Exposing the Limitations of Molecular Machine Learning with Activity Cliffs

Citation for published version (APA):
van Tilborg, D., Alenicheva, A., & Grisoni, F. (2022). Exposing the Limitations of Molecular Machine Learning with Activity Cliffs. *Journal of Chemical Information and Modeling*, *62*(23), 5938-5951. https://doi.org/10.1021/acs.jcim.2c01073

Document license:
CC BY

DOI:
10.1021/acs.jcim.2c01073

Document status and date:
Published: 01/12/2022

Document Version:
Publisher’s PDF, also known as Version of Record (includes final page, issue and volume numbers)

JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

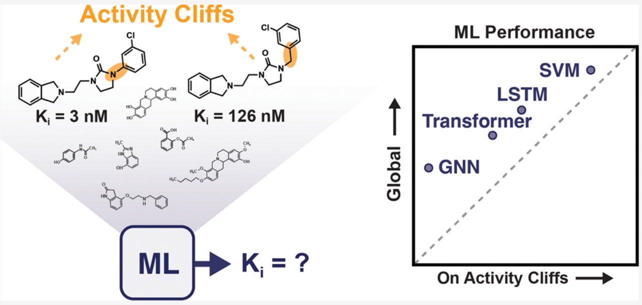
Exposing the Limitations of Molecular Machine Learning with Activity Cliffs

Derek van Tilborg, Alisa Alenicheva, and Francesca Grisoni*

Cite This: J. Chem. Inf. Model. 2022, 62, 5938−5951

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



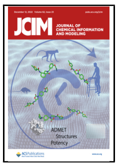
ABSTRACT: Machine learning has become a crucial tool in drug discovery and chemistry at large, *e.g.*, to predict molecular properties, such as bioactivity, with high accuracy. However, activity cliffs—pairs of molecules that are highly similar in their structure but exhibit large differences in potency—have received limited attention for their effect on model performance. Not only are these edge cases informative for molecule discovery and optimization but also models that are well equipped to accurately predict the potency of activity cliffs have increased potential for prospective applications. Our work aims to fill the current knowledge gap on best-practice machine learning methods in the presence of activity cliffs. We benchmarked a total of 24 machine and deep learning approaches on curated bioactivity data from 30 macromolecular targets for their performance on activity cliff compounds. While all methods struggled in the presence of activity cliffs, machine learning approaches based on molecular descriptors outperformed more complex deep learning methods. Our findings highlight large case-by-case differences in performance, advocating for (a) the inclusion of dedicated "activity-cliff-centered" metrics during model development and evaluation and (b) the development of novel algorithms to better predict the properties of activity cliffs. To this end, the methods, metrics, and results of this study have been encapsulated into an open-access benchmarking platform named MoleculeACE (Activity Cliff Estimation, available on GitHub at: https://github.com/molML/MoleculeACE). MoleculeACE is designed to steer the community toward addressing the pressing but overlooked limitation of molecular machine learning models posed by activity cliffs.

■ INTRODUCTION

In the last decade, artificial intelligence (AI) in the form of machine learning has permeated many domains of science. The chemical sciences have particularly benefited from the AI renaissance.[1−3] In multiple applications, machine learning has performed *on par* or even outperformed existing approaches, *e.g.*, for computer-assisted synthesis planning,[4−6] protein structure prediction,[7,8] and *de novo* molecular design.[9−11] Most AI breakthroughs in chemistry have been driven by deep learning—based on neural networks with multiple processing layers.[12−14] However, there is currently no consensus on whether deep learning models outperform simpler machine learning approaches when it comes to molecular property prediction.[15−17] The identification of current gaps in machine and deep learning approaches would allow the development of more reliable and widely applicable models to accelerate molecule discovery.

Molecular property prediction has the principle of similarity at its heart[18]—postulating that similar compounds are likely to have similar properties. Notably, one particular exception to this principle holds great insights into the underlying structure−activity (or structure−property) relationships.[19] Such an exception is constituted by activity cliffs[20]—pairs of structurally similar molecules that exhibit a large difference in their biological activity. Activity cliffs may cause machine learning models to remarkably mispredict the activity of certain molecules, even with an overall high model predictivity. Although generally constituting a source of "disappointment",[20] activity cliffs also encode valuable information for many applications[19] (*e.g.*, hit-to-lead optimization,[21,22] structural alert development[23]) since the large change in activity is induced by small structural changes.[24,25] Activity cliffs are particularly relevant in the context of virtual screening, with the number of highly similar molecules in commonly used commercial libraries varying between 10,000 and 170,000 (Supporting Table S1). While numerous studies have focused on defining activity cliffs,[19,24,26,27] their detrimental effect on machine learning models has been disproportionately under-investigated.[25] Arguably, models that can provide better predictions on activity cliffs are overall better, as they capture

Received: August 23, 2022
Published: December 1, 2022



https://doi.org/10.1021/acs.jcim.2c01073

the underlying "structure−activity landscape"[20] more accurately. Finally, although (macromolecular) structure-based approaches can aid in identifying discontinuities in the activity landscape,[28] ligand-based methods are routinely employed "out of the box" for virtual screening without incorporating considerations on activity cliffs.

Stemming from these considerations, the presented work has a threefold goal: (1) benchmark the performance of several machine and deep learning methods on activity cliffs, (2) quantify the effect of activity cliffs on the overall performance of machine learning, and (3) identify promising approaches and future directions in the field of molecular machine learning. To this end, we compared sixteen "traditional" machine learning methods—based on human-engineered features ("molecular descriptors"[29])—with seven deep learning approaches based on molecular strings or graphs to predict the biological activity of more than 35,000 molecules over 30 macromolecular targets. Our results highlight a generally poor performance of machine learning approaches on activity cliff compounds (particularly evident for deep learning), thereby further underscoring the relevance of assessing structure−activity "discontinuities" during model training and selection.

To further steer the community's efforts toward the relevant topic of activity cliffs, the results of our study were encapsulated in a dedicated benchmarking platform called MoleculeACE ("Activity Cliff Estimation"). MoleculeACE complements existing benchmarks and data sets for molecular property prediction[30−33] by providing a novel framework specifically focused on identifying activity cliffs and quantifying the corresponding model performance. MoleculeACE positions itself in a broader movement within the machine learning community[34−36] and aims to survey the landscape of existing AI approaches systematically for molecular property prediction.[37]

■ RESULTS AND DISCUSSION

Study Setup. *Data Sets and Activity Cliff Definition.* To ensure a comprehensive analysis of model performance, we collected and curated data on 30 macromolecular targets from ChEMBL[38] v29 (Table 1). Acknowledging known limitations of public data, we tried to rule out the presence of significant sources of error as much as possible and curated molecules according to best practices.[39−41] In particular, we checked for (a) the presence of duplicates, salts, and mixtures; (b) the consistency of structural annotations (*i.e.*, molecular validity and "sanity", charge standardization, and stereochemistry definition); and (c) the reliability of the reported experimental values in terms of annotated validity, the standard deviation of multiple entries, and the presence of outliers (see Materials and Methods section). The curated collection contains a total of 48,707 molecules (of which 35,632 were unique) and mimics typical drug discovery data sets, as it (a) includes several target families relevant for drug discovery (*e.g.*, kinases, nuclear receptors, G-protein-coupled receptors, transferases, and proteases) and (b) spans different training scenarios, from small (*e.g.*, 615 molecules for Janus Kinase 1 [JAK1]) to large (*e.g.*, 3657 molecules, dopamine D3 receptor [DRD3]) data sets (Table 1).

For each macromolecular target, activity cliffs were identified by considering pairwise structural similarities and differences in potency. We quantified molecular similarity between any pairs of molecules belonging to the same data set with three distinct approaches:

Table 1. Data Set Overview, with Response Type (Inhibition [Inhibitory Constant, K_i] or Agonism [Half-Maximal Effective Concentration, EC_{50}]), the Number of Total and Test Set Molecules (n and n_{TEST}, Respectively), along with the Percentage of Total and Test Activity Cliffs (%cliff and %cliff$_{test}$)[a]

target name	type	n (n_{TEST})	%cliff (% cliff$_{TEST}$)
androgen receptor (AR)	K_i	659 (134)	24 (23)
cannabinoid receptor 1 (CB1)	EC_{50}	1031 (208)	36 (36)
coagulation factor X (FX)	K_i	3097 (621)	44 (43)
delta opioid receptor (DOR)	K_i	2598 (521)	37 (37)
dopamine D3 receptor (D3R)	K_i	3657 (734)	39 (40)
dopamine D4 receptor (D4R)	K_i	1859 (374)	38 (38)
dopamine transporter (DAT)	K_i	1052 (213)	25 (25)
dual specificity protein kinase CLK4	K_i	731 (149)	9 (9)
farnesoid X receptor (FXR)	EC_{50}	631 (128)	39 (39)
ghrelin receptor (GHSR)	EC_{50}	682 (139)	48 (49)
glucocorticoid receptor (GR)	K_i	750 (152)	31 (31)
glycogen synthase kinase-3 β (GSK3)	K_i	856 (173)	18 (18)
histamine H1 receptor (HRH1)	K_i	973 (197)	23 (23)
histamine H3 receptor (HRH3)	K_i	2862 (574)	38 (38)
janus kinase 1 (JAK1)	K_i	615 (126)	7 (8)
janus kinase 2 (JAK2)	K_i	976 (197)	12 (13)
kappa opioid receptor (KOR) agonism	EC_{50}	955 (193)	42 (42)
kappa opioid receptor (KOR) inhibition	K_i	2602 (521)	36 (36)
mu-opioid receptor (MOR)	K_i	3142 (630)	35 (35)
orexin receptor 2 (OX2R)	K_i	1471 (297)	52 (52)
peroxisome proliferator-activated receptor alpha (PPARα)	EC_{50}	1721 (344)	41 (41)
peroxisome proliferator-activated receptor gamma (PPARγ)	EC_{50}	2349 (470)	38 (38)
peroxisome proliferator-activated receptor delta (PPARδ)	EC_{50}	1125 (225)	42 (42)
PI3-kinase p110-α subunit (PIK3CA)	K_i	960 (193)	37 (36)
serine/threonine-protein kinase PIM1	K_i	1456 (294)	33 (33)
serotonin 1a receptor (5-HT1A)	K_i	3317 (666)	35 (35)
serotonin transporter (SERT)	K_i	1704 (342)	35 (35)
sigma opioid receptor (SOR)	K_i	1328 (267)	35 (35)
thrombin (F2)	K_i	2754 (553)	36 (36)
tyrosine-protein kinase ABL1	K_i	794 (161)	32 (32)

[a] An extensive description of the data sets can be found in Supporting Table S2.

1. *Substructure similarity*. We computed the Tanimoto coefficient[42] on extended connectivity fingerprints[43] (ECFPs) to capture the presence of shared radial, atom-centered substructures among pairs of molecules. This approach captures "global" differences between molecules by considering the entire set of substructures they contain (Figure 1a).

2. *Scaffold similarity*, determined by computing ECFPs on atomic scaffolds[44] and calculating the respective Tanimoto similarity coefficient. The scaffold similarity allows identifying pairs of compounds that have minor differences in their molecular cores or differ based on their scaffold decoration (Figure 1b).

3. *Similarity of SMILES strings*, captured by the Levenshtein distance.[45] This metric detects character insertions, deletions, and translocations (Figure 1c).

