ELSEVIER

Point of view of the current state of the activity cliff phenomenon focusing on the rationale, effects and potential solutions to handle the influence of activity cliffs in drug discovery.

Activity cliffs in drug discovery: Dr Jekyll or Mr Hyde?

Maykel Cruz-Monteagudo[1,2,3,4], José L. Medina-Franco[5], Yunierkis Pérez-Castillo[4], Orazio Nicolotti[6], M. Natália D.S. Cordeiro[2] and Fernanda Borges[1]

[1] CIQ, Department of Chemistry and Biochemistry, Faculty of Sciences, University of Porto, 4169-007 Porto, Portugal
[2] REQUIMTE, Department of Chemistry and Biochemistry, Faculty of Sciences, University of Porto, 4169-007 Porto, Portugal
[3] Centro de Estudios de Química Aplicada (CEQA), Faculty of Chemistry and Pharmacy, Central University of Las Villas, Santa Clara 54830, Cuba
[4] Molecular Simulation and Drug Design Group, Centro de Bioactivos Químicos (CBQ), Central University of Las Villas, Santa Clara 54830, Cuba
[5] Mayo Clinic, 13400 East Shea Boulevard, Scottsdale, AZ 85259, USA
[6] Dipartimento di Farmacia-Scienze del Farmaco, Università degli Studi di Bari 'Aldo Moro', Via Orabona 4, 70125 Bari, Italy

The impact activity cliffs have on drug discovery is double-edged. For instance, whereas medicinal chemists can take advantage of regions in chemical space rich in activity cliffs, QSAR practitioners need to escape from such regions. The influence of activity cliffs in medicinal chemistry applications is extensively documented. However, the 'dark side' of activity cliffs (i.e. their detrimental effect on the development of predictive machine learning algorithms) has been understudied. Similarly, limited amounts of work have been devoted to propose potential solutions to the drawbacks of activity cliffs in similarity-based approaches. In this review, the duality of activity cliffs in medicinal chemistry and computational approaches is addressed, with emphasis on the rationale and potential solutions for handling the 'ugly face' of activity cliffs.

Introduction

Activity cliffs or, more generally, property cliffs are pairs of compounds with high structural similarity but unexpectedly high activity (or property) difference [1]. For medicinal chemists the existence and applications of activity cliffs is obvious [2–4]. Every experienced medicinal chemist is aware of pairs of molecules with high structural similarity but very different activity [5,6]. However, efforts to measure and detect activity cliffs systematically in screening datasets using computational methods have raised the question as to whether activity cliffs exist or if they are just artifacts of the computational methods employed [7,8].

Corresponding author:. Cruz-Monteagudo, M. (gmailkelcm@yahoo.es), (maikelcm@uclv.edu.cu)

Maykel Cruz-Monteagudo is currently a postdoctoral researcher of CIQUP at the Department of Chemistry and Biochemistry, Faculty of Sciences, University of Porto. He received the Bachelor of Science degree in pharmaceutical sciences from the Central University of Las Villas, Cuba, in 2003; and his PhD (toxicology) in pharmaceutical sciences from the Faculty of Pharmacy, University of Porto, in 2010. His current research is devoted to the development and application of chemoinformatics approaches to drug discovery, focusing on the application of system chemical biology concepts to multitarget and multi-objective drug discovery. He has authored more than 30 publications in peer reviewed journals and two international book chapters.

José L. Medina-Franco received a Bachelor of Science degree in chemistry from the National Autonomous University of Mexico (UNAM) in 1998. That same year, he joined Procter & Gamble in Mexico City, working in the research and development department. He received a Master of Science degree in 2002 and a PhD degree in 2005, both from UNAM. During his doctoral studies, José conducted research at a major research center in Madrid, Spain; as well as at the University of North Carolina at Chapel Hill and at the University of Arizona. In 2005, he joined the University of Arizona as a postdoctoral fellow. José was named Assistant Member at the Torrey Pines Institute for Molecular Studies in Florida in August 2007 and was named Full Research Professor at UNAM in October 2012. In June 2013 he joined the Mayo Clinic. His research is focused on drug discovery. He is also interested in developing and applying chemoinformatic methods for the systematic analysis of structure–activity relationships and the classification and data mining of compound libraries.

Fernanda Borges is Associate Professor of the Department of Chemistry and Biochemistry, Faculty of Sciences, University of Porto, and Senior Researcher of CIQUP. She received her MSc and PhD (pharmaceutical chemistry) in pharmaceutical sciences from the Faculty of Pharmacy, University of Porto, Portugal. Her current research is focused on medicinal chemistry, namely in the design and development of drugs to be used in the prevention and/or therapy of neurodegenerative diseases. She has authored more than 170 publications in peer reviewed journals, eight international book chapters and three patents.

1359-6446/06/$ - see front matter http://dx.doi.org/10.1016/j.drudis.2014.02.003

Although there are numerous publications focused on defining or applying the concept of activity cliffs [6], the detrimental effects of activity cliffs on the application of QSAR and similarity-based approaches have been disproportionally understudied. The positive and negative effects of activity cliffs in drug discovery are well known – a 'duality' that deserves a closer look. In this review, we discuss such duality in medicinal chemistry and computational applications. We propose strategies to overcome the limitations and issues associated with activity cliffs in QSAR and similarity-based approaches. This paper exposes the authors' shared opinion on activity cliffs but it is not intended to provide definitive right or wrong answers. Instead, we aim to present a snapshot of the current state of the controversial duality of activity cliffs discussing, in an integrated manner, the perspectives of experts in the field of activity landscape modeling.

Activity landscapes and activity cliffs

Most of the widely accepted activity cliff definitions rely on the activity landscape concept. Activity landscapes and activity cliffs are rich in information for SAR studies [9–16]. An activity landscape represents a hypersurface in a biologically relevant chemical space resembling geographical maps where compound potency is added as a third dimension to a 2D projection of the chemical space [13]. In activity landscapes, smooth regions are associated with continuous SAR and represent areas where gradual changes in chemical structure induce moderate changes in biological activity. By contrast, rugged regions are associated with discontinuous SAR where small chemical modifications drastically change the biological response [11]. The extreme forms of SAR discontinuity are termed activity cliffs, which are formed by pairs of structurally similar compounds with large differences in potency [5,17].

Whereas discontinuous SAR and activity cliffs provide the basis for lead optimization [5,18], continuous and smooth SAR regions are prerequisites for the successful application of QSAR and similarity-based methods for scaffold hopping or simply as predictive tools [11,19]. These quantitative approaches rely on the similarity property principle (SPP) [20], which states that similar molecules should have similar activity, thus assuming the presence of continuous SAR. By contrast, in rugged and discontinuous SAR regions, the application of similarity-based methods is meaningless.

Several numerical analysis functions including the SAR index (SARI) [21] or the structure–activity landscape index (SALI) [22] have been introduced to quantify SAR discontinuity and to identify activity cliffs. These functions are useful for directly comparing compound and activity similarities [19]. The SALI approach is particularly suitable to detect activity cliffs in a dataset. However, the magnitude of the activity cliffs is not determined by the SALI metric because its values are compared on a relative scale. Consequently, a disadvantage of this approach is that cliffs detected at a certain cutoff might be irrelevant (shallow or pseudo cliffs) [5]. Subsequently, Stumpfe and Bajorath [5] highlighted the need for using discrete criteria to define activity cliffs including the applied similarity criterion, the potency measure and the magnitude of the potency difference. These experts recommend considering a pair of compounds as an activity cliff only if: (i) a pre-established similarity criterion is satisfied; (ii) one compound in the pair has potency in the nanomolar range; and (iii) there is at least a 100-fold difference in potency between the two compounds. Yet, these criteria can be modified depending on the goals of the study.

Even so, discrete definitions of activity cliffs have limitations. The type (i.e. IC_{50} or K_i) and quality of experimental measurements [23], the molecular representations chosen and the similarity metrics [24] can significantly influence the assessment of activity cliffs [14]. Thus, activity cliffs identified in a given chemical and biological space might not be conserved in a different reference space [25].

