Evolving Concept of Activity Cliffs

Dagmar Stumpf, Huabin Hu, and Jürgen Bajorath*

Department of Life Science Informatics, b-it, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Endenicher Allee 19c, D-53115 Bonn, Germany

ABSTRACT: Activity cliffs (ACs) are generally defined as pairs or groups of structurally similar compounds that are active against the same target but have large differences in potency. Accordingly, ACs capture chemical modifications that strongly influence biological activity. Therefore, they are of particular interest in structure–activity relationship (SAR) analysis and compound optimization. The AC concept is much more complex than it may appear at a first glance, especially if one aims to represent ACs computationally and identify them systematically. To these ends, molecular similarity and potency difference criteria must be carefully considered for AC assessment. Furthermore, ACs are often perceived differently in medicinal and computational chemistry, depending on whether they are studied on a case-by-case basis or systematically. For practical applications, intuitive access to AC information plays a major role. Over the years, the AC concept has been further refined and extended. Herein, we review the evolution of the AC concept, emphasizing new analysis schemes and findings that help to better understand ACs and extract SAR knowledge from them.

1. INTRODUCTION

Activity cliffs (ACs) have been discussed in computational and medicinal chemistry for nearly three decades.1–4 The first mention of the term probably dates back to a book chapter published by Michael Lajiness in 1991.1 ACs were generally defined as pairs of structurally similar active compounds with a large difference in potency.1,2 As such, ACs received strong attention in computational chemistry and drug design because they represented instances of structure–activity relationship (SAR) discontinuity that were detrimental for quantitative SAR (QSAR) modeling.2 In QSAR, AC compounds were often falsely considered outliers, as pointed out in a milestone reference by Gerry Maggiora that raised awareness of ACs in the field.3 However, SAR discontinuity also translates into high SAR information content and ACs identify small chemical modifications that determine the potency of active compounds.4 This explains the strong interest in ACs in medicinal chemistry to aid in compound optimization.3,4 ACs and the SAR discontinuity they capture are usually focal points during early stages of compound optimization efforts when potency must be improved. By contrast, encountering ACs is less desirable during late stages when multiple compound properties must be balanced.5 This is the case because ACs and the underlying “steep” SARs complicate multiple property optimization while attempting to retain sufficiently high compound potency.5 In medicinal chemistry and structure-based compound design, individual ACs are frequently found,6,7 including instances revealing dramatic effects of minute chemical modifications such as “magic methyl” effects.8 Moreover, large volumes of SAR information become available when ACs are systematically identified across bioactive compounds.8 This is an area where computational and medicinal chemistry meet, providing a large knowledge base for compound optimization.5,9 While ACs might be intuitively appreciated from a chemical viewpoint, by subjectively considering compound series and judging SARs, the systematic study of ACs requires careful consideration of two criteria that are essential for defining ACs, the similarity criterion and potency difference criterion.1,2,3 Investigating these criteria and their interplay has largely determined the way in which the AC concept has evolved over the years and continues to evolve,5 as discussed in the following.

2. SIMILARITY

2.1. Molecular Graph-Based Similarity. Compound similarity can be assessed in a variety of ways. For computational representation of ACs, molecular fingerprints (bit string representations of chemical structure and/or properties) have originally been used as descriptors to calculate Tanimoto similarity,5 a standard procedure in chemoinformatics. Tanimoto coefficient (Tc)-based ACs are shown in Figure 1. While convenient computationally, the use of the Tc, which ranges from 0 (no fingerprint overlap) to 1 (identical fingerprints), requires the choice of a threshold value for classifying compounds as similar, which is largely subjective.3 Furthermore, different fingerprints produce different Tc values, which changes similarity criteria. Accordingly, attempts have been made to identify ACs that would be formed...
independently of different molecular representations and similarity calculations, leading to the search for “consensus ACs.” Moreover, although straightforward to calculate for computational screening of compounds, Tc values are sometimes difficult to interpret from a chemical perspective because they are calculated as a whole-molecule similarity measure and do not take any substituent rules or reaction information into account. Therefore, as an alternative to calculated similarity values, substructure-based similarity criteria can be applied, which lead to a binary assessment of similarity, i.e., either two compounds share a predefined substructure and are classified as similar or not. Substructure-based similarity can also be assessed in different ways, for example, by using formally defined molecular scaffolds (core structures) and distinguishing different scaffold/substituent (R-group) relationships, as illustrated in Figure 1.

