Taking a Respite from Representation Learning for Molecular Property Prediction

Jianyuan Deng[1], **Zhibo Yang**[2], **Hehe Wang**[3], **Iwao Ojima**[3], **Dimitris Samaras**[2], and **Fusheng Wang**[1,2,*]

[1]Stony Brook University, Department of Biomedical Informatics, Stony Brook, 11790, United States
[2]Stony Brook University, Department of Computer Science, Stony Brook, 11790, United States
[3]Stony Brook University, Department of Chemistry, Stony Brook, 11790, United States

ABSTRACT

Artificial intelligence (AI) has been widely applied in drug discovery with a major task as molecular property prediction. Despite the boom of AI techniques in molecular representation learning, some key aspects underlying molecular property prediction haven't been carefully examined yet. In this study, we conducted a systematic evaluation on three representative models, random forest, MolBERT and GROVER, which utilize three major molecular representations, extended-connectivity fingerprints, SMILES strings and molecular graphs, respectively. Notably, MolBERT and GROVER, are pretrained on large-scale unlabelled molecule corpuses in a self-supervised manner. In addition to the commonly used MoleculeNet benchmark datasets, we also assembled a suite of opioids-related datasets for downstream prediction evaluation. We first conducted dataset profiling on label distribution and structural analyses; we also examined the activity-cliffs issue in the opioids-related datasets. Then, we trained 4,320 predictive models and evaluated the usefulness of the learned representations. Furthermore, we explored into the model evaluation by studying the effect of statistical tests, evaluation metrics and task settings. Finally, we dissected the chemical space generalization into inter-scaffold and intra-scaffold generalization and measured prediction performance to evaluate model generalizbility under both settings. By taking this respite, we reflected on the key aspects underlying molecular property prediction, the awareness of which can, hopefully, bring better AI techniques in this field.

1 Introduction

Drug discovery is an expensive process in both time and cost with a daunting attrition rate. As revealed by a recent study[1], the average cost of developing a new drug was around 1 billion dollars and has been ever increasing[2]. In the past decade, the practice of drug discovery has been undergoing radical transformations in light of the advancements in artificial intelligence (AI)[3–5], which, at its core, is molecular representation learning. Molecules are usually represented by: 1) chemical descriptors, which reflect the existence of certain fragments, such as hashed fingerprints, Extended-Connectivity FingerPrints (ECFP), 2) linear notations, such as Simplified Molecular Input Line Entry System (SMILES) strings and 3) molecular graphs[6]. With the advent of the deep-learning era, various neural networks have been proposed for molecular representation learning, such as convolutional neural networks (CNNs), recurrent neural networks (RNNs) and graph neural networks (GNNs), among others[5]. One major task for AI in drug discovery is molecular property prediction, which seeks to learn a function that maps a structure to a property value. In the literature, deep representation learning has been widely reported to show great potential in molecular property prediction over fixed molecular representations[7,8]. More recently, to address the lack of labeled data in drug discovery, self-supervised learning has been proposed to leverage large-scale unlabeled corpus for both SMILES strings[9–11] and molecular graphs[12–15], which has enabled state-of-the-art (SOTA) performance based on the MoleculeNet benchmark datasets[16].

Nevertheless, AI-driven drug discovery is not without its critiques despite its current prosperity. Usually, when a new technique is developed for molecular property prediction, improved metrics by experimenting on the MoleculeNet benchmark datasets[16] are used to substantiate the argument that the model outperforms previous ones. Although novel techniques are indeed proposed along with impressive metrics, most often they do not suffice to solve the molecular property prediction need in real-world drug discovery settings. In fact, the existing practice of representation learning for molecular property prediction can be dangerous yet quite rampant[17]. The detailed issues are elaborated as follows.

First, there is a heavy reliance on the benchmark datasets, which may be of little relevance to real-world drug discovery[18]. Moreover, despite the wide adoption of the benchmark datasets, the split folds can vary in the literature, which entail unfair performance comparison[19]. Very often, achieving the SOTA performance becomes the major pursuit of new model development, which also leads to little focus on the statistical rigor and the model applicability[17]. For instance, when reporting prediction performance for a newly developed model, most papers just used the mean value averaged over 3-fold[7,13,20] or 10-fold[11,12,19,21] splits. The seeds for dataset splitting may or may not be explicitly provided. Sometimes, it may just come from some arbitrary split with a few individual runs. The inherent variability of dataset splitting is simply ignored. The caveat is that, without rigorous analysis, the seemingly improved metric values may just be some statistical noise[17]. For the model applicability, besides the limited relevance of the heavily used benchmark datasets, there is

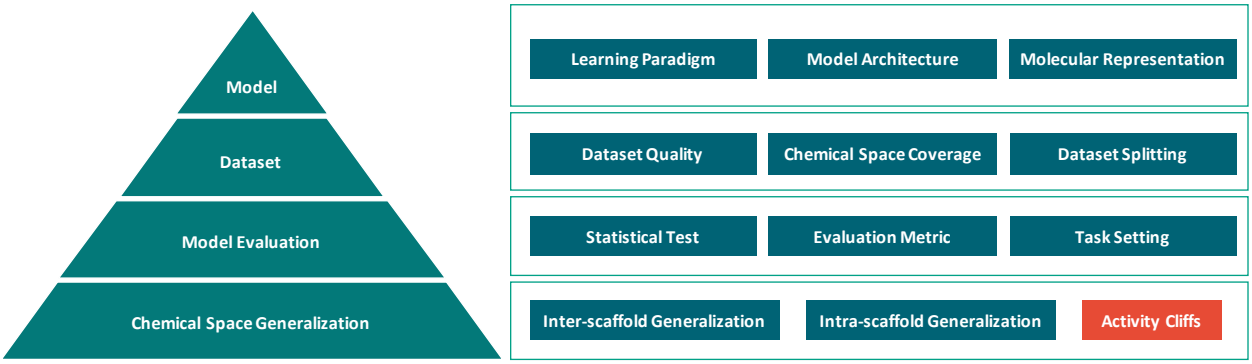

Figure 1. Key Aspects underlying Molecular Property Prediction. There are four aspects underlying molecular property prediction: model, dataset, model evaluation and chemical space generalization. Usually in the literature, the focus is more on the model, which aims at developing novel learning paradigms or model architectures on specific molecular representation formats. However, other aspects, which are pertaining to 1) what the model is built upon, 2) how the model is evaluated and 3) eventually what the model is capable of, should also be considered. For the dataset, its quality (such as duplicates, contradictories and uncertainty in labels), the chemical space coverage (w.r.t. both structures and labels) and the splitting method should be carefully checked before building a model applicable to a certain property prediction task. For the model evaluation, the use of statistical test, evaluation metric and the task setting also affect the observed prediction performance and may even alter the conclusion. For the chemical space generalization, there is a necessity to clarify what type of generalization the model is capable of and whether the activity-cliffs issue is also addressed.

also a lack of practical relevance for the recommended evaluation metrics. One example is AUROC, which, as opined by Robinson *et al.*[17], cannot well capture the true positive rate, a more relevant metric in virtual screening. To address these existing issues, we took a respite to rethink the key aspects underlying molecular property prediction. We conducted a systematic evaluation on representative models in molecular property prediction by training 4,320 models, with a focus on: 1) dataset profiling, including label distribution and structural analysis; 2) model evaluation, on the effect of statistical test, metrics and task setting; 3) chemical space generalization, w.r.t. inter-scaffold and intra-scaffold generalization.

The outline of the paper is as follows. We first discussed the preliminaries for molecular property prediction, including molecular representation formats, model architectures and learning paradigms[5]. We then discussed in detail the rationale of this study and proposed our experiment schemes. Secondly, we elaborated on the methods, including datasets collection, evaluation metrics, model training and statistical analyses. To further evaluate the usefulness of the molecular representation learning models, we also assembled a suite of opioids-related datasets from ChEMBL[22]. Thirdly, we presented the results, where we checked on three aspects: 1) dataset profiling, 2) model evaluation and 3) chemical space generalization. Finally, we discussed our thoughts on how to advance the representation learning for molecular property prediction. During this respite, we summarized the key aspects underlying molecular property prediction, which should be considered in future model development (Figure 1). Similar to the quote from Bender et al[23,24] "a method cannot save an unsuitable representation which cannot remedy irrelevant data for an ill-thought-through question.", our central thesis is that *"a model cannot save an unqualified dataset which cannot remedy an improper evaluation for an ambiguous chemical space generalization claim"*.

2 Preliminaries

2.1 Molecular representation formats

Over the years, various formats have been used to represent small molecules[5,6]. Arguably, the simplest format are the 1D descriptors which represent a molecule based on its formula, such as atom counts, types and molecular weight. Moreover, molecules can also be represented using 2D fingerprints, such as structural keys and path-based or circular fingerprints[25]. These fingerprints can be either 1) bit vectors, which are binary vectors with each dimension tracking the presence or absence of a particular substructure or 2) count vectors, which track the frequency of each substructure appearance. One of the most widely used molecular fingerprints is the ECFP bit-vector based on the Morgan algorithm, firstly proposed to solve the molecular isomorphism issue, which is to identify if two molecules, with different atom numberings, are the same[26]. There are three sequential stages during ECFP generation: 1) an initial assignment stage when each atom has an integer identifier assigned to it, 2) an iterative updating stage when each atom identifier is updated to reflect the identifiers of each atom's neighbors, including identification of whether it is a structural duplicate of other features, 3) a duplicate identifier removal stage when multiple occurrences of the same feature are reduced to single representative in the final feature list (the occurrence count may be retained if a set of counts are required instead of a standard binary fingerprint). ECFP has been the de facto standard circular fingerprint and

Table 1. Commonly Used Node and Edge Features

Type	Feature	Notes
Node	Atom type	Element type
Node	Formal charge	Assigned charges
Node	Implicit Hs	Number of bonded hydrogens
Node	Chirality	*R* or *S* configuration
Node	Hybridization	Orbital hybridization
Node	Aromaticity	Aromatic atom or not
Edge	Bond type	Single, double, triple, aromatic
Edge	Conjugated	Conjugated or not
Edge	Stereoisomers	cis or trans, (*E* or *Z*), none, any

is still valuable in drug discovery[25]. The vector size of ECFP is usually 1024 or 2048. The radius size of ECFP can either be 2 or 3, corresponding to ECFP4 and ECFP6, respectively, which are the common variants of ECFP in the literature. For instance, Yang *et al.*[8] used ECFP4 while Mayr *et al.*[7], Robinson *et al.*[17] and Skinnider *et al.*[27] used ECFP6.

Intuitively, small molecules can also be represented as graphs, with atoms as nodes and bonds as edges. Formally, a graph is defined as $G = (V, E)$, where V and E represent nodes (atoms) and edges (bonds), respectively. The attributes of atoms can be represented by a node feature matrix X and each node v can be represented by an initial vector $x_v \in R^D$ and a hidden vector $h_v \in R^D$. Similarly, the attributes of bonds can also be represented by a feature matrix. In addition, an adjacency matrix A is used to represent the pairwise connections between nodes. For every two nodes v_i and v_j, $A_{ij} = 1$ if there is a bond connecting nodes v_i and v_j; $A_{ij} = 0$ otherwise. Usually, the edge feature matrix and the adjacency matrix can be combined to form an adjacency tensor. Table 1 summarizes the commonly used node and edge features.

Although the graph representation carries more structural information, one drawback is that graphs are often storage- and memory-demanding[6]. Alternatively, a more computationally efficient representation of molecules is the SMILES string[28], where atoms are represented by the atomic symbols and bonds are represented by symbols like "-", "=", "#" and ":", corresponding to single, double, triple and aromatic bonds, respectively. Notably, single bonds and aromatic bonds are usually omitted. Moreover, parentheses are used to denote the branches in a molecule. For the cyclic structure, a single or aromatic bond is first broken down in the ring and the bonds are then numbered in any order with the ring-opening bonds by a digit following the atomic symbol at each ring. Notably, one molecule can have multiple SMILES-string representations[6]. Thus, the canonicalized SMILES strings[29] are more often used in the literature. To be machine-readable, SMILES strings are usually converted into one-hot vectors.

2.2 Model architectures

So far, various model architectures have been proposed for molecular property prediction, such as RNNs, GNNs and transformers[5]. Originally designed for handling sequential data (e.g., text and audio), RNNs can be naturally used to model molecules formatted as the SMILES strings, such as SMILES2Vec[30] and SmilesLSTM[7]. GNNs, on the other hand, can be applied on the molecular graphs, such as graph convolutional networks (GCN)[31], graph attention network (GAT)[32], message passing neural networks (MPNN)[33], directed MPNN (D-MPNN)[8], graph isomorphism networks (GIN)[12,34]. Recently, to address the lack of annotated data in drug discovery, self-supervised learning has been proposed for pretraining on large-scale unlabeled molecule corpuses before downstream finetuning[5]. In this study, we mainly utilized two self-supervised models: MolBERT[11] and GROVER[13] which take SMILES strings and molecular graphs as input, respectively. Random forest (RF) on ECFP4 is used as the baseline.

