Chemical Science

EDGE ARTICLE

View Article Online
View Journal | View Issue

Cite this: *Chem. Sci.*, 2019, **10**, 8374

All publication charges for this article have been paid for by the Royal Society of Chemistry

A deep neural network model for packing density predictions and its application in the study of 1.5 million organic molecules†

Mohammad Atif Faiz Afzal,*[a] Aditya Sonpal,[a] Mojtaba Haghighatlari,[a] Andrew J. Schultz[a] and Johannes Hachmann *[abc]

The process of developing new compounds and materials is increasingly driven by computational modeling and simulation, which allow us to characterize candidates before pursuing them in the laboratory. One of the non-trivial properties of interest for organic materials is their packing in the bulk, which is highly dependent on their molecular structure. By controlling the latter, we can realize materials with a desired density (as well as other target properties). Molecular dynamics simulations are a popular and reasonably accurate way to compute the bulk density of molecules, however, since these calculations are computationally intensive, they are not a practically viable option for high-throughput screening studies that assess material candidates on a massive scale. In this work, we employ machine learning to develop a data-derived prediction model that is an alternative to physics-based simulations, and we utilize it for the hyperscreening of 1.5 million small organic molecules as well as to gain insights into the relationship between structural makeup and packing density. We also use this study to analyze the learning curve of the employed neural network approach and gain empirical data on the dependence of model performance and training data size, which will inform future investigations.

Received 2nd June 2019
Accepted 8th July 2019

DOI: 10.1039/c9sc02677k

rsc.li/chemical-science

I. Introduction

The packing of atoms, molecules, and polymers in a given volume – either in crystalline or amorphous form – is a fundamental and long-standing issue that has been considered by various disciplines for over a century.[1] The packing density directly impacts properties such as the ionic conductivity,[2] mobility in solvents,[3] mechanical[4] and optical behavior,[5,6] and numerous other physical and chemical properties.[7] Today, the packing density has gained renewed attention in the context of developing advanced materials that fulfill very specific property requirements. Molecular materials and polymers are of particular interest as their packing in the bulk is directly affected by their molecular structure.[8] Manipulating and tailoring the latter offers many opportunities (and challenges) to achieve targeted density values.

Traditional, experimentally-driven trial-and-error searches for new compounds with desired sets of properties have proved to be time consuming and resource intensive. The advent of powerful modeling and simulation techniques as well as readily available time on high-performance computing systems have brought computational and computationally-guided studies to the forefront of the chemical and materials domain. These studies allow us to make increasingly accurate predictions for compounds of interest and uncover promising leads for experimentalist partners to follow up on (see, *e.g.*, ref. 9–20). An even more recent development has been the emergence of machine learning techniques that empower us to advance, augment, correct, or even replace physics-based modeling and simulation.[21] In the latter scenario, machine learning is used to create data-derived property prediction models that serve as surrogates for physics-based models in order to dramatically accelerate the compound characterization and thus the overall discovery process. In addition to enabling hyperscreening studies, we can employ machine learning to gain a better understanding of the structure–property relationships that determine the behavior of

[a]*Department of Chemical and Biological Engineering, University at Buffalo, The State University of New York, Buffalo, NY 14260, USA. E-mail: m27@buffalo.edu; hachmann@buffalo.edu*

[b]*Computational and Data-Enabled Science and Engineering Graduate Program, University at Buffalo, The State University of New York, Buffalo, NY 14260, USA*

[c]*New York State Center of Excellence in Materials Informatics, Buffalo, NY 14203, USA*

† Electronic supplementary information (ESI) available: It provides the SMILES of the virtual compound library, SAscores of the 15 building blocks, and details of the computational and experimental data underlying the figures and tables throughout this paper and that were used in the statistical analysis. We also provide the list of descriptors used to develop the DNN model and the trained model in scikit-learn's pickle (pkl) format. We note that this and other trained density models as well as the corresponding ML workflows are available as part of the *ChemML* package's template and model collection (*e.g.*, for retraining, customization, or transfer learning).[51] Finally, we give detailed definitions of all statistical metrics employed in this work. See DOI: 10.1039/c9sc02677k

compounds in the corresponding domains of chemical space. The creation of new machine learning prediction models for various target properties and the advancement of the underlying methodology is an active field of research.[22] Key considerations are accuracy, cost, robustness, and range of applicability. Artificial neural networks are a popular and efficient machine learning approach.[23] Multi-layer 'deep' neural networks (DNNs) yield particularly flexible models that have been used to predict an array of chemical properties, including refractive indices,[24] dielectric constants,[25] atomization energies,[26] chemical reactivities,[27] melting points,[28,29] viscosities,[30] solubilities,[31] and others.

In this work, we develop a DNN prediction model for the packing density of small organic molecules in an amorphous bulk phase and conduct a hyperscreening of 1.5 million candidate compounds. Our interest in this target property originates from our ongoing *in silico* discovery and design efforts for polymers with high refractive index (RI)[32–34] to be used in optic and optoelectronic applications.[35,36] We previously established an RI modeling protocol based on the Lorentz–Lorenz equation and parametrized with the polarizability and number density.[32,33] For the number density, we introduced a hybrid physics-based/data-derived prediction model using the van der Waals volume computed *via* the Slonimskii method and the packing coefficient from a support vector regression machine learning model.[37] An alternative and commonly employed route to computing the (number) density, is the use of molecular dynamics (MD) simulations, which we recently started exploring in our study of high-RI polymers. However, as these MD calculations are computationally expensive and technically challenging, they are not particularly well suited for the large-scale assessment of compounds in the course of high-throughput screening studies. To bypass this problem, we develop a DNN surrogate model for the MD density predictions. It allows us to rapidly and accurately compute the density values of the 1.5 million molecules of a virtual screening library we create for proof of concept. For this, we perform MD simulations on a subset of 100 000 compounds, use the results to train our DNN model, and subsequently employ it to compute the packing density of the remaining 1.4 million molecules. We mine the density results to identify patterns that lead to desirable outcomes (*i.e.*, different density regimes). We also evaluate the learning curve for the density prediction to assess the dependence of training set size and model accuracy.

In Sec. II, we detail the methods employed in our work. We describe the MD modeling protocol we use to compute the density values (Sec. IIA), discuss the molecular design space we consider and the application of our virtual high-throughput screening tools on the resulting compound library (Sec. IIB), introduce our DNN prediction model (Sec. IIC), and establish our pattern analysis approaches to mine the obtained results (Sec. IID). Sec. III presents and discusses the outcomes of our study, in particular the density predictions from MD and DNN (Sec. IIIA), the efficiency of the DNN approach (Sec. IIIB), and the emerging structure–property relationships (Sec. IIIC). Our findings are summarized in Sec. IV.