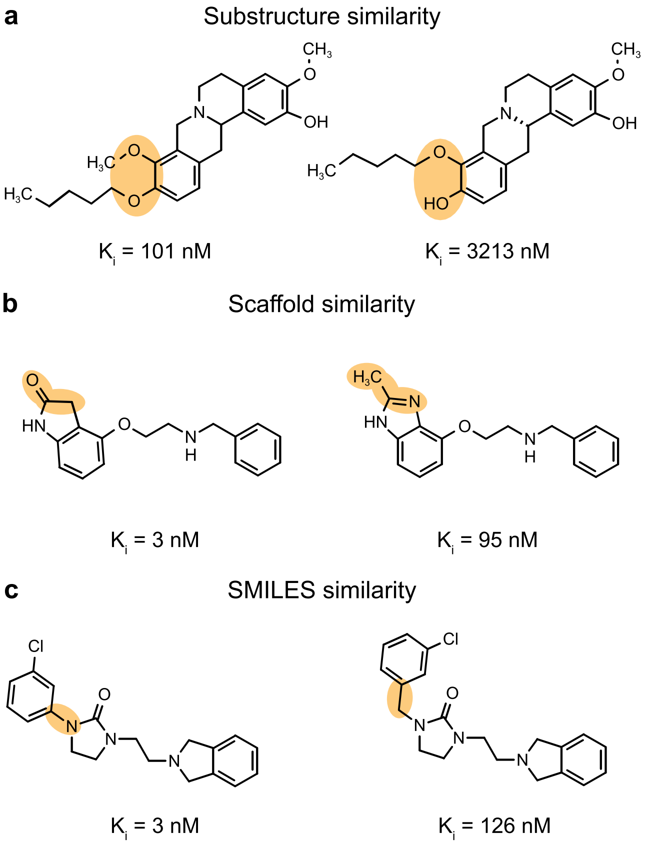


Figure 1. Selected examples of activity cliffs (on dopamine D3 receptor, D3R). (a) General substructure similarity (Tanimoto coefficient on ECFP). (b) Scaffold similarity that quantifies the similarity between molecular cores or scaffold decorations (Tanimoto coefficient on scaffold ECFP). (c) SMILES similarity that detects string insertions, deletions, and translocations (scaled Levenshtein distance).

Although there is no widely accepted definition of activity cliffs[19,46,47] and each similarity metric captures only part of the underlying "chemical reality," these three definitions were chosen to cover different types of structural differences relevant to medicinal chemistry. Moreover, they are in line with existing literature on activity cliffs (Supporting Table S3). The so-called "chirality cliffs"[28] were not considered in this study. Pairs of molecules that had a computed similarity larger than 90% with at least one of the three methods were considered as "highly similar" in structure. We specifically use a "soft" consensus to retain the unique properties the different similarity measures capture. Such pairs of compounds were then checked for their difference in reported potency. In agreement with previous studies,[21] a onefold (10×) or larger difference in bioactivity (*i.e.*, on reported K_i or EC_{50} values) was used to identify activity cliff pairs. Compounds that formed at least one activity cliff pair were labeled as "activity cliff compounds". The percentage of activity cliff compounds identified with our approach varied from 7% (JAK1) to 52% (OX2R, Table 1). Although widespread in their usage, we did not consider matched molecular pairs,[47,48] as they almost doubled the number of cliff compounds compared to our initial approach while covering 86.6% of cliff compounds identified by our approach.

Data Splitting Strategy. The nature of activity cliffs complicates data splitting into training and test sets. Having high structural similarity but vastly differing bioactivities makes it infeasible to evenly distribute activity cliff molecules across sets by both their structure and activity. Besides, multiple molecules are often involved in the same activity cliff series: across all data sets, molecules have on average 2.7 ± 0.9 activity cliff "partners" identified by our approach (Supporting Table S4). In this work, we set out to ensure (a) a proportional representation of the number of activity cliff compounds in the train and test set (to avoid an over/underestimation of their effect on the performance) and (b) preserving structural similarity between training and test molecules, as previously suggested.[49]

To this end, for each data set, molecules were clustered based on substructure similarity using spectral clustering[50] on extended connectivity fingerprints (ECFPs).[43] For each cluster, molecules were split into a training (80%) and test set (20%) by stratified random sampling using their activity cliff label (see Materials and Methods section). This method ensured that, even in the case where all activity cliff "partners" end up in the test set (9.1 ± 5.3% of activity cliff molecules on average), highly similar molecules (in terms of substructure [0.80 ± 0.03], scaffold [0.93 ± 0.02], and SMILES [0.95 ± 0.01] similarity) are still present in the training set (Supporting Table S4).

To rule out any potential bias in favor of ECFPs, we set out to compare the similarities of different molecular descriptors in the training and test sets for each macromolecular target (see Materials and Methods section). An FDR-adjusted Mann−Whitney U test ($\alpha = 0.05$) revealed no statistical difference between the distributions of the two sets across all descriptors and all targets. This indicates that the train−test similarity is also preserved when using different molecular descriptors.

Traditional Machine Learning Strategies. In this work, we considered four traditional machine learning algorithms that are commonly used for structure−activity relationship prediction (Figure 2), as follows:

1. *K-nearest neighbor* (KNN),[51] a nonparametric approach that uses the k most similar training molecules to predict the response of a new molecule (as the average of the response values). Since KNN operates directly on similarity, it is expected to struggle on activity cliff molecules and was considered a baseline.
2. *Random forest* (RF),[52] based on an ensemble of t distinct decision trees, each trained on various subsamples of the training set (built by bootstrapping). The molecule's response is predicted as average over t predictions.
3. *Gradient boosting machine* (GBM).[53] Like RF, this algorithm uses multiple decision trees. However, each next decision tree is optimized to minimize the residuals of the previous tree.
4. *Support vector regression* (SVM),[54] which maps data into higher dimensions *via* a kernel function (a radial basis function in this work) to fit an optimal hyperplane to the training data.

Each algorithm was combined with four types of molecular descriptors[29] (Figure 2), *i.e.*, human-engineered numerical features designed to capture predetermined chemical information. We explored molecular descriptors with several levels of complexity: (1) extended connectivity fingerprints[43] (ECFPs), encoding atom-centered radial substructures[43] in the form of a binary array; (2) Molecular ACCess System[55] (MACCS) keys, which encode the presence of predefined substructures in a

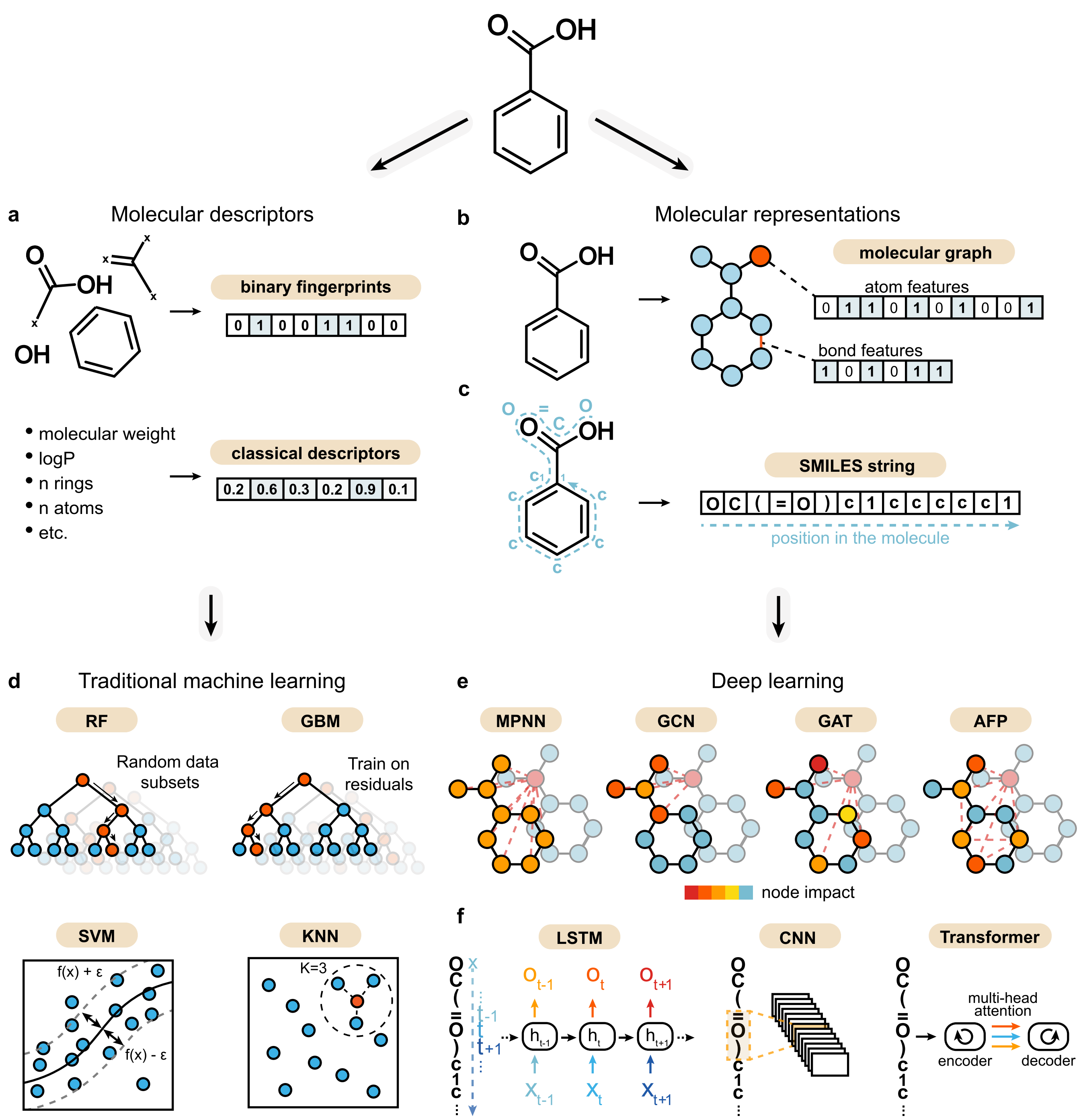


Figure 2. Machine learning strategies. (a) Simplified representation of molecular descriptors, which capture predefined molecular features. Both binary fingerprints and traditional molecular descriptors are used in this work. (b) Molecular graph, in which atoms are represented as nodes (with corresponding node features) and bonds are represented as edges (with corresponding edge features, if any). (c) SMILES strings, which capture two-dimensional information (atom and bond type and molecular topology) into a string. (d) Selected traditional machine learning algorithms that are trained on molecular descriptors: random forest (RF), gradient boosting (GBM), support vector regression (SVM), and *K*-nearest neighbor (KNN). (e) Deep learning methods. Four graph neural networks that can learn from molecular graphs were used: message passing neural network (MPNN), graph convolutional network (GCN), graph attention network (GAT), and attentive fingerprint (AFP). Node colors indicate the impact of other nodes during feature aggregation (indicated by dashed lines). Three SMILES-based methods that can learn from sequential data were used: long short-term memory networks (LSTM), one-dimensional (1D) convolutional neural networks (CNN), and transformers.

binary array; (3) weighted holistic invariant molecular (WHIM) descriptors,[56] capturing three-dimensional molecular size, shape, symmetry, and atom distribution; and (4) 11 physicochemical properties relevant for drug-likeness[57] (see Materials and Methods section), used as a baseline. This selection is not comprehensive (owing to the high number of existing molecular descriptors[29]), but we believe that it constitutes a good overview of different types of descriptors used in the medicinal chemistry domain.

