Guanghui Zhu, Chiho Kim, Anand Chandrasekarn, Joshua D. Everett, Rampi Ramprasad* and Ryan P. Lively*

Polymer genome–based prediction of gas permeabilities in polymers

https://doi.org/10.1515/polyeng-2019-0329
Received November 8, 2019; accepted December 17, 2019

Abstract: Predicting gas permeabilities of polymers a priori is a long-standing challenge within the membrane research community that has important applications for membrane process design and ultimately widespread adoption of membrane technology. From early attempts based on free volume and cohesive energy to more recent group contribution methods, the ability to predict membrane permeability has improved in terms of accuracy. However, these models usually stay "within the paper", i.e. limited model details are provided to the wider community such that adoption of these predictive platforms is limited. In this work, we combined an advanced polymer chemical structure fingerprinting method with a large experimental database of gas permeabilities to provide unprecedented prediction precision over a large range of polymer classes. No prior knowledge of the polymer is needed for the prediction other than the repeating unit chemical formula. In addition, we have incorporated this model into the existing Polymer Genome project to make it open to the membrane research community.

Keywords: gas separation; machine learning; membrane.

1 Introduction

Membrane-based gas separation technologies and promising platforms are emerging and gradually being integrated into industries such as fine chemical, pharmaceutical, and petroleum production [1–4]. The demand for high-performance polymer membranes requires new polymer material discovery and process innovations to provide solutions to different types of separation tasks. Although advanced processing technologies such as hollow fiber spinning are pushing the limit of productivity and economics, the membrane performance is ultimately bound by the properties of the polymers [5–8].

There are two intrinsic material properties related to the performance of gas separation membranes: gas permeability and permselectivities between gas pairs. Robeson [9, 10] reported a trade-off between these two properties (known as "the polymer upper bound"). The polymer upper bound dictates that high permeability polymers usually possess reduced selectivity and *vice versa*. However, this upper bound has not proven to be a hard boundary and continued research into membrane materials has resulted in many re-evaluations of the location of the polymer upper bound. Designing membrane materials and processes, especially for novel gas separation challenges, requires utilizing or creating polymers that have the desired combinations of permeability and permselectivity. Currently, it is possible to make initial guesses on polymer permeability based on the type of polymer and empirical knowledge; however, it is challenging to predict gas selectivity by intuition alone. Moreover, establishing a data-driven relationship between polymer structure and polymer membrane performance will accelerate artificial intelligence–based approaches focused on designing custom polymers for specific separation tasks.

Early works in permeability prediction focused on correlating gas permeability to three empirical factors: F, which depends on the nature of the polymer; G, which depends on the nature of the gas; and γ, which represents the interactions between polymer and gas [11]. This method provided a rough correlation for permeabilities. However, the F and G values are empirically estimated and importantly are not available for new polymers. Another study by Dow Chemicals found that CO_2 and O_2 permeabilities correlated with specific free volume of polymers (V_f), which can be experimentally measured [12]. The author of this work suggested several functional groups that will provide small permeabilities suitable for barrier design based on experience, which raises the idea

***Corresponding authors: Rampi Ramprasad**, Georgia Institute of Technology, 771 Ferst Drive, Northwest Atlanta, Atlanta, GA 30332, USA, e-mail: rampi.ramprasad@mse.gatech.edu; and **Ryan P. Lively,** Georgia Institute of Technology, 311 Ferst Drive, Northwest Atlanta, Atlanta, GA 30332, USA, e-mail: ryan.lively@chbe.gatech.edu. https://orcid.org/0000-0002-8039-4008
Guanghui Zhu and Joshua D. Everett: Georgia Institute of Technology, 311 Ferst Drive, Northwest Atlanta, Atlanta, GA 30332, USA
Chiho Kim and Anand Chandrasekarn: Georgia Institute of Technology, 771 Ferst Drive, Northwest Atlanta, Atlanta, GA 30332, USA

for correlating permeation properties to polymer building blocks. Several studies added cohesive energy density (E_{coh}) of the polymer to the correlation. Salame [13] devised and assigned Permachor values for backbone segments based on free volume and cohesive energy contributions. The Permachor values of each segment were then averaged and used to predict gas permeability. In another study, it was found that there is a linear relationship between log of permeability with V_f/E_{coh} [14].

