PCCP

PAPER

View Article Online
View Journal | View Issue

Cite this: *Phys. Chem. Chem. Phys.*, 2020, **22**, 19687

A machine-learning-assisted study of the permeability of small drug-like molecules across lipid membranes

Guang Chen, Zhiqiang Shen and Ying Li *

Study of the permeability of small organic molecules across lipid membranes plays a significant role in designing potential drugs in the field of drug discovery. Approaches to design promising drug molecules have gone through many stages, from experiment-based trail-and-error approaches, to the well-established avenue of the quantitative structure–activity relationship, and currently to the stage guided by machine learning (ML) and artificial intelligence techniques. In this work, we present a study of the permeability of small drug-like molecules across lipid membranes by two types of ML models, namely the least absolute shrinkage and selection operator (LASSO) and deep neural network (DNN) models. Molecular descriptors and fingerprints are used for featurization of organic molecules. Using molecular descriptors, the LASSO model uncovers that the electro-topological, electrostatic, polarizability, and hydrophobicity/hydrophilicity properties are the most important physical properties to determine the membrane permeability of small drug-like molecules. Additionally, with molecular fingerprints, the LASSO model suggests that certain chemical substructures can significantly affect the permeability of organic molecules, which closely connects to the identified main physical properties. Moreover, the DNN model using molecular fingerprints can help develop a more accurate mapping between molecular structures and their membrane permeability than LASSO models. Our results provide deep understanding of drug–membrane interactions and useful guidance for the inverse molecular design of drug-like molecules. Last but not least, while the current focus is on the permeability of drug-like molecules, the methodology of this work is general and can be applied for other complex physical chemistry problems to gain molecular insights.

Received 16th June 2020,
Accepted 27th July 2020

DOI: 10.1039/d0cp03243c

rsc.li/pccp

1 Introduction

Permeability of small drug-like molecules across lipid membranes characterizes one of the most important physicochemical properties of potential drugs.[1–6] Study of the passive permeation of drug molecules, driven by a concentration gradient, is of great significance to understand the molecular mechanism behind, and most importantly, to facilitate new drug design in pharmaceutical applications.[7–9]

The most widely adopted metric to evaluate permeability is the partition coefficient of a molecule, which is physically related to the potential of mean force (PMF) and local diffusivity across lipid membranes by the following inhomogeneous solubility-diffusion model:[2,10]

$$P^{-1} = \int_z \frac{\exp(G(z)/k_B T)}{D(z)} dz \quad (1)$$

where P is the permeability coefficient, k_B and T are the Boltzmann constant and absolute temperature, and $G(z)$ and $D(z)$ are the PMF profile and local diffusivity distribution along the direction of membrane thickness z, respectively. Another common metric is to evaluate PMF alone since diffusion is a physical process mainly driven by concentration gradient while relatively insensitive to molecular types.[11,12] Therefore, in computer simulations, the distribution of diffusivity across lipid membranes is assumed to be the same for small organic molecules for simplicity.[12]

There are several different ways to quantify membrane permeability. Experimental measurements of the partition coefficient of small organic molecules across certain membranes can be carried out by, for example, high-performance liquid chromatography (HPLC)[14] and the shake-flask method.[15] However, experimental measurements are very time- and cost-consuming, which makes them intractable by doing one-by-one screening of massive candidates of molecules. Additionally, they can hardly provide a passive transport mechanism at the molecular level.[16] Furthermore, since the chemical space of potential drug-like molecules is extremely large, as is approximated

Department of Mechanical Engineering and Polymer Program, Institute of Materials Science, University of Connecticut, Storrs, CT 06269, USA.
E-mail: yingli@engr.uconn.edu; Fax: +1-860-486-5088; Tel: +1-860-486-7110

to contain 10^{60}–10^{100} molecules,[17,18,19,20] a study of a small range of organic molecules in the whole chemical space is not universal and heuristic. As a result, experimental methods are not suitable for high-throughput screening (HTS),[12] which is the most common approach to predict the pharmacokinetic properties and screen potential drugs in pre-clinical drug development.[21]

On the other hand, physics-based molecular dynamics (MD) simulation provides a tractable way to study drug permeability. Through MD simulations, the PMF profile and local diffusivity can be obtained simultaneously, which (when necessary) can be used to compute the permeability coefficient P using the inhomogeneous solubility-diffusion model. For example, using all-atom MD (AAMD) simulations of the PMF, Kim *et al.* were able to verify the selectivity of certain antibiotics to target only bacteria's membrane which does no harm to the mammalian membrane in their molecular design of antibiotics.[9] However, AAMD modeling of the permeation of drug-like molecules across lipid membranes is extremely computationally expensive to explore the vast chemical space.[2] To deal with this issue, coarse-grained molecular dynamics (CGMD),[22,23,24,25,26] by reducing the complexity and degree of freedom of the simulation system through coarse-graining, enables accelerated exploration of the large chemical space with reasonable computational cost. A brief illustration of the CGMD simulation framework is given in Fig. 1a. With this simulation technique, the publicly available data of membrane permeability is enriched. For example, a recent work by Menichetti *et al.*[12] explored the chemical space of 511 427 small drug-like molecules using CGMD simulations by considering the coarse-grained degree and hydrophobicity of these molecules.

In addition to experimental measurements and MD simulations, statistical methods, such as linear regression, have long been used to study the quantitative structure–activity relationship (QSAR) in pharmaceutical engineering.[27,28] Thanks to the recent advancements of machine learning (ML) and artificial intelligence techniques, especially deep learning (DL), deep neural network (DNN) based methods have been another important workhorse for permeability prediction. Using several layers of perceptrons, DNN can learn any continuous functions,[29] which is one of the reasons why it is so powerful and popular nowadays. Therefore, DL has the potential to find the complex and underlying relationships between a molecular structure and its permeability with high accuracy and efficiency. Though DNN-based models are very advantageous in some situations, other ML methods such as the LASSO model can still be beneficial in revealing molecular insights into drug–membrane interactions, as will be presented in this study.

