Evaluating Polymer Representations via Quantifying Structure-Property Relationships

Ruimin Ma[1], Zeyu Liu[1], Quanwei Zhang[1], Zhiyu Liu[1], Tengfei Luo[1,2*]

1 Department of Aerospace and Mechanical Engineering, University of Notre Dame, Notre Dame, Indiana 46556, United States

2 Department of Chemical and Biomolecular Engineering, University of Notre Dame, Notre Dame, Indiana 46556, United States

* Corresponding email: tluo@nd.edu

Abstract

Machine learning techniques are being applied in quantifying structure-property relationships for a wide variety of materials, where the properly representing materials plays key roles. Although algorithms for representation learning are extensively studied, their applications to domain-specific areas, such as polymer, are limited largely due to the lack of benchmark databases. In this work, we investigate different types of polymer representations, including Morgan Fingerprint (MF), molecular embedding (ME) and molecular graph (MG), based on a benchmark database from a subset of PolyInfo. We evaluate the quality of different polymer representations via quantifying the relationships between the representations and polymer properties, including density, melting temperature and glass transition temperature. Different representation learning schemes, such as supervised learning, semi-supervised learning and transfer learning, are investigated. It is found that ME outperforms the other representations for structure-property relationship quantification in all cases studied, and MG is shown to be much inferior than ME and MF, likely due to the relatively small volumes of training data available. For MEs, it is found that the similarities of substructure MEs under different learning schemes (e.g., SL, SSL and TL) are differently estimated, thus leading to different performance scores in structure-property relation quantification. Several ME mixtures have shown to outperform the single MEs in the corresponding regression tasks, and this is attributed to the information gain when mixing different ME.

INTRODUCTION

With the emergence of big data and machine learning,[1-3] the data-driven science, which unifies theory, experiment and computation, is being advanced rapidly. Data-driven science has great advantage in identifying patterns rapidly, and is being broadly applied in fields like speech and image recognition[4], bioinformatics[5] and economics.[6] Researchers in the materials science community have started to adopt the data-driven approaches to quantify the structure-property relationships, where material informatics becomes increasingly popular.[7] Fischer et al. adopted early data mining techniques combined with quantum mechanics approaches to design stable crystal structure of materials in 2006.[8] In 2013, Koji et al. used data-driven approach to rapidly design lithium superionic conductors based on the data calculated from fist-principle molecular dynamics.[9] Zhan et al. predicted the thermal boundary resistance using data-driven method based on experimental data.[10] Blay et al. predicted up to eight different properties of zeolites using machine learning and perturbation theory, which enabled the data-driven design of zeolites as inorganic catalysts.[11] Shi et al. developed machine learning models to predict specific surface area (SSA) of ABO_3-type perovskite so that users can search for additional perovskite materials with high SSA using their model.[12] Hachmann et al. also built a highly diverse database for designing the next generation of organic photovoltaics and understanding the structure-property relationship in the domain of organic electronics.[13] These data-driven machine learning approaches can potentially help the rapid screening of materials with properties of interest and provide useful guidance for *de novo* material design.

The success of machine learning algorithms can greatly depends on how data is represented, since different representations can have different explanatory factors of variation behind the data.[14] Different representation learning techniques have been applied to fields like natural language processing[15], image recognition[16], and even music[17], taking advantage of the big data in those fields. Recently, with the exponential increase in the volume of data in materials science, representation learning techniques for materials have been studied, so as to improve the accuracy of quantifying the structure-property relationships.[18-22] However, in the organic materials field, the techniques being developed for materials representation are mainly based on the drug-like small molecules.[19-22] We are still not very clear about

what might be a better representation for polymers, even though a few works[23-27] have applied some of the aforementioned techniques to represent polymers.

In this work, we explore different kinds of mathematical representations for polymers, including Morgan fingerprint (MF), molecular embedding (ME) and molecular graph (MG), and evaluate their quality via quantifying the structure-property relationships for properties like density, melting temperature, glass transition temperature, and performing similarity studies. It is found that ME outperform the other two kinds of representations for polymer in structure-property relationship regression. Different ME learning schemes like supervised learning (SL), semi-supervised learning (SSL) and transfer learning (TL) are also studied, and it is found SSL and TL can have slightly better performance than SL likely due to the larger data set used in the training. The reduced-dimension visualization of MEs of substructures is used to examine the neighbor list of the substructures and measure the similarity between different substructures. The differently estimated similarities for MEs under different learning schemes might explain the different performance scores in structure-property relationship quantification. Finally, mixing of MEs is explored and identified as a way to further improve such a representation learning process.

METHODS

Dataset: The benchmark dataset used in our work are built based on the well-known web-based polymer database, PolyInfo.[28] 1442 homopolymer structures from different polymer classes are collected and all property data studied are for neat polymers (i.e., non-composite). Each homopolymer is represented as a two-monomer structure in this study, since the two-monomer structure contains all the chemical information of a polymer while the one-monomer structure will leave out the bonding information between neighboring monomers. This can be important for correctly capturing the polymer structure-property relationship since many polymer properties are inherently related to the polymer conformation, which depends on the bonding characteristics.[29, 30] Out of the 1442 homopolymers, 318 of them are labeled by the density (ρ), 641 of them are labeled by the melting temperature (T_m), and 1034 of them are labeled by

the glass transition temperature (T_g). If multiple values are recorded for a homopolymer in each label, they are reduced to the mean value. The distributions of density, melting temperature and glass transition temperature in the database and the correlation plots among them are shown in **Figure 1**. Based on the plots, the properties distributions are nearly normally distributed, and a relatively strong correlation is found between glass transition temperature and melting temperature, but the correlations between density and the other two are weak as indicated by the R^2 values.

