molecular informatics

DOI: 10.1002/minf.200((full DOI will be filled in by the editorial staff))

Ligand-based Activity Cliff Prediction Models with Applicability Domain

Shunsuke Tamura,[a] Tomoyuki Miyao[a,b] and Kimito Funatsu*[a,b,c]

Abstract: Activity cliffs (ACs) are formed by pairs of structurally similar compounds with large differences in potency. Predicting ACs is of high interest in lead optimization for drug discovery. Previous AC prediction models that focused on matched molecular pair (MMP) cliffs produced adequate performances. However, the extrapolation ability of these models is unclear because the main scaffold for MMPs, the core structure, could exist in both training and test data sets. Also, representation of MMPs did not consider the attachment points where the core and R-group substituents are connected.

In this study, we aimed to improve a ligand-based AC prediction method using molecular fingerprints. We incorporated applicability domain, which was defined using R-path fingerprints to consider the local environment around an attachment point.

Rigorous evaluation of the extrapolation ability of AC prediction models showed that MMP-cliffs were accurately predicted for nine biological targets. Furthermore, incorporation of training MMPs with cores distinct from those of test MMPs improved the predictability compared with using training MMPs with only similar cores.

Keywords: activity cliff prediction, applicability domain, drug design, ligand-based approach, structure-activity relationships

1 Introduction

Activity cliffs (ACs) are formed by pairs of structurally similar compounds with large differences in potency.[1,2] ACs can be found during lead optimization in drug discovery because a series of analogous compounds are synthesized. Modifying R-group substituents to find compounds with better characteristics, including absorption, distribution, metabolism, and toxicity,[3,4] is a common optimization approach. A useful representation of structural similarity among such analogous compounds is the matched molecular pair (MMP) formalism. An MMP is a pair of compounds that share a core substructure and differ at a single site.[5] ACs based on MMPs, termed MMP-cliffs, are used to associate potency change with the corresponding substructure change.[6]

The existence of MMP-cliffs in a series of analogous compounds indicates discontinuity of structure–activity relationships (SARs). Although this might impair rational compound development, MMP-cliffs help to understand molecular interactions between a compound and a protein target by focusing on the key interactions suggested by structural differences among MMP-cliffs.[7] Hence, prior to synthesis of compounds, prediction of individual compounds forming an MMP-cliff is of great importance.

In ligand-based approaches, predicting ACs is fundamentally inapplicable in quantitative structure–activity relationship (QSAR) models.[8] The similarity principle in QSAR, which states that similar compounds have similar properties,[9] is not valid in AC prediction. Therefore, ordinal descriptors, such as molecular weight, number of hydrogen bond donor/acceptor atoms, and lipophilicity (logP), cannot be used for this purpose.

Previous approaches to the prediction of MMP-cliffs have taken both core and substituents information into account by introducing a kernel function or a set of engineered descriptors.[10,11] Combining support vector machine (SVM) models[12] with a well-designed kernel function (MMP kernel) specified for AC prediction gave sufficient overall predictivity for test MMP-cliffs. However, because of the random split of training and test data, the extrapolation ability of these models, *i.e.*, their predictive performance when applied to an MMP with a core structure that is not in the training data set, has not been investigated. Furthermore, because the MMP kernel was the multiplication of individual kernels for a core and substituents, attachment points, which are the connection point between a core and a substructure inside the compound structure, were not considered at all. Another approach to MMP-cliff prediction used the reaction-based representation for MMPs.[11] The prediction performance of this approach was similar to those of the SVM models that used an MMP kernel. Although a validation scheme for the AC prediction models was carefully set up, the same core structures between training and test compounds might exist, leading to a limited estimation of the extrapolation ability of the models.

[a] *Graduate School of Science and Technology, Nara Institute of Science and Technology*
8916-5 Takayama-cho, Ikoma, Nara 630-0192, Japan

[b] *Data Science Center, Nara Institute of Science and Technology*
8916-5 Takayama-cho, Ikoma, Nara 630-0192, Japan

[c] *Department of Chemical System Engineering, School of Engineering, The University of Tokyo*
7-3-1 Hongo. Bunkyo-ku, Tokyo 113-8656, Japan
**e-mail: funatsu@chemsys.t.u-tokyo.ac.jp,*
phone/fax: +81-3-5841-7751

Supporting Information for this article is available on the WWW under www.molinf.com

In this study, we aimed to improve the AC prediction method previously proposed by Heikamp et al.[10] and investigated the extrapolation ability of AC prediction schemes. The proposed AC prediction method that uses R-path fingerprints[13] to consider the local environment around attachment points performed better than a conventional one when applicability domain of the prediction models was introduced. Because R-path fingerprints can quantify similarity focusing on an attachment point, MMP cores with distinct attachment points can be correctly quantified (*vide infra*). For rigorous validation of the extrapolation ability of AC prediction models, MMPs were split into training and test data sets so that no MMP cores and compounds overlapped between the two data sets.