2.2.1 Random forest

Random Forest (RF) is an ensemble of decision tree predictors, which can be applied for classification and regression tasks[35]. RF has been widely adopted in drug discovery prior to the "deep-learning" era[4]. By assembling a certain number of de-correlated decision trees, RF establishes a strong baseline for deep learning models in many tasks[36]. Therefore, we chose to use RF as the baseline for evaluating advanced molecular representation learning models (i.e., MolBERT and GROVER). Note that RF can be applied to both descriptor-based representation[37] and fingerprint-based representation[8]. Here, we focused on the circular fingerprint, ECFP4.

2.2.2 MolBERT

The SMILES strings can be viewed as a "chemical" language. Language models, therefore, have been widely applied in molecular representation learning for molecular property prediction, molecule generation and retro-synthesis prediction[5]. Related model architectures include RNNs and transformers[5]. Recently, inspired by Bidirectional Encoder Representation from Transformers (BERT) in natural language processing, Fabian *et al.*[11] exploited the architecture of BERT for molecular property prediction. Using transformers as the building block, MolBERT is pretrained on a corpus of c.a. 1.6M SMILES strings[38], which improves prediction performance on 6 benchmark datasets in both classification (BACE, BBBP, HIV) and regression (ESOL, FreeSolv, Lipop) settings[16].

The abstracted architecture of MolBERT is depicted in Figure 2 (a). An input SMILES string is firstly tokenized and embedded into a sequence of d-dimensional vectors. Unlike RNNs, which handle the input sequentially, a positional embedding layer is added to the input tokens to capture the sequential information. A stack of n BERT encoder layers is then added on top of the embedding layers to learn the latent representations of the input sequence. During pretraining, different pretext self-supervised tasks, such as masked language modeling, are designed to utilize the output embeddings after the pooler layer. During finetuning, new task heads can be appended by attaching a single linear layer to the pooled output for downstream property prediction. The learned weights of the backbone model during pretraining are fixed, which pro-

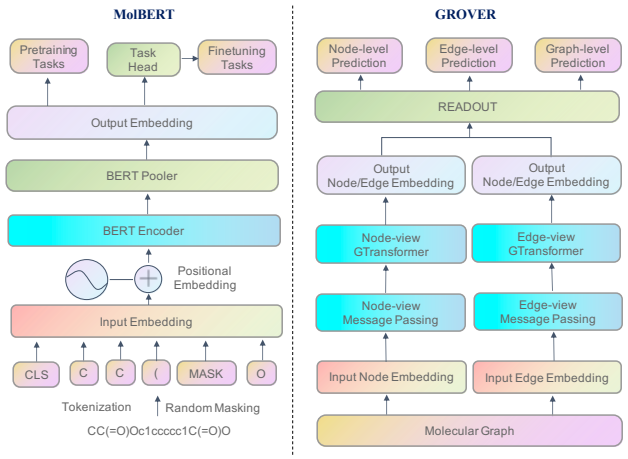

Figure 2. Abstracted Model Architectures Modified from MolBERT[11] and GROVER[13].

vide a better model initialization and can also reduce training burden during finetuning, especially when the model is large.

MolBERT is pretrained on a fixed vocabulary of 42 tokens and a maximum sequence length of 128 characters. To support the arbitrary length of input SMILES strings at inference time (for instance, we adopted the maximum sequence length of 400 during finetuning), relative positional encoding[39] is used. Following the original BERT model, MolBERT uses the BERT-Base architecture with an output embedding size of 768, 12 BERT encoder layers, 12 attention heads and a hidden size of 3,072, resulting in c.a. 85M parameters. In this study, we used the pretrained MolBERT provided by Fabian et al[11].

2.2.3 GROVER

In addition to the SMILES strings, molecules can also be intuitively abstracted as graphs. GNNs, therefore, have been widely applied in molecular representations learning[5]. The core operation in GNNs is message passing, a.k.a. neighborhood aggregation[33]. During message passing, a node's hidden state is iteratively updated by aggregating the hidden states of its neighboring nodes and edges, which involves multiple hops. After each iteration, the message vectors can be integrated using certain AGGREGATE functions, such as sum, mean, max pooling or graph attention[40]. The AGGREGATE function is essentially a trainable layer together with some activation functions, which is shared by different hops within an iteration. When the message passing is completed, the hidden states of the last hop from the last iteration are used as the nodes' embeddings, followed by a READOUT function to obtain the graph-level embedding. Two tasks are commonly used to pre-train the GNNs[12]: 1) node-level self-supervised training, such as atom type prediction and 2) graph-level supervised training, such as molecular label prediction. However, supervised pretraining may cause "negative transfer", where downstream performance can be deteriorated therein.

Recently, Rong *et al.*[13] proposed GROVER with delicately-designed, self-supervised pretraining tasks at the node-, edge- and graph-level, respectively. GROVER is pretrained on c.a. 10M unlabeled molecules, which is then evaluated on 11 benchmark datasets (classification setting: BACE, BBBP, ClinTox, SIDER, Tox21, ToxCast; regression setting: ESOL, FreeSolv, Lipop, QM7, QM8) and achieves SOTA performance in all datasets. The abstracted model architecture of GROVER is depicted in Figure 2 (b). The input node and edge embeddings are first learned via message passing[33], which are then passed to the node-view GTransformer and edge-view GTransformer to output the node and edge embeddings from the two views, respectively. With some READOUT function, the final embeddings can be used for node-level, edge-level or graph-level prediction tasks. For downstream tasks, following the practice in Chemprop[8], GROVER extracts 200 global molecular features using RDKit[41], which are concatenated with the learned embedding vector, i.e., output of the READOUT function, and then passes through a linear layer, i.e., a task head, for molecular property prediction.

Notably, GROVER has two configurations: GROVER_{base} and GROVER_{large}, corresponding to c.a. 48M and c.a. 100M model parameters, respectively. With a corpus of c.a. 10M unlabeled molecules for pretraining, GROVER demands very intensive computational resources. As stated in the paper[13], GROVER_{base} costs 2.5 days, and GROVER_{large} costs 4 days on 250 NVIDIA V100 GPUs for pretraining. Due to the large number of experiments to be conducted in this study, we only used the pretrained GROVER_{base}. To examine the actual power of GROVER and the effect of the additional RDKit features, we distinguished GROVER (without RDKit features) and GROVER_RDKit, which are also denoted as "GROVER 1" and "GROVER 2" later in the main text.

2.3 Opioids with reduced overdose effects

To further evaluate the molecular representation learning models, we collected a suite of opioids-related datasets besides the MoleculeNet benchmark datasets[16]. Opioid overdose is a leading cause of injury-related death in the United States[42] and there is an increasing interest in developing opioid analgesics with reduced overdose effects. As pointed out by a large-scale observational study[43], reduced overdose effects can potentially be addressed from the pharmacokinetic (PK) perspective and pharmacodynamic (PD) perspective, which correspond to 1) decreasing the overdose events by avoiding excessive amount of opioids in the action site and 2) alleviating the overdose outcome by avoiding off-target effects, respectively. The PK targets include multi-drug resistance protein 1 (MDR1), cytochrome P450 (CYP) 2D6, and CYP3A4 whereas the PD targets include μ opioid receptor (MOR), δ opioid receptor (DOR) and κ opioid receptor (KOR). Details on the opioids-related datasets collection are in Section 3.1.

2.4 Study rationale and experiment design

2.4.1 How useful are the learned representations?

The first question that our study aims to answer is: how useful are the learned representations for molecular property prediction? Previously, although deep neural networks have been

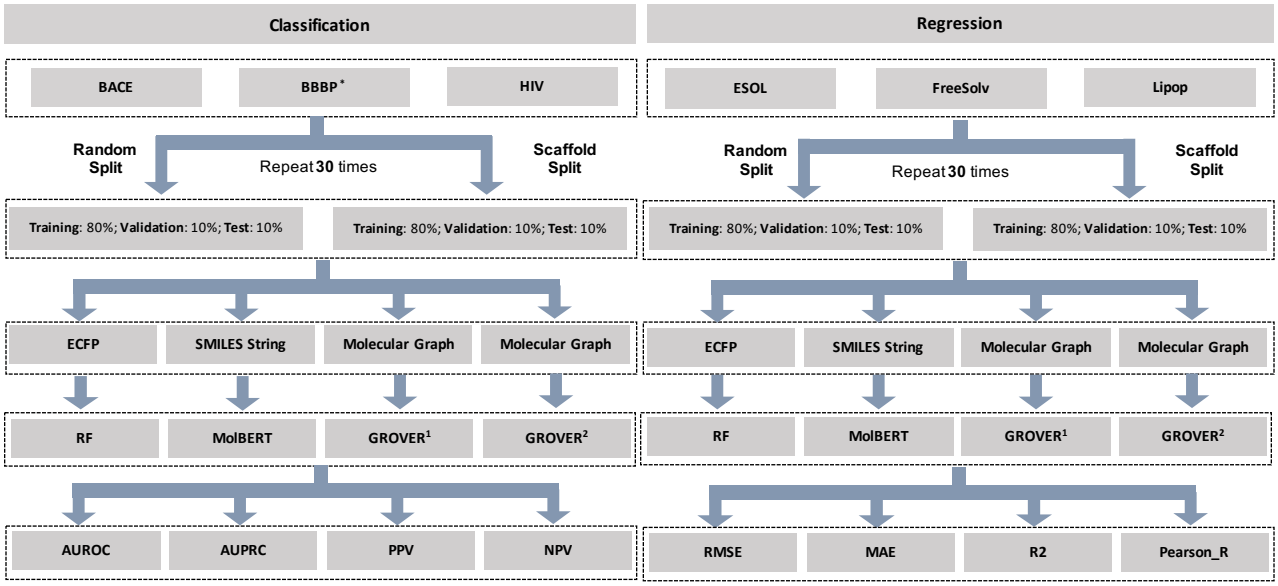

Figure 3. Experiment Scheme on the Benchmark Datasets.

reported to outperform traditional machine learning models, such as RF and support vector machine (SVM) on ECFP in a large-scale bioactivity prediction study[7], Robinson *et al.*[17] have re-analyzed the results and found that SVM is still competitive with the deep learning models. Indeed, whether the learned representations can supersede fixed representations, such as ECFP, still needs further interrogation. To this end, we selected two representative models for molecular property prediction after an extensive literature review[5]. In total, we investigated 4 models, namely, RF, MolBERT, GROVER and GROVER_RDKit (see Section 2.2). The experiment scheme on the benchmark datasets is in Figure 3. We also evaluated the models using the opioids-related datasets (Figure 4).

2.4.2 Are the models properly evaluated?

As suggested by MoleculeNet[16], each benchmark dataset comes with a recommended evaluation metric, a common practice widely adopted by subsequent studies. For example, for the classification datasets, the area under the receiver operating characteristic curve (AUROC) is always used; whereas for the regression datasets, the most common metric is the root mean square error (RMSE). However, these recommended metrics can be problematic. As opined by Robinson *et al.*[17], AUROC for the classification task may be of little relevance in real-world drug discovery applications, such as virtual screening. In the case of imbalanced datasets, which is often the case in reality where only a small portion of test molecules are actives, AUROC can be biased[44]. In fact, AUROC can be seen as the expected true positive rate averaged over all classification thresholds (false positive rates). Thus, if two ROC curves cross, even if one curve has higher AUROC, it may perform much worse (lower true positive rate) under certain thresholds of interest. One alternative is the area under the precision-recall curve (AUPRC)[17,44], which only focuses on the minor class, often the actives.

In this study, we further argue that the evaluation metric should be contingent on the question of interest during drug discovery. For instance, one popular sub-task involved in virtual screening is target fishing, which is to identify all possible targets that a molecule can bind to. According to Hu *et al.*[45], an active PubChem compound can interact with c.a. 2.5 targets and consequently, off-target effects can be pervasive and may lead to undesired adverse drug reactions. Thus, one research area is to identify potential targets for a molecule during early stage[46]. In this scenario, the question is not just about predicting whether a molecule can bind to a specific target. Instead, we would care more about questions like: 1) given a set of predicted drug targets k, what is the fraction of correct predictions among the predicted positives, i.e., $recall@k$; 2) given a set of predicted drug targets k, what is the fraction of correct predictions among the annotated positives, i.e., $precision@k$. Even in the single-target virtual screening scenario, we may care more about the *precision*, which is the positive predictive value (PPV), inasmuch it is imperative to ensure that there is a sufficient amount of true positives out of the predicted positives. On the contrary, if the goal is to exclude molecules inactive against certain targets which are related to adverse reactions, the negative predictive value (NPV) is of more interest. Thus, for each experiment, we evaluated a set of metrics (Figure 3 & Figure 4). More details on the evaluation metrics are in Section 3.2.