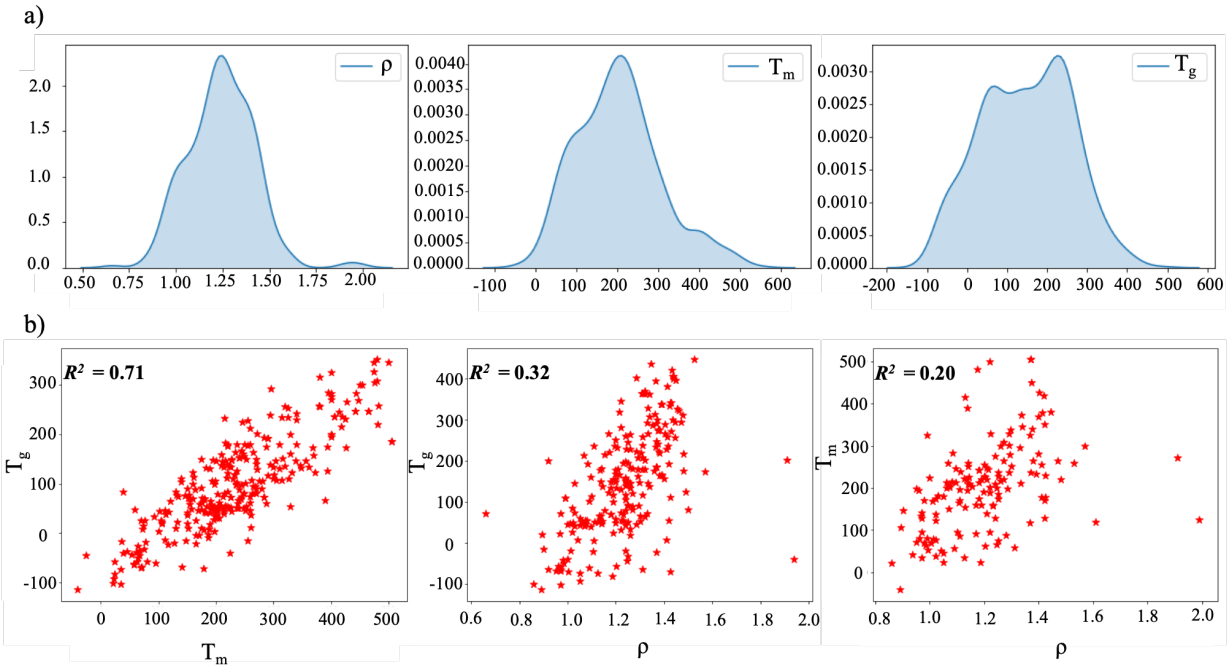


Figure 1. a) Distributions of density, melting temperature and glass transition temperature data in the studied polymer database; b) Correlation plots between different properties of interest.

Representations: Three mathematical representations are employed for the polymer structures, including MF, ME and MG. MF, also known as extended-connectivity fingerprints, is the most commonly used mathematical representation in organic molecular activity predictions.[31-37] To generate a MF, all substructures around all non-hydrogen atoms of a molecule within a defined radius are generated and converted to unique identifiers.[31] These identifiers are then usually hashed to high-dimensional and sparse vectors with a fixed length. A disadvantage is that these vectors are likely to contain bit collisions. Although also based on the identifiers calculated by the Morgan algorithm, ME, a continuous-value vector

obtained by the Mol2vec[21] model, is the post-representation of each identifier, which avoids bit collision. The MEs of each identifier is obtained through machine learning. In this work, the MEs are obtained using the package implemented in Ref. [21]. The substructures in MF are represented as one-hot vectors, so the similarity between different substructures, which is measured by the dot product of two vectors, is 0, but the ME representation can achieve the similarity measurement between different substructures as enabled by the continuous-value feature of this representation. **Figure 2a** shows a representative two-monomer polymer structure and the schematic diagram of MF and ME are shown in **Figure 2b**. MG, on the other hand, treats molecules as an undirected graph with attributed nodes and edges, and it is represented as a molecular tensor as implemented by Coley et al.[20] **Figure 2c** shows an example MG of propane and its corresponding molecular tensor. The propane is first treated as an undirected graph, each heavy atom with its surrounding hydrogen atoms are treated as a node and the bonds between different atom groups are treated as an edge. Finally, the atom feature, such as atomic identification of heavy atoms, and the bond feature, such as bond order, are encoded and used to populate the molecular tensor. The atom feature and bond feature are partially visualized here as an example, and the rest features can be found in Ref [20]. RDKit,[38] an open-source cheminformatics package, is used in this study for molecular file preprocessing, identifier calculation and MF generation.

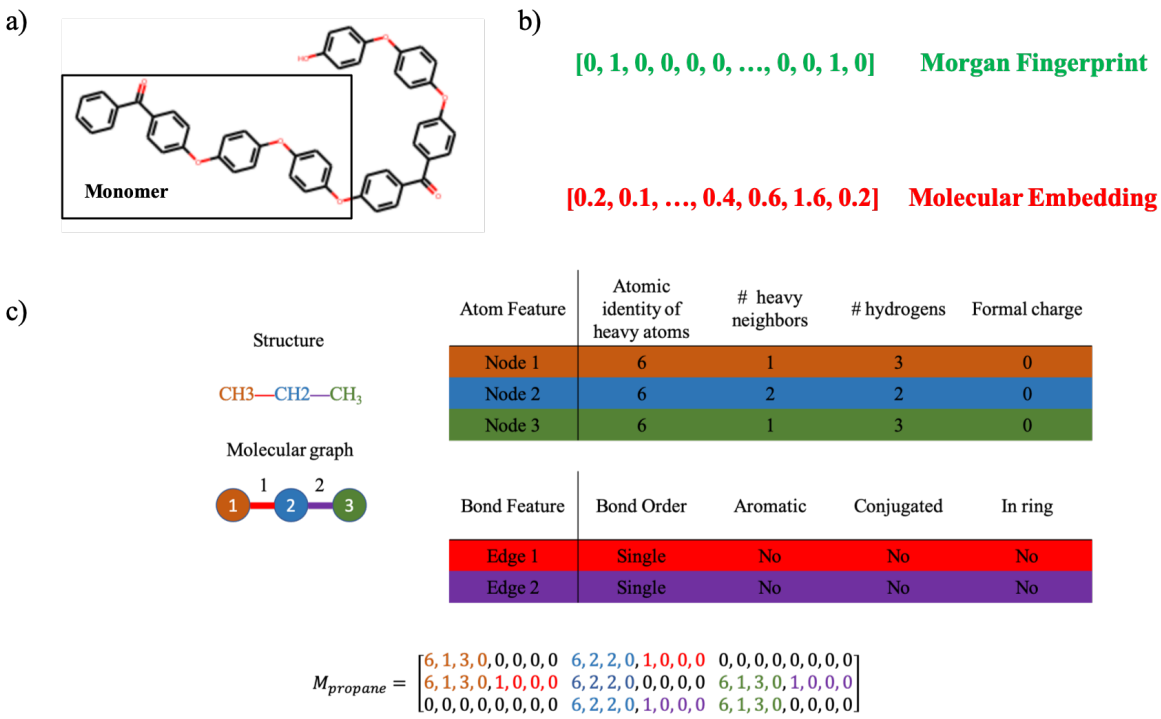


Figure 2. a) An example of a two-monomer polymer structure; b) The schematic diagram of MF and ME: MF is a binary vector, a "1" in MF indicates a unique substructure in the polymer structure, the "1s" at different positions represent different substructures. MEs the substructures are continuous-value vectors, and the ME of the polymer structure is the summation of all MEs of the substructures in the polymer; c) An example MG of propane and its corresponding molecular tensor. The atom and bond attributes of propane are extracted and individually color coded to indicate how their features are used to populate the molecular tensor, *M*.

