



Available online at
www.heca-analitika.com/hjas

Heca Journal of Applied Sciences

Vol. 4, No. 2, 2026



Activity Cliffs and Machine Learning Generalization in QSAR: Why Similar Molecules Remain Difficult to Predict

Teuku Rizky Noviandy ¹, Mohsina Patwekar ², and Faheem Patwekar ³

¹ Department of Information Systems, Faculty of Engineering, Universitas Abulyatama, Aceh Besar 23372, Indonesia; trizkynoviandy@gmail.com (T.R.N.)

² Department of Pharmacology, College of Medicine, Gulf Medical University, Ajman, United Arab Emirates; mohsina.patwekar@gmail.com (M.P.)

³ Department of Pharmacognosy, Luqman College of Pharmacy, Gulbarga, India; ifaheemp@gmail.com (F.P.)

* Correspondence: rizky_si@abulyatama.ac.id

Article History

Received 18 July 2026

Revised 4 September 2026

Accepted 13 September 2026

Available Online 30 September 2026

Keywords:

Activity cliffs

QSAR

Molecular representation

Drug discovery

Abstract

Machine learning (ML) has substantially expanded the predictive capability of quantitative structure-activity relationship (QSAR) modeling, yet reliable generalization remains difficult in chemically challenging regions. Activity cliffs—pairs of highly similar molecules that exhibit large differences in biological activity—represent a particularly stringent test because they violate the local smoothness assumed by many similarity-based and representation-learning approaches. This narrative review examines why activity cliffs remain difficult for contemporary QSAR models and how they expose limitations that may be concealed by conventional global performance metrics. We discuss the definition, detection, and chemical origins of activity cliffs; their relationship to discontinuities in structure-activity landscapes; and the influence of molecular representation, training-test partitioning, structural redundancy, and data leakage on apparent model performance. Evidence from fingerprint-based ML, deep neural networks, graph neural networks, and recent molecular foundation models indicates that increasing model complexity does not consistently resolve activity-cliff prediction. Models with strong overall QSAR performance may still exhibit substantially larger errors or poor sensitivity for cliff-forming compounds. Recent approaches that explicitly incorporate activity-cliff information through pairwise, contrastive, or activity-aware representation learning provide a promising alternative to conventional compound-level regression. We therefore propose an activity-cliff-aware QSAR perspective integrating rigorous bioactivity curation, activity-landscape characterization, leakage-resistant validation, global and cliff-specific evaluation, applicability-domain analysis, uncertainty quantification, and cliff-aware learning objectives. Activity cliffs should ultimately be regarded not merely as problematic outliers, but as information-rich stress tests of molecular ML generalization and local structure-activity understanding.



Copyright: © 2026 by the authors. This is an open-access article distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License. (<https://creativecommons.org/licenses/by-nc/4.0/>)

1. Introduction

Quantitative structure-activity relationship (QSAR) modeling is founded on the premise that variations in molecular structure can be systematically related to variations in biological activity [1, 2]. Early QSAR

approaches formalized this relationship using relatively small sets of physicochemical descriptors, including hydrophobic, electronic, and steric terms, establishing a quantitative basis for connecting chemical structure with biological response [3]. Over subsequent decades, QSAR

evolved from conventional statistical models toward increasingly flexible machine learning (ML) approaches capable of learning nonlinear relationships from high-dimensional molecular representations [4–8]. Random forests (RF) [9], support vector machines (SVM) [10], gradient-boosting (GB) methods [11], artificial neural networks, and, more recently, graph-based deep-learning architectures have consequently become important components of ligand-based drug discovery.

Benchmark initiatives such as MoleculeNet further accelerated this transition by providing standardized molecular datasets and evaluation protocols for comparing learning algorithms across physicochemical, biological, and pharmacological prediction tasks [12]. Large-scale studies using bioactivity databases such as ChEMBL have also demonstrated that modern ML and deep-learning methods can achieve substantial predictive performance across diverse drug-target prediction problems [13].

Despite these advances, high predictive performance does not necessarily imply that a model has learned a generalizable structure-activity relationship (SAR). A central distinction exists between interpolation within densely sampled chemical space and genuine generalization to compounds whose structural or activity relationships are poorly represented during training [14, 15]. This distinction is especially relevant in medicinal chemistry, where datasets are commonly organized into congeneric series and contain substantial structural redundancy. Randomly allocating molecules between training and test sets can therefore place closely related analogues on both sides of the split, allowing a model to obtain apparently strong predictive performance while relying heavily on local chemical similarity [16].

Prospective and time-split evaluations have shown that random validation may provide optimistic estimates of performance relative to the prediction of genuinely future compounds [14]. Similarly, analyses of ligand-based benchmarks have demonstrated that structural redundancy between training and validation sets can reward memorization rather than generalization [15]. Consequently, the reliability of a QSAR model depends not only on its algorithmic complexity but also on the chemical-space coverage, validation design, and relationship between training and test compounds.

This problem becomes particularly pronounced in the presence of activity cliffs. Activity cliffs are generally understood as pairs of structurally highly similar molecules that nevertheless exhibit unexpectedly large differences in biological activity or potency [17, 18]. They represent an extreme form of SAR discontinuity and

directly challenge the commonly invoked molecular similarity principle, according to which similar structures are expected, on average, to possess similar properties. Maggiora highlighted this issue nearly two decades ago when discussing why compounds that appear as QSAR 'outliers' may instead reflect important discontinuities in the underlying activity landscape [17].

Subsequent analyses established that activity cliffs are not merely statistical curiosities: they can contain highly informative SAR signals because a relatively small chemical modification may identify a structural feature that is critical for molecular recognition and potency [18, 19]. At the same time, experimental uncertainty can generate or exaggerate apparent cliffs, emphasizing the importance of high-quality and consistently measured bioactivity data when studying SAR discontinuity [20].

The dual nature of activity cliffs creates a fundamental difficulty for predictive modeling. From a medicinal-chemistry perspective, a cliff can be highly valuable because it identifies a small structural transformation associated with a major potency change. From an ML perspective, however, the same observation contradicts the local smoothness that many molecular representations and learning procedures effectively exploit. If two molecules occupy neighboring regions of descriptor, fingerprint, or latent representation space but possess very different activities, a model must resolve a steep local change in the target function from comparatively little structural information.

Studies of dataset modelability have accordingly shown that local inconsistencies between chemical similarity and activity labels are strongly associated with reduced QSAR predictability [21]. Sheridan et al. further demonstrated that activity-cliff-related properties of datasets are important determinants of predictive performance and that multiple modern QSAR methods remain poor at predicting molecules located in such cliff regions [20]. Thus, increasing model flexibility alone does not necessarily resolve the underlying problem.

The rapid development of deep learning has intensified this question. Neural networks and graph neural networks (GNNs) can learn nonlinear molecular representations directly from descriptors, fingerprints, strings, or molecular graphs, raising the expectation that more expressive architectures might capture subtle structural transformations that conventional QSAR approaches overlook [22]. However, evidence increasingly suggests that model complexity and activity-cliff sensitivity are not synonymous.

Van Tilborg et al. benchmarked a broad range of machine- and deep-learning approaches across multiple

bioactivity datasets and showed that prediction errors increased substantially for compounds involved in activity cliffs; importantly, sophisticated deep-learning architectures did not consistently eliminate this limitation [23]. Likewise, large-scale activity-cliff prediction across 100 activity classes showed that predictive accuracy did not increase monotonically with methodological complexity and that apparently strong performance could be substantially influenced by compound overlap and memorization between training and test sets [24]. Dablander et al. similarly found that commonly used QSAR models frequently fail to identify activity cliffs when the activities of both molecules are unknown, with performance becoming especially challenging when structurally related information is absent from the training data [25].

These observations expose an important limitation of conventional model evaluation. Global metrics such as the coefficient of determination (R^2), root-mean-square error (RMSE), mean absolute error (MAE), accuracy, or area under the receiver-operating-characteristic curve summarize predictive behavior across an entire test set. They do not necessarily reveal whether a model can reproduce sharp local differences between highly related molecules. A model may therefore obtain an excellent aggregate score by accurately predicting the majority of chemically smooth regions while systematically smoothing over a comparatively small but chemically important population of activity cliffs. The resulting model can appear robust according to conventional benchmarks yet fail precisely where medicinal chemists often require the greatest sensitivity: distinguishing close analogues during lead optimization.

This limitation also connects activity-cliff analysis to the broader concept of the QSAR applicability domain. Regulatory guidance emphasizes that predictive reliability depends on the chemical and response space represented by model-development data and that extrapolative predictions require particular caution [26]. Activity cliffs add an additional complication because poor prediction may occur even for molecules that appear structurally close to training examples, demonstrating that structural proximity alone does not guarantee local predictability.

Accordingly, activity cliffs should not be viewed solely as problematic observations to be removed from QSAR datasets. Their presence provides a stringent test of whether a model has learned chemically meaningful differences rather than merely broad similarity patterns. This perspective is beginning to influence the design of molecular-learning methods. Recent approaches explicitly incorporate activity-cliff information into

representation learning and loss functions rather than treating cliffs as conventional training examples. For example, activity-cliff-aware graph learning has been proposed to reshape molecular latent spaces so that subtle structural modifications associated with large activity changes are more effectively distinguished [27]. Such developments signal a transition from simply measuring performance on cliff-containing datasets toward designing models that explicitly recognize SAR discontinuity.

Against this background, the key question is no longer simply whether ML can produce accurate QSAR models, but whether these models generalize when structural similarity ceases to be a reliable proxy for biological similarity. This narrative review therefore examines activity cliffs as a fundamental challenge to ML generalization in QSAR. It first considers how activity cliffs are defined, detected, and generated within molecular activity landscapes; subsequently evaluates why they challenge conventional assumptions of similarity-based learning; compares the behavior of classical ML, deep-learning, and graph-based approaches in cliff-containing chemical space; and discusses validation strategies, molecular representations, uncertainty estimation, applicability-domain concepts, and activity-cliff-aware learning strategies.