Medicinal chemists sometimes question activity cliffs defined using similarity approaches because of their limited chemical interpretability [5,14]. To address this issue, Bajorath and colleagues [26] have used the matched molecular pair (MMP) formalism [27]. A MMP is defined as a pair of compounds that only differ at a single site (represented by a substructure) such as a ring or an R group. Thus, to classify a molecule pair as an MMP-cliff the potency difference required remains essentially the same as that applied in similarity-based definitions. Instead, the difference in size of the exchanged fragments and their size is restricted to a predefined maximum number of non-hydrogen atoms that guarantee the level of structural similarity expected for an activity cliff [26]. Note that the MMP-cliff definition is not free from using predefined thresholds. Activity cliffs have also been defined on the basis of consistently defined scaffolds and the presence of different scaffold–R-group relationships [3,28,29]; or by calculating the 3D similarity between compound binding modes observed in the X-ray structure of ligand–target complexes [4,30]. An updated review of the existing activity cliff definitions can be found in [6,31], and references therein.

Activity cliffs: facts or artifacts?

As mentioned above, it has been argued that activity cliffs could be artifacts as a result of a structural description that is not relevant to the specific dataset in the context of the biological problem [7]. An illustrative description of this point of view is offered by Horvath in a rich article that familiarizes experimental chemists with QSAR [8]. Horvath mentions that two molecules codified by a set of descriptors that fail to account for the actual difference between them will be close in the structure space but will display a spectacular and unexplained difference of activity, artificially generating an activity cliff [8]. Therefore, it is not straightforward to differentiate this situation from the presence of 'true' activity cliffs. After all, even for the 'genuine' cases, activity cliffs can appear simply as a result of pharmacokinetic differences or even measurement or annotation errors [7].

Because activity cliffs might be produced by the selection of inadequate descriptors, it is tempting to try to remove activity cliffs from a dataset by searching for appropriate descriptors that smooth out the activity landscape. However, identifying appropriate descriptors is not straightforward because this approach involves effective feature selection or mapping methods, and/or the selection of adequate evaluation criteria.

Addressing activity cliffs with chemoinformatics

Finding the 'appropriate' descriptors to smooth an activity landscape is a dataset-dependent problem. Therefore, the search for a universal molecular representation is not a realistic expectation [32]. A potential solution to find suitable descriptors for a given

dataset could be the integration of feature-selection methods commonly used in chemoinformatics with numerical analysis functions to quantify SAR discontinuity such as SARI [21] or SALI [22]. So, the relevant descriptor subset is searched by trying to minimize and maximize the discontinuity and continuity of the SAR present in the training set.

The rationale behind the use of numerical analysis functions to quantify SAR discontinuity as objective functions for feature selection relies on taking advantage of two well-known facts in similarity analysis: (i) the representation dependence of activity landscapes; and (ii) the distinct characteristic of scalar descriptors commonly used in QSAR modeling to account for ligand–target interactions indirectly. In this way, the search is directed toward finding structural features rendering continuous SAR spaces. Prospective applications of QSAR and similarity-based analyses applied over such a continuous reference space should be more meaningful and predictive than those based on heterogeneous reference spaces.

The milestone work of Guha and Van Drie, directed to assess how well a modeling protocol captures a structure–activity landscape by means of SALI curves [19], is a documented example of the potential offered by numerical analysis functions as evaluation criteria to quantify SAR discontinuity. Unlike previous retrospective applications of the SALI approach, in this work it is demonstrated that these SALI curves can be prospectively applied to measure quantitatively the capability of a given model to capture the activity cliffs that are inherent in SAR.

Concerning mapping algorithms, one solution could be the use of different strategies to include information of activity cliffs in *k* nearest neighbor (*k*-NN) models to identify the *k* nearest neighbors reflecting the applicability domain (AD) of a prediction. In this case, the well-known limitations associated with the AD definition need to be considered [33,34]. Alternatively, we can split the training dataset into structural clusters to build more-reliable local or multidomain models, controlling the risks associated with the use of such models [35]. Finally, an alternative solution to search for the 'best' set of descriptors is to develop improved molecular representations capable of encoding relevant information and the ability of generating static or dynamic robust descriptor spaces. That is, once a robust descriptor space has been found for a given dataset it can be regarded as dynamic if proven to be robust enough to explain the SAR of a new dataset augmented with new screening data. Otherwise, the descriptor space found is robust but static. The introduction of *in silico* design and data analysis (ISIDA) property-labeled fragment descriptors is a good example pointing in this direction [36].

Global versus local molecular similarity

The sometimes divergent opinions of medicinal chemists and chemoinformaticians on the utility of activity cliffs (i.e. whether activity cliffs are desirable or undesirable features in datasets) are rooted on the application of different concepts of molecular similarity and the respective molecular similarity methods required for each task. Chemoinformaticians and, more specifically, practitioners of QSAR and pharmacophore-based methods are used to working with 'local' similarities, commonly expressed as scalar molecular descriptors that encode, for example, topological, constitutional or functional aspects of the molecular structure. By contrast, medicinal chemists are used to envision molecular similarity using 'global' or 'holistic' molecular representations, which are frequently encoded using fingerprints such as the MACCS keys [37] or the extended connectivity fingerprints [38]. Probably, this is also the reason why medicinal chemists easily accept the existence of activity cliffs, whereas not every chemoinformatician does. This rationale has been discussed in the recent paper of Stumpfe and Bajorath [39], which essentially states that every holistic similarity method must recognize two closely related analogs as being 'similar', even if one is active and the other one is not, owing to the violation of crucial receptor–ligand contact(s). This means that methods conceptually based on the SPP are truly ligand-centric and do not take interaction criteria into account. Although local and global views of similarity are employed in contemporary similarity searching, holistic molecular representations are the most frequently used so far. The most important reason is the scaffold hopping potential [40] of holistic similarity searching, which falls exactly into the applicability domain of the SPP.

Addressing molecular representation dependence

It is well-accepted that molecular representation is the most important parameter for defining activity cliffs [25,31]. As discussed above, one approach to address such dependence is using MMPs or discrete R-group substitutions around a core scaffold. Another approach proposed by Medina-Franco and colleagues [25,41–43] is using multiple representations [44] to derive a consensus activity landscape model [25]. Using this strategy, the SALI approach [22] was extended to compute consensus SALI values obtained over multiple, diverse and orthogonal 2D and 3D representations. This approach led to the identification of molecule pairs concurrently classified as activity cliffs by multiple representations encoding determinant topological, conformational and pharmacophore-based information. A similar data fusion strategy combining multiple 2D and 3D representations can be extended to other activity landscape methods such as the SARI approach proposed by Peltason and Bajorath [21] to derive a consensus SAR index.

Activity cliffs in medicinal chemistry

As elaborated above, in medicinal chemistry the activity cliff concept can be conveniently used in lead optimization where it is highly relevant for identifying small structural modifications associated with significant potency changes. A study devoted to capture SAR progression by comparing activity-cliff-dependent and -independent pathways [45] evidences the potential of exploiting the activity cliff concept in lead optimization projects. The study showed that most potent compounds were identified through activity-cliff-dependent pathways in comparison to cliff-independent ones. Specifically, pathways originating from 54% of all activity cliffs included the most potent dataset compounds, whereas pathways originating from only 28% of compounds not involved in activity cliffs successfully detected potent compounds, supporting the advantages of exploiting activity cliff information.

Additionally, the concept can be implemented on different types of molecular similarity-based computational analyses because it relies on just two types of similarity relationships (structural and potency similarities) that can be easily quantified.

However, in specific computational applications that rely on continuous SARs, activity cliffs impose a limitation because they represent exceptions to the SPP.

Specific applications and reviews of activity cliffs in SAR characterization and lead optimization have been extensively presented and discussed elsewhere. Table 1 summarizes representative and practical applications of the activity cliff concept. Although the list of examples is not exhaustive, Table 1 illustrates the evolution and the current state of investigations focused on the development and application of the activity cliff concept in medicinal chemistry and drug discovery.

Activity cliffs or instances that should be misclassified

Despite the fact that there is a large number of molecular descriptors available and the machine learning techniques for developing QSAR models are growing, their predictive capability is still limited. Significant mis-predictions of activity still arise among similar molecules even in cases where the overall predictivity is high. This observation made by Maggiora in 2006 [9] still holds and probably will in the near future. Maggiora points out in his editorial [9] that the reason why QSAR often disappoints is related to the nature of the underlying SAR. That is, the main assumption of QSAR and similarity-based approaches is SAR continuity, where the structure–activity landscape looks like gently rolling hills. Therefore, gradual changes in structure should necessarily lead to gradual changes in activity. However, systematic quantitative profiling of many different sets of active compounds has shown that the majority of global SARs are heterogeneous [21], that is: their activity landscapes contain gently sloped regions but also sharp and shallow cliffs.