We found that for the nine selected biological targets, MMP-cliffs were accurately predicted after taking applicability domain into account. In addition, instead of using only MMPs with similar cores as the training set, incorporating MMPs with distinct cores from the test MMPs into the training data set improved the prediction performance.

2 Computational Methods

2.1 Data Sets

Nine protein targets were selected based on the number of active compounds. Compounds active against the protein targets were extracted from the ChEMBL database (version 24)[14] using the following criteria: maximal assay confidence score of 9, interaction relationship type 'D', and equilibrium constant (K_i) as a potency measurement. When multiple K_i values were available for a compound, their geometric mean (the mean of pK_i) was calculated to obtain its potency value, otherwise the compound was discarded. In the present study, MMP-cliffs were employed as a definition of ACs.[6] MMPs with differences in potency >2 were defined as ACs and <1 as non-ACs. The other MMPs (differences in potency 1–2) were discarded. Target-wise MMPs were generated systematically as described previously.[10] MMPs were generated computationally using the efficient algorithms[15] implemented by Wawer and Bajorath.[16] For MMP generation, the following size restriction criteria were applied: substructure exchange no more than 13 heavy atoms, and maximum difference between substructures of a compound no more than eight heavy atoms. MMPs that shared the same core were collected as a matching molecular series (MMS).[17] In our analysis, MMSs containing at least one AC MMP and one non-AC MMP were used. Data set profiles for the nine targets are provided in **Table 1**.

2.2 MMP Fingerprints

MMP fingerprints were assembled as described previously[10] as shown in **Figure 1**. From each MMP, fingerprints were calculated individually for one core and a pair of substructures. Extended connectivity fingerprints of bond diameter 4 (ECFP4)[18] were used as the fingerprint algorithm. The calculated identifiers of ECFP4 were folded into a 4096-bit vector using the modulo operation. The individual vectors were concatenated after changing bits with one to zero when the bits were both on for the substructures, to focus on the transformation.

Table 1. Profiles of the Data Sets Used in This Study.

ChEMBL ID	Target	Abbreviation	#CPDs	#MMPs	#MMSs	Potency (pK_i)		MW		#HA	
						Max	Min	Max	Min	Max	Min
244	Factor Xa	fxa	632	4728	62	11.40	3.90	711.91	361.46	50	25
259	Melanocortin receptor 4	mc4	444	2877	39	9.30	4.07	672.57	447.06	46	28
237	Kappa opioid receptor	kor	555	6590	50	10.82	4.09	605.64	271.4	44	20
204	Thrombin	thr	221	840	28	10.30	2.81	693.86	280.75	50	19
217	Dopamine D2 receptor	dd2	612	6574	56	10.24	2.85	588.33	205.3	40	15
226	Adenosine A1 receptor	aa1	621	2624	75	12.23	4.60	616.47	215.21	40	12
251	Adenosine A2a receptor	aa2	989	5192	105	11.38	4.00	713.87	230.23	53	13
261	Carbonic anhydrase I	ca1	293	818	51	11.00	-4.18	702.66	154.12	50	11
205	Carbonic anhydrase II	ca2	362	994	69	11.00	0.70	678.36	153.14	42	11

For each data set, the CHEMBL ID, number of compounds (CPDs), number of matched molecular pairs (#MMPs), number of matching molecular series (#MMSs), maximum and minimum potencies (pK_i values), maximum and minimum molecular weight (MW), and heavy atom counts (#HA) are listed.

