Predicting drug properties with parameter-free machine learning: pareto-optimal embedded modeling (POEM)

MACHINE LEARNING
Science and Technology

PAPER

Predicting drug properties with parameter-free machine learning: pareto-optimal embedded modeling (POEM)

Andrew E Brereton[1], **Stephen MacKinnon**[1], **Zhaleh Safikhani**[1,2], **Shawn Reeves**[1], **Sana Alwash**[1], **Vijay Shahani**[1] **and Andreas Windemuth**[1]

[1] Cyclica Inc., 207 Queens Quay W Suite 420, Toronto, ON M5J 1A7, Canada
[2] Vector Institute for Artificial Intelligence, Toronto, ON, Canada

E-mail: windemut@yahoo.com

Keywords: POEM, small molecule, ADMET, prediction, molecular graph convolution, SVM, random forest

Supplementary material for this article is available online

OPEN ACCESS

RECEIVED
5 December 2019

REVISED
6 April 2020

ACCEPTED FOR PUBLICATION
14 April 2020

PUBLISHED
19 May 2020

Abstract

The prediction of absorption, distribution, metabolism, excretion, and toxicity (ADMET) of small molecules from their molecular structure is a central problem in medicinal chemistry with great practical importance in drug discovery. Creating predictive models conventionally requires substantial trial-and-error for the selection of molecular representations, machine learning (ML) algorithms, and hyperparameter tuning. A generally applicable method that performs well on all datasets without tuning would be of great value but is currently lacking. Here, we describe pareto-optimal embedded modeling (POEM), a similarity-based method for predicting molecular properties. POEM is a non-parametric, supervised ML algorithm developed to generate reliable predictive models without need for optimization. POEM's predictive strength is obtained by combining multiple different representations of molecular structures in a context-specific manner, while maintaining low dimensionality. We benchmark POEM relative to industry-standard ML algorithms and published results across 17 classifications tasks. POEM performs well in all cases and reduces the risk of overfitting.

1. Introduction

Chemical activity predictions continue to present a longstanding challenge with practical importance during pharmaceutical research and development. In particular, predictive tasks that associate chemical structures to their activity are known as quantitative structure activity relationships (QSARs) [1]. In modern drug design programs, QSARs are used for modeling specific target biological activities or broader pharmacokinetic behaviours, including the absorption, distribution, metabolism, excretion, and toxicity of drug candidate molecules, collectively referred to as *ADMET* [2–4]. All machine learning (ML) approaches to predict chemical activity have three fundamental requirements [1]: a *library* of diverse reference molecules with a known property to predict (labeled examples) [2], a form of molecular representation and [3] a discriminative supervised learning algorithm. In practical drug discovery applications, multiple algorithms and molecular representations are explored and optimized on a trial-and-error basis [5], since their performance varies considerably from predictive task to predictive task. This process generates final models which can vary considerably, even between extremely similar predictive tasks, in a manner that is seemingly arbitrary and often difficult to interpret.

Molecular representations are a key component for predictive modeling of chemical activity. Different representations can be more or less relevant to different predictive tasks, not unlike the context-dependent use of different molecular representations in a laboratory. For example, acetylsalicylic acid can be described by several different common names: IUPAC name, chemical formula ($C_9H_8O_4$), a simplified molecular-input line-entry system (SMILES) string (e.g. 'O=C(C)Oc1ccccc1C(=O)O') [6, 7], or a drawing of a 2-D structure (e.g. as in figure 1). Each of these representations is valid, but describe acetylsalicylic acid with differing levels of information content. For example, acetozone has the same chemical formula as