Machine Learning Methods: Three different machine learning methods — random forest (RF), multi-layer perceptron (MLP) and support vector machine (SVM) — are used for the structure-property relationship training to evaluate the performance of MF and ME. Scikit-learn[39] is used for all the three learning schemes. RF is an ensemble learning method that fits a number of decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and mitigate over-fitting,[40, 41] and the number of trees is set to 500 in the RF training. MLP, also known as the feed-forward neural-network, consists of a system of simple interconnected neurons.[42] Two hidden layers are used here, which

contain 20 and 15 neurons respectively. All the hidden layers have the rectified linear unit (ReLU) activation function[43] and the Adam optimizer[44] is used to minimize the mean squared error for regression. The basic idea of SVM is to first map the data into a high dimensional input space and then construct an optimal separating hyperplane in this space.[45] The SVM is used for regression task here, in which the radial basis function is used, the penalty parameter (C) of the error term is set to 20 and the epsilon value that specifies the penalty-free area is set to 0.2. All the hyper-parameters mentioned above have been optimized before producing the final results.

RESULTS

Firstly, we perform a limited scope parametric study for different radii in generating substructures, different sizes of MEs, and different lengths of MFs. In this parametric study, all three machine learning models are trained using the 318 polymers labeled with densities in a 5-fold cross-validation manner, and the coefficient of determination (R^2) is used for performance evaluation. Based on the results shown in **Figure 3**, we choose a radius size of 1, ME size of 300 and the MF length of 1024 and 2048 for the rest of the study.

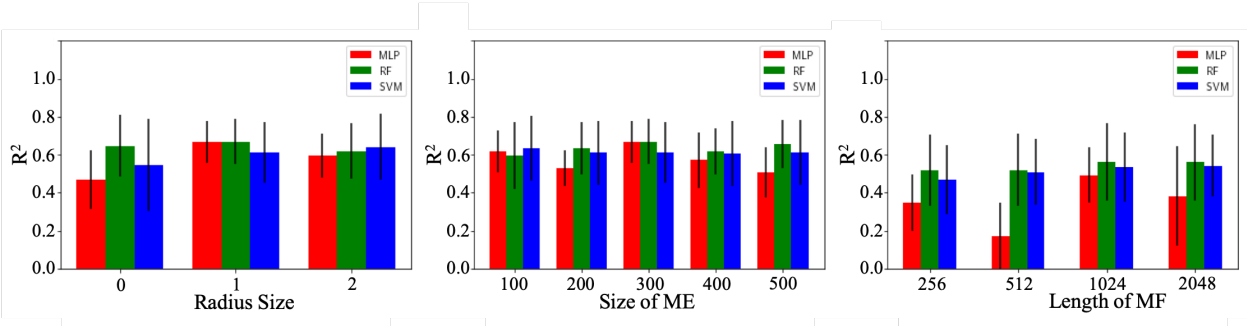


Figure 3. Coefficient of determination (R^2) under different tests: the radius size, the size of ME and the length of MF. In these tests, while changing the radius size, we keep the size of ME constant, and while changing the size of ME and length of MF, we keep the radius size constant as 1.

After obtaining the appropriate size of ME and lengths of MF, the relationships between polymer representations and their corresponding properties are quantified using all three machine learning methods

(i.e., RF, MLP and SVM). The 5-fold cross-validation is used and three different performance metrics — coefficient of determination (R^2), mean squared error (MSE) and mean absolute error (MAE) — are employed for performance evaluation. The quantitative structure-property relationships between the polymer representations and different properties are shown in **Table 1**, **Table 2**, and **Table 3**. Based on the results, we find that ME outperforms MF in all cases as the polymer representation in quantifying the structure-property relationships, as indicated by the higher R^2, lower MSE and lower MAE. The MF can only capture the substructure counts as a binary fingerprint, while the ME cannot only capture the substructure counts but also indicate substructure importance via the vector amplitude, which provides more information in learning the structure-property relationships, and thus leading to better performance. To visually capture the different performances of MF and ME, we show a few representative "predictions vs. ground truths" plots on validation datasets for all three properties in **Figure 4**, and the data plotted here correspond to the best performance of MF and ME.

Table 1. Quantitative structure-property relationships between MF, ME and density (ρ).

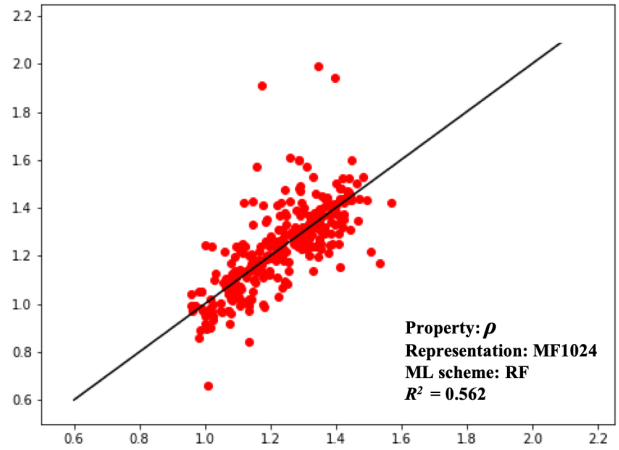
Regressor	Metrics	Morgan Fingerprint (MF)		Molecular Embedding (ME)		
		Length of 1024	Length of 2048	SLME	SSLME	TLME
MLP	R^2	0.492±0.146	0.380±0.262	0.667±0.108	0.696±0.064	**0.724±0.068**
	MSE	0.016±0.006	0.019±0.009	0.011±0.005	**0.009±0.002**	0.009±0.003
	MAE	0.083±0.011	0.086±0.015	0.074±0.014	0.068±0.009	**0.063±0.007**
RF	R^2	0.562±0.204	0.561±0.202	0.668±0.118	0.701±0.093	0.648±0.137
	MSE	0.014±0.008	0.014±0.008	0.011±0.005	0.010±0.004	0.011±0.0066
	MAE	0.077±0.019	0.077±0.019	0.070±0.013	0.065±0.012	0.072±0.014
SVM	R^2	0.534±0.182	0.542±0.163	0.610±0.161	0.659±0.085	0.720±0.103
	MSE	0.015±0.007	0.015±0.006	0.013±0.007	0.011±0.003	0.009±0.004
	MAE	0.080±0.015	0.081±0.014	0.074±0.017	0.071±0.009	0.063±0.012