Building on this synthesis, the review culminates in a proposed activity-cliff-aware QSAR framework that integrates rigorous bioactivity curation, activity-landscape characterization, leakage-resistant validation, global and cliff-specific evaluation, applicability-domain analysis, uncertainty quantification, and cliff-aware learning objectives for more reliable prospective QSAR modeling. By connecting activity-cliff research with the broader problem of molecular ML generalization, this review aims to clarify why increasingly powerful algorithms can still struggle with apparently simple questions: when two molecules are nearly identical, why can their activities be dramatically different, and why do predictive models so often fail to anticipate that difference?

2. Activity Cliffs: Definition, Detection, and Chemical Origins

2.1. Activity Cliffs as Discontinuities in Structure-Activity Landscapes

The activity-cliff concept provides a quantitative representation of one of the most challenging characteristics of molecular SARs: small structural changes can occasionally produce unexpectedly large changes in biological activity. In its most general form, an activity cliff is defined as a pair of compounds that are

highly similar in molecular structure, act against the same biological target, but exhibit a substantial difference in potency [17–19]. Accordingly, this review uses the term activity cliff primarily in its pairwise sense; larger cliff-containing groups or networks are regarded as collections of interconnected pairwise activity-cliff relationships rather than as a separate formal definition.

Although this definition appears straightforward, identifying activity cliffs computationally requires explicit decisions regarding what constitutes high structural similarity and what magnitude of activity difference should be considered sufficiently large to form a cliff. Consequently, activity-cliff identification is inherently dependent on molecular representation, similarity measurement, biological assay quality, and the selected potency threshold [18, 19].

A useful conceptual framework for understanding these relationships is the structure-activity landscape. In this representation, molecular similarity defines proximity within chemical space, whereas biological activity can be viewed as an additional dimension of the landscape [28, 29]. Regions in which structurally similar molecules exhibit similar activities form relatively smooth surfaces and correspond to continuous SARs. In contrast, structurally neighboring molecules with markedly different activities generate steep or discontinuous regions—the activity cliffs. Such landscape topology highlights an important asymmetry between conventional QSAR modeling and medicinal-chemistry interpretation. Smooth regions are comparatively favorable for interpolation because molecular similarity provides useful information about activity, whereas cliffs contain disproportionately high SAR information. After all, a small chemical change identifies a structural feature capable of dramatically altering potency [18, 29].

Activity cliffs should therefore not automatically be interpreted as noise or anomalous observations. Maggiora originally emphasized that compounds appearing as QSAR outliers may instead reveal genuine discontinuities in the underlying SAR [17]. Subsequent studies strengthened this interpretation by showing that cliff-forming modifications can identify substitutions, atomic positions, or interaction patterns that are particularly important for ligand potency [18, 19]. Nevertheless, apparent cliffs may also originate from inconsistent measurements or experimental uncertainty, making differentiation between genuine SAR discontinuity and assay-derived noise essential [20]. Thus, a meaningful activity-cliff analysis requires both structural and biological criteria.

2.2. Quantifying Molecular Similarity and Potency Differences

The first component of an activity-cliff definition is molecular similarity. One of the most widely used approaches represents molecules as binary molecular fingerprints and calculates pairwise similarity using the Tanimoto coefficient. For two molecular fingerprints, A and B, the coefficient can be written as shown in Equation (1):

$$T(A, B) = \frac{c}{a + b - c} \quad (1)$$

where a and b are the numbers of features present in molecules A and B, respectively, and c represents features shared by both molecules. The coefficient ranges from 0, representing no shared fingerprint features, to 1, representing identical fingerprint representations. Importantly, a Tanimoto value is not an intrinsic property of a molecular pair because its magnitude depends on the fingerprint used to represent the molecules. Thus, the same compounds can appear more or less similar under alternative molecular representations [19].

Extended-connectivity fingerprints (ECFPs), particularly ECFP4-type representations, have become widely used for this purpose because they encode local circular substructures around individual atoms and were developed specifically for structure-activity modeling [30]. When used together with Tanimoto similarity, ECFPs provide an efficient means of screening large molecular datasets for highly similar compound pairs. However, fingerprint similarity remains representation dependent. Two molecules may receive a high similarity score because they share most of their topology even though a small substituent, heteroatom, stereochemical feature, or positional modification critically alters biological recognition. Conversely, chemically meaningful analogues may sometimes receive lower numerical similarity values depending on fingerprint construction. Consequently, no universally valid Tanimoto threshold can define an activity cliff across all datasets and molecular representations [19].

The second component is the difference in biological activity between the compounds. For continuous potency measurements, activities are commonly converted to logarithmic units using Equation (2):

$$pActivity = -\log_{10}(Activity, [M]) \quad (2)$$

where Activity may correspond to K_i , K_d , IC_{50} , or another relevant measurement. The difference in biological activity between compounds *i* and *j* is then calculated as shown in Equation (3):

$$\Delta pActivity_{(i,j)} = |pActivity_i - pActivity_j| \quad (3)$$

A difference of two logarithmic units corresponds to a 100-fold potency difference. A frequently applied conservative activity-cliff definition therefore requires highly similar compounds to differ in potency by at least two orders of magnitude [19, 31]. However, this threshold should not be interpreted as universal. Thresholds of one or two log units have both appeared in activity-cliff studies, and their suitability depends on the target, potency distribution, assay uncertainty, and purpose of the analysis [19].

The type and quality of bioactivity measurement are equally important. Equilibrium constants such as K_i or K_d are particularly attractive for systematic cliff analysis because they more directly characterize molecular binding affinity than assay-dependent end points such as IC_{50} [19]. Nevertheless, large public resources frequently contain IC_{50} measurements, making such data unavoidable in many ML studies. Mixing measurements obtained under substantially different experimental conditions can create artificial potency differences, whereas experimental errors can amplify or obscure genuine cliffs. Sheridan et al. demonstrated that activity-cliff metrics are strongly associated with dataset predictability but also emphasized that large activity uncertainty makes it difficult to distinguish genuine cliffs from uncertainty-induced apparent cliffs [20]. Careful activity-data curation should therefore precede computational cliff assignment.

2.3. Computational Strategies for Activity-Cliff Detection

Activity cliffs can be identified using several complementary strategies, each capturing a somewhat different interpretation of molecular similarity. A straightforward approach performs pairwise fingerprint comparisons and selects compound pairs satisfying predefined similarity and activity-difference thresholds. Although computationally simple and scalable, this approach strongly depends on the choice of fingerprint and numerical similarity cutoff.

A more quantitative approach is the Structure-Activity Landscape Index (SALI) introduced by Guha and Van Drie [28]. For compounds i and j , SALI can be expressed as Equation (4):

$$SALI_{(i,j)} = \frac{|A_i - A_j|}{1 - Sim(i,j)} \quad (4)$$

where A_i and A_j represent the biological activities of compounds i and j , respectively, and $Sim(i,j)$ denotes their structural similarity. A high SALI value is obtained when two molecules are highly similar but differ strongly in activity, directly identifying steep regions of the

structure-activity landscape. Unlike a binary cliff classification, SALI provides a continuous measure of SAR discontinuity and can therefore rank molecular pairs according to cliff severity. It has also been proposed as a means of evaluating how effectively QSAR models reproduce the topology of experimentally observed activity landscapes [28].

A particularly chemically interpretable alternative is matched molecular pair (MMP) analysis. An MMP consists of two molecules differing by a well-defined localized structural transformation while retaining the remainder of the molecular framework [32]. Hussain and Rea developed an efficient fragmentation-based algorithm that enabled systematic identification of MMPs in very large molecular collections [32]. This approach was subsequently adapted specifically for activity-cliff detection. Hu et al. introduced MMP-cliffs, in which compounds related through a localized structural transformation are screened for substantial potency differences [31]. Compared with global fingerprint similarity, MMP analysis explicitly identifies the chemical transformation associated with the potency change and is therefore particularly suitable for medicinal-chemistry interpretation.

A conservative formulation uses size-restricted MMPs, limiting the magnitude of the exchanged molecular fragment so that paired compounds remain close structural analogues [19, 31]. Combined with a potency difference of approximately two log units, this approach reduces ambiguity in determining whether paired molecules should truly be considered structurally similar. It also provides an intuitive answer to an important medicinal-chemistry question: which local structural modification transformed one potency level into another?

The activity-cliff concept has also been extended beyond two-dimensional molecular graphs. Hu and Bajorath introduced three-dimensional activity cliffs by comparing experimentally determined ligand binding modes from protein-ligand crystal structures [33]. Similarity was evaluated using information concerning ligand conformation, spatial position, and atomic properties. Their analysis of BACE1 and factor Xa inhibitors revealed limited correspondence between cliffs identified using two-dimensional and three-dimensional representations, demonstrating that cliff assignment can change substantially with molecular representation [33]. Thus, an apparent structural neighbor in a 2D fingerprint space does not necessarily display an equivalent three-dimensional binding mode, and the reverse can also occur.

These approaches should consequently be considered complementary rather than interchangeable. Fingerprint-based cliffs are scalable and convenient for large ML datasets; SALI characterizes the magnitude of SAR discontinuity; MMP-based cliffs maximize chemical interpretability; and 3D cliffs can connect activity discontinuities directly to experimentally observed binding configurations.

2.4. Chemical and Biological Origins of Activity Cliffs

The mechanistic origin of an activity cliff is ultimately associated with the nonlinear nature of molecular recognition. Ligand binding depends on a combination of intermolecular contacts, conformational preferences, electrostatic interactions, hydrogen bonding, steric complementarity, desolvation, and dynamic behavior. Therefore, a structural change that appears small according to a molecular fingerprint does not necessarily represent a small perturbation in binding free energy. A single modification may create an additional favorable interaction, disrupt an essential contact, alter ligand orientation, change conformational accessibility, or modify the physicochemical environment of a critical binding-site region.

Structure-based studies provide direct evidence for this behavior. Analyses of three-dimensional cliffs derived from experimentally determined protein-ligand complexes have shown that highly related molecules can differ in interaction patterns despite maintaining substantial overall structural similarity [33]. In some cases, observed differences in ligand-target contacts provide plausible explanations for large potency changes; in others, even high-resolution structures do not provide an obvious explanation, indicating that static structural representations may fail to capture all energetic and dynamic contributions underlying cliff formation.