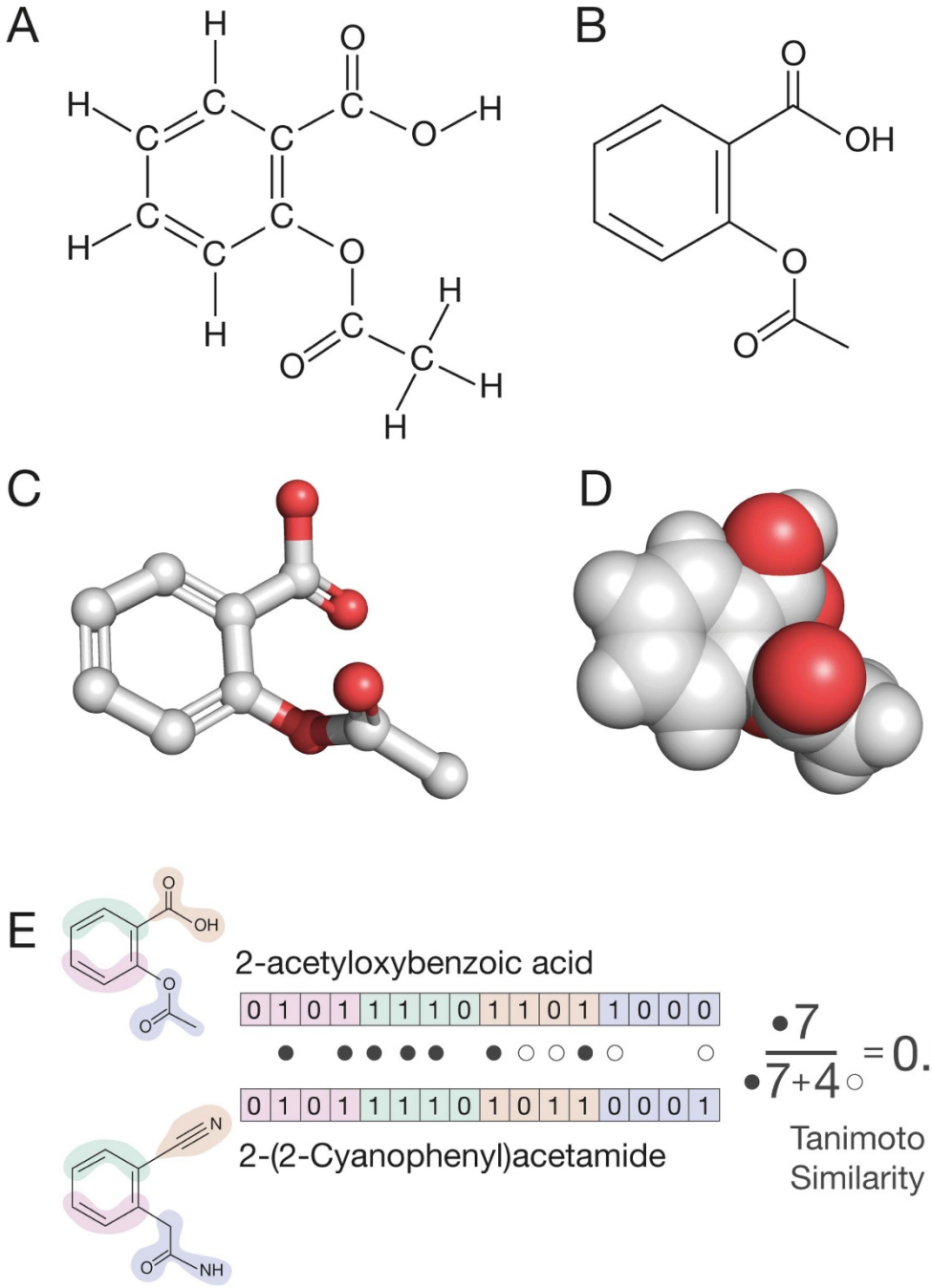

Figure 1. Multiple representations of acetylsalicylic acid with varying levels of information. (A) A 2D representation that explicitly shows all atoms and bonds within ASA, (B) a 2D line-diagram of ASA with details removed but isomorphic to the previous diagram, (C) a 3D depiction of ASA which highlights the geometry between each atom, and (D) a 3D space-filled depiction of ASA, which gives a sense of relative atom sizes. (E) Simplified depiction of a hypothetical molecular fingerprint showing two molecules converted into unique bits that can be directly compared (e.g. bits are used to calculate Tanimoto similarity [8] or distance (1—Tanimoto similarity)).

acetylsalicylic acid, but not the same 2-D structure; this is consistent with the observation that the chemical formula is a higher entropy description of a molecule than a drawing of its 2-D structure. For most ML applications, the molecular representation is defined in a way that is much simpler for a computer to extract useful information from, such as a binary or numeric vector (e.g. figure 1(E)).

Physicochemical molecular descriptors are one conceptually simple way to create such a vector: by measuring a series of known properties (e.g. mass, number of heteroatoms, charge, etc) [9]. While physicochemical descriptors have favorable interpretability, these overly simplistic representations of a molecule can lead to poor predictive power or a host of other problems associated with incorrect descriptor selections [10]. Thus, an approach relying on physicochemical molecular descriptors may be robust for a single predictive task, but often will not be generalizable.

Alternatively, molecules may be converted into a vector format in a process called molecular fingerprinting [11]. Unlike physicochemical molecule descriptors, fingerprints do not need to have any human-obvious relationship to the properties of the molecule; they are generated by following strict algorithmic rules that associate individual positions in the vector to the presence/absence of specific substructures and substructure relationships. A primary advantage gained by using fingerprints, as opposed to purely physicochemical molecular descriptors, is that they are usually much more generalizable; a fingerprint of a molecule will often be useful to some degree across many different problems. Unfortunately, this generality comes with a lack of specificity, and risks training a model on syntactic features of the representation ('noise') rather than true details of the structure [12].

Most fingerprinting methods do lose information upon encoding, resulting in vectors that tentatively represent several different molecules [13, 14]. Typically, the loss of information is coupled to specific advantages such as improved speed, lower memory usage, algorithmic advantages, or increased density of useful information. For instance, Daylight fingerprints can exist as long, variable length fingerprints, or they can be 'folded' into a fixed-length representation, gaining speed and memory efficiency at the cost of reversibility [11]. In addition to losing information, compressed (folded) fingerprints are prone to introducing algorithmic noise, leading to potential false equivalencies. While reversibility may be desirable in some applications, its is not always necessary.

The relationship between molecular features and chemical activity is established during the training stages of supervised learning algorithms. Chemical activities are predicted using industry-standard ML approaches for supervised learning, such support vector machines (SVMs), random forest modeling, ridge classification and neural networks. Typically, each algorithm has its own specific set of *hyperparameters* that influence the performance of the trained models. The problem of knowing which hyperparameters to use is itself an interesting and non-trivial topic of research [15]. In practice, the optimal combination of hyperparameter values also varies across tasks and must be calibrated through a trial-and-error based optimization process. This high computational cost of retraining models can be a deterrent for incorporating new training data at a later stage, although this drawback is partially mitigated in some neural network based approaches through recent advances in transfer learning [16, 17]. When a supervised learning algorithm successfully establishes relationships between molecular features and chemical activities, the resulting model can be used to predict the chemical activity of new, previously unseen molecules.

Model interpretability is another important factor in the selection of supervised learning algorithms. Some algorithms, for example decision trees, provide human interpretable justifications for the associations between molecular features and chemical activities. For other algorithms, such as neural networks, the reasons for any given prediction can often be extremely hard to determine. In practice, the 'black box' nature of these models can inhibit trust and possibly reduce the real-world utility of the model. Understanding this rationale can lead to new strategies for reducing bias, overfitting, and possibly even improve theory [18]. Ideally, each prediction generated by the model has a clear rationale behind it, which can be understood by the researcher using the model.

Unfortunately, the combination of variable context-dependent molecular representations, choice of supervised learning algorithms and hyperparameterization is typically inconsistent between predictive tasks. Any specific combination of these three factors optimized for a specific chemical activity model will not be applicable to predicting other activities. In practice, prediction of seemingly related chemical properties may end up requiring models with different molecular features, algorithms, and/or parameters following substantial development effort by expert chemoinformaticians.

One approach to addressing inconsistency in the choice of molecular fingerprints and tentatively boosting performance is to use multiple fingerprints simultaneously as features. It is clear that distinct molecular representations possess varying amounts of useful information relating to specific problems (i.e. each molecular representation discards information, but not every method discards *the same* information). So, two fingerprints of the same molecule produced by two different low-information fingerprint methods will share some redundant information about the structure, and contain some unique information. By combining molecular representations, a greater amount of 'true' information can be captured from the complete structure of the molecule. A naive way to combine molecular representations is to append the vectors generated by each fingerprint directly onto each other, to create a new, longer fingerprint, of multiple types. This approach has at least two major disadvantages [1]: for some supervised learning algorithms, the computational costs may become prohibitive when concatenating multiple fingerprints, due to increased length of the vector; and [2], the concatenation approach is prone to the creatively named 'curse of dimensionality' [19–21]. This phenomenon occurs when the ratio of training data to features is low, and overfitting becomes not just possible, but likely. Other methods approach this issue by using voting schemes to try to weigh consensus between models built using different fingerprints [22], or by nesting the selection of a fingerprint with hyperparameter selection to empirically determine the most effective fingerprint [23].