In the development of ML models, the efficiency of the ML models to make new predictions instantaneously with less cost is of great importance in the mind of the developers. For example, though it is found that[12] acidity pK_a and bulk partitioning free energy barrier from water to membrane midplane ΔG have an extremely high correlation with membrane permeability $\log P$, these two descriptors are not readily available to make instantaneous predictions for new molecules. Rather, expensive computer simulations are needed to obtain these two descriptors in order to feed into ML models. Thus, in this work, we aim to link molecular descriptors and fingerprints[30] to the permeability of small organic molecules, which are easily obtained using popular cheminformatics packages, *e.g.* RDKit.[31]

Moreover, taking advantage of ML techniques to study drug–membrane interactions, there are two main questions we are trying to answer. Firstly, what are the main features determining the permeability of small organic molecules across lipid membranes? Secondly, can an accurate structure–property relationship be constructed using ML methods? We find that LASSO and DNN models are able to answer these two questions, respectively.

In this paper, we adopt two types of ML methods and two different representations of organic molecules to study the drug–membrane interaction problem as illustrated in Fig. 1b. In the first scenario, a linear regression method called LASSO[32] is adopted to quickly find the main features (descriptors and chemical substructures) of the molecules. The most important

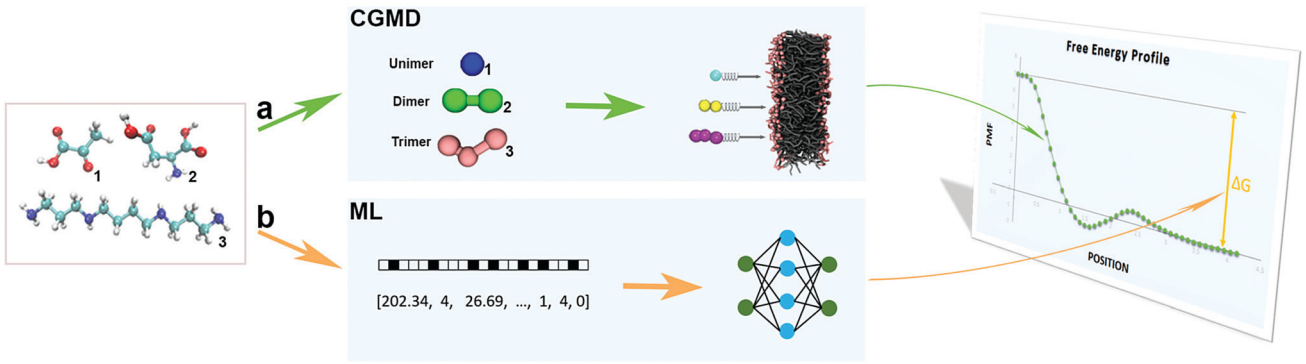

Fig. 1 Computational methods for drug–membrane interactions. (a) Coarse-grained molecular dynamics (CGMD) simulations in which three examples of organic molecule are first mapped to their corresponding Martini representations, as indicated by the numeric indexes. The umbrella sampling method[13] is then used to quantify their potential of mean force (PMF) profiles across lipid membranes (for simplicity only one profile is displayed); (b) the machine learning (ML)-assisted approach in which the organic molecules are firstly converted to their molecular fingerprints and/or descriptors. These features are then used by ML models to predict free energy barriers across lipid membranes. It is assumed that ML models are established by training on the existing CGMD data.

molecular properties and substructures are revealed by the LASSO model. In the second scenario, the DNN method is employed to build a more accurate predictive model, in comparison with the LASSO model, linking molecular structures to the permeability property for organic molecules. We expect that our results can be further applied to design drug-like molecules with different membrane permeabilities in the near future.

2 Computational method

In a ML-based study, the database, featurization, and ML models are key ingredients,[30] which are described in detail in this section.

2.1 Database

The database of this study comes from the published literature,[12,33,34] where data containing small drug-like molecules are publicly available. Through the coarse-grained Martini model,[35] the authors were able to perform high-throughput CG simulations of drug-like molecules across a 1,2-dioleoyl-*sn-glycero*-3-phosphocholine (DOPC) lipid bilayer through the umbrella sampling technique.[36] In this way, a large number and range of small organic molecules have been studied in detail.

The organic molecules, in the form of coarse-grained Martini beads, are divided into three categories, namely unimers, dimers, and trimers, each of which has one, two, and three Martini beads, respectively. The number of molecules in each category is shown in Fig. 2a. In the current database, unimer and dimer data are taken from CGMD simulations,[12] while trimer data are from taken CG Monte Carlo (CGMC) simulations.[33] One issue with the public data is that in the original CGMD database, the label is bulk partitioning free energy of water/octanol for unimers and dimers, but not that of water/membrane. However, these data can be easily converted to the bulk partitioning free energy of water/membrane since a linear relation exists between bulk partitioning of water/octanol and water/membrane for unimers and dimers.[37] Therefore, a big and valid chemical database containing a large number of molecules and associated permeability values (ΔG) is obtained for the current ML-based study.

To have direct visualization of the chemical space and PMF distribution, using representative data points is a more cost-effective way. Visualization of the total database of 770 231 data points would otherwise be very time-consuming and exhausting. Therefore, 8000 data points are extracted from the total database using the *K*-means clustering technique[38] based on similarities between molecules in the form of Morgan fingerprints in 1024 bit.[39] Specifically, 1000, 4000, and 3000 data points are selected from unimers, dimers, and trimers, respectively, as shown in Fig. 2b. During clustering of these molecules, the total data are firstly divided into N clusters, and then one data point from each cluster is selected to form the selected database. Note that we employed *K*-means not for finding clustering features of them, but for data reduction or selection. That is, selecting representative and enough candidates for ML model development, which will be justified more in the following part.

