–6	5
organic chlorides	$C_nH_{2n+1}Cl$	1, 3–6	5
benzene derivatives	$C_6H_6/C_6H_5OCH_3/C_6H_5OCH_2CH_3/C_6H_5CH_2OCH_3/C_6H_5C_nH_{2n+1}$	1, 2, 4	7

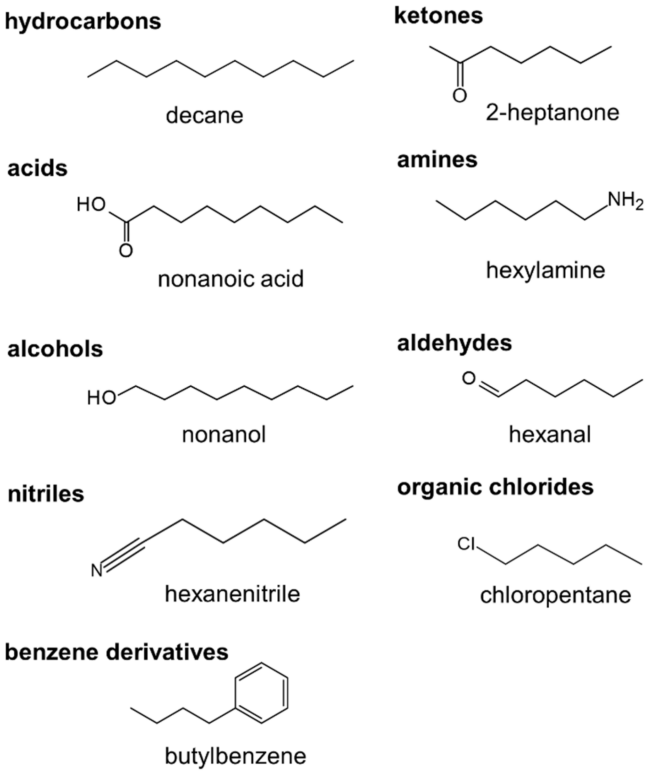


Figure 2. The largest molecule of each type used in the training dataset: decane ($C_{10}H_{22}$), nonanoic acid ($C_8H_{17}COOH$), nonanol ($C_9H_{19}OH$), 2-heptanone ($C_7H_{14}O$), hexylamine ($C_6H_{13}NH_2$), hexanal ($C_5H_{11}CHO$), hexanenitrile ($C_5H_{11}CN$), chloropentane ($C_5H_{11}Cl$) and butylbenzene ($C_6H_5C_4H_9$).

The organic molecules in both datasets cover a wide range of ΔH_{vap} values, from 8.19 kJ/mol (methane) to 69.00 kJ/mol (heptanoic acid), and represent different chemical structures. The experimental values of ΔH_{vap} for all molecules, measured around a normal boiling point, were collected from the literature (see Supplementary Materials).

2.2. Computational Details

All density functional theory calculations were performed as a gas phase using the Turbomole program package [16]. The Becke 3-parameter Lee–Yang–Parr (B3-LYP) [17] exchange–correlation functional was employed, along with triple zeta valence plus po-

larization (def2-TZVP) [18] basis sets and Grimme dispersion correction (DFT-D3) [19]. The geometry convergence criteria for DFT calculations were 10^{-6} hartree for the energy change and 10^{-3} hartree/bohr for the gradient norm. Count descriptors and topological descriptors were calculated with the PaDEL-Descriptor program [20]. Python 3 with the Scikit-learn package was used for building all machine learning models [21].

Table 2. Summary of polymer repeating elements (REs) included in the validation set with experimental ΔH_{vap} (see Supplementary Materials, Tables S2 and S7).