The chemical subtlety of these effects has been demonstrated particularly clearly for single-atom activity cliffs. Hu and Bajorath systematically investigated compound pairs distinguished only by single-atom modifications, including heteroatom replacements and changes in heteroatom positions [34]. More than 1,500 such cliffs were identified across multiple targets, demonstrating that major potency changes do not necessarily require large molecular transformations. When protein-ligand crystal structures were available, individual atomic modifications could in some cases be related to differences in ligand-target interactions [34]. These observations are especially important for ML because many molecular representations encode two such molecules as nearly identical despite their substantially different activities.

Stereochemical and positional differences create a related challenge. Two molecules can share the same or nearly identical connectivity while adopting different three-dimensional arrangements or interaction geometries. Depending on whether chirality, conformation, or local atomic environments are explicitly represented, a computational descriptor may either preserve or obscure these distinctions. Hence, an activity cliff is not solely a property of two molecules; computationally, it emerges from the combination of the molecular structures, the representation used to compare them, and the biological activity measurements assigned to them [19, 33].

An additional source of apparent cliff formation is experimental uncertainty. Bioactivity measurements obtained in different assay formats, laboratories, protein constructs, or experimental conditions may not be directly comparable. A nominally large activity difference may consequently reflect biological or experimental variability rather than a genuine chemical effect. This problem is particularly relevant when assembling large QSAR datasets from public databases because compounds may originate from heterogeneous studies. Data standardization, target-confidence filtering, consistent activity types, removal of ambiguous measurements, and consideration of replicate variability are therefore essential for constructing reliable activity-cliff benchmarks [20].

Taken together, activity cliffs should be interpreted as local violations of an otherwise useful similarity-activity relationship, rather than as evidence that molecular similarity itself is meaningless. Most molecular datasets contain extensive regions in which structurally related compounds display correlated activities, permitting QSAR models to learn meaningful patterns. Cliffs represent comparatively localized regions where this relationship becomes sharply nonlinear. Indeed, large-scale analyses have found that although a considerable fraction of bioactive molecules may participate in at least one cliff, cliff-forming pairs constitute only a relatively small fraction of all qualifying similar compound pairs [35]. Their rarity, however, does not reduce their importance. Because they concentrate large activity changes into very small structural differences, they represent some of the most information-rich and predictive-demanding regions of chemical space.

This dual character makes activity cliffs particularly relevant to ML generalization. The same feature that makes them useful to medicinal chemists—a small structural modification with a large biological consequence—makes them difficult for algorithms that learn predominantly from broader correlations between

molecular similarity and activity. Accordingly, the next question is not merely whether an ML model achieves good performance across an entire molecular dataset, but whether it can reproduce these steep local transitions when confronted with compounds that are almost indistinguishable in representation space.

3. Why Activity Cliffs Challenge Machine Learning Generalization

3.1. From Molecular Similarity to a Local Generalization Problem

Most ML approaches to QSAR implicitly rely on a degree of continuity between molecular structure and biological activity. Given a molecular representation x , a model estimates an activity value according to Equation (5):

$$\hat{y} = f(x) \quad (5)$$

where f is learned from experimentally characterized molecules. Successful prediction is therefore facilitated when nearby molecules in representation space tend to possess similar activity values. In smooth regions of the structure–activity landscape, this assumption is reasonable: local structural similarity provides information that allows a model to interpolate between previously observed examples. Activity cliffs violate this local regularity.

For an activity-cliff pair (i,j) , the prediction problem can be expressed conceptually as follows, as shown in Equation (6):

$$d(x_i, x_j) \rightarrow 0 \quad \text{while simultaneously} \quad |y_i - y_j| \quad (6)$$

where $d(x_i, x_j)$ denotes molecular distance and y_i and y_j denote measured activities. The model is consequently required to map a small displacement in molecular representation space to a large displacement in biological activity. This is fundamentally more difficult than interpolating through a smooth region of an SAR landscape.

The difficulty is compounded by the way many molecular representations are constructed. Fingerprints intentionally encode shared substructures, physicochemical descriptors compress molecules into a finite set of global properties, and message-passing neural networks aggregate local atomic environments into learned molecular representations. These approaches are highly useful for identifying broad molecular patterns but can place close analogues near one another even when a subtle modification is responsible for a large potency shift. Thus, the central challenge posed by activity cliffs is not simply that the compounds are unusual; rather, the structural signal that

differentiates the compounds may be small relative to the magnitude of the biological response that must be predicted.

Activity-cliff compounds are therefore particularly difficult to predict when neither member of the pair is represented in the training data. Dablander et al. found consistently low cliff sensitivity under this condition, whereas sensitivity improved when the measured activity of one analogue was already available [25]. This suggests that predicting an unseen cliff is substantially more difficult than estimating the effect of modifying an experimentally characterized analogue.

3.2. Interpolation Is Not Equivalent to Generalization

The distinction between interpolation and generalization has long been recognized in QSAR validation. Random splitting assigns individual molecules to training and test sets without explicitly considering chemical relatedness. Consequently, members of the same analogue series—or even nearly identical compounds—may occur in both partitions. Under such conditions, a model can predict a test molecule largely by learning relationships represented by close neighbors in its training set. The resulting performance is useful for assessing interpolation within known chemistry but may overestimate performance on genuinely novel chemical space.

Sheridan demonstrated this problem using temporal validation, showing that conventional random cross-validation could provide overly optimistic estimates relative to genuinely prospective predictions and that time-split validation more closely reproduced prospective performance [14]. Wallach and Heifets later examined seven ligand-based classification benchmarks and showed that train-validation redundancy was strongly correlated with apparent model performance across different fingerprints, similarity measures, and algorithms. They argued that substantial benchmark performance could be explained by memorization of redundant chemical information rather than genuine prospective generalization [15].

A common response has been the use of scaffold-based splitting, in which compounds sharing the same molecular framework are kept within the same partition. The approach derives from the molecular-framework concept of Bemis and Murcko [36] and aims to create a more demanding test of transfer to previously unseen chemotypes. Yang et al., in a benchmark spanning 19 public and 16 proprietary molecular-property datasets, found random splits to be systematically easier than scaffold and chronological splits and concluded that scaffold splitting better reflects generalization to new

chemical space when temporal information is unavailable [16]. They nevertheless regarded chronological splitting as preferable when a reliable time axis exists.

Activity cliffs expose an additional weakness in the train-test splitting problem. Because a cliff is defined by a pair of compounds, splitting pairs randomly can result in the same individual compound appearing in a training pair and a test pair. The model may therefore recognize one or both molecules rather than infer the underlying SAR discontinuity. Tamura et al. demonstrated this directly in a large-scale comparison of activity-cliff prediction across 100 activity classes. When overlapping compounds were allowed between training and test MMPs, performance was markedly higher. When compound overlap was explicitly removed, balanced accuracy generally declined to approximately 0.6 or below and Matthews correlation coefficients frequently fell below 0.25 [24]. This finding illustrates that data leakage can transform an intrinsically difficult activity-cliff problem into an apparently successful prediction task.

Thus, evaluation of activity-cliff generalization requires independence at more than the level of individual database rows. Independence must also be considered at the level of molecular identity, analogue relationships, scaffolds, MMPs, and local chemical neighborhoods.

3.3. Global Performance Can Conceal Local Failure

Another reason activity cliffs remain underappreciated is that conventional QSAR metrics aggregate performance across the complete test set. For regression, common metrics include MAE, as shown in Equation (7), and RMSE, as shown in Equation (8):

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (7)$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (8)$$

Global error metrics can obscure poor performance on activity cliffs because cliff-associated compounds may represent only a small fraction of the test set. A model can therefore achieve a low overall MAE while producing substantially larger errors for these compounds. Cliff-specific error should be reported alongside global metrics to determine whether the model captures local SAR discontinuities.

Van Tilborg et al. provided direct evidence for such behavior by benchmarking multiple molecular ML and deep-learning approaches on activity datasets containing

cliffs. Across evaluated models, compounds associated with activity cliffs were generally harder to predict than non-cliff compounds, and the use of more sophisticated neural architectures did not eliminate the performance deficit [23]. Dablander et al. similarly observed a relationship between overall QSAR error and activity-cliff classification performance. Still, they showed that models remained poorly sensitive to previously unseen cliffs even when their conventional QSAR predictions were comparatively strong [25].

These findings indicate that model evaluation should assess both overall predictive accuracy and performance within local SAR discontinuities. This distinction is particularly relevant to lead optimization, where medicinal chemists often compare close analogues rather than structurally unrelated compounds. A model may perform well across broad chemical space yet fail to rank highly similar compounds correctly. Dablander et al. found that predicting the potency direction between similar compounds was especially difficult when neither compound had been observed during training [25]. Therefore, local ranking accuracy and activity-cliff sensitivity should complement global regression metrics when QSAR models are intended to support analogue optimization.

3.4. Increasing Model Complexity Does Not Necessarily Solve Activity Cliffs

The development of deep learning has encouraged the expectation that sufficiently flexible architectures can learn subtle nonlinear SARs that traditional QSAR models cannot capture. GNNs are particularly appealing because they operate directly on molecular graphs and can learn task-specific representations rather than relying solely on predefined fingerprints. However, the available evidence does not support a simple relationship between architectural complexity and successful cliff prediction.

Tamura et al. compared approaches of increasing complexity for activity-cliff classification and found that predictive performance did not improve monotonically with model sophistication [24]. More importantly, removing compound overlap between training and test pairs significantly reduced performance regardless of the algorithm used, demonstrating that data composition and evaluation design could exert a stronger influence than increasing model complexity.

Dablander et al. reached a complementary conclusion. Graph-isomorphism-network representations showed advantages for activity-cliff classification and tended to achieve higher cliff sensitivity than classical representations. However, ECFPs remained stronger for overall QSAR prediction in their evaluated datasets [25].

Thus, a molecular representation optimized for broad predictive accuracy is not necessarily the representation that most effectively discriminates subtle cliff-forming transformations.

This observation is consistent with broader molecular-property benchmarking. Yang et al. found that learned graph representations can provide strong molecular prediction performance, particularly when supplemented with additional molecular descriptors, but also showed that performance depends strongly on the validation scenario [16]. More recent systematic benchmarking across EGFR, DRD2, BACE1, and hERG data further demonstrates that generalization gaps persist across algorithms and targets. Alshehade et al. reported target-dependent reductions in R^2 under scaffold-based evaluation and substantially poorer performance for compounds outside similarity-defined applicability domains [37].

