Dissertation for the degree of Doctor of Medicine and Health Sciences in the discipline of pharmaceutical sciences

Synthesis and study of anticancer and antibacterial properties of new 4-thiazolidinone derivatives

MPharm Tomasz Szostek

Supervisor: Associate Professor Daniel Szulczyk

Assistant supervisor: PhD Dagmara Otto-Ślusarczyk

Chair and Department of Biochemistry

Faculty of Medicine

Medical University of Warsaw

Doctoral dissertation defense before the Pharmaceutical Sciences Council of the Medical University of Warsaw

Warsaw, 2026

Rozprawa na stopień doktora nauk medycznych i nauk o zdrowiu w dyscyplinie nauki farmaceutyczne

Synteza i badanie właściwości przeciwnowotworowych i przeciwbakteryjnych nowych pochodnych 4-thiazolidinonu

mgr farm. Tomasz Szostek

Promotor: Dr hab. Daniel Szulczyk

Promotor pomocniczy: dr Dagmara Otto-Ślusarczyk

Katedra i Zakład Biochemii

Wydział Lekarski

Warszawski Uniwersytet Medyczny

Obrona rozprawy doktorskiej przed Radą Dyscypliny Nauk Farmaceutycznych Warszawskiego Uniwersytetu Medycznego

Warszawa, 2026

Dedicated to my beloved Grandmother, whose memory will always remain with me, and to my beloved Mother, with heartfelt gratitude for your endless love, unwavering support, and constant belief in me.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisor, Dr. hab. Daniel Szulczyk, for his invaluable guidance, mentorship, and unwavering support throughout my doctoral journey. I am profoundly grateful not only for the scientific knowledge he shared with me, but also for shaping me both as a researcher and as a person. He taught me that hard work, perseverance, and determination are the foundations of achieving meaningful goals, lessons that will remain with me throughout my career.

I would also like to sincerely thank my co-supervisor, Dr. Dagmara Otto-Ślusarczyk, for introducing me to experimental biological research and for patiently teaching me laboratory techniques. Her willingness to help, openness, and constant support whenever I needed advice have been invaluable throughout my doctoral studies.

My sincere thanks also go to Professor Marta Struga and all members of the Chair and Department of Biochemistry, Medical University of Warsaw, for creating an inspiring scientific environment, for their support, and for providing me with the opportunity to develop as a scientist while conducting my research.

I am deeply grateful to all my friends for their encouragement, understanding, and support throughout this journey. Their presence made both the challenging and the rewarding moments of my doctoral studies much easier to navigate.

Finally, I would like to express my deepest gratitude to my family, especially my beloved Mother and Stepfather, and to the memory of my Grandmother, for their love, endless encouragement, and for always being by my side throughout this journey. Their presence, strength, and support have meant more to me than words can express.

Table of contents

ACKNOWLEDGEMENTS 7

TABLE OF CONTENTS 9

1. SCIENTIFIC PROFILE OF THE DOCTORAL CANDIDATE 11

2. LIST OF PUBLICATIONS INCLUDED IN THE DOCTORAL THESIS 13

3. COMPLETE LIST OF PUBLICATIONS 14

4. SCIENTIFIC INTERNSHIPS 16

5. OTHER SIGNIFICANT INFORMATION 17

6. SUMMARY OF SCIENTIFIC ACHIEVEMENTS 18

7. ABSTRACTS 19

7.1. Abstract 19
7.2. Abstract in Polish 21

8. INTRODUCTION WITH JUSTIFICATION OF THE RESEARCH TOPIC 23

8.1. Drug discovery 23
8.2. Thiazolidinones as privileged scaffolds in medicinal chemistry 24
8.3. Navigating the underexplored regions of Structure-Activity Landscape 26
8.4. Artificial Intelligence in modern drug discovery 29

9. AIM OF THE THESIS 32

10. REALIZATION OF THE SCIENTIFIC OBJECTIVES 33

11. MATERIALS AND METHODS 34

11.1. Literature analysis of 4-thiazolidinone derivatives 34
11.2. Chemistry, biological evaluation, and molecular modelling 34
11.3. Activity cliff-aware approaches in small-molecule drug discovery 36

12. RESULTS 39

12.1. 4-thiazolidinone chemical space 39
12.2. Novel 4-thiazolidinone derivatives 39
12.3. Activity cliff-aware strategies for improving molecular modelling 41
12.4. *De novo* generation of the new thiazolidinone derivatives 42