Polymer	Formula	RE	Formula of RE
poly(acrylic acid)	$(C_3H_4O2)_n$	propanoic acid	CH_3CH_2COOH
poly(allyl cyanide)	$(C_4H_5N)_n$	butanenitrile	CH_3CH_2CH2CN
polyacrylonitrile	$(C_3H_3N)_n$	propanenitrile	CH_3CH_2CN
polybutylene	$(C_4H_8)_n$	butane	$CH_3CH_2CH_2CH_3$
polyethylene (HDPE)	$(C_2H_4)_n$	ethane	CH_3CH_3
poly(ethylene glycol)	$(C_2H_4O)_n$	dimethyl ether	CH_3OCH_3
cis-1,4-polyisoprene	$(C_5H_8)_n$	2-methyl-2-butene	$CH_3CHC(CH_3)_2$
polyisobutene	$(C_4H_8)_n$	isobutane	$(CH_3)_2CHCH_3$
polymethacrylonitrile	$(C_4H_5N)_n$	isobutyronitrile	$(CH_3)_2CHCN$
poly(methyl methacrylate)	$(C_5H_8O_2)_n$	methyl butyrate	$CH_3CH_2CH_2COOCH_3$
polypropylene	$(C_3H_6)_n$	propane	$CH_3CH_2CH_3$
polystyrene	$(C_8H_8)_n$	ethylbenzene	$C_6H_5C_2H_5$
poly(vinyl alcohol)	$(C_2H_4O)_n$	ethanol	CH_3CH_2OH
poly(vinyl acetate)	$(C_4H_6O2)_n$	ethyl acetate	$CH_3COOCH_2CH_3$
poly(vinyl chloride)	$(C_2H_3Cl)_n$	chloroethane	CH_3CH_2Cl
poly(vinyl ethyl ether)	$(C_4H_6O)_n$	diethyl ether	$CH_3CH_2OCH_2CH_3$

2.3. Molecular Descriptors

In the current work, four quantum chemical descriptors were obtained using DFT calculations: atomization energy (AE), quadrupole moment (QM), chemical hardness η and electronegativity χ. There are different definitions for the quadrupole moment [22,23]. In the present work, the quadrupole moment was defined as the second moment of charge [23], and QM was taken as

$$\mathrm{QM} = \frac{1}{3}\left(Q_{xx} + Q_{yy} + Q_{zz}\right) \tag{3}$$

where Q_{xx}, Q_{yy} and Q_{zz} are diagonal elements of the second moment of the charge tensor.

Chemical hardness η and electronegativity χ are chemical reactivity descriptors that were applied in an artificial neural network for predicting solvation energies [24]. They are defined as

$$\eta \simeq E_{\mathrm{LUMO}} - E_{\mathrm{HOMO}} \tag{4}$$

and

$$\chi = -\frac{1}{2}\left(E_{\mathrm{LUMO}} + E_{\mathrm{HOMO}}\right) \tag{5}$$

where E_{HOMO} and E_{LUMO} denote the energies of the highest occupied (HOMO) and lowest unoccupied molecular orbitals (LUMO), respectively.

The remaining descriptors, such as number of aromatic bonds (nAromBond) and number of heavy atoms (nHeavyAtom), were generated using the PaDEL-Descriptor [20] program based on the atomic structures of molecules. The descriptors were obtained using exactly the same method for both datasets (Tables 1 and 2). The full specification of the descriptors is given in the Supplementary Materials.

2.4. Machine Learning Models

Two supervised machine learning models were used: multilinear regression (MLR) and kernel ridge regression (KRR) [25]. MLR is the simplest ML model using the least

square method and has been widely applied in data analysis. The MLR model can be represented as a linear combination of all descriptors

$$y_{\text{prediction}} = \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_i x_i + \theta_0 \quad (6)$$

where θ_i is the coefficient of each descriptor x_i, and θ_0 is the intercept. The training of the MLR model involves the determination of the best $(\theta_1, \theta_2, \ldots, \theta_i)$ and θ_0. All hyperparameters used for MLR training used the default values implemented in the Scikit-learn package.