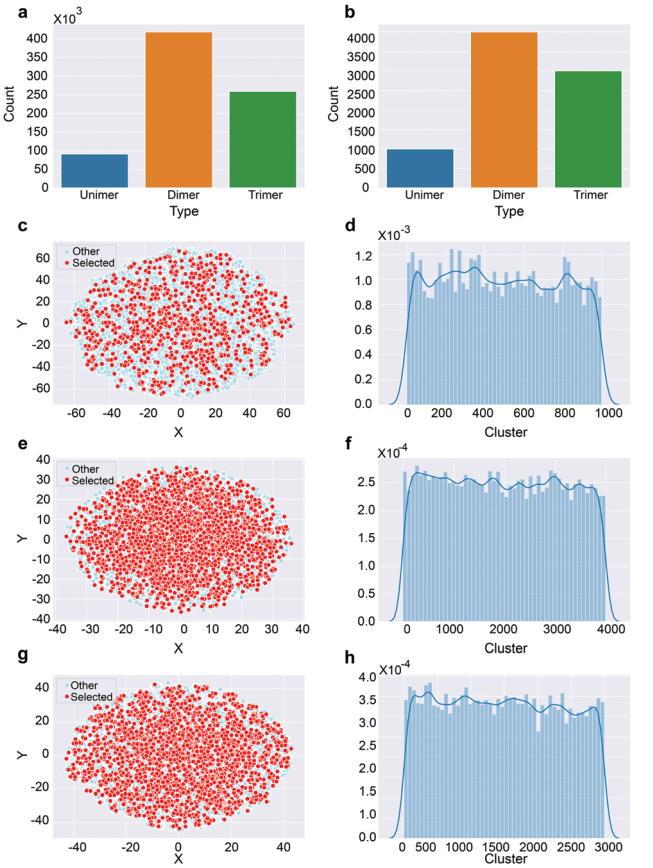

Fig. 2 Database information. (a) Count of unimers, dimers, and trimers in the total database; (b) count of unimers, dimers, and trimers in the selected database (8000 data points in total); (c) *t*-SNE plot of selected and all unimers in the database; (d) distribution of divided clusters of unimers; (e) *t*-SNE plot of selected and all dimers in the database; (f) distribution of divided clusters of dimers; (g) *t*-SNE plot of selected and all trimers in the database; and (h) distribution of divided clusters of trimers.

The adopted selection method is superior to a random selection of data points, which cannot guarantee that the selected data are uniform and representative of the total data points. Fig. 2c, e and g show the *t*-SNE plot[40] of selected and unselected data points from each category of unimers, dimers, and trimers. Fig. 2d, f and h show the cluster distributions of unimers, dimers, and trimers. One can see that the distribution is nearly uniform in each category, which confirms that the selected data are uniform and representative. In plotting the *t*-SNE figures using the scikit-learn package,[41] the fingerprints of organic molecules are transformed by principal component analysis (PCA) using 50 dimensions first and then followed by the nonlinear *t*-SNE transformation to ensure reasonable overall variance.[42]

Fig. 3 plots the PMF distribution of the selected data points. Fig. 3a shows the distribution of all categories, while Fig. 3b shows the individual distribution of each category. There are two main observations. Firstly, the overall PMF distribution for each category is nearly symmetric about the zero value and in a bimodal shape; secondly, the range and shape of the distribution for each category are different. Unimers are in a narrow

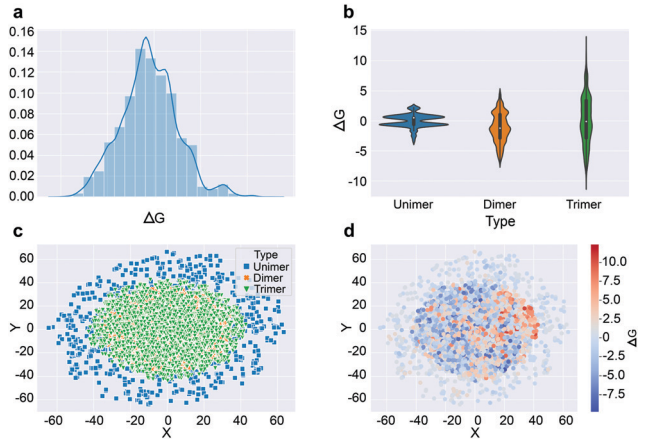

Fig. 3 Potential of mean force (PMF) distribution. (a) The density distribution of PMF of the selected data points; (b) the distribution of PMF of the selected unimers, dimers, and trimers; (c) *t*-SNE plot of the selected data points; and (d) spatial distribution of the PMF of selected data points.

and concentrated range, while the distribution curves are wider and more uniform for dimers and trimers. Fig. 3c shows the *t*-SNE plot of all the selected data points. One can see that unimers are more widely distributed than dimers and trimers; while Fig. 3d gives the corresponding PMF distribution, which is consistent with the observations in Fig. 3b.

2.2 Feature representation

In terms of molecular featurization, different methods have been widely used to predict partition coefficients, which mainly fall into two classes: substructure-based and property-based features.[43] Though the graph-based representation, an emerging molecular representation, seems to have huge potential in ML model development, it still uses atomic and pair properties to form the molecular featurization.[44] Thus, we classify it into property-based features.

In a substructure-based representation method, the simplified molecular-input line-entry system (SMILES) notation[45] for an organic molecule is usually adopted, which is then used to generate molecular fingerprints carrying the substructure information of the molecule. Note that SMILES is a specification in the form of a line notation for describing the structure of chemical species using short ASCII strings.[46] The advantage of this representation is that it converts the molecular formula into a text form that can be processed by a computer. For example, the SMILES form can not only be converted into fingerprint vectors[39] or molecular-graph-based vectors;[44] but also be employed directly as input features in ML models, *e.g.* a grammar variational autoencoder.[47]

On the other hand, property-based features or so-called molecular descriptors,[48,49] which can be obtained either by experiments or theoretical computations, are mathematical representations of spatial, physical, and/or chemical information of organic molecules. Utilization of molecular descriptors has facilitated the development of QSAR in the pharmaceutical engineering field.[50,51,52,53] In this work, molecular descriptors and fingerprints are chosen as molecular features for small organic molecules.

Currently, there are many cheminformatics packages that can easily generate molecular fingerprints and descriptors from the SMILES form of a molecule, such as RDKit.[31] RDKit is used in this work to generate Morgan fingerprints in 1024 bits and all descriptors that are available in the package (200 in total) for molecular featurization.

2.3 ML algorithms and model training

To develop QSAR, ML regression algorithms have been widely applied which define a mapping function $f = f(x;w)$ from input variables x (molecular featurization) to output variables f (free energy barrier values in the present work), where w are the associated weights in this regression function. In general, the input and output variables can be scalar, vector or tensor. A ML algorithm determines the weights by minimizing the loss function, such as the mean squared error (MSE) between predicted values and true values, on given data named 'training data'. The minimization process is the process of training, in which the model keeps updating the weights until the minimum loss is obtained. In different ML models, the mapping functions and loss functions may be designed differently, which differentiates various ML models.