Current materials discovery is in general based on "chemical intuition" or high-throughput screening. Several steps are needed to assess gas permeation properties for novel polymer materials in this process, including polymer synthesis, membrane fabrication, and permeation testing. The first two steps require substantial optimization for each polymer before permeation experiments can be conducted. This initial intensive investment hinders high-throughput research focused on discovering new polymers for permeation applications, and more focus has thus been placed on process optimization with commercially available polymers. The ability to predict the separation capability of a new polymer with minimal information input thus has the potential to accelerate research in the area of membrane-based gas separations.

Machine learning (ML) methods have been utilized in chemistry and materials discovery with high levels of success. ML techniques applied to chemistry have resulted in synthesis outcome prediction with text extraction from the literature [15], and retrosynthetic routes design [16], and polymers among others. Recently, Ramprasad et al. [17] developed Polymer Genome, which is an expanding tool that uses ML methods to model various polymer properties according to the fingerprints of the polymer. With the input of just the polymer chemical formula [in terms of a simplified molecular input line entry system (SMILES) string], the suite of algorithms can predict properties including glass transition, dielectric properties, solvent interactions, refractive index, among others [18–26].

Machine learning has been applied to predict the gas permeability of polymers. In 1994, Wessling et al. reported a method to map polymer permeability onto polymer infrared spectra using a neural network approach [27]. This interesting trial showed only limited correlations between predicted and experimental values but was nevertheless the first known attempt to develop an ML-based approach to polymer membrane performance prediction. Another approach is to use a group contribution method by dividing the polymer chemical structure into small fragments. Park and Paul [28] used the group contribution method to calculate the fractional free volume (FFV) and density of polymers from a database of 102 polymers and then used these values to calculate gas permeability. Robeson et al. [29] proposed another group contribution approach that avoids the calculation of FFV and density by correlating gas permeability to the permeability contribution of the structural units. Normalization with molar volume was applied to compensate for size differences in repeating units. Yampolskii et al. [30] further evaluated the effect of normalization method (no normalization, and normalization by the number of atoms and molar volume.) based on a database of 300 polymers. The same group descriptors were used by Hasnaoui et al. [31] to build artificial neural network model for prediction of polymer permeability, and the results were compared with Park and Paul, Permachor, and Yampolskii's methods. Ryzhikh et al. [32] reported a comparison between the correlation used atomic contributions and bond contribution methods for the prediction of the gas transport parameters (permeability [P] and diffusion [D] coefficients) of 900 amorphous glassy polymers. To develop a more general and implementable model for permeability prediction, we aim to build on the bases of Polymer Genome and develop a universal prediction model that covers a wide range of polymer classes. The model will be integrated into the current Polymer Genome platform and will be available to other researchers.

2 Materials and methods

2.1 Data sets

In this work, we used experimentally collected gas permeability data for six different gases (CH_4, CO_2, H_2, He, N_2, and O_2) from more than 300 publications with more than 1000 entries [10]. The testing conditions vary across the literature with most of the tests executed at 25°C to 35°C and 1–10 atm on dense membranes with a thickness of 10–200 μm. This extensive collection of experimental gas permeability data enables us to construct a universal model for polymers spanning across polyimides, polyamides, poly(amide-imides), polyarylates, polycarbonates, polyesters, polypyrrolones, polynorbornenes, vinyl and vinylidene polymers, poly(aryl ethers) and poly(aryl ether ketones), polyphosphazenes, perfluorinated polymers, polysulfones, parylenes, high-temperature polymers, polypropynes, substituted polyacetylenes, and polypentynes. A total of 315 polymers were considered in this work of which the number of permeability of measurements recorded for CH_4, CO_2, H_2, He, N_2, and O_2 was 258, 293, 174, 183, 293, and 300, respectively. This collection represents a total of 1501 permeability measurements.