Collectively, these results argue against interpreting activity cliffs as a problem that will disappear automatically with larger neural architectures. Model complexity increases the capacity to fit nonlinear relationships, but the relevant local relationship must still be supported by training data and adequately represented in molecular features. If two molecules are encoded nearly identically and the training set contains little evidence explaining why one should be much more potent than the other, a highly expressive model may have no reliable basis for learning the discontinuity. In such circumstances, additional capacity may increase fitting flexibility without necessarily improving prospective cliff generalization.

3.5. Chemical-Space Coverage, Applicability Domain, and Uncertainty

Activity-cliff errors can arise from two conceptually different forms of insufficient information. The first occurs when a test compound lies far from the training distribution, representing conventional out-of-domain extrapolation. The second occurs when the compound lies close to known molecules but the local SAR is steep or discontinuous. Activity cliffs are particularly interesting because they demonstrate that chemical proximity and predictive reliability are not equivalent.

The applicability-domain concept attempts to identify whether a prediction is supported by the chemical space represented during model development. Similarity-based applicability domains generally assume that confidence should decrease as a query compound becomes more distant from the training set. This assumption remains useful: recent multi-target QSAR benchmarking found marked reductions in predictive performance outside

Tanimoto-defined applicability domains [37]. However, activity cliffs reveal an important limitation. A query compound may be extremely similar to training molecules and therefore appear safely inside a structural applicability domain while still being difficult to predict because the local SAR is discontinuous.

The same problem arises in uncertainty quantification. Ideally, a predictive system should assign greater uncertainty to molecules for which large errors are likely. Yet uncertainty methods can themselves struggle in cliff-rich regions. Kötter et al. examined commonly used ML uncertainty approaches on molecular activity datasets and found that poor predictions arose both from insufficient training-set representation and from steep SAR regions containing large property differences between similar molecules. Several uncertainty methods had difficulty consistently identifying prediction failures associated specifically with steep SAR behavior [38]. Consequently, low structural distance and low estimated uncertainty should not automatically be interpreted as evidence that a prediction is locally reliable.

Recent activity-cliff-focused work also reinforces the importance of dataset composition. López-Pérez and Miranda-Quintana examined how the distribution of activity cliffs across training and test partitions influences ML prediction errors and found that nonuniform distributions of cliffs and chemical space can produce poorer models, emphasizing that cliff-rich datasets require case-specific analysis rather than reliance on a single splitting rule [39].

More broadly, generalization failure can arise from two distinct sources: chemical-space extrapolation and local SAR discontinuity. The former occurs when the model has not encountered sufficiently similar chemistry during training. In contrast, the latter may occur even when highly similar chemical structures are present in the training set. Conventional QSAR validation primarily assesses the former through measures of chemical-space coverage and dataset splitting. Activity-cliff analysis, however, highlights the importance of the latter by revealing prediction errors that arise from abrupt changes in biological activity among structurally similar compounds.

Activity cliffs provide a demanding test of molecular ML models because they expose failures that may be hidden by global performance metrics. Prediction errors can arise from local SAR discontinuities, structural overlap between training and test compounds, and inadequate chemical-space coverage, and these problems are not necessarily resolved by increasing model complexity.

4. Machine Learning Approaches and Molecular Representations for Activity-Cliff Prediction

4.1. Classical Machine Learning and Fixed Molecular Representations

Classical ML approaches remain important baselines for QSAR because they combine relatively simple learning algorithms with chemically engineered molecular representations [40–42]. RF, SVM, GB methods, k-nearest-neighbor (kNN) models, and multilayer perceptron (MLP) trained on fingerprints or physicochemical descriptors continue to achieve competitive performance across many molecular-property prediction tasks. Their continued relevance is particularly important in activity-cliff research because increasing model complexity has not consistently translated into superior cliff prediction.

ECFPs are among the most widely used molecular representations in ligand-based QSAR. ECFPs encode circular atomic environments within a defined neighborhood and transform molecular structure into a fixed-dimensional representation suitable for conventional ML algorithms [30]. For ordinary QSAR prediction, this representation remains remarkably effective. Dablander et al., for example, compared ECFPs, physicochemical-descriptor vectors, and graph-isomorphism-network representations across RF, kNN, and MLP regression models. Across dopamine D2 receptor, factor Xa, and SARS-CoV-2 main-protease datasets, ECFP-based models generally achieved the lowest QSAR prediction errors, with the ECFP-MLP combination consistently performing particularly well [25].

The strength of ECFPs is also closely connected to their potential weakness in activity-cliff regions. Fingerprints encode the presence or absence of local structural environments but do not explicitly represent the magnitude of biological consequences associated with those environments. Two compounds differing by only a small substituent may therefore retain extremely similar fingerprints despite exhibiting a substantial potency difference. Similarity-based algorithms such as kNN are especially vulnerable to this problem because their basic prediction rule assumes that nearby compounds should have similar responses. Van Tilborg et al. explicitly included kNN as an activity-cliff baseline for this reason [23].

Nevertheless, the performance of classical models on activity cliffs should not be underestimated. In a systematic benchmark of 24 machine- and deep-learning approaches across 30 biological targets, van Tilborg et al. found that all evaluated model classes experienced

performance deterioration for activity-cliff compounds, but descriptor-based conventional ML approaches often outperformed more sophisticated deep-learning alternatives [23]. This result challenges the assumption that learned molecular representations necessarily provide superior sensitivity to subtle SAR discontinuities.

Different conventional algorithms also display different error characteristics. Dablander et al. observed that RF models tended to achieve comparatively high activity-cliff precision but lower sensitivity, suggesting relatively conservative predictions of extreme potency differences. In contrast, other model-representation combinations identified more cliffs but also produced more false-positive predictions [25]. Thus, model comparison should not rely solely on aggregate regression performance; sensitivity, precision, Matthews correlation coefficient, and potency-direction accuracy provide complementary information about how an algorithm behaves specifically around activity cliffs.

These findings suggest an important principle: a representation that performs well for global QSAR is not necessarily optimized for recognizing SAR discontinuities. ECFPs remain strong general-purpose representations because they preserve rich local substructural information. Still, they do not explicitly encode whether a small structural transformation should correspond to a smooth or abrupt biological response.

4.2. Graph Neural Networks and Learned Molecular Representations

GNNs offer an alternative representation paradigm in which molecules are treated directly as graphs containing atoms as nodes and bonds as edges. Rather than relying entirely on predetermined fingerprint rules, GNNs iteratively aggregate information from neighboring atoms and learn molecular representations optimized for the target prediction task. This flexibility has made graph-based architectures one of the dominant approaches in modern molecular ML.

Early large-scale benchmarks provided substantial motivation for this transition. MoleculeNet showed that learned graph representations frequently outperformed conventional methods across a variety of molecular-property datasets. However, the advantage was strongly task dependent and decreased in settings characterized by limited data or other challenging conditions [12]. Yang et al. subsequently benchmarked learned molecular representations across 19 public and 16 proprietary datasets. They introduced a directed message-passing neural network that matched or exceeded several established descriptor- and graph-based approaches across many endpoints [16].

In principle, learned graph representations should be attractive for activity-cliff prediction because they can adapt molecular embeddings according to the property being modeled. Instead of defining similarity exclusively through structural overlap, training can potentially place two structurally related molecules farther apart in latent space when their activities differ substantially. Dablander et al. found evidence consistent with this possibility. Graph-isomorphism-network features were competitive with or superior to conventional representations for activity-cliff classification, and graph isomorphism network-based models displayed greater cliff sensitivity when both compounds were unseen during training [25].

However, this advantage did not translate into universal superiority. In the same experiments, ECFPs remained consistently stronger for conventional compound-level QSAR prediction [25]. Van Tilborg et al. similarly found that deep-learning models were not systematically superior to traditional descriptor-based approaches for cliff-containing compounds [23]. These results reveal an important distinction between two objectives: overall molecular-property prediction is not equivalent to resolution of local activity discontinuities.

One reason is the neighborhood-aggregation mechanism underlying many GNNs. Message-passing architectures repeatedly combine atomic information from local graph neighborhoods to construct increasingly global molecular embeddings. This process is powerful for capturing molecular context, but standard training objectives still primarily minimize average prediction error across molecules. If cliff-forming compounds constitute a small fraction of training observations, their contribution to the global loss may be insufficient to force the latent representation to distinguish them. Structurally related molecules can consequently remain close in learned representation space even when their activities diverge.

Pretraining does not automatically remove this limitation. Self-supervised graph pretraining can improve molecular prediction when labeled data are limited by exposing models to larger unlabeled chemical collections. Hu et al. showed that carefully designed node- and graph-level pretraining improved generalization across downstream molecular tasks, while naïve pretraining strategies could provide little benefit or even lead to negative transfer [43]. The implication for activity cliffs is important: learning a richer general representation of molecular structure is valuable, but such a representation is not necessarily optimized to distinguish the specific structural perturbations responsible for abrupt potency changes.

4.3. SMILES-Based and Transformer Representations

Molecular representation learning has also increasingly adopted ideas from natural-language processing. In this framework, molecules are represented as token sequences, typically using simplified molecular-input line-entry system (SMILES) strings, and sequence models learn relationships among atoms, bonds, branches, ring closures, and larger structural patterns. Transformer architectures are particularly attractive because self-attention can model dependencies between distant tokens and pretraining can exploit very large unlabeled molecular corpora.

The principal advantage of this approach is that molecular representations need not be restricted to handcrafted descriptors or fixed-radius neighborhoods. Pretrained sequence models can learn contextualized embeddings and subsequently be fine-tuned for target-specific property prediction [44]. Early work such as SMILES Transformer demonstrated that unsupervised pretraining on large molecular collections could generate representations useful for downstream molecular prediction, especially in data-limited settings [45]. More recent transformer architectures continue to improve molecular-property benchmarks through multiscale, multimodal, and domain-adaptation strategies.

Strong performance of molecular language models on conventional property-prediction benchmarks does not necessarily imply reliable activity-cliff prediction. A cliff-sensitive representation must preserve small structural changes when those changes produce large differences in biological activity. This capability should therefore be evaluated directly using activity-cliff benchmarks rather than inferred from overall predictive performance.