In addition to the evaluation metrics, another crucial but often missing part in previous studies is the statistical test, despite that the benchmark datasets are small-sized[17,18,47]. Most often, when a new model is developed, some arbitrary split or 3/10-fold splits are applied to calculate the mean of some metric for rudimentary comparison. The reality is, however, due to lack of rigorous statistical tests, such result is

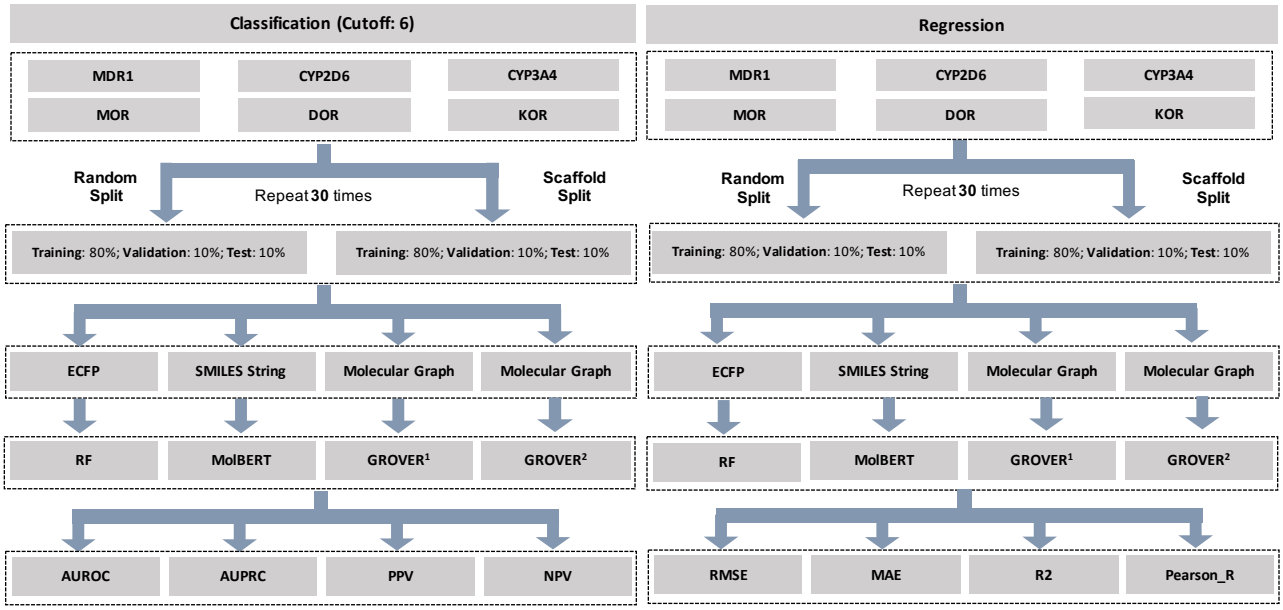

Figure 4. Experiment Scheme on the Opioids-related Datasets.

insufficient to justify whether there is a true improvement. Besides, the task setting may also affect the evaluation. Usually, after collecting the pIC50 values, a cutoff value such as 5 or 6 is used to dissect the collected molecules into actives and inactives. Nevertheless, how the classification with an arbitrary cutoff value affects the final prediction compared to directly regressing the pIC50 values hasn't been unexplored yet. To study the influence of task settings, we conducted experiments under both classification and regression settings for the opioids-related datasets (Figure 4).

2.4.3 What does chemical space generalization mean?

In the representation learning for molecular property prediction, the ultimate goal is to build models that can generalize from seen molecules to unseen ones. To mimic chronological split in the real-world setting, MoleculeNet[16] proposed scaffold split as a proxy which ensures that molecules in the test set are equipped with unseen scaffolds during training, posing a more challenging prediction task. In the literature, most papers have adopted the scaffold-split practice and claimed chemical space generalization capability when evaluation metrics are seen improved. The assumption, here, is that chemical space generalization is approximated as generalization between different scaffolds, which further assumes each scaffold is associated with certain property values, for instance, similar activity. Nevertheless, one scaffold may not necessarily map to a narrow range of property values. In this case, the use of scaffold split does not suffice to guarantee chemical space generalization. Moreover, it may lead to ambiguity.

Formally, the chemical space is defined as the set of all possible organic molecules, more specifically, the biologically relevant molecules[48]. In the chemical space, usually there are some structural constellations, which is populated with certain property values and can be identified using scaffold-based analysis[49]. Since the constellations have diverse scaffolds, two molecules with different scaffolds can have disparate properties, which is termed as the "scaffold cliff"[49]. Using the widely-adopted scaffold split, we argue that it actually addresses the "scaffold cliff", which is essentially doing the inter-scaffold generalization. Meanwhile, another challenge in drug discovery is the "activity cliffs" (see Section 4.1.2), where a tiny structural change causes a drastic activity change between a pair of molecules with similar structures, usually with the same scaffold[50]. On the contrary to inter-scaffold generalization, it is the intra-scaffold generalization needed in the case of the activity cliffs. Unfortunately, while activity cliff is prevalent and has been discussed in computational and medicinal chemistry for nearly three decades[50], it has not been emphasized in most molecular property prediction papers. In this study, we adopted both scaffold split and random split to check inter-scaffold generalization capability (Figures 3 & 4). Furthermore, to check intra-scaffold generalization, we filtered out molecules with scaffolds observed with the activity-cliffs issue, denoted as the AC molecules (see Section 4.1.2), and then calculated prediction performance on the AC and non-AC molecules (see Section 4.3.2), respectively.

3 Methods

3.1 Datasets assembly

3.1.1 Benchmark datasets

In 2018, a suite of benchmark datasets from MoleculeNet for molecular property prediction was proposed[16], which have been widely used to develop new molecular representation learning models. Among them, we selected three classification datasets (BACE, BBBP, HIV) and three regression

datasets (ESOL, FreeSolv, Lipop), which are used in MolBERT[11] and GROVER[13] (except for HIV) as well as a recent comparison study Jiang et al[36]. Note that these datasets are for single-task purpose and were downloaded from MolMapNet[19], which has a more updated version. Each dataset and its task type, number of molecules, maximum string length and number of scaffolds are summarized in Table 2. Since MolBERT needs to pad the input SMILES strings to the maximum length, we only kept the molecules with length no longer than 400. As shown in Table 2, all selected benchmark datasets have maximum length less than 400 except for HIV, where c.a. 0.01% molecules were removed.

Dataset	Task	#Mol.	Max. Len.	#Scaff.
BACE	CLS	1,513	198	737
BBBP	CLS	2,039	400	1,101
HIV	CLS	41,127	580	19,085
ESOL	REG	1,128	98	268
FreeSolv	REG	642	82	62
Lipop	REG	4,200	267	2,443

Table 2. Summary of the Benchmark Datasets.

As for dataset splitting, there are several options, such as random split, scaffold split, stratified split and time split[16]. Each method has its own purpose. For example, time split makes the model trained on older data points and tested on newer molecules, which mimics the real-world scenario where models are built on existing data points and are used to predict properties of newly synthesized molecules[16]. The most widely adopted method in the literature is scaffold split, which addresses the inter-scaffold generalization (see Section 2.4.3). However, the actual split folds can still vary across studies. For instance, for the regression datasets, MolBERT[11] used the random split folds provided in MolMapNet[19] while GROVER adopted scaffold split. For the classification datasets, both MolBERT and GROVER adopted scaffold split, although the seeds are not provided and the split folds may not be identical.

In this study, we adopted both scaffold and random split, following a 80:10:10 ratio for training/validation/test sets (Figures 3 & 4). Additionally, to ensure sufficient statistical rigor, we repeated the dataset split procedure 30 times with 30 different seeds $(0, 1, 2, \cdots, 29)$ using GROVER's implementation for dataset split, which were saved and kept the same for all experiments to ensure fair comparison. All the used benchmark datasets and split folds are provided in our Github repository.

3.1.2 Opioids-related datasets

To check the usefulness of the representation learning models for molecular property prediction, we also assembled a suite of opioids-related datasets (see Section 2.3). Specifically, we collected the binding affinity for the following pharmacological components[43,51]: MDR1 (ChEMBL ID: 4302), CYP2D6 (ChEMBL ID: 289), CYP3A4 (ChEMBL ID: 340), MOR (ChEMBL ID: 233), DOR (ChEMBL ID: 236) and KOR (ChEMBL ID: 237). The binding affinity data is retrieved from ChEMBL27[22] using *in vitro* potency measures, namely, IC50, EC50, Ki and Kd. Assay type is set as "Binding", standard relationship is set as "=" with the standard unit as "nM" and the organism as "Homo Sapiens". The raw binding affinity data is further converted into the negative log 10 scale, which is denoted as pIC50. Lastly, we cleaned the datasets by removing contradictory entries and duplicates.

Notably, IC50/EC50/Ki/Kd are often heteroscedastic[52]. Consequently, measurement errors are not equally distributed throughout the range of activity and, therefore, regression of the raw pIC50 values may not be favorable. Thus, one common practice is to convert direct regression into a binary classification task. For the active vs. inactive threshold, 1 μM (pIC50 at 6) is usually used as the default cutoff[53]. In this study, we also adopted pIC50 at 6 as the cutoff (Figure 4).

Dataset	Task	#Mol.	Max. Len.	#Scaff.
MDR1	CLS-REG	1,438	252	602
CYP2D6	CLS-REG	2,293	217	1,330
CYP3A4	CLS-REG	3,671	244	2,022
MOR	CLS-REG	3,553	373	1,623
DOR	CLS-REG	3,223	373	1,531
KOR	CLS-REG	3,326	373	1,660

Table 3. Summary of the Opioids-related Datasets.

The opioids-related datasets, together with the task type, number of molecules, maximum SMILES string length and number of scaffolds, are summarized in Table 3. Since each dataset has a maximum length less than 400, the collected molecules are 100% retained. For the opioids-related datasets, we also applied both scaffold and random split (Figure 4), each of which was repeated 30 times with 30 different seeds $(0, 1, 2, \cdots, 29)$ in GROVER's implementation for dataset split. The split folds were also saved and kept the same in all subsequent experiments. The opioids-related datasets and the split folds are also provided in our Github repository.

3.2 Evaluation metrics

In Section 2.4.2, we discussed why the recommended metric may not be proper for model evaluation. Details on the set of basic metrics used in this study are illustrated as follows. Note there other enrichment metrics in early discovery, such as Boltzmann-Enhanced Discrimination of ROC (BEDROC)[11].

3.2.1 Classification metrics

In a binary classification task setting, a probability for each molecule as the positive (or active) class is usually assigned. When the estimated probability is greater than a threshold (a value between 0 and 1), the molecule is classified as positive (or active), otherwise negative (or inactive). In total, there are four possible outcomes: true positive (TP), false positive (FP), true negative (TN) and false negative (FN). Based on the TP and FP rates across different probability thresholds, the receiver operating characteristic curve can be plotted with the area under the ROC curve as AUROC. Similarly, based on

precision and recall, the precision-recall curve can be plotted and AUPRC is calculated likewise. AUROC usually ranges from 0.5 and 1, with 0.5 for random classification and 1 for perfect classification; in the case of a classifier worse than random guessing, AUROC can be lower than 0.5. Compared to accuracy, AUROC is more robust in the case of imbalanced datasets[17]. Nonetheless, by assuming equal importance of positive and negative classes, AUROC may suffer from imbalanced datasets, especially when the minor class is more of interest[17]. Thus, AUPRC is proposed as an alternative when datasets are imbalanced[44], with the baseline value as the proportion of the minor class.

$$PPV = \frac{TP}{TP+FP} \tag{1}$$

$$NPV = \frac{TN}{TN+FN} \tag{2}$$

Despite the usefulness of AUROC and AUPRC, these "collective" metrics may not be directly pertinent to virtual screening[17], a common application for molecular property prediction[5]. In fact, the primary goal of early drug discovery is to rank molecules based on the predicted activity so as to avoid the intractable number of false positives or false negatives to be filtered out experimentally[54]. Given a set of predicted actives or inactives, depending on the screening goal, we argue that positive predictive value (PPV; Equation 1) and negative predictive value (NPV; Equation 2) are more relevant to virtual screening and drug design, as discussed in Section 2.4.2. Notably, unlike AUROC and AUPRC which are averaged across different probability thresholds, a decision threshold is determined first before deriving the TP, FN, TN and FP, based on which PPV and NPV are calculated. When the datasets are balanced, the decision threshold for positive vs. negative is set as 0.5. However, for imbalanced datasets, the threshold needs to be adjusted. Here, we used Youden's J statistic[55], the distance between the ROC curve and the random chance line, to adjust threshold which maximizes the J statistic.