II. Methods and computational details

A. Molecular dynamics modeling protocol

We employ the following MD modeling protocol to generate the data for the training and testing of the DNN density prediction model at the center of this work. Starting from the simplified molecular-input line-entry system (SMILES)[38] string of a given compound, we employ the OpenBabel code[39] to create a 3-dimensional structure guess, and then pre-optimize it using the MMFF94s force field[40] *via* steepest descent. We then compute the packing density with the general Amber force field (GAFF).[41] For this, we obtain the GAFF parameters in automated fashion[42] using the Antechamber toolkit that is part of AmberTools,[43] and carry out the MD simulations within the GROMACS package.[44] We employ GROMACS' solvate tool to create a $(10\ \mathrm{nm})^3$ simulation box and fill it with the pre-optimized target molecules. The number of molecules in the simulation box depends on the given molecule size, but a typical system contains around 1000 molecules (*e.g.*, 972 for benzylcyclopentane). The system is first subjected to a minimization of the internal energy, which is associated with the relaxation of bonds, bond angles, and dihedral bond angles. This is followed by NVT and NPT equilibration steps for 100 and 240 ps, respectively. Both NVT and NPT ensembles use a Nosé–Hoover thermostat at 298.15 K for temperature control. The NPT ensemble uses the Parinello–Rahman barostat for pressure control. We conclude the MD protocol with a final 40 ps NPT production run. We use an MD timestep of 0.2 fs. We obtain the density by averaging the density values of the system at intervals of 0.2 ps during this final run. We note that this protocol is expected to yield kinetically stable amorphous phases rather than thermodynamically stable crystal structures or meta-stable polymorphs. GAFF is known to underestimate the density values compared to those from experiment, especially for high-density compounds.[45] We employ a linear fit between the calculated and experimental values to account for the systematic differences and empirically calibrate the MD results.

B. Candidate library generation and high-throughput screening

We create a virtual library of 1.5 million small organic molecules using our library generator code *ChemLG*[46,47] in constrained combinatorial mode. This library is constructed based on the sequential combinatorial linking of the 15 molecular building blocks shown in Fig. 1 for four generations, while enforcing certain constraints, *i.e.*, a molecular weight within the range of 150 to 400 Dalton and limiting the number of ring-moieties to four. The hydrogen atoms in each building block are used as linker handles. Our proof-of-principle library is designed to feature different connections between simple moieties, most of which are commonly used in organic materials (except **B5**, **B8**, and **B9**). We use the eToxPred software to compute the synthetic accessibility score (SAscore) of all 15 building blocks[48] and obtain similarly favorable values between 2.4 and 2.9 on the 1–10 scale (with 1 being the most

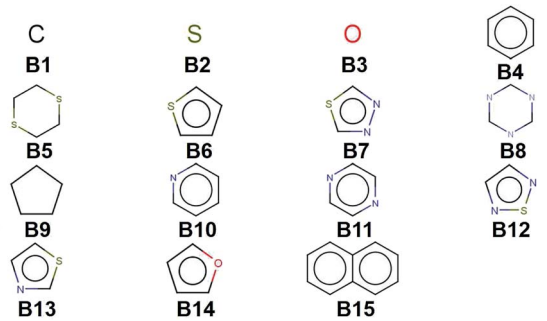

Fig. 1 Molecular building blocks used to create the candidate library of 1.5 million compounds studied in this work.

synthetically accessible). We provide the details of the accessibility analysis in the ESI.† Our generation approach limits the size of the library while yielding both a diverse set of compounds as well as candidates with more subtle differences for the model to distinguish. The complete library is provided in the ESI.†

To facilitate the density evaluation for a large number of compounds *via* the MD modeling protocol introduced in Sec. IIA, we employ our automated virtual high-throughput screening framework *ChemHTPS*.[46,49] *ChemHTPS* creates inputs for the MD simulations, executes the modeling protocol, monitors the calculations, parses and assesses the results, and extracts and processes the information of interest. Of the 1.5 million compounds in our screening library, we randomly select a subset of 100 000 for study at the MD level.

C. Neural network prediction model

We use the MD results for these 100,000 molecules as the ground truth for our data-derived density prediction model. For this, we pursue a DNN approach within a feature space of molecular descriptors. We build the DNN model using *ChemML*,[46,50,51] our program suite for machine learning and informatics in chemical and materials research. In this work, *ChemML* employs the scikit-learn 0.18.2 library for the multilayer perceptron regressor 1.17.1 (ref. 52) and 197 descriptors from Dragon 7.[53] These descriptors include constitutional indices and functional group counts. If two descriptors are mutually correlated, they are not independent and thus redundant. In the cases where the Pearson correlation coefficient R is >95%, Dragon removes one of them, *i.e.*, the one that shows more correlation with the rest of the descriptors. (A detailed list of the descriptors is provided in the ESI.†) We apply the grid search method for a coarse optimization of the DNN model hyperparameters. The hyperparameter search space includes a number of activation functions (identity, tanh, rectified linear unit, and logistic), L2 regularization parameters (0.1, 0.01, 0.001, 0.0001, and 0.00001), solvers for the optimization of the weights (sgd and adam), and learning rate types (constant, invscaling, and adaptive). The best model from the hyperparameter optimization features the rectified linear unit as the activation function, 'adam' solver, adaptive learning rate, and an L2 regularization parameter of 0.0001. The final DNN has two fully connected hidden layers with 100 neurons each. For the initial model evaluation, we randomly divide the 100 000-molecule data set into 80% training and 20% test set. To assess the learning curve, we evaluate the model performance for incrementally increasing training set size from 0.05% to 100% of the entire data set (*i.e.*, from 50 to 100 000 data points). We apply the bootstrapping method, *i.e.*, for each training set size, we obtain the training set by randomly sampling the entire data set. The remaining data points serve as test set. For every training set size, we repeat the process (with replacement) 50 times, *i.e.*, all 50 repetitions are independent of each other. We subsequently calculate statistics over the results of the 50 models that are based on these training sets for each training set size.

