Journal of Cheminformatics

RESEARCH

Open Access

Exploring QSAR models for activity-cliff prediction

Markus Dablander[1], Thierry Hanser[2], Renaud Lambiotte[1] and Garrett M. Morris[3*]

Abstract

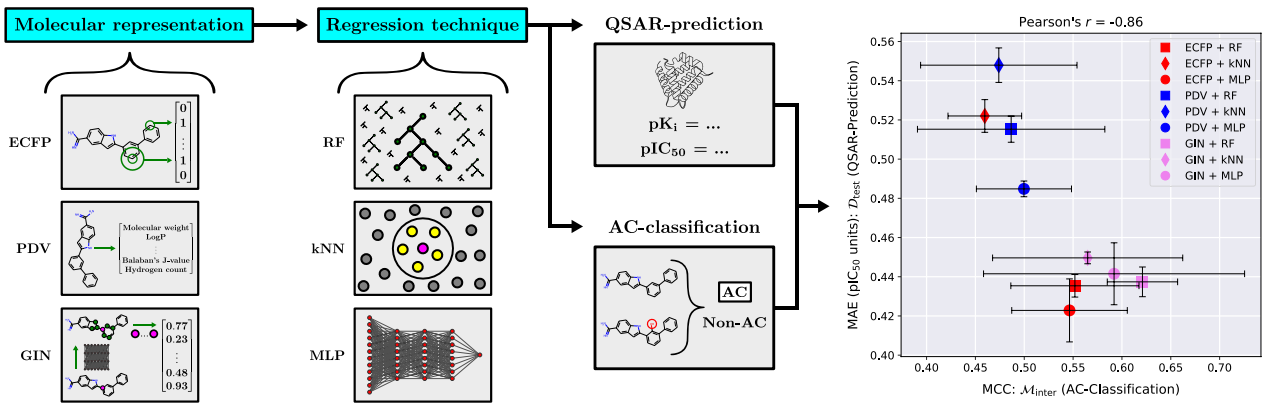
Introduction and methodology Pairs of similar compounds that only differ by a small structural modification but exhibit a large difference in their binding affinity for a given target are known as activity cliffs (ACs). It has been hypothesised that QSAR models struggle to predict ACs and that ACs thus form a major source of prediction error. However, the AC-prediction power of modern QSAR methods and its quantitative relationship to general QSAR-prediction performance is still underexplored. We systematically construct nine distinct QSAR models by combining three molecular representation methods (extended-connectivity fingerprints, physicochemical-descriptor vectors and graph isomorphism networks) with three regression techniques (random forests, k-nearest neighbours and multilayer perceptrons); we then use each resulting model to classify pairs of similar compounds as ACs or non-ACs and to predict the activities of individual molecules in three case studies: dopamine receptor D2, factor Xa, and SARS-CoV-2 main protease.

Results and conclusions Our results provide strong support for the hypothesis that indeed QSAR models frequently fail to predict ACs. We observe low AC-sensitivity amongst the evaluated models when the activities of both compounds are unknown, but a substantial increase in AC-sensitivity when the actual activity of one of the compounds is given. Graph isomorphism features are found to be competitive with or superior to classical molecular representations for AC-classification and can thus be employed as baseline AC-prediction models or simple compound-optimisation tools. For general QSAR-prediction, however, extended-connectivity fingerprints still consistently deliver the best performance amongs the tested input representations. A potential future pathway to improve QSAR-modelling performance might be the development of techniques to increase AC-sensitivity.

Keywords QSAR modelling, Activity cliffs, Activity cliff prediction, Machine learning, Deep learning , Molecular representation, Physicochemical descriptors, Extended-connectivity fingerprints, Graph isomorphism networks, Binding affinity prediction

*Correspondence:
Garrett M. Morris
garrett.morris@stats.ox.ac.uk
Full list of author information is available at the end of the article

Graphical Abstract



Introduction

Activity cliffs (ACs) are pairs of small molecules that exhibit high structural similarity but at the same time show an unexpectedly large difference in their binding affinity against a given pharmacological target [9, 37, 50, 51, 54–56]. The existence of ACs directly defies the intuitive idea that chemical compounds with similar structures should have similar activities, often referred to as the *molecular similarity principle*. An example of an AC between two inhibitors of blood coagulation factor Xa [33] is depicted in Fig. 1; a small chemical modification involving the addition of a hydroxyl group leads to an increase in inhibition of almost three orders of magnitude.

For medicinal chemists, ACs can be puzzling and confound their understanding of structure-activity relationships (SARs) [13, 54, 62]. ACs reveal small compound-modifications with large biological impact and thus represent rich sources of pharmacological information. Mechanisms by which a small structural transformation can give rise to an AC include a drastic change in 3D-conformation and/or the switching to a different binding mode or even binding site. ACs form discontinuities in the SAR-landscape and can therefore have a crucial impact on the success of lead-optimisation programmes. While knowledge about ACs can be powerful when trying to escape from flat regions of the SAR-landscape, their presence can be detrimental in later stages of the drug development process, when multiple molecular properties beyond mere activity need to be balanced carefully to arrive at a safe and effective compound [9, 54].

In the field of computational chemistry, ACs are suspected to form one of the major roadblocks for successful quantitative structure-activity relationship (QSAR) modelling [9, 18, 37, 50]; abrupt changes in potency are expected to negatively influence machine learning algorithms for pharmacological activity prediction. During the development of QSAR models, ACs are sometimes dismissed as measurement errors [39], but simply removing ACs from a training data set can result in a loss of precious SAR-information [10].

Golbraikh et al. [18] developed the MODI metric to quantify the smoothness of the SAR-landscape of binary molecular classification data sets and showed that the SAR-landscape smoothness is a strong determinant for downstream QSAR-modelling performance. In a related work, Sheridan et al. [50] found that the density of ACs in a molecular data set is strongly predictive of its overall modelability by classical descriptor- and fingerprint-based QSAR methods. Furthermore, they found that such methods incur a significant drop in performance when the test set is restricted to "cliffy" compounds that form a large number of ACs. In a more extensive study, van Tilborg et al. [60] observed a similar drop in performance when testing classical and graph-based QSAR techniques on compounds involved in ACs. Notably, in both studies this performance drop was also observed for highly nonlinear and adaptive deep learning models. In fact, van Tilborg reports that descriptor-based QSAR methods even outperform more complex deep learning models on "cliffy" compounds associated with ACs. This runs counter to earlier hopes expressed in the literature that the approximation power of deep neural networks might ameliorate the problem of ACs [64].

While these works provide valuable insights into the detrimental effects of SAR discontinuity on QSAR models, they consider ACs mainly indirectly by focussing on *individual* compounds involved in ACs. Arguably, a

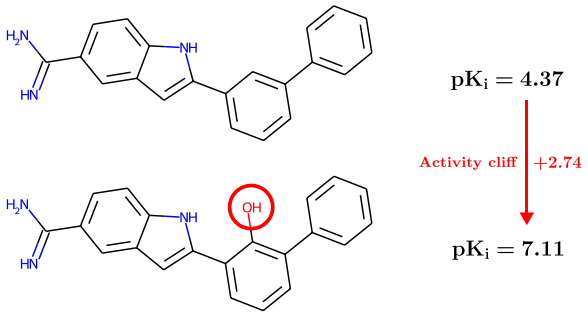


Fig. 1 Example of an activity cliff (AC) for blood coagulation factor Xa. A small structural transformation in the upper compound leads to an increase in inhibitory activity of almost three orders of magnitude. Both compounds were identified in the same ChEMBL assay with ID 658338

distinct and more natural approach would be to investigate ACs directly at the level of compound *pairs*. This approach has been followed in the AC-prediction field which is concerned with developing techniques to classify whether a pair of similar compounds forms an AC or not. An effective AC-prediction method would be of high value for drug development with important applications in rational compound optimisation and automatic SAR-knowledge acquisition.

The AC-prediction literature is still very thin compared to the QSAR-prediction literature. An attempt to conduct an exhaustive literature review on AC-prediction techniques revealed a total number of 15 methods [3, 5, 7, 11, 19, 21, 24, 26, 30, 34, 41–43, 46, 57], all of which have been published since 2012. Current AC-prediction methods are often based on creative ways to extract features from pairs of molecular compounds in a manner suitable for standard machine learning pipelines. For example, Horvath et al. [21] used condensed graphs of reactions [20, 27], a representation technique originally introduced for modelling of chemical reactions, to encode pairs of similar compounds and subsequently predict ACs. Another method was recently described by Iqbal et al. [26] who investigated the abilities of convolutional neural networks operating on 2D images of compound pairs to distinguish between ACs and non-ACs. Interestingly, none of the AC-prediction methods we identified employ feature extraction techniques built on modern graph neural networks (GNNs) [14, 17, 31, 61, 66] with the exception of Park et al. [43] who recently applied graph convolutional methods to compound-pairs to predict ACs.

In spite of the existence of advanced AC-prediction models there are significant gaps left in the current AC-prediction literature. Note that any QSAR model can immediately be repurposed as an AC-prediction model by using it to individually predict the activities of two structurally similar compounds and then thresholding the predicted absolute activity difference. Nevertheless, at the moment there is no study that uses this straightforward technique to investigate the potential of current QSAR models to classify whether a pair of compounds forms an AC or not. Importantly, this also entails that the most salient AC-prediction models [19, 21, 26, 34, 57] have not been compared to a simple QSAR-modelling baseline applied to compound pairs. It is thus an open question to what extent (if at all) these tailored AC-prediction techniques outcompete repurposed QSAR methods in the detection of ACs. This is especially relevant in light of the fact that several published AC-prediction models [19, 26, 34] are evaluated via compound-pair-based data splits which incur a significant overlap between training set and test set at the level of individual molecules; this type of data split should strongly favour standard QSAR models for AC-prediction, yet a comparison to such baseline methods is lacking.