KRR is a combination of the kernel function and ridge regression, which is an improvement on the ordinary linear regression method [26,27]. There are several kernel functions available for different tasks, and in the present work, the polynomial kernel function was applied

$$k(x, x') = (x \cdot x' + c)^d, \quad (7)$$

where x and x' are descriptors, and hyperparameters c and d are the soft margin constant and degree of the polynomial kernel, respectively. The accuracy and performance of the model usually depend on the choice of hyperparameters. Since the dataset was relatively small, changes in parameters other than c and d had little effect on the model accuracy and were therefore set to constant values (alpha (regularization strength) = 0.001 and gamma = none). In the current work, c and d in Equation (7) were determined using the grid search function of Scikit-learn. Based on the grid search results, the value of c had little effect on the model and was finally set to 1. The models with d = 1 and d = 2 showed similar performance, and both models were retained for further study. Changes in parameters other than c and d had little effect on the model accuracy and were therefore left at the default values implemented in the Scikit-learn package.

Due to the limited size of the dataset (61 molecules in total), the leave-one-out cross-validation (LOOCV) method was used in the current work, aiming to make optimal use of each sample and to obtain a more justified model. LOOCV is an extreme case of cross-validation, in which only one sample is selected for testing in each cycle, and the other samples are used to train the model until all samples have been selected once. The final model is optimized by averaging the LOOCV results. MLR and KRR models were trained with the same dataset, and LOOCV was applied for all models. After training and LOOCV, all models were used to predict ΔH_{vap} of polymer REs (16 molecules in Table 2), and the performance of all models was analyzed [28] using root-mean-square error (RMSE)

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (8)$$

mean absolute error (MAE)

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (9)$$

and the average of relative error (ARE)

$$\text{ARE} = \frac{1}{n}\sum_{i=1}^{n}\left|1 - \frac{\hat{y}_i}{y_i}\right| \quad (10)$$

where y_i and $\hat{y}_i$ are the reference and predicted values, respectively. In addition, the coefficient of determination (R^2) was used to describe the proportion of variability in a dataset that can be explained by the model [29].

3. Discussion

3.1. Selection of Molecular Descriptors

Molecular descriptors were manually selected and filtered by analyzing multicollinearity based on correlation coefficients. For this, descriptors were selected in addition to the quantum

chemical descriptors that showed low multicollinearity (correlation coefficient within ±0.75) and a high degree of correlation with ΔH_{vap}. The 15 descriptors finally selected are shown in Figure 3.

Figure 3. Correlation coefficients among 15 descriptors. AE: atomization energy, QM: quadrupole moment, nAromBond: number of aromatic bonds; nHeavyAtom: number of heavy atoms (all but hydrogen); SsOH: sum of (-OH) E-States; SssO: sum of (-O-) E-States; nAcid: number of acidic groups; SHdsCH: sum of (=CH-) E-States; SsNH2: sum of (-NH2) E-States; SsCH3: sum of (-CH3) E-States; SHBa: sum of E-States for hydrogen bond acceptors; SHBd: sum of E-States for hydrogen bond donors; SssCH2: sum of (-CH2) E-States (see Supplementary Materials); η: chemical hardness; χ: electronegativity.

Figure 3 shows that there is only weak correlation among most descriptors, which can reduce the risk of collinearity problems [30]. Reducing redundant and irrelevant descriptors also lowers the cost of training and reduces the possibility of an overfitting problem [14,31].

3.2. Predictions of ΔH_{vap} for Small Organic Molecules

The final MLR model for predicting ΔH_{vap} (in kJ/mol) is given as

$$\begin{aligned}\Delta H_{vap} = {} & -23.723\text{AE} + 0.234\text{QM} - 3.303\text{nAromBond} + 3.601\text{SsOH} - 0.33\text{SssO} + 0.477\text{SsCH3} + 2.753\text{SsNH2} \\ & -0.65\text{SHBa} + 1.301\text{SHdsCH} - 5.580\text{nAcid} - 0.618\text{SssCH2} - 15.805\text{SHBd} \\ & +15.53\text{nHeavyAtom} + 10.667\end{aligned} \quad (11)$$