D. Data mining and pattern recognition

In addition to identifying candidates with particular density values from our MD screening and DNN hyperscreening studies, we mine the compiled results to better understand the correlation between molecular structure and packing density. One pattern recognition approach we pursue is the hypergeometric distribution analysis, in which we determine the Z-scores (Z_i) of each building block i used in the creation of the molecular library as

$$Z_i = \frac{k_i - m\frac{K_i}{M}}{\sigma_i},$$

with

$$\sigma_i = \left[\frac{mK_i}{M} \times \left(\frac{M - K_i}{M}\right) \times \left(\frac{M - m}{M - 1}\right)\right]^{\frac{1}{2}},$$

where M is the total number of molecules in the entire library, m is the subset of molecules under consideration (*e.g.*, the compounds in a certain density regime), K_i is the number of occurrences of building block i in M molecules and k_i its occurrences in the subset of m molecules. A large Z-score indicates that a building block appears more frequently in that subset compared to the rest of the library (or a random sample). By applying the hypergeometric distribution analysis, we can thus identify the building blocks with the largest impact on the target property and the degree to which they correlate with desired density values. Furthermore, we identify the building blocks that are prominent in particular density regimes and assess Z-score trends in density-ordered candidate subsets across the entire density range. In addition, we compute the average density values of the candidates derived from each building block, and analyze this data for trends. We employ the *ChemML* package for all data mining and pattern recognition tasks.

The following metrics are used in the error analyses of our modeling approaches: mean absolute error (MAE), mean absolute percentage error (MAPE), root mean squared error (RMSE), root mean squared percentage error (RMSPE), mean error (ME), mean percentage error (MPE), maximum absolute error (MaxAE), and maximum absolute percentage error (MaxAPE).

Aside from providing these direct measures, we also quantify the extent of correlations and systematic biases between results of different methods by listing the correlation coefficients R^2, slopes, and offsets of linear regressions.

III. Results and discussion

A. Density predictions

To test the accuracy of the MD density modeling protocol introduced in Sec. IIA, we compare its results against the experimentally known density values of 175 small organic molecules.[54] This collection of compounds exhibits densities between 600 and 2000 kg m^{-3}. As shown in the first column of Table 1, the calibrated MD results obtained for this collection are in good agreement with the experimental data, in particular considering that the density is also affected by factors other than the molecular structure (*e.g.*, processing and ambient conditions) that are not accounted for in the simulations. We also note that the experimental data set is structurally more diverse than the screening library, *i.e.*, it includes non-aromatic and aromatic moieties, halogens, and different functional groups such as OH, C=C, C≡C, C=O, N=O, *etc*. Despite this diversity, the comparison shows an R^2 of 0.95, which underscores the utility of the employed MD approach. (Note that the very small ME/MPE as well as the ideal offset and slope are due to the empirical calibration scheme introduced in Sec. IIA. The calculated slope and offset of the linear fit between MD and experimental data is 0.84 and 121 kg m^{-3}, respectively.)

The trained DNN model mimics – by design – the MD simulations it is derived from. The second column of Table 1 shows the DNN predictions for the MD test data. The benchmark analysis reveals a very good agreement between the MD and DNN results. The correlation coefficient is $R^2 = 0.98$, both MAPE and RMSPE are around 1%, and MaxAPE is just 5.5%. Most importantly, we find that the DNN prediction errors are significantly smaller than the intrinsic MD errors (by a factor of 4 to 6), which means that the DNN and MD results are statistically indistinguishable. The prediction quality for the MD training data set is essentially identical to the test set shown in Table 1, indicating that our DNN model is not overfitted and that its predictions are sound and reliable. Fig. 2 shows the comparison of the MD and DNN results for the 100 000 compounds for which both MD and DNN results were computed with training data shown in blue and test data in red. We note that the inset of Fig. 2 shows clusters of molecules for which the DNN model predicts very similar density values (suggested by the horizontal lines of data points). This indicates a certain incompleteness of the selected feature representation and associated information loss, which suggests that even more accurate models can be achieved using other, more comprehensive descriptor sets (such as 3-dimensional molecular descriptors,[53] extended connectivity,[55] hashed topological torsion (HTT),[56] hashed atom pair (HAP),[57] or Morgan[58] fingerprints, or a combination thereof). We also test our DNN model on the experimental data set. As the latter exhibits a structural diversity that goes well beyond the relatively narrow scope of the training data used to create the former, it yields unsurprisingly large errors (MAPE = 9.7%, RMSPE = 13.5%). Nonetheless, the $R^2 = 0.89$ shows that the DNN model still captures the structure–property relationships to a certain degree, and given appropriate training data, DNN should deliver predictive models for those compound pools as well.

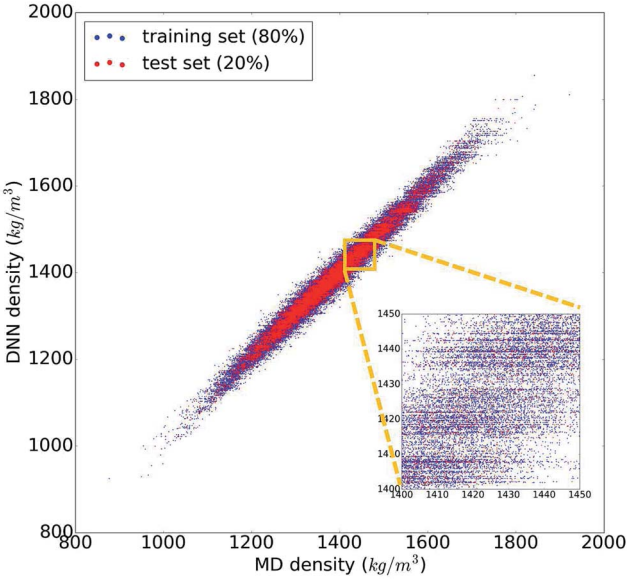

Fig. 2 Comparison of the 100 000 calculated density values from molecular dynamics (MD) and the deep neural network (DNN) prediction model. Data points of the DNN training set are shown in blue and those of the test set in red.