Graph-Based Deep Learning. Molecular graphs are a mathematical representation of molecular topology, with nodes and edges representing atoms and chemical bonds,

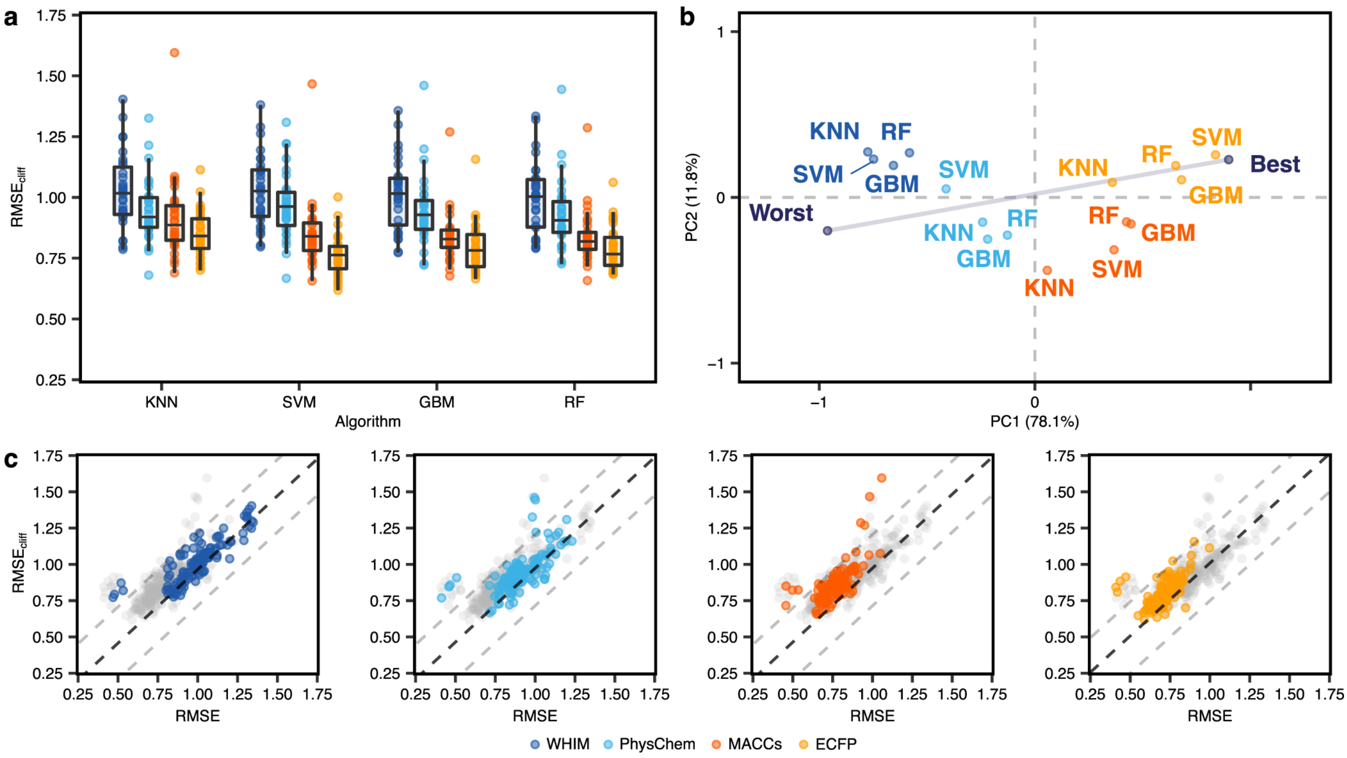


Figure 3. Performance of traditional machine learning methods. (a) RMSE on activity cliff compounds using different machine learning algorithms and molecular descriptors (indicated by colors). (b) Global ranking of all methods using PCA (first two principal components, PC1 and PC2), scaled between best and worst performance. Every point captures a different combination of the machine learning method and the descriptor it relied on and is obtained by considering the corresponding $\mathrm{RMSE}_{\mathrm{cliff}}$ on all data sets. "Worst" and "Best" indicated the worst and best performance obtained across all data sets, respectively. Percentages represent the variance explained by each principal component. (c) Comparison between the error on activity cliff compounds ($\mathrm{RMSE}_{\mathrm{cliff}}$) and the error on all compounds (RMSE) for all methods. Black dashed lines indicate RMSE = $\mathrm{RMSE}_{\mathrm{cliff}}$, while gray dashed lines indicate a difference of ±0.5 log units between $\mathrm{RMSE}_{\mathrm{cliff}}$ and RMSE.

respectively (Figure 2b). Neural networks that can learn directly from graphs are becoming increasingly popular for molecular property prediction.[14,58−61] In this work, we explored four neural network architectures that can directly operate on molecular graphs (Figure 2d), as follows:

1. *Message passing neural network* (MPNN).[62] For every node in the molecular graph, information (the "message") from neighboring nodes is aggregated by transforming it with a learnable function.
2. *Graph attention network* (GAT).[63] Instead of a message passed across edges, this algorithm also learns attention coefficients that determine the importance of features.
3. *Graph convolutional network* (GCN),[64] which aggregates information from neighboring nodes using a fixed convolution.
4. *Attentive fingerprint* (AFP),[59] which uses attention mechanisms at both the atom and molecule level, allowing it to better capture subtle substructure patterns.

SMILES-Based Deep Learning Methods. As an additional representation, we employed the simplified molecular input line entry system (SMILES) strings,[65] which have recently become particularly popular for *de novo* molecular design,[9−11] and captured two-dimensional molecular information in a textual format (Figure 2c). Here, we explored three types of neural networks suitable to learn from SMILES strings:

1. *Convolutional neural network* (CNN).[66] This neural network architecture uses a learnable convolutional filter to aggregate information from neighboring positions in a SMILES string with a sliding window approach.
2. *Long short-term memory* (LSTM)[67] networks. LSTM—a type of recurrent neural network—can learn from string sequences by keeping track of long-range dependencies. As in a previous study,[68] LSTM models were pretrained on SMILES obtained by merging all training sets with no repetitions (36,281 molecules) using next-character prediction before applying transfer learning for bioactivity prediction.
3. *Transformer model.* Transformers process the whole sequence at once in a graphlike manner using positional embedding to capture positional information.[69] Transformers implement the so-called attention,[69] which enables the model to learn which portions of the sequence are more relevant for a given task. The pretrained ChemBERTa[70] architecture (10M compounds) was used in combination with transfer learning for bioactivity prediction.

In agreement with previous studies[66,71,72] and thanks to the nonunivocal character of SMILES strings, we used tenfold SMILES augmentation to artificially increase the number of training samples for all approaches.

Model Performance with Activity Cliffs. *Traditional Machine Learning Methods.* First, we evaluated the ability of "traditional" machine learning approaches to predict bioactivity (expressed as pEC_{50} or $\mathrm{p}K_{\mathrm{i}}$) in the presence of activity cliffs. The performance was quantified using the root-mean-square error on test set molecules (RMSE—the lower, the better; eq

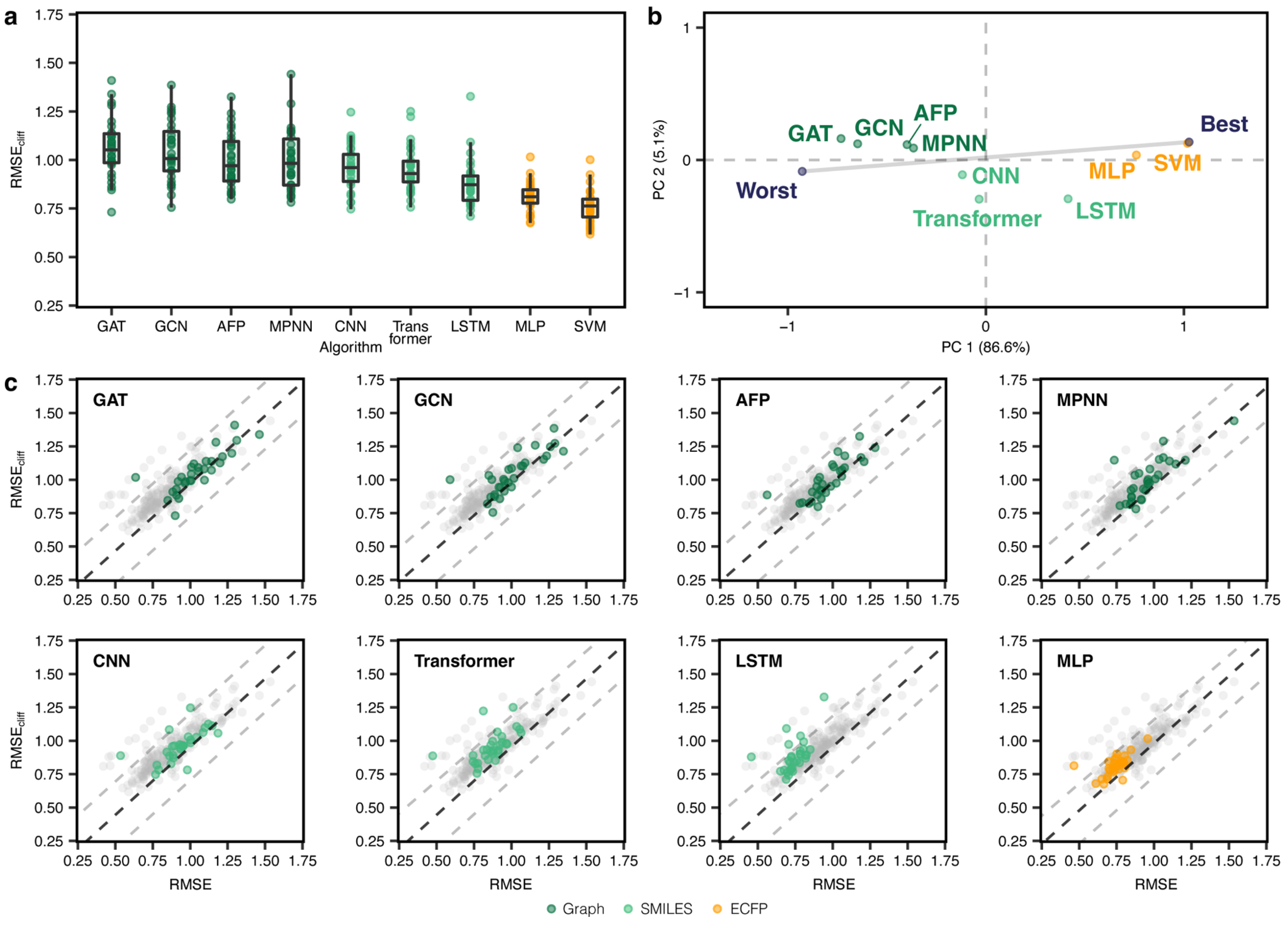


Figure 4. Performance of deep learning methods. (a) RMSE on activity cliff compounds on different deep learning strategies. SVM is reported as a reference. (b) Global ranking of all methods using PCA (first two principal components, PC1 and PC2), scaled between best and worst performance. Every point captures the performance of a different machine learning approach obtained by considering the corresponding RMSE$_{cliff}$ on all data sets. "Worst" and "Best" indicated the worst and best performance obtained across all data sets, respectively. Percentages indicate the explained variance by each principal component. (c) Prediction error on activity cliff compounds (RMSE$_{cliff}$) compared to all compounds (RMSE) for all methods.