Activity cliffs, machine learning and chemoinformatics

SAR continuity provides the fundamental basis for QSAR analyses. By contrast, SAR discontinuity has a direct detrimental effect on the prediction ability of QSAR models [9]. Although statistical learning methods played a protagonist role in QSAR development, at present machine learning algorithms are the most extended tools in chemoinformatics applications [46–48]. Similar to QSAR and similarity-based approaches, machine learning methods also rely on continuous SAR. The two general purposes for which machine learning is used in chemoinformatics are classification and data generalization. Here, machine learning is used to extract regularity from data (that is: the process to get a view of trends and patterns). In drug discovery, machine learning algorithms use SAR knowledge to generate classifications and generalizations that are conceptually meaningful [49]. In this context, activity cliffs represent exceptions or contradictions to the assumed continuous SAR of the dataset.

So, if the classification mechanism in machine learning is understood as a function that maps a description of an example (chemical structure encoded by molecular descriptors) to its label (i.e. a continuous value or a class membership) the negative effect of activity cliffs is clear. Most machine learning techniques just capture major trends ('rolling hills') and fail to recognize activity cliffs reducing the reliability of prospective predictions. Even for advanced techniques capable of handling nonlinear relationships such as neural networks or support vector machines (SVMs), it is difficult to identify activity cliffs. But even if the machine learning model succeeds in capturing most of the relevant activity cliffs it comes at a cost. A model that learns from a training dataset including a significant number of activity cliffs is prone to overfitting [50]. Finally, although it is highly desirable to strive for a machine learning model efficiently accounting for such variability, it is important to be aware that it is unrealistic to find such a learning algorithm. The current chemical and biological knowledge is still immature and on the basis of such incomplete information one cannot expect a perfect model. Thus, a better question could be how to deal with chemoinformatics data and the lack of generalization ability of prediction models trained on this type of data. This issue raises the question – how to develop predictive models with heterogeneous SARs.

Activity cliffs, outliers, noise and instances that should be misclassified

Finding a parallelism for the activity cliff phenomenon in the machine learning area and establishing a rationale for the negative effect of activity cliffs over the predictive capabilities of machine learning models is not a trivial task. However, Smith and Martinez introduced preprocessing instances that should be misclassified (PRISM), a novel filtering method that identifies 'instances that should be misclassified' (ISMs) using heuristics that predict how likely is that an instance will be misclassified (Box 1) [51]. ISM, the basic concept behind PRISM, seems to be the closest analog of the activity cliff concept in the machine learning arena. An instance is recognized as an ISM if, restricted to the information provided in the training data, the label assigned by the learning algorithm to that instance is more likely to be correct, even if its actual label is different. Unlike traditional outliers and class noise, ISMs exhibit a high degree of class overlap. That is: an ISM is close in the task space to other instances of different class.

Although ISMs have been defined in a classification context, a good analogy can be established between ISMs and activity cliffs. In machine learning, ISMs are similar instances with different labels in a region of the task space; in activity landscape modeling, activity cliffs represent molecules close in the chemical space with a large difference in activity. However, this analogy becomes almost perfect if, as proposed by Bajorath and colleagues [52], the activity cliff concept is extended to include inactive compounds, reinterpreting the activity landscape as an active or inactive classification task. In this context, activity cliffs can be understood as special cases of ISMs. The main difference relies on the explicit consideration of the degree of similarity between instances. In addition to class overlapping, a high degree of similarity is required to label a pair of instances with opposite classes as an activity cliff.

It is important to note that the rationale behind the PRISM algorithm is different from that of traditional QSAR where outliers are frequently removed after the models have been built. In PRISM, ISMs are identified and removed from the training set before any modeling effort. Outlier detection methods aim at finding anomalies in the data, whereas noise reduction methods attempt to identify and remove mislabeled instances. However, noise and outlier detection and removal are difficult because there is no universal definition of what an outlier actually is or if an instance is noisy or not. So, ISMs have to be differentiated from outliers and noise by resorting to their basic features as considered by the

TABLE 1

Leading publications reporting activity cliffs aimed at medicinal chemistry tasks.

Publication title	Pub. year	Pub. type[a]	Activity-cliff-related medicinal chemistry task	Approach	Refs
Structure–activity landscape index: identifying and quantifying activity cliffs	2008	M	Identification and quantification of activity cliffs.	Numerical quantification of activity cliffs in terms of the structure–activity landscape index (SALI).	[22]
Characterization of activity landscapes using 2D and 3D similarity methods: consensus activity cliffs	2009	M	Activity landscape characterization. Addressing the representation dependence of activity cliffs.	Definition of consensus activity cliffs by means of a consensus model of the activity landscape based on multiple 2D and 3D representations.	[25]
From structure–activity to structure–selectivity relationships: quantitative assessment, selectivity cliffs, and key compounds	2009	M	Structure–selectivity relationship (SSR) characterization.	Compound similarity and selectivity data are analyzed with the aid of network-like similarity graphs (NSGs).	[70]
Structural interpretation of activity cliffs revealed by systematic analysis of structure–activity relationships in analog series	2009	A	Structural interpretation of activity cliffs.	Different compound series are analyzed in combinatorial analog graphs to determine substitution patterns that introduce activity cliffs of varying magnitude.	[71]
Molecular scaffolds with high propensity to form multi-target activity cliffs	2010	A	Identification of molecular scaffolds with high propensity to form multitarget activity cliffs. Identification of potentially promiscuous candidate scaffolds during compound optimization efforts.	Exhaustive analysis of scaffolds and associated compound activity data in the ChemblDB and BindingDB databases.	[72]
Chemical substitutions that introduce activity cliffs across different compound classes and biological targets	2010	A	Identification of defined chemical changes with high propensity to introduce activity cliffs.	Application of the concept of matched molecular pairs to analyze systematically the ability of defined chemical changes to introduce activity cliffs.	[73]
Computational analysis of activity and selectivity cliffs	2011	M	Identification of local structure–activity relationships (SAR) and SSR environments. Identification of key compounds involved in the formation of activity and/or selectivity cliffs displaying structural features responsible of molecular selectivity.	Integrative computational approach combining a numerical scoring scheme and graphical visualization of molecular networks for the systematic analysis of SARs and SSRs of small molecules.	[17]
Design of multitarget activity landscapes that capture hierarchical activity cliff distributions	2011	M	Identification of single-, dual- and triple-target activity cliffs. Selection of compounds forming complex activity cliffs.	First activity landscape design integrating compound potency relationships across multiple targets in a formally consistent manner.	[74]
From activity cliffs to activity ridges: informative data structures for SAR analysis	2011	M	Extension of the activity cliff concept by introducing the concept of 'activity ridges' (two subsets of highly and weakly potent structurally analogous compounds that form all possible pairwise activity cliffs between them). Identification of compound subsets having high priority for SAR analysis.	Systematic analysis of 242 compound datasets by means of an information-theoretic approach devised to characterize the structural composition of activity ridges.	[75]
Comprehensive analysis of single- and multi-target activity cliffs formed by currently available bioactive compounds	2011	A	First systematic survey of single- and multi-target activity cliffs contained in currently available bioactive compounds. Identification of compounds providing a rich source of SAR information across many different target families.	The three major public domain compound repositories were analyzed – PubChem (http://pubchem.ncbi.nlm.nih.gov), Binding DB [76] and ChEMBL [77].	[78]
Multitarget structure–activity relationships characterized by activity-difference maps and consensus similarity measure	2011	M	Identification of multitarget activity cliffs. Multitarget scaffold hopping.	Dual and triple activity-difference (DAD/TAD) maps are employed for the systematic characterization of the SAR of a benchmark set of 299 compounds screened against dopamine, norepinephrine and serotonin transporters. Consensus activity cliffs and scaffold hops were quantified and represented using the mean SALI and consensus structure–activity similarity (SAS) maps.	[43]

TABLE 1 (*Continued*)