This model does not contain any unreasonably small or large factors for descriptors, which indicates that there are no irrelevant or redundant descriptors. Figure 4 shows that MLR performed well for the training set of small molecules and the LOOCV, according to the R^2 score and other metrics. ARE for training (0.071) and LOOCV (0.105) showed the same trend as the other metrics. The MLR model showed good accuracy for predicting ΔH_{vap} of molecules, with a maximum deviation of 12.43 kJ/mol for ethanoic acid. However, there are large disparities in the values of ΔH_{vap} for ethanoic acid across the literature (from 23.7 (at 391.1 K) to 42 (at 305 K) kJ/mol) [32,33]. The overall deviation is within experimental accuracy. For the LOOCV, both the RMSE of 5.291 kJ/mol and MAE of 3.607 kJ/mol are within ranges indicative of good accuracy.

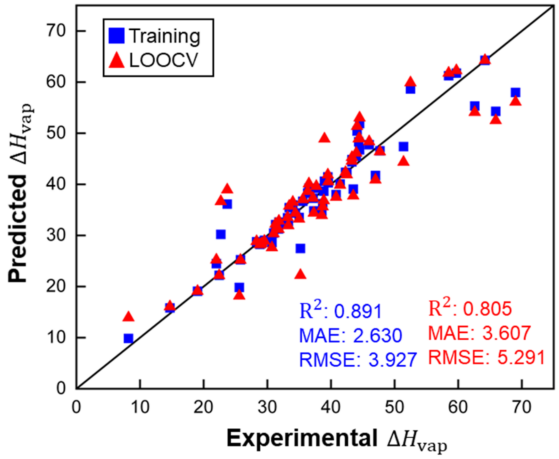
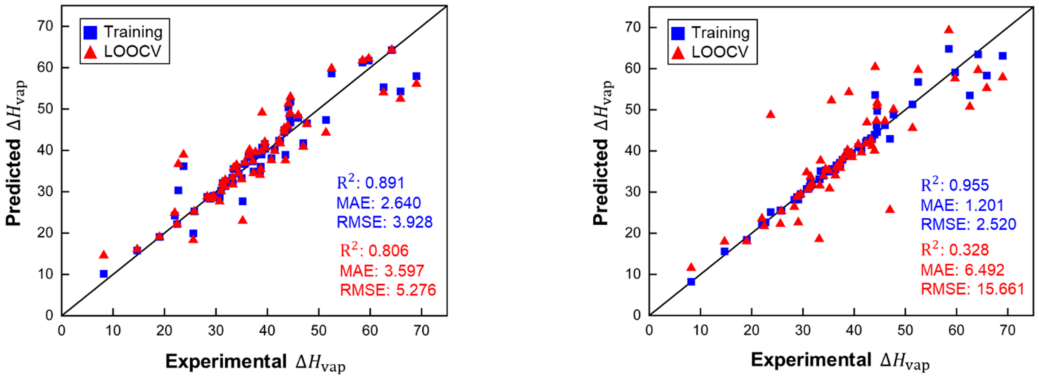


Figure 4. Performance of the MLR model for ΔH_{vap} predictions of small molecules (ARE for the training set: 0.072 and for LOOCV: 0.105). ΔH_{vap}, MAE and RMSE in kJ/mol.

The performance comparison of the final KRR (d = 1) and KRR (d = 2) models is shown in Figure 5.



(**a**) (**b**)

Figure 5. Performance of the KRR models for ΔH_{vap} predictions of small molecules: (**a**) d = 1 (ARE for the training set: 0.072 and for LOOCV: 0.106), (**b**) d = 2 (ARE for the training set: 0.026 and for LOOCV: 0.184). ΔH_{vap}, MAE and RMSE in kJ/mol.

Compared to the MLR model, the KRR model (d = 1) did not perform better in training, but all metrics had a small lead in cross-validation, which showed slightly better stability. KRR (d = 2) performed best during training but was the worst in LOOCV, and this case was most likely due to the overfitting. Considering the size of the datasets used in the current work, high-scoring ML models trained with small datasets can often suffer from overfitting.