4.4. Why Representation Matters as Much as the Algorithm

Comparisons across model families reveal that activity-cliff prediction is not simply an algorithm-selection problem. The interaction between molecular representation, learning algorithm, training objective, and chemical-space coverage determines whether a model can resolve local SAR discontinuities.

Consider two highly similar compounds represented by embedding vectors z_i and z_j . A conventional molecular representation may produce Equation (9):

$$|z_i - z_j| \approx 0 \quad (9)$$

because the molecules share most of their structural features. If the experimental activities satisfy Equation (10):

$$|y_i - y_j| \gg 0 \quad (10)$$

the representation becomes locally inconsistent with activity space. A downstream model must then generate substantially different predictions from nearly indistinguishable inputs. Increasing the capacity of the predictor may help, but it does not remove the underlying representational conflict.

Ideally, a property-aware representation would instead learn a geometry in which molecular distances reflect relevant biological differences, as expressed in Equation (11):

$$|z_i - z_j| \propto |y_i - y_j| \quad (11)$$

This objective differs fundamentally from maximizing structural representation quality alone. In conventional QSAR training, representation learning is usually driven by compound-level prediction loss. For example, the mean absolute error can be expressed as Equation (12):

$$L_{\text{reg}} = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i| \quad (12)$$

or through its squared-error equivalent. Such objectives penalize incorrect activity predictions but do not explicitly instruct the model to reorganize molecular neighborhoods when close structural analogues form activity cliffs.

No single molecular representation appears to perform best under all evaluation settings. ECFPs often provide strong overall regression performance, whereas trainable GNN representations may provide greater sensitivity to some activity cliffs. Conventional descriptor-based models can also outperform more complex architectures [23–25]. Activity-cliff prediction should therefore be treated as a specific representation-learning problem rather than as a simple comparison of model complexity.

4.5. From Generic Representation Learning to Activity-Cliff-Aware Models

Recent work provides strong support for explicitly incorporating activity-cliff information into model training rather than expecting a generic GNN to discover these rare discontinuities automatically. Cui et al. introduced an activity-cliff awareness framework that modifies the latent geometry learned by GNNs [27]. Rather than optimizing compound-level regression alone, the method combines conventional regression loss with a triplet-based contrastive objective designed to identify inconsistent molecular neighborhoods.

Conceptually, the method considers an anchor compound and its molecular neighbors, whose relationships in structural and activity spaces may differ.

The additional learning objective encourages structurally similar compounds with compatible activities to maintain consistent representations while accounting for structurally close compounds that exhibit substantially different activities. The overall training objective therefore combines conventional molecular-property prediction loss with an activity-cliff-aware loss component. The relative contribution of each component is controlled by a weighting parameter, allowing the model to balance overall prediction accuracy with sensitivity to local SAR discontinuities.

The main contribution of this approach is the modified learning objective rather than the specific GNN architecture. Standard regression minimizes average prediction error, whereas activity-cliff-aware learning also encourages the latent representation to separate structurally similar compounds with large activity differences.

Cui et al. evaluated this principle across 52 datasets, including low-sample congeneric series, mixed-scaffold molecular benchmarks, matched-pair cliff classification, and tasks related to absorption, distribution, metabolism, excretion, and toxicity. Incorporating activity-cliff awareness improved prediction relative to conventional ECFP and GNN baselines and transferred across multiple GNN backbones [27]. Importantly, the work also found that ordinary regression training left many high-value activity-cliff inconsistencies in latent space largely unresolved, supporting the hypothesis that conventional global loss functions provide insufficient learning pressure on cliff regions.

These results shift the focus from selecting the best conventional QSAR algorithm to using activity-cliff information directly during representation learning. The learning objective may therefore be as important as the choice of predictive architecture.

Current evidence does not identify a universally superior representation for activity-cliff prediction. Fingerprint- and descriptor-based models remain strong baselines, while learned graph representations can improve cliff sensitivity in some datasets. Pretraining alone does not guarantee sensitivity to SAR discontinuities; more consistent improvements may require activity-cliff information to be incorporated directly into the training objective.

The practical question is therefore whether a representation preserves the small structural differences associated with large biological effects while maintaining reliable performance across the broader chemical space.

4.6. Comparative Evidence from Major Activity-Cliff Studies

The preceding evidence indicates that the activity-cliff problem has evolved from an early conceptual concern regarding SAR discontinuity into a broader challenge involving molecular representation, validation strategy, data leakage, and ML generalization. Table 1 summarizes representative studies that have shaped this progression, highlighting how the field has moved from defining and quantifying activity cliffs toward evaluating their impact on modern ML models and developing activity-cliff-aware learning strategies explicitly.

5. Toward Activity-Cliff-Aware QSAR: Evaluation, Reliability, and Future Directions

5.1. Activity Cliffs Should Be Treated as a Model-Diagnostic Problem

The evidence reviewed indicates that activity cliffs should not be considered merely problematic observations within otherwise well-behaved QSAR datasets. Instead, they provide a particularly demanding diagnostic of whether a model has learned fine-grained SARs or primarily broad correlations across chemical space. A model capable of accurately predicting chemically dissimilar compounds may still fail to distinguish two close analogues whose activities differ substantially. Consequently, conventional estimates of QSAR performance should be complemented by explicit characterization of the activity landscape on which the model is trained and evaluated.

This principle shifts model development from an exclusively algorithm-centered perspective toward a data-representation-model-evaluation framework. Before fitting a predictive model, the dataset itself should be assessed for chemical-space coverage, activity distribution, structural redundancy, assay consistency, and local SAR discontinuity. Activity-cliff frequency alone may not fully describe the difficulty of a dataset because cliffs can differ in magnitude, density, distribution, and chemical context. A dataset containing a small number of strategically located cliffs may present a greater modeling challenge than one containing a larger number of cliffs concentrated within well-sampled analogue series.

Quantitative measures of activity-landscape roughness provide one route toward evaluating this problem before model development. Aldeghi et al. introduced the Roughness Index (ROGI), which quantifies how molecular-property dispersion changes as molecules are progressively grouped according to structural similarity [46]. ROGI was found to correlate strongly with out-of-sample ML prediction error across multiple regression tasks, supporting the general expectation that

increasingly rugged molecular landscapes are more difficult to model. Such measures complement pairwise activity-cliff detection because they characterize the global geometry of a property landscape rather than individual exceptional compound pairs.

In practical terms, modelability assessment should therefore precede model comparison. If the activity landscape is intrinsically rough, modest predictive performance may reflect the difficulty of the underlying SAR rather than inadequate algorithm selection. Conversely, exceptional test performance on a highly discontinuous landscape should motivate careful examination of train-test redundancy, analogue overlap, or other potential sources of information leakage.

This concept becomes particularly important as increasingly complex molecular foundation models are applied to QSAR. A 2026 benchmark evaluating classical ML, GNNs, pretrained molecular models, AutoML, and foundation-model-based approaches under several realistic drug-discovery challenges found that all evaluated categories continued to struggle with activity cliffs [47]. Thus, improvements in generic representation learning do not eliminate the need to characterize dataset modelability explicitly.

5.2. Validation Should Separate Chemical Interpolation from Prospective Generalization

A central implication of activity-cliff research is that no single train-test split adequately characterizes molecular-model generalization. Random splitting evaluates performance under a largely interpolation-oriented scenario, whereas scaffold, similarity, cluster, temporal, and other distribution-shifted partitions examine progressively different forms of extrapolation.

This distinction should be made explicit rather than treating one split as universally correct. Random splitting remains informative when the intended application is prediction within an existing chemical series. Scaffold-based validation is more appropriate when the objective involves transferring predictions toward new core structures. Temporal splitting can approximate prospective drug-discovery conditions when compound synthesis or assay chronology is available. Similarity- or distance-controlled splitting can further expose performance as test molecules move away from known training chemistry.

Importantly, activity cliffs introduce another dimension beyond conventional structural extrapolation. Validation should ensure that cliff-related compound pairs do not create unintended information leakage between training and test sets. If compound A forms activity-cliff pairs with

Table 1. Representative studies shaping activity-cliff research and machine-learning generalization.

Focus / approach	Key finding	Relevance	Refs.
Conceptual perspective on QSAR outliers and SAR discontinuity	QSAR outliers may represent genuine steep activity-landscape regions rather than noise.	Established the conceptual basis of activity cliffs.	[17]
Introduced the Structure-Activity Landscape Index (SALI)	Quantified discontinuities between highly similar molecules with large activity differences.	Provided a quantitative framework for cliff detection.	[28]
Review of cliff definitions, detection, and medicinal-chemistry applications	Activity cliffs contain valuable SAR information for compound optimization.	Positioned cliffs as informative rather than undesirable outliers.	[18]
Roughness Index (ROGI) for molecular-property landscapes	Greater landscape roughness was associated with poorer out-of-sample predictability.	Linked activity-landscape complexity with modelability.	[46]
24 ML/DL approaches across 30 targets; MoleculeACE	Cliff compounds were harder to predict; complex DL models were not consistently superior.	Showed that model complexity alone does not solve cliffs.	[23]
Large-scale ML/DL activity-cliff classification	Train-test compound overlap inflated performance; stricter separation reduced accuracy.	Highlighted leakage and memorization risks.	[24]
ECFP, descriptors, and GIN with RF, kNN, and MLP	Unseen cliffs showed low sensitivity; GIN aided cliff detection, whereas ECFP remained strong for general QSAR.	Separated global QSAR accuracy from cliff-prediction ability.	[25]
Similarity- and SALI-based data splitting	Uneven chemical-space and cliff distributions reduced prediction quality.	Showed that validation design influences apparent generalization.	[39]
Foundation models, GNNs, AutoML, and classical ML	Modern model classes continued to struggle with activity cliffs.	Demonstrated persistence of the problem in contemporary molecular AI.	[47]
Cliff-aware GNN with triplet/contrastive learning; 52 datasets	Explicit cliff-aware representation learning improved prediction over standard baselines.	Represents the move toward activity-cliff-aware learning.	[27]

B, C, and D, placing A-B in training and A-C in testing does not constitute a fully independent evaluation of cliff generalization because the model has already been exposed to the anchor compound. Tamura et al. demonstrated that eliminating such compound overlap substantially reduced activity-cliff classification performance [24].