3.2.2 Regression metrics

For the regression task, RMSE (Equation 3) and MAE (Equation 4) are the recommended metrics, which indicate how apart the predicted values are from the labels: a lower value stands for a better model fit. MAE is a natural measure of average error whereas RMSE gives more weight to large errors and is more sensitive to the outliers. In addition, two other metrics can also measure the regression performance[8,16,56], namely, Pearson_R and R2, which are scale independent.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2} \tag{3}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i-\hat{y}_i| \tag{4}$$

$$Pearson\_R = \frac{\sum_{i=1}^{N}(y_i-\bar{y}_{obs})(\hat{y}_i-\bar{y}_{pred})}{\sqrt{\sum_{i=1}^{N}(y_i-\bar{y}_{obs})^2\sum_{i=1}^{N}(\hat{y}_i-\bar{y}_{pred})^2}} \tag{5}$$

$$R2 = 1-\frac{\sum_{i=1}^{N}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{N}(y_i-\bar{y}_{obs})^2} \tag{6}$$

Pearson_R is an intuitive measure of the linear correlation between the predicted values and labels[57], which is the ratio between the covariance of two variables and the product of their standard deviations (Equation 5), ranging from -1 to 1. An absolute value of 1 indicates a perfect linear relationship between the predicted values and labels. Notably, some studies used Pearson_R[56] while other papers used the square of Pearson_R[16,19], i.e., Pearson_R2, which always ranges from 0 to 1. On the other hand, R2, also known as the coefficient of determination, is not based on correlation. Instead, R2 calculates the proportion of the variance in the predicted values that can be explained by the labels (Equation 6). Usually, R2 ranges from 0 to 1. A higher R2 corresponds to a better model fit. The best scenario is that the predicted values exactly match the observed values and R2 would be 1. On the contrary, the baseline case is that the model always predicts $\bar{y}_{obs}$ and R2 would be 0. Worse still, when models have poorer predictions than the baseline, R2 can even be negative. Presumably due to naming similarity, R2 and Pearson_R2 can sometimes be messed with each other. In this study, we included both Pearson_R and R2, which are calculated with the scipy package and the scikit-learn package, respectively.

3.3 Model training

For RF, we followed Chemprop[8] and set the number of trees as 500. For the hyperparameters in MolBERT and GROVER, we adopted the optimal hyperparameters reported in the original papers[11,13]. All experiments with the neural networks were run on a single NVIDIA V100 GPU. For the HIV dataset, MolBERT takes around 3 hours to complete a 100-epochs training in each split fold when the batch size is 32. Since GROVER takes even more time, we set the batch size as 256 when applying GROVER on HIV, which still takes around 5 hours to complete a 100-epoch training. To ensure fair comparison, we saved the raw predictions, based on which prediction performance were calculated using same evaluation codes, which are also provided in our Github repository.

3.4 Statistical analyses

To examine if the prediction performance is significantly improved by the representation learning models, we conducted statistical analyses on the prediction performance. Usually, two major categories of analyses can be applied: parametric and non-parametric tests[58] (Table 4). Among them, the parametric t tests examine whether two groups have equal means, which can be further categorized into paired t test and unpaired t test. For the paired t test, the null hypothesis

Statistical Test	Alias	Parametric	Normality	Equal Variance	Equal Size
Paired t test	Dependent t test	✓	✓	✓	✓
Unpaired t test	Independent or Welch's t test	✓	✓	×	×
Wilcoxon signed-rank test	-	×	×	✓	✓
Wilcoxon rank-sum test	Mann-Whitney U test	×	×	×	×

Table 4. Summary of the Statistical Tests.

is that the means of two populations are equal, underlying by the equal-variance assumption. When two samples have unequal variances and/or unequal sample sizes, the unpaired or independent t test, a.k.a. Welch's t test, should be used. Notably, the paired and independent t tests are parametric tests with the normality assumption. When the sample size is large, the t tests can still be robust to moderate violations of the normality assumption. Nonetheless, if the normal assumption is violated and the sample size is small, non-parametric tests should be used instead, namely, Wilcoxon signed-rank test and Wilcoxon rank-sum test. Wilcoxon signed-rank test uses the signed rank to compare the medians of two populations, which is a non-parametric version of the paired t test. Wilcoxon rank-sum test, a.k.a. Mann-Whitney U test, also compares the medians but is robust to the violations of homoscedasticity. Neither Wilcoxon signed-rank test nor Wilcoxon rank-sum test requires the normality distribution. However, when the data are normally distributed, the non-parametric tests may lead to less statistical power, which corresponds to a higher probability of making the type II error (false negative)[17,58].

After examining the distribution of evaluation metrics, we adopted the non-parametric Mann-Whitney U test to compare the prediction performance. The statistical significance is defined as the two-sided p value less than 0.05.

4 Results

4.1 On the datasets

4.1.1 Dataset profiling

To have a clear grasp on the datasets, we first presented the label profiling for the benchmark datasets and the opioids-related datasets (Figure 5). As shown in Figure 5 (a), BACE is balanced with a positive rate of 45.7%, whereas BBBP is imbalanced towards the positives (76.5%) and HIV has much less positive instances (3.5%). The labels of ESOL, FreeSolv and Lipop all exhibit left-skewed distribution, especially for FreeSolv. On the other hand, the pIC50 distribution for the opioids-related datasets is right-skewed as shown in Figure 5 (b), presumably because most screened molecules exhibited low activity to the targets. To construct the opioids-related datasets in the classification setting, we applied a cutoff at 6 on the raw pIC50 values to convert them into binary values, namely active and inactive, abiding by the rule that pIC50 less than 6 inactive otherwise active. As shown in Figure 5 (b), the resultant datasets are all imbalanced. For MOR, DOR and KOR, the positive rates are 29.7%, 23.3% and 27.8%, respectively. For MDR1, CYP2D6, CYP3A4, the positive rates are even lowered to 9.1%, 1.4% and 2.2%.

To quantify the difference of label distributions, we calculated the Kolmogorov D statistic[59] among the training, validation and test sets (Sup. Figure 1 (a)). When using scaffold split, the D statistic is more dispersed with a higher median than that when using random split, suggesting that scaffold split leads to larger gaps in the label distributions in addition to separating molecules by the scaffolds, which, to some extent, manifests that molecules with same scaffolds tend to have similar properties. Random split, on the other hand, results in a more compact distribution of the D statistic with a lower median, indicating that training and test sets are more likely to have molecules with close labels. To quantify the structural similarity among the training, validation and test sets, we also calculated the Tanimoto similarity[60], shown in Sup. Figure 1 (b). Likewise, Tanimoto similarity exhibits more compact distribution under random split, the median of which are also higher, showing that training and test molecules are more structurally similar compared to those under scaffold split.

We also calculated the percentage of top fragments, i.e., heterocycles and functional groups, for both benchmark and opioids-related datasets, which are summarized in Sup. Figures 2 & 3. The top heterocycles vary across different datasets, which also manifest their unique pharmacological properties. For instance, in Sup. Figure 3 (a), the top heterocycle is piperidine for MOR, DOR and KOR, which is commonly seen in opioid analgesics[61]. For the functional groups, all datasets share top functional groups such as benzenes and amines, which are common components to facilitate the interaction with drug targets, typically proteins with abundant amino acid residues, via forming hydrogen bonds or π-π stacking interactions[62]. Other structural traits, such as the molecular weight (MW), number of rotatable bonds and the number of rings, are summarized in Sup. Figure 4.

4.1.2 Activity cliffs

For molecular property prediction, the essential goal is to achieve chemical space generalization, which, ideally, encompasses both inter-scaffold generalization and intra-scaffold generalization (see Section 2.4.3). Scaffold split, which ensures no scaffold overlap among training, validation and test sets, addresses the inter-scaffold challenge and has been widely adopted in the literature. However, the intra-scaffold generalization, especially in the case of activity cliffs, has not been well addressed yet. Thus, we proposed to evaluate the intra-scaffold generalizability of the representation learning models using the opioids-related datasets.

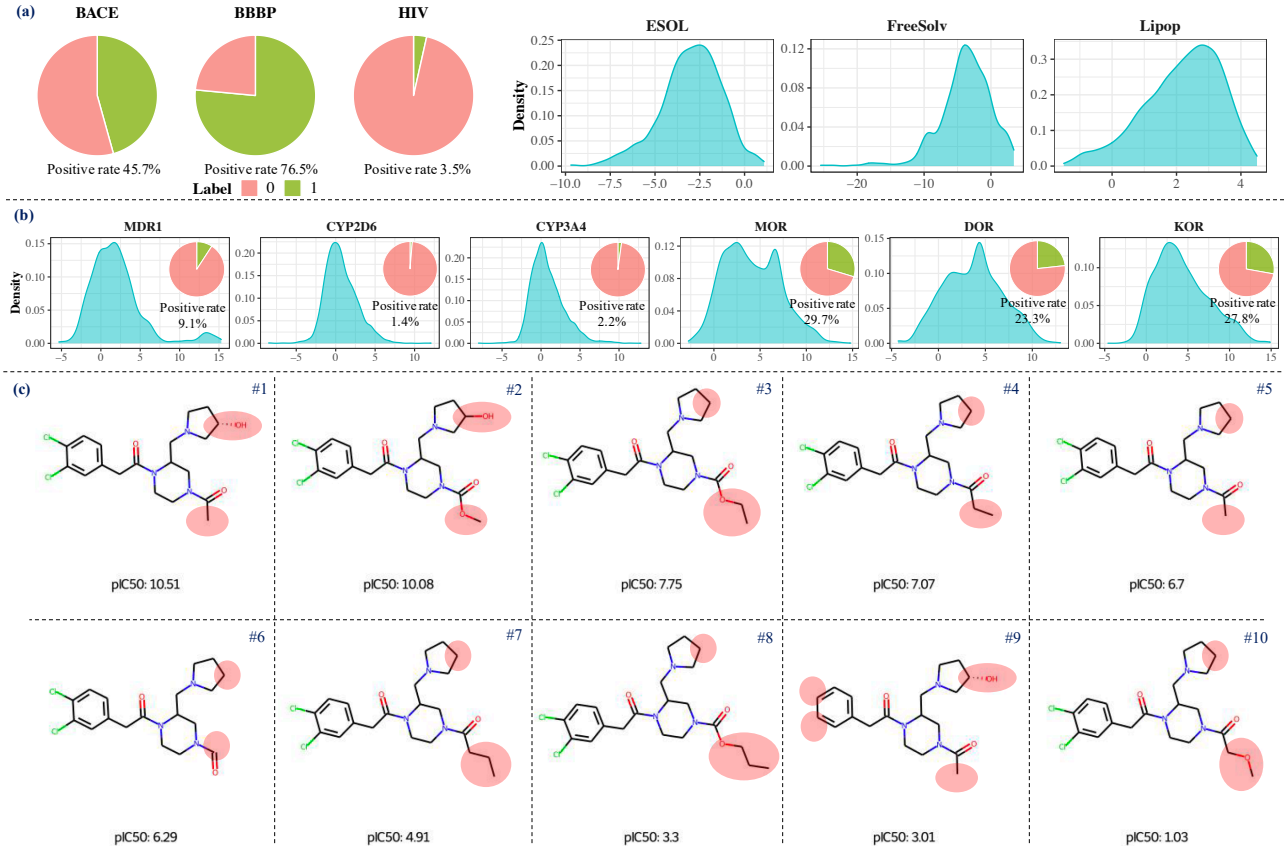

Figure 5. Datasets Profiling for the Benchmark and Opioids-related Datasets. (a). Label distribution for the selected MoleculeNet benchmark datasets. (b). Label distribution for the opioids-related datasets. (c). Activity-cliffs showcase on a series of molecules with the KOR Top 14 scaffold (Sup. Figure 10).

For each dataset, we visualized the top 30 scaffolds along with its pIC50 distribution (Sup. Figures 5-10). To showcase the activity cliff where analogs exhibit drastic difference in potency, we used the KOR Top 14 scaffold (Sup. Figure 10) for illustration. In Figure 5 (c), the replacement of the two hydrogen atoms with the chlorine atoms at the phenyl ring from molecule #9 to molecule #1 results in a drastic activity increase by 7 orders of magnitude, which, presumably, is because the chlorine atoms help the ligand better occupy the hydrophobic space in the binding pocket, an important contributor for binding. When comparing molecule #1 to molecule #5, the introduction of the hydroxyl group at the pyrrolidine ring increases the potency by 4 orders of magnitude, which also indicates that a potential H-bond interaction with the receptor is crucial for binding. Meanwhile, although shortening the acetyl group to the aldehyde group poses a minor reduction in activity when contrasting molecule #5 to molecule #6, longer side chains (molecule #7 & #8) can undermine activity, which suggests limited space around the binding site.