In this work, the LASSO model and DNN model are adopted. The LASSO model, implemented using the scikit-learn package,[42] is able to find the main features (molecular descriptors and fingerprints) since the use of regularization can shrink the unimportant features and leave only important features. Thus, it can prevent the model from overfitting the data. Moreover, it helps to explain the PMF distribution given in Fig. 3 and to provide molecular insights into drug permeability across lipid membranes; the DNN model, implemented using the Tensorflow platform[54] and taking advantage of a large data set, is able to build a more accurate link between molecular structures and their permeabilities. In training the LASSO model, since increasing the data size does not necessarily improve the performance of regular ML models such as linear regression,[55] only the selected 8000 data points are used for model development. While in training the DNN model, all data points are used.

In the development of the ML models, it is found that the order of training data significantly affects the performance of ML models in terms of stability and robustness, as shown in the box plot of Fig. 4. It is important to see that the performance of the LASSO model using ordered data varies significantly, compared to the shuffled data. Note that the original data in the referenced literature were stored orderly from simple to

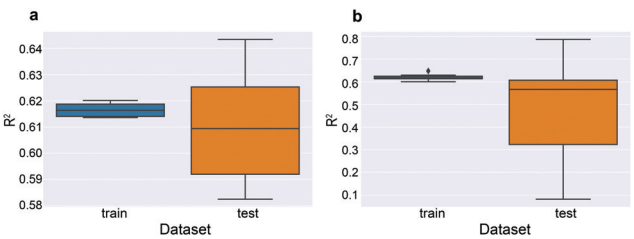

Fig. 4 Performance comparisons of the LASSO model using molecular descriptors. (a) Using shuffled selected data points and (b) using original ordered data points.

complex molecules. Thus, the data shuffle technique is adopted in training both ML models. To avoid overfitting of the data and construct an accurate ML model, various advanced techniques are selected. Specifically, in building the LASSO model, the *n*-fold cross-validation technique[56] is employed to make sure stable models with less variance of prediction ability are developed; while in building the DNN model, train/test/validation data split, dropout, early stopping[57] and checkpoint techniques are applied to select the best model during training.

3 Results

3.1 LASSO model reveals molecular features affecting membrane permeability

In training the LASSO models, molecular descriptors and fingerprints are used for molecular featurization separately. The 10-fold cross-validation technique is applied to ensure the development of a stable model with less variance of prediction capability.

The R^2 correlation and mean squared error (MSE) score of the LASSO model using molecular descriptors are plotted in Fig. 5a and b. It can be seen that the performance of the LASSO model is comparable on the training dataset and test dataset. Additionally, the deviations of scores are very small. Thus, the LASSO model obtained is stable and robust.

With the trained model at hand, important molecular descriptors are found by analyzing the weights of the LASSO model. There are 200 total weights associated with 200 molecular descriptors. All weights are ordered by the ratio of their absolute weights to the total absolute weights. It is found that only 39 weights are nonzero, which means that they have roles in the permeation process. Among these nonzero weights, 16 weights take up over 80% (about 81.13%) of the total absolute weights, the molecular descriptors associated with which are called main molecular descriptors, as listed in Table 1.

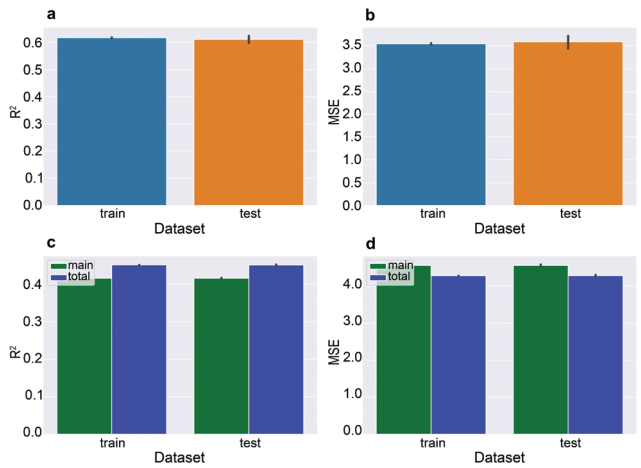

Fig. 5 Performance of the LASSO model using molecular descriptors and 10-fold cross-validation. (a and b) The R^2 correlation score and MSE score of the LASSO model using only selected database; (c and d) R^2 correlation score and MSE score for the LASSO model using only main molecular descriptors and total molecular descriptors on the total database.

Table 1 The main molecular descriptors and associated weights found by the LASSO model

Mol. des.	Abs. wt ratio	Mol. des.	Abs. wt ratio
FpDensityMorgan3	0.212	PEOE_VSA2	0.027
SMR_VSA3	0.086	PEOE_VSA6	0.026
VSA_EState9	0.070	fr_allylic_oxid	0.026
PEOE_VSA1	0.066	VSA_EState8	0.025
SlogP_VSA2	0.065	PEOE_VSA8	0.023
VSA_EState4	0.065	EState_VSA5	0.019
SlogP_VSA5	0.038	SMR_VSA5	0.018
EState_VSA8	0.030	EState_VSA1	0.018

To confirm the validity of these main molecular descriptors, the results of using only the 16 main molecular descriptors and 200 molecular descriptors on the total database are compared, as shown in Fig. 5c and d. As can be seen that using only 8% of the total molecular descriptors gives comparable performance of the LASSO model, indicating that using main features for ML model development is a cost-effective way, especially when a very large database is involved.

Further analysis of these main molecular descriptors[58] indicates that the electro-topological, electrostatic, polarizability, and hydrophobic/hydrophilic properties of organic molecules are the crucial factors influencing the permeation process. This is consistent with the like-likes-like principle.[59] Since the lipid bilayer core is hydrophobic, small molecules are less favorable to permeate if they are more polarized and hydrophilic. Therefore, when doing large scale screening, these four molecular properties can be selected as the most important values to assess their permeability properties.

These main molecular descriptors derive certain understanding of the drug–membrane interaction problem to some extent. However, they are too general to reveal molecular level insights. Specifically, it would be better if decisive substructures can be known. Towards this goal, the LASSO model is again adopted while molecular fingerprints are used as the input features of the ML model. Morgan fingerprints with 1024 bits and 2-bond length as the radius are generated for the selected 8000 data points. 10-Fold cross-validation is also employed to obtain a stable model. The R^2 and MSE scores are plotted in Fig. 6. Again, comparable performances of the LASSO model on both the training dataset and test dataset are observed for R^2 and MSE metrics, which confirms the stability of the trained LASSO model.