The substructure-based classification scheme depicted in Figure 1 establishes different categories of ACs that are chemically intuitive including “chirality clffs” where compounds differ only by the configuration at a single stereo center but display large potency differences. The study of such chirality clffs (or chiral clffs) capturing subtle structural differences was recently further extended by systematically evaluating if pairs of enantiomers tested in the same assay formed an AC or not. In this study, chiral clffs were represented by a variety of chirality-sensitive descriptors and subjected to machine learning to predict chiral clffs and distinguish them from pairs of enantiomers that did not form ACs.

Another substructure-based similarity criterion for ACs is provided by the formation of matched molecular pairs (MMPs), i.e., pairs of compounds that are only distinguished by a chemical modification at a single site. MMPs can be efficiently identified algorithmically in large compound databases. The size of chemical modifications can be restricted to mostly limit compounds forming MMPs to structural analogs typically generated in medicinal chemistry. Application of this similarity criterion led to the introduction of MMP-based ACs that were termed “MMP-clffs.” For MMP generation, algorithmic fragmentation of exocyclic single bonds can also be replaced with bond fragmentation according to retro-synthetic rules, giving rise to the formation of retrosynthetic MMPs (RMMPs) and “RMMP-clffs” (in analogy to MMP-clffs), as further discussed below. By definition, MMPs and RMMPs are characterized by substitutions at a single site. Hence, MMP- and RMMP-clffs cannot contain different R-groups at multiple sites, which is often the case for lead optimization series. Recently, a computational method has been developed for systematically identifying analog series with single or multiple substitution sites in compound data sets. This approach, termed the compound-core relationship (CCR) method, is applicable to enumerate analog pairs (APs) from a given series with single or multiple substitution sites, hence providing an extension of the (R)MMP formalism to multiple
points of chemical modification (R-group replacements)\textsuperscript{15} and another similarity criterion for AC formation.

\textbf{2.2. From Two to Three Dimensions.} ACs can not only be defined using molecular graph-based (2D) similarity but also on the basis of three-dimensional (3D) structures. To these ends, bound ligands in complex structures with a given target protein must be spatially aligned based upon protein superposition and 3D similarity is calculated taking conformational and positional differences between ligands into account.\textsuperscript{16} For this purpose, numerical 3D similarity functions are available.\textsuperscript{4} Then, if ligands with similar binding modes exceed a predefined potency difference threshold, “3D-cliffs” are obtained. For this purpose, numerical 3D similarity functions are available.\textsuperscript{4} Then, if ligands with similar binding modes exceed a predefined potency difference threshold, “3D-cliffs” are obtained. As illustrated in Figure 2, compounds forming 3D-cliffs often (but not always) display characteristic interaction differences in experimental structures that are likely to explain differences in their potency. As such, 3D-cliffs provide informative examples for structure-based compound design and excellent test cases for the calibration of computational methods to calculate binding energies.

In 2015, a systematic analysis of publicly available X-ray structures of ligands in complex with human targets identified 630 3D-cliffs for which high confidence activity data were available in the medicinal chemistry literature.\textsuperscript{16} These ACs involved 61 human targets. MMP search in ChEMBL,\textsuperscript{17} the major public repository of compounds and activity data for medicinal chemistry, identified a total of 1980 structural analogs of 268 3D-cliff compounds with activity against 50 human targets.\textsuperscript{16} These analogs complemented 414 3D-cliffs and hence bridged between 3D- and 2D-ACs, providing ample opportunities for advanced SAR analysis.