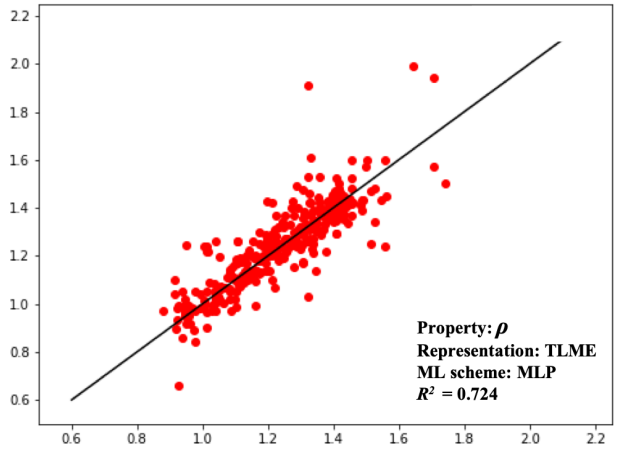
Table 2. Quantitative structure-property relationships between MF, ME and melting temperature (T_m).

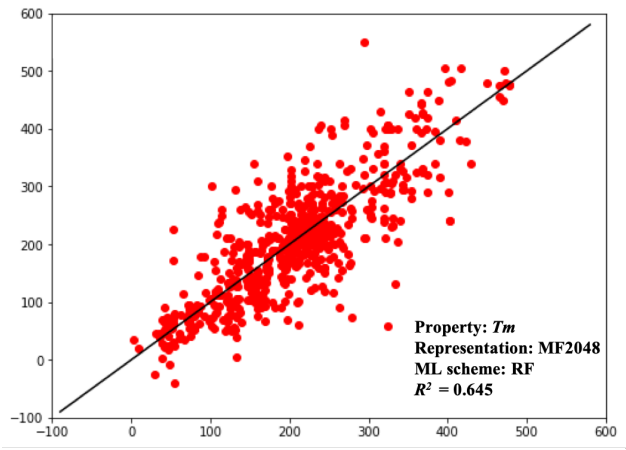
Regressor	Metrics	Morgan Fingerprint (MF)		Molecular Embedding (ME)		
		Length of 1024	Length of 2048	SLME	SSLME	TLME
MLP	R^2	0.516±0.150	0.542±0.092	0.597±0.050	0.599±0.087	0.624±0.054
	MSE	4940.075±1087.783	4714.405±637.154	4199.681±575.565	4118.422±658.782	3885.808±386.265
	MAE	48.902±4.896	48.156±3.737	48.612±3.841	48.004±4.631	45.806±2.745
RF	R^2	0.645±0.079	0.645±0.075	0.662±0.059	**0.684±0.054**	0.681±0.055
	MSE	3618.983±386.075	3633.643±379.554	3489.120±447.132	**3268.919±468.430**	3329.932±659.739
	MAE	45.072±2.744	45.221±2.753	42.124±3.178	**40.464±2.652**	41.146±4.948
SVM	R^2	0.597±0.097	0.617±0.085	0.651±0.069	0.651±0.074	0.676±0.051
	MSE	4121.232±609.482	3934.450±577.552	3604.422±628.983	3597.295±641.081	3360.045±508.099
	MAE	46.864±3.988	45.779±3.901	42.731±3.726	42.361±3.835	41.001±3.196

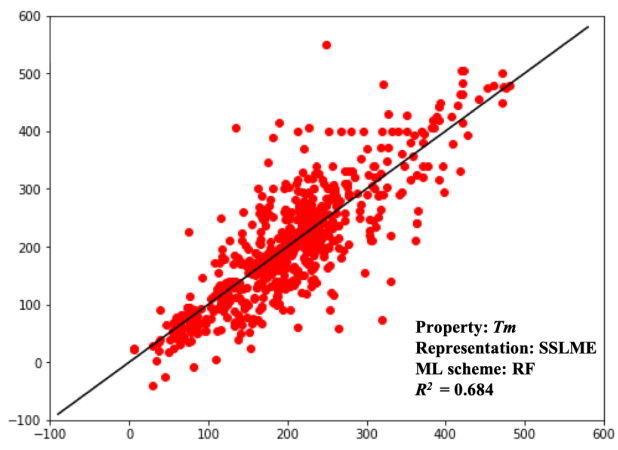
Table 3. Quantitative structure-property relationships between MF, ME and glass transition temperature (T_g).

Regressor	Metrics	Morgan Fingerprint (MF)		Molecular Embedding (ME)		
		Length of 1024	Length of 2048	SLME	SSLME	TLME
MLP	R^2	0.751±0.053	0.807±0.032	0.819±0.023	0.832±0.028	0.827±0.027
	MSE	3100.174±558.6112	2431.716±461.369	2270.717±239.209	2117.769±388.884	2164.443±237.657
	MAE	36.794±2.520	34.573±3.388	34.018±0.947	31.833±3.002	31.985±1.735
RF	R^2	0.848±0.031	0.849±0.033	0.863±0.034	0.865±0.030	0.861±0.033
	MSE	1910.702±412.941	1904.136±436.917	1722.556±475.903	1709.877±445.498	1752.278±454.228
	MAE	30.255±2.621	30.193±2.816	28.284±3.218	28.012±3.300	28.188±2.848
SVM	R^2	0.820±0.035	0.832±0.025	0.858±0.034	0.863±0.031	**0.865±0.026**
	MSE	2261.821±434.207	2105.846±280.078	1799.059±507.899	1737.997±460.278	**1699.999±355.951**
	MAE	34.218±3.284	33.096±2.790	28.297±3.145	28.183±3.279	**27.493±2.752**