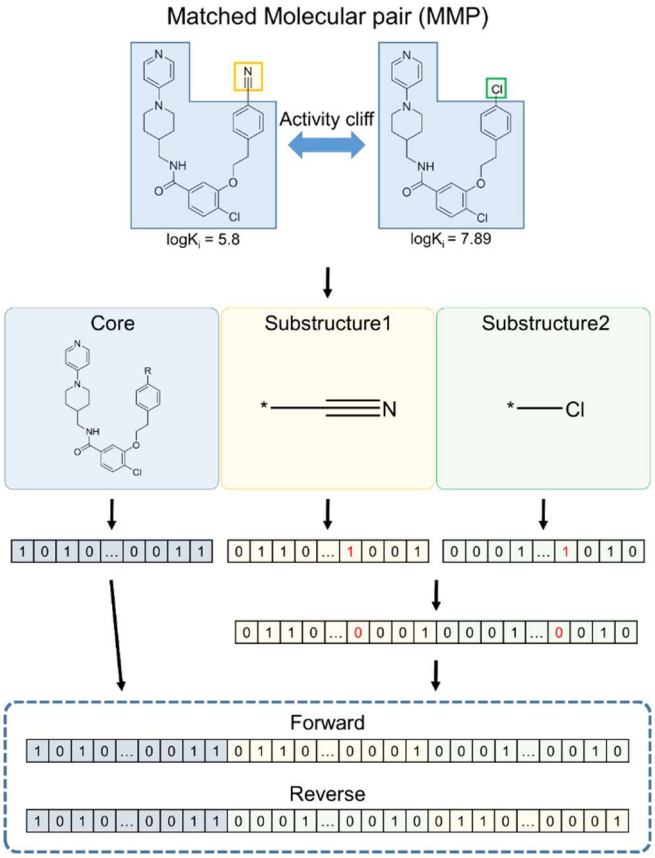
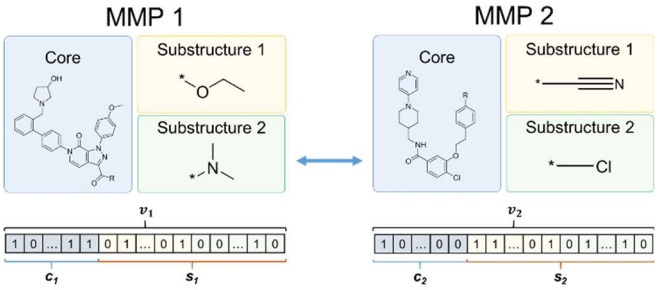


Figure 1. Overview of Fingerprint Generation for MMP based Activity Cliff Pairs. A matched molecular pair (MMP) is formed by two compounds with an activity cliff relationship. The two compounds share the main scaffold (core) and differ only in their substructures. Generally, one core and two substructures are present in MMPs. MMP fingerprint generation starts from individual fingerprints for these three parts. The fingerprints for the two substructures were concatenated after changing bits with one to zero when the bits were both on, to focus on the molecular transformation. Then the core fingerprint was appended to the concatenated vector. Two types of MMP fingerprints (*Forward* and *Reverse*) were generated in opposite concatenation order of substructure fingerprints.

Two MMP fingerprints were generated in the opposite concatenation order of substructure fingerprints, termed *Forward* and *Reverse*. In the model construction phase, a substructure for the lower potent compound was assigned to be the first and for the higher potent compound it was assigned to be the last (*Forward* in **Figure 1**).

2.3 AC Prediction Models

AC prediction models have been constructed using support vector machine (SVM), which is a supervised learning algorithm that aims to identify a hyperplane for separating two classes while maximizing the margin from the hyperplane.[12] SVM was proposed originally as a linear classification method but it can be extended to a nonlinear classification method using a kernel function. In this study, we used an MMP kernel based on Tanimoto similarity[19] of MMP fingerprints for nonlinear projection.[10] A schematic representation of an MMP kernel calculation is shown in **Figure 2**. Two individual Tanimoto kernel functions for the core and substructures were



$$K(\boldsymbol{v_1}, \boldsymbol{v_2}) = K(\boldsymbol{c_1}, \boldsymbol{c_2}) \times K(\boldsymbol{s_1}, \boldsymbol{s_2})$$

Figure 2. Schematic Representation of the Kernel Function for Matched Molecular Pairs (MMPs) for Activity Cliff Prediction. Following the procedure shown in **Figure 1**, MMP fingerprints were generated for the MMPs. The kernel function is the multiplication of two individual kernels for cores (represented as *c*) and substructures (represented as *s*).

calculated and combined by multiplication. Hyperparameter *C* in SVM was set to 1.0, which is the default setting in scikit-learn, a machine learning library,[20] because of the limited number of MMPs in the training data sets.