These molecules show that major activity change can happen even with minor structural changes. More formally, the activity cliff is defined as IC50 values spanning at least two orders of magnitude within the same scaffold[50,63]. Notably, there are also studies which adopted one order of magnitude[50]. In this study, the scaffolds observed with activity cliffs are termed as the AC scaffolds and the molecules with the AC scaffolds are denoted as the AC molecules. The number (percentage) of AC scaffolds and AC molecules are summarized in Table 5. Notably, among MDR1, MOR, DOR and KOR, nearly half molecules are equipped with the AC scaffolds, although the percentages of AC scaffolds are around 10% in all opioids-related datasets.

Dataset	#AC. Scaff. (%)	#Mol. AC. Scaff. (%)
MDR1	62 (10.2)	594 (41.3)
CYP2D6	124 (9.3)	710 (31.0)
CYP3A4	146 (7.2)	926 (25.2)
MOR	213 (13.1)	1,627 (46.1)
DOR	178 (11.6)	1,342 (41.6)
KOR	218 (13.1)	1,502 (45.2)

Table 5. Summary of Activity Cliffs in the Opioids Datasets.

4.2 On the model evaluation

In this section, we focus on the prediction under scaffold split.

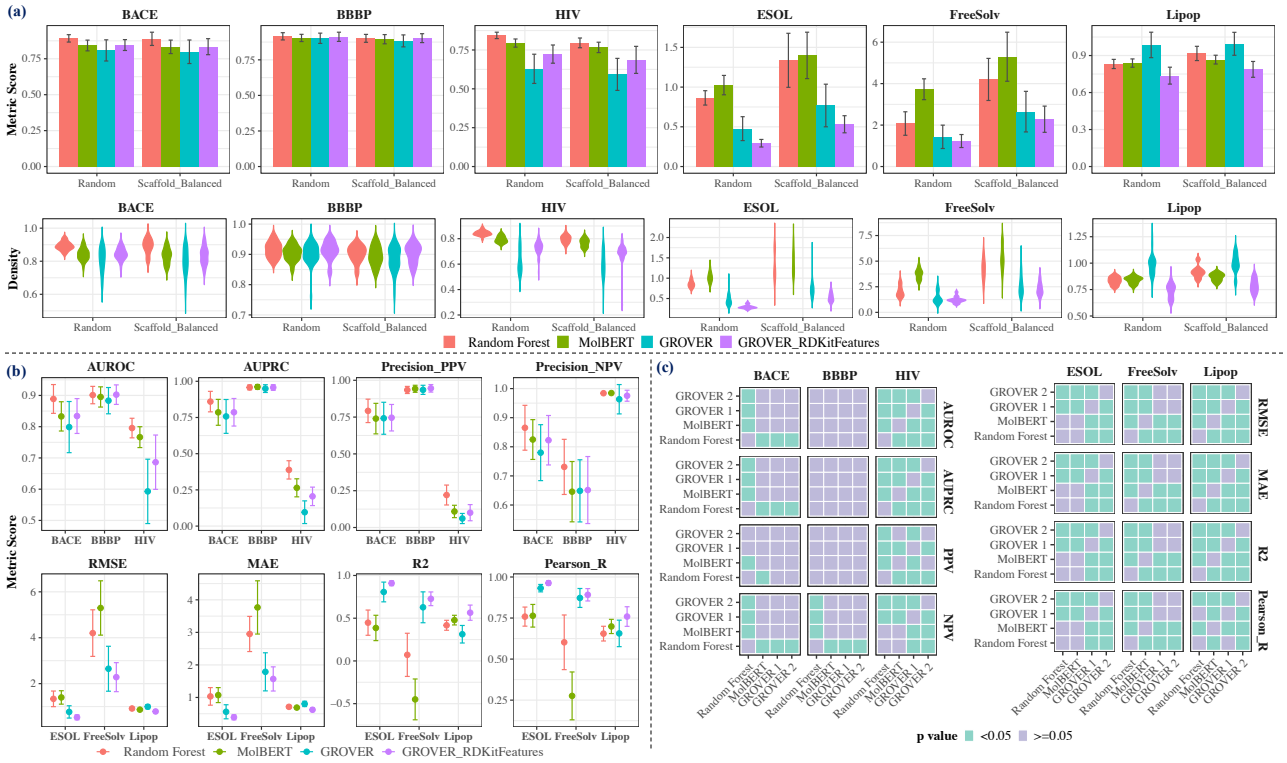

Figure 6. Comparison of Molecular Property Prediction Performance Using Benchmark Datasets. (a). Bar plot (mean±standard deviation) and violin plot of prediction performance using recommended metrics (AUROC for classification; RMSE for regression) under random and scaffold split. (b). Point plot (mean±standard deviation) of prediction performance under scaffold split. (c). Statistical significance using Mann-Whitney U test in prediction performance comparison between different models under scaffold split.

4.2.1 Does the learned representation surpass ECFP?

We first compared molecular property prediction performance of different models, namely, RF, MolBERT, GROVER and GROVER_RDKit, using the recommended evaluation metrics[16]. For ECFP, we followed Chemprop[8] and set radius as 2 and number of bits as 2048.

As shown in Figure 6 (a), RF achieves the highest AUROC in BACE and HIV ($p < 0.05$). MolBERT achieves similar performance as GROVER and GROVER_RDKit in BACE whereas in HIV, MolBERT outperforms both of them ($p < 0.05$). In fact, GROVER_RDKit also outperforms GROVER in HIV ($p < 0.05$). As for BBBP, all models achieve comparable AUROC. Thus, when evaluated by the benchmark classification datasets, RF on ECFP can be more useful than the representation learning models. On the other hand, when using the benchmark regression datasets, GROVER_RDKit can beat RF, MolBERT and GROVER, as indicated by the lowest RMSE ($p < 0.05$). Notably, without concatenating the RDKit features, the regression performance can be harmed, as shown by the significantly increased RMSE of GROVER over GROVER_RDKit in ESOL and Lipop. However, even without the concatenated features, GROVER still outperforms RF and MolBERT in ESOL and FreeSolv, which manifests the prediction power of GNNs. One exception, though, is in Lipop, where GROVER shows the highest RMSE ($p < 0.05$). Given the fact that ESOL and FreeSolv have less data points and shorter SMILES lengths compared to other benchmark datasets (Table 2), one viable hypothesis is, therefore, that GNNs can fit better when dataset size is small and molecules are relatively simple with shorter sequences.

To further examine the effectiveness of the learned representations, we also conducted evaluation using the opioids-related datasets under the classification setting, as shown in Figure 7 (a). For MDR1, RF achieves the best performance under scaffold split whereas MolBERT, GROVER and GROVER_RDKit show similar performance. For CYP2D6, which is imbalanced with a positive rate of 1.4% at cutoff 6, all models exhibit highly variant yet very limited performance with mean AUROC around 0.5, or even lower, when using GROVER and GROVER_RDKit. For CYP3A4, RF and GROVER_RDKit achieve similarly high performance under scaffold split ($p \geq 0.05$) whereas GROVER exhibits lowest AUROC among all models ($p < 0.05$). Notably, these PK-related datasets are all imbalanced, which could be why the AUROC values exhibit large variability across the 30-fold splits, especially for CYP2D6. On the contrary, for the PD-related datasets, i.e., MOR, DOR and KOR, the metric variability is smaller but RF still outperforms MolBERT,

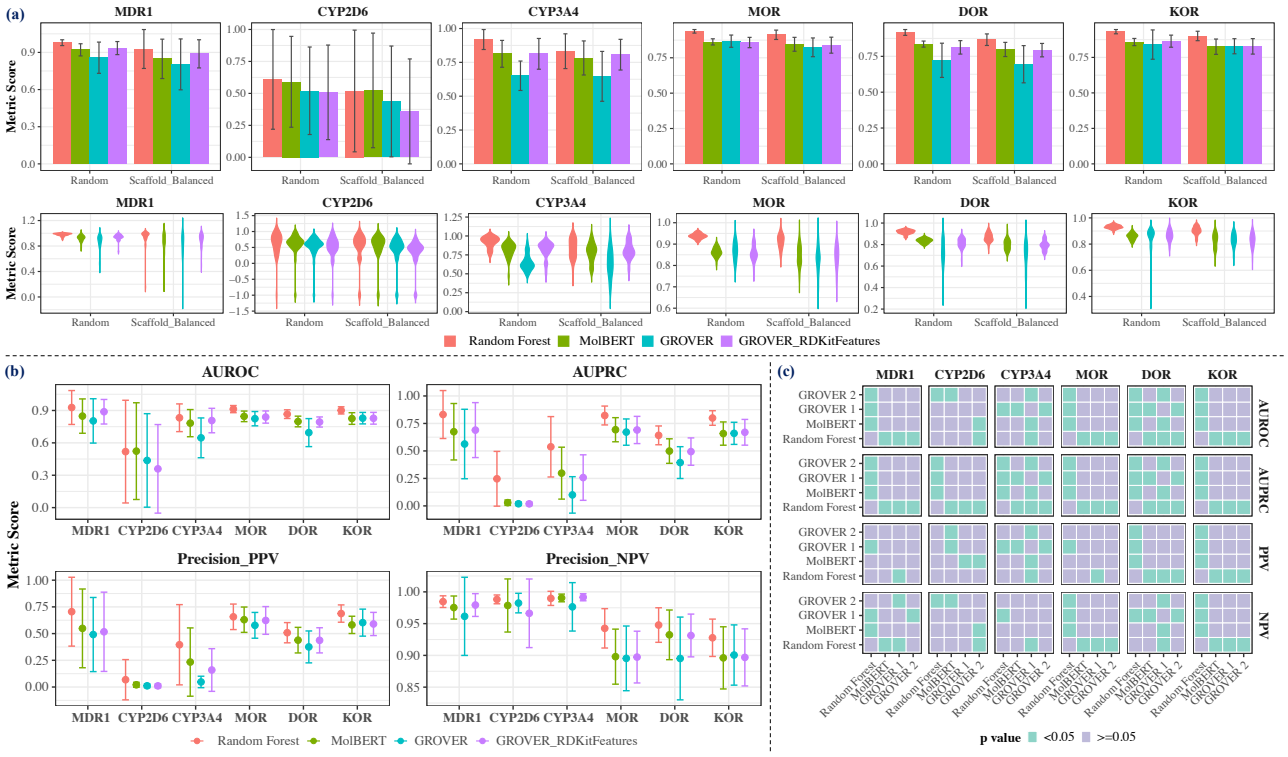

Figure 7. Molecular Property Prediction Performance using Opioids-related Datasets in Classification Setting. (a). Bar plot (mean±standard deviation) and violin plot of prediction performance using recommended metrics (AUROC for classification) under random and scaffold split. (b). Point plot (mean±standard deviation) of prediction performance under scaffold split. (c). Statistical significance using Mann-Whitney U test in prediction performance comparison between different models under scaffold split.

GROVER and GROVER_RDKit ($p < 0.05$), which, together with the observation on the PK-related datasets, suggest that ECFP can still be very useful.

Moreover, we observed that the RDKit-derived descriptors, originally proposed to concatenate with the learned representations in Chemprop[8], have a non-negligible role in GROVER's prediction performance. For instance, compared to GROVER, GROVER_RDKit achieves significantly higher AUROC in BACE and HIV while exhibits significantly lower RMSE in ESOL and Lipop in Figure 6 (a). In the opioids-related datasets, compared to GROVER, GROVER_RDKit achieves significantly higher AUROC in CYP3A4 and DOR (Figure 7). Thus, the concatenation of RDKit descriptors to the learned representations may be misleading when evaluating the real power of the representation learning models.

However, with the concatenation of RDKit descriptors, the variability of AUROC associated with GROVER can be reduced, as indicated by the violin plots in Figure 6 (a) and Figure 7 (a). For example, different models not only generate varying mean AUROC values but also exhibit different extents of variation in AUROC. Generally, RF on ECFP shows the lowest extent of dispersion, whereas GROVER on molecular graphs generates AUROC varied greatly across split folds, indicating high prediction uncertainty. Again, ECFP corresponds to not only better but also more stable prediction performance, which can be, therefore, applied more reliably.

4.2.2 Are the statistical tests necessary?

To demonstrate the necessity of statistical tests, we conducted a simple analysis using the benchmark datasets, which tries to answer: under scaffold split, the widely-adopted practice, how many 1-fold or 3-fold split(s) out of the 30 split folds are there for a certain model to be concluded as the best based on the average metric value alone? Note that GROVER_RDKit is removed from this analysis since concatenation of RDKit descriptors can significantly bias the comparison of the representation learning models (see Section 4.2.1).

Model	RANDOM FOREST	MOLBERT	GROVER[1]
BACE	28	2	0
BBBP	12	9	9
HIV	26	4	0
ESOL	0	0	30
FreeSolv	1	0	29
Lipop	5	23	2

Table 6. Number of Single Fold where a Model Achieves the Best Performance under Scaffold Split.