1) and activity cliff molecules in the test set (RMSE$_{cliff}$—the lower, the better; eq 2). Overall, large differences in predictive performance on activity cliff compounds can be observed among data sets, with RMSE$_{cliff}$ values ranging from 0.62 to 1.60 log units (Figure 3a). This effect was also observed in the overall performance of test set molecules, with RMSE values ranging from 0.41 to 1.35 log units (Supporting Figure S1a), in line with previous works.[73−75] Differences in performance relate mostly to the chosen molecular descriptor rather than the machine learning algorithm ($p < 0.05$, Wilcoxon rank-sum test with Benjamini−Hochberg correction, Supporting Figure S4), with ECFPs yielding the lowest average prediction error on average. Nonbinary descriptors (WHIM and physicochemical properties) performed considerably worse overall than binary fingerprints (ECFPs and MACCS), with a higher variation among data sets.

To provide a global assessment of methods across the analyzed data sets, we performed a principal component analysis (PCA) on the obtained RMSE$_{cliff}$ values (Figure 3b and Supporting Figure S2a). PCA is a multivariate analysis technique used for data visualization and dimensionality reduction, which linearly combines the original variables into new orthogonal variables (principal components), sorted by the variance they explain. To enhance the interpretability, rows capturing the best and worst RMSE$_{cliff}$ for each data set were added to stretch the PCA results along the direction of the best and worst results as in previous studies.[76,77] This PCA allows considering each method on a data set-basis and to account for the presence of targets more difficult to "model". The closer a method is to the "best" point, the better its overall performance. The higher the orthogonal deviation from the best-worst line, the higher the variability of a method's performance based on the data set. For instance, methods based on MACCS fingerprints show a higher dependency on the chosen targets than those based on ECFPs. KNN methods show the highest dependency on the chosen target overall. Our analysis confirms the higher impact of molecular descriptors than the chosen machine learning algorithm on the model performance.[78,79] SVM coupled with ECFPs resulted in the best method on activity cliffs on average, in agreement with a previous study.[80] However, no statistical difference was found between SVM, GBM, or RF coupled with ECFPs (Wilcoxon rank-sum test, Supporting Figure S4). In the case of our results, however, the superior performance of ECFPs is somewhat surprising, given that they were used for the definition of

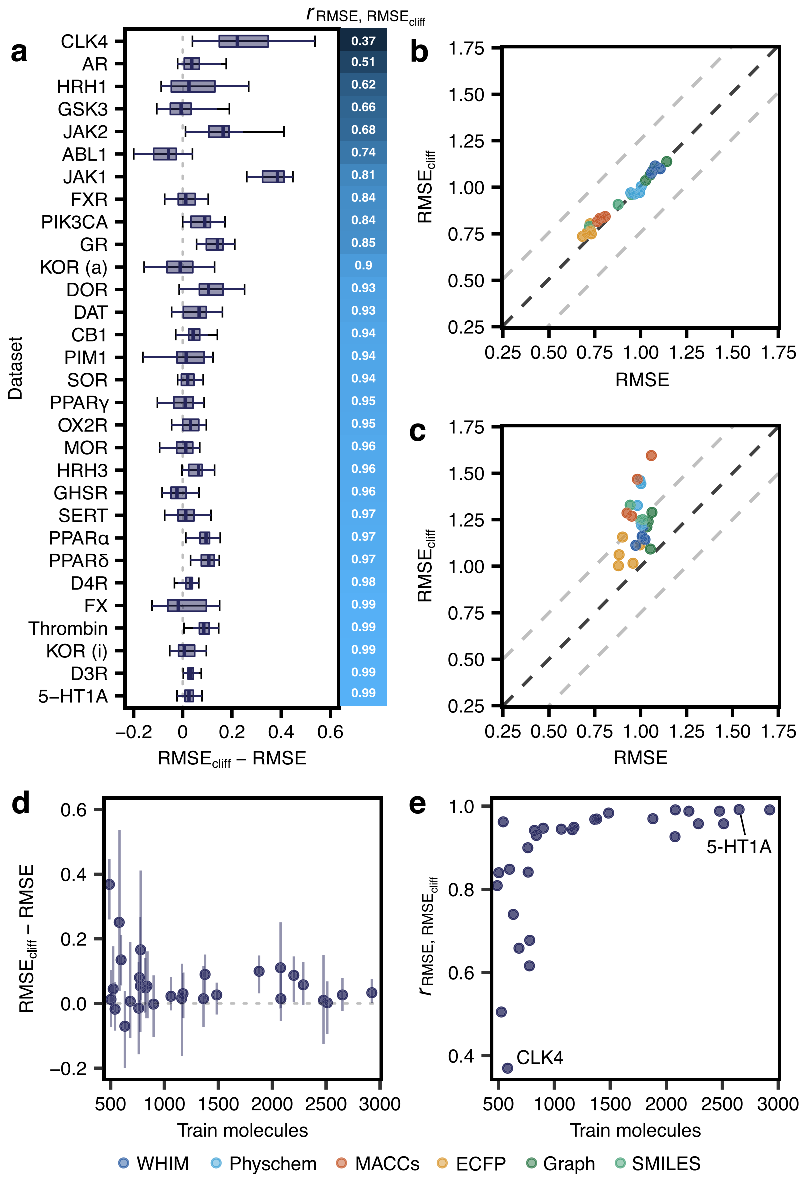


Figure 5. Comparing overall model performance and performance on activity cliff compounds. (a) Method-wide differences between overall RMSE and RMSE$_\text{cliff}$ for all targets ordered by Pearson correlation (r) between RMSE and RMSE$_\text{cliff}$. Error bars indicate the lowest and highest RMSE$_\text{cliff}$. (b) Comparison between RMSE and RMSE$_\text{cliff}$ of all methods on 5-HT1A. (c) Comparison between RMSE and RMSE$_\text{cliff}$ of all methods on CLK4. (d) Effect of the number of training molecules on the difference between RMSE and RMSE$_\text{cliff}$. Error bars indicate the lowest and highest RMSE$_\text{cliff}$. (e) Relationship between the number of training molecules and the Pearson correlation (r) of RMSE and RMSE$_\text{cliff}$.

activity cliffs (criteria 1 and 2, Figure 1a), which was expected to introduce an unfavorable bias.

To further investigate the relevance of considering activity cliffs for model assessment, we compared RMSE$_\text{cliff}$ with the overall error on the test set molecules (Figure 3c and Supporting Figure S3). As expected, activity cliff compounds tend to yield higher prediction errors, regardless of the considered approach.[81] Although in most of the cases RMSE and RMSE$_\text{cliff}$ are highly correlated (r = 0.81 on average), the model performance on activity cliff compounds might be overestimated when considering RMSE alone, up to 0.54 log units. For instance, SVM coupled with ECFP descriptors—resulting in the best performance on average—ranged greatly in its ability to handle activity cliffs. While the mean difference between RMSE and RMSE$_\text{cliff}$ for this method was only 0.094 log units, large differences were observed in certain data sets (*e.g.*, up to 0.39 log units for the JAK1 receptor). This underscores that strategies with a low overall prediction error might not necessarily be the best ones at handling activity cliffs, thereby hampering their potential for prospective applications.

Deep Learning Methods. In contrast to traditional machine learning algorithms, neural networks allow bypassing human-constructed molecular descriptors and can learn directly from "unstructured" representations of chemical structures. Deep learning approaches trained on either graphs or SMILES strings were compared with (a) a multilayer perceptron (MLP) based on ECFPs and (b) the best-performing traditional machine learning method (SVM with ECFP fingerprints), both serving as a reference point (Figure 4a).

Transfer learning[82]—applying a models' previously learned knowledge to a new, related problem by further training—was applied to the LSTM and transformer models in agreement with previous studies.[68,70,83,84] In a preliminary analysis, we explored transfer learning approaches for graph neural networks using self-supervision (context prediction,[85] infomax,[86] edge prediction,[87] and masking[85]). Since, in line with a recent study,[88] no approach yielded a notable increase in predictive performance, we did not consider transfer learning further for graph neural networks. When comparing the performance of all tested deep learning methods, we found large differences in predictive performance across data sets—like with traditional machine learning approaches—with $\text{RMSE}_{\text{cliff}}$ values ranging from 0.68 to 1.44 log units (Figure 4a). Among the graph-based neural networks, MPNN models resulted in the lowest error on activity cliff compounds on average, although no differences were statistically significant (Wilcoxon rank-sum test with Benjamini−Hochberg correction, Supporting Figure S4). SMILES-based methods outperformed graph-based methods on average, with LSTM models outperforming all other deep learning methods, including the SMILES-based CNN and transformer models. For CNNs, we did not implement any transfer learning strategy, which could explain their poor(er) performance compared to the other SMILES-based methods. Notably, despite transformers being pretrained on a larger corpus of SMILES strings (10M compounds[70]), they did not perform better than LSTMs, which were pretrained on 36,281 molecules only.

When inspecting the PCA performed on the obtained $\text{RMSE}_{\text{cliff}}$ values for each target (Figure 4b), the multilayer perceptron coupled with ECFPs outperformed all other neural networks based on SMILES or graphs. This is surprising to a certain extent, considering that ECFPs and SMILES are constructed from a molecular graph. This aspect further underscores a current gap in learning efficient features from "raw" molecular representations in the small-data regimes typical of drug discovery. Compared to most traditional machine learning approaches, deep neural networks seem to fall short at picking up subtle structural differences (and the corresponding property change) that give rise to activity cliffs. Similar results were obtained when comparing graph networks for (a) feature attribution with activity cliffs,[89] and (b) bioactivity prediction.[30] A recent analysis on physicochemical-property cliffs highlights an opposite trend, with deep learning methods performing better than simpler machine learning approaches[90]—potentially due to the higher number of training samples (approx. 20,000 molecules).

Interestingly, no deep learning method was stable across data sets, as shown by the large deviation from the worst-best line (Figure 4b and Supporting Figure S2b). This highlights the need to evaluate the usage of such methods on a case-by-case basis.

Failure Modes of Machine Learning on Activity Cliffs. The systematic training and assessment of 720 machine learning models allowed us to investigate the potential "failure modes" of machine learning approaches on activity cliffs. All methods tend to struggle in the presence of activity cliffs (Figures 3 and 4). Our first analysis addressed the variation of $\text{RMSE}_{\text{cliff}}$ across methods and data sets in search of causes of poor performance. Although small-data regimes are known to affect the performance of machine and deep learning methods, no correlation was found between the number of molecules in the training set and the prediction error on activity cliffs (Supporting Figure S5). Furthermore, no relationship between the percentage of activity cliff compounds in the data and model performance was found, except for differences between RMSE and $\text{RMSE}_{\text{cliff}}$. This relates to the fact that the higher the percentage of activity cliffs, the more the $\text{RMSE}_{\text{cliff}}$ values (computed on a subset of molecules, eq 2) will approach RMSE values (Supporting Figure S6). At the same time, the drug target family did not seem to affect $\text{RMSE}_{\text{cliff}}$ either (Supporting Figure S7), further highlighting the difficulties in forecasting the performance of machine learning on activity cliffs.

We then compared the overall prediction error (RMSE on test set molecules) with the performance on activity cliffs ($\text{RMSE}_{\text{cliff}}$ on test set molecules). While RMSE and $\text{RMSE}_{\text{cliff}}$ tend to correlate to a high degree ($r > 0.70$ for 25 data sets out of 30, Figure 5a), we observed large case-by-case variations. In most cases, the difference between $\text{RMSE}_{\text{cliff}}$ and RMSE is similar among methods (Figure 5a,b). This implies that, when choosing a method for its overall error on test set molecules, the performance on activity cliff compounds will be implicitly accounted for. However, for some targets (*e.g.*, CLK4), methods with comparable RMSE scores can exhibit large differences in $\text{RMSE}_{\text{cliff}}$ scores (Figure 5c). This indicates that, in these specific cases, choosing a model based on only RMSE might lead to poor prospective performance, *e.g.*, for hit-to-lead optimization or virtual screening in the presence of congeneric compounds (Supporting Table S1). These "islands" of poor performance on activity cliffs were observed across the whole spectrum of machine learning strategies, independently of the reported average performance.