Activity-cliff-aware evaluation should therefore distinguish between at least two scenarios: known-analogue prediction and unseen-pair prediction. The first reflects lead optimization in which experimental information is already available for one member of an analogue relationship. The second tests whether a model can anticipate an activity cliff when both molecules are previously unseen. Dablander et al. showed that the latter setting is substantially more difficult, with modern QSAR models displaying poor cliff sensitivity [25].

Recent work also indicates that the relationship between splitting strategy and activity-cliff performance is not trivial. López-Pérez and Miranda-Quintana compared splitting procedures designed around chemical similarity and activity-landscape characteristics and found that nonuniform distributions of chemical space and cliffs between training and test sets generally produced poorer models [39]. However, random splitting remained competitive or superior in several evaluated settings,

emphasizing that splitting strategies should not be ranked independently of the scientific prediction problem. Instead, multiple validation scenarios should be reported when the intended application includes both interpolation and chemical-space expansion.

Model performance should therefore be assessed across random, scaffold-based, out-of-domain, and activity-cliff-focused evaluation settings. Comparing performance across these increasingly demanding scenarios can provide useful information about model reliability that would not be visible from a single test-set score.

In particular, the decline in performance from random splitting to out-of-domain evaluation can be used to characterize the model's generalization gap. A large decrease would indicate that strong performance under conventional random splitting does not necessarily translate into reliable predictions for compounds located outside the training chemical space.

Similarly, performance on activity-cliff compounds should be compared with performance on non-cliff compounds. The difference between these two groups can reveal the additional loss in predictive accuracy associated specifically with steep SAR regions. Such reporting would provide a more informative picture of model robustness than relying on a single global performance metric.

5.3. Global QSAR Metrics Should Be Complemented by Cliff-Specific Evaluation

Because activity cliffs occupy localized regions of chemical space, their impact can be obscured when model performance is averaged across a large dataset. Future QSAR benchmarks should therefore evaluate ordinary compounds and cliff-associated compounds separately.

For regression models, conventional metrics such as R^2 , RMSE, and MAE remain necessary, but they should be complemented by the same metrics calculated specifically for activity-cliff compounds. This allows researchers to determine whether overall model performance masks a disproportionate increase in prediction error within steep SAR regions.

A useful additional comparison is the ratio between prediction error for activity-cliff compounds and that for non-cliff compounds. Values close to one would indicate similar predictive reliability across smooth and cliff-containing regions, whereas substantially larger values would suggest that the model performs disproportionately poorly around activity cliffs. This ratio is proposed here as a simple interpretive quantity rather than as an established or standardized metric.

Pairwise evaluation can provide further insight into model behavior. For structurally related compound pairs, a reliable model should ideally reproduce not only the activities of the individual molecules but also the magnitude and direction of the observed difference in potency between them. The discrepancy between the observed and predicted activity differences can therefore be used as a direct measure of how well the model captures the local SAR transformation.

This type of evaluation is particularly relevant for activity cliffs because it assesses whether the model correctly represents the sharp activity change associated with a small structural modification. It therefore complements conventional compound-level metrics by examining the relational information that defines an activity cliff, rather than evaluating each molecule independently.

For cliff/non-cliff classification, metrics such as sensitivity, specificity, precision, F1 score, balanced accuracy, and Matthews correlation coefficient are particularly informative because activity cliffs may represent a minority class. Accuracy alone can become misleading in substantially imbalanced pairwise datasets. The low sensitivity observed in several contemporary activity-cliff studies reinforces the importance of reporting multiple metrics [24, 25].

Evaluation should also consider potency-direction accuracy. During medicinal-chemistry optimization, the practical question is frequently not whether an exact pK_i or pIC_{50} value can be predicted, but whether modifying compound A into compound B is expected to increase or decrease potency. A model that predicts approximate activities accurately but repeatedly reverses potency ordering among close analogues could provide misleading guidance during lead optimization.

Activity-cliff benchmarking should therefore move toward a hierarchical evaluation framework, progressing from global prediction to local prediction, cliff recognition, potency direction assessment, and ultimately prospective validation. This hierarchy would better reflect the increasingly difficult decisions encountered during molecular optimization.

5.4. Applicability Domain and Uncertainty: Need a Local SAR Perspective

Applicability-domain analysis has long been considered a central requirement for trustworthy QSAR modeling. Organization for Economic Co-operation and Development (OECD) guidance states that a QSAR should possess a defined domain of applicability describing the structural, physicochemical, and response space within which model predictions can reasonably be considered reliable [26]. Traditionally, applicability domains are commonly estimated using molecular similarity, descriptor distance, leverage, density, or related measures.

Activity cliffs reveal an important limitation of purely structural definitions of applicability. A query compound may be highly similar to its nearest training compound and would therefore be considered reliable by a conventional similarity-based applicability-domain method. However, if the two compounds are located in a steep SAR region, small structural changes may correspond to large differences in biological activity. In such cases, high structural similarity does not necessarily imply high predictive reliability.

A more robust applicability-domain framework should therefore consider not only how far a compound lies from known chemical space, but also the local characteristics of the surrounding activity landscape and the uncertainty associated with the model prediction. In particular, predictions made in locally smooth SAR regions should be distinguished from those made near known activity discontinuities, even when both compounds are well represented in the training chemical space.

This perspective highlights two distinct forms of high-risk prediction. The first occurs when a molecule is chemically distant from the training data and therefore represents a conventional extrapolation problem. The second occurs when a molecule lies within familiar chemical space but is located near a region of steep activity variation. Both situations may lead to unreliable predictions, although the underlying causes of failure are different.

Uncertainty quantification (UQ) is therefore particularly important in this context. Ideally, estimated uncertainty should increase when the probability of prediction error is high and should provide an indication of when model outputs require cautious interpretation. This is especially relevant for activity-cliff compounds, where predictions may appear chemically well supported despite substantial local uncertainty in biological activity.

However, available evidence suggests that existing UQ methods do not consistently identify activity-cliff-related prediction failures. Kötter et al. examined several commonly used uncertainty approaches and distinguished poor predictions caused by insufficient training-set representation from those occurring in steep SAR regions [38]. Several uncertainty methods were more successful at identifying conventional extrapolation errors than errors associated with steep structure-activity relationships. Consequently, QSAR models may remain highly confident even in regions where activity cliffs make prediction particularly difficult.

Future UQ approaches should therefore be evaluated separately across out-of-domain, activity-cliff, and locally smooth regions. Calibration performance and empirical coverage should also be reported for these different regions whenever sufficient data are available. Such analyses would clarify whether uncertainty estimates remain informative across different forms of generalization failure rather than appearing reliable only when performance is averaged across the entire test set.

5.5. Future Models Should Learn Molecular Differences, Not Only Individual Molecules

Conventional QSAR models are generally trained using compound-level observations in which each molecule is associated with an individual activity value. However, the defining information contained in an activity cliff is relational rather than purely compound-specific. The scientifically relevant signal lies in the structural transformation between two highly similar molecules and the corresponding difference in their biological activities. This suggests that future activity-cliff-aware modeling may benefit from shifting part of the learning objective from absolute property prediction toward relative property learning.

Pairwise, contrastive, triplet, ranking, and delta-learning strategies provide natural mechanisms for incorporating this relational information. Instead of training a model only to predict the absolute activity of individual compounds, the learning process could also require the model to reproduce the activity differences observed between structurally related molecules. Such an approach would explicitly encourage the model to recognize local SAR changes rather than treating each compound as an independent training example.

A combined learning strategy may therefore be particularly useful. The model could retain the conventional objective of predicting individual molecular activities while simultaneously learning whether structural modifications between related compounds lead to increases or decreases in activity and how large those changes are. The relative contribution of these two objectives could then be adjusted during model training depending on the importance assigned to global prediction accuracy and local SAR sensitivity.

This type of formulation may be especially valuable for activity-cliff prediction because it places greater emphasis on the molecular differences that generate abrupt changes in biological activity. By learning both absolute activities and local activity changes, future QSAR models may become better equipped to distinguish smooth SAR regions from discontinuous ones and to generalize more reliably when encountering closely related compounds with substantially different activities.

Recent activity-cliff-aware graph learning provides strong empirical support for this direction. Cui et al. introduced an activity-cliff-awareness framework using dynamically mined triplets and a contrastive objective that reshapes molecular latent space according to activity relationships [27]. Across 52 datasets, incorporating cliff-awareness improved prediction relative to conventional fingerprint and GNN baselines, while the approach remained transferable across different graph-learning backbones. The study also showed that ordinary regression objectives left many high-value cliff inconsistencies unresolved in learned latent spaces [27].

This is an important conceptual development. Activity cliffs may not require only a better molecular encoder; they may require different supervision. If the training objective treats every compound independently, the relative relationships that define cliffs are provided only indirectly. Explicit pairwise or contrastive supervision can make these relationships part of the learning target itself.

Several extensions deserve systematic investigation. Cliff-aware objectives could be integrated with molecular foundation models, graph transformers, multimodal

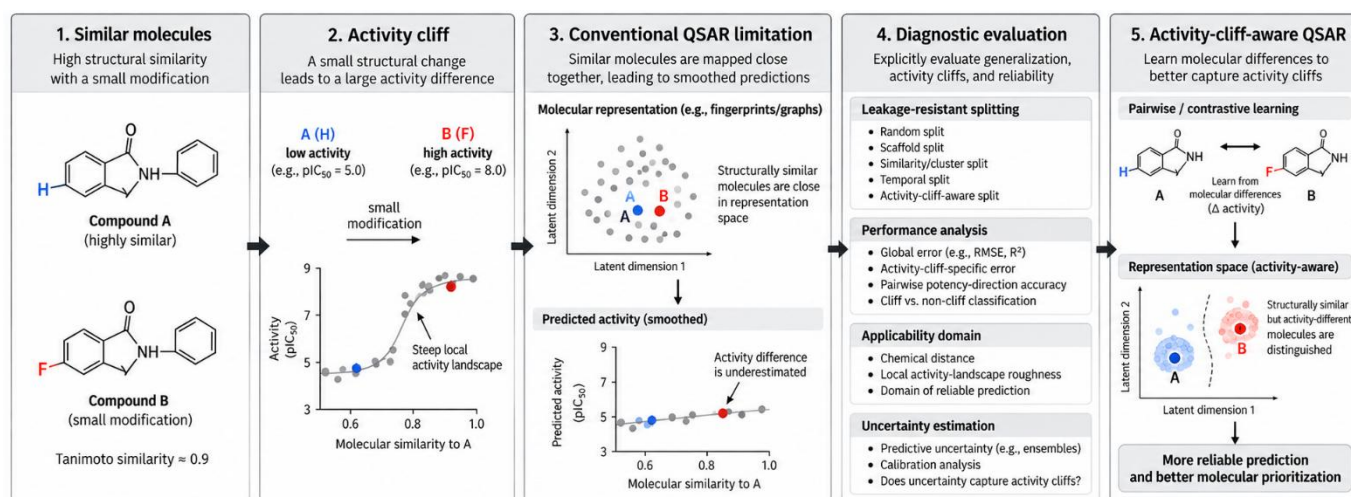


Figure 1. Conceptual framework illustrating how activity cliffs challenge QSAR generalization and how activity-cliff-aware modeling can improve model development, validation, and prospective prediction.

protein-ligand models, three-dimensional molecular representations, and multitask learning. Three-dimensional information may be particularly relevant when cliffs arise from conformational differences, stereochemistry, orientation within a binding site, or rearrangements of intermolecular interactions that are incompletely represented by 2D graphs.