We address these gaps by systematically investigating the abilities of nine frequently used QSAR models to classify pairs of similar compounds as ACs or non-ACs within three pharmacological data sets: dopamine receptor D2, factor Xa, and SARS-CoV-2 main protease. Each QSAR model is constructed by combining a molecular representation method (physicochemical-descriptor vectors (PDVs) [58], extended-connectivity fingerprints (ECFPs) [47], or graph isomorphism networks (GINs) [66]) with a regression technique (random forests (RFs), k-nearest neighbours (kNNs), or multilayer perceptrons (MLPs)). All models are used for two distinct prediction tasks: QSAR-prediction at the level of individual molecules, and AC-classification at the level of compound-pairs. The main contribution of this study is to shed light on the following questions:

- What is the relationship between the ability of a QSAR model to predict the activities of individual compounds, versus its ability to classify whether pairs of similar compounds form ACs?
- When (if at all) are common QSAR models capable of predicting ACs?
- When (if at all) are common QSAR models capable of predicting which of two similar compounds is the more active one?
- Which QSAR model shows the strongest AC-prediction performance, and should thus be used as a baseline against which to compare tailored AC-prediction models?

- Do trainable GINs outperform classical precomputed ECFPs and PDVs as molecular representations for QSAR- and/or AC-prediction?
- How could ACs potentially be used to improve QSAR-modelling performance?

Experimental methodology

Molecular data sets

We built three binding affinity data sets of small-molecule inhibitors of dopamine receptor D2, factor Xa, and SARS-CoV-2 main protease. Factor Xa is an enzyme in the coagulation cascade and a canonical target for blood-thinning drugs [33]. Dopamine receptor D2 is the main site of action for classic antipsychotic drugs which act as antagonists of the D2 receptor [49]. SARS-CoV-2 main protease is one of the key enzymes in the viral replication cycle of the SARS coronavirus 2, that recently caused the unprecedented COVID-19 pandemic; it is one of the most promising targets for antiviral drugs against this coronavirus [59]. Note that while we focus on three target-based data sets in this work, it might be worthwhile and interesting to extend the methodology of this study to also include ADMET-based data sets in the future.

For dopamine receptor D2 and factor Xa, data was extracted from the ChEMBL database [35] in the form of SMILES strings with associated K_i [nM] values. For SARS-CoV-2 main protease, data was obtained from the COVID moonshot project [1] in the form of SMILES strings with associated IC_{50} [μM] values. SMILES strings were standardised and desalted via the ChEMBL structure pipeline [6]. This step also removed solvents and all isotopic information. Following this, SMILES strings that produced error messages when turned into an RDKit mol object were deleted. Finally, a scan for duplicate molecules was performed: If the activities in a set of duplicate molecules were within the same order of magnitude then the set was unified via geometric averaging. Otherwise, the measurements were considered unreliable and the corresponding set of duplicate molecules was removed. This procedure reduced the data set for dopamine receptor D2 / factor Xa / SARS-CoV-2 main protease from 8883 / 4116 / 1926 compounds to 6333 / 3605 / 1924 unique compounds whereby 174 / 21 / 0 sets of duplicate SMILES were removed and the rest was unified.

Activity cliffs: definition of binary classification tasks

The exact definition of an AC hinges on two concepts: structural similarity and large activity difference. An elegant technique to measure structural similarity in the context of AC analysis is given by the matched molecular pair (MMP) formalism [23, 29]. An MMP is a pair of compounds that share a common structural core but differ by a small chemical transformation at a specific site. Figure 1 depicts an example of an MMP whose variable parts are formed by a hydrogen atom and a hydroxyl group. To detect MMPs algorithmically, we used the `mmpdb` `Python`-package provided by Dalke et al. [12]. We restricted ourselves to the commonly used definition of MMPs [19, 21, 57] which employs relatively generous size constraints: the MMP core was required to contain at least twice as many heavy atoms as either of the two variable parts; each variable part was required to contain no more than 13 heavy atoms; the maximal size difference between both variable parts was set to eight heavy atoms; and bond cutting was restricted to single exocyclic bonds. To guarantee a well-defined mapping from each MMP to a unique structural core, we canonically chose the core that contained the largest number of heavy atoms whenever there was ambiguity.

Based on the ratio of the activity values of both MMP compounds, each MMP was assigned to one of three classes: "AC", "non-AC" or "half-AC". In accordance with the literature [4, 19, 21, 42, 62] we assigned an MMP to the "AC"-class if both activity values differed by at least a factor of 100. If both activity values differed by no more than a factor of 10, then the MMP was assigned to the "non-AC"-class. In the residual case the MMP was assigned to the "half-AC"-class. To arrive at a well-separated binary classification task, we labelled all ACs as positives and all non-ACs as negatives. The half-ACs were removed and not considered further in our experiments. It is relevant to know the direction of a potential activity cliff, i.e. which of the compounds in the pair is the more active one. We thus assigned a binary label to each MMP indicating its potency direction (PD). PD-classification is a balanced binary classification task. Table 1 gives an overview of all our curated data sets.

Table 1 Sizes of our curated data sets and their respective numbers of matched molecular pairs (MMPs), activity cliffs (ACs), half-activity-cliffs (half-ACs) and non-activity-cliffs (non-ACs)

Data Set	Dopamine Receptor D2	Factor Xa	SARS-CoV-2 Main Protease
Compounds	6333	3605	1924
MMPs	35484	21292	12594
ACs	461	1896	521
Half-ACs	3804	4693	1762
Non-ACs	31219	14703	10311
ACs : Non-ACs	$\approx 1:68$	$\approx 1:8$	$\approx 1:20$

Data splitting technique

ACs are molecular pairs rather than single molecules; it is thus not obvious how best to split up a chemical data set into non-overlapping training- and test sets for the fair evaluation of an AC-prediction method. There seems to be no consensus about which data splitting strategy should be canonically used. Several authors [19, 26, 34] have employed a random split at the level of compound pairs. While this technique is conceptually straightforward, it must be expected to incur a significant overlap between training- and test set at the level of individual molecules. For example, randomly splitting up a set of three MMPs $\{\{s,\tilde{s}\},\{s,\hat{s}\},\{\tilde{s},\hat{s}\}\}$ into a training- and a test set might lead to $\{s,\tilde{s}\}$ and $\{s,\hat{s}\}$ getting assigned to the training- and $\{\tilde{s},\hat{s}\}$ getting assigned to the test set which leads to a full inclusion of the test set in the training set at the level of individual molecules. This molecular overlap is problematic for at least three reasons: Firstly, it likely leads to overly optimistic results for AC-prediction methods since they will have already encountered some of the test compounds during training. Secondly, it does not model the natural situation encountered by medicinal chemists who we assume will not know the activity value of at least one compound in a test-set pair. Thirdly, the mentioned molecular overlap should lead to strong AC-prediction results for standard QSAR models, but to the best of our knowledge, no such control experiments have been run in the literature.

Horvath et al. [21] and Tamura et al. [57] have made efforts to address the shortcomings of a compound-pair-based random split. They came up with advanced data splitting algorithms designed to mitigate the molecular-overlap problem by either managing distinct types of test sets according to compound membership in the training set or by designing splitting techniques based on the structural cores of MMPs. However, their data splitting schemes exhibit a relatively high degree of complexity which can make their implementation and interpretation difficult.

We propose a novel data splitting method which represents a favourable trade-off between rigour, interpretability and simplicity. Our technique shares some of its concepts with the methods proposed by Horvath et al. [21] and Tamura et al. [57] but might be simpler to implement and interpret. We first split the data into a training- and test set at the level of individual molecules and then use this basic split to distinguish several types of test sets at the level of compound pairs. Let

$$\mathcal{D} = \{s_1, s_2, \ldots\}$$

be the given data set of individual compounds. Let $\mathcal{M}$ be the set of all MMPs in $\mathcal{D}$ that have been labelled as either ACs or non-ACs. Note that $\mathcal{M}$ is a subset of the set of general compound pairs in $\mathcal{D}$, i.e.

$$\mathcal{M} \subseteq \{\{s_i, s_j\} \mid i \neq j \text{ and } s_i, s_j \in \mathcal{D}\}.$$

In the following, we use the notation $\{s,\tilde{s}\}$ to represent MMPs in $\mathcal{M}$. Each MMP $\{s,\tilde{s}\} \in \mathcal{M}$ shares a common structural core denoted as $\text{core}(\{s,\tilde{s}\})$. We use a random split to partition $\mathcal{D}$ into a training set $\mathcal{D}_{\text{train}}$ and a test set $\mathcal{D}_{\text{test}}$ and then define the following MMP-sets:

$$\begin{aligned}
\mathcal{M}_{\text{train}} &= \{\{s,\tilde{s}\} \in \mathcal{M} \mid s,\tilde{s} \in \mathcal{D}_{\text{train}}\},\\
\mathcal{M}_{\text{inter}} &= \{\{s,\tilde{s}\} \in \mathcal{M} \mid s \in \mathcal{D}_{\text{train}},\ \tilde{s} \in \mathcal{D}_{\text{test}}\},\\
\mathcal{M}_{\text{test}} &= \{\{s,\tilde{s}\} \in \mathcal{M} \mid s,\tilde{s} \in \mathcal{D}_{\text{test}}\},\\
\mathcal{M}_{\text{cores}} &= \{\{s,\tilde{s}\} \in \mathcal{M}_{\text{test}} \mid \text{core}(\{s,\tilde{s}\}) \notin \mathcal{C}_{\text{train}}\}.
\end{aligned}$$

Here,

$$\mathcal{C}_{\text{train}} = \{\text{core}(\{s,\tilde{s}\}) \mid \{s,\tilde{s}\} \in \mathcal{M}_{\text{train}} \cup \mathcal{M}_{\text{inter}}\},$$

which describes the set of MMP-cores that appear in $\mathcal{D}_{\text{train}}$.

Note that $\mathcal{M}_{\text{train}} \cup \mathcal{M}_{\text{inter}} \cup \mathcal{M}_{\text{test}} = \mathcal{M}$. The pair $(\mathcal{D}_{\text{train}}, \mathcal{M}_{\text{train}})$ describes the training space at the level of individual molecules and MMPs, and can be used to train a QSAR- or AC-prediction method. A trained method can then classify MMPs in $\mathcal{M}_{\text{test}}$, $\mathcal{M}_{\text{inter}}$ and $\mathcal{M}_{\text{cores}}$. $\mathcal{M}_{\text{test}}$ models an AC-prediction setting where the activities of both MMP-compounds are unknown. $\mathcal{M}_{\text{cores}}$ represents the subset of MMPs in $\mathcal{M}_{\text{test}}$ whose structural cores do not appear in $\mathcal{M}_{\text{train}} \cup \mathcal{M}_{\text{inter}}$; $\mathcal{M}_{\text{cores}}$ thus models the difficult task of predicting ACs within MMPs that do not contain near analogs to MMP-compounds in the training set. Finally, $\mathcal{M}_{\text{inter}}$ represents an AC-prediction scenario where the activity of one MMP-compound is given *a priori*; this can be interpreted as a compound-optimisation task where one strives to predict small AC-inducing modifications of a query compound with known activity. An illustration of our data splitting strategy is given in Fig. 2.