One of the benefits of using experimental data is that hidden information on polymer processing and testing is integrated into the data, which will be used in model regression. In this work, the model input only contains a chemical structure. Other factors that will affect polymer performance, such as polymer process history and testing method, are not explicitly used as parameters in the model. However, because experimental data in the database were generated under realistic processing and testing methods, the effect of these hidden factors will be correlated in ML. This process also indicates that the model we generated from ML will predict an experimental value of the polymer permeability that is expected with current "averaged" processing and testing method.

2.2 Polymer fingerprinting

Gas permeation properties of a polymer are related to complex contributions that can be broadly categorized into physical and chemical interactions. Chemical interactions will affect both diffusion and sorption of guest molecules in the polymer, which will be captured by the building blocks of the polymer. Physical interactions originate from the contributions of free volume and polymer chain arrangement that will affect the diffusion of guest molecules in the polymer. We believe that the packing of the polymer affected by the guest molecules can be captured by using experimental data and higher level descriptors such as morphological descriptors.

A hierarchical polymer fingerprinting scheme was used to comprehensively capture the descriptors that may control the gas permeability of polymers [23]. Briefly, the fingerprint building process consists of four hierarchical levels of descriptors. The first level is at the atomic scale wherein the occurrence of atomic triples, which is a set of three contiguous atoms (e.g. C2–C3–C4, made up of a twofold coordinated oxygen, a threefold coordinated carbon, and a fourfold coordinated carbon), was calculated [33]. For the 315 polymers considered in this study, there are 93 such components. The second set of fingerprint components captures a population of predefined chemical building blocks (e.g. $-C_6H_4-$, $-CH_2-$, $-C(=O)-$). We include a fixed set of 148 block level components. The third hierarchical level deals with quantitative structure-property relationship descriptors, such as van der Waals surface area [34], topological surface area [35], and the fraction of rotatable bonds, implemented in the RDKit cheminformatics library [36]. Such descriptors, 39 in total, form the next set of components of the overall fingerprint. To these third level descriptors, morphological features such as the topological distance between rings, fraction of atoms that are part of side chains and length of the largest side chain were added [17]. We include a fixed set of 19 such morphological descriptors.

As a result of the fingerprinting method, some limitations on polymer structures that can be included and predicted by the model were introduced. As the fingerprinting method digests a single repeating unit of the polymer, polymers with structures that cannot be represented by a single repeating unit with single end points are not included (i.e. polymer blends, random copolymers, random substituted polymers, ladder polymers, among others). Chemical bonds across repeating units are also limited (i.e. crosslinked polymer, hyperbranched polymer, thermal rearranged polymer, among others).

2.3 Machine learning model development

In this work, we utilize Gaussian process regression, which generates a probabilistic surrogate model of a specific property. We used a radial basis function kernel defined as

$$k(\mathbf{x}_i, \mathbf{x}_j) = \sigma^2 \exp\left[\frac{-(\mathbf{x}_i - \mathbf{x}_j)^2}{2l^2}\right] + \sigma_n^2 \delta(\mathbf{x}_i, \mathbf{x}_j)$$

where σ, l, and σ_n are hyperparameters to be determined during the training process (in the ML parlance, these hyperparameters are referred to as signal variance, length scale parameter, and noise level parameter, respectively). $\mathbf{x}_i$ and $\mathbf{x}_j$ are the fingerprint vectors for two polymers i and j. Performance of the model was evaluated based on the coefficient of determination (R^2).