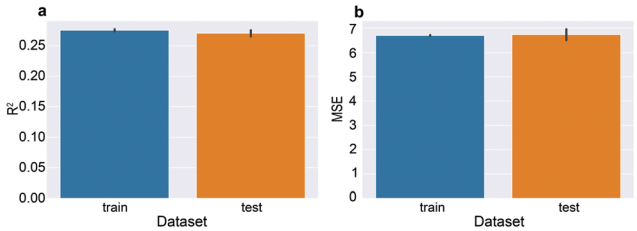

Fig. 6 Performance of the LASSO model using molecular fingerprints and 10-fold cross-validation. (a and b) The R^2 correlation score and MSE score of the LASSO model using only selected data points.

Following a similar procedure to the previous analysis of molecular descriptors, we can also extract important information of molecular substructures influencing the permeability of small organic molecules across lipid membranes. It should be noted that since a radius of two is used to generate the Morgan fingerprints, the substructures here only represent the chemical environments with at most two-bond length. Namely, only 0-bond (atom), 1-bond, and 2-bond connectivity information is captured. For hydrophobicity property evaluation, this radius is large enough; while for other properties, the size of the radius may need to change accordingly.

From the above LASSO model, only 17 weights are nonzero, which correspond to 17 bit positions, the bit indices of which are 33, 90, 128, 147, 222, 283, 294, 342, 378, 623, 650, 656, 694, 725, 807, 881, and 935. Using appropriate molecules which have a bit-on value, *viz.* 1, in these 17 bit positions by the RDKit package, the main substructures can be directly visualized. It is found that although the mapping from the bit position to substructure is not strictly one-to-one, there is still a prevailing substructure for each bit position with very high percentages of presence. Here we only plot these substructures that are shared by most of the eligible molecules. For example, there are 5939 molecules in the selected database (8000 in total) that have values of 1 at bit position 33. Among them, 5897 molecules (about 99.3% presence rate) map to the primary carbon atom as shown in Fig. 7.

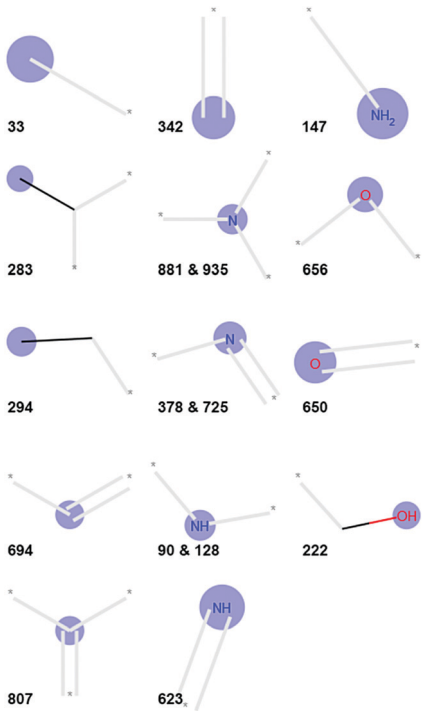

Fig. 7 Main substructures found by the LASSO model in certain bit positions (the labeled integer value), in which blue dots without label in the center and non-labeled nodes are carbon atoms. Atoms with grey bonds are 0-bond (atoms) substructures indicating their bonding information; while atoms with black bonds are the captured 1-bond (first nearest neighbor) information.

By the same procedure, 14 unique substructures are identified, including 11 distinct chemical environments of carbon, nitrogen, and oxygen atoms and three different bonds (bit position: 283, 294, and 222 which are two methyl groups connected to the secondary and tertiary carbon atoms, and one alcohol group), as shown in Fig. 7. These substructures indicate that they are critical to determine the membrane permeability of small organic molecules. This is consistent with domain knowledge[2] and can be explained qualitatively in the following way. Molecules dominant in nitrogen- and oxygen-based substructures have more significant polarizability and hydrophilicity. Therefore, they are more similar to water and are restricted to pass through lipid membranes. Thus, these molecules should have high free energy barriers. On the other hand, alkanes, which are dominant in carbon-based substructures, are favorable to permeate across lipid membranes and consequently have low (negative) free energy barriers.

To further demonstrate the importance of these substructures, 10 molecules with the highest free energy barriers and another 10 with the lowest free energy barriers are taken from the selected database, as shown in Fig. 8. The left two columns of molecules have the highest free energy barriers (difficult to permeate); while the right two columns of molecules have the lowest free energy barriers (easy to permeate), which are clearly differentiated by the main substructures found using the LASSO model. One can see that molecules in the left two columns are dominated by oxygen and nitrogen substructures;

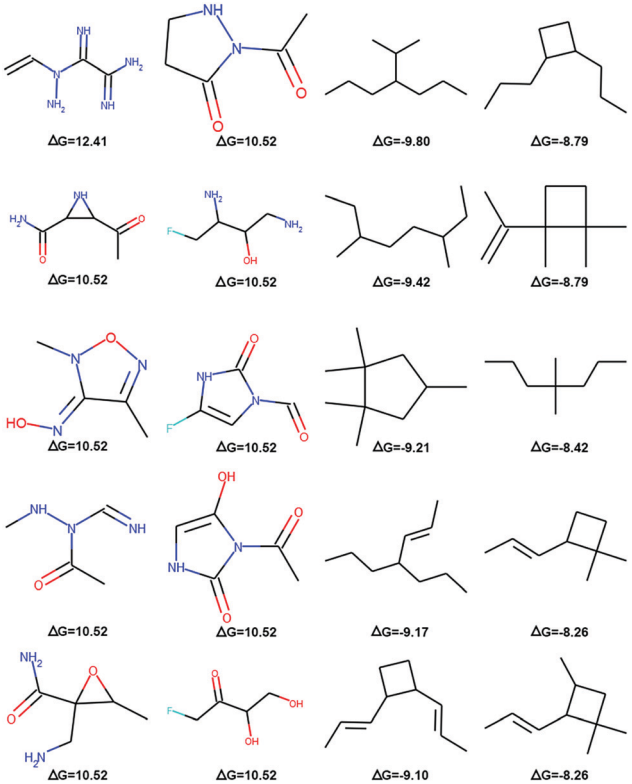

Fig. 8 Organic molecules with the highest (left two columns) and lowest (right two columns) free energy barriers from the selected database.

while the molecules in the right column have only carbon substructures. This observation is in excellent agreement with the findings of the substructures. The reason for selecting these 20 molecules rather than a random selection of molecules is to differentiate the roles of these main substructures. That is, when random molecules comprised of both nitrogen (or oxygen) and carbon atoms are selected (their free energy barriers are intermediate), it would be difficult to tell that these substructures are important since their effects are neutralized.