With the accuracy of the trained DNN model established, we apply it to the remaining 1.4 million compounds of the screening library introduced in Sec. IIB with the expectation of obtaining similar results as MD would yield. The DNN density predictions are summarized in Fig. 3. The density values of the molecules at hand range from 902 to 1851 kg m^{-3} with an average of 1384 kg m^{-3}. The results show a *t*- or Gaussian-like distribution and most of the compounds in the library have density values between 1200 to 1600 kg m^{-3}, with only very few examples at the extreme high and low density regime. It is worth

Table 1 Performance comparison of our density predictions approaches. The column labeled MD (exp) compares the calibrated MD predictions with the experimental values of our collection of 175 compounds. The column labeled DNN (MD) compares the DNN predictions with the 20 000 MD results of our test set: all errors (MAE, RMSE, ME, MaxAE) and the offset are given in kg m^{-3}

	MD (exp)	DNN (MD)
R^2	0.95	0.98
Slope	1.00	0.97
Offset	0.00	38.49
MAE (MAPE)	50.8 (4.7%)	10.8 (0.9%)
RMSE (RMSPE)	69.5 (6.2%)	13.6 (1.1%)
ME (MPE)	0.0 (0.0%)	−3.5 (−0.3%)
MaxAE (MaxAPE)	225.0 (20.4%)	59.2 (5.5%)

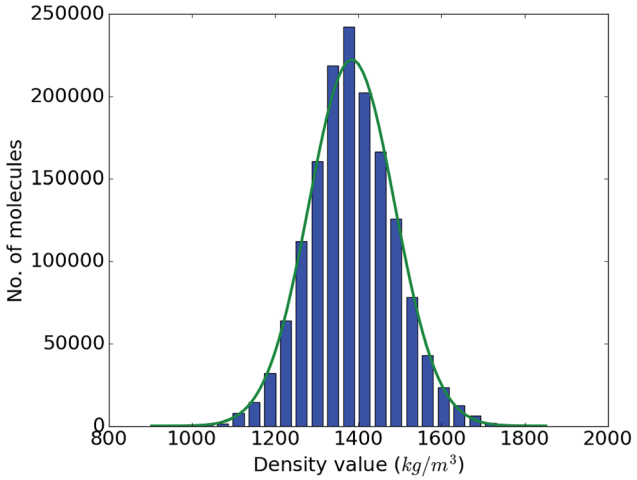

Fig. 3 Range and distribution of the DNN density predictions for our proof-of-concept screening library of 1.5 million small organic molecules with a corresponding normal distribution overlayed.

noting that these extreme packing density values may be desirable for certain material applications (*e.g.*, light-weight plastics with large strength-to-density ratio or rigid, impact-resistant thermo-plastics). The sparsity of instances for extreme density values emphasizes the valuable role that high-throughput screening studies *via* physics-based modeling and/or data-derived prediction models can play in the discovery of suitable materials candidates.

B. Neural network efficiency

After confirming that the DNN prediction model can accurately reproduce MD-level results (which we in turn showed to accurately reproduce experimental data), we now investigate its efficiency, in particular relative to MD. Our MD calculations for the subset of 100 000 molecules took a total of 5 million core hours of compute time on a modern high-performance computing cluster. For the entire screening library, this extrapolates to approximately 75 million core hours (In addition to the compute time, there is generally a considerable amount of human time required for the setup and execution of these calculations. In our study, many of these tasks were performed by *ChemHTPS* without manual intervention.) The demand on disk space is another issue, and we estimate a need for 120 terabytes for the entire library (15 terabytes without trajectories). The DNN prediction model produces essentially the same results in less than 10 core hours of compute time (without performance optimization), with all but 10 minutes of the time required to generate the feature matrix of the compound library. Disk use is marginal. This corresponds to a speed-up of about seven orders of magnitude, with negligible loss in accuracy. A speed-up of that magnitude allows a corresponding increase in the scale and scope that is affordable for screening studies.

The bottleneck of our DNN prediction model is the generation of the training data needed for its creation. It is worth noting, though, that this is a fixed cost rather than an effort that scales with the number of compounds studied. The size of the employed training set (100 000 compounds corresponding to 5 million core hours) was originally chosen *ad hoc*. We now assess the learning curve as a function of training set size to gain insights into the actual data needs of our DNN model, which is one of the key questions in applying machine learning to any given problem setting. Our goal is to establish, how many data points are necessary to converge the learning process and/or achieve a desired accuracy. By minimizing the training set size requirement, we minimize the investment in computational resources needed to perform the expensive MD simulations. To address this question, we successively increase the size of the training set from 50 to 100 000 molecules. The resulting learning curve is shown in Fig. 4. We observe that all models trained on fewer than 2000 data points (*i.e.*, 2% of the available data) perform poorly. Models based on 2000 to 4000 data points offer acceptable accuracy. Those based on 4000 to 6000 data points offer very good accuracy, and at 10 000 data points, the training is essentially saturated and the learning curve plateaus off. Additional training data does not lead to an improvement of the DNN model and is essentially wasted. Thus, we do not require a large data set of 100 000 molecules to learn the packing density of organic molecules. We can develop an accurate model using MD data of just 5000 molecules (or more conservatively 10 000). This reduces our demand of computing time from 5 million to less than 0.25 (or 0.5) million core hours (including additional data for the test set), which has significant implications for the cost-benefit analysis and viability of this approach.

We stress that the data demand is highly dependent on the nature of the data and the employed machine learning approach (including the feature representation), and there are distinct limits to generalizing our findings. Instead of a *post-mortem* analysis of the learning curve as provided here, we will use an on-the-fly assessment of the learning curve combined

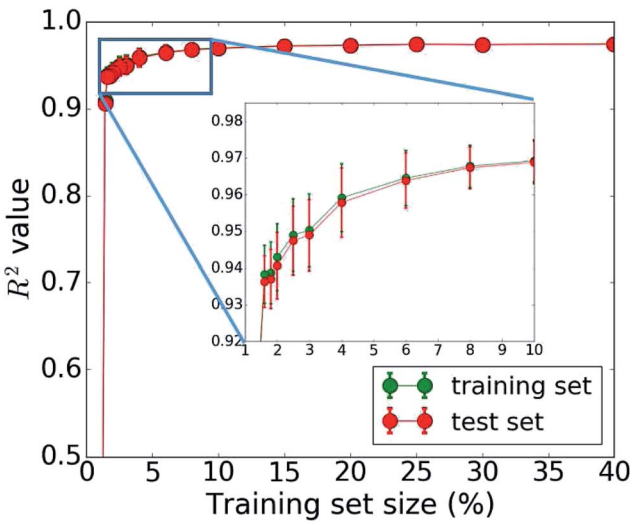

Fig. 4 Dependence of the model accuracy (measured by the correlation coefficient R^2) on the training set size (1% corresponds to 1000 data points). The learning curve shows the mean R^2 from 50 bootstrap repetitions and the standard deviation is given in the error bars. The accuracy for the training set is plotted in green, that for the test set in red.

with a just-in-time termination of the training data generation to minimize our data footprint in future studies.