Publication title	Pub. year	Pub. type[a]	Activity-cliff-related medicinal chemistry task	Approach	Refs
From activity cliffs to target-specific scoring models and pharmacophore hypotheses	2011	M	Identification of structure-based activity cliffs. Supporting for the elucidation of key interacting atoms of the binding site. Development of pharmacophore hypotheses.	A new approach for the identification of structure-based activity cliffs (ISAC) by analyzing interaction energies of protein–ligand complexes.	[79]
Exploration of 3D activity cliffs on the basis of compound binding modes and comparison of 2D and 3D cliffs	2012	M	Exploration of 3D activity cliffs derived from comparisons of experimentally determined compound binding modes and comparison with 2D activity cliffs to aid in SAR analysis and mapping of crucial binding determinants in cases where sufficient structural information is available.	Crystallographic binding modes of beta-secretase 1 (BACE1) and factor Xa (FXa) inhibitors were systematically compared using a 3D similarity method taking conformational, positional and atomic property differences into account.	[30]
Searching for coordinated activity cliffs using particle swarm optimization	2012	M	Identification of 'coordinated activity cliffs' (compounds within groups of structural neighbors that form multiple cliffs with different partners, giving rise to local networks of cliffs in a dataset; representing centers of high SAR discontinuity and information content).	A systematic search of coordinated activity cliffs in different compound sets was conducted by using particle swarm optimization.	[65]
MMP-Cliffs: systematic identification of activity cliffs on the basis of matched molecular pairs	2012	M	Identification of chemically intuitive activity cliffs by considering well-defined substructure replacements instead of calculated similarity values.	Activity cliffs are defined on the basis of the matched molecular pair (MMP) formalism (MMP-cliffs). MMPs were systematically derived from public domain compounds, and MMP-cliffs were extracted from them.	[26]
Systematic identification and classification of three-dimensional activity cliffs	2012	A	Categorization of 3D activity cliffs (3D-cliffs) into different categories on the basis of crystallographic interaction patterns. Rationalization of activity cliffs at the level of ligand–target interactions.	Activity cliffs were systematically extracted from public domain X-ray structures of targets for which complexes with multiple ligands were available, following the concept of 3D-cliffs. Binding modes of ligands with well-defined potency measurements were compared in a pairwise manner, and their 3D similarity was calculated using a previously reported property density function-based method taking conformational, positional and chemical differences into account.	[4]
Extending the activity cliff concept: structural categorization of activity cliffs and systematic i dentification of different types of cliffs in the ChEMBL database	2012	A	Structural categorization of activity cliffs and systematic identification of different types of cliffs.	Assignment of activity cliffs on the basis of well-defined structural criteria.	[3]
Frequency of occurrence and potency range distribution of activity cliffs in bioactive compounds	2012	A	Analysis of the frequency of occurrence and potency range distribution of activity cliffs in bioactive compounds. General definition of activity cliffs for data mining.	Cliff formation was studied across the global potency range observed for qualifying bioactive compounds.	[80]
Prediction of activity cliffs using support vector machines	2012	M	Prediction of activity cliffs.	Activity cliffs are predicted by support vector machine (SVM) models in test calculations on different datasets.	[66]
Identification of multitarget activity ridges in high-dimensional bioactivity spaces	2012	M	Extension of the activity ridge concept to the multitarget case. SAR exploration of high-dimensional activity spaces.	Systematic analysis of a high-dimensional kinase inhibitor dataset released by Abbott Laboratories by means of a new representation format designed for these ridges based on a scaffold–target matrix and a scoring scheme developed to identify compounds that were most variably distributed across a multitarget ridge and displayed target differentiation potential.	[81]

TABLE 1 (*Continued*)

Publication title	Pub. year	Pub. type[a]	Activity-cliff-related medicinal chemistry task	Approach	Refs
Matched molecular pair analysis of small molecule microarray data identifies promiscuity cliffs and reveals molecular origins of extreme compound promiscuity	2012	A	Identification of 'promiscuity cliffs' (pairs of structural analogs with single-site substitutions that lead to large-magnitude differences in apparent compound promiscuity) and the substructures or small substructure transformations that are generally responsible for introducing promiscuity.	Patterns of compound promiscuity in a publicly available small molecule microarray dataset involving between 50 and 97 unrelated targets were analyzed utilizing the matched molecular pair formalism.	[82]
Exploring SAR continuity in the vicinity of activity cliffs	2012	A	Exploration of SAR continuity in the vicinity of prominent activity cliffs.	Different compound datasets were mined for the presence of SAR continuity within the vicinity of prominent activity cliffs by using a computational approach using particle swarm optimization to examine the structural neighborhood of activity cliffs for continuous SAR components.	[83]
Bioactivity landscape modeling: chemoinformatic characterization of structure–activity relationships of compounds tested across multiple targets	2012	M	Chemoinformatics characterization of multitarget SARs.	Structure multiple activity similarity (SmAS) maps and the structure multiple activity landscape index (SmALI) were employed for characterizing the SAR of three benchmark sets of compounds screened with different target families.	[42]
Identifying activity cliff generators of PPAR ligands using SAS maps	2012	M	Identification of activity cliff generators (molecular structure that has a high probability to form activity cliffs with molecules tested in the same biological assay).	SAS maps were used systematically to identify and analyze activity cliff generators present in a dataset of 168 compounds tested against three peroxisome-proliferator-activated receptor (PPAR) subtypes.	[61]
Scanning structure–activity relationships with structure–activity similarity and related maps: from consensus activity cliffs to selectivity switches	2012	R	Review of the development, practical applications, limitations and perspectives of the SAS and related maps that are intuitive and powerful informatic tools to analyze SPRs computationally.	–	[41]
Exploring activity cliffs in medicinal chemistry	2012	R	Detailed description and discussion of the multifaceted nature of activity cliffs, the underlying scientific concepts and the usefulness of the individual or systematic analysis of activity cliffs to extract useful information for medicinal chemistry programs.	–	[5]
Rapid scanning structure–activity relationships in combinatorial data sets: identification of activity switches	2013	M	Description of the SAR of combinatorial datasets with activity for two biological endpoints. Rapid identification of substitutions that have a large impact on activity and selectivity such as 'activity switches' (specific substitutions that have an opposite effect on the activity of the compounds against two targets) or single- and double-target 'R-cliffs' (compounds where a single or double substitution around the central scaffold dramatically modifies the activity for one or two targets, respectively).	Dual-activity difference (DAD) maps were applied for the visual and quantitative analysis of all pairwise comparisons of one, two or more substitutions around a molecular template on a set of 106 pyrrolidine bis-diketopiperazines tested against two formylpeptide receptors obtained from positional scanning deconvolution methods of mixture-based libraries.	[28]
Compound pathway model to capture SAR progression: comparison of activity cliff-dependent and -independent pathways	2013	M	Compound pathway model to monitor SAR progression in compound datasets designed to mimic compound optimization efforts. Determination of SAR information gain associated with activity cliffs.	Compound pathway model comprising different pre-defined pathway categories.	[45]
Introduction of target cliffs as a concept to identify and describe complex molecular selectivity patterns	2013	M	Description of complex molecular selectivity patterns by introducing and applying the concept of 'target cliff' (a pair of targets against which at least one compound displays a large difference in potency). Comparison of target cliffs and activity cliffs for the identification and prioritization of selective compounds revealing relevant SAR information.	Target cliffs and activity cliffs are systematically extracted from a data structure termed target-compound matrices.	[84]

TABLE 1 (*Continued*)