3 Results and discussion

3.1 Model performance

In the data collected from the literature, not all six gases (He, H_2, CO_2, O_2, N_2, and CH_4) were measured for every polymer. To increase the number of data available to the model, we assigned a distinct vector to each gas permeability data depending on the type of gas (i.e. one-hot encoding). As a result, a total number of 1501 data points can be used in the model fitting, which greatly improves the goodness of fitting. In addition, by doing this, we can easily predict the missing permeability data in the literature for gases that were not measured. The learning curve

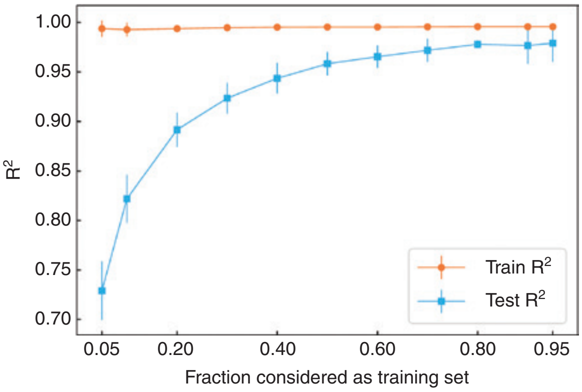

Figure 1: Learning curves constructed from the R^2 of the machine learning models for gas permeability. Total gas permeability datasets of 5%, 10%, 20%, …, 90% and 95% were used for training the model. For each model, the data were obtained from 50 independent runs with a different selection of train and test sets.

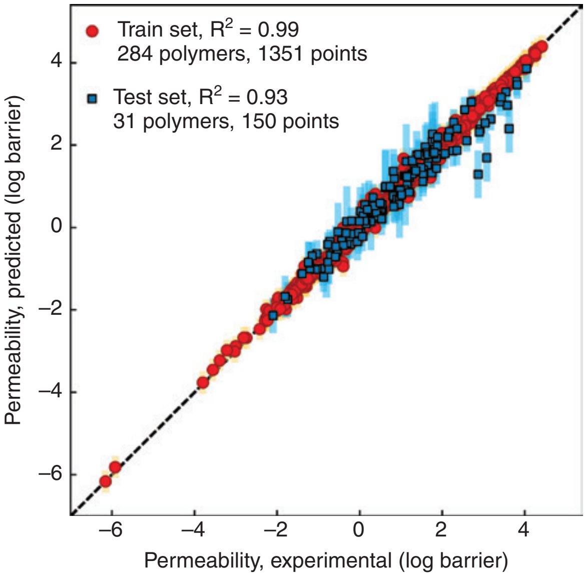

Figure 2: Comparison of predicted and experimental permeability values for 31 polymers that were not used for model training.

(Figure 1) shows that the size of the dataset (1501 permeability measurements) is sufficient to train the model as the R^2 of training set and test set converges.

To test the ability for predicting unknown polymers, we randomly selected polymers from each polymer class as the test set and used the remainder of the database as the training set (Supplementary Figures S1 and S2). The polymers selected are listed in Supplementary Table S1. The correlation between the predicted and experimental permeability values for polymers in the test set are shown in Figure 2. Good agreement was achieved in general with several outlier points, which include one case in each of polynorbornenes, vinyl and vinylidene polymers, and parylenes. The outliers can be attributed to either data scarcity for that polymer class during training, or uncertainties within the experimental data.

3.2 Comparison of experimental and calculated permeability coefficients

Figure 3 shows the correlations between experimental and predicted permeability values. The correlation coefficients for each individual gases range from 0.98 to 0.99. This is by far the most accurate prediction model for gas permeabilities based on fitted experimental data. Compared with earlier group contribution and bond contribution methods based on a much larger database (R^2 = 0.9), our fingerprinting method showed better prediction performance [32].

3.3 Analysis of individual polymer classes

The inclusion of different polymer classes in the same model sacrifices accuracy to a certain extent because of the distinct physical properties of different polymer classes. Some polymer classes in the database are presented in larger numbers and thus are more heavily weighted in the model. As a result, the ability to predict permeability from minority polymer classes (e.g. polynorbornenes and polyesters vide infra) can be questionable.