2.4 Evaluation of AC Prediction Models

The proposed AC prediction scheme illustrated in **Figure 3**, takes applicability domain of the SVM models into account. The applicability domain can be determined based on the distance to the training data set.[21] In our prediction scheme, the distance metric was Tanimoto similarity using R-path fingerprints to evaluate similarity by considering the attachment point. An MMS was iteratively selected as a test set and the other MMSs formed the training set to avoid using the same core in both data sets (leave one MMS out). All compounds in the test MMPs were different from those in the training data set simply by eliminating training MMPs consisting of duplicated compounds. For each test MMS, similarity values to the training MMS cores were calculated. When at least one training core with a similarity value over a set threshold was found, an SVM model was constructed and the AC was predicted accordingly. Otherwise, the test MMS was labelled as unpredictable and no prediction was made. All the MMPs, except those in the test MMS, were used in the training data set. MMPs in an MMS that had a core that was dissimilar to that of the test MMS were included in the training data set. The effect of incorporating MMPs distinct to test MMPs will be discussed in section 3.1. Each SVM model was trained using *Forward* MMP fingerprints and tested using both *Forward* and *Reverse* MMP fingerprints. Therefore, the AC prediction models learned whether an MMP was AC or non-AC for a chemical transformation from the first substructure to the second substructure (low to high). The interpretations of the model outputs are summarized in **Table 2**. In this study, true-positive (TP) was assigned only to the AC MMPs that were previously predicted as AC using *Forward* fingerprints and non-AC using *Reverse* fingerprints. True-negative (TN) was assigned to the non-AC MMPs, predicted as non-AC using both fingerprints. False-positive (FP) was assigned to the non-ACs initially predicted as AC using either *Forward* or *Reverse*

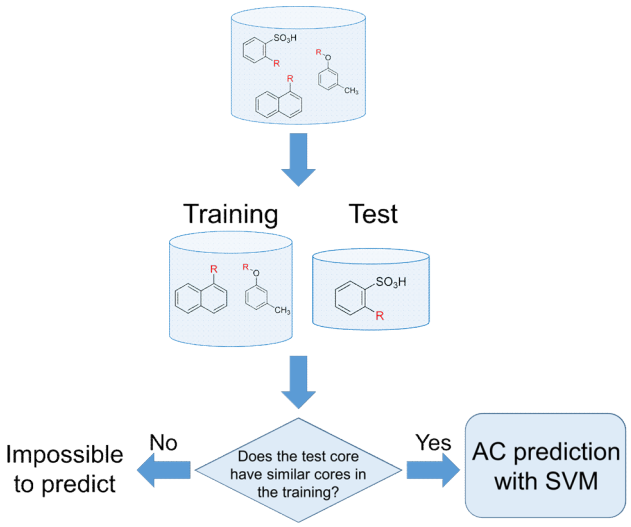


Figure 3. Model Construction and Prediction Scheme on the Basis of Core Structures. The activity cliff (AC) prediction scheme for testing the extrapolation ability of SVM models is shown. Matched molecular pairs (MMPs) with the same core were joined as a set of MMPs. Training compounds had cores different from that of the test set. Each MMP set (each core) was iteratively selected as a test set and the rest of the cores formed the training set. Training data selection was based on the similarity of cores to the test cores using R-path fingerprints. When at least one training core similar to the test core was found, an SVM model was constructed and the AC was predicted accordingly. Otherwise, the test data were labeled as unpredictable and the prediction was not performed.

Table 2. Evaluation of the Activity Cliff (AC) Prediction Results

Actual class	Predicted class		Interpretation
	Forward	*Reverse*	
AC	AC	AC	-
AC	AC	Non-AC	TP
AC	Non-AC	AC	FP
AC	Non-AC	Non-AC	FN
Non-AC	AC	AC	-
Non-AC	AC	Non-AC	FP
Non-AC	Non-AC	AC	FP
Non-AC	Non-AC	Non-AC	TN

Interpretation of the label assignment from the actual class and the predicted results is shown. TP, true-positive; FP, false-positive; TN, true-negative; FN, false-negative.

fingerprints and to the AC MMPs predicted as non-AC using *Forward* and AC using *Reverse* fingerprints. These MMPs were regarded as FP because the direction of potency was wrong. False-negative (FN) was assigned to the AC MMPs predicted as non-AC using both fingerprints. When predicted classes were AC using both types of fingerprints, the results were regarded as inconsistent and ignored. The AC prediction model performance was measured by Matthew's correlation coefficient (MCC) based on these labels.