3.3. Predictions of ΔH_{vap} for Polymer Repeating Elements

All three models were applied to predict ΔH_{vap} of molecules representing polymer REs. Table 3 and Figure 6 show that the MLR and KRR (d = 1) models provided the best accuracy. The KRR (d = 2) model failed to predict ΔH_{vap} of polymer RE, and the much larger error of the KRR (d = 2) suggests that the model was overfitted. As mentioned, KRR algorithms do not offer advantages on small datasets.

Figures 4–6 demonstrate that the MLR model showed slightly worse performance than the two KRR models during training and cross-validation, but the MAE of MLR for polymer RE was better than that of KRR (d = 1). Therefore, the MLR model and the KRR model (d = 1) in the current work have better extrapolation ability than the KRR (d = 2) models. However, the KRR algorithm with higher d could still yield better results for a larger dataset with more complex structures and chemical compositions.

Table 3. Performance comparison of MLR and KRR models (MAE and RMSE in kJ/mol).

	Training Set (LOOCV)				Polymer REs			
	R^2	MAE	RMSE	ARE	R^2	MAE	RMSE	ARE
MLR	0.805	3.607	5.291	0.105	0.819	3.533	4.348	0.118
KRR ($d = 1$)	0.806	3.597	5.276	0.106	0.821	3.540	4.348	0.118
KRR ($d = 2$)	0.328	2.520	15.661	0.184	0.311	23.472	36.536	0.820

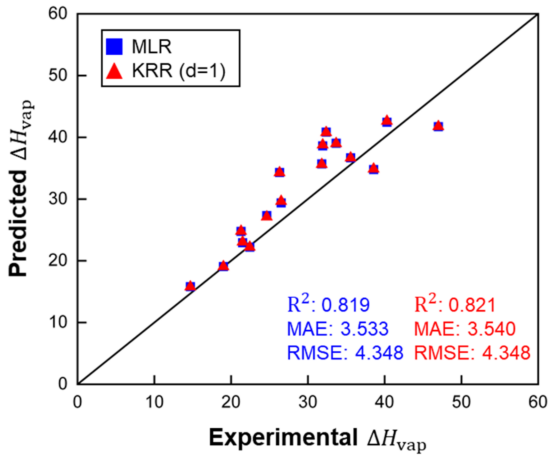


Figure 6. ΔH_{vap} predictions of polymer RE by MLR and KRR (d = 1), ARE for MLR: 0.118 and for KRR (d = 1): 0.118. ΔH_{vap}, MAE and RMSE in kJ/mol.

3.4. Hildebrand Solubility Parameter of Polymers

Our results show that the heat of vaporization of small molecules and polymeric REs, and thus their SPs, can be predicted with good accuracy using the MLR and KRR (d = 1) models. The question now is how well the Hildebrand SP of RE correlates with the SP of the corresponding polymers. For this, Hildebrand SPs of polymers were collected from the CROW polymer database [34] with recommended values, and SPs of REs were calculated from MLR-predicted ΔH_{vap} (see Supplementary Materials). Figure 7 shows the correlation of Hildebrand SPs between polymers and REs. The linear model yields an R^2 value of 0.855.

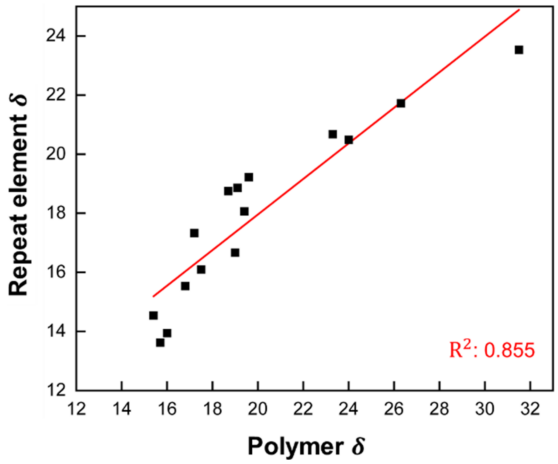


Figure 7. Correlation of Hildebrandt SP between polymers and REs. All values in $MPa^{1/2}$. Linear fit model: $\delta_{polymer} = 0.602\delta_{RE} + 5.915$.