Among the various polymer classes in the database, we picked two over-represented classes and two under-represented classes. As the over-represented classes, one is the combined group of polyimides and polypyrrolones; the other is the combined group of polypropynes, substituted polyacetylenes, and polypentynes. The two groups constitute approximately 50% of the data in the entire database. The two under-represented classes are polynorbornenes and polyesters, which only have six and three distinct structures, respectively. Figure 4A–D shows the parity plot of experimental and predicted permeability values from the four polymer classes. As expected, for the over-represented groups, there is good agreement between the predicted and experimental permeability values. Surprisingly, a similar agreement was observed for polynorbornenes and polyesters. This ability to unify drastically different polymers in one model shows the versatility and transferability characteristics of the Polymer Genome approach, which has proven successful for predicting other polymer properties from sparse data sets.

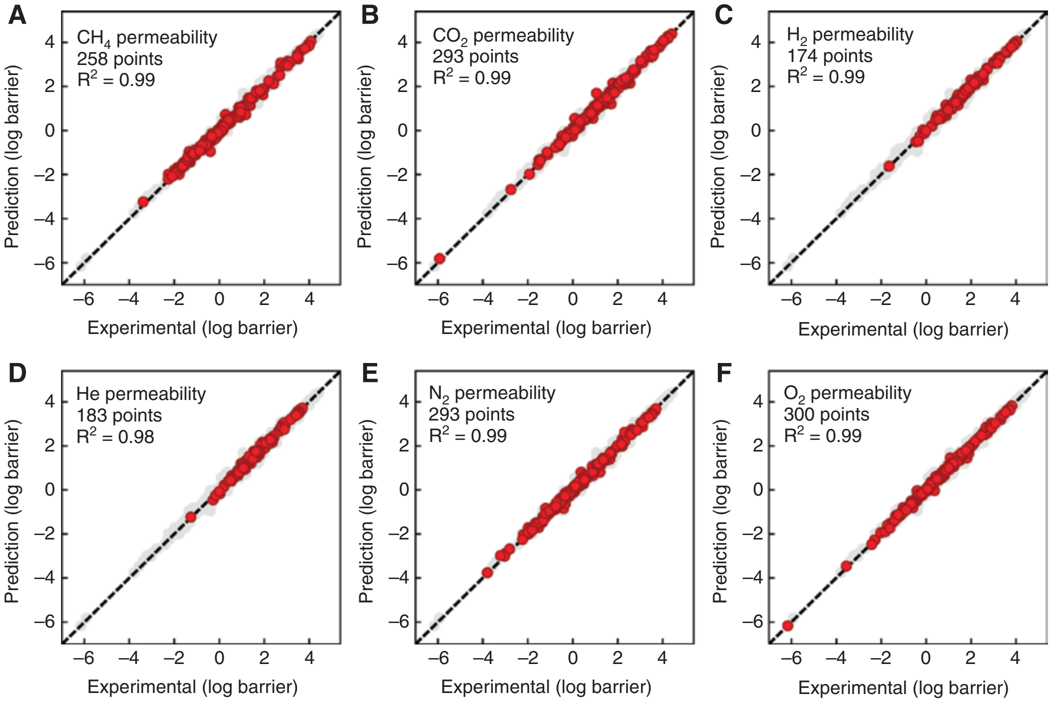

Figure 3: Performance of the permeability prediction model. Parity plots showing comparison between experimental and predicted gas permeability for (A) CH_4, (B) CO_2, (C) H_2, (D) He, (E) N_2, and (F) O_2 are generated using one unified predictive model for six gases trained on 100% dataset (total 1501 permeability data points associated with 315 polymers, illustrated in all figures as gray points).