2.5 Incorporation of MMS with Dissimilar Cores into Training Data Set

For accurate prediction of activity for a test compound, the similarity principle[9] requires that the compounds similar to the test compound are included in a training data set. We investigated whether using only the MMS with cores that were similar to those of the test MMS was sufficient or whether incorporating MMS with cores that were dissimilar to those of the test MMS might improve predictability of the models. For this investigation, two types of analyses were conducted.

First, the training MMS core was kept similarity to the test MMS core using R-path fingerprints. For each biological target, the thresholds were increased from 0.2 to 0.8 in steps of 0.1, and the performances of the AC prediction models were evaluated.

Second, two types of training data sets were prepared and AC predictions were conducted using the models built with these data sets. One data set contained only an MMS with cores similar to the test MMS core (*only similar* cores), and the other data set contained all the MMPs in the original training set (*filtered* cores). The performances of the models with these two types of training data sets were expected to reveal the effect of incorporating MMPs with distinct cores from the test set into the training data set. The threshold for defining similar core was set as 0.6 using R-path fingerprints.

3 Results and Discussion

3.1 Incorporation of MMS with Dissimilar Cores into Training Data Set

The leave one-MMS out MCC values for the nine biological targets are shown in **Figure 4**.

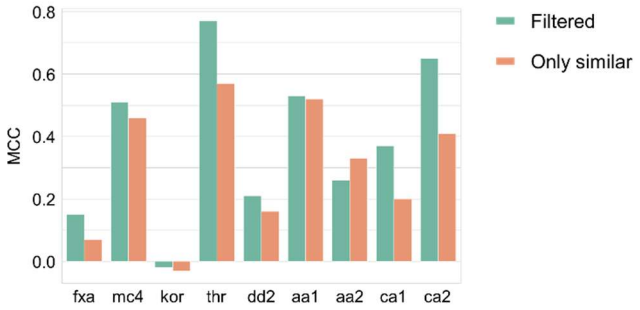


Figure 4. Effects of Incorporating Distinct Cores from a Test Data Set into the Training Data Set. Matthew's correlation coefficient (MCC) values indicate the model performance for each target. Only similar (orange), models built using matched molecular pairs (MMPs) with cores similar to those in the test data set; Filtered (green), models built using MMPs with both similar and dissimilar cores to those in the test data set. The data set abbreviations (fxa, mc4, kor, thr, dd2, aa1, aa2, ca1, ca2) are defined in **Table 1**.

For seven of the nine biological targets, the *filtered* model had higher MCC values than the *only similar* model, which indicated that to obtain AC prediction models with high predictability the training data set should contain MMPs with both similar and dissimilar cores. For the AC prediction models that were trained using solely MMPs with similar cores (*only similar*), the contribution of the substructure kernel became dominant in the MMP kernel function. Substructure fingerprints are usually very sparse because, by definition, substructures were smaller than

cores, which made extrapolation based on the substructure fingerprints difficult.

Conversely, the *filtered* method was able to build robust AC prediction models using information about both core and substructures. However, for the adenosine A2a receptor MMPs, the *only similar* method performed better than the *filtered* method; only 5 FP MMPs with the *only similar* method but 39 FP MMPs with the *filtered* method. We manually examined the FP MMPs in the test set and found that most of them shared substructure pairs with training MMPs. Two such typical MMPs are depicted in **Figure 5**. The similarity between these two MMPs in the MMP kernel was 0.49 (core part 0.49; substructure part 1.0), which was much higher than the average pairwise similarity of the training MMPs (0.0081). Furthermore, this was the maximum similarity value between the test MMP and training MMP data sets in terms of the kernel function. Because the training MMP was categorized as an AC, it was reasonable to speculate that the test MMP also was an AC. This situation might be avoided by introducing a stringent similarity threshold value for selecting the test MMS that can be predicted using AC prediction models in the proposed scheme (**Figure 3**).