C. Relationship between molecular structure and packing density

When considering the screening results, we are not only in a position to assess a large number of compounds, but we can also learn patterns from the data set in its entirety. Our analysis in Fig. 5 shows the average density values and distributions of all compounds containing a given building block (*cf.* Fig. 1). On the high density end, we find sulfur-heterocyclic moieties; the nitrogen- and oxygen-heterocycles yield medium density systems; and the low-density regime is dominated by carbon-based, non-heteroatomic building blocks. Molecules with **B7** (1,3,4-thiadiazole) and **B12** (1,2,5-thiadiazole) have the highest average densities, while those that incorporate **B1** (CH_2-linker) and **B9** (cyclopentane) exhibit the lowest values. Aside from the linker groups, there is a clear correlation between density value and the heteroatom type and fraction in a corresponding moiety.

Based on the construction of our library, more than 80% of the candidate compounds contain sulfur and more than 90% contain nitrogen. Fig. 6 demonstrates how the density values depend on the weight percentage of the sulfur and nitrogen atoms in the compounds at hand. Our library thus yields the highest density values for molecules that by weight contain 30 to 50% sulfur and 20 to 30% nitrogen.

While the average density values indicate the cumulative impact of a particular building block, we find relatively large standard deviations (*cf.* Fig. 5). For a more detailed picture of the occurrences of building blocks in a particular subset of the library, we perform the *Z*-score analysis introduced in Sec. IID. Fig. 7 shows the corresponding results for the molecules with the highest density values (*i.e.*, the top 10% subset) with clear and distinct trends. Consistent with our previous analysis, we observe very large *Z*-score values for and

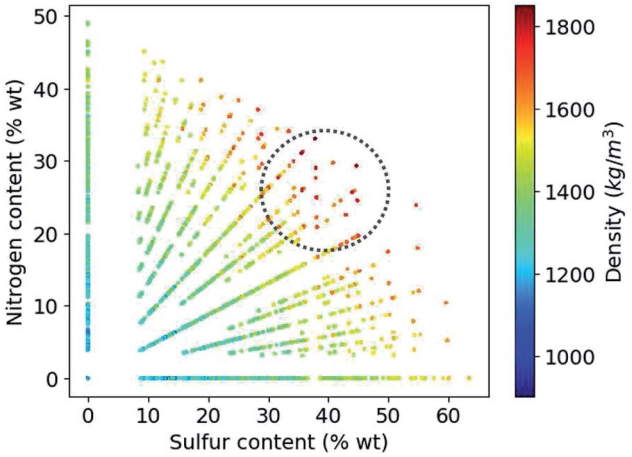

Fig. 6 Variation of density values as a function of weight percentage of sulfur and nitrogen in the molecules.

thus a strong overexpression of **B7** and **B12**. **B13** (thiazole) also shows a large *Z*-score, and so do to a lesser extent **B2** (S-linker) and **B3** (O-linker) as well. These moieties are clearly favorable if high-density compounds are desired. In addition to assessing the high-density regime, we employ the hypergeometric distribution analysis to identify the prevalence of building blocks in the complete spectrum of density values. For this, we sort our virtual library by increasing density values, divide it into ten equal segments, and perform our analysis within each of these subsets as shown in Fig. 8. Based on the data from this analysis, we can identify trends in the impact of individual building blocks on the density of organic molecules. The *Z*-score of building blocks **B2**, **B7**, **B12**, and **B13** increases with increasing density values, indicating a direct correlation, whereas it decreases for **B1**, **B4**, **B9**, **B10**, and **B15**, indicating an inverse correlation. The former are thus suitable to design organic molecules with

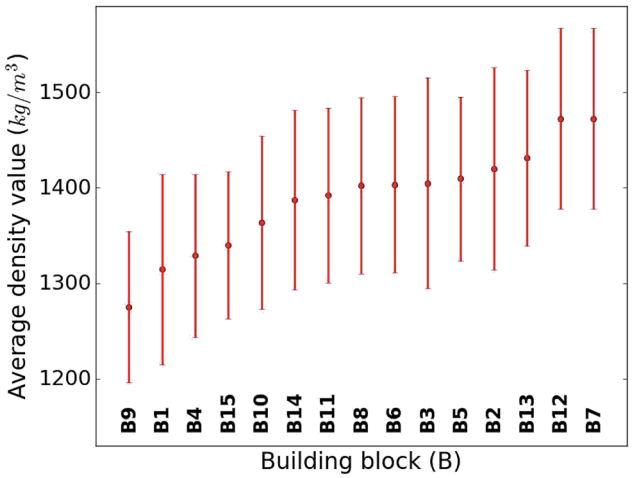

Fig. 5 Density value distribution around the respective average density values (points) of the molecules containing a given building block. The bands refer to one standard deviation.

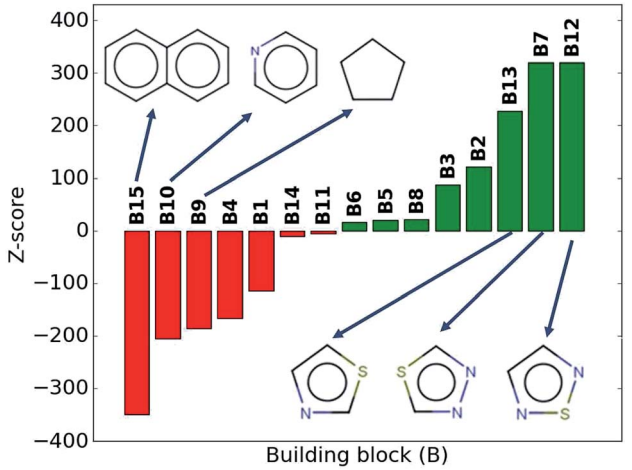

Fig. 7 *Z*-scores of each building block in the compounds with the highest density values (top 10% of the library). Green represents positive *Z*-scores, and negative ones are shown in red.