In this report, we describe pareto-optimal embedded modelling (POEM), a novel supervised learning method that reliably creates accurate models for predicting drug activity based on multiple representations of molecular structure. POEM evades the *curse of dimensionality* by using pareto multi-front optimization [24, 25] to massively shrink the number of dimensions that define molecular similarity, in a context-dependent manner. A Pareto optimization algorithm is a powerful general approach for identifying optimal solutions in cases where there are more than one metric to optimize, that may not always be in agreement with each other. It has found broad applications including protein structure minimization [26, 27] and lead optimization in drug design [28]. Effectively, the context-dependent dimensionality reduction introduced by POEM enables the use of multiple fingerprints to describe every molecule in the reference molecule library, without introducing a risk of overfitting. The specific use of a pareto-based approach to defining similarity ensures that all comparisons remain 'like-to-like' and avoids the need for heuristic transformations and weighting schemas. POEM also has a number of additional functional advantages: the rationale for predictions are each interpretable, the algorithm has no hyperparameters, and models can be easily updated with new labeled reference molecules. This approach was designed for the rapid generation of multiple predictive models, without the need for expert intervention. We demonstrate the generalizability and consistency of POEM across a broad range of predictive tasks by modeling 17 ADMET properties of interest to pharmaceutical drug development.

2. Materials and methods

2.1. Strategy

POEM uses a Pareto dominance definition to combine multiple definitions of molecular similarity, into a robust metric suitable for supervised learning tasks. The POEM method has four major steps [1]: fingerprinting [2], embedding the known molecules into a limited context based on similarity [3], calculating Pareto dominance relationships for the known molecules, and [4] converting these dominance relationships into similarity scores and predictions. A high-level flowchart describing this process and the outputs at each step is provided in figure S1 (available online at stacks.iop.org/MLST/1/025008/mmedia).

2.1.1. Fingerprinting

N molecular fingerprints techniques are applied to the target molecule and the reference library of **M** labeled compounds. Step 1 results in a matrix of $\mathbf{M} \times \mathbf{N}$ fingerprints for reference library and a vector of **N** fingerprints for the reference molecule. For this study, 10 diverse and widely-used fingerprints were chosen. Detailed parameters and references for these 10 fingerprints are provided in table S1.

2.1.2. Embedding known molecules

The fingerprint representations of the target molecule are embedded onto the chemical landscape of the reference library via a non-reversible transformation. For each fingerprint, Tanimoto distances [8] (figure 1(E)) are calculated between the target molecule and all reference molecules in the library. Step 2 results in a matrix of $\mathbf{M} \times \mathbf{N}$ distance values, centered on the target molecule.

2.1.3. Calculating dominance relationships

Although each molecule in the reference library is represented by a set of **N** distances, the specific values of the distances are not directly comparable. A distance of 0.4 may represent a significant match for one fingerprint, but random noise for another. To resolve this issue, Pareto dominance relationships are used to establish an overall distance. Here, the target molecule is selected as the *ideal objective* for multi-front optimization and the set of fingerprints represents **N**-dimensional space [28, 29]. Dominance relationships between molecules from the reference library defined on the basis of which molecule is *closer* to the target. One reference library molecule may be closer than another to the target for all **N** distances, or a subset of the **N** dimensions. When evaluating the dominance of one reference library molecule (**A**) to another labeled molecule (**B**) across all 10 distances, closer distances are assigned a value of 1, ties are assigned a value of 0.5, and further distances are assigned a value of 0. A comparison vector $\mathbf{AB} = [1, 0, 1, 0.5, 0, 0.5, 0.5, 1, 1, 1]$ would indicate that molecule A is more similar to the target molecule in five fingerprint representations, tied in three representations, and more dissimilar using two fingerprint representations. Step 3 results in an $\mathbf{M} \times \mathbf{M}$ symmetric matrix of dominance relationships.

In a naive Pareto scheme, a molecule would *dominate* another molecule all its distances were as close or closer to the target molecule, and at least one distance was closer [27]. POEM relaxes the naive definition of dominance, allowing a molecule to claim dominance over another, even if $\leq$10% of its distance comparisons remain further from the target. In practice, this relaxation yields more dominance relationships overall when a larger number of fingerprints is used. The added dominance relationships reduce the likelihood of ties,

which helps establish a complete ranking of all molecules in the labeled library relative to the target. Sample code for evaluating these relationships is provided in supplementary pseudocode 1.

2.1.4. Calculating fitness scores and final prediction

Labeled reference molecules are ranked by their similarity to the target molecule by converting dominance relationships to a single-value *fitness score*. For a given molecule, its fitness is defined as:

$$Fitness_i = MeanDominance_i \cdot \frac{(NumDominating_i + 0.05)}{(NumSubmitting_i + 0.05)}$$

This schema favors labeled reference molecules which compare favorably 'on average' for all fingerprints (*MeanDominance*), which dominate many other molecules (*NumDominating*), and which are not being dominated by others (*NumSubmitting*). This approach favors labeled molecules that are 'best-of-class' across all metrics of similarity to the target molecule. Sample code for ranking molecules is provided in supplementary pseudocode 2.

Labeled reference molecules are ranked according to their fitness scores and summed to provide a 'total fitness' value. In practice, fitness values vary by orders of magnitude between the most similar and dissimilar molecules. In some cases, the top few molecules could contribute the vast majority of the weight towards the summed fitness value. Alternatively, contribution towards the total fitness value may be more broadly distributed across the reference molecule library. The relative contribution of each labeled reference molecule to the total fitness score is then used as a weight towards each observed class label. Weighted averages are treated as probabilities of the target having any given label. Sample code evaluating probabilities is provided in supplementary pseudocode 3. Step 4 results in a Length **M** fitness vector of similar molecules, which is used to assign probabilities to each label class.

2.2. Benchmarking POEM to standard approaches with 17 ADMET property predictions

POEM was benchmarked relative to five standard supervised learning algorithms, across 17 predictive tasks related to drug ADMET properties. All 17 ADMET datasets were taken from public sources and range between 522 and 6505 labeled reference molecules. Each molecule was represented by a SMILES string. RDKit release 2018.09.1 was used to parse, canonicalize, and featurize small molecules. Molecules that could not be automatically processed by RDKit and molecules that have both positive and negative data labels were excluded from each dataset. Redundant data points representing experimental replicates were also removed. The *supplementary text* contains references and descriptions for each dataset used in this study, including the total number of training examples for each label after dataset cleaning. Python's scikit-learn package v0.19.1 [30] was used to build models for each of the five standard supervised learning algorithms: Gradient Boosting Classifier, Random Forest, Ridge Classifier, Stochastic Gradient Descent Classifier, and Support Vector Machine, representing a range of industry-standard supervised classifier types, which are suitable for the dataset sizes in this benchmark study. Each of the five standard supervised learning algorithms was trained using a grid search strategy for hyperparameter optimization, with an added nested layer evaluating performance separately for each of the fingerprints listed in table S1. In contrast, POEM has no hyperparameters and considers all fingerprints simultaneously. In addition, a Molecule Graph Convolution [31, 32] model was trained using the GraphConv model in the DeepChem python package [33], on all 17 datasets, to provide a comparison to state-of-the-art deep neural network methods. Results are reported both from a naive model with default parameters (2 convolution layers of size 64, one dense layer of 128, 75 features per atom, 10 training epochs), and a hyperparameter optimized model (parameters selected after 20 rounds of 5-fold cross-validated Bayesian optimization, for reference see table S4). Finally, we also evaluated POEM models that were limited to using only individual fingerprints, to provide an added comparison to a non-consensus, similarity-based classifier approach.