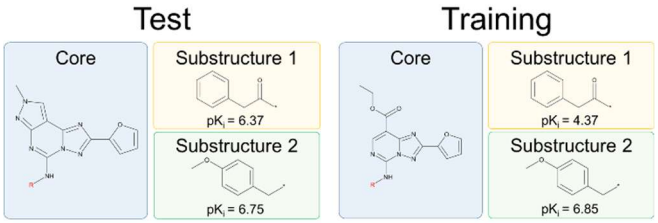


Figure 5. Typical MMPs for which the AC Prediction Model Failed Using the *filtered* Data Set. The activity cliff (AC) prediction failed for the matched molecular pair (MMP) in the test data set (left) using the *filtered* data set. The training MMP (right) was not included in the *only similar* data set but was included in the *filtered* data set.

The three targets: thrombin (thr), carbonic anhydrase I (ca1), and carbonic anhydrase II (ca2), for which the *filtered* method outperformed the *only similar* method by the greatest margins, were chosen to further investigate the rationale of introducing MMPs with distinct cores in training data sets. An additional method, termed *"only dissimilar"* was tested. For these three targets, AC prediction models were trained only on MMPs with cores dissimilar to the test. Similarity thresholds varied from 0.2 to 0.6 by 0.2 using R-path fingerprints.

When using a similarity threshold of 0.6, the MCC values were 0.46 and 0.60 for ca1 and ca2, respectively. For thr, models predicted non-cliff for all the test MMPs, failing to calculate the MCC values. By reducing the similarity threshold to 0.4 (0.2), the MCC values decreased to 0.33 (0.28) and 0.51 (0.39) for ca1 and ca2, respectively. Therefore, the *only dissimilar* method was inferior to the *filtered* method for the three targets.

During the validation process in this study, the same training compounds as in the test set were eliminated, as mentioned in 2.4 section. In this situation, diverse MMPs with both distinct and similar cores were found to be needed for more accurate AC prediction.

3.2 AC Prediction in MMS

All the studies so far focused on predicting ACs for MMPs with cores different from in training data sets. AC prediction within an MMS (termed shared core) was also conducted along with adding MMPs with dissimilar cores into training. Similarity threshold was set to 0.6 and adding MMPs with dissimilar cores to the training was controlled by adjusting the number of training MMPs in the MMS to be a certain ratio. Three ratios: 0.25, 0.5 and 0.75 were tested. Training MMPs with only dissimilar cores were also tested. In this case, the number of MMPs were set to the same as for the ratio of 0.25.

The MCC values were calculated by leave-one-out cross validation. For each test MMP, the same compounds in the training data set were eliminated before model construction. Each prediction was performed three times considering the effect of random sampling of training MMPs with dissimilar core except for a ratio of 1.0 (within MMS). The MCC values for the three targets are reported in **Figure 6**. AC prediction for the shared core showed the least MCC values for the three targets. Performances improved as the number of MMPs with dissimilar cores in training set increased. When there are no shared compounds between training and test, increasing the size of the training data set is important by adding MMPs with dissimilar cores. The procedure of eliminating overlapped compound from training set could explain this result.

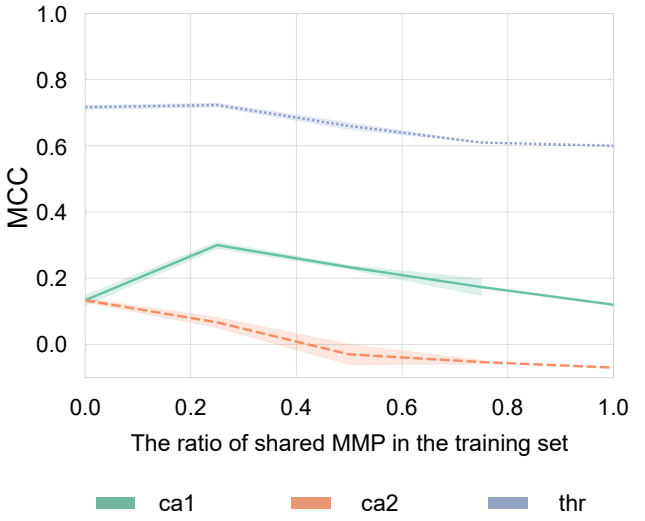


Figure 6. Performance of the AC Prediction Models and Ratios of Shared MMPs in the Training Set. The graphs show the Matthew's correlation coefficient (MCC) values for the test activity cliff (AC) pairs in the selected three biological targets. The ratios of training matched molecular pairs (MMPs) sharing the core with test MMPs are indicated on the x-axis. The ratio decreased when the training set included more MMPs with cores dissimilar to test cores. The data set abbreviations (thr, ca1, ca2) are defined in **Table 1**.