Property: ρ
Representation: MF1024
ML scheme: RF
R^2 = 0.562


Property: ρ
Representation: TLME
ML scheme: MLP
R^2 = 0.724


Property: Tm
Representation: MF2048
ML scheme: RF
R^2 = 0.645


Property: Tm
Representation: SSLME
ML scheme: RF
R^2 = 0.684

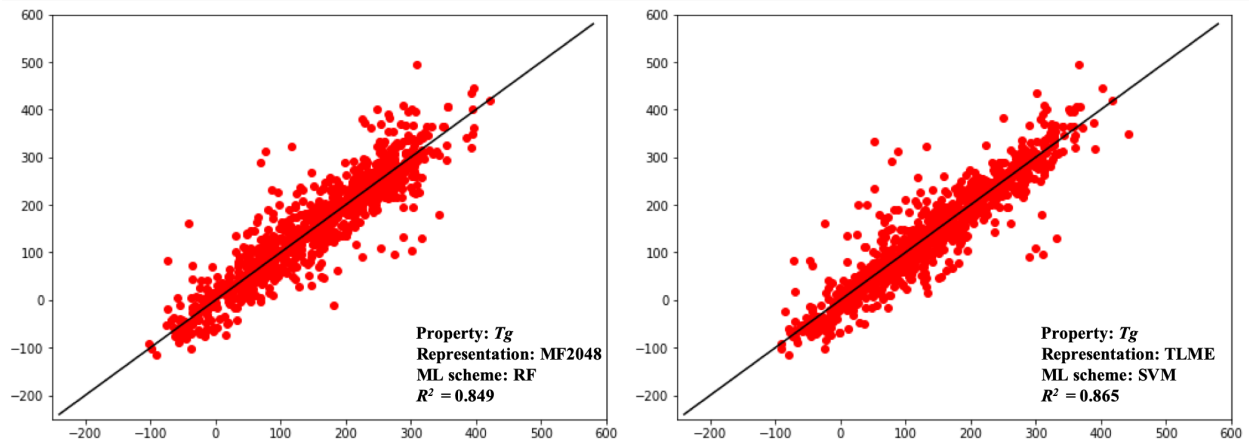


Figure 4. "Predictions vs. ground truths" plots on the validation sets for all three properties, the data for the plots correspond to the best performance of MF and ME. X-axes are for the predictions and y-axes are for ground truths.

To evaluate the performance of MG, we also quantify the same structure-property relationships. The package that we adopt to encode the MG and quantify the structure-property relationships are implemented by Ref. [20]. The regression method used is deep neural network (DNN) and the results in **Table 4** are produced after the hyper-parameter optimization for the DNN. Based on the results, MG shows a much inferior performance compared to MF and ME. The potential reasons can be the following three: 1) the molecular tensor used to encode MG are in high dimension and sparse even for such small molecule like propane in **Figure 2c**. For our two-monomer polymer structures, which are long-chain molecules in nature, the sizes of molecule tensors are even larger (32 × 8) and highly sparse, which will lead to low signal-to-noise ratio; 2) The structure-property relationship quantifying process is based on deep learning method, which highly depends on the volume of training data, thus several hundred to one thousand data may not be enough for obtaining accurate results here; 3) It is challenging to find the global optimal in DNN, which can be another attribute to the inferior performance of MG.

Table 4. The quantitative structure-property relationships between MG and different properties.

MG				
Regressor	Metrics	ρ	T_m	T_g
DNN	R^2	0.260±0.229	-0.149±0.235	0.711±0.017
	MSE	0.017±0.003	9560.946±2015.798	3566.943±605.063
	MAE	0.102±0.012	79.561±7.339	45.440±2.469

Furthermore, we study the ME under different representation learning schemes, including supervised learning (SL), semi-supervised learning (SSL) and transfer learning (TL). We name the representations learned under supervised learning SLME, that learned under semi-supervised learning SSLME, and that learned under transfer learning TLME. To quantify the relationships between SLME and properties, 318 polymers labeled with densities are used for obtaining the SLME in **Table 1**, 641 polymers labeled with melting temperatures are used for obtaining the SLME in **Table 2**, and 1034 polymers labeled with glass transition temperatures are used for obtaining the SLME in **Table 3**. To quantify the relationships between SSLME and properties, all 1442 polymers are used for training the SSLMEs shown in all three tables. For the relationships between TLME and properties, we leverage 20 million organic molecular structures from the ZINC version 15[46] and the ChEMBL version 23[47, 48] databases to train the TLMEs in all three tables. The best performance scores are marked in bold in **Tables 1, 2** and **3**, which all happen in SSLME or TLME. Comparing the ME's best performances (bold numbers) in **Tables 1** and **3**, the improved performance in **Table 3** can be attributed to the increased training samples. However, a counterintuitive case is shown in **Table 2**. Even though the number of training samples in **Table 2** is more than that in **Table 1**, the performance is degraded. A potential explanation could be that when quantifying the relationships between polymer structures and fundamental properties such as density, no additional information needs to be specified; but for those application properties like melting temperature, who is sensitive to the measurement methods, additional information besides ME, such as the measurement method, may need to be added to the representation in order to accurately quantify the structure-property relationships.[49]

To investigate the difference in MEs of substructures under different representation learning schemes, we find a way to visualize the MEs of substructures. We use the t-SNE technique[50] to project the three sets (each from a learning scheme) of 268 MEs, whose corresponding substructures are shared by the training molecules for obtaining SLME, SSLME and TLME used in **Table 1**, into a 2D space, and we name them T1-SLME-268, T1-SSLME-268 and T1-TLME-268, respectively. In this way we can observe the local neighbor pattern for MEs of substructures. In **Figure 5a, b** and **c**, we can visually see that the local neighbor patterns of T1-SLME-268, T1-SSLME-268 and T1-TLME-268 are very different. To be more quantitative, we calculate the one-to-the-rest distance for all the MEs of substructures. In this calculation, we randomly pick a ME in the space, calculate the distances between this ME and all the rest ones, and then sum up all the calculated pair-wise distances as the total distance for this ME. We repeat this calculation for all the MEs and then plot the distribution of the total distances in **Figure 5d.** Based on **Figure 5**, the local-neighbor-patterns for MEs and their distributions of total distances change with different representation learning schemes, thus leading to different similarity estimation in MEs under different learning schemes. For the essence of structure-property relations, similar structures should have similar properties. However, if the similarities between structures are differently estimated among different learning schemes, the quantitative structure-property relationship will be different, thus leading to difference performance scores as seen in **Tables 1-3**.