To better elucidate the "drivers of failure" on activity cliffs, we investigated the effect of the training set size on (a) the difference between predictivity on the entire test set and on activity cliffs only ($\text{RMSE}_{\text{cliff}} - \text{RMSE}$) and (b) the correlation between the overall performance (RMSE) and the performance on activity cliffs ($\text{RMSE}_{\text{cliff}}$). The absolute difference between RMSE and $\text{RMSE}_{\text{cliff}}$ does not correlate with the number of training molecules ($r = -0.15$, Figure 5d). However, the number of training molecules is an important factor in determining the correlations between RMSE and $\text{RMSE}_{\text{cliff}}$ (Figure 5e). Data sets containing a sufficient number of training molecules (*e.g.*, larger than 1000) showed a high correlation between RMSE and $\text{RMSE}_{\text{cliff}}$ ($r > 0.80$). In other words, if the number of training molecules increases, the "relative difficulty" of predicting bioactivity on activity cliff molecules decreases. This implies that, with a sufficient number of training molecules, optimizing RMSE alone will improve $\text{RMSE}_{\text{cliff}}$, too. However, the problem of determining the targets on which $\text{RMSE}_{\text{cliff}}$ will be suboptimal remains, especially in small-data regimes, further underscoring the relevance of implementing activity-cliff-related evaluation approaches. Moreover, these results corroborate the need to develop more efficient machine and deep learning models for low-data regimes.

Bringing It All Together: The MoleculeACE Benchmark and Future Applications. Our results and systematic analyses expose current limitations of molecular machine learning and motivate the use of dedicated metrics and tools for assessing the model performance on activity cliffs, especially in low-data regimes. Hence, we collected the modeling and assessment strategies of this study into a dedicated, "activity-cliff-centered" benchmark tool, called MoleculeACE (available at: https://github.com/molML/MoleculeACE). All data sets and scripts to replicate this study can be found here as well. MoleculeACE integrates standardized data processing for molecular bio-

https://doi.org/10.1021/acs.jcim.2c01073
J. Chem. Inf. Model. 2022, 62, 5938−5951

activity data, a comprehensive approach to quantifying activity cliffs, and the tailored performance evaluation strategies presented in this work. Thanks to its modular character, MoleculeACE will allow researchers to

1. *systematically benchmark a model's performance* on activity cliffs compounds (*e.g.*, using different machine learning approaches or including additional molecular descriptors), in comparison with well-established machine and deep learning methods;
2. *evaluate the deck of chosen models on a new data set* not included in our benchmark, thanks to the data collection and curation pipeline; and
3. *further expand the definition of activity cliffs*[91−93] based on specific use cases.[19] It is possible to use custom thresholds for potency differences and structural similarity (*e.g.*, matched molecular pairs, which are already supported) in determining cliff compounds. As this work relies on public bioactivity data, which might be affected by undetectable experimental noise[81,94] (despite the best data curation efforts), we hope in the future to also see applications of MoleculeACE on more homogeneous data, *e.g.*, in terms of use *in vitro* assays and assay conditions.

We envision that MoleculeACE, along with the results of this benchmark study, will incentivize machine learning researchers to consider the crucial topic of activity cliffs in model evaluation and development pipelines. We envision that MoleculeACE will serve as a platform for the wider community to develop models that can more accurately capture complex structure−activity landscapes and ultimately boost the capabilities of machine learning for molecule discovery.

CONCLUSIONS AND OUTLOOK

While machine learning is increasingly often employed for early drug discovery, the topic of activity cliffs has received only limited attention from the scientific community. As shown by our results, not only do machine learning strategies struggle with activity cliffs compared to their overall performance but also deep learning methods are particularly challenged by the presence of such compounds. Approaches based on human-engineered molecular descriptors resulted in outperforming deep learning based on graphs or SMILES, with no machine learning strategy being consistently better at handling activity cliffs compared to their absolute performance. Our results corroborate previous evidence showing that deep learning methods do not necessarily hold up against simpler machine learning methods (yet) for drug discovery purposes.[15−17] Although our analysis does not allow us to identify mechanistic causes of the performance gap with activity cliffs, we speculate that current molecular representations and corresponding representation learning algorithms might not capture complex structure−activity information well enough.[95,96] We envision the development of deep learning strategies that are (a) more efficient in low-data scenarios (*e.g.*, self-supervised learning[97]) and (b) better-suited to capture structure−activity "discontinuities" to be key for future prospective applications. Structure-based deep learning approaches[28,98,99] (considering the structure of the macromolecular target in addition to ligand information) might be key to filling current performance gaps due to activity cliffs. However, to date, there is no consensus on the benefit of including structural information in machine learning for bioactivity prediction,[100] potentially due to undesirable bias in existing databases.[100−102]

In the framework of our study design, the model's performance on activity cliff compounds resulted in being highly data set-dependent, especially for deep learning methods in low-data scenarios. Although the overall prediction error often approximates the performance on activity cliffs, "islands" of poor performance on activity cliffs exist when different strategies are compared on the same data set. These results highlight the importance of evaluating machine learning models for their performance on activity cliffs, especially when prospective applications are envisioned (*e.g.*, virtual screening).[103]

To facilitate such an "activity cliff-centered" model evaluation and development, we developed MoleculeACE. By estimating a model's performance in the presence of activity cliffs alongside regular performance, MoleculeACE has the goal of incentivizing researchers in molecular machine learning to consider the long-standing issue of activity cliffs fully. Models that can accurately predict the effects of subtle structural changes on molecular properties will ultimately give rise to more effective hit-to-lead optimization and the identification of activity cliffs during lead optimization. We envision these improvements as key to propelling the potential of deep learning in drug discovery and beyond.

MATERIALS AND METHODS

Data Curation. *Data Collection and Preparation.* For each macromolecular target, compound bioactivity values were collected from ChEMBL[101] v29 *via* the "ChEMBL webresource" client (*Homo sapiens*). Molecules in the form of canonical SMILES strings were sanitized using RDKit[104] v. 2020.09.5[104] with default settings and neutralized if charged. Compounds with failed sanitization, annotated in the form of salts, and/or with doubtful data validity (as in the "data_validity_comment" entry of ChEMBL) were removed (4.74% on average). For each unique SMILES string, experimental bioactivity data (*i.e.*, K_i or EC_{50} values [nM]) were collected. Dixon's Q test[105] was used to detect the presence of outliers among multiple annotations of a given molecule ($\alpha = 0.05$, 0.78% of molecules on average). The mean K_i or EC_{50} value for each molecule was computed and subsequently converted into pEC_{50}/pK_i values (as the negative logarithm of molar concentrations). If the standard deviation of the multiple annotations used to compute the average was above 1 log unit, the corresponding molecule was removed (4.33% on average). To rule out errors due to inconsistent annotation of stereochemistry, pairs of compounds having different canonical SMILES but identical ECFPs were removed (9.74% on average).

Molecular Descriptors' Calculation. Molecular descriptors were computed from canonicalized SMILES strings using RDkit v. 2020.09.5.[104] (a) Extended connectivity fingerprints (ECFPs)[43] were computed with a length of 1024 bits and a radius of 2 bonds. (b) MACCS keys,[55] with a length of 166, were computed with default settings. (c) Weighted holistic invariant molecular (WHIM) descriptors[103] (114 descriptors) were computed on the minimum energy conformers generated with experimental-torsion knowledge distance geometry[106] and MMFF94[107] force field optimization. (d) "Physicochemical descriptors" included 11 properties of drug-likeness, *i.e.*, molecular weight; predicted octanol−water partitioning coefficient;[108] molar refractivity; topological polar surface area; formal charge; and the number of hydrogen bond donors, hydrogen bond acceptors, rotatable bonds, atoms, rings, and heavy atoms. Real-valued descriptors were standardized by Gaussian normalization using the training data mean and standard deviation values.

Detection of Activity Cliffs. Pairs of structurally similar molecules were detected with three approaches: (a) *substructure similarity*, computed via the Tanimoto coefficient on ECFP; (b) *scaffold similarity*, calculated on the ECFP of molecular graph frameworks[44] (Tanimoto coefficient); and (c) (canonical) *SMILES similarity*, computed using the Levenshtein distance[106,109] (scaled and subsequently converted into "1-distance"). Pairs of compounds having a computed similarity equal to or larger than 0.9 according to at least one of these metrics were checked for the fold difference in their respective bioactivity (in nM units). Pairs of highly similar compounds showing more than tenfold difference in their respective bioactivity were considered activity cliffs.

Train/Test Splitting. For each target, molecules were clustered by their molecular structure (described as ECFP) into five clusters using spectral clustering[107] implemented with sklearn v. 1.0.2[110] (using a Gaussian kernel and a precomputed affinity matrix of Tanimoto distances). For each cluster, 80% of molecules were assigned to the training data and 20% were assigned to the test data by stratified splitting (using their belonging to at least one activity cliff pair ["yes"/"no"] as a label).

Descriptor Similarity between Training and Test Sets. Similarity among molecular descriptors in the training set was calculated as the mean distance of each molecule in the training set to its five nearest neighbors in the training set. The similarity between molecular descriptors of each molecule in the test set was calculated for the five nearest neighbors in the train set. Graph representations were not considered, as computing graph distances is nontrivial and ECFPs are directly related to molecular graphs. A Mann−Whitney *U* test, corrected for a false discovery rate of 0.05, was performed using SciPy v. 1.8.1.[111]

Molecular Graph Featurization. For all methods, atom features were encoded as follows. (a) One-hot-encoded properties included atom type, orbital hybridization, atomic vertex degree, aromaticity, and ring membership. (b) Numerically encoded properties included atomic weight, partial charge (Gasteiger−Marsili[112]), number of valence electrons, and number of bound hydrogens. The atomic weight and partial charge were scale-transformed *via* a sigmoidal function. For MPNN and AFP architectures, bond features were included, *i.e.*, with bond type and conjugation (one-hot-encoded).

Model Implementation. *Hyperparameter Optimization.* Hyperparameter optimization was performed with Bayesian optimization using a Gaussian process (method-based specifics are mentioned below). For all models, a maximum of 50 hyperparameter combinations were evaluated using fivefold cross validation.

Traditional Machine Learning Algorithms. KNN, SVM, GBM, and RF regression models were implemented using sklearn v. 1.0.2.[110] For each approach, the model hyperparameters were optimized as follows: (a) KNN, optimization of the number of neighbors (k), k = [3, 5, 11, 21]; (b) SVM, optimization of the kernel coefficient (γ) and regularization parameter (C), γ = [1×10^{-6}, 1×10^{-5}, 1×10^{-4}, 1×10^{-3}, 1×10^{-2}, or 1×10^{-1}] and C = [1, 10, 100, 1000, 10,000]; (c) GBM, optimization of the number of boosting stages (n_b) and maximal model depth (m_d), n_b = [100, 200, 400] and m_d = [5, 6, 7]; and (d) RF, number of decision trees (t), t = [100, 250, 500, 1000].