3D-cliffs were also used to investigate the ability of different docking and scoring approaches to predict large potency differences between similar compounds.\textsuperscript{18} From 158 complex X-ray structures of nine target proteins, 146 3D-cliffs were isolated and subjected to docking calculations. A target-based scoring scheme was established to evaluate 3D-cliffs and predict cliff partners for compounds with known or modeled binding modes.\textsuperscript{18} Another study combined the analysis of corresponding 2D- and 3D-ACs using a database of more than 1700 X-ray structures of inhibitor complexes of 190 kinases.\textsuperscript{19} Compound similarity was assessed on the basis of 2D fingerprints and similarity of crystallographic inhibitor-kinase interactions using molecular interaction fingerprints (IFPs), hence providing a complementary assessment of molecular and interaction similarity, leading to the introduction of “interaction cliffs”.\textsuperscript{19} Only ~25% of 2D-ACs identified in this study also qualified as interaction cliffs, due to overall low interaction similarity detected with IFPs. However, interaction cliffs revealed ligand–target interaction “hot spots” and were proposed to aid in the design of new inhibitors,\textsuperscript{19} thus widening the spectrum of 3D-cliffs.

\textbf{3. CONCERTED FORMATION OF ACTIVITY CLIFFS}

Originally, ACs were investigated on the basis of compound pairs,\textsuperscript{2,3} consistent with how ACs were typically encountered during sequential compound optimization efforts.\textsuperscript{7} However, in...
compound data sets, ACs are rarely formed by “isolated” pairs of compounds (i.e., pairs without structural neighbors forming ACs). Rather, most ACs (>90%) are formed by groups of structural analogs with varying potency, involving compounds in multiple ACs. The formation of these “coordinated” ACs can be rationalized using AC network representations in which nodes represent compounds and edges pairwise AC relationships. Figure 3 shows a representative example. In networks, coordinated ACs give rise to clusters of varying size and topology. These clusters reveal more SAR information than ACs considered as isolated pairs.

AC clusters are frequently formed by a highly potent compound and multiple weakly potent analogs or vice versa, resulting in densely connected central nodes or “hubs” in AC networks. ACs can also be visualized in “activity landscapes”, which are generally defined as graphical representations that...
integrate compound similarity and potency relationships. Hubs in AC networks correspond to “AC generators” in activity landscape models, which were defined as compounds forming ACs with high frequency. AC generators identified using fingerprint Tanimoto similarity can be organized on the basis of scaffolds and enrichment factors can be calculated that account for the proportion of ACs containing a particular scaffold, thereby combining numerical and substructure-based similarity assessment.

4. RELEVANT POTENCY DIFFERENCES

While similarity criteria for AC formation have been intensely investigated, as discussed above, comparatively little attention has been paid to determining most relevant potency differences. In general, as potency measurements for AC assessment, assay-independent equilibrium constants ($K_i$ values) are strongly preferred over other assay-dependent measurements (such as $IC_{50}$ values). Although it is possible to represent ACs as a continuum of compound pairs with increasing potency differences, most studies have used a predefined and constant potency threshold across different compound activity classes (also termed target sets). A potency difference threshold of 2 orders of magnitude (100-fold) has often been applied, but also a threshold of only 1 order of magnitude. A constantly applied potency difference threshold is convenient for computational identification and analysis of ACs across different target sets but does not take target set-dependent differences in potency value distributions into account. However, these distributions significantly vary in target sets where AC formation is strongly influenced by the interplay between potency value distributions and structural relationships. Therefore, ACs have also been defined on the basis of variable target set-dependent potency difference thresholds. For a given target set, it is first investigated if compound potency variations are large enough to yield ACs, then all qualifying RMMPs or APs are identified and their potency differences determined. Finally, the set-dependent potency difference threshold for AC formation is calculated as the mean of the compound pair-based potency difference distribution plus two standard deviations. Figure 4 provides a representative example.

In a systematic analysis of ChEMBL (release 23), 212 target sets with available $K_i$ measurements and qualifying potency value distributions yielded a total of 16,096 set-dependent RMMP-cliffs. Most target set-dependent potency difference thresholds fell into the range $1 \leq \Delta pK_i \leq 2.5$. For comparison, when a general potency difference criterion of $\Delta pK_i \geq 2$ was applied, 11,773 RMMP-cliffs originating from 195 target sets were obtained. Hence, the target set-dependent potency difference criterion yielded more ACs in more target sets than a generally applied potency difference threshold of comparable magnitude and a more balanced distribution of ACs across target sets. The increase in the number of set-dependent ACs was often due to AC formation in different compound subsets,
which also further increased the amount of AC-associated SAR information.