The finding of substructures can also help explain the distribution of free energy barriers in Fig. 3b. For unimers with just one bead, the free energy barriers are mostly either negative (more lipid-like) or positive (more water-like) with small absolute values. Thus, the distribution of unimers tends to be in a narrow range and sharp bimodal. However, for dimers and trimers with more constituents of single beads, the combinations of unimers with small negative or positive values forming dimers and trimers give rise to a wider range and flattened bimodal distribution. It is seen that the range of dimers is approximately twice that of unimers. The distribution of trimers is even much wider and more flattened than that of dimers, since they are longer and have more combinations. But there is no three-time relation of the range for trimers and unimers, which indicates that simple linear combination alone is not enough to explain the properties of complex molecules by using the properties of single constituents.

3.2 DNN model leads to accurate prediction of membrane permeability

The DNN model is very powerful in learning latent complex structure–property relationships. Nevertheless, it can easily overfit data. To avoid overfitting and obtain a precise model, the early stopping (with certain epochs patience), and dropout (rate = 0.5) techniques are applied. The train–test–validation split ratio adopted in this work is 90–5–5 since the size of the database is very large. The DNN model is first trained and tested on the training and test dataset, and then validated by the unseen validation dataset. During model training, checkpoints are set to save the best model.

The input of the DNN model is 1024-bit fingerprints; the output is a single node of the free energy barriers ΔG. Two hidden layers with 600 and 100 nodes with the rectified linear unit (ReLU) being the activation function are employed. The loss function is MSE between train and test datasets, while the evaluation metric is the mean absolute error (MAE) between predicted ΔG and true value on the validation dataset. These values are recorded to log the learning curves of the DNN model.

Fig. 9 plots the performance of the DNN model. As seen from Fig. 9a the model trained at about epoch 11 has comparable ability on both the training dataset and test dataset, at which the model is saved as the best model. Fig. 9b shows the prediction of the ΔG values of the trained DNN model on the validation dataset. A higher correlation between molecular structures in the form of fingerprints and the permeability is established, compared to that obtained by LASSO models.

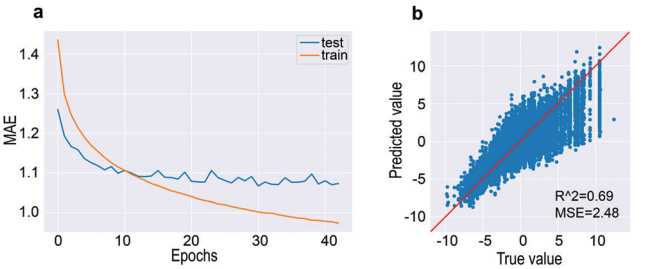

Fig. 9 Performance of the DNN model. (a) Loss evolution for training and test datasets during the training process; (b) the predictability of the trained DNN model on free energy barrier $\Delta\Delta G$ using the validation dataset.

4 Discussion

With LASSO and DNN models, we are able to identity main molecular features affecting the permeability of small drug-like molecules across the membrane, *i.e.*, important physical descriptors and substructures, and to develop a relatively accurate correlation function between a molecular structure and its permeability (ΔG). One can see that the model performance of LASSO using molecular descriptors is better than that using molecular fingerprints. This suggests that molecular descriptors are more suitable for ML model development than fingerprints in a linear regression model. In addition, one can notice that only eleven 0-bond (atoms) and three 1-bond substructures stand out by the LASSO model using molecular fingerprints. However, this never means that other substructures (*e.g.* 2-bond substructures) are not important at all. The LASSO model can actually capture 2-bond substructures, but their presence rate is much lower than these identified substructures, due to their less common presence. These fourteen substructures are discovered merely from the selected 8000 molecules, in terms of membrane permeability. If different numbers or types of molecules are adopted to feed into the LASSO model for evaluation of other properties, the main substructures may vary to some degree.

These newly identified main substructures can be further verified by commercial drugs qualitatively, as shown in Fig. 10. As seen from this figure nitrogen- and oxygen-based substructures are dominant in these drug molecules. Thus, these drugs have positive free energy barriers, *i.e.* difficult to pass through lipid membranes. This result is promising for the inverse molecular design of drug-like molecules. For inverse molecular designs, generative models are usually adopted for molecular generation.[60,61] In the case of molecular design with good

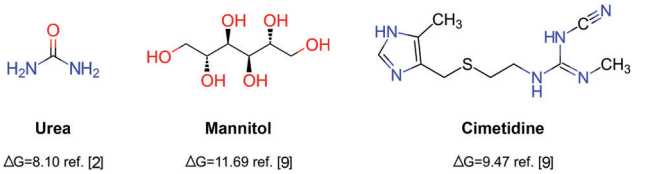

Fig. 10 Free energy barriers of three representative commercial drug molecules.

membrane permeability (lower free energy barrier ΔG), generated molecules with more carbon-based substructures will be given rewards, while molecules with more oxygen- or nitrogen-based substructures will be given penalties, through the reinforcement learning technique.[62] Consequently, new molecules are generated towards the desired range of membrane permeability. It has to be noted that for real drug design, other properties such as solubility, melting point, hydrophobicity, inhibitory activity and toxicity are also very important, and must be considered simultaneously as a multi-objective optimization problem.

Though the correlations in both LASSO models are not high, they can help identify important molecular features affecting the permeability of small molecules across lipid membranes as discussed above. Since no single method alone can solve all the questions, a method can be useful at certain aspects provided that the findings by the method are properly verified and consistent with domain knowledge. Therefore, we suggest that if one wants to find main features within the problems of interest in qualitative sense, the LASSO model might be useful; or if one wants to build a higher correlation model to make predictions, the DNN model is a better choice. Furthermore, one can even use LASSO as a pre-processing tool on the database to get some useful insights from the problem beforehand, and feed these findings by the LASSO model into the DNN model to develop a better predictive model.