Graph Neural Networks. All regression models were implemented using the PyTorch Geometric package v. 2.0.4.[113] In MPNN, GCN, and GAT models, global pooling was implemented with a graph multiset transformer[114] using eight attention heads, followed by a fully connected prediction head. For all models, we optimized the learning rate (lr), lr = [5×10^{-4}, 5×10^{-5}, or 5×10^{-6}]. The following hyperparameters were optimized:

(a) GCN, hidden atom features (h_a), number of convolutional layers (n_c), hidden multiset transformer nodes (h_t), hidden predictor features (h_p), h_a = [32, 64, 128, 256, 512], n_c = [1, 2, 3, 4, 5], h_t = [64, 128, 256, 512], h_p = [128, 256, 512];

(b) GAT, the hyperparameter search space used for GCN models and the use of GATv1[112] or GATv2[115] convolutions;

(c) MPNN, hidden atom features (h_a), hidden edge features (h_e), number of message passing steps (s_m), hidden multiset transformer nodes (h_t), hidden predictor features (h_p), h_a = [32, 64, 128, 256], h_e = [32, 64, 128, 256], s_m = [1, 2, 3, 4, 5], h = [64, 128, 256, 512], h_p = [128, 256, 512]; and

(d) AFP, number of attentive layers (n_a), timesteps (n_t), number of hidden predictor features (h_p), n_a = [1, 2, 3, 4, 5], n_t = [1, 2, 3, 4, 5], h_p = [32, 64, 128, 256].

All models were trained for 300 epochs using early stopping with a patience of ten epochs.

Feed-Forward Neural Network. A multilayer perceptron was implemented using Pytorch v. 1.11.0.[116] It was optimized for (a) the learning rate (lr = [5×10^{-4}, 5×10^{-5} 64, 5×10^{-6}]), (b) the number of hidden features (n_h = [256, 512, 1024]), and (c) the number of layers (n_l) = [1, 2, 3, 4, 5]. Models were trained for 500 epochs using early stopping with a patience of 10 epochs.

SMILES-Based Models. SMILES strings were encoded as one-hot vectors. SMILES strings longer than 200 characters were truncated (0.71% on average). Tenfold data augmentation was applied to all SMILES-based methods using a maximum of nine extra noncanonical SMILES strings for every SMILES string in the data set. Noncanonical SMILES strings were generated using RDKit.[104]

(a) LSTM models were pretrained on SMILES obtained by merging all training sets with no repetitions (36,281 molecules) using next-character prediction as in a recent study.[68] The network was composed of four layers comprising 5,820,515 parameters (layer 1, batch normalization; layer 2, LSTM with 1024 units; layer 3, LSTM with 256 units; layer 4, batch normalization). We used the Adam optimizer with a learning rate of 10^{-4} for 100 epochs. Regression models were then obtained by transfer learning (with weight freezing for layer no. 2) for 100 epochs with a regression head.

(b) 1D CNNs were adapted from a recent study.[66] We used a single 1D convolutional layer with a step size equal to 1, followed by a fully connected layer, with training for 500 epochs. It was optimized for the learning rate (lr), the number of hidden features in the fully connected layer (n_h), and convolution kernel size (n_k), lr = [5 ×

10^{-4}, 5×10^{-5},[64] 5×10^{-6}], n_h = [128, 256, 512, 1024], n_k = [4, 8, 10].

(c) Transformer models and the corresponding SMILES tokenization were based on the ChemBERTa[114] architecture. We used the pretrained ChemBERTa model weights based on 10M compounds from PubChem.[117] We fine-tuned the model by freezing its weights and replacing the final pooling layer with a regression head with one fully connected layer and trained for 100 epochs. We used the Adam optimizer with a learning rate of 5×10^{-4}. For all methods, we used early stopping with a patience of ten epochs.

Performance Evaluation. The overall model performance was quantified via the root-mean-square error (RMSE) computed on the bioactivity values (*i.e.*, pK_i or pEC$_{50}$), as follows (eq 1)

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n}} \quad (1)$$

where $\hat{y}_i$ is the predicted bioactivity of the *i*th compound, y_i is the corresponding experimental value, and *n* represents the number of considered molecules.

The performance on activity cliffs compounds was quantified by computing the root-mean-square error (RMSE$_{\mathrm{cliff}}$) only on compounds that belonged to at least one activity cliff pair, as follows (eq 2)

$$\mathrm{RMSE}_{\mathrm{cliff}} = \sqrt{\frac{\sum_{j=1}^{n_c} (\hat{y}_j - y_j)^2}{n_c}} \quad (2)$$

where $\hat{y}_j$ is the predicted bioactivity of the *j*th activity cliff compound, y_j is the corresponding experimental value, and n_c represents the total number of activity cliff compounds considered. R^2 and Q^2 metrics, or normalized RMSE values, were not considered to avoid the introduction of undesired biases related to the different range of the training/test set responses across data sets.[118,119]

■ ASSOCIATED CONTENT

Data Availability Statement

The MoleculeACE benchmark tool, together with the Python code to replicate and extend our study, is freely available on GitHub at the following URL: https://github.com/molML/MoleculeACE.The curated data sets are available at the following URL: https://github.com/molML/MoleculeACE/tree/main/MoleculeACE/Data/benchmark_data.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.2c01073.

Presence of highly similar compounds in commercially available libraries (Table S1); data set overview (Table S2); activity cliff definitions across different published studies (Table S3); training/test set analysis (Table S4); overall performance of machine learning methods on all targets (Figure S1); PCA loadings of all methods (Figure S2); relative prediction error of activity cliff compounds (Figure S3); statistical differences between the RMSE$_{\mathrm{cliff}}$ values obtained by different machine learning strategies (Figure S4); relationship between the number of training molecules on RMSE$_{\mathrm{cliff}}$ (Figure S5); relationship between the fraction of activity cliff compounds and model performance (Figure S6); relationship between drug target classes and RMSE$_{\mathrm{cliff}}$ (Figure S7) (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Francesca Grisoni − *Institute for Complex Molecular Systems and Dept. Biomedical Engineering, Eindhoven University of Technology, 5612AZ Eindhoven, The Netherlands; Centre for Living Technologies, Alliance TU/e, WUR, UU, UMC Utrecht, 3584CB Utrecht, The Netherlands;* orcid.org/0000-0001-8552-6615; Email: f.grisoni@tue.nl

Authors

Derek van Tilborg − *Institute for Complex Molecular Systems and Dept. Biomedical Engineering, Eindhoven University of Technology, 5612AZ Eindhoven, The Netherlands; Centre for Living Technologies, Alliance TU/e, WUR, UU, UMC Utrecht, 3584CB Utrecht, The Netherlands*

Alisa Alenicheva − *JetBrains Research, 194100 Saint Petersburg, Russia*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.2c01073

Author Contributions

Conceptualization: F.G. and D.v.T. Data curation: D.v.T. and F.G. Formal analysis: D.v.T. and A.A. Methodology: D.v.T., A.A., and F.G. Software: D.v.T. and A.A. Writing—original draft: D.v.T. Writing—review and editing: all authors. All authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank the group of Prof. Luc Brunsveld for valuable discussion and feedback, Dr. Jiménez-Luna and Luke Rossen for comments on the code, and Rıza Özçelik for feedback on the manuscript. F.G. acknowledges support from the Irène Curie Fellowship and the Centre for Living Technologies.

■ ABBREVIATIONS

AFP, attentive fingerprint; CNN, convolutional neural network; ECFP, extended connectivity fingerprints; GAT, graph attention network; GBM, gradient boosting machine; GCN, graph convolutional network; KNN, *K*-nearest neighbor; LSTM, long short-term memory network; MACCS, Molecular ACCess System; MLP, multilayer perceptron; MPNN, message passing neural network; RF, random forest; RMSE, root-mean-square error; SMILES, simplified molecular input line entry system; SVM, support vector machine; WHIM, weighted holistic invariant molecular

■ REFERENCES

(1) de Almeida, A. F.; de Moreira, R.; Rodrigues, T. Synthetic Organic Chemistry Driven by Artificial Intelligence. *Nat. Rev. Chem.* **2019**, *3*, 589−604.

(2) Baskin, I. I.; Winkler, D.; Tetko, I. V. A Renaissance of Neural Networks in Drug Discovery. *Expert Opin. Drug Discovery* **2016**, *11*, 785−795.

(3) Chen, H.; Engkvist, O.; Wang, Y.; Olivecrona, M.; Blaschke, T. The Rise of Deep Learning in Drug Discovery. *Drug Discovery Today* **2018**, *23*, 1241−1250.

(4) Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning Chemical Syntheses with Deep Neural Networks and Symbolic AI. *Nature* **2018**, *555*, 604−610.

(5) Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; Lee, A. A. Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Cent. Sci.* **2019**, *5*, 1572−1583.

(6) Coley, C. W.; Green, W. H.; Jensen, K. F. Machine Learning in Computer-Aided Synthesis Planning. *Acc. Chem. Res.* **2018**, *51*, 1281−1289.

(7) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596*, 583−589.

(8) Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G. R.; Wang, J.; Cong, Q.; Kinch, L. N.; Schaeffer, R. D.; Millán, C.; Park, H.; Adams, C.; Glassman, C. R.; DeGiovanni, A.; Pereira, J. H.; Rodrigues, A. V.; van Dijk, A. A.; Ebrecht, A. C.; Opperman, D. J.; Sagmeister, T.; Buhlheller, C.; Pavkov-Keller, T.; Rathinaswamy, M. K.; Dalwadi, U.; Yip, C. K.; Burke, J. E.; Garcia, K. C.; Grishin, N. V.; Adams, P. D.; Read, R. J.; Baker, D. Accurate Prediction of Protein Structures and Interactions Using a Three-Track Neural Network. *Science* **2021**, *373*, 871−876.

(9) Segler, M. H. S.; Kogej, T.; Tyrchan, C.; Waller, M. P. Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. *ACS Cent. Sci.* **2018**, *4*, 120−131.

(10) Merk, D.; Friedrich, L.; Grisoni, F.; Schneider, G. De Novo Design of Bioactive Small Molecules by Artificial Intelligence. *Mol. Inf.* **2018**, *37*, No. 1700153.

(11) Yuan, W.; Jiang, D.; Nambiar, D. K.; Liew, L. P.; Hay, M. P.; Bloomstein, J.; Lu, P.; Turner, B.; Le, Q.-T.; Tibshirani, R.; Khatri, P.; Moloney, M. G.; Koong, A. C. Chemical Space Mimicry for Drug Discovery. *J. Chem. Inf. Model.* **2017**, *57*, 875−882.

(12) Schmidhuber, J. Deep Learning in Neural Networks: An Overview. *Neural Networks* **2015**, *61*, 85−117.

(13) LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521*, 436−444.

(14) Atz, K.; Grisoni, F.; Schneider, G. Geometric Deep Learning on Molecular Representations. *Nat. Mach. Intell.* **2021**, 3, 1023−1032.

(15) Jiang, D.; Wu, Z.; Hsieh, C.-Y.; Chen, G.; Liao, B.; Wang, Z.; Shen, C.; Cao, D.; Wu, J.; Hou, T. Could Graph Neural Networks Learn Better Molecular Representation for Drug Discovery? A Comparison Study of Descriptor-Based and Graph-Based Models. *J. Cheminf.* **2021**, *13*, No. 12.

(16) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; Palmer, A.; Settels, V.; Jaakkola, T.; Jensen, K.; Barzilay, R. Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **2019**, *59*, 3370−3388.

(17) Valsecchi, C.; Collarile, M.; Grisoni, F.; Todeschini, R.; Ballabio, D.; Consonni, V. Predicting Molecular Activity on Nuclear Receptors by Multitask Neural Networks. *J. Chemom.* **2020**, No. e3325.

(18) Johnson, M. A.; Maggiora, G. M.*Concepts and Applications of Molecular Similarity*; Wiley, 1990.