### 5. DIFFERENT GENERATIONS OF ACTIVITY CLIFFS

On the basis of alternative similarity and potency difference criteria, as discussed above, we distinguish between different generations of ACs, as illustrated in Figure 5, which reflects the evolution of the AC concept in our work. The first generation of ACs was based upon fingerprint descriptors, numerical Tanimoto similarity, and predefined potency difference thresholds. The transition to second generation ACs was marked by the application of substructure-based similarity criteria such as the formation of MMPs or RMMPs and the use of general or target set-dependent potency difference criteria, leading to the introduction of MMP-cliffs and RMMP-cliffs respectively. These types of ACs contained a single substitution site, consistent with the underlying MMP formalism. Compared to numerical similarity measures, the use of MMPs/RMMPs often increased the interpretability and medicinal chemistry relevance of ACs and set-dependent potency difference criteria further increased their SAR information content.

Third generation ACs, as introduced recently, are exclusively based on variable target set-dependent potency difference thresholds and depend on APs enumerated from computationally identified or already available analog series such that single and multiple substitution sites are taken into account, as illustrated in Figure 6. Thus, application of this similarity criterion has led to the introduction of single-site and multisite ACs. A systematic search in ChEMBL (release 24.1) revealed that only 25.6% of currently available third generation ACs were multisite. For 297 dual-site ACs, pairs of single-site analogs were identified that contained the individual substitutions of the dual-site ACs. Comparison of the potency of dual-site AC partners with single-site analogs identified redundant dual-site ACs, which were represented by single-site cliffs, as well as the presence of additive, synergistic, and compensatory potency effects in confirmed dual-site ACs. Figure 6 illustrate a synergistic effect of individual substitutions.

### 6. MONITORING ACTIVITY CLIFF POPULATIONS ON A TIME SCALE

Some new findings are reported to illustrate how AC populations evolve over time. In ChEMBL (release 25), 96 target sets were identified that grew over a period of 10 years (2009–2018) and contained at least 100 APs from different analog series in 2009. For each year, the number of third generation ACs per set was determined. Figure 7 shows the cumulative growth in the number of compounds, APs, and ACs. While the total number of third generation ACs available in all sets increased from 5563 to 17 539 over 10 years, the proportion of APs forming target set-dependent ACs remained essentially constant at 4.9%. Over time, there only was a slight increase in the proportion of dual-site and multisite ACs with three or more substitution sites relative to single-site ACs. Hence, ACs were formed at a constant rate over time in different targets sets in the presence of stable compound structure-potency relationships and single-site ACs dominated AC populations.

### 7. CONCLUSIONS AND OUTLOOK

Herein, foundations of the AC concept and its evolution have been reviewed. ACs are of interest in both computational and
medicinal chemistry. They are explored on a case-by-case basis, especially during early stages of compound optimization, as well as on a large scale through systematic computational analysis of compound collections and activity data. In computational chemistry, there is also interest in predicting ACs using machine learning, and a few models have been reported so far, yielding promising prediction accuracy.¹¹,²⁶,²⁷

Most of our current knowledge about ACs has originated from compound data analysis. In growing compound data sets, ACs are formed at an essentially constant rate. Large numbers of ACs have been identified across current target sets, providing a wealth of SAR information for medicinal chemistry. In addition to applying molecular graph-based similarity measures, ACs can also be defined on the basis of compound binding modes and 3D similarity, providing interesting examples and test cases for drug design.

Several factors influence AC analysis. First and foremost, compound similarity and potency difference criteria are of critical importance for the definition, perception, and utilization of ACs.

Clearly, alternative AC definitions change AC populations and SAR information associated with them. Another critical aspect of AC analysis is that experimental accuracy and measurement characteristics must be carefully considered. For example, AC analysis is not reliable if incompatible measurement types such as $K_i$ and IC$_{50}$ values are used in combination. In general, high-confidence activity data are strongly preferred.