3.3 Effect of Controlling Applicability Domain on Model Performance

The most straightforward application of R-path fingerprints is to use them to represent MMP fingerprints. For the nine

targets, we compared the performances of the AC prediction models using R-path fingerprints to represent MMP fingerprints with their performances using ECFP4. We found that their performances using ECFP4 were slightly better than their performances using R-path fingerprints. The prediction results are shown in **Supporting information 1**. Therefore, in the following analysis, R-path fingerprints were used only for applicability domain definition and ECFP4 was used to generate MMP fingerprints.

The proposed AC prediction scheme considering applicability domain of models is shown in **Figure 3**. From the discussion in section 3.1, including MMPs with cores both similar and dissimilar to the test MMP was important to accurately predict the AC for the test MMP. Therefore, after deciding whether a test MMP was eligible or not, an AC prediction model was constructed using all the MMPs, except for those in the test MMS. The results are reported in **Figure 7.**

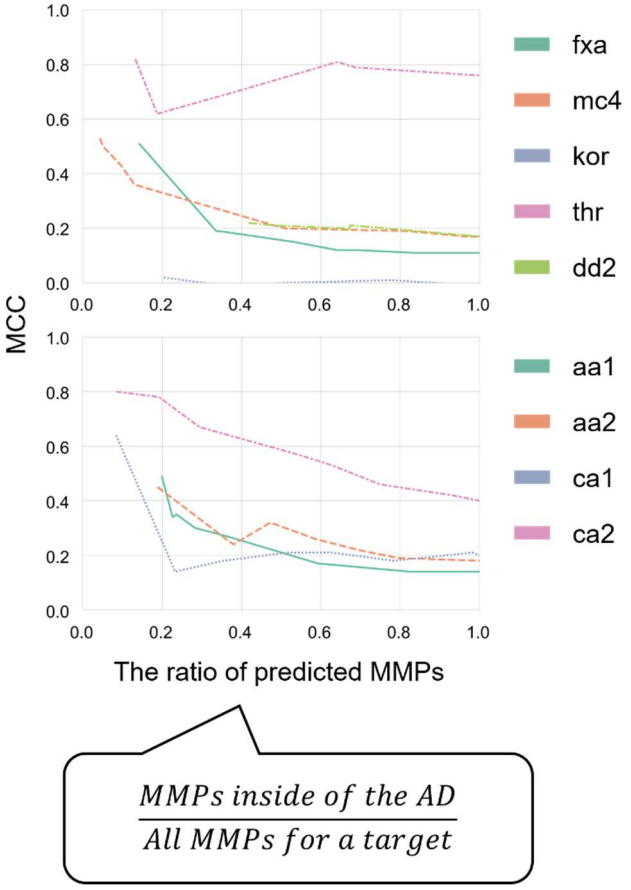


Figure 7. Performance of the AC Prediction Model and Number of Applicable Test Compounds. The graphs show the Matthew's correlation coefficient (MCC) values for the test activity cliff (AC) pairs in the nine biological targets. The ratios of predictable matched molecular pairs (MMPs) for all test MMPs are indicated on the x-axis. The ratio decreased when the thresholds of R-path similarity for selecting training cores became tighter. The data set abbreviations (fxa, mc4, kor, thr, dd2, aa1, aa2, ca1, ca2) are defined in **Table 1**.

Overall, by restricting eligibility for test MMS, the MCC values increased monotonically, except for thrombin. This shows that core distance using R-path fingerprints was a good indicator for understanding the applicability domain of the models. A previous study reported that limiting MMPs inside the applicability domain of an AC prediction model that was based on MMP-cliffs did not improve the performance of the model.[11] In that study, R-path-based distance to the training data set was used to define the applicability domain, whereas bounding box approaches were used to delineate the chemical space zone of the training MMPs.

For thrombin, restriction of test MMPs did not improve the model performance, and only using substructure information was sufficient to accurately predict ACs (MCC = 0.71). This indicated that the core information for thrombin was not important, and the applied restriction was not effective.

Our analysis indicated that to accurately predict the AC for a test MMP, incorporating MMPs (or MMS) with cores similar to that of the test MMP was necessary. Furthermore, the proposed prediction scheme worked better than the conventional AC prediction modelling methods by introducing training MMPs with cores dissimilar to those in the test MMPs.