2.2.1. Cross validation of predictive models

For each of the five standard supervised learning approaches, models were trained and evaluated with a five-fold cross-validation strategy, using 80% of the dataset for training and hyperparameter optimization and withholding 20% for blind performance evaluation. The same 80%/20% split was used to create a *POEM test set*, which provides a direct comparison to the five standard algorithms. POEM is also amenable to full 'leave-one-out' cross validation due to the lack of a computationally-expensive training process. This corresponding *POEM Full Set* evaluation provides an indication of predictive robustness with respect to dataset size. Additionally, nested cluster validation was performed as in Mayr *et al* [34], with the added restriction that the test set must contain at least one example of each class.

Table 1. A summary of the performance (ROC AUC score) of 17 ADMET properties of POEM and best-performing, standard classifier-fingerprint combination. POEM test: 80%/20% testing split for a direct comparison to traditional approaches, POEM full: 'leave-one-out' full cross-validation.

Property	POEM full (ROC AUC)	POEM test (ROC AUC)	Best traditional (ROC AUC)	Best classifier	Best fingerprint
AMES Toxicity	0.872	0.869	0.802	Gradient Boosting	Pattern
Androgen Receptor	0.864	0.744	0.687	Gradient Boosting	Layered
Blood Brain Barrier	0.979	0.981	0.916	Stochastic Gradient Descent	Pattern
Caco-2 permeability	0.828	0.822	0.726	Gradient Boosting	Pharm Gobbi
Carcinogenic	0.726	0.659	0.678	Gradient Boosting	Layered
CYP450 2C9 Inhibitor	0.801	0.840	0.594	Stochastic Gradient Descent	Pharm Base
CYP450 2C9 Substrate	0.690	0.612	0.507	Stochastic Gradient Descent	Morgan Rad:4
CYP450 2D6 Inhibitor	0.750	0.718	0.566	Stochastic Gradient Descent	Pharm Base
CYP450 2D6 Substrate	0.778	0.688	0.594	Ridge	Morgan Rad:2
CYP450 3A4 Inhibitor	0.700	0.677	0.613	Gradient Boosting	Pattern
CYP450 3A4 Substrate	0.676	0.586	0.532	Gradient Boosting	Morgan Rad:4
Estrogen Receptor alpha	0.963	0.961	0.855	Gradient Boosting	Pharm Base
Human Intestinal Absorption	0.949	0.954	0.883	Stochastic Gradient Descent	Layered
Human Oral Bioavailability	0.770	0.763	0.676	Gradient Boosting	Pharm Base
Human Pregnane × Receptor	0.895	0.876	0.646	Ridge	Layered
P-glycoprotein Inhibitor	0.945	0.954	0.885	Gradient Boosting	Pattern
P-glycoprotein Recognition	0.955	0.943	0.855	Ridge	Morgan Rad:2

3. Results

3.1. POEM outperforms five standard supervised learning methods for predicting 17 ADMET properties

Table 1 provides a comparison between POEM performance relative to the top performing fingerprint/algorithm combinations generated from standard supervised learning approaches, across 17 different ADMET tasks. Across all tasks, the *POEM Test Set* outperforms the standard supervised learning algorithms, as determined by ROC area under the curve (AUC) score, in some cases by more than 10% (figure 2). Additionally POEM performs approximately as well, or better, than the GraphConv neural network (figure 2). Validation scores associated with standard classifiers are consistently better than test scores (figure 3), indicating some degree of overfitting. In contrast, the *POEM Full Set* scores are sometimes higher and sometimes lower than the *POEM Test Set* scores. Generally, similarity between these scores is an indication of predictive robustness and resistance to overfitting. A notable exception in both cases is the AR dataset, which shows the largest score disparity between the *POEM Full Set, POEM Test Set.* This disparity and relatively low predictive performance may indicate an insufficient representation of chemical space in the underlying dataset.

To assess predictive robustness with regards to the random test set selection, 100 different 80%/20% testing splits were performed on the Blood Brain Barrier (BBB) and the Caco-2 Permeability (Caco2) datasets (figure S2). ROC AUC was computed for all 100 using both POEM and the previously determined best model and hyperparameters for each approach (as reported in table S2). Compared to the standard classifiers, the POEM performance is higher for each test set, and overall shows less variability.

To assess POEM's ability to generalize, a nested cluster validation approach was used. This approach provides insight into how well POEM can make predictions for molecules highly dissimilar to any known reference molecules. We see that POEM does generalize well, especially for certain datasets (e.g. BBB in figure S3), though in some other cases performance is poor on dissimilar molecules (e.g. AR in figure S3). This approach was also applied to the GraphConv neural network, with comparably good results (figure S4).

We also compared POEM performance with literature-reported results for models trained on the same datasets. Hansen *et al* report expert-optimized models for the Ames Mutagenicity dataset built using molecular descriptors and seven binary classification tools [36]. The top performing model in that study was a support vector machine (SVM) with ROC AUC of 0.86, whereas the non-parametric POEM automatically generated comparable models with 0.87 AUC. This study shares three common datasets with AdmetSAR, a predictive tool that uses substructure-based descriptors and support vector machines [35]. AdmetSAR