We implemented our data splitting strategy within a k-fold cross validation scheme repeated with m random seeds. This generated data splits of the form

$$\mathcal{S}^{n,l} = (\mathcal{D}^{n,l}_{\text{train}}, \mathcal{D}^{n,l}_{\text{test}}, \mathcal{M}^{n,l}_{\text{train}}, \mathcal{M}^{n,l}_{\text{test}}, \mathcal{M}^{n,l}_{\text{inter}}, \mathcal{M}^{n,l}_{\text{cores}})$$

for $n \in \{1,\ldots,m\}$ and $l \in \{1,\ldots,k\}$ where $(\mathcal{D}^{n,l}_{\text{train}}, \mathcal{D}^{n,l}_{\text{test}})$ represents the l-th split of $\mathcal{D}$ in the cross validation round with random seed n. The overall QSAR- and AC-prediction performance of each model was recorded as the average over the mk training- and test runs based on all data splits $\mathcal{S}^{1,1},\ldots,\mathcal{S}^{m,k}$. We chose the configuration $(m,k) = (3,2)$ which gave a good trade-off between computational costs and accuracy and reasonable numbers of MMPs in the compound-pair-sets. In particular, random

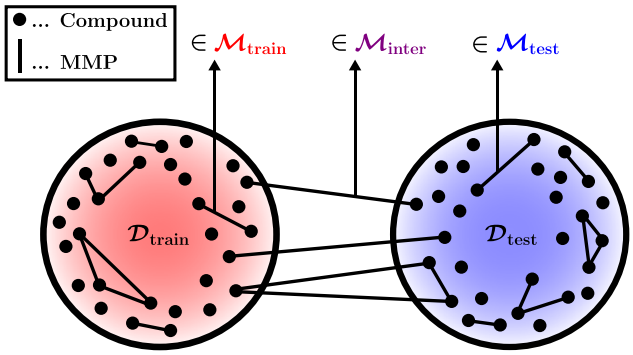


Fig. 2 Illustration of our data splitting strategy. We distinguish between three MMP-sets, $\mathcal{M}_{\text{train}}$, $\mathcal{M}_{\text{inter}}$ and $\mathcal{M}_{\text{test}}$, depending on whether both MMP-compounds are in $\mathcal{D}_{\text{train}}$, one MMP-compound is in $\mathcal{D}_{\text{train}}$ and the other one is in $\mathcal{D}_{\text{test}}$, or both MMP-compounds are in $\mathcal{D}_{\text{test}}$. We additionally consider a fourth MMP-set, $\mathcal{M}_{\text{cores}}$, consisting of the MMPs in $\mathcal{M}_{\text{test}}$ whose structural cores do not appear in $\mathcal{M}_{\text{train}} \cup \mathcal{M}_{\text{inter}}$

cross-validation with $k = 2$ gave expected relative sizes of:

$$|\mathcal{M}_{\text{train}}| : |\mathcal{M}_{\text{inter}}| : |\mathcal{M}_{\text{test}}| = 1 : 2 : 1.$$

On average, 12.7 %, 11.91 %, and 6.84 % of MMPs in $\mathcal{M}_{\text{test}}$ were also in $\mathcal{M}_{\text{cores}}$ for dopamine receptor D2, factor Xa, and SARS-CoV-2 main protease, respectively.

Prediction strategies and performance measures

In a data split of the form

$$\mathcal{S} = (\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{test}}, \mathcal{M}_{\text{train}}, \mathcal{M}_{\text{test}}, \mathcal{M}_{\text{inter}}, \mathcal{M}_{\text{cores}})$$

each individual compound, $s \in \mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{test}} = \mathcal{D}$, can be associated with an activity label $\mathrm{a}(s) \in \mathbb{R}$ which we define as the negative decadic logarithm of the experimentally measured activity of s. We stuck with the original units used in the ChEMBL database and the COVID moonshot project before applying the logarithm ([nM] for K_i and [μM] for IC_{50}); each activity label $a(s)$ thus represents a standard pK_i- or pIC_{50} value that was additively shifted towards 0 due to the use of [nM]- or [μM]-units instead of the usual [M]-units; this shift towards 0 might slightly benefit prediction techniques initialised around the origin.

We are interested in QSAR-prediction functions,

$$f : \mathcal{D} \to \mathbb{R},$$

that can map a chemical structure $s \in \mathcal{D}$ to an estimate of its binding affinity $a(s)$. The mapping f is found via an algorithmic training process on the labelled data set

$$\{(s, a(s)) \mid s \in \mathcal{D}_{\text{train}}\}$$

and can then either be used to predict the activity labels of compounds in $\mathcal{D}_{\text{test}}$, or it can be repurposed to classify whether an MMP forms an activity cliff (AC-classification) and what the potency direction of an MMP is (PD-classification).

If $\{s, \tilde{s}\} \in \mathcal{M}_{\text{inter}}$, then one can assume that the activity label of one of the compounds, say $a(s)$, is known; f is then used to classify $\{s, \tilde{s}\}$ via:

$$\{s, \tilde{s}\} \mapsto \begin{cases} \text{Non-AC} & \text{if } |a(s) - f(\tilde{s})| \leq d_{\text{crit}}, \\ \text{AC} & \text{if } |a(s) - f(\tilde{s})| > d_{\text{crit}}. \end{cases}$$

Here $d_{\text{crit}} \in \mathbb{R}_{>0}$ is a critical threshold above which an MMP is classified as an AC. Throughout this work we use $d_{\text{crit}} = 1.5$ (in pK_i- or pIC_{50} units) since this value represents the middle point between the intervals [0, 1] and $[2, \infty)$ which correspond to absolute activity-label differences associated with non-ACs and ACs respectively. If $\{s, \tilde{s}\} \in \mathcal{M}_{\text{test}} \cup \mathcal{M}_{\text{cores}}$ then the activities of both compounds are unknown and we classify $\{s, \tilde{s}\}$ via:

$$\{s, \tilde{s}\} \mapsto \begin{cases} \text{Non-AC} & \text{if } |f(s) - f(\tilde{s})| \leq d_{\text{crit}}, \\ \text{AC} & \text{if } |f(s) - f(\tilde{s})| > d_{\text{crit}}. \end{cases}$$

PD-classification for MMPs is performed in a straightforward manner: the activity labels of both MMP-compounds are predicted via f and then compared to classify which compound is the more active one. Since the employed AC- and PD-prediction strategies are based solely on activity predictions for individual compounds in MMPs, they can be generated entirely on the basis of single-molecule representations and do not require an additional representation method for compound pairs.

The performance of f for standard QSAR prediction in $\mathcal{D}_{\text{test}}$ is measured via the mean absolute error (MAE). Note that the MAE is measured over all compounds in $\mathcal{D}_{\text{test}}$ and not just over compounds involved in MMPs or ACs. For the balanced PD-classification problem we rely on accuracy as a suitable performance measure. For the

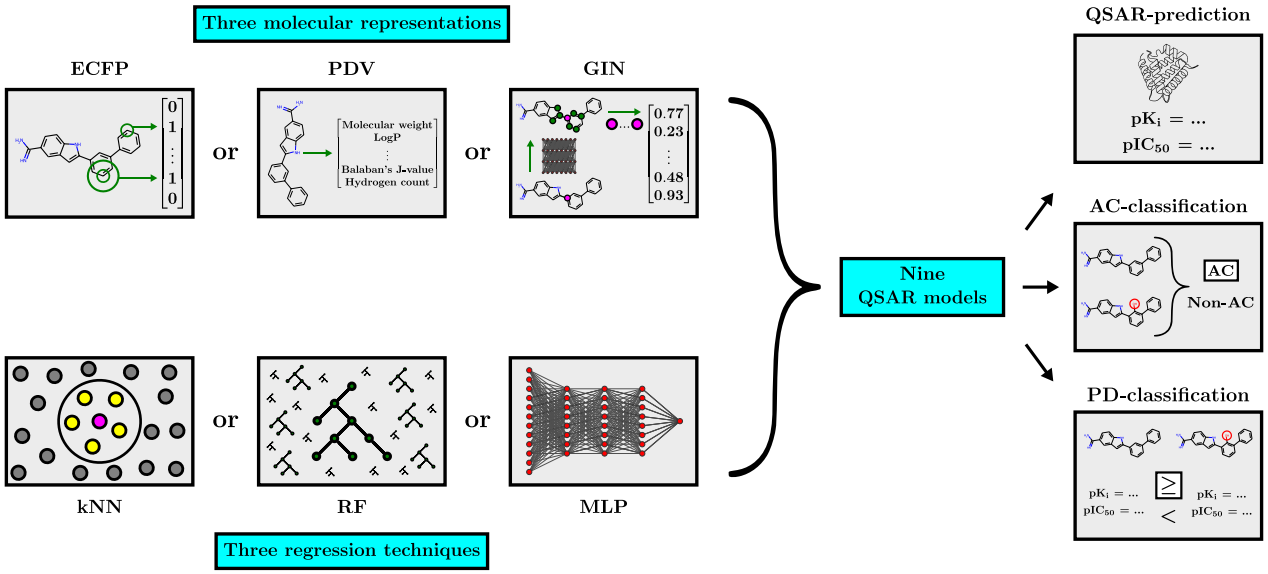


Fig. 3 Schematic showing the combinatorial experimental methodology used for the study. Each molecular representation method is systematically combined with each regression technique, giving a total of nine QSAR models. Each QSAR model is trained and evaluated for QSAR-prediction, AC-classification and PD-classification within a 2-fold cross validation scheme repeated with 3 random seeds. For each of the $2 * 3 = 6$ trials, an extensive inner hyperparameter-optimisation loop on the training set is performed for each QSAR model

highly imbalanced task of AC-classification, however, we use the Matthews correlation coefficient (MCC), as well as sensitivity and precision. For the relatively small SARS-CoV-2 main protease data set we sometimes encountered the edge case where there were no positive predictions; we then set MCC = 0 and ignored ill-defined precision measurements when averaging the performance metrics to obtain the final results.

