Proposal for a Large-scale High-quality Dataset of Activity Cliffs

Anonymous Author(s)

Abstract

Activity cliffs—pairs of structurally similar molecules that display large differences in binding affinity—pose a fundamental challenge in structure-based drug discovery. They highlight subtle yet critical determinants of protein-ligand recognition and provide stringent test cases for computational methods. This proposal aims to establish a large-scale, high-quality dataset of activity cliffs to enable systematic study of structure-activity discontinuities and to benchmark both affinity prediction and docking approaches. We have already curated a large-scale dataset containing over 16k activity cliff pairs across 50 human protein targets from ChEMBL. Future development will focus on validating affinity data under unified experimental conditions and integrating structural annotations of representative molecular pairs. The long-term goal is to develop an open, community-driven database of activity cliffs that will accelerate method development and provide actionable insights for drug discovery.

14 1 Introduction

2

3

4

5

6

9

10

11

12

13

30

31

15 In medicinal chemistry and structure-based drug 16 discovery (SBDD), a central challenge lies in accurately capturing the relationship between molec-17 ular structure and biological activity [2]. While 18 most structurally similar compounds tend to ex-19 20 hibit comparable activities, there exist cases where small, localized chemical modifications lead to dis-21 proportionately large differences in potency [7]. 22 These cases, termed activity cliffs (ACs), repre-23 sent abrupt discontinuities in the structure-activity 24 landscape [8]. Typically, an activity cliff is iden-25 tified from a matched molecular pair (MMP), i.e., 26 27 two compounds that differ only by a small chem-28 ical substitution while sharing a common scaffold, vet display a binding affinity difference exceeding 29

a predefined threshold [3], as shown in Figure 1.

Such cliffs are of particular interest because they

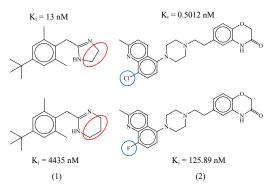


Figure 1: Two examples of activity cliffs shown in [4].

encode critical insights into protein-ligand recognition and highlight subtle interaction determinants that are often missed by global similarity-based models [9].

A high-quality dataset of activity cliffs is therefore invaluable for advancing SBDD. At a mechanistic level, ACs support understanding protein-ligand binding behaviors, by pinpointing local structural features that drive affinity changes[5]. Methodologically, ACs have already been leveraged for benchmarking machine learning-based affinity prediction models, highlighting their weaknesses in regions of sharp structure-activity discontinuities [10]. AC datasets also have the potential to open

- the door to benchmarking protein-ligand docking methods, where the task is not merely to reproduce experimental poses but to predict relative affinity differences between similar ligands [6].
- Despite these advances, current AC datasets suffer from important limitations [12]. Most existing
- resources provide only affinity labels derived from heterogeneous experimental reports. These values
- 43 are subject to assay-specific conditions, measurement uncertainties, and inter-laboratory variability,
- 44 which complicate the interpretation of cliffs. Moreover, while affinity data highlight the outcome of
- 45 binding, they do not capture the structural basis of the cliff. As a result, computational methods trained
- 46 or benchmarked on such data may inherit experimental noise and fail to identify the fine-grained
- interaction mechanisms underlying the cliffs.
- To address these gaps, future development of activity cliff datasets should move toward multi-modal,
- 49 higher-fidelity resources. Three complementary directions are particularly promising:
 - Wet-lab validation under unified conditions. Systematic measurement of MMP binding affinities using standardized assays would minimize inconsistencies and improve the reliability of cliff labels.
 - **High-precision binding structures.** Obtaining crystallographic or cryo-EM structures for representative MMPs would directly reveal how small structural perturbations alter binding geometry.
 - Molecular dynamics (MD) simulations. High-quality MD studies of AC pairs could
 provide atomistic insight into dynamic and entropic contributions that static docking and
 affinity labels cannot capture.
- Together, these directions would transform activity cliff datasets from affinity-only collections into comprehensive interaction-aware benchmarks, enabling deeper understanding of protein-ligand recognition, more robust evaluation of predictive models, and ultimately more effective structure-based drug discovery.

3 2 Data Curation

50

51

52

53

54

55

56

57

58

- We have already constructed a large-scale dataset of activity cliffs, curated from the ChEMBL 64 database [11]. Specifically, matched molecular pairs (MMPs) were extracted by applying strict 65 structural filters: pairs differ only by a localized chemical substitution while sharing a common 66 scaffold, ensuring chemical interpretability and medicinal relevance. Activity cliffs were defined as 67 MMPs exhibiting at least a 100-fold difference in binding affinity toward the same human protein 68 target, based on high-confidence, direct binding assay measurements (K_i values). To avoid ambiguity, 69 only pairs with <10-fold affinity differences were included as non-AC controls, while intermediate 70 cases were excluded. 71
- The resulting dataset comprises over 55 498 protein-ligand binding pairs across 50 diverse human protein targets, from which 16 386 activity cliff pairs and 271 196 non-AC MMPs are extracted to form the dataset. All ligand SMILES were standardized and validated for 3D structure generation, and protein structures were sourced from the AlphaFold Protein Structure Database [1], ensuring compatibility with structure-based modeling. The dataset is available at https://anonymous. 4open.science/r/ACDataset-1A4B, and more details are shown in Appendix A.
- Looking forward, the dataset can be further developed into a more powerful resource by incorporating validated affinity data under unified experimental conditions and structural information for representative MMPs, as discussed in Section 1. Such extensions will inevitably be costly, as they require new wet-lab measurements and high-resolution structural determination, but the investment is justified: a rigorously validated and structurally annotated AC dataset would provide unparalleled insights into protein-ligand recognition and set a new standard for evaluating computational methods in drug discovery.
- To maximize accessibility and impact, it is also important to establish a public, online open-source database of activity cliffs. By continuously updating the collection with newly validated cliffs, structural annotations, and simulation data, this platform would not only accelerate method development in docking and affinity prediction but also serve as a shared knowledge base for medicinal chemists and computational biologists.