A recent conceptual advance has been the introduction of target set-dependent ACs taking differences in potency value distributions into account and arriving at set-dependent potency difference thresholds. These ACs best reflect SAR characteristics of different compound classes and data sets. The evolution of the AC concept is well illustrated by distinguishing different generation of ACs and their characteristics, as discussed herein.

The AC concept is subject to further extensions. For example, an interesting task will be the evaluation of ACs during sequential optimization efforts, considering alternative compound series in parallel. Since most large-scale analyses of ACs have focused on data sets from heterogeneous sources, the frequency with which ACs occur during series-based compound optimization is currently unknown. A thorough analysis of this topic has thus far been prohibited by limited public availability of “real life” lead optimization series from...
drug discovery. Another interesting area for future research will be the use of the AC knowledge base to better understand structural features that determine compound potency across different target sets. This also provides further opportunities for machine learning and significance analysis of molecular features that determine successful predictions. Clearly, despite progress made in rationalizing ACs over the years, we are just beginning to explore their potential for practical applications. There is more to come.

■ AUTHOR INFORMATION

Corresponding Author
*Phone: 49-228-7369-100. E-mail: bajorath@bit.uni-bonn.de.

ORCID

Jürgen Bajorath: 0000-0002-0557-5714

Notes

The authors declare no competing financial interest.

Biographies

Dagmar Stumpfe studied Biology at the University of Bonn, Germany. In 2006, she joined the Department of Life Science Informatics headed by Prof. Jürgen Bajorath for her Ph.D. thesis, investigating methods for computer-aided chemical biology with a special focus on the exploration of compound selectivity. Since 2009, Dagmar has been a postdoctoral fellow and researcher in the Department. Her current research interests include computational chemical biology and large-scale structure–activity relationship analysis.

Huabin Hu obtained his M.D. degree in medicinal chemistry from Peking Union Medical College in 2017. He then joined the Department of Life Science Informatics at the University of Bonn as a Ph.D. student under the supervision of Prof. Jürgen Bajorath. Currently, his Ph.D. work mainly focuses on large-scale investigation of activity cliffs and structure–activity relationships.

Jürgen Bajorath is Professor and Chair of Life Science Informatics at the University of Bonn. He is also an Affiliate Professor in the Department of Biological Structure at the University of Washington, Seattle. His research interests include drug discovery, computer-aided medicinal chemistry, chemical biology, and chemoinformatics. For further information, please see http://bajorath.bit.uni-bonn.de.

■ ACKNOWLEDGMENTS

We thank OpenEye and Chemical Computing Group for academic software licenses. H.H. is supported by the China Scholarship Council.