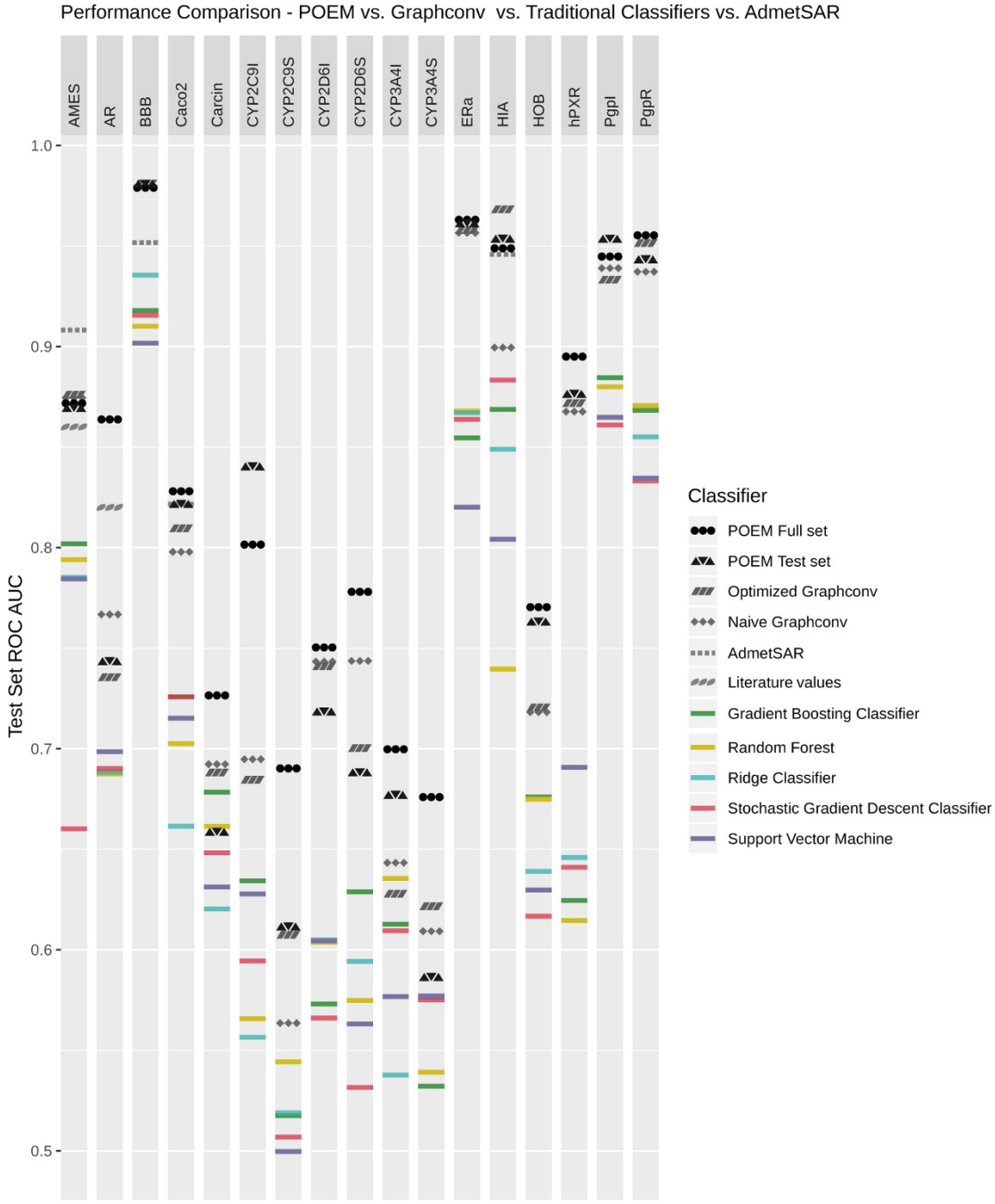

Figure 2. Comparison of POEM performance versus industry-standard classifiers and literature reported values. ROC area under the curve (AUC) scores for POEM and industry-standard ML methods tasked with predicting 17 ADMET properties are compared. Shown are the ROC AUC scores for full leave-one-out prediction with POEM (POEM full set, black circles), a 20% test set with POEM (POEM test set, black triangles), and the same 20% test set for best performing fingerprint and hyperparameters for each industry-standard ML method (colored straight lines). In addition, reported scores from AdmetSAR [35] (gray squares), an ADMET prediction tool, other literature values [31, 33] (grey ovals), and a naive (gray diamonds) and hyperparameter optimized (gray rhomboids) GraphConv [31] neural network are also shown when applicable.

reports five-fold cross-validation ROC AUC values for: Blood Brain Barrier (0.9517), Human Intestinal Absorption (0.9458), Caco-2 Permeability (0.8216). These models were generated using different combinations of three fingerprints and three binary classification algorithms [35]. The Androgen Receptor (AR) activity dataset was taken from the Tox21 Challenge, a federal collaboration involving NIH, EPA and the FDA aimed to develop better toxicity assessment methods. For this sub-challenge, 31 teams contributed different predictive models, with the leading ROC AUC at 0.828 [37]. Consistently, POEM matched or outperformed expert-developed models reported in the literature.

We also compared the standard POEM approach to a modified variation that uses individual fingerprints rather than a consensus of 10, reporting all *leave-one-out* cross validation results for each dataset in table S3.

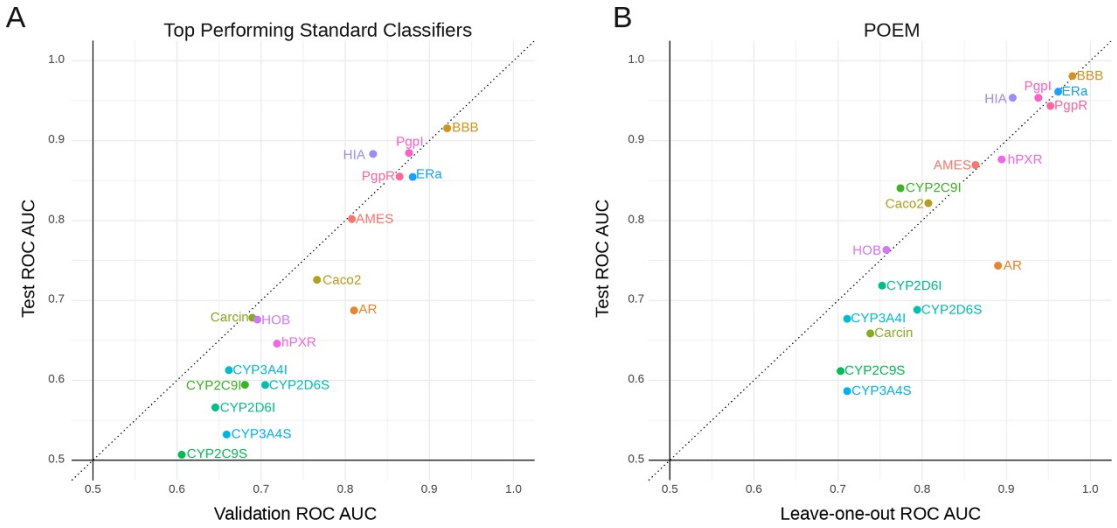

Figure 3. Comparison of ROC AUC scores on test data versus leave-one-out cross validation scores (for POEM) or pre-validation scores (for industry-standard classifiers). (A) ROC AUC scores are shown (*y*-axis) for each best performing classifier, hyperparameter, and fingerprint combination, for each of the 17 properties, for the same 20% test set as used in the POEM comparisons, and the mean of the 5-fold cross-validation score (pre-validation ROC AUC, *x*-axis), as assessed during model training. A dotted line is shown as a reference along the diagonal. (B) Shown are the POEM ROC AUC scores for each of the 17 properties being predicted, for a 20% test set (*y*-axis), and for leave-one-out cross validation on the remaining 80% set (the same set as in A) (*x*-axis). A dotted line is shown as a reference along the diagonal.