For RF, MolBERT and GROVER (denoted as GROVER[1]), we first calculated the number of single test fold where a model achieves best performance using the recommended metrics (Table 6). For the classification datasets, RF dominates AUROC in BACE and HIV across 28 and 26 folds, respectively, whereas GROVER does not excel in any test fold, which is consistent with the finding in Section 4.2.1. Still, there are a few folds where MolBERT achieves the highest AUROC in BACE and HIV, which means there is a chance to wrongly conclude MolBERT as best-performing, despite that RF has the highest AUROC over 30 folds (see Section 4.2.1). In BBBP, where RF, MolBERT and GROVER exhibit comparable AUROC, there are around 10 folds for each model to achieve the highest AUROC. This can be seen as a manifestation of statistical noise, which makes it prone to an erroneous conclusion if only tested on some random fold. To emulate the common 3-fold practice, we also calculated the number of 3-fold combinations out of the 30 folds, where a specific model achieves the best performance based on the algorithmic mean of the recommended metrics. Table 7 shows that there are quite some 3-fold combinations where one model can be mistaken as best performing using the recommended metrics. For the analysis results using other metrics, readers can refer to Sup. Tables 1 & 2.

Model	RANDOM FOREST	MOLBERT	GROVER[1]
BACE	3,924	136	0
BBBP	2,105	1,013	942
HIV	3,711	349	0
ESOL	0	0	4,060
FreeSolv	0	0	4,060
Lipop	397	3,597	66

Table 7. Number of 3-Fold Combinations where a Model Achieves the Best Performance under Scaffold Split.

Thus, without statistical tests, chances are that wrong conclusions can be drawn with regard to whether a new technique truly advances molecular property prediction performance. Moreover, since the benchmark datasets are publicly accessible, one caveat is that the test folds might be customized to cater to a specific model. Moreover, the test type for comparison also matters. As mentioned in Section 3.4, we adopted the Mann-Whitney U test to compare different models in a pairwise manner. Meanwhile, we also calculated statistical significance using other statistical tests. As shown in Sup. Figure 11&12, statistical significance can vary with the tests. For instance, RF and MolBERT show comparable RMSE in ESOL without statistically significant difference but when using the Wilcoxon signed-rank test, RF significantly outperforms MolBERT. Another example is that RF significantly outperforms MolBERT, GROVER and GROVER_RDKit in MDR1 but if using the unpaired t test, the difference among RF, MolBERT and GROVER is not significant. Therefore, when evaluating prediction performance, it is imperative to conduct statistical tests and report the test type.

4.2.3 Which evaluation metric is appropriate?

In MoleculeNet[16], each benchmark dataset comes with a recommended evaluation metric. However, in real-world drug discovery, the recommended metrics may not be pertinent (see Section 2.4.2). In this section, we compared the model performance using a set of evaluation metrics in addition to the recommended ones. For the classification datasets, we calculated AUROC, AUPRC, PPV and NPV (see Section 3.2.1); for the regression datasets, we calculated RMSE, MAE, R2 and Pearson_R (see Section 3.2.2). As shown in Figure 6 (b) & (c), when using the recommended AUROC, RF achieves similar performance with MolBERT, GROVER and GROVER_RDKit in BBBP ($p \geq 0.05$). However, if the evaluation metric is NPV, RF significantly outperforms all the other three models. Another example is that, when evaluated by AUROC, MolBERT achieves significantly higher performance in HIV compared to GROVER_RDKit; but when evaluated by PPV, MolBERT shows similar performance with GROVER_RDKit ($p \geq 0.05$). Besides, although MolBERT and GROVER have similar Pearson_R values in Lipop ($p \geq 0.05$), GROVER is significantly outperformed by MolBERT with R2 as the evaluation metric. Thus, different evaluation metrics may lead to disparate conclusions and caution is needed, especially for metrics with similar naming, such as R2 and Pearson_R.

As for which metrics are appropriate, we observed that for the opioids-related datasets, AUROC is generally over 0.75 (except for CYP2D6), whereas most AUPRC drops below 0.75 in Figure 7 (b). For MolBERT, GROVER and GROVER_RDKit, AUPRC drops to c.a. 0.25 in CYP3A4 (positive rate: 2.2%) and in CYP2D6 (positive rate: 1.4%), AUPRC is nearly zero. Thus, AUROC is over-optimistic and AUPRC can be a more proper metric for imbalanced datasets. Furthermore, despite the high AUROC (c.a. 0.90), AUPRC (c.a. 1.0) and PPV (c.a. 0.90) in BBBP, NPV drops to c.a. 0.75 for RF and c.a. 0.65 for the other three models in Figure 6 (b). In this case, although the collective metrics, namely AUROC and AUPRC, are nearly perfect, NPV can still be very limited. This can be an issue if the goal of virtual screening is to identify hits impermeable through the blood-brain barrier, since only around 65% are truly not permeable among the predicted negatives, . On the contrary, although the best AUROC in HIV achieves c.a. 0.80 as shown in Figure 6 (b), the best PPV is just c.a. 0.25. Also shown in Figure 7 (b), PPV is limited for the opioids-related datasets. For instance, the best PPV is c.a. 0.7 in MDR1, whereas in MOR, DOR and KOR, best PPV is around 0.6, let alone the extremely imbalanced CYP2D6 and CYP3A4, where PPV can drop to nearly zero. Thus, if the goal of virtual screening is to identify hits active for these targets, there can be a large proportion of false positives among the predicted actives. In short, precision metrics, such as PPV and NPV, can be more suitable for evaluating molecular property prediction under the classification setting, which is further contingent upon the emphasis on positives or negatives, i.e., the specific goal of virtual screening.

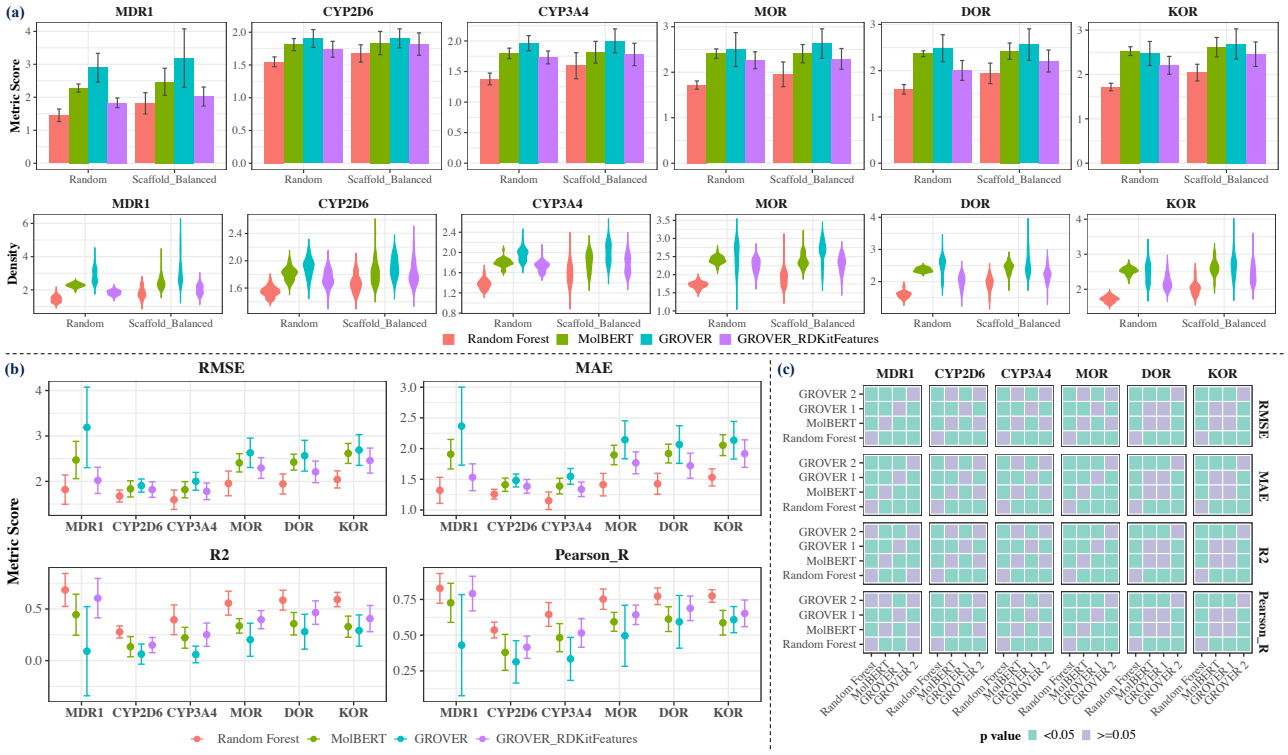

Figure 8. Evaluation of Molecular Property Prediction Using Opioids-related Datasets in Regression Setting. (a). Bar plot (mean±standard deviation) and violin plot of prediction performance using recommended metrics (RMSE for regression) under random and scaffold split. (b). Point plot (mean±standard deviation) of prediction performance under scaffold split. (c). Statistical significance using Mann-Whitney *U* test in prediction performance comparison between different models under scaffold split.

4.2.4 How does the task setting matter?

To study how task setting affects model evaluation, we conducted experiments to directly regress the pIC50 values in the opioids related datasets and the results are shown in Figure 8.

We first compared model performance using the recommended metric, RMSE. Despite the superior performance of GROVER_RDKit in two benchmark regression datasets, namely ESOL and FreeSolv (see Section 4.2.1), RF achieves the lowest RMSE in all raw opioids-related datasets compared to the other three models ($p < 0.05$). Even using other evaluation metrics, namely, MAE, R2 and Pearson_R, RF still achieves the best performance ($p < 0.05$). More specifically, MAE can be lowered to c.a. 1.5 by RF in all opioids-related datasets whereas for GROVER, MAE can be as high as c.a. 2.5 in MDR1. These observations, again, supports the effectiveness of ECFP for molecular property prediction.

Next, we compared the prediction performance under the classification and regression settings. As shown in Figure 7 (b), under the classification setting, all models achieve limited performance in CYP2D6, with particularly abysmal performance in PPV. However, under the regression setting, RMSE and MAE in CYP2D6 can be lowered to c.a. 1.5, which suggests that regression might be more suitable in predicting CYP2D6's activity, although the pIC50 labels can be noisy[52]. On the contrary, in MDR1, MOR, DOR and KOR, where the prediction performance is promising indicated by high AUROC values under the classification setting, RMSE and MAE are worsen when compared to the CYP2D6. One potential cause for the disparate performance between the classification and regression settings can be the arbitrary activity cutoff value. Since each dataset has a unique label distribution (Figure 5), the arbitrary cutoff value at 6 may cause prediction difficulty to different extents. For instance, similar molecular structures with close pIC50 values around 6 could be coerced to actives vs. inactives, which forms a challenging task and may act as a major source of misclassification.

4.3 On the chemical space generalization

In this section, we checked how the predictive models perform in terms of both inter-scaffold and intra-scaffold generalization, using the opioids-related datasets for illustration.

4.3.1 Inter-scaffolds generalization

As illustrated in Figure 4, we conducted experiments using both random and scaffold split for the opioids-related datasets. The prediction performance under scaffold split have been discussed in Section 4.2. Since there is no scaffold overlap among training, validation and test sets scaffold split, we com-