Interestingly, it is seen that the performance of the LASSO model using total database is worse than that using selected 8000 data points. We believe that this is due to the feature of the database. Though many molecules are present in the database, the free energy barriers are not numerically computed one by one. Rather, they are computed by Martini coarse-graining. For example, there are only 26 unique unimer bead types in the database, while there are 92 458 small molecules mapped to these 26 bead types. For molecules mapped to the same bead type, they have the same free energy barrier, *i.e.* there are 26 unique free energy barriers of these 92 458 small molecules. This is reasonable from a physical perspective since different molecules can be similar and thus have similar free energy barriers. However, from the perspective of ML model development, it is not a good thing since these molecules are like repeated data points. This issue can also be reflected from both Fig. 8 and 9. As shown in the left two columns of Fig. 8, there are many molecules with the same positive free energy barriers (ΔG = 10.52). Similarly, in Fig. 9b, there are many different predicted free energy barriers corresponding to the same true values of ΔG. This leads to very significant vertical patterns in the positive region of ΔG. This issue is especially worthy of notice since currently the CGMD simulation is a main computational source to provide big data for data-driven studies such as this work. Possible solution to improve the ML model includes pre-processing of the data before model development. For example, one can only use representative data, rather than the total data for model development; and perhaps it is better to use Martini bead types as input features of ML models. We leave it for future studies, as it is not the scope of the present study.

5 Conclusions

In this work, two types of ML models, namely LASSO and DNN models, are used to investigate the drug–membrane interaction problem. The LASSO model using molecular descriptors reveals that electro-topological, electrostatic, polarizability, and hydrophobic/hydrophilic properties of a molecule are critical properties to determine its membrane permeability. Additionally, using molecular fingerprints integrated with the LASSO model, 14 unique substructures are identified, which are in excellent agreement with the main molecular descriptors and domain knowledge. Last but not least, using the DNN model, a relatively higher correlation between molecular structures and membrane permeability is developed. These findings can help us understand the physical problem of drug–membrane interaction and provide guidance for the inverse molecular design of drug-like molecules in the near future.

Conflicts of interest

The authors declare no competing interests.

Acknowledgements

G. C., Z. S. and Y. L. are grateful for support by the National Science Foundation under the grant no. OAC-1755779 and Department of Mechanical Engineering at the University of Connecticut. This work was partially supported by a fellowship grant (to Z. S.) from GE's Industrial Solutions Business Unit under a GE-UConn partnership agreement. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of Industrial Solutions or UConn. This research benefited from the computational resources and staff contributions provided by the Booth Engineering Center for Advanced Technology (BECAT) at the University of Connecticut. Part of this work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by the National Science Foundation grant number ACI-1053575.

References

1 J. K. Seydel and M. Wiese, *Drug-membrane interactions: analysis, drug distribution, modeling*, John Wiley & Sons, 2009, vol. 15.
2 C. T. Lee, J. Comer, C. Herndon, N. Leung, A. Pavlova, R. V. Swift, C. Tung, C. N. Rowley, R. E. Amaro and C. Chipot, *et al.*, *J. Chem. Inf. Model.*, 2016, **56**, 721–733.
3 R. M. Venable, A. Krämer and R. W. Pastor, *Chem. Rev.*, 2019, **119**, 5954–5997.
4 A. Ghysels, A. Krämer, R. M. Venable, W. E. Teague, E. Lyman, K. Gawrisch and R. W. Pastor, *Nat. Commun.*, 2019, **10**, 1–12.
5 P. Chen, H. Yue, X. Zhai, Z. Huang, G.-H. Ma, W. Wei and L.-T. Yan, *Sci. Adv.*, 2019, **5**, eaaw3192.