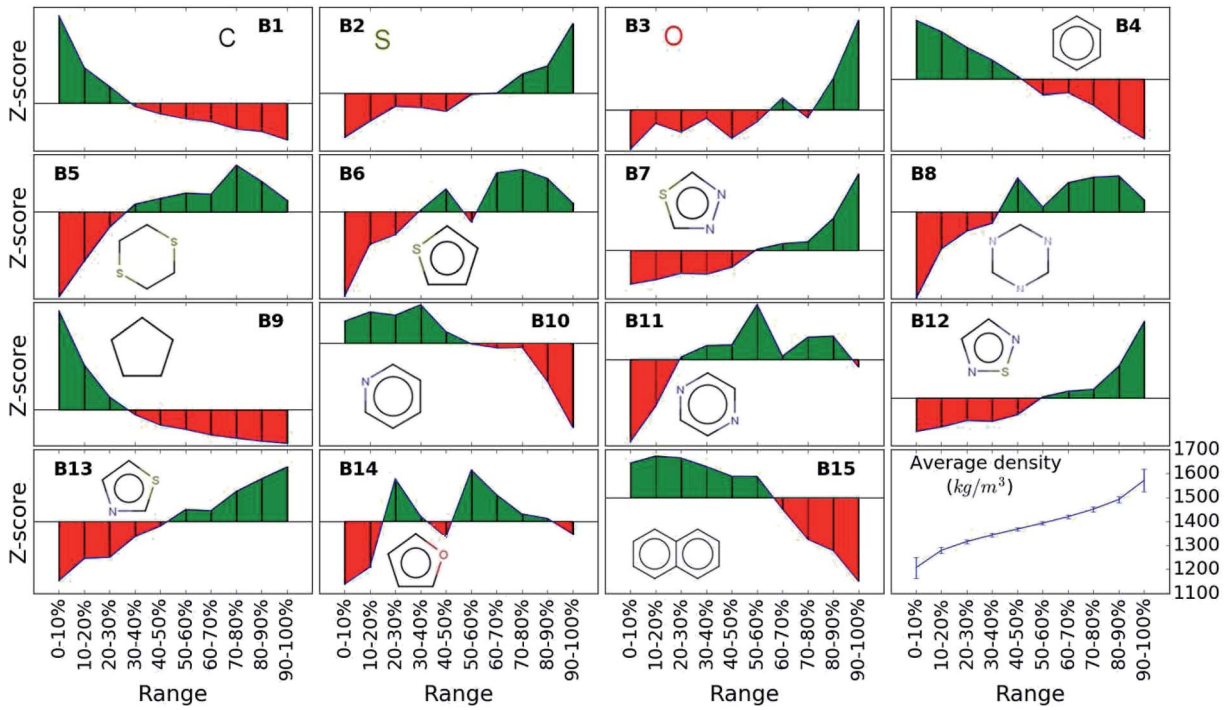

Fig. 8 *Z*-score of each building block in all library subsets with increasing density values. Green color indicates positive *Z*-scores and red negative values. The last cell shows the average density values in each of the ten segment with increasing trend from 1200 to 1600 kg m^{-3}.

higher density, and the latter could be used to achieve compounds with lower density. These findings are consistent with our prior analysis.

IV. Conclusions

The ability to predict the properties of novel compounds prior to synthesis, and to understand how these properties depend on their structure, is of considerable importance in materials discovery and design. In this paper, we showed that MD simulations can accurately predict experimental packing density values of small organic molecules and we provided corresponding benchmark results to quantify this finding. We conducted a high-throughput MD screening of 100 000 compounds, which allowed us to train a DNN density prediction model. This DNN model accurately reproduces the MD data within the margins of MD's intrinsic error, while being nearly seven orders of magnitude faster than MD. This exceedingly efficient approach allowed us to rapidly obtain the density values of a 1.5 million compound screening library, which would have been prohibitively time consuming and well out of reach for MD. By analysing the large data set resulting from this study, we could elucidate structure–property relationships that determine the density values. We identified prevalent moieties in the high and low density regime and could quantify the impact of heteroatoms (sulfur and nitrogen). Further, we evaluated the DNN learning curve for the density prediction with respect to the available training data and found a considerably lower data demand than we had anticipated. Following this lesson, we will in future studies employ an on-the-fly assessment of the learning curve and terminate the training data generation once we observe satisfactory saturation. This will allow us to alleviate the data generation bottleneck and make machine learning models an even more viable and attractive proposition. Overall, our study underscores the value of combining powerful machine learning approaches with traditional computational modeling for the generation of the necessary data. It also demonstrates the utility of our software ecosystem (including the *ChemLG* molecular library generator code, the *ChemHTPS* automated high-throughput *in silico* screening program, and the *ChemML* machine learning package) in facilitating and supporting research efforts of this nature.

Conflicts of interest

The authors declare to have no competing financial interests.

Acknowledgements

This work was supported by the National Science Foundation (NSF) CAREER program (grant No. OAC-1751161), and the New York State Center of Excellence in Materials Informatics (grants No. CMI-1140384 and CMI-1148092). Computing time on the high-performance computing clusters '*Rush*', '*Alpha*', '*Beta*', and '*Gamma*' was provided by the UB Center for Computational Research (CCR). The work presented in this paper is part of MAFA's PhD thesis.[59] MH gratefully acknowledges support by Phase-I and Phase-II Software Fellowships (grant No. ACI-1547580-479590) of the NSF Molecular Sciences Software

Institute (grant No. ACI-1547580) at Virginia Tech.[60,61] We are grateful to Prof. Chong Cheng for valuable discussions and insights.