References

- 91 [1] Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf 92 Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure 93 prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, 2024.
- [2] Leonardo G Ferreira, Ricardo N Dos Santos, Glaucius Oliva, and Adriano D Andricopulo.
 Molecular docking and structure-based drug design strategies. *Molecules*, 20(7):13384–13421,
 2015.
- [3] Xiaoying Hu, Ye Hu, Martin Vogt, Dagmar Stumpfe, and Jürgen Bajorath. Mmp-cliffs: system atic identification of activity cliffs on the basis of matched molecular pairs. *Journal of chemical information and modeling*, 52(5):1138–1145, 2012.
- 100 [4] Xiuyuan Hu, Guoqing Liu, Yang Zhao, and Hao Zhang. Activity cliff-aware reinforcement learning for de novo drug design. *Journal of Cheminformatics*, 17(1):54, 2025.
- Tiago Janela and Jürgen Bajorath. Anatomy of potency predictions focusing on structural
 analogues with increasing potency differences including activity cliffs. *Journal of Chemical Information and Modeling*, 63(22):7032–7044, 2023.
- [6] Zijing Liu, Xinni Zhang, Yankai Chen, Bin Feng, Mingjun Yang, Zenglin Xu, Yu Li, and Irwin King. Dockedac: Empowering deep learning models with 3d protein-ligand data for activity cliff analysis. *Openreview*, 2025.
- 108 [7] Dagmar Stumpfe and Jürgen Bajorath. Exploring activity cliffs in medicinal chemistry: miniper-109 spective. *Journal of medicinal chemistry*, 55(7):2932–2942, 2012.
- [8] Dagmar Stumpfe, Huabin Hu, and Jürgen Bajorath. Evolving concept of activity cliffs. ACS
 omega, 4(11):14360–14368, 2019.
- 112 [9] Dagmar Stumpfe, Ye Hu, Dilyana Dimova, and Jürgen Bajorath. Recent progress in understand-113 ing activity cliffs and their utility in medicinal chemistry: miniperspective. *Journal of medicinal* 114 *chemistry*, 57(1):18–28, 2014.
- [10] Derek van Tilborg, Alisa Alenicheva, and Francesca Grisoni. Exposing the limitations of
 molecular machine learning with activity cliffs. *Journal of Chemical Information and Modeling*,
 62(23):5938–5951, 2022.
- 118 [11] Barbara Zdrazil, Eloy Felix, Fiona Hunter, Emma J Manners, James Blackshaw, Sybilla Corbett,
 119 Marleen de Veij, Harris Ioannidis, David Mendez Lopez, Juan F Mosquera, et al. The chembl
 120 database in 2023: a drug discovery platform spanning multiple bioactivity data types and time
 121 periods. *Nucleic acids research*, 52(D1):D1180–D1192, 2024.
- 122 [12] Ziqiao Zhang, Bangyi Zhao, Ailin Xie, Yatao Bian, and Shuigeng Zhou. Activity cliff prediction: Dataset and benchmark. *arXiv preprint arXiv:2302.07541*, 2023.

124 A Dataset Details

Table 1: Schema of the dataset of activity cliffs. Each row represents a matched molecular pair (MMP) for a given protein target, including SMILES strings of the two ligands, their experimental K_i values from ChEMBL, and a binary activity cliff label.

Protein Target	Ligand 1	K_{i1}	Ligand 2	K_{i2}	AC Label
ID1	\mathfrak{m}_1	$K_i(\mathfrak{m}_1)$	\mathfrak{m}_2	$K_i(\mathfrak{m}_2)$	Yes
ID1	\mathfrak{m}_3	$K_i(\mathfrak{m}_3)$	\mathfrak{m}_4	$K_i(\mathfrak{m}_4)$	No
ID2	\mathfrak{m}_5	$K_i(\mathfrak{m}_5)$	\mathfrak{m}_6	$K_i(\mathfrak{m}_6)$	Yes
	•••	•••	•••	•••	•••

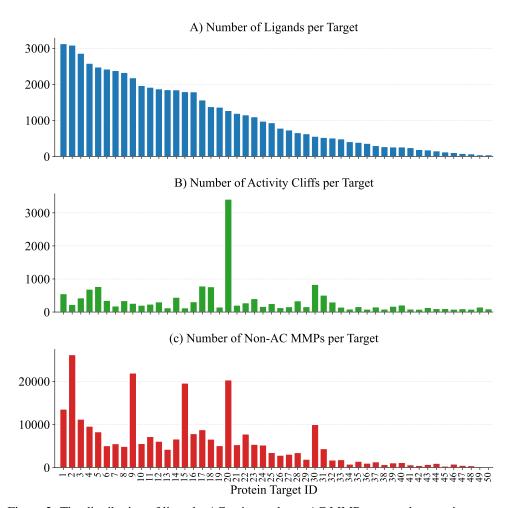


Figure 2: The distribution of ligands, AC pairs, and non-AC MMPs across the protein targets.