Molecular representation- and regression techniques

We constructed nine QSAR models via a robust combinatorial methodology that systematically combines three molecular representation methods with three regression techniques. This setup allows, for example, to compare the performance of molecular representation methods across regression techniques, data sets and predictions tasks.

For molecular representation, we used extended-connectivity fingerprints [47] (ECFPs), physicochemical molecular descriptor vectors [58] (PDVs), and graph isomorphism networks (GINs) [66]. Both ECFPs and PDVs were computed via RDKit [32]. The ECFPs were chosen to use a radius of two, a length of 2048 bits, and active chirality flags. The PDVs had a dimensionality of 200 and were constructed using the general list of descriptors from the work of Fabian et al. [15]. This list encompasses properties related to druglikeness, logP, molecular refractivity, electrotopological state, molecular graph-structure, fragment profile, charge, and topological surface properties. The GIN was implemented using PyTorch Geometric [16] and consisted of a variable number of graph convolutional layers, each with two internal hidden layers with ReLU activations and batch normalisation [25]. We further chose the maxpool operator which computes the component-wise maximum over all atom feature vectors in the final graph layer to obtain a graph-level representation.

Each molecular representation was used as an input featurisation for three regression techniques: random forests (RFs), k-nearest neigbours (kNNs) and multilayer perceptrons (MLPs). The RF- and kNN-models were implemented via scikit-learn [45] and the MLP-models via PyTorch [44]. The MLPs used ReLU activations and batch normalisation at each hidden layer.

The GIN was combined with the regression techniques as follows: For MLP regression, the GIN was trained with the MLP as a projection head after the pooling step in the usual end-to-end manner. For RF- or kNN-regression, the GIN was first trained with a single linear layer added after the global pooling step that directly mapped the graph-level representation to an activity prediction. After this training phase the weights of the GIN were frozen and it was used as a static feature extractor. The RF- or kNN-regressor was then trained on the features extracted by the frozen GIN. Figure 3 illustrates our combinatorial experimental methodology.

Model training and hyperparameter optimisation

All models were trained using full inner hyperparameter-optimisation loops. Hyperparameters of RFs and kNNs were optimised in scikit-learn [45] by uniformly random sampling of hyperparameters from a predefined grid. The hyperparameters of MLPs and GINs were sampled from a predefined grid via the tree-structured Parzen estimator algorithm implemented in Optuna [2]. Deep learning models were trained for 500 epochs on a single NVIDIA GeForce RTX 3060 GPU via the mean squared error loss function using AdamW optimisation [36]. Weight decay, learning rate decay and dropout [52] were employed at all hidden layers for regularisation. Batch size, learning rate, learning rate decay rate, weight decay rate, and dropout rate were treated as hyperparameters and subsequently optimised. Note that the training length (i.e. the number of gradient updates) was implicitly optimised by tuning the batch size for the fixed number of 500 training epochs. Further implementation details can be found in our public code repository.[1]

Results and discussion

The QSAR-prediction-, AC-classification- and PD-classification results for all three data sets are depicted in Figs. 4, 5, 6, 7, 8 and 9.

QSAR-prediction performance

When considering the results depicted in Figs. 4, 5, 6, 7, 8 and 9 with respect to QSAR-prediction performance, one can see that ECFPs tend to lead to better performance (i.e. a lower QSAR-MAE) compared to GINs, which in turn tend to lead to better performance compared to PDVs. In particular, the combination MLP-ECFP consistently produced the lowest QSAR-MAE across all three targets. These observations reinforce a growing corpus of literature that suggests that trainable GNNs have not yet reached a level of technical maturity by which they consistently and definitively outperform the much simpler task-agnostic ECFPs at important molecular property prediction tasks [8, 28, 38, 40, 48, 53, 65].

AC-classification performance

The AC-MCC plots in Figs. 4, 5, 6 reveal surprisingly strong overall AC-classification results on $\mathcal{M}_{\text{inter}}$. This type of MMP-set models a compound-optimisation scenario where a researcher strives to identify small structural modifications with a large impact on the activity of query compounds with known activities. For this task, a significant portion of our QSAR models exhibit an AC-MCC value greater than 0.5 across targets, which appears impressive considering the simplicity of the approach. Exchanging $\mathcal{M}_{\text{inter}}$ with either $\mathcal{M}_{\text{test}}$ or $\mathcal{M}_{\text{cores}}$ leads to a substantial drop in the AC-MCC to approximately 0.3 that appears to be mediated by a large drop in AC-sensitivity.

In most cases, GINs perform better than the other molecular representation methods with respect to the AC-MCC. Notably, the combination GIN-kNN consistently performs considerably better for AC-classification than the combinations ECFP-kNN and PDV-kNN. This supports the idea that GINs might have a heightened ability to resolve ACs by learning an embedding of chemical space in which the distance between two compounds is reflective of activity difference rather than structural difference. The combinations GIN-MLP, GIN-RF and ECFP-MLP exhibit particularly high AC-MCC values relative to the other methods. We recommend using at least one of these three models as a baseline against which to compare tailored AC-prediction models; the practical utility of any AC-prediction technique that cannot outperform these three common QSAR methods is questionable.

Across all three targets, AC-sensitivity is moderately high on $\mathcal{M}_{\text{inter}}$ but universally low on $\mathcal{M}_{\text{test}}$ and $\mathcal{M}_{\text{cores}}$. This is consistent with the hypothesis that ACs form one of the major sources of prediction error for QSAR models. The weak AC-sensitivity on $\mathcal{M}_{\text{test}}$ and $\mathcal{M}_{\text{cores}}$ indicates that modern QSAR methods are largely blind to ACs formed by two MMP-compounds outside the training set and thus lack essential chemical knowledge. GINs clearly outperform the other two more classical molecular representations across regression techniques with respect to AC-sensitivity. In particular, the GIN-MLP combination leads to the highest AC-sensitivity in all examined cases and thus discovers the most ACs. The highly parametric nature of GINs that makes them prone to overfitting could at the same time enable them to better model jagged regions of the SAR-landscape that contain ACs than classical task-agnostic representations.

There is a wide gap between distinct prediction techniques with respect to AC-precision: some models achieve a considerable level of AC-precision such that over 50% of positively predicted MMPs in $\mathcal{M}_{\text{test}}$ and $\mathcal{M}_{\text{cores}}$ are indeed actual ACs. Other QSAR models, however, seem to fail almost entirely with respect to this metric on $\mathcal{M}_{\text{test}}$ and $\mathcal{M}_{\text{cores}}$ and only deliver modest performance on $\mathcal{M}_{\text{inter}}$. RFs tend to exhibit the strongest AC-precision and the weakest AC-sensitivity. This might be as a result of their ensemble nature which should intuitively lead to conservative but trustworthy predictions of extreme effects such as ACs.

[1] https://github.com/MarkusFerdinandDablander/QSAR-activity-cliff-experiments.

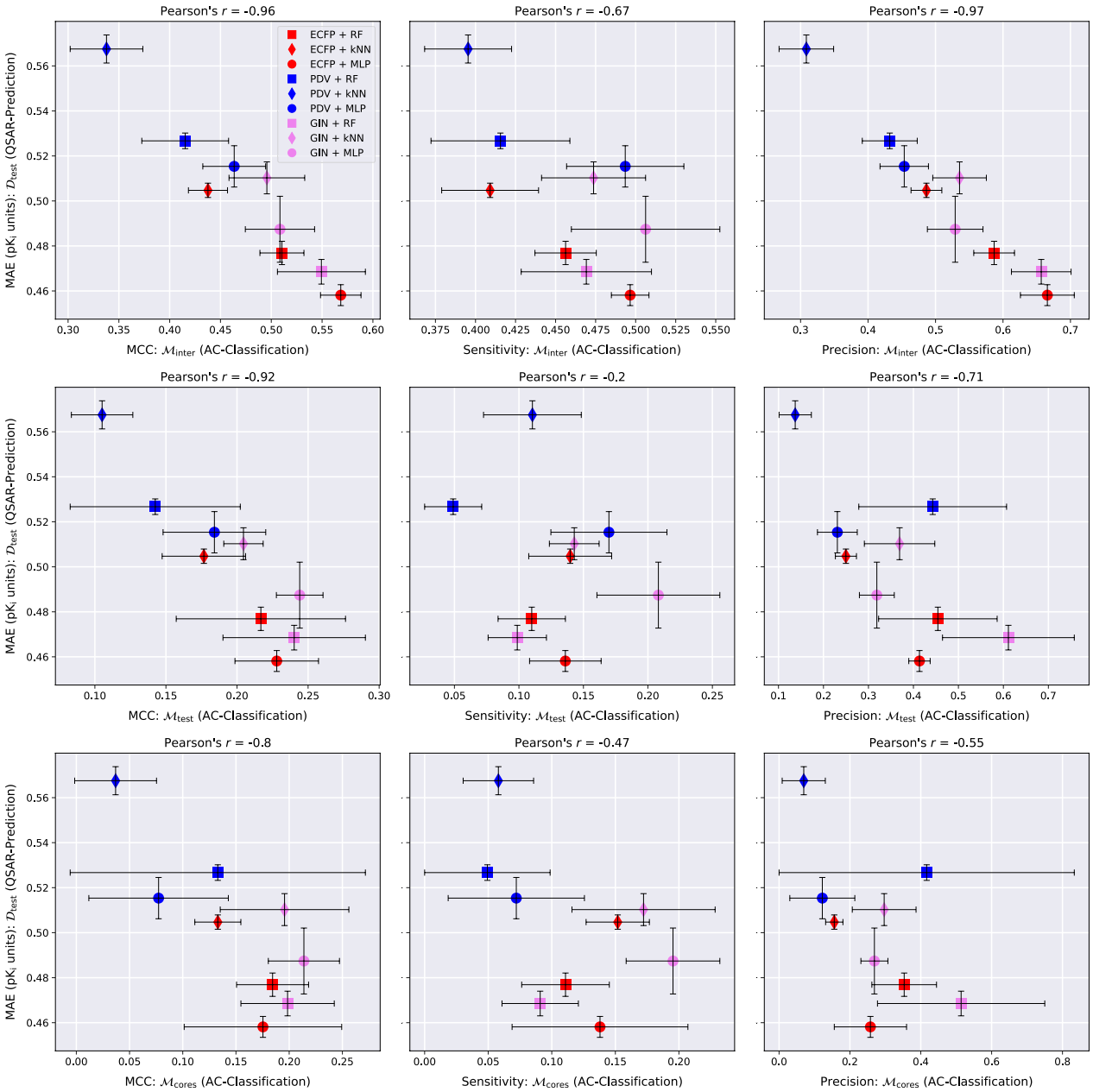


Fig. 4 QSAR-prediction- and AC-classification results for **dopamine receptor D2**. For each plot, the *x*-axis corresponds to a combination of MMP-set and AC-classification performance metric and the *y*-axis shows the QSAR-prediction performance on the molecular test set $\mathcal{D}_{\text{test}}$. The total length of each error bar equals twice the standard deviation of the performance metric measured over all $mk = 3 * 2 = 6$ hyperparameter-optimised models. For each plot, the lower right corner corresponds to strong performance at both prediction tasks

PD-classification performance

The abilities of the evaluated QSAR models to identify which is the more active compound in an MMP is universally weak, with PD-accuracies clustering around 0.7 on $\mathcal{M}_{\text{inter}}$ and around 0.6 on $\mathcal{M}_{\text{test}}$ and $\mathcal{M}_{\text{cores}}$, as can be seen in the top rows of Figs. 7, 8, 9. Predicting the potency direction for two compounds with similar structures and thus usually similar activity levels must be considered a challenging task. The combination ECFP-MLP reaches the strongest PD-accuracy in the majority of cases and we recommend starting with this model as a baseline for more advanced PD-prediction methods.