(19) Stumpfe, D.; Hu, H.; Bajorath, J. Advances in Exploring Activity Cliffs. *J. Comput. Aided Mol. Des.* **2020**, *34*, 929−942.

(20) Maggiora, G. M. On Outliers and Activity Cliffs–Why QSAR Often Disappoints. *J. Chem. Inf. Model.* **2006**, *46*, 1535.

(21) Stumpfe, D.; Bajorath, J. Exploring Activity Cliffs in Medicinal Chemistry. *J. Med. Chem.* **2012**, *55*, 2932−2942.

(22) Dimova, D.; Heikamp, K.; Stumpfe, D.; Bajorath, J. Do Medicinal Chemists Learn from Activity Cliffs? A Systematic Evaluation of Cliff Progression in Evolving Compound Data Sets. *J. Med. Chem.* **2013**, *56*, 3339−3345.

(23) Wedlake, A. J.; Folia, M.; Piechota, S.; Allen, T. E. H.; Goodman, J. M.; Gutsell, S.; Russell, P. J. Structural Alerts and Random Forest Models in a Consensus Approach for Receptor Binding Molecular Initiating Events. *Chem. Res. Toxicol.* **2020**, *33*, 388−401.

(24) Hu, Y.; Bajorath, J. Extending the Activity Cliff Concept: Structural Categorization of Activity Cliffs and Systematic Identification of Different Types of Cliffs in the ChEMBL Database. *J. Chem. Inf. Model.* **2012**, *52*, 1806−1811.

(25) Cruz-Monteagudo, M.; Medina-Franco, J. L.; Pérez-Castillo, Y.; Nicolotti, O.; Cordeiro, M. N. D. S.; Borges, F. Activity Cliffs in Drug Discovery: Dr Jekyll or Mr Hyde? *Drug Discovery Today* **2014**, *19*, 1069−1080.

(26) Stumpfe, D.; Hu, Y.; Dimova, D.; Bajorath, J. Recent Progress in Understanding Activity Cliffs and Their Utility in Medicinal Chemistry. *J. Med. Chem.* **2014**, *57*, 18−28.

(27) Bajorath, J.; Peltason, L.; Wawer, M.; Guha, R.; Lajiness, M. S.; Van Drie, J. H. Navigating Structure−Activity Landscapes. *Drug Discovery Today* **2009**, *14*, 698−705.

(28) Husby, J.; Bottegoni, G.; Kufareva, I.; Abagyan, R.; Cavalli, A. Structure-Based Predictions of Activity Cliffs. *J. Chem. Inf. Model.* **2015**, *55*, 1062−1076.

(29) Consonni, V.; Todeschini, R.*Molecular Descriptors for Chemoinformatics: Volume I: Alphabetical Listing / Volume II: Appendices, References*; John Wiley & Sons, 2009.

(30) Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: A Benchmark for Molecular Machine Learning. *Chem. Sci.* **2018**, *9*, 513−530.

(31) Feinberg, E. N.; Sur, D.; Wu, Z.; Husic, B. E.; Mai, H.; Li, Y.; Sun, S.; Yang, J.; Ramsundar, B.; Pande, V. S. PotentialNet for Molecular Property Prediction. *ACS Cent Sci* **2018**, *4*, 1520−1530.

(32) Hu, W.; Fey, M.; Zitnik, M.; Dong, Y.; Ren, H.; Liu, B.; Catasta, M.; Leskovec, J. Open Graph Benchmark: Datasets for Machine Learning on Graphs. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 22118−22133.

(33) Stanley, M.; Bronskill, J. F.; Maziarz, K.; Misztela, H.; Lanini, J.; Segler, M.; Schneider, N.; Brockschmidt, M. In *FS-Mol: A Few-Shot Learning Dataset of Molecules*, Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2), 2021.

(34) Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Fei-Fei, L. In *ImageNet: A Large-Scale Hierarchical Image Database*, 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009; pp 248−255.

(35) Wang, A.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; Bowman, S. R.GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. 2018, arXiv:1804.07461. arXiv.org e-Print archive. https://arxiv.org/abs/1804.07461.

(36) Brown, N.; Fiscato, M.; Segler, M. H. S.; Vaucher, A. C. GuacaMol: Benchmarking Models for de Novo Molecular Design. *J. Chem. Inf. Model.* **2019**, *59*, 1096−1108.

(37) Raji, I. D.; Bender, E. M.; Paullada, A.; Denton, E.; Hanna, A.AI and the Everything in the Whole Wide World Benchmark. 2021, arXiv:2111.15366. arXiv.org e-Print archive. https://arxiv.org/abs/2111.15366.

(38) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.* **2012**, *40*, D1100−D1107.

(39) Tiikkainen, P.; Bellis, L.; Light, Y.; Franke, L. Estimating Error Rates in Bioactivity Databases. *J. Chem. Inf. Model.* **2013**, *53*, 2499−2505.

(40) Mansouri, K.; Grulke, C. M.; Richard, A. M.; Judson, R. S.; Williams, A. J. An Automated Curation Procedure for Addressing

Chemical Errors and Inconsistencies in Public Datasets Used in QSAR Modelling. *SAR QSAR Environ. Res.* **2016**, *27*, 911−937.

(41) Fourches, D.; Muratov, E.; Tropsha, A. Trust, but Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research. *J. Chem. Inf. Model.* **2010**, *50*, 1189−1204.

(42) Cereto-Massagué, A.; Ojeda, M. J.; Valls, C.; Mulero, M.; Garcia-Vallvé, S.; Pujadas, G. Molecular Fingerprint Similarity Search in Virtual Screening. *Methods* **2015**, *71*, 58−63.

(43) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.

(44) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893.

(45) Yujian, L.; Bo, L. A Normalized Levenshtein Distance Metric. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 1091−1095.

(46) Hussain, J.; Rea, C. Computationally Efficient Algorithm to Identify Matched Molecular Pairs (MMPs) in Large Data Sets. *J. Chem. Inf. Model.* **2010**, *50*, 339−348.

(47) Hu, X.; Hu, Y.; Vogt, M.; Stumpfe, D.; Bajorath, J. MMP-Cliffs: Systematic Identification of Activity Cliffs on the Basis of Matched Molecular Pairs. *J. Chem. Inf. Model.* **2012**, *52*, 1138−1145.

(48) Bajorath, J. Representation and Identification of Activity Cliffs. *Expert Opin. Drug Discovery* **2017**, *12*, 879−883.

(49) Puzyn, T.; Mostrag-Szlichtyng, A.; Gajewicz, A.; Skrzyński, M.; Worth, A. P. Investigating the Influence of Data Splitting on the Predictive Ability of QSAR/QSPR Models. *Struct. Chem.* **2011**, *22*, 795−804.

(50) Stella, X. Y.; Shi, J. In *Multiclass Spectral Clustering*, IEEE International Conference on. Vol. 2. IEEE Computer Society, 2003.

(51) Fix, E.; Hodges, J. L. Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties. *Int. Stat. Rev.* **1989**, *57*, 238−247.

(52) Breiman, L. Bagging Predictors. *Mach. Learn.* **1996**, *24*, 123−140.

(53) Friedman, J. H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **2001**, *29*, 1189−1232.

(54) Cristianini, N.; Shawe-Taylor, J.Department of Computer Science Royal Holloway John Shawe-Taylor. In *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*; Cambridge University Press, 2000.

(55) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273−1280.

(56) Todeschini, R.; Gramatica, P.New 3D Molecular Descriptors: The WHIM Theory and QSAR Applications. In *3D QSAR in Drug Design*; Kluwer Academic Publishers: Dordrecht, 2005; pp 355−380.

(57) Walters, W. P.; Murcko, M. A. Prediction of ‘Drug-Likeness. *Adv. Drug Delivery Rev.* **2002**, *54*, 255−271.

(58) Stokes, J. M.; Yang, K.; Swanson, K.; Jin, W.; Cubillos-Ruiz, A.; Donghia, N. M.; MacNair, C. R.; French, S.; Carfrae, L. A.; Bloom-Ackermann, Z.; Tran, V. M.; Chiappino-Pepe, A.; Badran, A. H.; Andrews, I. W.; Chory, E. J.; Church, G. M.; Brown, E. D.; Jaakkola, T. S.; Barzilay, R.; Collins, J. J. A Deep Learning Approach to Antibiotic Discovery. *Cell* **2020**, *181*, 475−483.

(59) Xiong, Z.; Wang, D.; Liu, X.; Zhong, F.; Wan, X.; Li, X.; Li, Z.; Luo, X.; Chen, K.; Jiang, H.; Zheng, M. Pushing the Boundaries of Molecular Representation for Drug Discovery with the Graph Attention Mechanism. *J. Med. Chem.* **2020**, *63*, 8749−8760.

(60) Chen, C.; Ye, W.; Zuo, Y.; Zheng, C.; Ong, S. P. Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals. *Chem. Mater.* **2019**, *31*, 3564−3572.

(61) Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; Riley, P. Molecular Graph Convolutions: Moving beyond Fingerprints. *J. Comput. Aided Mol. Des.* **2016**, *30*, 595−608.

(62) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E.Neural Message Passing for Quantum Chemistry. In *Proceedings of the 34th International Conference on Machine Learning*; Precup, D.; Teh, Y. W., Eds.; Proceedings of Machine Learning Research; PMLR, 06--11 Aug **2017**; Vol. *70*, pp 1263−1272.

(63) Velickovic, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; Bengio, Y.*Graph Attention Networks*. 2017, arXiv:1710.10903, arXiv.org e-Print archive; Vol. *1050*, p 20. https://arxiv.org/abs/1710.10903.

(64) Kipf, T. N.; Welling, M.Semi-Supervised Classification with Graph Convolutional Networks. 2016, arXiv:1609.02907, arXiv.org e-Print archive. https://arxiv.org/abs/1609.02907.

(65) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.

(66) Kimber, T. B.; Gagnebin, M.; Volkamer, A. Maxsmi: Maximizing Molecular Property Prediction Performance with Confidence Estimation Using SMILES Augmentation and Deep Learning. *Artif. Intell. Life Sci.* **2021**, *1*, No. 100014.

(67) Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735−1780.

(68) Moret, M.; Grisoni, F.; Katzberger, P.; Schneider, G. Perplexity-Based Molecule Ranking and Bias Estimation of Chemical Language Models. *J. Chem. Inf. Model.* **2022**, *62*, 1199−1206.

(69) Vaswani, A.; Shazeer, N.; Parmar, N.et al.Attention Is All You Need. *Adv. Neural Inf. Process. Syst.*, **2017**.

(70) Chithrananda, S.; Grand, G.; Ramsundar, B.ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction. 2020, arXiv:2010.09885. arXiv.org e-Print archive. https://arxiv.org/abs/2010.09885.

(71) Arús-Pous, J.; Johansson, S. V.; Prykhodko, O.; Bjerrum, E. J.; Tyrchan, C.; Reymond, J.-L.; Chen, H.; Engkvist, O. Randomized SMILES Strings Improve the Quality of Molecular Generative Models. *J. Cheminf.* **2019**, *11*, No. 71.

(72) Bjerrum, E. J.SMILES Enumeration as Data Augmentation for Neural Network Modeling of Molecules. 2017, arXiv:1703.07076. arXiv.org e-Print archive. https://arxiv.org/abs/1703.07076.

(73) Sheridan, R. P. Three Useful Dimensions for Domain Applicability in QSAR Models Using Random Forest. *J. Chem. Inf. Model.* **2012**, *52*, 814−823.