Protein context also deserves greater attention. Many contemporary QSAR models remain ligand-centered, yet an activity cliff ultimately reflects a ligand-target interaction. The same molecular transformation may produce a cliff for one protein but a smooth SAR for another. Future models combining ligand representations with protein sequence, structure, binding-site descriptors, or protein-ligand interaction graphs may therefore help distinguish structural transformations whose biological effects are target-dependent.

5.6. A Proposed Activity-Cliff-Aware QSAR Framework

Based on the accumulated evidence, reliable QSAR development in cliff-containing datasets should be viewed as an integrated process rather than an algorithm-selection exercise. The proposed framework begins with rigorous assay and molecular curation, followed by explicit characterization of chemical-space redundancy and activity-landscape roughness. Activity cliffs should then be identified using reproducible structural and potency criteria, ideally supplemented by multiple definitions when sensitivity to the cliff threshold is scientifically relevant.

Model validation should include both interpolation-oriented and generalization-oriented splits. Compound overlap among activity-cliff pairs should be controlled explicitly, and results should distinguish situations in

which one analogue is already known from situations in which both members of a cliff are unseen.

Model performance should subsequently be reported on at least two levels: global QSAR performance and activity-cliff-specific performance. Where practical, this should be supplemented with pairwise potency-direction evaluation, applicability-domain assessment, and calibrated predictive uncertainty. Activity-landscape roughness should be interpreted alongside conventional structural distance so that chemically close but locally discontinuous regions are not incorrectly assumed to be reliable.

Finally, activity-cliff information should increasingly be incorporated during model training rather than used solely for post hoc error analysis. Contrastive learning, pairwise ranking, delta-property prediction, activity-aware representation learning, and related strategies offer routes for increasing learning pressure on precisely those regions that conventional global regression tends to underweight. The resulting conceptual workflow is illustrated in Figure 1.

6. Conclusions

Activity cliffs remain a major challenge for QSAR and molecular ML because small structural changes can produce large biological effects that are difficult to capture using conventional similarity-based representations and global learning objectives. Increasing model complexity alone does not consistently resolve this problem, as reliable performance also depends on molecular representation, training-data quality, chemical-space coverage, validation strategy, and learning objective. Activity-cliff-aware approaches that explicitly incorporate local SAR discontinuities into representation learning therefore represent a promising

direction for improving generalization. Future QSAR studies should complement global performance metrics with cliff-specific evaluation, applicability-domain analysis, and uncertainty assessment to identify where models remain unreliable. Overall, activity cliffs should be treated not merely as difficult outliers but as informative tests of whether a model can distinguish structurally similar compounds with substantially different biological activities and support more reliable molecular optimization.

Author Contributions: Conceptualization, T.R.N., M.P., and F.P.; methodology, T.R.N.; software, T.R.N.; validation, T.R.N., M.P., and F.P.; formal analysis, T.R.N.; investigation, T.R.N.; resources, T.R.N., M.P., and F.P.; data curation, T.R.N.; writing—original draft preparation, T.R.N.; writing—review and editing, T.R.N., M.P., and F.P.; visualization, T.R.N.; supervision, M.P. and F.P.; project administration, T.R.N.; funding acquisition, T.R.N. All authors have read and agreed to the published version of the manuscript.

Funding: This study does not receive external funding.

Ethical Clearance: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were generated or analyzed in this review article. The study is based exclusively on previously published literature and publicly available information cited in the manuscript.

Conflicts of Interest: All the authors declare no conflicts of interest.

References

- Keyvanpour, M. R., and Shirzad, M. B. (2021). An Analysis of QSAR Research Based on Machine Learning Concepts, *Current Drug Discovery Technologies*, Vol. 18, No. 1, 17–30. doi:10.2174/1570163817666200316104404.
- Noviandy, T. R., Idroes, G. M., Tallei, T. E., Handayani, D., and Idroes, R. (2024). QSAR Modeling for Predicting Beta-Secretase 1 Inhibitory Activity in Alzheimer's Disease with Support Vector Regression, *Malacca Pharmaceutics*, Vol. 2, No. 2, 79–85. doi:10.60084/mp.v2i2.226.
- Hansch, C., and Fujita, T. (1964). ρ - σ - π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure, *Journal of the American Chemical Society*, Vol. 86, No. 8, 1616–1626. doi:10.1021/ja01062a035.
- Noviandy, T. R., Maulana, A., Emran, T. B., Idroes, G. M., and Idroes, R. (2023). QSAR Classification of Beta-Secretase 1 Inhibitor Activity in Alzheimer's Disease Using Ensemble Machine Learning Algorithms, *Heca Journal of Applied Sciences*, Vol. 1, No. 1, 1–7. doi:10.60084/hjas.v1i1.12.
- Noviandy, T. R., Maulana, A., Idroes, G. M., Maulydia, N. B., Patwekar, M., Suhendra, R., and Idroes, R. (2023). Integrating Genetic Algorithm and LightGBM for QSAR Modeling of Acetylcholinesterase Inhibitors in Alzheimer's Disease Drug Discovery, *Malacca Pharmaceutics*, Vol. 1, No. 2, 48–54. doi:10.60084/mp.v1i2.60.
- Hesping, E., Chua, M. J., Pflieger, M., Qian, Y., Dong, L., Bachu, P., Liu, L., Kurz, T., Fisher, G. M., Skinner-Adams, T. S., Reid, R. C., Fairlie, D. P., Andrews, K. T., and Gorse, A.-D. J. P. (2022). QSAR Classification Models for Prediction of Hydroxamate Histone Deacetylase Inhibitor Activity against Malaria Parasites, *ACS Infectious Diseases*, Vol. 8, No. 1, 106–117. doi:10.1021/acsinfecdis.1c00355.
- Noviandy, T. R., Maulana, A., Idroes, G. M., Irvanizam, I., Subianto, M., and Idroes, R. (2023). QSAR-Based Stacked Ensemble Classifier for Hepatitis C NS5B Inhibitor Prediction, *2023 2nd International Conference on Computer System, Information Technology, and Electrical Engineering (COSITE)*, IEEE, 220–225. doi:10.1109/COSITE60233.2023.10250039.
- Noviandy, T. R., Maulana, A., Idroes, G. M., Suhendra, R., Afidh, R. P. F., and Idroes, R. (2024). An Explainable Multi-Model Stacked Classifier Approach for Predicting Hepatitis C Drug Candidates, *Sci*, Vol. 6, No. 4, 81. doi:10.3390/sci6040081.
- Breiman, L. (2001). Random Forests, *Machine Learning*, Vol. 45, No. 1, 5–32. doi:10.1023/A:1010933404324.
- Cortes, C., and Vapnik, V. (1995). Support-Vector Networks, *Machine Learning*, Vol. 20, No. 3, 273–297. doi:10.1007/BF00994018.
- Noviandy, T. R., Idroes, G. M., Maulana, A., Afidh, R. P. F., and Idroes, R. (2024). Optimizing Hepatitis C Virus Inhibitor Identification with LightGBM and Tree-Structured Parzen Estimator Sampling, *Engineering, Technology & Applied Science Research*, Vol. 14, No. 6, 18810–18817. doi:10.48084/etasr.8947.
- Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K., and Pande, V. (2018). MoleculeNet: A Benchmark for Molecular Machine Learning, *Chemical Science*, Vol. 9, No. 2, 513–530. doi:10.1039/C7SC02664A.
- Mayr, A., Klambauer, G., Unterthiner, T., Steijaert, M., Wegner, J. K., Cleulemans, H., Clevert, D.-A., and Hochreiter, S. (2018). Large-Scale Comparison of Machine Learning Methods for Drug Target Prediction on ChEMBL, *Chemical Science*, Vol. 9, No. 24, 5441–5451. doi:10.1039/C8SC00148K.
- Sheridan, R. P. (2013). Time-Split Cross-Validation as a Method for Estimating the Goodness of Prospective Prediction., *Journal of Chemical Information and Modeling*, Vol. 53, No. 4, 783–790. doi:10.1021/ci400084k.
- Wallach, I., and Heifets, A. (2018). Most Ligand-Based Classification Benchmarks Reward Memorization Rather than Generalization, *Journal of Chemical Information and Modeling*, Vol. 58, No. 5, 916–932. doi:10.1021/acs.jcim.7b00403.
- 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., and Barzilay, R. (2019). Analyzing Learned Molecular Representations for Property Prediction, *Journal of Chemical Information and Modeling*, Vol. 59, No. 8, 3370–3388. doi:10.1021/acs.jcim.9b00237.
- Maggiore, G. M. (2006). On Outliers and Activity Cliffs—Why QSAR Often Disappoints, *Journal of Chemical Information and Modeling*, Vol. 46, No. 4, 1535–1535. doi:10.1021/ci060117s.
- Stumpfe, D., Hu, Y., Dimova, D., and Bajorath, J. (2014). Recent Progress in Understanding Activity Cliffs and Their Utility in Medicinal Chemistry, *Journal of Medicinal Chemistry*, Vol. 57, No. 1, 18–28. doi:10.1021/jm401120g.
- Stumpfe, D., Hu, H., and Bajorath, J. (2019). Evolving Concept of Activity Cliffs, *ACS Omega*, Vol. 4, No. 11, 14360–14368. doi:10.1021/acsomega.9b02221.
- Sheridan, R. P., Karnachi, P., Tudor, M., Xu, Y., Liaw, A., Shah, F., Cheng, A. C., Joshi, E., Glick, M., and Alvarez, J. (2020). Experimental Error, Kurtosis, Activity Cliffs, and Methodology: What Limits the Predictivity of Quantitative Structure-Activity Relationship Models?, *Journal of Chemical Information and Modeling*, Vol. 60, No. 4, 1969–1982. doi:10.1021/acs.jcim.9b01067.
- Golbraikh, A., Muratov, E., Fourches, D., and Tropsha, A. (2014). Data Set Modelability by QSAR, *Journal of Chemical Information and Modeling*, Vol. 54, No. 1, 1–4. doi:10.1021/ci400572x.
- Chuang, K. V., Gunsalus, L. M., and Keiser, M. J. (2020). Learning Molecular Representations for Medicinal Chemistry, *Journal of*