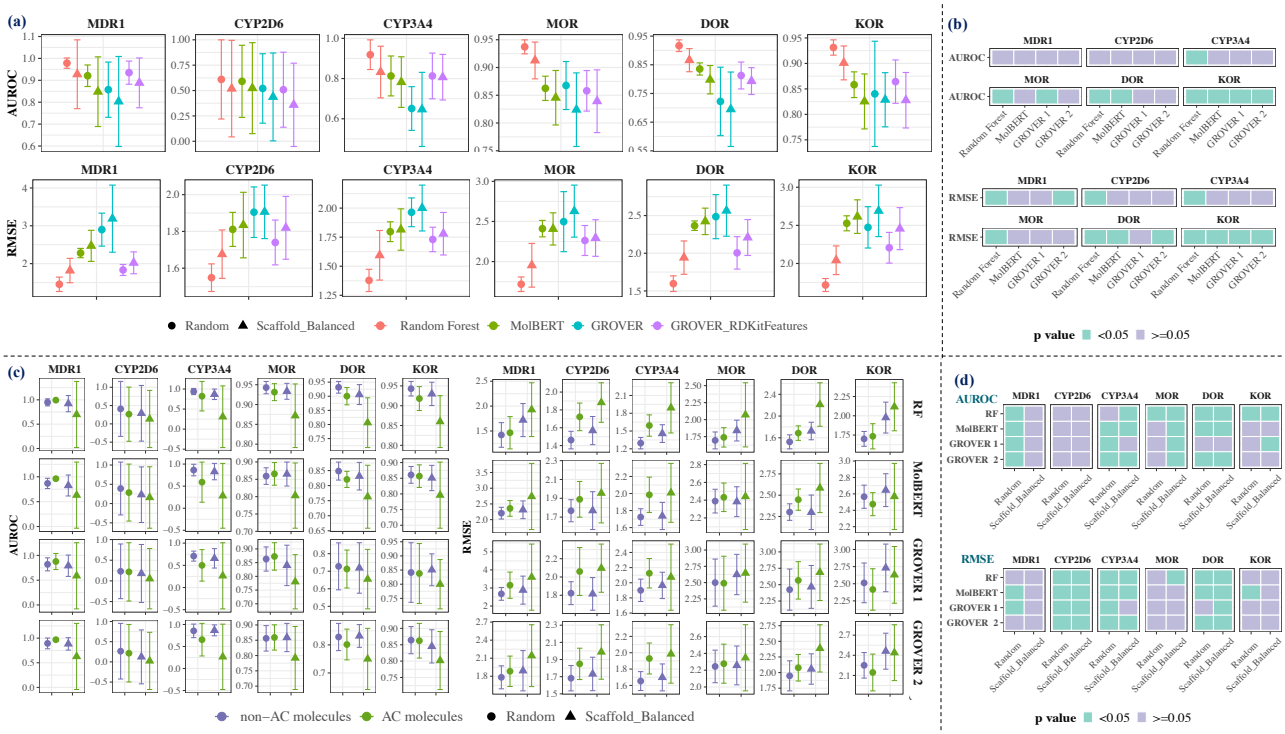

Figure 9. Prediction Performance Comparison on Chemical Space Generalization. (a). Point plot (mean±standard deviation) of prediction performance using recommended metrics (AUROC for classification; RMSE for regression) under random and scaffold split. (b). Statistical significance using Mann-Whitney *U* test in prediction performance comparison between random split and scaffold split. (c). Bar plot (mean±standard deviation) of prediction performance among AC and non-AC molecules using recommended metrics (AUROC for classification; RMSE for regression). (d). Statistical significance using Mann-Whitney *U* test in prediction performance comparison between AC and non-AC molecules.

pared the prediction performance between scaffold split and random split so as to evaluate how the models perform during the inter-scaffold generalization. To calculate the statistical significance, we adopted the Mann-Whitney *U* test to compare with the recommended metrics, which are summarized in Figure 9 (a) & (b). Compared to random split, prediction performance, as indicated by AUROC and RMSE, significantly deteriorates under scaffold split in most datasets using RF, manifesting the inter-scaffold generalization challenge. Meanwhile, using MolBERT, GROVER and GROVER_RDKit, similar AUROC and RMSE between the two split settings are more frequently observed, which can be due to their innate limited performance under both split schemes. Moreover, the dispersion extent of evaluation metrics is also higher under scaffold split, indicating higher uncertainty in prediction during inter-scaffold generalization.

4.3.2 Intra-scaffold generalization

To examine the intra-scaffold generalization, we compared the prediction performance for the test molecules with and without the AC scaffolds under both scaffold and random split, shown in Figure 9 (c) & (d). For the AC molecules, the prediction performance is generally worse compared to those of the non-AC molecules, as indicated by the lower AUROC and higher RMSE. The inferior performance for the AC molecules reflects the limited intra-scaffold generalization, especially in the case of activity cliffs. Clearly, there is a gap for intra-scaffold generalization. Nonetheless, for models exhibiting limited performance in Figure 7 (a), such as GROVER which exhibits low AUROC in DOR, the gap for intra-scaffold generalization may not be readily observed. For instance, there is a significant difference in prediction performance between AC and non-AC molecules in DOR when using RF, MolBERT and GROVER_RDKit, which, nonetheless, turns insignificant when using GROVER. Besides, the performance differences between AC and non-AC molecules are generally more prominent under scaffold split. In other words, random split can alleviate the intra-scaffold generalization in the case of activity cliffs, presumably due to that some AC scaffolds have been seen during training, which enables more accurate prediction at inference time. Again, this observation suggests the importance of scaffolds in molecular property prediction.

In fact, as pointed out by Robinson *et al.*[17], active molecules with different scaffolds can interact with the target with very different mechanisms. Thus, expecting a model to generalize by learning from unseen scaffolds can be somewhat unrealistic. Here, by interrogating the intra-scaffold generalization, we further substantiate this point with the activity-cliffs issue,

which has two-fold meanings: 1). exposing the predictive model to a set of diverse scaffolds during training phase can be conducive for inference, even holding potential for activity cliffs, although further study is needed; 2). when applying the trained molecular property prediction models, for example, in a deep reinforcement learning framework for drug design[64], if the generated molecule has a novel scaffold, which is further related to molecules with drastic activity changes, the predicted property for the molecule should be treated with lower confidence.

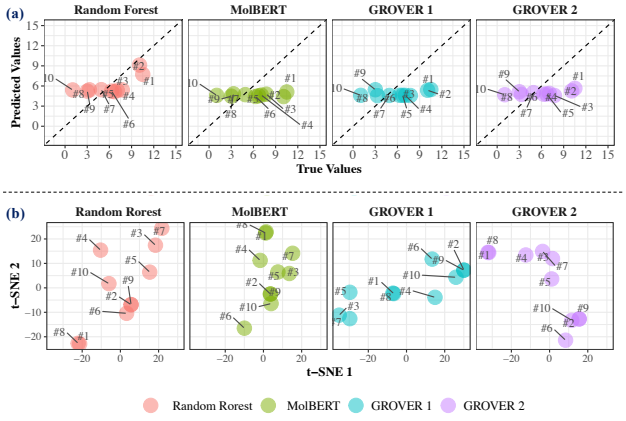

Figure 10. The Raw Predictions (a) and Learned Embeddings (b) for the KOR Activity-Cliffs Showcase Molecules from Figure 5 (c).

Furthermore, we also examined the learned representations for the AC showcase molecules in Figure 5 (c), taken from one test fold (seed: 4) under scaffold split. As shown in Figure 10 (a), the raw pIC50 predictions are abysmal, especially for MolBERT and GROVER, where the average pIC50 values seem to be simply imputed as the predicted values for the AC molecules, indistinguishable from each other. Unsurprisingly, the visualized embeddings, shown in Figure 10 (b), are also mixed despite their disparate pIC50 values, which is a manifestation of the structure-property mismatch[21].

Discussion

In this study, we take a respite from representation learning for molecular property prediction. Two self-supervised representation learning models based on SMILES strings and molecular graphs, MolBERT[11] and GROVER[13], are mainly investigated. Compared to supervised learning, self-supervised learning does not require heavy human annotations[65], which can be particularly expensive in drug discovery[1]. As shown in the work by Hu *et al.*[12], self-supervised pre-training helps avoid the negative transfer associated with supervised pre-training. Self-supervised learning can be roughly categorized into: generative, contrastive and generative-contrastive (adversarial)[66]. Pretraining tasks, such as masked language modeling in MolBERT and contextual property prediction in GROVER, are leaning towards the generative type.

Recently, the contrastive type of self-supervised pretraining has also been applied in molecular property prediction. For example, MolCLR[14] proposes three augmentation strategies, namely, atom masking, bond deletion and subgraph removal, on molecular graphs to pretrain GCN and GIN, respectively. For molecular graph pairs augmented from the same molecule, they are denoted as positive otherwise negative and the normalized temperature-scaled cross-entropy loss (NT-Xent) loss is applied, which maximizes the similarity between positive pairs and minimizes the similarity between negative pairs. More recently, iMolCLR[15] has also been proposed, where structural similarities between negative pairs and fragment-level contrasting substructures decomposed from molecules are integrated into the NT-Xent loss function. Notably, one pretraining task in MolBERT[11] is the SMILES equivalence prediction, where given an input of two SMILES strings (the second SMILES is either randomly sampled from the pretraining corpus or a synonymous permutation of the first SMILES), the model predicts whether they represent the same molecule. Based on the ablation study, however, the SMILES equivalence task slightly but consistently decreases downstream performance. Besides, MolBERT[11] and GROVER[13] also utilized RDKit[41] to calculate molecular descriptors values or extract graph-level motifs as domain-relevant labels for pretraining. As indicated by the ablation study in MolBERT, molecular descriptors values prediction has the highest impact on downstream performance. Also shown in this study, the 200 RDKit global molecular features play a crucial role in GROVER. We also observed that ECFP4 significantly outperforms the learned representations in many cases, which coincides with previous studies[36,67]. One potential direction would therefore be coming up with more suitable ways of exploiting molecular descriptors or fixed fingerprints.

Nonetheless, despite the advancement in AI techniques, whether AI can benefit real-world drug discovery is not without its concerns[23,24]. Guidelines for evaluating molecules generated by AI techniques have been suggested by Walters *et al.*[68]. Likewise, evaluation of molecular property prediction should also be standardized. Recently, Bender *et al.*[69] proposed a set of evaluation guidelines for machine learning tools, covering appropriate comparison methods and evaluation metrics, among others. In our study, we addressed three aspects of molecular property prediction on datasets profiling, model evaluation and chemical space generalization (Figure 1). For the datasets, each of them has unique label distribution and molecular structures, which poses different degrees of prediction difficulty. The molecular structures are dissected into scaffolds and structural traits, including fragments (functional groups and heterocycles) and other structural traits, such as MW and number of rings. Furthermore, under different dataset split schemes and with different random seeds, the structural and label divergence among the training, validation and test sets also vary, which acts as a source for the performance variance. For the model evaluation, we compared the two self-supervised representation learning

models, MolBERT and GROVER, with RF on the baseline fixed molecular features, ECFP4. With rigorous statistical analyses, ECFP4 is demonstrated to lead in the performance in many cases, which suggests there is still room for representation learning in molecular property prediction. Besides, we also found that different models exhibit different extents of dispersion in the evaluation metrics, which might reflect the inherent prediction uncertainty underlying each model. We envision that the metric variability should also be incorporated for model evaluation rather than just comparing the means. In a nutshell, there should be sufficient split folds and statistical analysis. Moreover, the use of evaluation metrics may also affect the conclusions drawn, which should be chosen based upon the question of interest. Arguably, the precision metrics should be reported together with the collective metrics such as AUROC, which is more pertinent to virtual screening. As a last component in the model evaluation, we also experimented on how the task setting affects the prediction performance for the opioids-related datasets. The disparate performance between the classification setting and regression setting suggests that applying one arbitrary activity cutoff value can be another source for prediction uncertainty due to potentially distinct structural traits and label distribution around the cutoff. For the chemical space generalization, we dissect it into inter-scaffold generalization and intra-scaffold generalization. The inferior prediction performance under scaffold split, compared to random split, suggests that better AI techniques are needed to improve w.r.t. inter-scaffold generalization. Similarly, the inferior performance for the AC molecules (see Section 2.4.3) suggests more efforts are needed for intra-scaffold generalization, especially in the case of activity cliffs. A more standard protocol on routine activity-cliffs evaluation is needed. One reason for the neglect of this real-world drug discovery challenge could be partially due to the heavy reliance on the MoleculeNet benchmark datasets.

Indeed, the widely-used benchmark datasets may be of little relevance to real-world drug discovery[18]. Besides, some benchmark datasets can pose unreasonable prediction tasks[24]. For example, SIDER[16] is a benchmark dataset for 1,427 marketed drugs and their side effects in 27 system organ classes. In addition to the molecular structures, there are many other factors underlying the side effects in humans, such as food-drug interactions[70], drug-drug interactions[51], among others[24]. Thus, it is unrealistic to expect a model to directly predict side effects from the chemical structures. Another example is ClinTox[16], which has a classification task for FDA approval status in addition to the clinical trial toxicity. Again, these two tasks can only be partially attributed to the chemical structures. Thus, to examine the usefulness of the advanced representation learning models, we also assembled a suite of opioids-related datasets from ChEMBL. As shown in Figure 5, the MOR, DOR and KOR datasets related to the PD aspect of opioid overdose are quite balanced. On the contrary, the CYP2D6 and CYP3A4 datasets related to the PK aspect of opioid overdose are extremely skewed to the left with an active rate less than 10% under the cutoff 6. Consequently, the PPV for these two metabolic enzymes are very limited. Indeed, the datasets in some domains are still lacking[24]. Besides, these collected datasets from public databases can be very noisy. For example, we found quite some duplicates and contradictory records in the opioids-related datasets, which were removed from further analysis. Sometimes, even the established benchmark datasets need an extra "washing" step to ensure the data quality[36].

Last but not least, there are still some limitations in this study. Firstly, the sources underlying molecular property prediction include dataset split uncertainty, experimental data uncertainty, and model training uncertainty[52]. Our experimental scheme of repeating datasets split 30 times with different random seeds only partially addresses the uncertainty. Moreover, there could also be variations introduced during model training, such as random weight initialization and random mini-batch shuffling[71]. As a result, the ensembling technique has been proposed to alleviate the uncertainty related to model training and improve prediction accuracy[8], which was not evaluated in this study due to heavy computation burden. Another underlying assumption, yet often neglected, is that the collected datasets are usually regarded as the gold standard without any experimental errors, which, however, may not hold true. Experimental uncertainty needs to be taken into consideration as well[52]. Secondly, the explainability of the molecular property prediction models is not covered in this study, which is related to the explainable AI, aiming to make the predictions more understandable by domain experts[72].