Publication title	Pub. year	Pub. type[a]	Activity-cliff-related medicinal chemistry task	Approach	Refs
Activity cliffs in PubChem confirmatory bioassays taking inactive compounds into account	2013	A	Extension of the activity cliff concept to take inactive compounds into consideration to provide an additional source of SAR information.	Activity cliffs formed between pairs of active compounds and pairs of active and inactive compounds were systematically analyzed on a per-assay basis. PubChem confirmatory bioassays were used as source of confirmed active and inactive compounds.	[52]
Do medicinal chemists learn from activity cliffs? A systematic evaluation of cliff progression in evolving compound data sets	2013	A	Inspection over time of the presence of analogs of activity cliffs in public databases to estimate the extent activity cliff information is utilized in practical medicinal chemistry.	Fifty-six compound datasets that evolved over time were assembled and searched for analogs of activity-cliff-forming compounds with further increased potency.	[18]
Recent progress in understanding activity cliffs and their utility in medicinal chemistry	2013	R	Updated review of the recent studies applying the activity cliff concept emphasizing those that are particularly relevant for medicinal chemistry applications. The general activity cliff definition as well as the aspects to consider in activity cliff analysis are revisited and detailed.	–	[6]
Activity cliffs: facts or artifacts?	2013	R	Integrated discussion of some of the major aspects that raise the question whether all the activity cliffs detected in compound datasets are facts or just artifacts attributed to the molecular representation and quantitative definition of 'high' structural similarity.	–	[7]
Advancing the activity cliff concept	2013	R	Updated and concise, yet detailed, discussion of the current understanding of activity cliffs. A refined activity cliff concept is also introduced to a general audience in drug development.	–	[31]
Prediction of individual compounds forming activity cliffs using emerging chemical patterns	2013	M	Prediction of individual compounds forming activity cliffs.	Single compounds having high or low potency are accurately predicted to participate in activity cliffs on the basis of emerging chemical patterns.	[85]

[a] M, A and R refer to whether the main results reported in the publication are focused on introducing a new method, application or review regarding the activity cliff issue.

respective detection methods. Box 1 shows a comparison of ISMs, outliers and noise instances in machine learning and activity landscape terms.

Smith and Martinez compared PRISM with three existing outlier detection methods and one noise reduction technique using 48 datasets and nine learning algorithms [51]. Removing instances identified by PRISM before training achieved the highest overall classification accuracy compared with the machine learning algorithms trained on the original datasets as well as with outliers removed by the other methods. Rather than focusing on correctly classifying the ISMs and arbitrarily adjusting the classification boundary, removing the ISMs before training allows the machine learning algorithm to focus on the instances that can be correctly classified. In other words, the removal of ISMs allows the learning process to focus on the observed patterns rather than memorizing samples following no pattern. In this work, the authors demonstrated that removing ISMs during training was the most effective strategy with a high percentage of instances being detected as ISMs. This approach can be extended to activity landscape modeling by removing activity cliffs to smooth out the activity landscape (*vide infra*).

In addition to the PRISM algorithm [51], other outlier and/or noise detection methods have been recently introduced demonstrating their ability to improve the prediction accuracy of machine learning models [53,54]. For example, Byeon *et al.* introduced a novel technique to enhance the quality of training data with a noisy dependent variable for binary classification [53]. The approach termed Genetic Algorithm Prototype Selection uses a genetic algorithm (GA) to create the set of suspicious noisy instances and prototype selection to identify the set of actual noisy instances. The authors compared the performance of GAPS with filtering methods implemented in Weka [55] on two synthetic datasets created from the machine learning repository of the University of California-Irvine (UCI; http://archive.ics.uci.edu/ml/datasets.html). Whereas the Weka-enhanced datasets achieve similar levels of classification accuracy to noisy datasets from the classifier, the GAPS approach reduced the classification error over noisy datasets by approximately 26% on average with different increasing noise levels. More recently, based on the borderline noise factor, Yang and Gao applied data-cleaning techniques to remove the classification borderline noise. This work compared three under-sampling methods to select the representative majority class examples and remove the distant samples, which are useless to form the decision boundary [54]. The experimental results on bench datasets showed that the proposed method can effectively improve the classification accuracy of minority classes while achieving better overall classification.

BOX 1

To identify each type of data exception, the corresponding method (Table I) takes into consideration the presence (+) or absence (−) of extreme data values, instance mislabeling or class overlapping. The latter is not considered (*) by outlier and noise detection methods. Based on these three attributes: outliers can be defined as data exceptions represented by extreme values on the descriptor and/or property spaces not attributable to mislabeling; noise represents data exceptions attributable to mislabeling; and ISMs are data exceptions not attributable to mislabeling characterized by a high degree of class overlap.

In Figure I the dot line represents a classification boundary established by a hypothetical classifier. Instances 1 and 2 are labeled as ISMs because they should be misclassified as a result of class overlapping. Traditional outlier approaches would detect instances 1 and 3 as outliers but would not consider instance 2 as an outlier because it is not sufficiently different from the (−) instances and class is not taken into account. Noise reduction techniques cannot remove instance 1 because it is sufficiently different from the (−) instances.

Assuming the 2D plot (plane) as the descriptor space. Let + be 'active' compounds and − be 'inactive' compounds. In this case, active compounds 1 and 2 are similar in the descriptor space to the inactive compounds in the dataset and therefore can be regarded as activity cliffs. Therefore, a SAR dataset can actually comprise compounds exhibiting significantly different molecular descriptor and/or potency values (outliers) that are not necessarily annotation and/or measurement errors (noise). By contrast, a SAR dataset can actually contain structurally similar compounds exhibiting significantly different potency values (activity cliffs or ISMs) that cannot necessarily be attributed to annotation errors (noise or pseudo cliffs).

TABLE I

Main attributes defining outliers, noise and ISMs considered by the respective detection methods.

	Extreme value	Mislabeling	Class overlapping
Outlier	+	−	*
Noise	−	+	*
ISM	±	−	+

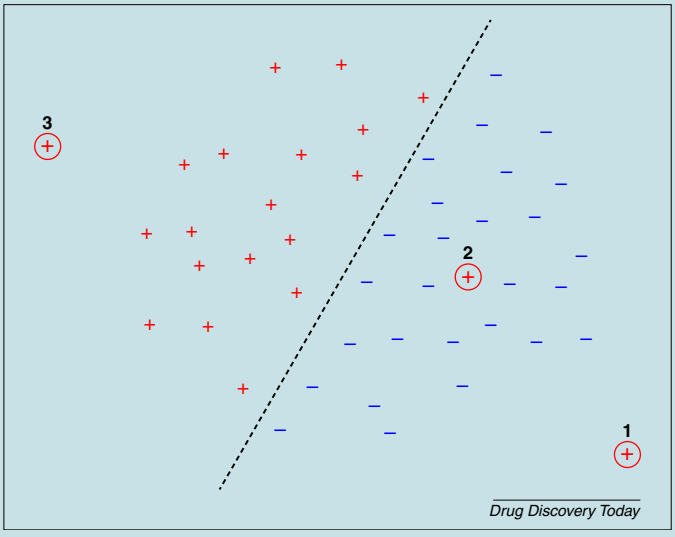

FIGURE I

A hypothetical 2D dataset illustrating ISMs, outliers and noise instances in machine learning and activity landscapes terms.

Because removing problematic instances (e.g. compounds) is a well-justified practice in machine learning to improve the prediction accuracy of models [51,53,54,56], it follows that it is reasonable to remove activity cliffs (smoothing out the activity landscape) for developing predictive QSAR models. The obvious drawback of this procedure is the inevitable loss of potentially crucial SAR information. However, this is the price that has to be paid in favor of the generalization ability of machine learning models, given that the primary goal is their use as prediction or virtual screening tools (i.e. in early hit identification stages). As such, although activity cliffs are not included in the predictive quantitative models, qualitatively it is still possible to interpret and take advantage of their rich SAR information content. This information can be complemented in further stages of drug discovery (i.e. lead optimization). In this way, the qualitative application of similarity analyses such as the network-like similarity graph (NSG) approach implemented on SARANEA [57] can complement previous SAR information derived from QSAR models. A recent example of the complementary use of QSAR and NSG approaches for SAR information mining is provided in Ref. [58]. Each stage of the drug discovery process imposes different priorities and such trade-off decisions highly influence its chance of success. Besides all this, other less obvious but also important issues on the ACG removal procedure will be addressed in the following section.

Activity cliffs and QSAR: resignation or remediation?

Despite the fact that the detrimental effect of activity cliffs in QSAR modeling is well accepted [59,60], to the best of our knowledge there are no reports directed at reducing SAR discontinuity on a dataset by removing activity cliffs and making the dataset amenable to QSAR and similarity-based methods. Similarly, there are no reports addressing whether the removal of activity cliffs from a dataset is beneficial, detrimental or nonsignificant for deriving general models of the SAR. So far, most of the chemoinformatics applications are focused on the description of the actual SAR and identification of activity cliffs [6,31]. As commented above, for practical purposes (i.e. for medicinal chemists) activity cliffs can provide key information to understand the SAR and guide lead optimization efforts [5,6].