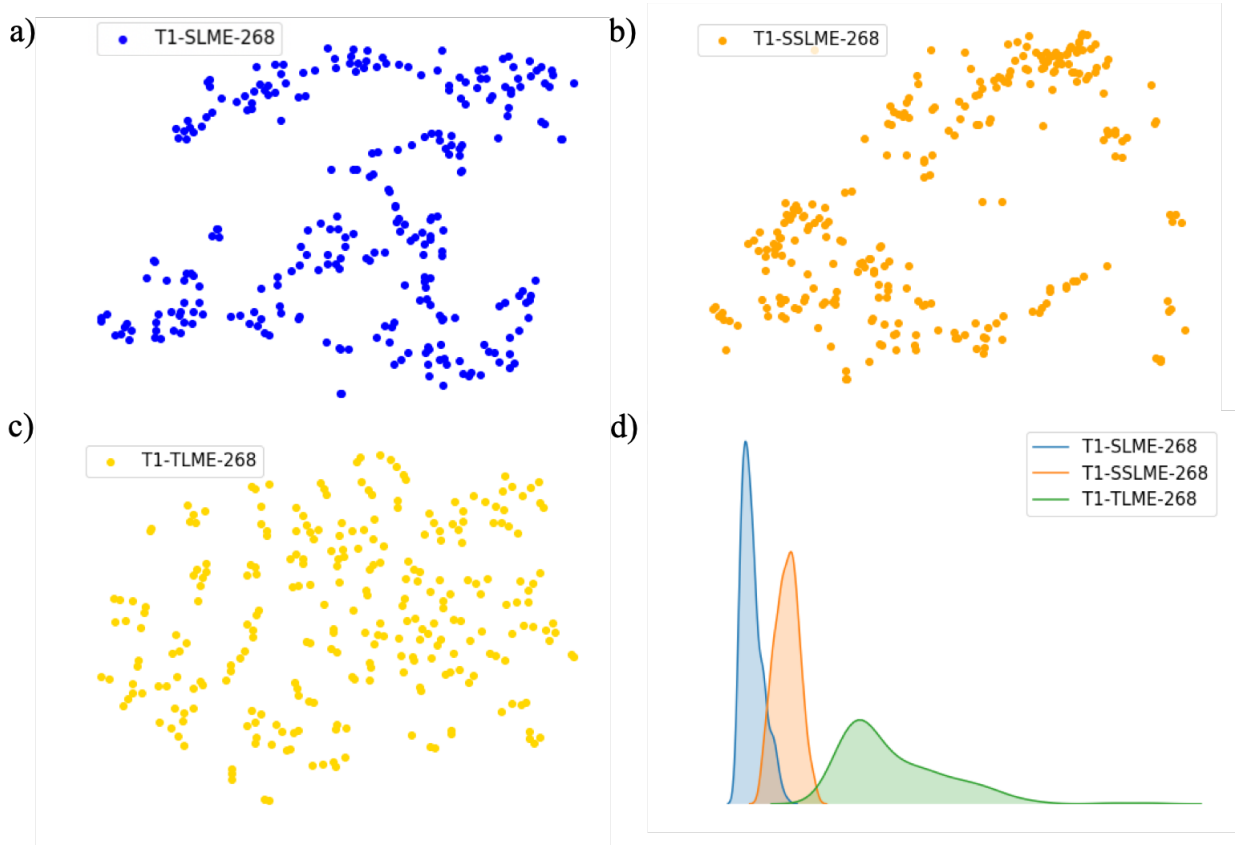


Figure 5. 2-D t-SNE visualization of a) T1-SLME-268, b) T1-SSLME-268 and c) T1-TLME-268, the number of data points in each graph are 268, which corresponds to the common substructures shared by the training molecules for obtaining SLME, SSLME and TLME used in **Table 1**; d) The distribution of total pair-wise distances of T1-SLME-268, T1-SSLME-268 and T1-TLME-268.

Finally, inspired by the idea of mixing word embeddings to capture more semantic information in language learning,[51] we mix our MEs here to see whether more information for estimating the substructure similarities can be captured in comparison to the strongest single MEs. We conduct the ME mixing in two ways: concatenation and weighted sum, which are schematically shown in **Figure 6**. In the concatenation mixing, two MEs are conjugated into one. In the weighted sum mixing, the weights ω_i is calculated as:

$$\omega_i = R_i^2 / \sum_j R_j^2 \quad (1)$$

where *i, j* is the type of ME (e.g., SLME, SSLME or TLME) and $R^2_{i,j}$ is the corresponding highest mean coefficient of determination (R^2) as shown in **Tables 1-3**. For example, when mixing SLME and SSLME according to the R^2 in **Table 1**, $\omega_{SLME} = 0.668/(0.668 + 0.701)$ for the SLME term. The weighted sum ME, ν_m, is then calculated as:

$$\nu_m = \sum_j \omega_j \cdot \nu_j \tag{2}$$

where *m* is the type of mixing (e.g., SLME + SSLME), *j* is the type of ME (e.g., SLME). Thus, the weighted sum ME of SLME and SSLME is calculated as $\nu_{SLME+SSLME} = \omega_{SLME} \cdot \nu_{SLME} + \omega_{SSLME} \cdot \nu_{SSLME}$.

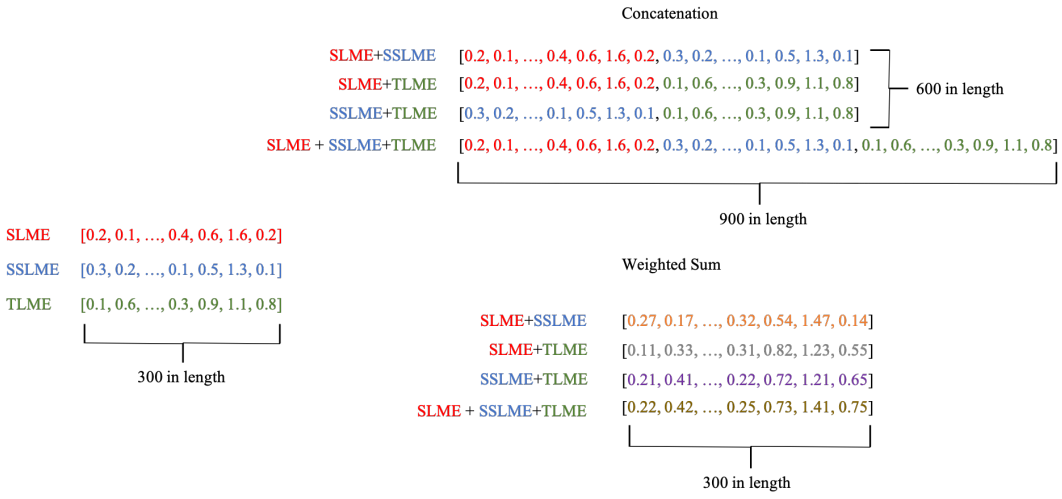


Figure 6. The schematics of mixed ME. Two different mixing schemes are studied: concatenation and weighed sum.