One can argue that the activity order of two similar compounds is of little interest if the true activity difference is small, as is often the case. We therefore also

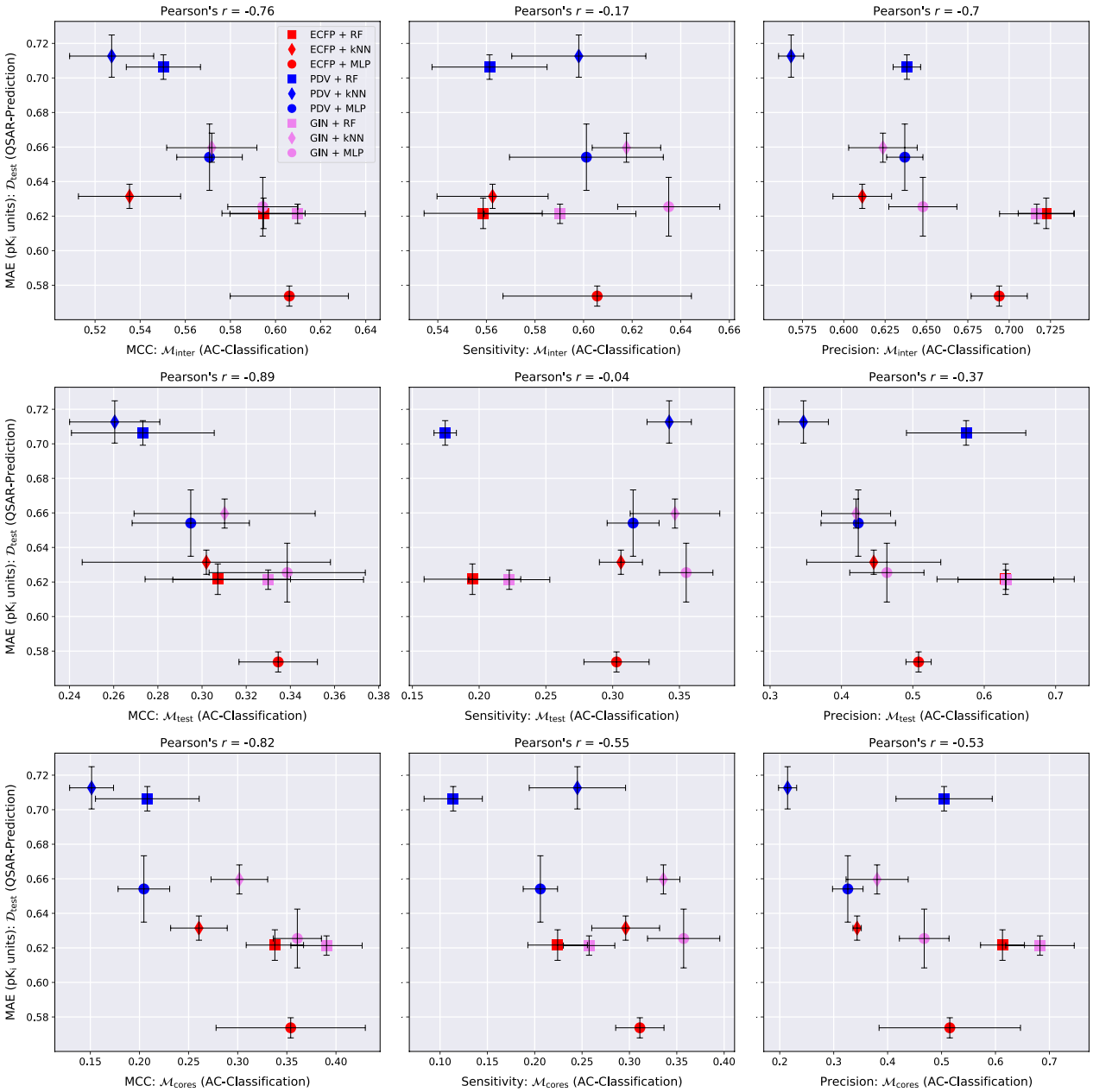


Fig. 5 QSAR-prediction- and AC-classification results for **factor Xa**. For each plot, the *x*-axis corresponds to a combination of MMP-set and AC-classification performance metric and the *y*-axis shows the QSAR-prediction performance on the molecular test set $\mathcal{D}_{\text{test}}$. The total length of each error bar equals twice the standard deviation of the performance metric measured over all $mk = 3 * 2 = 6$ hyperparameter-optimised models. For each plot, the lower right corner corresponds to strong performance at both prediction tasks

restricted PD-classification to predicted ACs. The three plots in the bottom rows of Figs. 7, 8, 9 depict the PD-accuracy of each QSAR model on the subset of MMPs that were also predicted to be ACs by the same model. In this practically more relevant scenario PD-prediction accuracy tends to exceed 0.9 on $\mathcal{M}_{\text{inter}}$ and 0.8 on $\mathcal{M}_{\text{test}}$ and $\mathcal{M}_{\text{cores}}$. The QSAR models investigated here are thus able to identify the correct activity order of MMPs if they also predict them to be ACs. The relatively rare instances in which the PD of a predicted AC is misclassified, however, reflect severe QSAR-prediction errors.

Linear relationship between QSAR-MAE and AC-MCC

Our experiments reveal a consistent linear relationship between the QSAR-MAE and the AC-MCC as can be seen in the left columns of Figs. 4, 5, 6. A potential

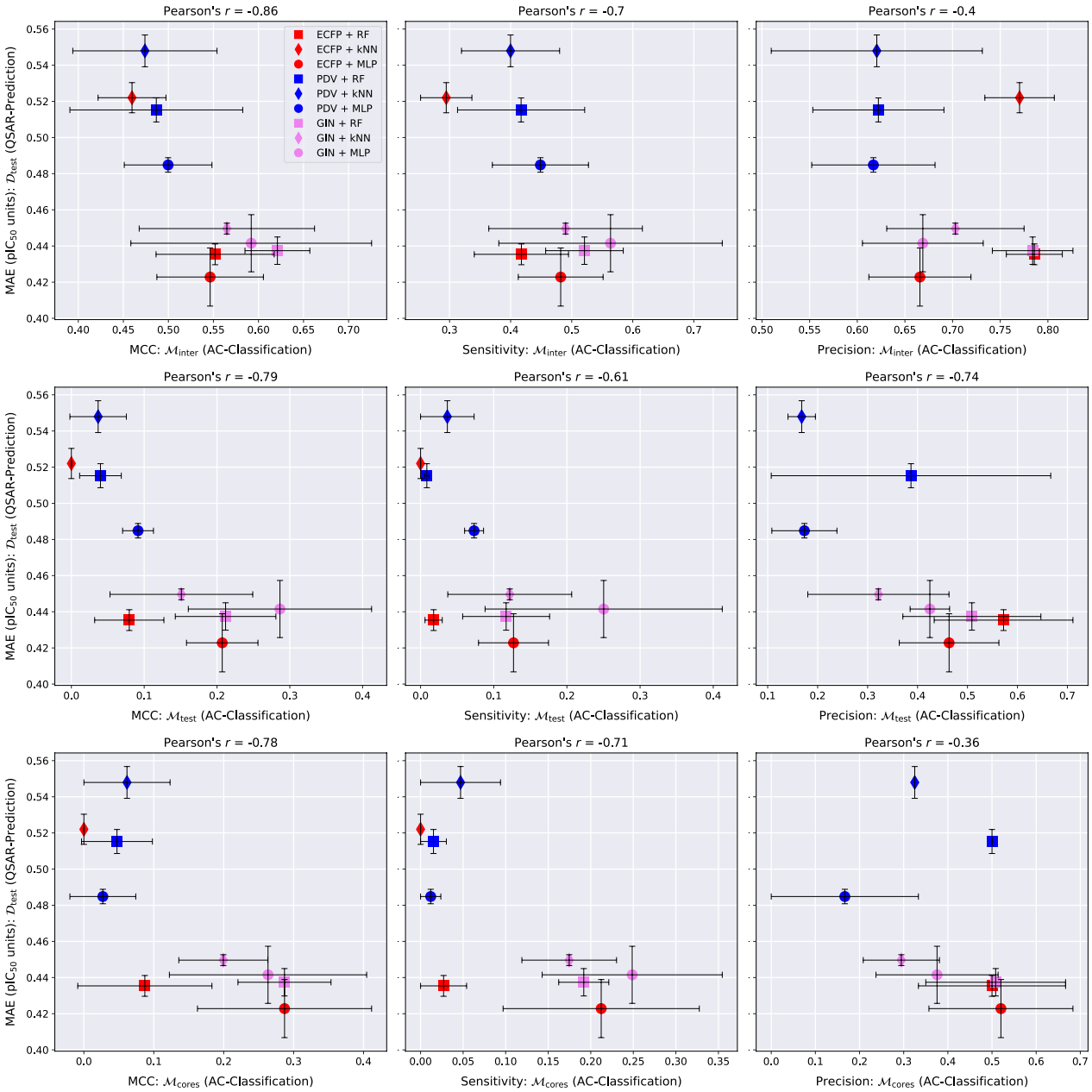


Fig. 6 QSAR-prediction- and AC-classification results for **SARS CoV-2 main protease**. For each plot, the *x*-axis corresponds to a combination of MMP-set and AC-classification performance metric and the *y*-axis shows the QSAR-prediction performance on the molecular test set $\mathcal{D}_{\text{test}}$. The total length of each error bar equals twice the standard deviation of the performance metric measured over all $mk = 3 * 2 = 6$ hyperparameter-optimised models. The precision of the AC-classification task is lacking for the ECFP + kNN technique on $\mathcal{M}_{\text{test}}$ and $\mathcal{M}_{\text{cores}}$ since this method produced only negative AC-predictions for all trials on this data set. For each plot, the lower right corner corresponds to strong performance at both prediction tasks

mechanism driving this effect could be that as the overall QSAR-MAE of a model improves, its accuracy at predicting activity differences between similar molecules might be expected to improve as well; previously misclassified MMPs whose predicted absolute activity differences were already close to the critical value $d_{\text{crit}} = 1.5$ might then gradually move to the correct side of the decision boundary and increase the AC-MCC. The results suggest that for real-world QSAR models the AC-MCC and the QSAR-MAE are strongly