- Medicinal Chemistry*, Vol. 63, No. 16, 8705–8722. doi:[10.1021/acs.jmedchem.0c00385](https://doi.org/10.1021/acs.jmedchem.0c00385).
23. van Tilborg, D., Alenicheva, A., and Grisoni, F. (2022). Exposing the Limitations of Molecular Machine Learning with Activity Cliffs, *Journal of Chemical Information and Modeling*, Vol. 62, No. 23, 5938–5951. doi:[10.1021/acs.jcim.2c01073](https://doi.org/10.1021/acs.jcim.2c01073).
 24. Tamura, S., Miyao, T., and Bajorath, J. (2023). Large-Scale Prediction of Activity Cliffs Using Machine and Deep Learning Methods of Increasing Complexity, *Journal of Cheminformatics*, Vol. 15, No. 1, 4. doi:[10.1186/s13321-022-00676-7](https://doi.org/10.1186/s13321-022-00676-7).
 25. Dablander, M., Hanser, T., Lambiotte, R., and Morris, G. M. (2023). Exploring QSAR Models for Activity-Cliff Prediction, *Journal of Cheminformatics*, Vol. 15, No. 1, 47. doi:[10.1186/s13321-023-00708-w](https://doi.org/10.1186/s13321-023-00708-w).
 26. Development, O. for E. C. and. (2014). *Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models, OECD Series on Testing and Assessment No. 69*, OECD, Paris. doi:[10.1787/9789264085442-en](https://doi.org/10.1787/9789264085442-en).
 27. Cui, C., Su, X., Zhang, Z., Velez-Arce, A., Wang, J., Shi, X., Zhang, Y., Wu, J., Chen, Y. Z., Zitnik, M., and Shen, W. (2026). Activity-Cliff Awareness Enables Robust Graph Learning for Molecular Property Prediction, *Nature Communications*, Vol. 17, No. 1, 9376. doi:[10.1038/s41467-026-75713-2](https://doi.org/10.1038/s41467-026-75713-2).
 28. Guha, R., and Van Drie, J. H. (2008). Structure–Activity Landscape Index: Identifying and Quantifying Activity Cliffs, *Journal of Chemical Information and Modeling*, Vol. 48, No. 3, 646–658. doi:[10.1021/ci7004093](https://doi.org/10.1021/ci7004093).
 29. Wassermann, A. M., Wawer, M., and Bajorath, J. (2010). Activity Landscape Representations for Structure–Activity Relationship Analysis, *Journal of Medicinal Chemistry*, Vol. 53, No. 23, 8209–8223. doi:[10.1021/jm100933w](https://doi.org/10.1021/jm100933w).
 30. Rogers, D., and Hahn, M. (2010). Extended-Connectivity Fingerprints, *Journal of Chemical Information and Modeling*, Vol. 50, No. 5, 742–754. doi:[10.1021/ci100050t](https://doi.org/10.1021/ci100050t).
 31. Hu, X., Hu, Y., Vogt, M., Stumpfe, D., and Bajorath, J. (2012). MMP-Cliffs: Systematic Identification of Activity Cliffs on the Basis of Matched Molecular Pairs, *Journal of Chemical Information and Modeling*, Vol. 52, No. 5, 1138–1145. doi:[10.1021/ci3001138](https://doi.org/10.1021/ci3001138).
 32. Hussain, J., and Rea, C. (2010). Computationally Efficient Algorithm to Identify Matched Molecular Pairs (MMPs) in Large Data Sets, *Journal of Chemical Information and Modeling*, Vol. 50, No. 3, 339–348. doi:[10.1021/ci900450m](https://doi.org/10.1021/ci900450m).
 33. Hu, Y., and Bajorath, J. (2012). Exploration of 3D Activity Cliffs on the Basis of Compound Binding Modes and Comparison of 2D and 3D Cliffs, *Journal of Chemical Information and Modeling*, Vol. 52, No. 3, 670–677. doi:[10.1021/ci300033e](https://doi.org/10.1021/ci300033e).
 34. Hu, H., and Bajorath, J. (2020). Activity Cliffs Produced by Single-Atom Modification of Active Compounds: Systematic Identification and Rationalization Based on X-Ray Structures, *European Journal of Medicinal Chemistry*, Vol. 207, 112846. doi:[10.1016/j.ejmech.2020.112846](https://doi.org/10.1016/j.ejmech.2020.112846).
 35. Stumpfe, D., and Bajorath, J. (2012). Frequency of Occurrence and Potency Range Distribution of Activity Cliffs in Bioactive Compounds, *Journal of Chemical Information and Modeling*, Vol. 52, No. 9, 2348–2353. doi:[10.1021/ci300288f](https://doi.org/10.1021/ci300288f).
 36. Bemis, G. W., and Murcko, M. A. (1996). The Properties of Known Drugs. 1. Molecular Frameworks, *Journal of Medicinal Chemistry*, Vol. 39, No. 15, 2887–2893. doi:[10.1021/jm9602928](https://doi.org/10.1021/jm9602928).
 37. Alshehade, S. A., Al Jabal, G., and Jebiril, I. H. (2026). Systematic Multi-Target QSAR Benchmarking: Machine Learning Algorithms, Molecular Descriptors, and Validation, *Journal of Chemical Information and Modeling*, Vol. 66, No. 15, 8923–8941. doi:[10.1021/acs.jcim.6c01237](https://doi.org/10.1021/acs.jcim.6c01237).
 38. Kötter, A., Singh, K., Matter, H., Hessler, G., and Grebner, C. (2025). Upgrading Reliability in Molecular Property Prediction by Robust Quantification of Uncertainty from Machine Learning Models, *Journal of Chemical Information and Modeling*, Vol. 65, No. 20, 10819–10831. doi:[10.1021/acs.jcim.5c00464](https://doi.org/10.1021/acs.jcim.5c00464).
 39. López-Pérez, K., and Miranda-Quintana, R. A. (2025). Extended Activity Cliffs-Driven Approaches on Data Splitting for the Study of Bioactivity Machine Learning Predictions, *Molecular Informatics*, Vol. 44, No. 1, e202400054. doi:[10.1002/minf.202400054](https://doi.org/10.1002/minf.202400054).
 40. Novianidy, T. R., Sufri, R., Setiawan, R., and Anisah, A. (2026). Comparative Analysis of Ensemble Machine Learning Models for QSAR-Based Prediction of Anticoagulant Activity in Thrombotic Disorders, *Heca Journal of Applied Sciences*, Vol. 4, No. 1, 88–95. doi:[10.60084/hjas.v4i1.393](https://doi.org/10.60084/hjas.v4i1.393).
 41. Novianidy, T. R., Patwekar, M., and Idroes, R. (2026). QSAR Modeling of Beta-2 Adrenergic Receptor Ligands Using Molecular Descriptor-Based Machine Learning, *Malacca Pharmaceutics*, Vol. 4, No. 1, 46–54. doi:[10.60084/mp.v4i1.394](https://doi.org/10.60084/mp.v4i1.394).
 42. Novianidy, T. R., Setiawan, R., Sufri, R., and Anisah, A. (2025). Optimizing Machine Learning Models for Cardiovascular Drug Candidate Screening with XGBoost and Successive Halving, *Open International Journal of Informatics*, Vol. 13, No. 2, 41–50. doi:[10.11113/oiji2025.13n2.339](https://doi.org/10.11113/oiji2025.13n2.339).
 43. Hu, W., Liu, B., Gomes, J., Zitnik, M., Liang, P., Pande, V., and Leskovec, J. (2019). Strategies for Pre-Training Graph Neural Networks, *ArXiv Preprint ArXiv:1905.12265*.
 44. Novianidy, T. R., Idroes, G. M., Patwekar, M., and Idroes, R. (2025). Fine-Tuning ChemBERTa for Predicting Activity of AXL Kinase Inhibitors in Oncogenic Target Modeling, *Grimsa Journal of Science Engineering and Technology*, Vol. 3, No. 2, 73–84. doi:[10.61975/gjset.v3i2.98](https://doi.org/10.61975/gjset.v3i2.98).
 45. Honda, S., Shi, S., and Ueda, H. R. (2019). Smiles Transformer: Pre-Trained Molecular Fingerprint for Low Data Drug Discovery, *ArXiv Preprint ArXiv:1911.04738*.
 46. Aldeghi, M., Graff, D. E., Frey, N., Morrone, J. A., Pyzer-Knapp, E. O., Jordan, K. E., and Coley, C. W. (2022). Roughness of Molecular Property Landscapes and Its Impact on Modellability, *Journal of Chemical Information and Modeling*, Vol. 62, No. 19, 4660–4671. doi:[10.1021/acs.jcim.2c00903](https://doi.org/10.1021/acs.jcim.2c00903).
 47. Zhao, D., Zhu, Y., Wu, Z., Wan, Y., Liu, X., Li, S., Xu, H., Hou, T., and Hsieh, C.-Y. (2026). Revisiting ADMET Prediction Reliability under Real-World Challenges in the Foundation Model Era, *Journal of Cheminformatics*, Vol. 18, No. 1, 95. doi:[10.1186/s13321-026-01217-2](https://doi.org/10.1186/s13321-026-01217-2).