There are several factors that can affect the accuracy of Hildebrand SP predictions for polymers. First, the experimental values of SPs for polymers can only be determined indirectly, and the accuracy of such values is essentially indeterminate. Second, the SPs

are determined not only by the internal structure of the polymer chains, reflected here in the descriptors derived from the polymer RE, but also by factors such as the degree of polymerization, polydispersity, and the nature of the end groups. Such factors cannot be determined from the properties of REs alone and must be derived from experimental data. How well the two descriptors, chemical hardness and electronegativity, actually help in the prediction of solubility parameters needs to be further investigated, as intuitively, the association between the two and polymer solubility is not strong. In addition, larger chemical structures, such as oligomers with several repeating units, may provide more information about inter- and intramolecular interactions and can improve the accuracy of machine learning models. Simulations of such structures are obviously less computationally expensive than simulations of polymers, but finding suitable descriptors may still be a challenge. This is also one of the pathways for future research studies.

4. Conclusions

In this work, descriptors derived from atomic structure and quantum chemical calculations for small molecules as potential polymer repeating elements were evaluated for machine learning models to predict the Hildebrand solubility parameters of the corresponding polymers. Since reliable cohesive energy density data and solubility parameters for polymers are difficult to obtain, the experimental heat of vaporization ΔH_{vap} of small molecules was used as a proxy property to evaluate the descriptors. The multilinear and kernel ridge regression model (with polynomial kernel degree = 1) showed good and very similar performance in training, cross-validation and the prediction of molecules representing polymer repeating elements. The kernel ridge regression model (degree = 2) was strongly overfitted, which was revealed by its poor performance in cross-validation and prediction. The Hildebrand solubility parameters derived from the multilinear regression model for the ΔH_{vap} of polymer repeating elements showed good correlation with the solubility parameters of the corresponding polymers collected from the CROW polymer database. However, atomistic descriptors derived from polymer repeating elements only reflect the internal structure of the polymer chains. More accurate models for predicting the Hildebrand solubility parameters of polymers must take into account additional relevant factors, such as the degree of polymerization, polydispersity and the nature of the polymer end groups. Such factors cannot be determined from the properties of the repeating elements of the polymer alone and must be derived from experimental data.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/polym14010026/s1, Table S1: Heat of vaporization ΔH_{vap} of small molecules; Table S2: Heat of vaporization ΔH_{vap} of polymer REs; Table S3: MLR predictions on polymer RE dataset; Table S4: KRR predictions on polymer RE dataset; Table S5: Hildebrand solubility parameter δ of polymers and calculated δ of REs; Table S6: Complete dataset of small organic molecules; Table S7: Complete dataset of polymer REs.