■ REFERENCES

(1) Lajiness, M. Evaluation of the Performance of Dissimilarity Selection Methodology. In QSAR: Rational Approaches to the Design of Bioactive Compounds; Slipo, C., Vittorià, A., Eds.; Elsevier: Amsterdam, Netherlands, 1991; pp 201–204.
(2) Maggiora, G. M. On Outliers and Activity Cliffs – Why QSAR often Disappoints. J. Chem. Inf. Model. 2006, 46, 1535–1535.
(3) Stumpfe, D.; Bajorath, J. Exploring Activity Cliffs in Medicinal Chemistry. J. Med. Chem. 2012, 55, 2932–2942.
(4) Stumpfe, D.; Hu, Y.; Dimova, D.; Bajorath, J. Recent Progress in Understanding Activity Cliffs and Their Utility in Medicinal Chemistry. J. Med. Chem. 2014, 57, 18–28.
(5) Bajorath, J. Duality of Activity Cliffs in Drug Discovery. Expert Opin. Drug Discovery 2019, 14, S17–S20.
(6) Pennington, L. D.; Moustakas, D. T. The Necessary Nitrogen Atom: A Versatile High-Impact Design Element for Multiparameter Optimization. J. Med. Chem. 2017, 60, 3552–3579.
(7) Leung, C. S.; Leung, S. S.; Tirado-Rives, J.; Jorgensen, W. L. Methyl Effects on Protein–Ligand Binding. J. Med. Chem. 2012, 55, 4489–4500.
(8) Stumpfe, D.; Bajorath, J. Monitoring Global Growth of Activity Cliff Information over Time and Assessing Activity Cliff Frequencies and Distributions. Future Med. Chem. 2015, 7, 1565–1579.
(9) Medina-Franco, J. L.; Martínez-Mayorga, K.; Bender, A.; Marin, R. M.; Giulanotti, M. A.; Pinilla, C.; Houghten, R. A. Characterization of Activity Landscapes using 2D and 3D Similarity Methods: Consensus Activity Cliffs. J. Chem. Inf. Model. 2009, 49, 477–491.
(10) Hu, Y.; Bajorath, J. Extending the Activity Cliff Concept: Structural Categorization of Activity Cliffs and Systematic Identification of Different Types of Cliffs in the ChEMBL Database. J. Chem. Inf. Model. 2012, 52, 1806–1811.
(11) Schneider, N.; Lewis, R. A.; Fechner, N.; Ertl, P. Chiral Cliffs: Investigating the Influence of Chirality on Binding Affinity. ChemMedChem 2018, 13, 1315–1324.
(12) Hu, X.; Hu, Y.; Vogt, M.; Stumpfe, D.; Bajorath, J. MMP-Cliffs: Systematic Identification of Activity Cliffs on the Basis of Matched Molecular Pairs. J. Chem. Inf. Model. 2012, 52, 1138–1145.
(13) Hussain, J.; Rea, C. Computationally Efficient Algorithm to Identify Matched Molecular Pairs (MMPs) in Large Data Sets. J. Chem. Inf. Model. 2010, 50, 339–348.

(14) de la Vega de León, A.; Bajorath, J. Matched Molecular Pairs Derived by Retrosynthetic Fragmentation. MedChemComm 2014, 5, 64–67.

(15) Naveja, J. J.; Vogt, M.; Stumpfe, D.; Medina-Franco, J. L.; Bajorath, J. Systematic Extraction of Analogue Series from Large Compound Collections Using a New Computational Compound–Core Relationship Method. ACS Omega 2019, 4, 1027–1032.

(16) Hu, Y.; Furtmann, N.; Bajorath, J. Extension of Three-Dimensional Activity Cliff Information through Systematic Mapping of Active Analogs. RSC Adv. 2015, 5, 43006–43015.

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

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

(19) Méndez-Lucio, O.; Kooistra, A. J.; de Graaf, C.; Bender, A.; Medina-Franco, J. L. Analyzing Multitarget Activity Landscapes Using Protein–Ligand Interaction Fingerprints: Interaction Cliffs. J. Chem. Inf. Model. 2015, 55, 251–262.

(20) Stumpfe, D.; Dimova, D.; Bajorath, J. Composition and Topology of Activity Cliff Clusters Formed by Bioactive Compounds. J. Chem. Inf. Model. 2014, 54, 451–461.

(21) Pérez-Villanueva, J.; Méndez-Lucio, O.; Soria-Arteche, O.; Medina-Franco, J. L. Activity Cliffs and Activity Cliff Generators Based on Chemotype-Related Activity Landscapes. Mol. Diversity 2015, 19, 1021–1035.

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

(23) Hu, H.; Stumpfe, D.; Bajorath, J. Rationalizing the Formation of Activity Cliffs in Different Compound Data Sets. ACS Omega 2018, 3, 7736–7744.

(24) Hu, H.; Stumpfe, D.; Bajorath, J. Second-Generation Activity Cliffs Identified on the Basis of Target Set-Dependent Potency Difference Criteria. Future Med. Chem. 2019, 11, 379–394.

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

(26) Heikamp, K.; Hu, X.; Yan, A.; Bajorath, J. Prediction of Activity Cliffs Using Support Vector Machines. J. Chem. Inf. Model. 2012, 52, 2354–2365.

(27) Horvath, D.; Marcou, G.; Varnek, A.; Kayastha, S.; de la Vega de León, A.; Bajorath, J. Prediction of Activity Cliffs Using Condensed Graphs of Reaction Representations, Descriptor Recombination, Support Vector Machine Classification, and Support Vector Regression. J. Chem. Inf. Model. 2016, 56, 1631–1640.