6 P. Chen, Z. Xu, G. Zhu, X. Dai and L.-T. Yan, *Phys. Rev. Lett.*, 2020, **124**, 198102.
7 W. Kim, A. D. Steele, W. Zhu, E. E. Csatary, N. Fricke, M. M. Dekarske, E. Jayamani, W. Pan, B. Kwon and I. F. Sinitsa, *et al.*, *ACS Infect. Dis.*, 2018, **4**, 1540–1545.
8 W. Kim, W. Zhu, G. L. Hendricks, D. Van Tyne, A. D. Steele, C. E. Keohane, N. Fricke, A. L. Conery, S. Shen and W. Pan, *et al.*, *Nature*, 2018, **556**, 103–107.
9 W. Kim, G. Zou, T. P. Hari, I. K. Wilt, W. Zhu, N. Galle, H. A. Faizi, G. L. Hendricks, K. Tori and W. Pan, *et al.*, *Proc. Natl. Acad. Sci. U. S. A.*, 2019, **116**, 16529–16534.
10 J. M. Diamond and Y. Katz, *J. Membr. Biol.*, 1974, **17**, 121–154.
11 T. S. Carpenter, D. A. Kirshner, E. Y. Lau, S. E. Wong, J. P. Nilmeier and F. C. Lightstone, *Biophys. J.*, 2014, **107**, 630–641.
12 R. Menichetti, K. H. Kanekal and T. Bereau, *ACS Cent. Sci.*, 2019, **5**, 290–298.
13 G. M. Torrie and J. P. Valleau, *J. Comput. Phys.*, 1977, **23**, 187–199.
14 L. R. Snyder, J. J. Kirkland and J. L. Glajch, *Practical HPLC method development*, John Wiley & Sons, 2012.
15 A. Andrés, M. Rosés, C. Ràfols, E. Bosch, S. Espinosa, V. Segarra and J. M. Huerta, *Eur. J. Pharm. Sci.*, 2015, **76**, 181–191.
16 W. Shinoda, *Biochim. Biophys. Acta*, 2016, **1858**, 2254–2265.
17 G. Schneider and U. Fechner, *Nat. Rev. Drug Discovery*, 2005, **4**, 649–663.
18 P. G. Polishchuk, T. I. Madzhidov and A. Varnek, *J. Comput.-Aided Mol. Des.*, 2013, **27**, 675–679.
19 H. S. Chan, H. Shan, T. Dahoun, H. Vogel and S. Yuan, *Trends Pharmacol. Sci.*, 2019, **40**, 592–604.
20 P. S. Gromski, A. B. Henson, J. M. Granda and L. Cronin, *Nat. Rev. Chem.*, 2019, **3**, 119–128.
21 C. Vraka, L. Nics, K.-H. Wagner, M. Hacker, W. Wadsak and M. Mitterhauser, *Nucl. Med. Biol.*, 2017, **50**, 1–10.
22 L. Monticelli, S. K. Kandasamy, X. Periole, R. G. Larson, D. P. Tieleman and S.-J. Marrink, *J. Chem. Theory Comput.*, 2008, **4**, 819–834.
23 M. Orsi and J. W. Essex, *Soft Matter*, 2010, **6**, 3797–3808.
24 S. J. Marrink and D. P. Tieleman, *Chem. Soc. Rev.*, 2013, **42**, 6801–6822.
25 S. M. Loverde, *J. Phys. Chem. Lett.*, 2014, **5**, 1659–1665.
26 A. Centi, A. Dutta, S. H. Parekh and T. Bereau, *Biophys. J.*, 2020, **118**, 1321–1332.
27 H. Kubinyi, *3D QSAR in drug design: volume 1: theory methods and applications*, Springer Science & Business Media, 1993, vol. 1.
28 H. Kubinyi, *Drug Discovery Today*, 1997, **2**, 457–467.
29 B. C. Csáji, *et al.*, *Faculty of Sciences*, Etvs Lornd University, Hungary, 2001, vol. 24, p. 7.
30 G. Chen, Z. Shen, A. Iyer, U. F. Ghumman, S. Tang, J. Bi, W. Chen and Y. Li, *Polymers*, 2020, **12**, 163.
31 RDKit, Open-source cheminformatics, http://www.rdkit.org, Online, accessed 11-November-2019.
32 R. Tibshirani, *J. R. Stat. Soc. B*, 1996, **58**, 267–288.
33 C. Hoffmann, R. Menichetti, K. H. Kanekal and T. Bereau, *Phys. Rev. E*, 2019, **100**, 033302.
34 C. Hoffmann, A. Centi, R. Menichetti and T. Bereau, *Sci. Data*, 2020, **7**, 1–7.
35 T. Bereau and K. Kremer, *J. Chem. Theory Comput.*, 2015, **11**, 2783–2791.
36 S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen and P. A. Kollman, *J. Comput. Chem.*, 1992, **13**, 1011–1021.
37 R. Menichetti, K. H. Kanekal, K. Kremer and T. Bereau, *J. Chem. Phys.*, 2017, **147**, 125101.
38 S. Lloyd, *IEEE Trans. Inf. Theory*, 1982, **28**, 129–137.
39 D. Rogers and M. Hahn, *J. Chem. Inf. Model.*, 2010, **50**, 742–754.
40 L. V. D. Maaten and G. Hinton, *J. Mach. Learn. Res.*, 2008, **9**, 2579–2605.
41 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss and V. Dubourg, *et al.*, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.
42 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, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.
43 R. Mannhold, G. I. Poda, C. Ostermann and I. V. Tetko, *J. Pharm. Sci.*, 2009, **98**, 861–893.
44 S. Kearnes, K. McCloskey, M. Berndl, V. Pande and P. Riley, *J. Comput.-Aided Mol. Des.*, 2016, **30**, 595–608.
45 D. Weininger, *J. Chem. Inf. Comput. Sci.*, 1988, **28**, 31–36.
46 E. Anderson, G. D. Veith and D. Weininger, *Environmental Research Laboratory-Duluth*, 1987.
47 P. B. Jørgensen, M. Mesta, S. Shil, J. M. Garca Lastra, K. W. Jacobsen, K. S. Thygesen and M. N. Schmidt, *J. Chem. Phys.*, 2018, **148**, 241735.
48 R. Todeschini and V. Consonni, *Handbook of molecular descriptors*, John Wiley & Sons, 2008, vol. 11.
49 R. Tauler, B. Walczak and S. D. Brown, *Comprehensive chemometrics: chemical and biochemical data analysis*, Elsevier, 2009.
50 M. Karelson, *Molecular descriptors in QSAR/QSPR*, Wiley-Interscience, New York, 2000, vol. 230.
51 T. Puzyn, J. Leszczynski and M. T. Cronin, *Recent advances in QSAR studies: methods and applications*, Springer Science & Business Media, 2010, vol. 8.
52 K. Varmuza, M. Dehmer and D. Bonchev, *Statistical modelling of molecular descriptors in QSAR/QSPR*, Wiley Online Library, 2012.
53 E. N. Muratov, J. Bajorath, R. P. Sheridan, I. V. Tetko, D. Filimonov, V. Poroikov, T. I. Oprea, I. I. Baskin, A. Varnek and A. Roitberg, *et al.*, *Chem. Soc. Rev.*, 2020, **49**, 3525–3564.
54 M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, *et al.*, 2016, arXiv preprint arXiv:1603.04467.
55 R. Gómez-Bombarelli, J. Aguilera-Iparraguirre, T. D. Hirzel, D. Duvenaud, D. Maclaurin, M. A. Blood-Forsythe, H. S. Chae, M. Einzinger, D.-G. Ha and T. Wu, *et al.*, *Nat. Mater.*, 2016, **15**, 1120–1127.

56 R. R. Picard and R. D. Cook, *J. Am. Stat. Assoc.*, 1984, **79**, 575–583.
57 R. Caruana, S. Lawrence and C. L. Giles, *Advances in neural information processing systems*, 2001, pp. 402–408.
58 P. Labute, *J. Mol. Graphics Modell.*, 2000, **18**, 464–477.
59 V. V. Yaminsky and E. A. Vogler, *Curr. Opin. Colloid Interface Sci.*, 2001, **6**, 342–349.
60 B. Sanchez-Lengeling and A. Aspuru-Guzik, *Science*, 2018, **361**, 360–365.
61 D. C. Elton, Z. Boukouvalas, M. D. Fuge and P. W. Chung, *Mol. Syst. Des. Eng.*, 2019, **4**, 828–849.
62 M. Popova, O. Isayev and A. Tropsha, *Sci. Adv.*, 2018, **4**, eaap7885.