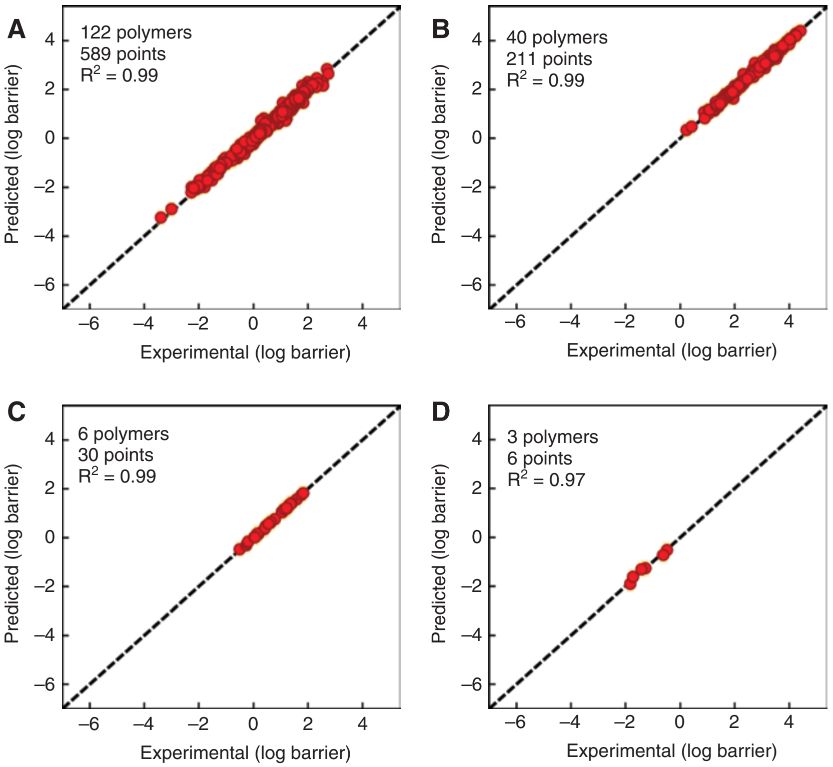

Figure 4: Performance of model to predict permeability for different over- and under-represented polymer classes including (A) polyimides and polypyrrolones, (B) polypropynes, substituted polyacetylenes, polypentynes, (C) polynorbornenes, and (D) polyesters.