Restoring SAR continuity by identification and removal of activity cliff generators

Arguably, the closest concept to an ISM in activity landscape modeling is that of an activity cliff generator (ACG), which is defined as a molecular structure that has a high probability of forming activity cliffs with molecules tested in the same biological assay [61]. In analogy with machine learning methods identifying and removing ISMs, one can propose identifying and removing ACGs. As noted above, removal of ACGs is different from removal of outliers in traditional QSAR.

Ideally, the concepts of consensus activity cliffs [25] or MMP-cliffs [26] should be applied for the identification of ACGs. The consensus ACGs identified using several representations should behave as such, irrespective of the reference space used or, at least, for most of the possible reference spaces. As discussed above, fingerprints are truly ligand-centric representations based on a global similarity definition and therefore better suited to codify global structure similarities. By contrast, scalar molecular descriptors codify local aspects of the molecular structure and can indirectly take into account ligand–target interaction information, and so might be better suited for an efficient SAR modeling [39].

Once the consensus ACGs previously identified have been removed, the goal is that the original training set fulfills the assumption made by the machine learning algorithm to be used for model construction, which is also the main premise of QSAR modeling. Additional curation [62] and balancing procedures [63,64] should also be applied to match the goals of the QSAR paradigm [9] and the machine learning algorithm [48]. As discussed above, removal of outliers in classical QSAR is an *a posteriori* procedure directed at improving the performance of a QSAR model and it depends on the reference space. By contrast, ACG detection and removal is done to optimize the training set for machine learning modeling before any modeling effort and it is independent of the reference space.

The essence of this solution is to remove from the training process those compounds responsible for the SAR discontinuity, and consequently restore the SAR continuity required for deriving reliable and predictive QSAR models (Fig. 1). However, the question that remains is to what extent the learning process is affected by the loss of the information encoded in the activity cliff pairs and therefore the generalization ability of the pattern found. In fact, according to Maggiora: 'some of the outliers may, in fact, be activity cliffs. Thus, removing such points would severely prejudice model's predictive capabilities' [9].

This observation brings some questions outlined below.

- Is it possible to develop models that predict activity cliffs? The answer to this question is yes. The prediction of coordinated activity cliffs was addressed by Namasivayam and Bajorath [65] using machine learning techniques; specifically, a particle swarm optimization guided by subset discontinuity scoring. Compounds forming the largest coordinated activity cliffs were automatically extracted from large compound datasets. Similarly, SVM models were derived to predict activity cliffs successfully [66]. In test calculations on different datasets, activity cliffs were accurately predicted using specifically designed structural representations and kernel functions.
- Is it feasible to preserve the model capabilities to predict SAR without removing the activity cliffs? There is not a definitive answer to this question. As can be noted (and expected), the very nature of activity cliffs hinders any possible consensus on how to deal with its negative impact on QSAR modeling. Even more discouraging, Maggiora concluded in his editorial [9] (referring to activity cliffs and other problems inherent to the QSAR approach) that: 'addressing all of these problems is a daunting task at best, and it may not be possible to treat some of them in any substantive way'. So, we have no other option than to try a procedure supposed to be valid but checking the ability of a model trained under these conditions to predict ACGs correctly.
- Will ACG removal enhance the performance of the models in prospective applications? Can this procedure help a medicinal chemist to prioritize compounds for synthesis and/or screen? A trade-off should be found when removing compounds because this implies a restored SAR continuity but also a reduction of AD. A possible solution could be to find a balance between the predictive ability of the newly developed QSAR models and their AD. By contrast, according to Guha [50] even if a machine learning model captures the most significant activity cliffs

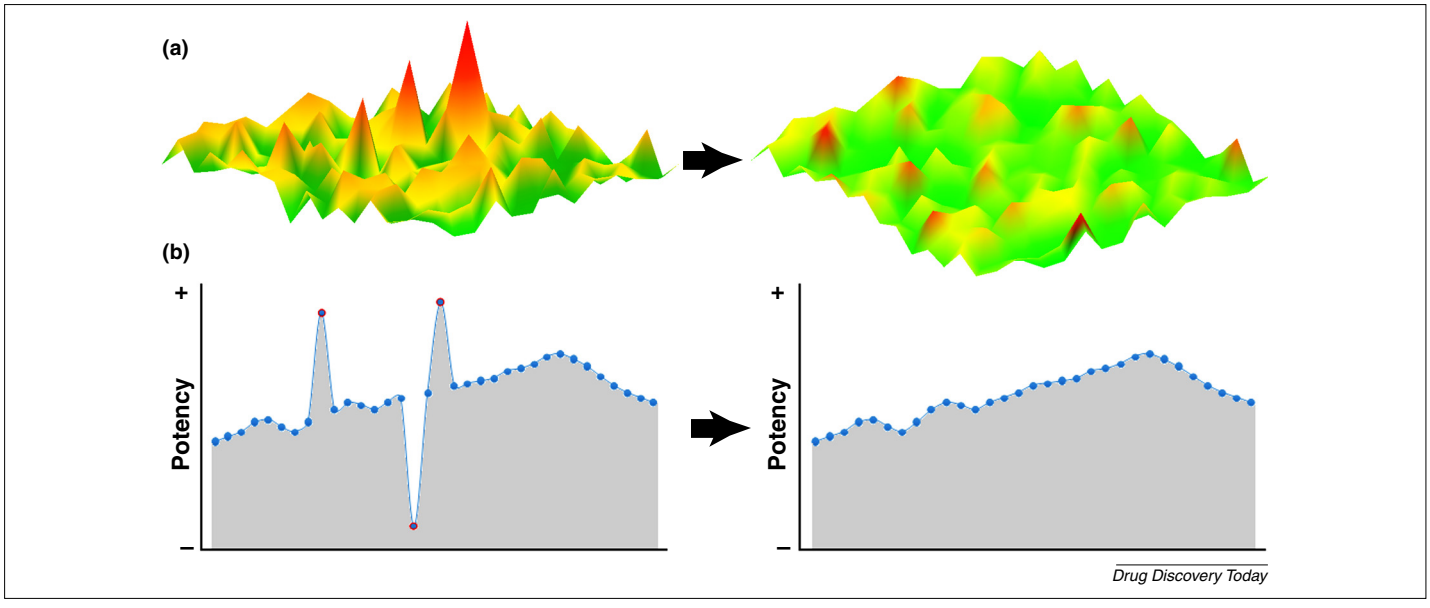

FIGURE 1

Representation of the SAR continuity restoration by detection and removal of activity cliff generators (ACGs) from the training set to be employed for machine learning modeling. **(a)** A hypothetical rugged activity landscape is smoothed out to restore SAR continuity. **(b)** 2D schematic representation of a hypothetical activity landscape and how a rugged one is smoothed out by detecting and removing ACGs. Compounds are represented by blue dots and ACGs are highlighted with a red circle.

present in the SAR landscape it comes at a cost. By requiring that the model encodes at least some of the significant activity cliffs, this introduces some degree of overfitting, because the most significant cliffs would correspond to discontinuities that would have to be memorized by the model. So, we might be forced to choose between a predictive model with a certain loss of applicability domain (by removing ACGs) or a model efficiently capturing the SAR but at the cost of a certain degree of overfitting, which hampers its predictive ability (by keeping ACGs).

A remedial measure to diminish the loss of applicability domain can be to develop several diverse machine learning models and implement a consensus classifier [67,68]. It is well known that multiple, ensemble or consensus classifiers are effective mainly because they span the decision space because each base classifier covers a different region of the decision space (chemical space or SAR) and the union of all the base classifiers produces a common region that results in a wider coverage of chemical space or applicability domain [67,69]. It is the authors' opinion that it is worth testing the hypothesis of ACG removal because reduction of the applicability domain seems to have a remedial solution, whereas overfitting does not.

Concluding remarks: Dr Jekyll or Mr Hyde?