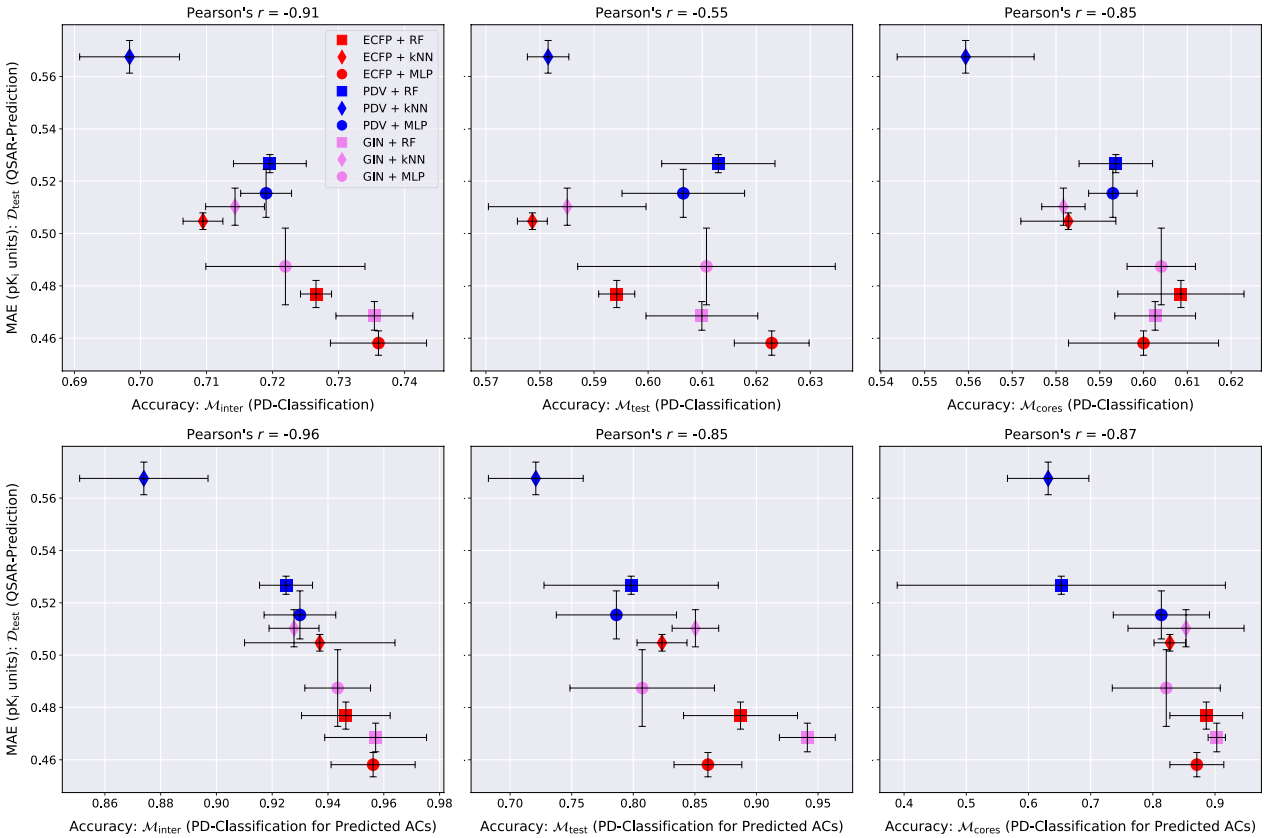


Fig. 7 QSAR-prediction- and PD-classification results for **dopamine receptor D2**. Each column corresponds to an upper plot and a lower plot for one of the MMP-sets $\mathcal{M}_{inter}$, $\mathcal{M}_{test}$ or $\mathcal{M}_{cores}$. The x-axis of each upper plot indicates the PD-classification accuracy on the full MMP-set; the x-axis of each lower plot indicates the PD-classification accuracy on a restricted MMP-set only consisting of MMP predicted to be ACs by the respective method. The *y*-axis of each plot shows the QSAR-prediction performance on the molecular test set $\mathcal{D}_{test}$. The total length of each error bar equals twice the standard deviation of the performance metrics measured over all $mk = 3 * 2 = 6$ hyperparameter-optimised models. For each plot, the lower right corner corresponds to strong performance at both prediction tasks

predictive of each other; while this observation only rests on nine models, it is highly consistent across MMP-sets and pharmacological targets.

Conclusions

To the best of our knowledge this is the first study to investigate the capabilities of QSAR models to classify between ACs and non-ACs. It is also the first work to explore the quantitative relationship between QSAR-prediction at the level of individual molecules and AC-prediction at the level of compound-pairs. As part of our methodology we have additionally introduced a simple, interpretable, and rigorous data-splitting technique for pair-based prediction problems.

When the activities of both MMP-compounds are unknown (i.e. absent from the training set) then common QSAR models exhibit low AC-sensitivity which limits their utility for AC-prediction. This strongly supports the hypothesis that QSAR methods do indeed regularly fail to predict ACs which might thus form a major source of prediction errors in QSAR modelling [9, 18, 37, 50]. However, if the activity of one MMP-compound is known (i.e., present in the training set) then AC-sensitivity increases substantially; for query compounds with known activities, QSAR methods can therefore be used as simple AC-prediction-, compound-optimisation- and SAR-knowledge-acquisition tools. Furthermore, based on the observed potency-direction (PD) classification

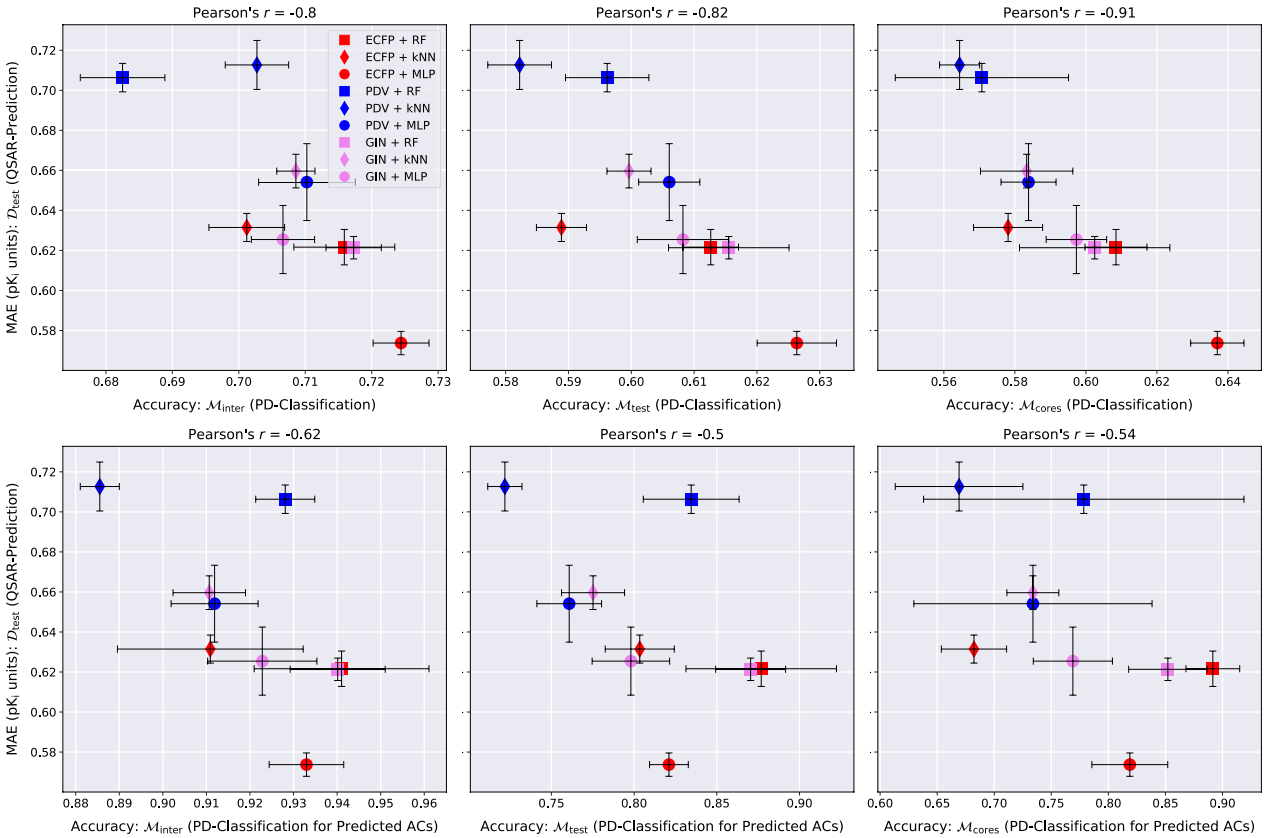


Fig. 8 QSAR-prediction- and PD-classification results for **factor Xa**. Each column corresponds to an upper plot and a lower plot for one of the MMP-sets $\mathcal{M}_{inter}$, $\mathcal{M}_{test}$ or $\mathcal{M}_{cores}$. The x-axis of each upper plot indicates the PD-classification accuracy on the full MMP-set; the x-axis of each lower plot indicates the PD-classification accuracy on a restricted MMP-set only consisting of MMP predicted to be ACs by the respective method. The *y*-axis of each plot shows the QSAR-prediction performance on the molecular test set $\mathcal{D}_{test}$. The total length of each error bar equals twice the standard deviation of the performance metrics measured over all $mk = 3 * 2 = 6$ hyperparameter-optimised models. For each plot, the lower right corner corresponds to strong performance at both prediction tasks

results we can expect the estimated activity direction of predicted ACs to have a high degree of accuracy.