13. DISCUSSION 46

13.1. Fragmented thiazolidinones SAR defined the problem 46
13.2. Structural certainty is essential 47
13.3. Experiments shaped future modelling 48
13.4. Turning limitations into advantage 49
13.5. Study limitations 50

14. ABBREVIATIONS 51

15. REFERENCES 53

16. PUBLICATION I 59

17. PUBLICATION II 71

18. PUBLICATION III 101

19. PUBLICATION IV 116

20. AUTHORS' CONTRIBUTION STATEMENTS 139

1. Scientific profile of the doctoral candidate

Tomasz Szostek obtained his Master of Pharmacy degree from the Medical University of Warsaw in 2023 and has been a PhD candidate at the Doctoral School of the Medical University of Warsaw since October 2022. To strengthen his computational expertise, he completed postgraduate studies in Bioinformatics at the Polish-Japanese Academy of Information Technology (PJATK) in 2024, gaining advanced knowledge in cheminformatics, molecular modelling, and data science applied to drug discovery.

His research focuses on medicinal chemistry and computer-aided drug discovery, integrating synthetic organic chemistry with computational approaches for the design of biologically active small molecules. His work combines molecular design, chemical synthesis, biological evaluation, QSAR modelling, molecular docking, predictive toxicology, explainable artificial intelligence, *de novo* molecular generation, and chemical space exploration to develop innovative methodologies supporting early-stage drug discovery.

Since the beginning of his doctoral studies in 2022, his research has centered on the discovery of novel 4-thiazolidinone derivatives as potential anticancer agents, combining synthetic medicinal chemistry with computational modelling and experimental biological validation. This research was supported by a Young Researcher ("Młody Badacz") grant awarded by the Medical University of Warsaw in 2023 and has resulted in multiple peer-reviewed publications forming the scientific foundation of his doctoral dissertation. In 2026, he was awarded a second Young Researcher grant to support the next stage of his research and the development of new computational approaches to drug discovery.

Since 2025, he has also been involved in the INTEGRA collaborative research project, jointly conducted by the Medical University of Warsaw and the Warsaw University of Technology, where he contributes to the development of next-generation methodologies for predicting biological activity by integrating molecular representations with proteome-wide interaction data. This interdisciplinary project aims to bridge molecular, protein, and cellular levels of biological information to improve phenotypic prediction models for drug discovery.

His current scientific interests focus on the development of new interpretable and robust methods for predictive toxicology, with particular emphasis on integrating computational

modelling and experimental validation to accelerate the identification of safe and effective drug candidates.

2. List of publications included in the doctoral thesis

The application for the doctoral degree in Medical and Health Sciences, within the discipline of Pharmaceutical Sciences, is based on a series of four scientific papers published between 2024-2026 in peer-reviewed journals, with a total Impact Factor of **24.0** and a combined score of **420** according to the Ministry of Education and Science.

- **Publication 1 (P.I) – Review Article**

  IF = 4.2; MNiSW = 20

  Szostek, T., Otto-Ślusarczyk, D., Roszkowski, P., Struga, M., & Szulczyk, D. (2024). Exploring the bioactive potential of (2-imino-4-oxo-1,3-thiazolidin-5-yl)acetic acid derivatives: A comprehensive review. *Results in Chemistry*, *11*, 101828. https://doi.org/10.1016/J.RECHEM.2024.101828

- **Publication 2 (P.II) – Original Research Article**

  IF = 3.2; MNiSW = 100

  Szostek, T., Otto-Ślusarczyk, D., Dobrowolski, M. A., Wiśniewski, M., Roszkowski, P., & Szulczyk, D. (2025). Synthesis, Cytotoxicity Evaluation, and Molecular Dynamics of 4-Thiazolidinone Derivatives: A Potential Path to EGFR-Targeted Anticancer Therapy. *Current Medicinal Chemistry, 33.* https://doi.org/10.2174/0109298673400295250913044755

- **Publication 3 (P.III)– Review Article**

  IF = 8.7; MNiSW = 200

  Szostek, T., & Szulczyk, D.

  From obstacle to design advantage: activity cliff aware modeling for small-molecule drug discovery. *Drug Discovery Today, 31(1).*

  https://doi.org/10.1016/J.DRUDIS.2025.104589

- **Publication 4 (P.IV)– Original Research Article**

  IF = 7.9; MNiSW = 100

  Szostek, T., Wiśniewski, M., Ballabio, D., Consonni, V., Plewczyński, D., & Szulczyk, D.