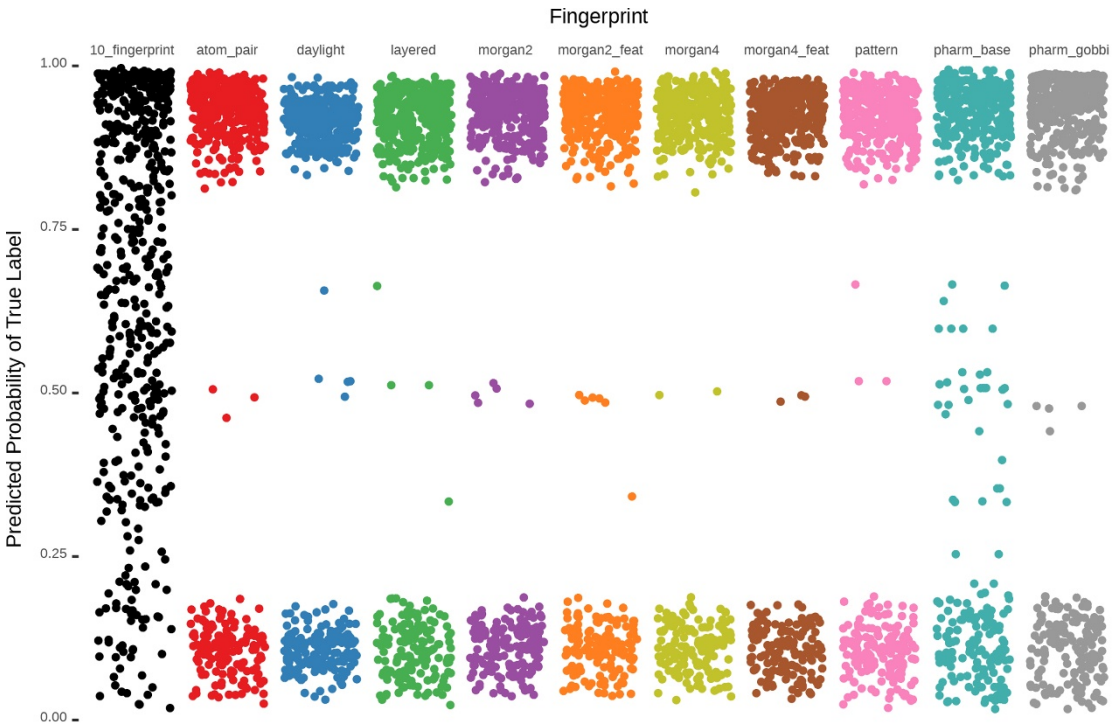

Figure 4. Comparison of individual leave-one-out POEM predictions of Caco-2 permeability using 10-fingerprint consensus and 10 single fingerprints individually. Each point represents the POEM prediction probability of the correct label for a reference labelled molecule withdrawn from the training data for a leave-one-out complete cross-validation. A probability of 1.0 represents 100% certainty of the correct prediction, and 0.0 represents 100% certainty of the incorrect prediction.

As expected, the standard consensus approach performs reliably well, appearing among the top models for each task and clearly outperforming any individual single-fingerprint across all tasks. The use of multiple fingerprints consensus also helps establish meaningful confidence scores associated with the predictive tasks. Figure 4 presents the distribution of POEM-predicted probability for the 'correct' label for the Caco-2 permeability dataset, across the consensus 10FP POEM, and each of the 10 single-fingerprint models. In this figure, data points that lie below 0.5 represent an incorrect prediction and points closer to 100% and 0% represent higher confidence predictions. The Caco-2 dataset was chosen as an example dataset to

demonstrate this principle, since the consensus approach ranked lower than three other single-fingerprint models. Performance varied across all fingerprinting methods, and was of approximately average predictive power overall. This figure demonstrates that the consensus model is better able to capture the confidence of a given prediction, as most of the observed incorrect predictions were made with lower confidence than for the single-fingerprint models. The single-fingerprint models almost exclusively produce highly confident correct predictions, or highly confident incorrect predictions.

4. Discussion

At its core, POEM is a method based on measuring similarity, conceptually similar to a K-nearest neighbor approach [21]. The predictive power of the method is based on the assumption that molecules with similar structures have similar properties. This approach has an established tradition [38], and is not in itself novel. POEM differs from these other approaches by intentionally restricting comparisons to relative similarity (through the embedding stage). Information is lost during the transformation of quantitative distances to relative similarity, such that similarity relationships evaluated are only meaningful in the context of the specific target molecule and predictive task. Specifically, the magnitude of any given distance has variable significance across fingerprints and predictive tasks. By ignoring the quantitative distribution of Tanimoto distances and operating only on less/greater comparisons between like fingerprints, POEM's treatment of reference molecule data is statistically *non-parametric*. The transformation to a 'distribution-free' representation of distances sidesteps the need for fingerprint-related transformation functions, weights or voting schemas. In turn, the entire landscape of labeled reference molecules can be used in making each prediction in a consistent and systematic manner, without the need to introduce fingerprint-related hyper-parameters to optimize from task-to-task. This approach leads to fast and objective model building, two major functional advantages of the algorithm.

While POEM's treatment of input data is statistically non-parametric, modification to the algorithm itself may impact performance. For example, adding new fingerprints may further improve performance. Future studies on larger datasets may identify new such modifications that produce significantly different optimal configurations from task-to-task, leading to optional hyper-parameters and a potential for optimization. Nonetheless, the results in this study demonstrate consistent performance using a static algorithm configuration and fingerprint selection. In this context, POEM is a powerful general purpose supervised machine learning approach that does not require hyperparameter optimization. POEM was designed to reduce the need for highly-tuned models, crafted by ML experts. Conceptually, the POEM algorithm is also applicable to problems outside of chemistry, as long as the object of the prediction has multiple representations which have metrics to define similarity.