With respect to molecular representation, we have found robust evidence that precomputed task-agnostic ECFPs still outcompete differentiable GINs at general QSAR-prediction. This adds to a growing awareness that standard message-passing GNNs might need to be improved further to definitively beat classical molecular featurisations such as ECFPs [8, 28, 38, 40, 48, 53, 65]. One potential angle to achieve this could be self-supervised GNN-pretraining, which has recently shown promising results in the molecular domain [22, 63]. However, while GINs appear to be inferior to ECFPs for QSAR-prediction, they tend to be advantageous for AC-classification; their highly parametric nature might simultaneously lead to increased overfitting but to a better modelling of the more jagged regions of the SAR-landscape. We thus recommend using GINs as an AC-classification baseline since such an agreed-upon baseline is currently lacking.

The low AC-sensitivity of the tested QSAR models when the activities of both MMP-compounds are unknown suggests that such methods are still lacking essential SAR knowledge. On the flip side, it might be possible to considerably boost the performance of common QSAR models by developing techniques to increase their AC-sensitivity which could potentially provide a fruitful direction for future research.

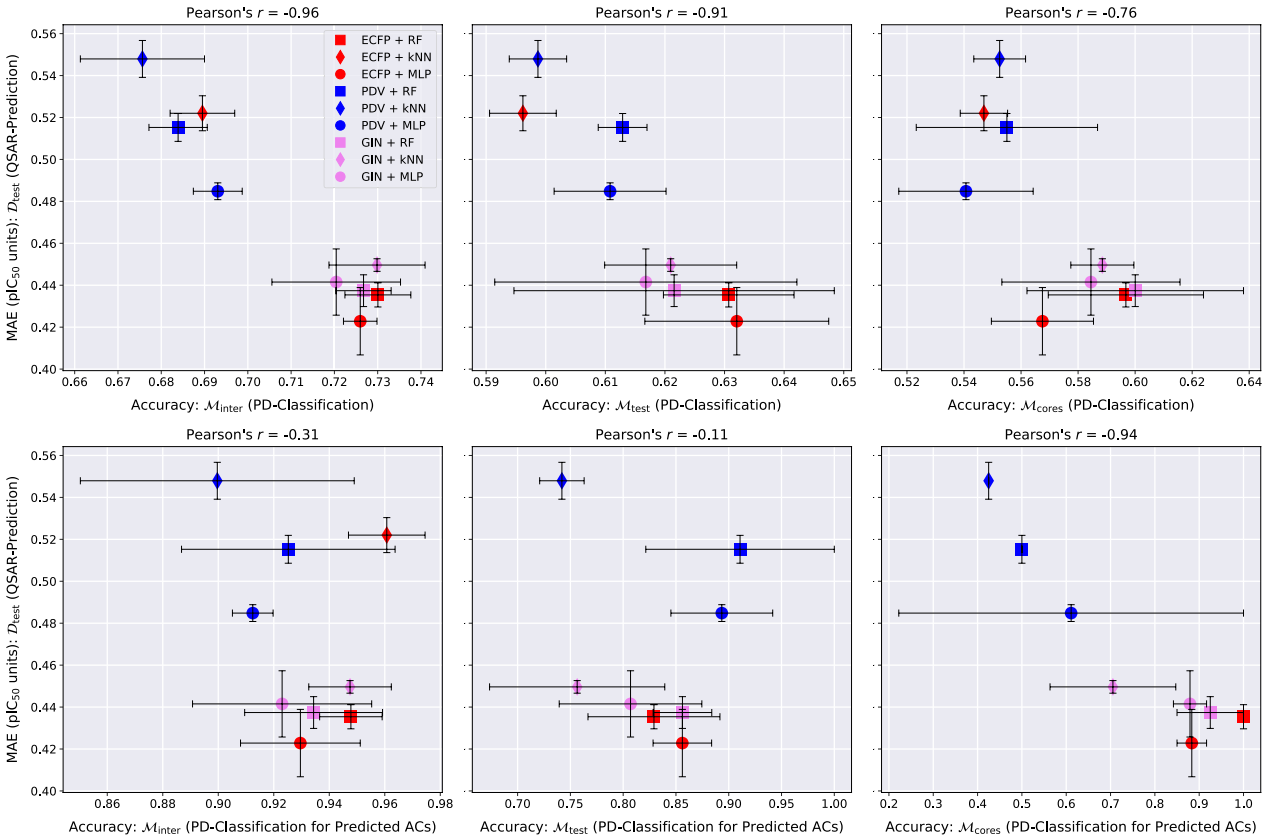


Fig. 9 QSAR-prediction- and PD-classification results for **SARS-CoV-2 main protease**. Each column corresponds to an upper plot and a lower plot for one of the MMP-sets $\mathcal{M}_{\text{inter}}$, $\mathcal{M}_{\text{test}}$ or $\mathcal{M}_{\text{cores}}$. The x-axis of each upper plot indicates the PD-classification accuracy on the full MMP-set; the x-axis of each lower plot indicates the PD-classification accuracy on a restricted MMP-set only consisting of MMP predicted to be ACs by the respective method. The *y*-axis of each plot shows the QSAR-prediction performance on the molecular test set $\mathcal{D}_{\text{test}}$. The total length of each error bar equals twice the standard deviation of the performance metrics measured over all $mk = 3 * 2 = 6$ hyperparameter-optimised models. The accuracy of the PD-classification task for predicted ACs is lacking for the ECFP + kNN technique on $\mathcal{M}_{\text{test}}$ and $\mathcal{M}_{\text{cores}}$ since this method produced only negative AC-predictions for all trials on this data set. For each plot, the lower right corner corresponds to strong performance at both prediction tasks

Abbreviations

AC	Activity cliff
ECFP	Extended-connectivity fingerprint
GIN	Graph isomorphism network
GNN	Graph neural network
kNN	k-nearest neighbour
MAE	Mean absolute error
MCC	Matthews correlation coefficient
MLP	Multilayer perceptron
MMP	Matched molecular pair
PD	Potency direction
PDV	Physicochemical-descriptor vector
QSAR	Quantitative structure-activity relationship
RF	Random forest
SAR	Structure-activity relationship

Author contributions

The computational study was designed, implemented, conducted and interpreted by the first author MD. The research was supervised by GMM, RL, and TH who gave valuable scientific advice during weekly meetings. The computer code was written by MD. The paper manuscript was written by MD. Feedback was provided by RL, GMM and TH during the writing process. The novel data splitting technique for MMP-data and the QSAR-modelling-based activity cliff prediction strategies were developed by MD. All scientific figures were designed by MD, with input from GMM, RL and TH. All chemical data sets were gathered and cleaned by MD. All authors read and approved the final manuscript.

Funding

This research was supported by the University of Oxford's UK EPSRC Centre For Doctoral Training in Industrially Focused Mathematical Modelling (EP/L015803/1) and by the not-for-profit organisation and educational charity Lhasa Limited (https://www.lhasalimited.org/).

Availability of data and materials

All used data sets, the code to reproduce and visualise the experimental results, and the exact numerical results generated by the original experiments are available in our public code repository https://github.com/MarkusFerdinandDablander/QSAR-activity-cliff-experiments.

Declarations

Competing interests

The authors declare that they have no competing interests.

Author details
[1]Mathematical Institute, University of Oxford, Andrew Wiles Building, Radcliffe Observatory Quarter (550), Woodstock Road, Oxford OX2 6GG, UK. [2]Lhasa Limited, Granary Wharf House, 2 Canal Wharf, Leeds LS11 5PS, UK. [3]Department of Statistics, University of Oxford, 24–29 St Giles', Oxford OX1 3LB, UK.

Received: 26 January 2023 Accepted: 10 March 2023
Published online: 17 April 2023