(74) Guha, R.; Dutta, D.; Jurs, P. C.; Chen, T. Local Lazy Regression: Making Use of the Neighborhood to Improve QSAR Predictions. *J. Chem. Inf. Model.* **2006**, *46*, 1836−1847.

(75) Subramanian, G.; Ramsundar, B.; Pande, V.; Denny, R. A. Computational Modeling of β-Secretase 1 (BACE-1) Inhibitors Using Ligand Based Approaches. *J. Chem. Inf. Model.* **2016**, *56*, 1936−1949.

(76) Todeschini, R.; Ballabio, D.; Cassotti, M.; Consonni, V. N3 and BNN: Two New Similarity Based Classification Methods in Comparison with Other Classifiers. *J. Chem. Inf. Model.* **2015**, *55*, 2365−2374.

(77) Grisoni, F.; Merk, D.; Byrne, R.; Schneider, G. Scaffold-Hopping from Synthetic Drugs by Holistic Molecular Representation. *Sci. Rep.* **2018**, *8*, No. 16469.

(78) Tetko, I. V.; Sushko, I.; Pandey, A. K.; Zhu, H.; Tropsha, A.; Papa, E.; Öberg, T.; Todeschini, R.; Fourches, D.; Varnek, A. Critical Assessment of QSAR Models of Environmental Toxicity against Tetrahymena Pyriformis: Focusing on Applicability Domain and Overfitting by Variable Selection. *J. Chem. Inf. Model.* **2008**, *48*, 1733−1746.

(79) Zhu, H.; Tropsha, A.; Fourches, D.; Varnek, A.; Papa, E.; Gramatica, P.; Oberg, T.; Dao, P.; Cherkasov, A.; Tetko, I. V. Combinatorial QSAR Modeling of Chemical Toxicants Tested against Tetrahymena Pyriformis. *J. Chem. Inf. Model.* **2008**, *48*, 766−784.

(80) de la Vega de León, A.; Bajorath, J. Prediction of Compound Potency Changes in Matched Molecular Pairs Using Support Vector Regression. *J. Chem. Inf. Model.* **2014**, *54*, 2654−2663.

(81) Sheridan, R. P.; Karnachi, P.; Tudor, M.; Xu, Y.; Liaw, A.; Shah, F.; Cheng, A. C.; Joshi, E.; Glick, M.; Alvarez, J. Experimental Error, Kurtosis, Activity Cliffs, and Methodology: What Limits the Predictivity of Quantitative Structure-Activity Relationship Models? *J. Chem. Inf. Model.* **2020**, *60*, 1969−1982.

(82) Cai, C.; Wang, S.; Xu, Y.; Zhang, W.; Tang, K.; Ouyang, Q.; Lai, L.; Pei, J. Transfer Learning for Drug Discovery. *J. Med. Chem.* **2020**, *63*, 8683−8694.

(83) Gupta, A.; Müller, A. T.; Huisman, B. J. H.; Fuchs, J. A.; Schneider, P.; Schneider, G. Generative Recurrent Networks for De Novo Drug Design. *Mol. Inf.* **2018**, *37*, No. 1700111.

(84) Awale, M.; Sirockin, F.; Stiefl, N.; Reymond, J.-L. Drug Analogs from Fragment-Based Long Short-Term Memory Generative Neural Networks. *J. Chem. Inf. Model.* **2019**, *59*, 1347−1356.

(85) Hu, W.; Liu, B.; Gomes, J.; Zitnik, M.; Liang, P.; Pande, V.; Leskovec, J.Strategies for Pre-Training Graph Neural Networks. 2019, arXiv:1905.12265. arXiv.org e-Print archive. https://arxiv.org/abs/1905.12265.

(86) Veličković, P.; Fedus, W.; Hamilton, W. L.; Liò, P.; Bengio, Y.; Hjelm, R. D.Deep Graph Infomax. 2018, arXiv preprint arXiv:1809.10341.

(87) Hamilton, W. L.; Ying, R.; Leskovec, J.Inductive Representation Learning on Large Graphs. 2017, arXiv:1706.02216. arXiv.org e-Print archive.

(88) Wang, H.; Kaddour, J.; Liu, S.; Tang, J.; Kusner, M.; Lasenby, J.; Liu, Q.*Evaluating Self-Supervised Learning for Molecular Graph Embeddings*. 2022, arXiv:2206.08005. arXiv.org e-Print archive. https://arxiv.org/abs/2206.08005.

(89) Jiménez-Luna, J.; Skalic, M.; Weskamp, N. Benchmarking Molecular Feature Attribution Methods with Activity Cliffs. *J. Chem. Inf. Model.* **2022**, *62*, 274−283.

(90) Kwapien, K.; Nittinger, E.; He, J.; Margreitter, C.; Voronov, A.; Tyrchan, C. Implications of Additivity and Nonadditivity for Machine Learning and Deep Learning Models in Drug Design. *ACS Omega* **2022**, *7*, 26573−26581.

(91) Stumpfe, D.; Hu, H.; Bajorath, J. Introducing a New Category of Activity Cliffs with Chemical Modifications at Multiple Sites and Rationalizing Contributions of Individual Substitutions. *Bioorg. Med. Chem.* **2019**, *27*, 3605−3612.

(92) Hu, H.; Bajorath, J. Introducing a New Category of Activity Cliffs Combining Different Compound Similarity Criteria. *RSC Med Chem* **2020**, *11*, 132−141.

(93) Guha, R.; Van Drie, J. H. Structure–Activity Landscape Index: Identifying and Quantifying Activity Cliffs. *J. Chem. Inf. Model.* **2008**, *48*, 646−658.

(94) Gogishvili, D.; Nittinger, E.; Margreitter, C.; Tyrchan, C. Nonadditivity in Public and Inhouse Data: Implications for Drug Design. *J. Cheminf.* **2021**, *13*, No. 47.

(95) Vogt, M. Progress with Modeling Activity Landscapes in Drug Discovery. *Expert Opin. Drug Discovery* **2018**, *13*, 605−615.

(96) Fourches, D.; Ash, J. 4D- Quantitative Structure-Activity Relationship Modeling: Making a Comeback. *Expert Opin. Drug Discovery* **2019**, *14*, 1227−1235.

(97) Zbontar, J.; Jing, L.; Misra, I.; LeCun, Y.; Deny, S.Barlow Twins: Self-Supervised Learning via Redundancy Reduction. 2021, arXiv:2103.03230, arXiv.org e-Print archive. https://arxiv.org/abs/2103.03230.

(98) Wallach, I.; Dzamba, M.; Heifets, A.AtomNet: A Deep Convolutional Neural Network for Bioactivity Prediction in Structure-Based Drug Discovery. 2015, arXiv:1510.02855. arXiv.org e-Print archive. https://arxiv.org/abs/1510.02855.

(99) Gentile, F.; Agrawal, V.; Hsing, M.; Ton, A.-T.; Ban, F.; Norinder, U.; Gleave, M. E.; Cherkasov, A. Deep Docking: A Deep Learning Platform for Augmentation of Structure Based Drug Discovery. *ACS Cent. Sci.* **2020**, *6*, 939−949.

(100) Volkov, M.; Turk, J.-A.; Drizard, N.; Martin, N.; Hoffmann, B.; Gaston-Mathé, Y.; Rognan, D. On the Frustration to Predict Binding Affinities from Protein−Ligand Structures with Deep Neural Networks. *J. Med. Chem.* **2022**, *65*, 7946−7958.

(101) Sieg, J.; Flachsenberg, F.; Rarey, M. In Need of Bias Control: Evaluating Chemical Data for Machine Learning in Structure-Based Virtual Screening. *J. Chem. Inf. Model.* **2019**, *59*, 947−961.

(102) Chen, L.; Cruz, A.; Ramsey, S.; Dickson, C. J.; Duca, J. S.; Hornak, V.; Koes, D. R.; Kurtzman, T. Hidden Bias in the DUD-E Dataset Leads to Misleading Performance of Deep Learning in Structure-Based Virtual Screening. *PLoS One* **2019**, *14*, No. e0220113.

(103) Sheridan, R. P.; Culberson, J. C.; Joshi, E.; Tudor, M.; Karnachi, P. Prediction Accuracy of Production ADMET Models as a Function of Version: Activity Cliffs Rule. *J. Chem. Inf. Model.* **2022**, *62*, 3275−3280.

(104) *RDKit: Open-source cheminformatics*. http://www.rdkit.org.

(105) Rorabacher, D. B. Statistical Treatment for Rejection of Deviant Values: Critical Values of Dixon's "Q" Parameter and Related Subrange Ratios at the 95% Confidence Level. *Anal. Chem.* **1991**, *63*, 139−146.

(106) Riniker, S.; Landrum, G. A. Better Informed Distance Geometry: Using What We Know To Improve Conformation Generation. *J. Chem. Inf. Model.* **2015**, *55*, 2562−2574.

(107) Halgren, T. A. Merck Molecular Force Field. I. Basis, Form, Scope, Parameterization, and Performance of MMFF94. *Comput. Chem.* **1996**, *17*, 490−519.

(108) Wildman, S. A.; Crippen, G. M. Prediction of Physicochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868−873.

(109) *python-Levenshtein*. https://pypi.org/project/python-Levenshtein/ (accessed Nov 23, 2021).

(110) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.

(111) Virtanen, P.; Gommers, R.; Oliphant, T. E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; van der Walt, S. J.; Brett, M.; Wilson, J.; Millman, K. J.; Mayorov, N.; Nelson, A. R. J.; Jones, E.; Kern, R.; Larson, E.; Carey, C. J.; Polat, İ.; Feng, Y.; Moore, E. W.; VanderPlas, J.; Laxalde, D.; Perktold, J.; Cimrman, R.; Henriksen, I.; Quintero, E. A.; Harris, C. R.; Archibald, A. M.; Ribeiro, A. H.; Pedregosa, F.; van Mulbregt, P. SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods* **2020**, *17*, 261−272.

(112) Gasteiger, J.; Marsili, M. Iterative Partial Equalization of Orbital Electronegativity—a Rapid Access to Atomic Charges. *Tetrahedron* **1980**, *36*, 3219−3228.

(113) Fey, M.; Lenssen, J. E.Fast Graph Representation Learning with PyTorch Geometric. 2019, arXiv:1903.02428. arXiv.org e-Print archive. https://arxiv.org/abs/1903.02428.

(114) Baek, J.; Kang, M.; Hwang, S. J.Accurate Learning of Graph Representations with Graph Multiset Pooling. 2021, arXiv:2102.11533, arXiv.org e-Print archive. https://arxiv.org/abs/2102.11533.

(115) Brody, S.; Alon, U.; Yahav, E.How Attentive Are Graph Attention Networks?. 2021, arXiv:2105.14491. arXiv.org e-Print archive. https://arxiv.org/abs/2105.14491.

(116) Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Kopf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; Chintala, S.PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Adv. Neural Inf. Process. Syst.***2019**, 32.

(117) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E. PubChem 2019 Update: Improved Access to Chemical Data. *Nucleic Acids Res.* **2019**, *47*, D1102−D1109.

(118) Alexander, D. L. J.; Tropsha, A.; Winkler, D. A. Beware of R2: Simple, Unambiguous Assessment of the Prediction Accuracy of QSAR and QSPR Models. *J. Chem. Inf. Model.* **2015**, *55*, 1316−1322.

(119) Consonni, V.; Todeschini, R.; Ballabio, D.; Grisoni, F. On the Misleading Use of Q2 F3 for QSAR Model Comparison. *Mol. Inf.* **2019**, *38*, No. e1800029.