We calculate the coefficient of determination (R^2) in a 5-fold cross-validation manner when quantifying the relationships between different ME mixtures and properties, and compare the results to the R^2 of the strongest single MEs in **Tables 1**, **2** and **3**. Based on the results in **Table 5**, several ME mixtures, which are marked in bold, outperform their corresponding strongest single MEs in quantifying the structure-property relationships. Such improvements, even though not very significant, may be attributed to the information gain when mixing different MEs.

Table 5. Quantitative structure-property relationships between different ME mixtures and properties.

Label	Regressor	Molecular Embedding (ME)								Strongest Single ME
		Concatenation				Weighted Sum				
		SLME+SSLME	SLME+TLME	SSLME+TLME	SLME+SSLME+TLME	SLME+SSLME	SLME+TLME	SSLME+TLME	SLME+SSLME+TLME	
ρ	MLP	0.685±0.044	0.706±0.094	0.692±0.083	0.651±0.140	0.644±0.104	0.683±0.116	0.717±0.066	0.708±0.068	0.724±0.068
	RF	0.705±0.095	0.663±0.124	0.694±0.105	0.700±0.106	0.664±0.097	0.661±0.122	0.691±0.113	0.667±0.103	
	SVM	0.684±0.078	0.719±0.098	0.712±0.085	0.714±0.082	0.671±0.085	**0.733±0.097**	0.698±0.092	0.711±0.089	
Tm	MLP	0.560±0.135	0.577±0.083	0.598±0.063	0.525±0.089	0.612±0.086	0.624±0.077	0.626±0.079	0.617±0.071	0.684±0.054
	RF	0.683±0.050	0.681±0.053	**0.689±0.047**	**0.689±0.047**	0.671±0.063	**0.690±0.051**	0.680±0.062	0.671±0.062	
	SVM	0.663±0.073	0.679±0.062	0.680±0.060	0.680±0.066	0.666±0.072	0.681±0.062	0.680±0.056	0.676±0.064	
Tg	MLP	0.816±0.028	0.833±0.022	0.833±0.019	0.822±0.040	0.825±0.022	0.831±0.038	0.832±0.038	0.843±0.023	0.865±0.026
	RF	**0.867±0.030**	**0.867±0.032**	**0.869±0.032**	**0.870±0.031**	0.853±0.038	0.863±0.034	0.861±0.031	0.856±0.036	
	SVM	0.860±0.034	**0.867±0.028**	**0.867±0.027**	**0.867±0.029**	0.861±0.035	**0.866±0.026**	**0.868±0.030**	**0.866±0.029**	

CONCLUSION

In summary, we have explored different representations for polymers and evaluated their quality by quantifying the relationships between them and three different polymer properties. It is found that ME, which carrying more information about substructures, outperforms the commonly used MF as the polymer representation, as indicated by better performance scores from the regression of the structure-property relationships. On the other hand, the MG representation is shown to be much inferior than ME and MF in the polymer structure-property relationship quantification, likely due to the relatively small volumes of training data available. For MEs, it is found that the similarities of substructure MEs under different learning schemes (e.g., SL, SSL and TL) are differently estimated, thus leading to different performance scores in structure-property relation quantification. Several ME mixtures have shown to outperform the single MEs in the corresponding regression tasks, and this is attributed to the information gain when mixing different ME.

ACKNOWLEDGEMENT

The authors acknowledge the financial support from the DuPont Young Professor Award program. T.L. would also like to thank the Dorini Family for the endowed professorship in energy studies. The computation is supported in part by the University of Notre Dame, Center for Researching Computing, and NSF through XSEDE resources provided by TACC Stampede II under a grant number TG-CTS100078.

REFERENCE

1. Marx, V., *Biology: The big challenges of big data.* 2013.
2. Jones, N., *Computer science: The learning machines.* Nature News, 2014. **505**(7482).
3. Jordan, M.I., and Tom M. Mitchell, *Machine learning: Trends, perspectives, and prospects.* Science, 2015. **349**(6245): p. 255-260.
4. Hinton, G., et al, *Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups.* IEEE Signal processing magazine, 2012. **29**(6): p. 82-97.
5. Kell, D.B., *Metabolomics, modelling and machine learning in systems biology; understanding complex systems using genetic programming to produce simple interpretable rules. The Theodor Bücher Lecture and Medal: P3-001.* The Febs Journal, 2005. **272**(2).
6. Parkes, D.C., and Michael P. Wellman, *Economic reasoning and artificial intelligence.* Science, 2015. **349**(6245): p. 267-272.
7. Ramprasad, R., et al, *Machine learning in materials informatics: recent applications and prospects.* npj Computational Materials 2017. **3**(1).
8. Fischer, C.C., et al., *Predicting crystal structure by merging data mining with quantum mechanics.* Nature materials, 2006. **5**(8).
9. Fujimura, K., et al, *Accelerated Materials Design of Lithium Superionic Conductors Based on First-Principles Calculations and Machine Learning Algorithms.* Advanced Energy Materials, 2013. **3**(8): p. 980-985.
10. Zhan, T., Lei Fang, and Yibin Xu, *Prediction of thermal boundary resistance by the machine learning method.* Scientific reports, 2017. **7**(1).

11. Blay, V., Toshiyuki Yokoi, and Humbert González-Díaz, *Perturbation Theory–Machine Learning Study of Zeolite Materials Desilication.* Journal of chemical information and modeling, 2018. **58**(12): p. 2414-2419.

12. Shi, L., et al., *Using Data Mining To Search for Perovskite Materials with Higher Specific Surface Area.* Journal of chemical information and modeling, 2018. **58**(12): p. 2420-2427.

13. Hachmann, J., et al., *The Harvard Clean Energy Project: Large-Scale Computational Screening and Design of Organic Photovoltaics on the World Community Grid.* The Journal of Physical Chemistry Letters, 2011. **2**(17): p. 2241-2251.

14. Bengio, Y., A. Courville, and P. Vincent, *Representation learning: a review and new perspectives.* IEEE Trans Pattern Anal Mach Intell, 2013. **35**(8): p. 1798-828.