References

1. Achdout H, Aimon A, Bar-David E, Barr H, Ben-Shmuel A, Bennett J, Bilenko VA, Bilenko VA, Boby ML, Borden B, Bowman GR, Brun J, et al (2022) Open science discovery of oral non-covalent SARS-CoV-2 main protease inhibitor therapeutics. BioRxiv. https://www.biorxiv.org/content/early/2022/01/30/2020.10.29.339317. Accessed 19 Jan 2023
2. Akiba T, Sano S, Yanase T, Ohta T, Koyama M (2019) Optuna: a next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp 2623–2631
3. Asawa Y, Yoshimori A, Bajorath J, Nakamura H (2020) Prediction of an MMP-1 inhibitor activity cliff using the SAR matrix approach and its experimental validation. Sci Rep 10(1):14710
4. Bajorath J (2014) Exploring activity cliffs from a chemoinformatics perspective. Mol Inf 33(6–7):438–442
5. Beck JM, Springer C (2014) Quantitative structure-activity relationship models of chemical transformations from matched pairs analyses. J Chem Inf Model 54(4):1226–1234
6. Bento AP, Hersey A, Félix E, Landrum G, Gaulton A, Atkinson F, Bellis LJ, de Veij M, Leach AR (2020) An open source chemical structure curation pipeline using RDKit. J Cheminformatics 12(1):1–16
7. Chen H, Vogt M, Bajorath J (2022) DeepAC - conditional transformer-based chemical language model for the prediction of activity cliffs formed by bioactive compounds. Dig Discov 1:898–909
8. Chithrananda S, Grand G, Ramsundar B (2020) ChemBERTa: large-scale self-supervised pretraining for molecular property prediction. http://arxiv.org/abs/2010.09885
9. Cruz-Monteagudo M, Medina-Franco JL, Pérez-Castillo Y, Nicolotti O, Cordeiro MNDS, Borges F (2014) Activity cliffs in drug discovery: Dr Jekyll or Mr Hyde? Drug Discov Today 19(8):1069–1080
10. Cruz-Monteagudo M, Medina-Franco L, J, Perera-Sardiña Y, Borges F, Tejera E, Paz-y Mino C, Pérez-Castillo Y, Sánchez-Rodríguez A, Contreras-Posada Z, Cordeiro ND, (2016) Probing the hypothesis of SAR continuity restoration by the removal of activity cliffs generators in QSAR. Curr Pharm Des 22(33):5043–5056
11. Dablander M, Lambiotte R, Morris GM, Hanser T (2021) Siamese neural networks work for activity cliff prediction. In: Poster presented at the 4th RSC-BMCS / RSC-CICAG Artificial Intelligence in Chemistry Symposium. https://www.researchgate.net/publication/362875964_Siamese_Neural_Networks_Work_for_Activity_Cliff_Prediction. Accessed 19 Jan 2023
12. Dalke A, Hert J, Kramer C (2018) mmpdb: an open-source matched molecular pair platform for large multiproperty data sets. J Chem Inf Model 58(5):902–910
13. Dimova D, Stumpfe D, Hu Y, Bajorath J (2015) Activity cliff clusters as a source of structure-activity relationship information. Expert Opin Drug Discov 10(5):441–447
14. Duvenaud DK, Maclaurin D, Iparraguirre J, Bombarell R, Hirzel T, Aspuru-Guzik A, Adams RP (2015) Convolutional networks on graphs for learning molecular fingerprints. In: Advances in Neural Information Processing Systems, pp 2224–2232
15. Fabian B, Edlich T, Gaspar H, Segler M, Meyers J, Fiscato M, Ahmed M (2020) Molecular representation learning with language models and domain-relevant auxiliary tasks. http://arxiv.org/abs/2011.13230
16. Fey M, Lenssen JE (2019) Fast graph representation learning with PyTorch Geometric. http://arxiv.org/abs/1903.02428
17. Gilmer J, Schoenholz SS, Riley PF, Vinyals O, Dahl GE (2017) Neural message passing for quantum chemistry. In: International Conference on Machine Learning, PMLR, pp 1263–1272
18. Golbraikh A, Muratov E, Fourches D, Tropsha A (2014) Data set modelability by QSAR. J Chem Inf Model 54(1):1–4
19. Heikamp K, Hu X, Yan A, Bajorath J (2012) Prediction of activity cliffs using support vector machines. J Chem Inf Model 52(9):2354–2365
20. Hoonakker F, Lachiche N, Varnek A, Wagner A (2011) Condensed graph of reaction: considering a chemical reaction as one single pseudo molecule. Int J Artif Intell Tools 20(2):253–270
21. Horvath D, Marcou G, Varnek A, Kayastha S, de la Vega de León A, Bajorath J, (2016) Prediction of activity cliffs using condensed graphs of reaction representations. J Chem Inf Model 56(9):1631–1640
22. Hu W, Liu B, Gomes J, Zitnik M, Liang P, Pande V, Leskovec J (2019) Strategies for pre-training graph neural networks. http://arxiv.org/abs/1905.12265
23. Hu Y, Bajorath J (2012) 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 52(7):1806–1811
24. Husby J, Bottegoni G, Kufareva I, Abagyan R, Cavalli A (2015) Structure-based predictions of activity cliffs. J Chem Inf Model 55(5):1062–1076
25. Ioffe S, Szegedy C (2015) Batch normalization: accelerating deep network training by reducing internal covariate shift. In: Proceedings of Machine Learning Research, pp 448–456
26. Iqbal J, Vogt M, Bajorath J (2021) Prediction of activity cliffs on the basis of images using convolutional neural networks. J Comput Aided Mol Des 35:1157–1164
27. Jauffret P, Tonnelier C, Hanser T, Kaufmann G, Wolff R (1990) Machine learning of generic reactions: 2. Toward an advanced computer representation of chemical reactions. Tetrahedron Comput Methodol 3(6):335–349
28. Jiang D, Wu Z, Hsieh CY, Chen G, Liao B, Wang Z, Shen C, Cao D, Wu J, Hou T (2021) 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):1–23
29. Kenny PW, Sadowski J (2005) Structure modification in chemical databases. Chemoinformatics Drug Discov 23:271–285
30. Keyvanpour MR, Barani Shirzad M, Moradi F (2021) PCAC: a new method for predicting compounds with activity cliff property in QSAR approach. Int J Inf Technol 13(6):2431–2437
31. Kipf TN, Welling M (2016) Semi-supervised classification with graph convolutional networks. arXiv preprint. https://arxiv.org/abs/1609.02907 [cs.LG]
32. Landrum G (2006) RDKit: open-source cheminformatics
33. Leadley J (2001) Coagulation factor Xa inhibition: biological background and rationale. Curr Top Med Chem 1(2):151–159
34. la Vega De, de León A, Bajorath J (2014) Prediction of compound potency changes in matched molecular pairs using support vector regression. J Chem Inf Model 54(10):2654–2663
35. Liu T, Lin Y, Wen X, Jorissen RN, Gilson MK (2007) BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. Nucleic Acids Res 35:D198–D201
36. Loshchilov I, Hutter F (2017) Decoupled weight decay regularization. http://arxiv.org/abs/1711.05101
37. Maggiora GM (2006) On outliers and activity cliffs: why QSAR often disappoints. J Chem Inf Model 46(4):1535–1535
38. Mayr A, Klambauer G, Unterthiner T, Steijaert M, Wegner JK, Ceulemans H, Clevert DA, Hochreiter S (2018) Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. Chem Sci 9(24):5441–5451
39. Medina-Franco JL (2013) Activity cliffs: facts or artifacts? Chem Biol Drug Design 81(5):553–556
40. Menke J, Koch O (2021) Using domain-specific fingerprints generated through neural networks to enhance ligand-based virtual screening. J Chem Inf Model 61(2):664–675
41. Namasivayam V, Bajorath J (2012) Searching for coordinated activity cliffs using particle swarm optimization. J Chem Inf Model 52(4):927–934
42. Namasivayam V, Iyer P, Bajorath J (2013) Prediction of individual compounds forming activity cliffs using emerging chemical patterns. J Chem Inf Model 53(12):3131–3139

43. Park J, Sung G, Lee S, Kang S, Park C (2022) ACGCN: graph convolutional networks for activity cliff prediction between matched molecular pairs. J Chem Inf Model 62(10):2341–2351. https://doi.org/10.1021/acs.jcim.2c00327
44. 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 (2019) PyTorch: an imperative style, high-performance deep learning library. In: Wallach H, Larochelle H, Beygelzimer A, d' Alché-Buc F, Fox E, Garnett R (eds) Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 32. https://proceedings.neurips.cc/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf. Accessed 19 Jan 2023
45. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V et al (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825–2830
46. Pérez-Benito L, Casajuana-Martin N, Jiménez-Rosés M, van Vlijmen H, Tresadern G (2019) Predicting activity cliffs with free-energy perturbation. J Chem Theory Comput 15(3):1884–1895
47. Rogers D, Hahn M (2010) Extended-connectivity fingerprints. J Chem Inf Model 50(5):742–754
48. Sabando MV, Ponzoni I, Milios EE, Soto AJ (2021) Using molecular embeddings in QSAR modeling: does it make a difference? http://arxiv.org/abs/2104.02604
49. Seeman P (1987) Dopamine receptors and the dopamine hypothesis of schizophrenia. Synapse 1(2):133–152
50. Sheridan RP, Karnachi P, Tudor M, Xu Y, Liaw A, Shah F, Cheng AC, Joshi E, Glick M, Alvarez J (2020) Experimental error, kurtosis, activity cliffs, and methodology: what limits the predictivity of quantitative structure-activity relationship models. J Chem Inf Model 60(4):1969–1982
51. Silipo C, Vittoria A (1991) QSAR, rational approaches to the design of bioactive compounds. In: Proceedings of European Symposium on Quantitative Structure-Activity Relationships, Distributors for the US and Canada, Elsevier Science
52. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 15(1):1929–1958
53. Stepišnik T, Škrlj B, Wicker J, Kocev D (2021) A comprehensive comparison of molecular feature representations for use in predictive modeling. Comput Biol Med 130(104):197
54. Stumpfe D, Hu Y, Dimova D, Bajorath J (2014) Recent progress in understanding activity cliffs and their utility in medicinal chemistry: miniperspective. J Med Chem 57(1):18–28
55. Stumpfe D, Hu H, Bajorath J (2019) Evolving concept of activity cliffs. ACS Omega 4(11):14360–14368
56. Stumpfe D, Hu H, Bajorath J (2020) Advances in exploring activity cliffs. J Comput Aided Mol Des 34(9):929–942
57. Tamura S, Miyao T, Funatsu K (2020) Ligand-based activity cliff prediction models with applicability domain. Mol Inform. https://doi.org/10.1002/minf.202000103
58. Todeschini R, Consonni V (2008) Handbook of molecular descriptors. John Wiley & Sons, New York
59. Ullrich S, Nitsche C (2020) The SARS-CoV-2 main protease as drug target. Bioorg Med Chem Lett 30(17):127377
60. Van Tilborg D, Alenicheva A, Grisoni F (2022) Exposing the limitations of molecular machine learning with activity cliffs. ChemRxiv. https://chemrxiv.org/engage/chemrxiv/article-details/623de3fbab0051148698fbcf. Accessed 19 Jan 2023
61. Veličković P, Cucurull G, Casanova A, Romero A, Lio P, Bengio Y (2017) Graph attention networks. http://arxiv.org/abs/1710.10903
62. Vogt M, Huang Y, Bajorath J (2011) From activity cliffs to activity ridges: informative data structures for SAR analysis. J Chem Inf Model 51(8):1848–1856
63. Wang Y, Wang J, Cao Z, Farimani AB (2021) MolCLR: molecular contrastive learning of representations via graph neural networks. http://arxiv.org/abs/2102.10056
64. Winkler DA, Le TC (2017) Performance of deep and shallow neural networks, the universal approximation theorem, activity cliffs, and QSAR. Mol Inform 36(1–2):1600118
65. Winter R, Montanari F, Noé F, Clevert DA (2019) Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. Chem Sci 10(6):1692–1701
66. Xu K, Hu W, Leskovec J, Jegelka S (2018) How powerful are graph neural networks? http://arxiv.org/abs/1810.00826

Publisher's Note