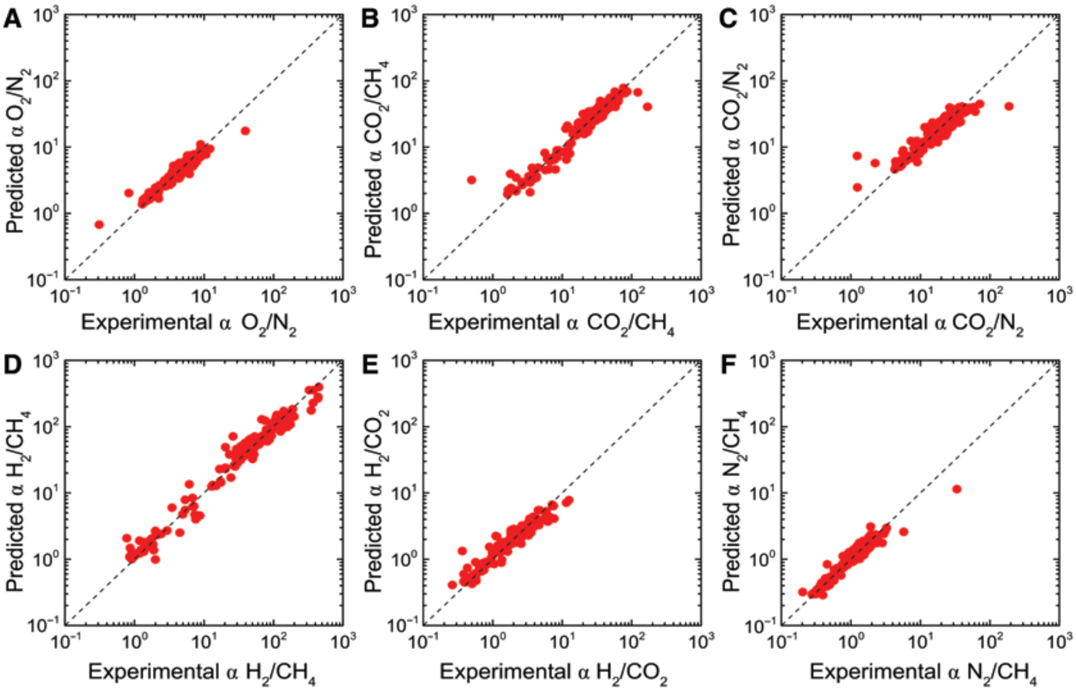

Figure 5: Comparison of experimental and predicted selectivities for (A) O_2/N_2, (B) CO_2/CH_4, (C) CO_2/N_2, (D) H_2/CH_4, (E) H_2/CO_2, and (F) N_2/CH_4.

3.4 Comparison of experimental and predicted selectivities

Although accurately predicting permeabilities is useful, the permselectivity of a potential polymer membrane is often the final arbiter in whether that polymer will go forward in the design process or not. To this end, we calculated the predicted selectivities ($\alpha_{ij} = P_i/P_j$) and compared with experimental selectivities in Figure 5A–F.

We also compared the model results with the 2008 Robeson upper bound (Supplementary Figure S3). It can be seen that the predicted gas pair permeability and selectivity closely resemble the experimental data and the 2009 upperbound [10].

4 Conclusions

We have built a gas permeability prediction model for polymer membranes with high accuracy by using a comprehensive fingerprinting method to represent polymer structures as input into ML algorithms. This model can be applied to a broad range of polymer classes bearing different functional groups. We have also made the prediction results available on the current Polymer Genome platform. The online platform is available at www.polymergenome.org. User inputs into the platform can be the polymer name, common abbreviation, the polymer structure, or a SMILES string. We believe these predictive capabilities will accelerate membrane materials research and membrane process engineering design. In this implemented model, the gas species are considered as a discrete parameter. As future work, we will implement the gas species with chemical descriptors to enable prediction for uncommon gases.

Funding: This work was supported by the Center for Understanding and Control of Acid Gas-Induced Evolution of Materials for Energy (UNCAGE-ME), an Energy Frontier Research Center, funded by U.S. Department of Energy (U.S. DoE), Office of Science, Basic Energy Sciences (BES) under Award Funder Id: http://dx.doi.org/10.13039/100000015, DE-SC0012577. This work was performed in part at the Georgia Tech Institute for Electronics and Nanotechnology, a member of the National Nanotechnology Coordinated Infrastructure, which is supported by the National Science Foundation (Grant ECCS-1542174). The PALS measurements performed at ORNL was supported by the U.S. Department of Energy Office of Fusion Energy Science under grant DE-AC05-00OR22725 with UT-Battelle LLC.

References

[1] Sholl DS, Lively RP. *Nature News* 2016, 532, 435.

[2] Padaki M, Surya Murali R, Abdullah MS, Misdan N, Moslehyani A, Kassim MA, Hilal N, Ismail AF. *Desalination* 2015, 357, 197–207.

[3] Galizia M, Chi WS, Smith ZP, Merkel TC, Baker RW, Freeman BD. *Macromolecules* 2017, 50, 7809–7843.

[4] Khalilpour R, Mumford K, Zhai H, Abbas A, Stevens G, Rubin ES. *J. Clean. Prod.* 2015, 103, 286–300.

[5] Ma C, Koros WJ. *J. Memb. Sci.* 2018, 551, 214–221.

[6] Liang CZ, Yong WF, Chung T-S. *J. Memb. Sci.* 2017, 541, 367–377.

[7] Liu G, Li N, Miller SJ, Kim D, Yi S, Labreche Y, Koros WJ. *Angew. Chem. Int. Ed. Engl.* 2016, 55, 13754–13758.