In conclusion, we expect that by taking this respite, increased awareness of the key aspects underlying molecular property prediction can be raised, which will eventually bring even better AI techniques in this field.

References

1. Wouters, O. J., McKee, M. & Luyten, J. Estimated research and development investment needed to bring a new medicine to market, 2009-2018. *JAMA* **323**, 844–853 (2020).
2. Simoens, S. & Huys, I. R&d costs of new medicines: A landscape analysis. *Front. Med.* **8** (2021).
3. Chen, H., Engkvist, O., Wang, Y., Olivecrona, M. & Blaschke, T. The rise of deep learning in drug discovery. *Drug Discov. Today* **23**, 1241–1250 (2018).
4. Vamathevan, J. *et al.* Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discov.* **18**, 463–477 (2019).
5. Deng, J., Yang, Z., Ojima, I., Samaras, D. & Wang, F. Artificial intelligence in drug discovery: applications and techniques. *Brief. Bioinforma.* **23**, bbab430 (2022).
6. David, L., Thakkar, A., Mercado, R. & Engkvist, O. Molecular representations in ai-driven drug discovery: a review and practical guide. *J. Cheminformatics* **12**, 1–22 (2020).

7. Mayr, A. *et al.* Large-scale comparison of machine learning methods for drug target prediction on chembl. *Chem. Sci* **9**, 5441–5451 (2018).

8. Yang, K. *et al.* Analyzing learned molecular representations for property prediction. *J. Chem. Inf. Model.* **59**, 3370–3388 (2019).

9. Honda, S., Shi, S. & Ueda, H. R. Smiles transformer: pre-trained molecular fingerprint for low data drug discovery. *arXiv preprint arXiv:1911.04738* (2019).

10. Chithrananda, S., Grand, G. & Ramsundar, B. Chemberta: Large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885* (2020).

11. Fabian, B. *et al.* Molecular representation learning with language models and domain-relevant auxiliary tasks. *arXiv preprint arXiv:2011.13230* (2020).

12. Hu, W. *et al.* Strategies for pre-training graph neural networks. *arXiv preprint arXiv:1905.12265* (2019).

13. Rong, Y. *et al.* Grover: Self-supervised message passing transformer on large-scale molecular data. *arXiv preprint arXiv:2007.02835* (2020).

14. Wang, Y., Wang, J., Cao, Z. & Farimani, A. B. Molclr: Molecular contrastive learning of representations via graph neural networks. *arXiv preprint arXiv:2102.10056* (2021).

15. Wang, Y., Magar, R., Liang, C. & Barati Farimani, A. Improving molecular contrastive learning via faulty negative mitigation and decomposed fragment contrast. *J. Chem. Inf. Model.* (2022).

16. Wu, Z. *et al.* Moleculenet: a benchmark for molecular machine learning. *Chem. Sci* **9**, 513–530 (2018).

17. Robinson, M. C., Glen, R. C. *et al.* Validating the validation: reanalyzing a large-scale comparison of deep learning and machine learning models for bioactivity prediction. *J. Comput. Aided Mol.* 1–14 (2020).

18. Walters, W. P. & Barzilay, R. Critical assessment of ai in drug discovery. *Expert. Opin Drug Discov* 1–11 (2021).

19. Shen, W. X. *et al.* Out-of-the-box deep learning prediction of pharmaceutical properties by broadly learned knowledge-based molecular representations. *Nat. Mach. Intell.* **3**, 334–343 (2021).

20. Xiong, Z. *et al.* Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *J. Med. Chem.* **63**, 8749–8760 (2019).

21. Na, G. S., Chang, H. & Kim, H. W. Machine-guided representation for accurate graph-based molecular machine learning. *Phys. Chem. Chem. Phys.* **22**, 18526–18535 (2020).

22. Mendez, D. *et al.* Chembl: towards direct deposition of bioassay data. *Nucleic Acids Res.* **47**, D930–D940 (2019).

23. Bender, A. & Cortes-Ciriano, I. Artificial intelligence in drug discovery: what is realistic, what are illusions? part 1: Ways to make an impact, and why we are not there yet. *Drug Discov. Today* (2020).

24. Bender, A. & Cortes-Ciriano, I. Artificial intelligence in drug discovery: what is realistic, what are illusions? part 2: a discussion of chemical and biological data used for ai in drug discovery. *Drug Discov. Today* (2021).

25. Gao, K. *et al.* Are 2d fingerprints still valuable for drug discovery? *Phys. Chem. Chem. Phys.* **22**, 8373–8390 (2020).

26. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).

27. Skinnider, M. A., Stacey, R. G., Wishart, D. S. & Foster, L. J. Deep generative models enable navigation in sparsely populated chemical space. *Nat Mach Intell* (2021).

28. Weininger, D. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *J Chem Inf. Comput. Sci* **28**, 31–36 (1988).

29. Weininger, D., Weininger, A. & Weininger, J. L. Smiles. 2. algorithm for generation of unique smiles notation. *J Chem Inf. Comput. Sci* **29**, 97–101 (1989).

30. Goh, G. B., Hodas, N. O., Siegel, C. & Vishnu, A. Smiles2vec: An interpretable general-purpose deep neural network for predicting chemical properties. *arXiv preprint arXiv:1712.02034* (2017).

31. Kipf, T. N. & Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).

32. Veličković, P. *et al.* Graph attention networks. *arXiv preprint arXiv:1710.10903* (2017).

33. Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O. & Dahl, G. E. Neural message passing for quantum chemistry. In *ICML*, 1263–1272 (PMLR, 2017).

34. Xu, K., Hu, W., Leskovec, J. & Jegelka, S. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826* (2018).

35. Breiman, L. Random forests. *Mach. learning* **45**, 5–32 (2001).

36. Jiang, D. *et al.* Could graph neural networks learn better molecular representation for drug discovery? a comparison study of descriptor-based and graph-based models. *J. Cheminformatics* **13**, 1–23 (2021).

37. Cano, G. *et al.* Automatic selection of molecular descriptors using random forest: Application to drug discovery. *Expert. Syst. with Appl.* **72**, 151–159 (2017).

38. Brown, N., Fiscato, M., Segler, M. H. & Vaucher, A. C. Guacamol: benchmarking models for de novo molecular design. *J. Chem. Inf. Model.* **59**, 1096–1108 (2019).

39. Shaw, P., Uszkoreit, J. & Vaswani, A. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 464–468, DOI: 10.18653/v1/N18-2074 (Association for Computational Linguistics, New Orleans, Louisiana, 2018).

40. Wieder, O. *et al.* A compact review of molecular property prediction with graph neural networks. *Drug Discov. Today* **37**, 1–12 (2020).

41. Landrum, G. Rdkit: Open-source cheminformatics software. *RDKit* (2016).

42. for Disease Control, C., Prevention *et al.* Drug overdose deaths in the united states, 1999–2018. *NCHS Data Brief: Natl. Cent. for Heal. Stat.* (2020).

43. Deng, J. *et al.* A large-scale observational study on the temporal trends and risk factors of opioid overdose: Real-world evidence for better opioids. *Drugs-Real World Outcomes* 1–14 (2021).

44. Saito, T. & Rehmsmeier, M. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one* **10**, e0118432 (2015).

45. Hu, Y. & Bajorath, J. What is the likelihood of an active compound to be promiscuous? systematic assessment of compound promiscuity on the basis of pubchem confirmatory bioassay data. *AAPS J.* **15**, 808–815 (2013).

46. Wale, N. & Karypis, G. Target fishing for chemical compounds using target-ligand activity data and ranking based methods. *J. Chem. Inf. Model.* **49**, 2190–2201 (2009).

47. Patrick Walters, W. Comparing classification models—a practical tutorial. *J. Comput. Aided Mol. Des.* 1–9 (2021).

48. Dobson, C. M. *et al.* Chemical space and biology. *Nature* **432**, 824–828 (2004).

49. Naveja, J. J. & Medina-Franco, J. L. Finding constellations in chemical space through core analysis. *Front. Chem.* **7**, 510 (2019).

50. Stumpfe, D., Hu, H. & Bajorath, J. Evolving concept of activity cliffs. *ACS Omega* **4**, 14360–14368 (2019).

51. Deng, J. & Wang, F. An informatics-based approach to identify key pharmacological components in drug-drug interactions. *AMIA Summits on Transl. Sci. Proc.* **2020**, 142 (2020).

52. Mervin, L. H. *et al.* Probabilistic random forest improves bioactivity predictions close to the classification threshold by taking into account experimental uncertainty. *J. Cheminformatics* **13**, 1–17 (2021).

53. Lagunin, A. A. *et al.* Comparison of quantitative and qualitative (q) sar models created for the prediction of ki and ic50 values of antitarget inhibitors. *Front. Pharmacol.* **9**, 1136 (2018).

54. Shoichet, B. K. Virtual screening of chemical libraries. *Nature* **432**, 862–865 (2004).

55. Schisterman, E. F., Faraggi, D., Reiser, B. & Hu, J. Youden index and the optimal threshold for markers with mass at zero. *Stat Med* **27**, 297–315 (2008).

56. Cortés-Ciriano, I. & Bender, A. Kekulescope: prediction of cancer cell line sensitivity and compound potency using convolutional neural networks trained on compound images. *J. Cheminformatics* **11**, 1–16 (2019).

57. Lu, J., Deng, K., Zhang, X., Liu, G. & Guan, Y. Neural-ode for pharmacokinetics modeling and its advantage to alternative machine learning models in predicting new dosing regimens. *Iscience* **24**, 102804 (2021).

58. Sedgwick, P. A comparison of parametric and non-parametric statistical tests. *BMJ* **350** (2015).

59. Massey Jr, F. J. The kolmogorov-smirnov test for goodness of fit. *J Am Stat Assoc* **46**, 68–78 (1951).

60. Todeschini, R. *et al.* Similarity coefficients for binary chemoinformatics data: overview and extended comparison using simulated and real data sets. *J. Chem. Inf. Model.* **52**, 2884–2901 (2012).

61. Smith, M. T., Kong, D., Kuo, A., Imam, M. Z. & Williams, C. M. Analgesic opioid ligand discovery based on nonmorphinan scaffolds derived from natural sources. *J. Med. Chem* (2022).

62. Bissantz, C., Kuhn, B. & Stahl, M. A medicinal chemist's guide to molecular interactions. *J. Med. Chem* **53**, 5061–5084 (2010).

63. Hu, Y., Stumpfe, D. & Bajorath, J. Advancing the activity cliff concept. *F1000Research* **2** (2013).

64. Deng, J., Yang, Z., Li, Y., Samaras, D. & Wang, F. Towards better opioid antagonists using deep reinforcement learning. *arXiv preprint arXiv:2004.04768* (2020).

65. Jing, L. & Tian, Y. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE PAMI* **43**, 4037–4058 (2020).

66. Liu, X. *et al.* Self-supervised learning: Generative or contrastive. *arXiv preprint arXiv:2006.08218* **1** (2020).

67. Lane, T. R. *et al.* Bioactivity comparison across multiple machine learning algorithms using over 5000 datasets for drug discovery. *Mol. Pharm.* **18**, 403–415 (2020).

68. Walters, W. P. & Murcko, M. Assessing the impact of generative ai on medicinal chemistry. *Nat. Biotechnol.* **38**, 143–145 (2020).

69. Bender, A. *et al.* Evaluation guidelines for machine learning tools in the chemical sciences. *Nat. Rev. Chem.* 1–15 (2022).

70. Deng, J. *et al.* A review of food–drug interactions on oral drug absorption. *Drugs* **77**, 1833–1855 (2017).

71. Fort, S., Hu, H. & Lakshminarayanan, B. Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757* (2019).

72. Jiménez-Luna, J., Grisoni, F. & Schneider, G. Drug discovery with explainable artificial intelligence. *Nat. Mach. Intell.* **2**, 573–584 (2020).

Acknowledgements

The experiments in this study are conducted using the computational resources provided by the AI institute in Stony Brook University. The authors also want to acknowledge the explicit codes and pretrained models from MolBERT and GROVER.

Author contributions statement

J.D. and Z.Y. conceived and designed the study. J.D. collected the datasets, conducted the experiments and analyzed the results. J.D. wrote the first draft of the manuscript. H. W. curated the chemistry-related contents. I. O., D.S. and F.W. supervised the project. F. W. provided funding support. All authors proofread the manuscript and made critical revisions.

Additional information

Accession codes The code, data and raw results are available in the Github repository: https://github.com/dengjianyuan/Respite_MPP. **Competing interests** The authors have no competing interest.