4 Conclusions

Predicting whether an analogous compound forms an MMP-cliff is of importance in the lead optimization phase of drug discovery. Previous ligand-based AC prediction methods using MMP-cliffs were inefficient at evaluating the extrapolation ability of AC models with MMPs in training and test data sets that shared the same core.

In this study, we improved the AC prediction method by incorporating applicability domain using R-path fingerprints that correctly represented topological information around attachment points of MMP cores.

The proposed method achieved higher prediction performance for the MMPs inside the applicability domain of models by discarding MMPs outside of the applicability domain. Incorporating MMPs with cores that were distinct from the core of a test MMP into a training data set further improved the predictability of the models in terms of the MCC.

Acknowledgments

We thank Prof. Jürgen Bajorath for providing the codes for MMP fragmentation and OpenEye Scientific Software, Inc., for providing a free academic license for the OpenEye Toolkits. We thank Dr. Swarit Jasial from our group for carefully proofreading the manuscript. We thank Margaret Biswas, PhD, from Edanz Group (https://en-author-services.edanzgroup.com/ac) for editing a draft of this manuscript.

References

[1] D. Stumpfe, J. Bajorath, *J. Med. Chem.* **2012**, *55*, 2932–2942.
[2] A. M. Wassermann, M. Wawer, J. Bajorath, *J. Med. Chem.* **2010**, *53*, 8209–8223.
[3] S. Kumar, K. Kassahun, R. A. Tschirret-Guth, K. Mitra, T. A. Baillie, *Curr. Opin. Drug Discovery Dev.* **2008**, *11*, 43–52.
[4] A. T. Maynard, C. D. Roberts, *J. Med. Chem.* **2016**, *59*, 4189–4201.

[5] P. W. Kenny, J. Sadowski, in *Chemoinformatics in Drug Discovery* (Ed.: T. I. Opera), Wiley - VCH Verlag GmbH & Co. KGaA, Weinheim, **2005**, pp 271–285.
[6] X. Hu, Y. Hu, M. Vogt, D. Stumpfe, J. Bajorath, *J. Chem. Inf. Model.* **2012**, *52*, 1138–1145.
[7] N. Furtmann, Y. Hu, M. Gütschow, J. Bajorath, *Chem. Biol. Drug Des.* **2015**, *86*, 1458–1465.
[8] G. M. Maggiora, *J. Chem. Inf. Model.* **2006**, *46*, 1535.
[9] M. A. Johnson, G. M. Maggiora, *Concepts and Applications of Molecular Similarity*, John Wiley & Sons., New York, **1990**.
[10] K. Heikamp, X. Hu, A. Yan, J. Bajorath, *J. Chem. Inf. Model.* **2012**, *52*, 2354–2365.
[11] D. Horvath, G. Marcou, A. Varnek, S. Kayastha, A. de la Vega de León, J. Bajorath, *J. Chem. Inf. Model.* **2016**, *56*, 1631–1640.
[12] V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer, Berlin, **2000**.
[13] S. Tamura, T. Miyao, K. Funatsu, *J. Chem. Inf. Model.* **2019**, *59*, 2656–2663.
[14] D. Mendez, A. Gaulton, A. P. Bento, J. Chambers, M. D. Veij, E. Félix, M. P. Magariños, J. F. Mosquera, P. Mutowo, M. Nowotka, M. Gordillo-Marañón, F. Hunter, L. Junco, G. Mugumbate, M. Rodriguez-Lopez, F. Atkinson, N. Bosc, C. J. Radoux, A. Segura-Cabrera, A. Hersey, A. R. Leach, *Nucleic. Acids. Res.* **2019**, *47*, D930–D940.
[15] J. Hussain, C. Rea, *J. Chem. Inf. Model.* **2010**, *50*, 339–348.
[16] M. Wawer, J. Bajorath, *J. Med. Chem.* **2011**, *54*, 2944–2951.
[17] A. Ghosh, D. Dimova, J. Bajorath, *Med. Chem. Commun.* **2016**, *7*, 237–246.
[18] D. Rogers, M. Hahn, *J. Chem. Inf. Model.* **2010**, *50*, 742–754.
[19] L. Ralaivola, S. J. Swamidass, H. Saigo, P. Baldi, *Neural Networks*. **2005**, *18*, 1093–1110.
[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
[21] K. Roy, S. Kar, P. Ambure, *Chemom. Intell. Lab. Syst.* **2015**, *145*, 22–29.