References

1 A. K. H. Kwan and W. W. S. Fung, Packing density measurement and modelling of fine aggregate and mortar, *Cem. Concr. Compos.*, 2009, **31**, 349–357.
2 J. Swenson and L. Börjesson, Correlation between free volume and ionic conductivity in fast ion conducting glasses, *Phys. Rev. Lett.*, 1996, **77**, 3569–3572.
3 Y. Shen, X. He and F. R. Hung, Structural and dynamical properties of a deep eutectic solvent confined inside a slit pore, *J. Phys. Chem. C*, 2015, **119**, 24489–24500.
4 H. E. H. Meijer and L. E. Govaert, Mechanical performance of polymer systems: The relation between structure and properties, *Prog. Polym. Sci.*, 2005, **30**, 915–938.
5 S. Ando, DFT calculations on refractive index dispersion of fluoro-compounds in the DUV-UV-visible region, *J. Photopolym. Sci. Technol.*, 2006, **19**, 351–360.
6 Y. Terui and S. Ando, Coefficients of molecular packing and intrinsic birefringence of aromatic polyimides estimated using refractive indices and molecular polarizabilities, *J. Polym. Sci., Part B: Polym. Phys.*, 2004, **42**, 2354–2366.
7 F. R. Sheu and R. T. Chern, Effects of packing density on the gas-transport properties of poly (phenolphthalein phthalate) s, *J. Polym. Sci., Part B: Polym. Phys.*, 1989, **27**, 1121–1133.
8 N. Tanio and M. Irie, Refractive index of organic photochromic dye-amorphous polymer composites, *Jpn. J. Appl. Phys., Part 1*, 1994, **33**, 3942.
9 J. Hachmann, R. Olivares-Amaya, S. Atahan-Evrenk, C. Amador-Bedolla, R. S. Sánchez-Carrera, A. Gold-Parker, L. Vogt, A. M. Brockway and A. Aspuru-Guzik, The Harvard Clean Energy Project: Large-scale computational screening and design of organic photovoltaics on the world community grid, *J. Phys. Chem. Lett.*, 2011, **2**, 2241–2251.
10 R. Olivares-Amaya, C. Amador-Bedolla, J. Hachmann, S. Atahan-Evrenk, R. S. Sánchez-Carrera, L. Vogt and A. Aspuru-Guzik, Accelerated computational discovery of high-performance materials for organic photovoltaics by means of cheminformatics, *Energy Environ. Sci.*, 2011, **4**, 4849–4861.
11 C. Amador-Bedolla, R. Olivares-Amaya, J. Hachmann and A. Aspuru-Guzik, Organic Photovoltaics, in *Informatics for materials science and engineering: Data-driven discovery for accelerated experimentation and application*, ed. Krishna Rajan, Butterworth-Heinemann, Amsterdam, 2013, ch. 17, pp. 423–442.
12 J. Hachmann, R. Olivares-Amaya, A. Jinich, A. L. Appleton, M. A. Blood-Forsythe, L. R. Seress, C. Roman-Salgado, K. Trepte, S. Atahan-Evrenk, S. Er, S. Shrestha, R. Mondal, A. Sokolov, Z. Bao and A. Aspuru-Guzik, Lead candidates for high-performance organic photovoltaics from high-throughput quantum chemistry – the Harvard Clean Energy Project, *Energy Environ. Sci.*, 2014, **7**, 698–704.
13 E. O. Pyzer-Knapp, C. Suh, R. Gómez-Bombarelli, J. Aguilera-Iparraguirre and A. Aspuru-Guzik, What is high-throughput virtual screening? A perspective from organic materials discovery, *Annu. Rev. Mater. Res.*, 2015, **45**, 195–216.
14 S. A. Lopez, R. O. Pyzer-Knapp, G. N. Simm, T. Lutzow, K. Li, L. R. Seress, J. Hachmann and A. Aspuru-Guzik, The Harvard organic photovoltaic dataset, *Sci. Data*, 2016, **3**, 160086.
15 S. Curtarolo, G. L. Hart, M. B. Nardelli, N. Mingo, S. Sanvito and O. Levy, The high-throughput highway to computational materials design, *Nat. Mater.*, 2013, **12**, 191.
16 P. Strasser, Q. Fan, M. Devenney, W. H. Weinberg, P. Liu and J. K. Nørskov, High throughput experimental and theoretical predictive screening of materials- a comparative study of search strategies for new fuel cell anode catalysts, *J. Phys. Chem. B*, 2003, **107**, 11013–11021.
17 A. Jain, Y. Shin and K. A. Persson, Computational predictions of energy materials using density functional theory, *Nat. Rev. Mater.*, 2016, **1**, 15004.
18 G. Hautier, C. Fischer, V. Ehrlacher, A. Jain and G. Ceder, Data mined ionic substitutions for the discovery of new compounds, *Inorg. Chem.*, 2010, **50**, 656–663.
19 S. Li, Y. G. Chung and R. Q. Snurr, High-throughput screening of metal–organic frameworks for CO_2 capture in the presence of water, *Langmuir*, 2016, **32**, 10368–10376.
20 K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev and A. Walsh, Machine learning for molecular and materials science, *Nature*, 2018, **559**, 547.
21 J. Hachmann, T. L. Windus, J. A. McLean, V. Allwardt, A. C. Schrimpe-Rutledge, M. A. F. Afzal and M. Haghighatlari, Framing the role of big data and modern data science in chemistry, *Tech. Rep.*, NSF CHE Workshop Report, 2018.
22 M. Haghighatlari and J. Hachmann, Advances of machine learning in molecular modeling and simulation, *Curr. Opin. Chem. Eng.*, 2019, **23**, 51–57.
23 X. Yao, Evolving artificial neural networks, *Proc. IEEE*, 1999, **87**, 1423–1447.
24 A. Alexandridis, E. Chondrodima, K. Moutzouris and D. Triantis, A neural network approach for the prediction of the refractive index based on experimental data, *J. Mater. Sci.*, 2012, **47**, 883–891.
25 A. Mannodi-Kanakkithodi, G. Pilania, T. D. Huan, T. Lookman and R. Ramprasad, Machine learning strategy for accelerated design of polymer dielectrics, *Sci. Rep.*, 2016, **6**, 20952.
26 G. Montavon, K. Hansen, S. Fazli, M. Rupp, F. Biegler, A. Ziehe, A. Tkatchenko, A. V. Lilienfeld and K. Müller, Learning invariant representations of molecules for atomization energy prediction, in *Advances in neural information processing systems 25*, ed. F. Pereira, C. J. C. Burges, L. Bottou and K. Q. Weinberger, Curran Associates, Inc., 2012, pp. 440–448.
27 V. Simon, J. Gasteiger and J. Zupan, A combined application of two different neural network types for the prediction of chemical reactivity, *J. Am. Chem. Soc.*, 1993, **115**, 9148–9159.
28 M. Karthikeyan, R. C. Glen and A. Bender, General melting point prediction based on a diverse compound data set

and artificial neural networks, *J. Chem. Inf. Model.*, 2005, **45**, 581–590.

29 A. Sonpal, *Predicting melting points of deep eutectic solvents*, Master's thesis, University at Buffalo, 2018.

30 F. Gharagheizi, QSPR analysis for intrinsic viscosity of polymer solutions by means of GA-MLR and RBFNN, *Comput. Mater. Sci.*, 2007, **40**, 159–167.

31 J. Huuskonen, M. Salo and J. Taskinen, Aqueous solubility prediction of drugs based on molecular topology and neural network modeling, *J. Chem. Inf. Comput. Sci.*, 1998, **38**, 450–456.

32 M. A. F. Afzal, C. Cheng and J. Hachmann, Combining first-principles and data modeling for the accurate prediction of the refractive index of organic polymers, *J. Chem. Phys.*, 2018, **148**, 241712.

33 M. A. F. Afzal and J. Hachmann, Benchmarking DFT approaches for the calculation of polarizability inputs for refractive index predictions in organic polymers, *Phys. Chem. Chem. Phys.*, 2019, **21**, 4452–4460.