In summary, the analyses provided in this review enable confirmation that the influence of activity cliffs on medicinal chemistry tasks has been extensively studied and documented. Activity cliffs are constantly faced in drug discovery and optimization efforts having a large influence on medicinal chemists to interpret SAR and decide what to synthesize and/or screen next. In addition, chemoinformatics approaches have been developed to detect activity cliffs in datasets systematically and efforts are being made to predict the presence of activity cliffs. However, the detrimental effects on the application of QSAR and similarity-based approaches have been disproportionally understudied. So, to fill this gap regarding the activity cliff problem, more studies must be conducted not only to study the effect of activity cliffs over the generalization ability of machine learning models but also to provide potential solutions to overcome such limitation. Therefore, from a chemoinformatics point of view and paraphrasing the fictional character Sheldon Cooper from the popular television series *The Big Bang Theory*: 'It's high time to address the *tweepadock* in the room'.

The impact of activity cliffs on chemoinformatics and medicinal chemistry applications has two 'faces'. Activity cliffs provide medicinal chemists with fundamental information to understand the underlying SAR of the dataset, which could significantly contribute to lead optimization efforts. However, the presence of activity cliffs seriously affects the generalization ability of machine learning models. One can work with such 'duality' of the activity cliffs depending on the goals of the project (e.g. using activity cliffs to retrieve fundamental SAR information or remove activity cliffs to smooth out the landscape and develop quantitative models for prospective applications). Therefore, activity cliffs, like Dr Henry Jekyll in the famous novel by Robert Louis Stevenson, will unavoidably have to co-exist with their alter ego. But, unlike the novel, in real life drug discovery there is no miraculous formula to isolate the good from the dark side of activity cliffs. Finally, we conclude that the role of activity cliffs in drug discovery is neither akin to Dr Jekyll nor Mr Hyde – it is more likely to be a combination of the two.

Conflicts of interest

The authors declare no conflicts of interest.

Acknowledgments

M.C-M. acknowledges the Fundação para a Ciência e a Tecnologia (FCT), Portugal, for grant: [SFRH/BPD/90673/2012], co-financed by the European Social Fund.

References

1 Medina-Franco, J.L. *et al.* (2012) Consensus models of activity landscapes. In *Statistical Modelling of Molecular Descriptors in QSAR/QSPR* (Dehmer, M. *et al.* eds), pp. 307–326, Wiley-VCH Verlag

2 Hu, Y. and Bajorath, J. (2013) Activity profile relationships between structurally similar promiscuous compounds. *Eur. J. Med. Chem.* 69C, 393–398

3 Hu, Y. and Bajorath, J. (2012) Extending the activity cliff concept: structural categorization of activity cliffs and systematic identification of different types of cliffs in the ChEMBL database. *J. Chem. Inf. Model.* 52, 1806–1811

4 Hu, Y. *et al.* (2012) Systematic identification and classification of three-dimensional activity cliffs. *J. Chem. Inf. Model.* 52, 1490–1498

5 Stumpfe, D. and Bajorath, J. (2012) Exploring activity cliffs in medicinal chemistry. *J. Med. Chem.* 55, 2932–2942

6 Stumpfe, D. *et al.* (2013) Recent progress in understanding activity cliffs and their utility in medicinal chemistry. *J. Med. Chem.* 57, 18–28

7 Medina-Franco, J.L. (2013) Activity cliffs: facts or artifacts? *Chem. Biol. Drug. Des.* 81, 553–556

8 Horvath, D. (2010) Quantitative structure–activity relantionships: *in silico* chemistry or high tech alchemy? *Rev. Roumaine Chim.* 55, 783–801

9 Maggiora, G.M. (2006) On outliers and activity cliffs – why QSAR often disappoints. *J. Chem. Inf. Model.* 46, 1535

10 Peltason, L. and Bajorath, J. (2007) Molecular similarity analysis uncovers heterogeneous structure–activity relationships and variable activity landscapes. *Chem. Biol.* 14, 489–497

11 Bajorath, J. *et al.* (2009) Navigating structure-activity landscapes. *Drug Discov. Today* 14, 698–705

12 Peltason, L. and Bajorath, J. (2009) Systematic computational analysis of structure–activity relationships: concepts, challenges and recent advances. *Future Med. Chem.* 1, 451–466

13 Peltason, L. *et al.* (2010) Rationalizing three-dimensional activity landscapes and the influence of molecular representations on landscape topology and the formation of activity cliffs. *J. Chem. Inf. Model.* 50, 1021–1033

14 Wassermann, A.M. *et al.* (2010) Activity landscape representations for structure–activity relationship analysis. *J. Med. Chem.* 53, 8209–8223

15 Iyer, P. *et al.* (2011) SAR monitoring of evolving compound data sets using activity landscapes. *J. Chem. Inf. Model.* 51, 532–540

16 Bajorath, J. (2012) Modeling of activity landscapes for drug discovery. *Expert Opin. Drug Discov.* 7, 463–473

17 Peltason, L. and Bajorath, J. (2011) Computational analysis of activity and selectivity cliffs. *Methods Mol. Biol.* 672, 119–132

18 Dimova, D. *et al.* (2013) Do medicinal chemists learn from activity cliffs? A systematic evaluation of cliff progression in evolving compound data sets. *J. Med. Chem.* 56, 3339–3345

19 Guha, R. and Van Drie, J.H. (2008) Assessing how well a modeling protocol captures a structure–activity landscape. *J. Chem. Inf. Model.* 48, 1716–1728

20 Johnson, M.A. and Maggiora, G.M., eds) (1990) *Concepts and Applications of Molecular Similarity*, John Wiley & Sons

21 Peltason, L. and Bajorath, J. (2007) SAR index: quantifying the nature of structure–activity relationships. *J. Med. Chem.* 50, 5571–5578

22 Guha, R. and Van Drie, J.H. (2008) Structure–activity landscape index: identifying and quantifying activity cliffs. *J. Chem. Inf. Model.* 48, 646–658

23 Stumpfe, D. and Bajorath, J. (2011) Assessing the confidence level of public domain compound activity data and the impact of alternative potency measurements on SAR analysis. *J. Chem. Inf. Model.* 51, 3131–3137

24 Dimova, D. *et al.* (2013) Quantifying the fingerprint descriptor dependence of structure–activity relationship information on a large scale. *J. Chem. Inf. Model.* 53, 2275–2281

25 Medina-Franco, J.L. *et al.* (2009) Characterization of activity landscapes using 2D and 3D similarity methods: consensus activity cliffs. *J. Chem. Inf. Model.* 49, 477–491

26 Hu, X. *et al.* (2012) MMP-cliffs: systematic identification of activity cliffs on the basis of matched molecular pairs. *J. Chem. Inf. Model.* 52, 1138–1145

27 Kenny, P.W. and Sadowski, J. (2004) Structure modification in chemical databases. In *Chemoinformatics in Drug Discovery* (Oprea, T.I., ed.), pp. 271–285, Wiley-VCH

28 Medina-Franco, J.L. *et al.* (2013) Rapid scanning structure-activity relationships in combinatorial data sets: identification of activity switches. *J. Chem. Inf. Model.* 53, 1475–1485

29 Aguayo-Ortiz, R. *et al.* (2013) Chemoinformatic characterization of activity and selectivity switches of antiprotozoal compounds. *Future Med. Chem.* http://dx.doi.org/10.4155/fmc.13.173

30 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. *J. Chem. Inf. Model.* 52, 670–677

31 Hu, Y. *et al.* (2013) Advancing the activity cliff concept. *F1000Res.* 2, 1–11

32 Wolpert, D.H. (1996) The lack of a prior distinctions between learning algorithms and the existence of a priori distinctions between learning algorithms. *Neural Comput.* 8, 1341–1390

33 Jaworska, J. *et al.* (2005) QSAR applicabilty domain estimation by projection of the training set descriptor space: a review. *Altern. Lab. Anim.* 33, 445–459

34 Gissi, A. *et al.* (2014) An alternative QSAR-based approach for predicting the bioconcentration factor for regulatory purposes. *ALTEX* 31, 23–36

35 Helgee, E.A. *et al.* (2010) Evaluation of quantitative structure–activity relationship modeling strategies: local and global models. *J. Chem. Inf. Model.* 50, 677–689

36 Ruggiu, F. *et al.* (2010) ISIDA property-labelled fragment descriptors. *Mol. Inf.* 29, 855–868

37 Durant, J.L. *et al.* (2002) Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.* 42, 1273–1280

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

39 Stumpfe, D. and Bajorath, J. (2011) Similarity searching. *WIREs Comput. Molec. Sci.* 1, 260–282

40 Schneider, G. *et al.* (1999) 'Scaffold hopping' by topological pharmacophore search: a contribution to virtual screening. *Angew. Chem. Int. Edit.* 38, 2894–2896