[8] Jue ML, Breedveld V, Lively RP. *J. Memb. Sci.* 2017, 530, 33–41.

[9] Robeson LM. *J. Memb. Sci.* 1991, 62, 165–185.

[10] Robeson LM. *J. Memb. Sci.* 2008, 320, 390–400.

[11] Stannett V, Szwarc M. *J. Polym. Sci.* 1955, 16, 89–91.

[12] Lee WM. *Polym. Eng. Sci.* 1980, 20, 65–69.

[13] Salame M. *Polym. Eng. Sci.* 1986, 26, 1543–1546.

[14] Jia L, Xu J. *Polym. J.* 1991, 23, 417–425.

[15] Kim E, Huang K, Saunders A, McCallum A, Ceder G, Olivetti E. *Chem. Mater.* 2017, 29, 9436–9444.

[16] Segler MHS, Preuss M, Waller MP. *Nature* 2018, 555, 604.

[17] Kim C, Chandrasekaran A, Huan TD, Das D, Ramprasad R. *J. Phys. Chem. C* 2018, 122, 17575–17585.

[18] Jha A, Chandrasekaran A, Kim C, Ramprasad R. *Model. Simul. Mat. Sci. Eng.* 2019, 27, 024002.

[19] Kim C, Chandrasekaran A, Jha A, Ramprasad R. *MRS Commun.* 2019, 9, 1–7.

[20] Kim C, Pilania G, Ramprasad R. *Chem. Mater.* 2016, 28, 1304–1311.

[21] Mannodi-Kanakkithodi A, Chandrasekaran A, Kim C, Huan TD, Pilania G, Botu V, Ramprasad R. *Mater. Today* 2018, 21, 785–796.

[22] Mannodi-Kanakkithodi A, Huan TD, Ramprasad R. *Chem. Mater.* 2017, 29, 9001–9010.

[23] Mannodi-Kanakkithodi A, Pilania G, Huan TD, Lookman T, Ramprasad R. *Sci. Rep.* 2016, 6, 20952.

[24] Mannodi-Kanakkithodi A, Pilania G, Ramprasad R. *Comput. Mater. Sci.* 2016, 125, 123–135.

[25] Mannodi-Kanakkithodi A, Treich GM, Huan TD, Ma R, Tefferi M, Cao Y, Sotzing GA, Ramprasad R. *Adv. Mater.* 2016, 28, 6277–6291.

[26] Ramprasad R, Batra R, Pilania G, Mannodi-Kanakkithodi A, Kim C. *NPJ Comput. Mater.* 2017, 3, 54.

[27] Wessling M, Mulder MHV, Bos A, van der Linden M, Bos M, van der Linden WE. *J. Memb. Sci.* 1994, 86, 193–198.

[28] Park JY, Paul DR. *J. Memb. Sci.* 1997, 125, 23–39.

[29] Robeson LM, Smith CD, Langsam M. *J. Memb. Sci.* 1997, 132, 33–54.

[30] Yampolskii Y, Shishatskii S, Alentiev A, Loza K. *J. Memb. Sci.* 1998, 149, 203–220.

[31] Hasnaoui H, Krea M, Roizard D. *J. Memb. Sci.* 2017, 541, 541–549.

[32] Ryzhikh V, Tsarev D, Alentiev A, Yampolskii Y. *J. Memb. Sci.* 2015, 487, 189–198.

[33] Pankajakshan P, Sanyal S, de Noord OE, Bhattacharya I, Bhattacharyya A, Waghmare U. *Chem. Mater.* 2017, 29, 4190–4201.

[34] Labute PA. *J. Mol. Graphics Modell.* 2000, 18, 464–477.

[35] Ertl P, Rohde B, Selzer P. *J. Med. Chem.* 2000, 43, 3714–3717.

[36] Landrum G. *RDKit: Open-Source Cheminformatics*, 2006.

Supplementary Material: The online version of this article offers supplementary material (https://doi.org/10.1515/polyeng-2019-0329).