We have shown that the increased performance and consistency of POEM is attributed to the use of multiple fingerprints simultaneously, as previously observed in other similarity approaches such as the *Similarity Ensemble Approach* [38]. POEM models created using 10 fingerprints outperform those created by individual fingerprints with few exceptions, which may be attributed to random variation. This is further supported by the observation that the optimal fingerprint differs on a task-by-task basis when evaluating the standard supervised learning methods (table 1) or single fingerprint POEM models (table S3). Even seemingly related tasks, such as Cytochrome P450 activity predictions, are best addressed with different fingerprints for different isoforms. If algorithmic noise is responsible for variation between models generated using different fingerprints or hyperparameters, then model performance may be compromised in subsequent real world applications. The use of multiple fingerprints simultaneously in POEM side-steps this issue, providing reliable performance across a range of predictive tasks. Efforts were initially made to also build predictive models for benchmarking using concatenated fingerprints, but computational runtime was deemed cost-prohibitive, demonstrating instead a distinct speed advantage to the POEM strategy for combining representations.

POEM is also seen to generalize well, in terms of the ability to make predictions for molecules unlike any known reference molecules (figure S3). More specifically, we observe that it performs as well or better than the GraphConv neural network (figure S4). This network in particular was chosen because it is known to be highly performative [31], and was also not too cost-prohibitive (though we were forced to limit the amount of nested cluster validation to two properties due to compute cost concerns).

While this study limits the scope of POEM to classification problems, the fundamental relationship between molecular similarity and activity established by POEM may provide a suitable framework for developing regression models. Preliminary findings applying POEM to four standard benchmark datasets presented in supplementary figure S5 demonstrate favorable performance over leading deep learning frameworks to develop regression models for chemical activity. Future studies using a broader range of datasets would provide a broader understanding on POEM's utility towards regression problems.

We have identified a number of functional advantages to POEM, including model building speed, reliability, predictive power, objectivity, ease of use, and model interpretability. POEM is particularly well suited to applications requiring automation due to its objective nature and reduced risk of overfitting. Highly automated applications may include: models built upon large-scale data mining expeditions, datasets with frequent updates, or model building by subject-area field experts without first-hand experience developing ML models. The similarity-based nature of this algorithm helps provide model interpretability, as each prediction is coupled with the list of reference molecules, their relative similarity, and their labels.

The above notwithstanding, there are important trade-offs associated with the POEM approach. Mainly, POEM has high algorithmic complexity associated with the generation of an $\mathbf{M} \times \mathbf{M}$ dominance matrix. Due to POEM's instance-based learning nature, predicting each unlabeled molecule scales proportionally with the square of the number of molecules in the reference molecule library. Algorithmically, this prediction stage is significantly slower when compared to other standard supervised learning methods. In practice however, POEM can handle datasets up to 100 000 training examples on modern personal computers (4 Core CPU, 16Gb RAM, Ubuntu 18.04). Applying POEM to ligand-based virtual screening using libraries with millions of molecules may also impose a technical challenge, requiring distributed computing solutions. Future heuristic approximations may improve POEM dataset scalability, including batching the dominance calculations (which would move POEM to O(nlogn) time rather than $O(n^2)$, at the cost of completely sorting all reference molecules). As it stands, we have not made these changes at present, as datasets of fewer than 10 000 reference molecules are still capable of performing predictions in the range of milliseconds to seconds. In our experience, these timescales are suitable for most practical applications in drug discovery. Nonetheless, even with regard to larger datasets, the lack of dedicated 'training' and 'optimization' stages make up for limitations in speed in applications where fewer predictions need to be made. For instance, POEM models can be easily improved over time, without added computational cost, simply by adding new data into the set of labeled reference molecules. Additionally, we have observed that highly unbalanced datasets can behave poorly when using POEM to make predictions, and additional dataset balancing might be desirable for producing highly performant models.

As is always the case with machine learning approaches, the main determinant for predictive performance is the nature and quality of the data used for training. This is observed here in the contrast between good models (BBB, ERa and HIA) and bad models (Carcin, CYP), especially when looking at generalizability (figure S2). Unfortunately, in the world of drug-property prediction, many of the best datasets are privately held, despite exciting but limited recent efforts to make some data available to researchers and the public [39].

In spite of the above limitations, we consider POEM to be a valuable addition to the roster of methods for supervised learning available today, especially given its lack of hyperparameters, high generalizability, and low cost.

Acknowledgments

We thank Lenny Morayniss and Joseph C Somody for challenging discussions and feedback. We thank the RDKit open source project and SciKit-learn for significant contributions to the scientific community and much useful code. We thank James Crompton and Marc Laforet for assisting with the benchmark standard classifiers and hyperparameter grid searches. We thank Naheed Kurji, CEO of Cyclica, for his unwavering support of this work without which it would not have been possible.

Author contributions

The method proposed in this paper was conceived by A E B, and developed by A E B and A W; writing was done by A E B; data collection and experiments were carried out by A E B, Z S, S A, S R., and S M; and all authors were involved in planning, discussion, and preparation of the manuscript.

Conflict of interest

The algorithms publicly disclosed in this investigation were developed and commercialized by Cyclica Inc.

Data and materials availability

All data are publicly available, with information about data location reported in the supplementary information. POEM is available commercially through Cyclica's Ligand Express platform, and free of charge under the Cyclica Academic Partnership Program (CAPP), upon request.

Funding

This work is partially supported through the Ontario Centre of Excellence TalentEdge Data Analytics Internship.

ORCID iDs

Andrew E Brereton https://orcid.org/0000-0002-6716-7177
Stephen MacKinnon https://orcid.org/0000-0003-4117-0340
Andreas Windemuth https://orcid.org/0000-0002-6985-9919