41 Mendez-Lucio, O. *et al.* (2012) Activity landscape modeling of PPAR ligands with dual-activity difference maps. *Bioorg. Med. Chem.* 20, 3523–3532

42 Waddell, J. and Medina-Franco, J.L. (2012) Bioactivity landscape modeling: chemoinformatic characterization of structure–activity relationships of compounds tested across multiple targets. *Bioorg. Med. Chem.* 20, 5443–5452

43 Yongye, A.B. *et al.* (2011) Consensus models of activity landscapes with multiple chemical, conformer, and property representations. *J. Chem. Inf. Model.* 51, 1259–1270

44 Sheridan, R.P. and Kearsley, S.K. (2002) Why do we need so many chemical similarity search methods? *Drug Discov. Today* 7, 903–911

45 Stumpfe, D. *et al.* (2013) Compound pathway model to capture SAR progression: comparison of activity cliff-dependent and -independent pathways. *J. Chem. Inf. Model.* 53, 1067–1072

46 Aguiar-Pulido, V. *et al.* (2013) Evolutionary computation and QSAR research. *Curr. Comput. Aided Drug Des.* 9, 206–225

47 Ivanciuc, O. (2009) Drug design with machine learning. In *Encyclopedia of Complexity and System Science* (Meyers, R.A., ed.), pp. 2159–2196, Springer-Verlag

48 Witten, I.H. and Frank, E., eds) (2005) *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann

49 Rose, J. (2003) Methods for data analysis. In *Handbook of Chemoinformatics*, (vol. 3) (Gasteier, J., ed.), pp. 1081–1097, Wiley-VCH

50 Guha, R. (2011) The ups and downs of structure–activity landscapes. *Methods Mol. Biol.* 672, 101–117

51 Smith, M.R. and Martinez, T. (2011) Improving classification accuracy by identifying and removing instances that should be misclassified. *The 2011 International Joint Conference on Neural Networks, IEEE* pp. 2690–2697

52 Hu, Y. *et al.* (2013) Activity cliffs in PubChem confirmatory bioassays taking inactive compounds into account. *J. Comput. Aided Mol. Des.* 27, 115–124

53 Byeon, B. *et al.* (2008) Enhancing the quality of noisy training data using a genetic algorithm and prototype selection. *The 2008 International Conference on Artificial Intelligence, IEEE*

54 Yang, Z. and Gao, D. (2013) Classification for imbalanced and overlapping classes using outlier detection and sampling techniques. *Appl. Math. Inf. Sci.* 7, 375–381

55 Frank, E. *et al.* (2004) Data mining in bioinformatics using Weka. *Bioinformatics* 20, 2479–2481

56 Smith, M.R. *et al.* (2014) An instance level analysis of data complexity. *Mach Learning* http://dx.doi.org/10.1007/s10994-013-5422-z

57 Lounkine, E. *et al.* (2010) SARANEA: a freely available program to mine structure–activity and structure–selectivity relationship information in compound data sets. *J. Chem. Inf. Model.* 50, 68–78

58 Cruz-Monteagudo, M. and Cordeiro, M.N. (2014) Chemoinformatics profiling of ionic liquids – uncovering structure–cytotoxicity relationships with network-like similarity graphs. *Toxicol. Sci.* http://dx.doi.org/10.1093/toxsci/kft210

59 Tropsha, A. (2010) Best practices for QSAR model development, validation, and Exploitation. *Mol. Inf.* 29, 476–488

60 Scior, T. *et al.* (2009) How to recognize and work around pitfalls in QSAR studies: a critical review. *Curr. Med. Chem.* 16, 4297–4313

61 Méndez-Lucio, O. *et al.* (2012) Identifying activity cliff generators of PPAR ligands using SAS maps. *Mol. Inf.* 31, 837–846

62 Fourches, D. *et al.* (2010) Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. *J. Chem. Inf. Model.* 50, 1189–1204

63 Japkowicz, N. (2000) Learning from imbalanced data sets: a comparison of various solutions. *In AAAI'2000 Workshop on Learning from Imbalanced Data Sets*

64 Japkowicz, N. (2000) The class imbalance problem: significance and strategies. *International Conference on Artificial Intelligence (ICAI'2000)*

65 Namasivayam, V. and Bajorath, J. (2012) Searching for coordinated activity cliffs using particle swarm optimization. *J. Chem. Inf. Model.* 52, 927–934

66 Heikamp, K. *et al.* (2012) Prediction of activity cliffs using support vector machines. *J. Chem. Inf. Model.* 52, 2354–2365

67 Polikar, R. (2006) Ensemble based systems in decision making. *IEEE Circuit Syst. Mag.* 6, 21–44

68 Zhang, L. *et al.* (2013) Discovery of novel antimalarial compounds enabled by QSAR-based virtual screening. *J. Chem. Inf. Model.* 53, 475–492

69 Kuncheva, L.I., ed. (2004) *Combining Pattern Classifiers, Methods and Algorithms*, Wiley Interscience

70 Peltason, L. *et al.* (2009) From structure–activity to structure–selectivity relationships: quantitative assessment, selectivity cliffs, and key compounds. *ChemMedChem* 4, 1864–1873

71 Sisay, M.T. *et al.* (2009) Structural interpretation of activity cliffs revealed by systematic analysis of structure–activity relationships in analog series. *J. Chem. Inf. Model.* 49, 2179–2189

72 Hu, Y. and Bajorath, J. (2010) Molecular scaffolds with high propensity to form multi-target activity cliffs. *J. Chem. Inf. Model.* 50, 500–510

73 Wassermann, A.M. and Bajorath, J. (2010) Chemical substitutions that introduce activity cliffs across different compound classes and biological targets. *J. Chem. Inf. Model.* 50, 1248–1256

74 Dimova, D. *et al.* (2011) Design of multitarget activity landscapes that capture hierarchical activity cliff distributions. *J. Chem. Inf. Model.* 51, 258–266

75 Vogt, M. *et al.* (2011) From activity cliffs to activity ridges: informative data structures for SAR analysis. *J. Chem. Inf. Model.* 51, 1848–1856

76 Liu, T. *et al.* (2007) Binding-DB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.* 35, D198–D201

77 Overington, J. (2009) ChEMBL. An interview with John Overington, team leader, chemogenomics at the European Bioinformatics Institute Outstation of the European Molecular Biology Laboratory (EMBL. EBI). Interview by Wendy A. Warr. *J. Comput. Aided. Mol. Des.* 23, 195–198

78 Wassermann, A.M. *et al.* (2011) Comprehensive analysis of single- and multi-target activity cliffs formed by currently available bioactive compounds. *Chem. Biol. Drug. Des.* 78, 224–228

79 Seebeck, B. *et al.* (2011) From activity cliffs to target-specific scoring models and pharmacophore hypotheses. *ChemMedChem* 6, 1630–1639

80 Stumpfe, D. and Bajorath, J. (2012) Frequency of occurrence and potency range distribution of activity cliffs in bioactive compounds. *J. Chem. Inf. Model.* 52, 2348–2353

81 Gupta-Ostermann, D. and Bajorath, J. (2012) Identification of multitarget activity ridges in high-dimensional bioactivity spaces. *J. Chem. Inf. Model.* 52, 2579–2586

82 Dimova, D. *et al.* (2012) Matched molecular pair analysis of small molecule microarray data identifies promiscuity cliffs and reveals molecular origins of extreme compound promiscuity. *J. Med. Chem.* 55, 10220–10228

83 Namasivayam, V. *et al.* (2012) Exploring SAR continuity in the vicinity of activity cliffs. *Chem. Biol. Drug. Des.* 79, 22–29

84 Hu, Y. and Bajorath, J. (2013) Introduction of target cliffs as a concept to identify and describe complex molecular selectivity patterns. *J. Chem. Inf. Model.* 53, 545–552

85 Namasivayam, V. *et al.* (2013) Prediction of individual compounds forming activity cliffs using emerging chemical patterns. *J. Chem. Inf. Model.* 53, 3131–3139