Author Contributions: Conceptualization, M.S.; methodology, M.S.; investigation, M.C., R.G. and T.S.; writing—original draft preparation, M.C., R.G., M.S., K.D. and R.M.; writing—review and editing, M.C., R.G., M.S., K.D. and R.M.; supervision, M.S. and R.M.; project administration, M.S.; funding acquisition, M.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) within the Collaborative Research Center PolyTarget (SFB 1278, project number 316213987, project A01).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All related data and results can be found in Supplementary Materials.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Qiu, L.Y.; Bae, Y.H. Polymer architecture and drug delivery. *Pharm. Res.* **2006**, *23*, 1–30. [CrossRef]
2. Hansen, C.M. 50 Years with solubility parameters—Past and future. *Prog. Org. Coat.* **2004**, *51*, 77–84. [CrossRef]
3. Venkatram, S.; Kim, C.; Chandrasekaran, A.; Ramprasad, R. Critical Assessment of the Hildebrand and Hansen Solubility Parameters for Polymers. *J. Chem. Inf. Model.* **2019**, *59*, 4188–4194. [CrossRef]
4. Erlebach, A.; Ott, T.; Otzen, C.; Schubert, S.; Czaplewska, J.; Schubert, U.S.; Sierka, M. Thermodynamic compatibility of actives encapsulated into PEG-PLA nanoparticles: In Silico predictions and experimental verification. *J. Comput. Chem.* **2016**, *37*, 2220–2227. [CrossRef]
5. Erlebach, A.; Muljajew, I.; Chi, M.Z.; Buckmann, C.; Weber, C.; Schubert, U.S.; Sierka, M. Predicting Solubility of Small Molecules in Macromolecular Compounds for Nanomedicine Application from Atomistic Simulations. *Adv. Theor. Simul.* **2020**, *3*, 2000001. [CrossRef]
6. Belmares, M.; Blanco, M.; Goddard, W.A.; Ross, R.B.; Caldwell, G.; Chou, S.H.; Pham, J.; Olofson, P.M.; Thomas, C. Hildebrand and Hansen solubility parameters from molecular dynamics with applications to electronic nose polymer sensors. *J. Comput. Chem.* **2004**, *25*, 1814–1826. [CrossRef]
7. Carvalho, S.P.; Lucas, E.F.; Gonzalez, G.; Spinelli, L.S. Determining Hildebrand Solubility Parameter by Ultraviolet Spectroscopy and Microcalorimetry. *J. Brazil. Chem. Soc.* **2013**, *24*, 1998–2007. [CrossRef]
8. Forster, A.; Hempenstall, J.; Tucker, I.; Rades, T. Selection of excipients for melt extrusion with two poorly water-soluble drugs by solubility parameter calculation and thermal analysis. *Int. J. Pharm.* **2001**, *226*, 147–161. [CrossRef]
9. Constantinou, L.; Gani, R. New Group-Contribution Method for Estimating Properties of Pure Compounds. *AIChE J.* **1994**, *40*, 1697–1710. [CrossRef]
10. Stefanis, E.; Constantinou, L.; Panayiotou, C. A group-contribution method for predicting pure component properties of biochemical and safety interest. *Ind. Eng. Chem. Res.* **2004**, *43*, 6253–6261. [CrossRef]
11. Walden, D.M.; Bundey, Y.; Jagarapu, A.; Antontsev, V.; Chakravarty, K.; Varshney, J. Molecular Simulation and Statistical Learning Methods toward Predicting Drug-Polymer Amorphous Solid Dispersion Miscibility, Stability, and Formulation Design. *Molecules* **2021**, *26*, 182. [CrossRef]
12. Cailliez, F.; Bourasseau, A.; Pernot, P. Calibration of Forcefields for Molecular Simulation: Sequential Design of Computer Experiments for Building Cost-Efficient Kriging Metamodels. *J. Comput. Chem.* **2014**, *35*, 130–149. [CrossRef]
13. Karelson, M.; Lobanov, V.S.; Katritzky, A.R. Quantum-chemical descriptors in QSAR/QSPR studies. *Chem. Rev.* **1996**, *96*, 1027–1043. [CrossRef]
14. Cano, G.; Garcia-Rodriguez, J.; Garcia-Garcia, A.; Perez-Sanchez, H.; Benediktsson, J.A.; Thapa, A.; Barr, A. Automatic selection of molecular descriptors using random forest: Application to drug discovery. *Expert. Syst. Appl.* **2017**, *72*, 151–159. [CrossRef]
15. Gopinathan, N.; Saraf, D.N. Predict heat of vaporization of crudes and pure components—Revised II. *Fluid Phase Equilibr.* **2001**, *179*, 277–284. [CrossRef]
16. Perdew, J.P.; Burke, K.; Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868, Erratum: *Phys. Rev. Lett.* **1997**, *78*, 1396. [CrossRef]
17. Van Nhu, N.; Singh, M.; Leonhard, K. Quantum mechanically based estimation of perturbed-chain polar statistical associating fluid theory parameters for analyzing their physical significance and predicting properties. *J. Phys. Chem. B* **2008**, *112*, 5693–5701. [CrossRef]
18. Zheng, J.J.; Xu, X.F.; Truhlar, D.G. Minimally augmented Karlsruhe basis sets. *Theor. Chem. Acc.* **2011**, *128*, 295–305. [CrossRef]
19. Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the Damping Function in Dispersion Corrected Density Functional Theory. *J. Comput. Chem.* **2011**, *32*, 1456–1465. [CrossRef]
20. Yap, C.W. PaDEL-Descriptor: An Open Source Software to Calculate Molecular Descriptors and Fingerprints. *J. Comput. Chem.* **2011**, *32*, 1466–1474. [CrossRef]
21. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
22. Applequist, J. Traceless Cartesian Tensor Forms for Spherical Harmonic-Functions—New Theorems and Applications to Electrostatics of Dielectric Media. *J. Phys. A-Math. Gen.* **1989**, *22*, 4303–4330. [CrossRef]
23. Buckingham, A.D.; Disch, R.L.; Dunmur, D.A. Quadrupole Moments of Some Simple Molecules. *J. Am. Chem. Soc.* **1968**, *90*, 3104–3107. [CrossRef]
24. Yang, J.; Knape, M.J.; Burkert, O.; Mazzini, V.; Jung, A.; Craig, V.S.J.; Miranda-Quintana, R.A.; Bluhmki, E.; Smiatek, J. Artificial neural networks for the prediction of solvation energies based on experimental and computational data. *Phys. Chem. Chem. Phys.* **2020**, *22*, 24359–24364. [CrossRef]
25. Butler, K.T.; Davies, D.W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* **2018**, *559*, 547–555. [CrossRef]
26. Douak, F.; Melgani, F.; Benoudjit, N. Kernel ridge regression with active learning for wind speed prediction. *Appl. Energ.* **2013**, *103*, 328–340. [CrossRef]
27. Zhang, Y.C.; Duchi, J.; Wainwright, M. Divide and Conquer Kernel Ridge Regression: A Distributed Algorithm with Minimax Optimal Rates. *J. Mach. Learn. Res.* **2015**, *16*, 3299–3340.