34 M. A. F. Afzal, M. Haghighatlari, S. P. Ganesh, C. Cheng and J. Hachmann, Accelerated discovery of high-refractive-index polyimides *via* first-principles molecular modeling, virtual high-throughput screening, and data mining, *J. Phys. Chem. C*, 2019, **123**, 14610–14618.

35 T. Higashihara and M. Ueda, Recent progress in high refractive index polymers, *Macromolecules*, 2015, **48**, 1915–1929.

36 E. K. Macdonald and M. P. Shaver, Intrinsic high refractive index polymers, *Polym. Int.*, 2015, **64**, 6–14.

37 G. L. Slonimskii, A. A. Askadskii and A. I. Kitaigorodskii, The packing of polymer molecules, *Polym. Sci.*, 1970, **12**, 556–577.

38 D. Weininger, SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules, *J. Chem. Inf. Model.*, 1988, **28**, 31–36.

39 N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch and G. R. Hutchison, Open Babel: An open chemical toolbox, *J. Cheminf.*, 2011, **3**, 33.

40 T. A. Halgren, MMFF VI MMFF94s option for energy minimization studies, *J. Comput. Chem.*, 1999, **20**, 720–729.

41 J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman and D. A. Case, Development and testing of a general amber force field, *J. Comput. Chem.*, 2004, **25**, 1157–1174.

42 J. Wang, W. Wang, P. A. Kollman and D. A. Case, Automatic atom type and bond type perception in molecular mechanical calculations, *J. Mol. Graphics Modell.*, 2006, **25**, 247–260.

43 D. A. Case, I. Y. Ben-Shalom, S. R. Brozell, D. S. Cerutti, T. E. Cheatham, V. W. D. Cruzeiro III, T. A. Darden, R. E. Duke, D. Ghoreishi, M. K. Gilson, H. Gohlke, A. W. Goetz, D. Greene, R. Harris, N. Homeyer, S. Izadi, A. Kovalenko, T. Kurtzman, T. S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, D. J. Mermelstein, K. M. Merz, Y. Miao, G. Monard, C. Nguyen, H. Nguyen, I. Omelyan, A. Onufriev, F. Pan, R. Qi, D. R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C. L. Simmerling, J. Smith, R. Salomon-Ferrer, J. Swails, R. C. Walker, J. Wang, H. Wei, R. M. Wolf, X. Wu, L. Xiao, D. M. York and P. A. Kollman, *Amber 2018*, 2018.

44 H. J. C. Berendsen, D. van der Spoel and R. van Drunen, GROMACS: A message-passing parallel molecular dynamics implementation, *Comput. Phys. Commun.*, 1995, **91**, 43–56.

45 C. Caleman, P. J. van Maaren, M. Hong, J. S. Hub, L. T. Costa and D. van der Spoel, Force field benchmark of organic liquids: density, enthalpy of vaporization, heat capacities, surface tension, isothermal compressibility, volumetric expansion coefficient, and dielectric constant, *J. Chem. Theory Comput.*, 2011, **8**, 61–74.

46 J. Hachmann, M. A. F. Afzal, M. Haghighatlari and Y. Pal, Building and deploying a cyberinfrastructure for the data-driven design of chemical systems and the exploration of chemical space, *Mol. Simul.*, 2018, **44**, 921–929.

47 M. A. F. Afzal, G. Vishwakarma, J. A. Dudwadkar, M. Haghighatlari and J. Hachmann, *ChemLG – A Program Suite for the Generation of Compound Libraries and the Survey of Chemical Space*, 2019.

48 L. Pu, M. Naderi, T. Liu, H. Wu, S. Mukhopadhyay and M. Brylinski, eToxPred: a machine learning-based approach to estimate the toxicity of drug candidates, *BMC Pharmacol. Toxicol.*, 2019, **20**, 2.

49 Y. Pal, W. S. Evangelista, M. A. F. Afzal, M. Haghighatlari and J. Hachmann, *ChemHTPS – An Automated Virtual High-Throughput Screening Platform*, 2019.

50 M. Haghighatlari, G. Vishwakarma, D. Altarawy, R. Subramanian, B. U. Kota, A. Sonpal, S. Setlur and J. Hachmann, *ChemML: A Machine Learning and Informatics Program Package for the Analysis, Mining, and Modeling of Chemical and Materials Data*, ChemRxiv, 8323271, 2019.

51 M. Haghighatlari, G. Vishwakarma, D. Altarawy, R. Subramanian, B. U. Kota, A. Sonpal, S. Setlur and J. Hachmann, *ChemML – A Machine Learning and Informatics Program Package for the Analysis, Mining, and Modeling of Chemical and Materials Data*, 2019.

52 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, Scikit-learn: Machine learning in python, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.

53 *Talete srl, DRAGON (Software for molecular descriptor calculation)*, 2011.

54 G. Piacenza, G. Legsai, B. Blaive and R. Gallo, Molecular volumes and densities of liquids and solids by molecular mechanics estimation and analysis, *J. Phys. Org. Chem.*, 1996, **9**, 427–432.

55 D. Rogers and M. Hahn, Extended-Connectivity Fingerprints, *J. Chem. Inf. Model.*, 2010, **50**, 742–754.

56 N. Ramaswamy, N. Bauman, J. S. Dixon and R. Venkataraghavan, Topological torsion: a new molecular descriptor for SAR applications. Comparison with other descriptors, *J. Chem. Inf. Comput. Sci.*, 1987, **27**, 82–85.

57 R. E. Carhart, D. H. Smith and R. Venkataraghavan, Atom pairs as molecular features in structure–activity studies: definition and applications, *J. Chem. Inf. Comput. Sci.*, 1985, **25**, 64–73.

58 H. L. Morgan, The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service, *J. Chem. Doc.*, 1965, **5**, 107–113.

59 M. A. F. Afzal, *From virtual high-throughput screening and machine learning to the discovery and rational design of polymers for optical applications*, Ph.D. thesis, University at Buffalo, 2018.

60 A. Krylov, T. L. Windus, T. Barnes, E. Marin-Rimoldi, J. A. Nash, B. Pritchard, D. G. A. Smith, D. Altarawy, P. Saxe, C. Clementi, T. D. Crawford, R. J. Harrison, S. Jha, V. S. Pande and T. Head-Gordon, Perspective: Computational chemistry software and its advancement as illustrated through three grand challenge cases for molecular science, *J. Chem. Phys.*, 2018, **149**, 180901.

61 N. Wilkins-Diehr and T. D. Crawford, NSF's inaugural software institutes: The science gateways community institute and the molecular sciences software institute, *Comput. Sci. Eng.*, 2018, **20**, 26–38.