References

[1] Craig P N 1984 QSAR—origins and present status: a historical perspective *Drug Inf. J.* **18** 123–30
[2] Sanders J M *et al* 2017 Informing the selection of screening hit series with *in silico* absorption, distribution, metabolism, excretion, and toxicity profiles *J. Med. Chem.* **60** 6771–80
[3] Clark P 2000 Computational methods for the prediction of drug-likeness *Drug Discovery Today* **5** 49–58
[4] Waring M J *et al* 2015 An analysis of the attrition of drug candidates from four major pharmaceutical companies *Nat. Rev. Drug Discovery* **14** 475–86
[5] Wu Z, Ramsundar B, Feinberg E N, Gomes J, Geniesse C, Pappu A S, Leswing K and Pande V 2018 MoleculeNet: a benchmark for molecular machine learning *Chem. Sci.* **9** 513–30
[6] OpenSMILES specification (available at: http://opensmiles.org/opensmiles.html)
[7] O'Boyle N M 2012 Towards a Universal SMILES representation—a standard method to generate canonical SMILES based on the InChI *J. Cheminform.* **4** 22
[8] Bajusz D, Rácz A and Héberger K 2015 Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminform.* **7** 20
[9] Lipinski C A, Lombardo F, Dominy B W and Feeney P J 2001 Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings *Adv. Drug Deliv. Rev.* **46** 3–26
[10] Gómez-Bombarelli R, Wei J N, Duvenaud D, Hernández-Lobato J M, Sánchez-Lengeling B, Sheberla D, Aguilera-Iparraguirre J, Hirzel T D, Adams R P and Aspuru-Guzik A 2018 Automatic chemical design using a data-driven continuous representation of molecules *ACS Cent. Sci.* **4** 268–76
[11] Daylight Chemical Information Systems 2011 Daylight Theory Manual (available at: http://www.daylight.com/dayhtml/doc/theory/)
[12] Winter R, Montanari F, Noé F and Clevert D-A 2019 Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations *Chem. Sci.* **10** 1692–701
[13] O'Boyle N M and Sayle R A 2016 Comparing structural fingerprints using a literature-based similarity benchmark *J. Cheminform.* **8** 36
[14] Riniker S and Landrum G A 2013 Open-source platform to benchmark fingerprints for ligand-based virtual screening *J. Cheminform.* **5** 26
[15] Bergstra J S, Bardenet R, Bengio Y and Kégl B 2011 *Advances in Neural Information Processing Systems 24*, ed J Shawe-Taylor, R S Zemel, P L Bartlett, F Pereira and K Q Weinberger (Cambridge, MA: MIT Press) pp 2546–54
[16] Pan S J and Yang Q 2010 A survey on transfer learning *IEEE Trans. Knowl. Data Eng.* **22** 1345–59
[17] Raina R, Battle A, Lee H, Packer B and Ng A Y 2007 Self-taught learning: transfer learning from unlabeled data *Proc. 24th Int. Conf. on Machine Learning* pp 759–66
[18] Mordvintsev A, Olah C and Tyka M 2015 Deepdream-a code example for visualizing neural networks *Google Res.* **2** 5
[19] Aggarwal C C 2005 On k-anonymity and the curse of dimensionality *Proc. 31st Int. Conf. on Very Large Data Bases* vol 5 pp 901–9
[20] Friedman J H and Bias O 1997 Variance, 0/1—loss, and the curse-of-dimensionality *Data Min. Knowl. Discovery* **1** 55–77
[21] Indyk P and Motwani R 1998 Approximate nearest neighbors: towards removing the curse of dimensionality *Proc. 30th Annual ACM Symp. on Theory of Computing* pp 604–13
[22] Wang Z, Liang L, Yin Z and Lin J 2016 Improving chemical similarity ensemble approach in target prediction *J. Cheminform.* **8** 20
[23] Dixon S L, Duan J, Smith E, Von Bargen C D, Sherman W and Repasky M P 2016 AutoQSAR: an automated machine learning tool for best-practice quantitative structure-activity relationship modeling *Future Med. Chem.* **8** 1825–39
[24] Magill M and Quinzii M 2002 *Theory of Incomplete Markets* (Cambridge, MA: MIT Press)
[25] Ngatchou P, Zarei A and El-Sharkawi A 2005 Pareto multi objective optimization *Proc. 13th Int. Conf. on Intelligent Systems Application to Power Systems* pp 84–91
[26] Li Y and Yaseen A 2013 Pareto-based optimal sampling method and its applications in protein structural conformation sampling *Workshops at the 27th AAAI Conf. on Artificial Intelligence* pp 32–37
[27] Li Y, Rata I and Jakobsson E 2011 Sampling multiple scoring functions can improve protein loop structure prediction accuracy *J. Chem. Inf. Model.* **51** 1656–66
[28] Besnard J *et al* 2012 Automated design of ligands to polypharmacological profiles *Nature* **492** 215–20
[29] Zhang W, Pei J and Lai L 2017 Computational multitarget drug design *J. Chem. Inf. Model.* **57** 403–12
[30] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R and Dubourg V 2011 SciKit-learn: machine learning in Python *J. Mach. Learn. Res.* **12** 2825–30
[31] Duvenaud D, Maclaurin D, Aguilera-Iparraguirre J, Gómez-Bombarelli R L, Hirzel T, Aspuru-Guzik A and Adams R P 2015 Convolutional networks on graphs for learning molecular fingerprints *Advances in Neural Information Processing Systems 28 (NIPS 2015)* pp 2224–32
[32] Kearnes S, McCloskey K, Berndl M, Pande V and Riley P 2016 Molecular graph convolutions: moving beyond fingerprints *J. Comput. Aided Mol. Des.* **30** 595–608
[33] Ramsundar B, Eastman P, Walters P, Pande V, Leswing K and Wu Z 2019 *Deep Learning for the Life Sciences* (Sebastopol, CA: O'Reilly Media)

[34] Mayr A, Klambauer G, Unterthiner T, Steijaert M, Wegner J K, Ceulemans H, Clevert D-A and Hochreiter S 2018 Large-scale comparison of machine learning methods for drug target prediction on ChEMBL *Chem. Sci.* **9** 5441–51
[35] Cheng F, Li W, Zhou Y, Shen J, Wu Z, Liu G and Lee P W 2012 admetSAR: a comprehensive source and free tool for assessment of chemical ADMET properties *J. Chem. Inf. Model.* **52** 3099–105
[36] Hansen K, Mika S, Schroeter T, Sutter A, Ter Laak A, Steger-Hartmann T, Heinrich N and Müller K-R 2009 Benchmark data set for in silico prediction of Ames mutagenicity *J. Chem. Inf. Model.* **49** 2077–81
[37] Huang R, Xia M, Nguyen D-T, Zhao T, Sakamuru S, Zhao J, Shahane S A, Rossoshek A and Simeonov A 2016 Tox21Challenge to build predictive models of nuclear receptor and stress response pathways as mediated by exposure to environmental chemicals and drugs *Front. Environ. Sci.* **3** 85
[38] Keiser M J, Roth B L, Armbruster B N, Ernsberger P, Irwin J J and Shoichet B K 2007 Relating protein pharmacology by ligand chemistry *Nat. Biotechnol.* **25** 197–206
[39] Ekins S and Williams A J 2010 When pharmaceutical companies publish large datasets: an abundance of riches or fool's gold? *Drug Discovery Today* **15** 812–15