15. Hinton, G.E., *Learning distributed representations of concepts.* Proceedings of the eighth annual conference of the cognitive science society, 1986. **1**.

16. Simonyan, K., and Andrew Zisserman, *Very deep convolutional networks for large-scale image recognition.* arXiv preprint arXiv:1409.1556, 2014.

17. Blumensath, T., and Mike Davies, *Sparse and shift-invariant representations of music.* IEEE Transactions on Audio, Speech, and Language Processing, 2006. **14**(1): p. 50-57.

18. Schütt, K.T., et al, *Quantum-chemical insights from deep tensor neural networks.* Nature communications, 2017. **8**.

19. Gomez-Bombarelli, R., et al., *Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules.* ACS Cent Sci, 2018. **4**(2): p. 268-276.

20. Coley, C.W., et al, *Convolutional embedding of attributed molecular graphs for physical property prediction.* Journal of chemical information and modeling, 2017. **57**(8): p. 1757-1772.

21. Jaeger, S., S. Fulle, and S. Turk, *Mol2vec: Unsupervised Machine Learning Approach with Chemical Intuition.* J Chem Inf Model, 2018. **58**(1): p. 27-35.

22. Altae-Tran, H., et al, *Low data drug discovery with one-shot learning.* ACS central science, 2017. **3**(4): p. 283-293.

23. Mannodi-Kanakkithodi, A., et al., *Machine Learning Strategy for Accelerated Design of Polymer Dielectrics.* Sci Rep, 2016. **6**: p. 20952.

24. Mannodi-Kanakkithodi, A., G. Pilania, and R. Ramprasad, *Critical assessment of regression-based machine learning methods for polymer dielectrics.* Computational Materials Science, 2016. **125**: p. 123-135.

25. Jabeen, F., et al., *Refractive indices of diverse data set of polymers: A computational QSPR based study.* Computational Materials Science, 2017. **137**: p. 215-224.

26. Cravero, F., et al., *Feature Learning applied to the Estimation of Tensile Strength at Break in Polymeric Material Design.* J Integr Bioinform, 2016. **13**(2): p. 286.

27. Sumpter, B.G., and Donald W. Noid, *Neural networks and graph theory as computational tools for predicting polymer properties.* Macromolecular theory and simulations, 1995. **3**(2): p. 363-378.

28. Otsuka, S., et al, *PoLyInfo: Polymer database for polymeric materials design.* Emerging Intelligent Data and Web Technologies (EIDWT), 2011 International Conference on. IEEE, 2011.

29. Ma, R., et al., *Determining influential descriptors for polymer chain conformation based on empirical force-fields and molecular dynamics simulations.* Chemical Physics Letters, 2018. **704**: p. 49-54.

30. Zhang, T. and T. Luo, *Role of Chain Morphology and Stiffness in Thermal Conductivity of Amorphous Polymers.* J Phys Chem B, 2016. **120**(4): p. 803-12.

31. Rogers, D., and Mathew Hahn, *Extended-connectivity fingerprints.* Journal of chemical information and modeling, 2010. **50**(5): p. 742-754.

32. Riniker, S., and Gregory A. Landrum, *Open-source platform to benchmark fingerprints for ligand-based virtual screening.* Journal of cheminformatics, 2013. **5**(1).

33. O'Boyle, N.M. and R.A. Sayle, *Comparing structural fingerprints using a literature-based similarity benchmark.* J Cheminform, 2016. **8**: p. 36.

34. Riniker, S., N. Fechner, and G.A. Landrum, *Heterogeneous classifier fusion for ligand-based virtual screening: or, how decision making by committee can be a good thing.* J Chem Inf Model, 2013. **53**(11): p. 2829-36.

35. Mayr, A., et al., *DeepTox: Toxicity Prediction using Deep Learning.* Frontiers in Environmental Science, 2016. **3**.

36. Merget, B., et al., *Profiling Prediction of Kinase Inhibitors: Toward the Virtual Assay.* J Med Chem, 2017. **60**(1): p. 474-485.

37. Sorgenfrei, F.A., S. Fulle, and B. Merget, *Kinome-Wide Profiling Prediction of Small Molecules.* ChemMedChem, 2018. **13**(6): p. 495-499.

38. Landrum, G., *RDKit: Open-Source Cheminformatics Software.(2016).* URL http://www.rdkit.org/, https://github.com/rdkit/rdkit, 2016.

39. Pedregosa, F., et al, *Scikit-learn: Machine learning in Python.* Journal of machine learning research, 2011. **12**(10): p. 2825-2830.

40. Breiman, L., *Random forests.* Machine Learning, 2001. **45**(1): p. 5-32.

41. Geurts, P., Damien Ernst, and Louis Wehenkel, *Extremely randomized trees.* Machine learning, 2006. **63**(1): p. 3-42.

42. Hinton, G.E., *Connectionist learning procedures.* Machine learning 1990. **Morgan Kaufmann**: p. 555-610.

43. Nair, V., and Geoffrey E. Hinton, *Rectified linear units improve restricted boltzmann machines.* Proceedings of the 27th international conference on machine learning, 2010. **ICML**(10).

44. Kingma, D.P., and Jimmy Ba, *Adam: A method for stochastic optimization.* arXiv preprint arXiv:1412.6980, 2014.

45. Chang, C.-C., and Chih-Jen Lin, *LIBSVM: a library for support vector machines.* ACM transactions on intelligent systems and technology (TIST), 2011. **2**(3).

46. Irwin, J.J., et al, *ZINC: a free tool to discover chemistry for biology.* Journal of chemical information and modeling, 2012. **52**(7): p. 1757-1768.

47. Gaulton, A., et al, *ChEMBL: a large-scale bioactivity database for drug discovery.* Nucleic acids research, 2011. **40**(D1): p. D1100-D1107.

48. Bento, A.P., et al, *The ChEMBL bioactivity database: an update.* Nucleic acids research, 2014. **42**(D1): p. D1083-D1090.

49. Audus, D.J., and Juan J. de Pablo, *Polymer informatics: opportunities and challenges.* 2017: p. 1078-1082.

50. Maaten, L.v.d., and Geoffrey Hinton, *Visualizing data using t-SNE.* Journal of machine learning research, 9.Nov 2008: p. 2579-2605.

51. Li, J., et al., *Learning distributed word representation with multi-contextual mixed embedding.* Knowledge-Based Systems, 2016. **106**: p. 220-230.