28. Chai, T.; Draxler, R.R. Root mean square error (RMSE) or mean absolute error (MAE)?—Arguments against avoiding RMSE in the literature. *Geosci. Model. Dev.* **2014**, *7*, 1247–1250. [CrossRef]
29. Szlek, J.; Paclawski, A.; Lau, R.; Jachowicz, R.; Kazemi, P.; Mendyk, A. Empirical search for factors affecting mean particle size of PLGA microspheres containing macromolecular drugs. *Comput. Meth. Prog. Bio.* **2016**, *134*, 137–147. [CrossRef]
30. Dormann, C.F.; Elith, J.; Bacher, S.; Buchmann, C.; Carl, G.; Carre, G.; Marquez, J.R.G.; Gruber, B.; Lafourcade, B.; Leitao, P.J.; et al. Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography* **2013**, *36*, 27–46. [CrossRef]
31. Dietterich, T. Overfitting and undercomputing in machine learning. *ACM Comput. Surv.* **1995**, *27*, 326–327. [CrossRef]
32. Stephenson, R.M.; Malanowski, S. *Handbook of the Thermodynamics of Organic Compounds*, 1st ed.; Stephenson, R.M., Ed.; Springer Science & Business Media: Berlin/Heidelberg, Germany, 1987; p. 552.
33. Majer, V.; Svoboda, V. *Enthalpies of Vaporization of Organic Compounds: A Critical Review and Data Compilation*; Blackwell Scientific Publications: Oxford, UK, 1986.
34. Chemical Retrieval on the Web (CROW). Available online: http://www.polymerdatabase.com/ (accessed on 24 October 2021).