  De novo design of anticancer 4-thiazolidinone derivatives: a generative framework shaped by activity cliffs. *Journal of Cheminformatics, 18(1), 88.* https://doi.org/10.1186/s13321-026-01216-3

3. Complete List of Publications

WARSZAWSKI
UNIWERSYTET
MEDYCZNY

BIBLIOTEKA UCZELNIANA

Nr referencyjny
BIBG/Punktacja/ 389 /2026/KK

Warszawa, 04.08.2026

BIBLIOMETRIC ANALYSIS OF THE COMPLETE SCIENTIFIC PUBLICATION OUTPUT
OF **MR TOMASZ SZOSTEK**
FOR THE DOCTORAL DEGREE PROCEEDINGS

No.	Bibliographic description	Impact Factor	MNiSW	Highest quartile[1]
I. Articles published in scientific journals or peer-reviewed conference proceedings included in the current list of the Polish Ministry of Science and Higher Education (MNiSW) [2]				
1.	Olejarz W, Augustynowicz-Kopeć E, Głogowska A, Napiórkowska M, **Szostek T**, Filipek A, Szulczyk D. In Vitro Anticancer Studies of New Derivatives based on the Furanocoumarin Scaffold. Scientific Reports. 2026;16:1-18 *Publication type: Original article*	4,900	140	Q1
2.	Szymańska-Majchrzak J, Głogowska A, Augustynowicz-Kopeć E, Greber K, Ciura K, Olejarz W, **Szostek T**, Struga M, Szulczyk D. Discovery of a tetrazole-thiourea derivative as a potential active agent against multidrug-resistant Staphylococcus aureus and Mycobacterium tuberculosis. Pharmacological Reports. 2026;78:862-874 *Publication type: Original article*	4,500	100	Q2
3.	**Szostek T** [corresponding author], Szulczyk D. From Obstacle to Design Advantage: Activity Cliff Aware Modeling for Small-Molecule Drug Discovery. Drug Discovery Today. 2026;31(1):1-14 *Publication type: Review article*	8,700	200	Q1
4.	**Szostek T** [corresponding author], Wiśniewski M, Ballabio D, Consonni V, Plewczyńska D, Szulczyk D. De novo design of anticancer 4-thiazolidinone derivatives: a generative framework shaped by activity cliffs. Journal of Cheminformatics. 2026;18:1-22 *Publication type: Original article*	7,900	100	Q1

[1] Quartile in the year of publication, according to the Journal Impact Factor.
[2] The list was prepared in accordance with the regulations issued pursuant to Article 267(2)(2)(b) of the Act of 20 July 2018 – Law on Higher Education and Science (Journal of Laws of 2022, item 574, as amended). The list constitutes an appendix to the announcement of the Ministry of Science and Higher Education of 5 January 2024 on the list of scientific journals and peer-reviewed conference proceedings

ul. Żwirki i Wigury 63
02-091 Warszawa
www.biblioteka.wum.edu.pl
tel.: +48 22 116 60 11
biblioteka@wum.edu.pl

5.	**Szostek T**, Otto-Ślusarczyk D, Dobrowolski M, Wiśniewski M, Roszkowski P, Szulczyk D. Synthesis, Cytotoxicity Evaluation, and Molecular Dynamics of 4-Thiazolidinone Derivatives: A Potential Path to EGFR-Targeted Anticancer Therapy. Current Medicinal Chemistry. 2025:1-29 *Publication type: Original article*	3,200	100	Q2
6.	**Szostek T**, Otto-Ślusarczyk D, Roszkowski P, Struga M, Szulczyk D. Exploring the bioactive potential of (2-imino-4-oxo-1,3-thiazolidin-5-yl) acetic acid derivatives: A comprehensive review. Results in Chemistry. 2024;11:1-11 *Publication type: Review article*	4,200	20	Q2
7.	Lasek P, Kosikowska U, Kołodziej P, Kubiak-Tomaszewska G, Krzyżanowska N, **Szostek T**, Struga M, Feldo M, Bogucka-Kocka A, Wujec M. New Thiosemicarbazide Derivatives with Multidirectional Biological Action. Molecules. 2024;29(7):1-18 *Publication type: Original article*	4,600	140	Q2
8.	**Szostek T**, Szulczyk D, Szymańska-Majchrzak J, Koliński M, Kmiecik S, Otto-Ślusarczyk D, Zawodnik A, Rajkowska E, Chaniewicz K, Struga M, Roszkowski P. Design and Synthesis of Menthol and Thymol Derived Ciprofloxacin: Influence of Structural Modifications on the Antibacterial Activity and Anticancer Properties. International Journal of Molecular Sciences. 2022;23(12):1-19 *Publication type: Original article*	5,600	140	Q1
	Total:	43,600	940	-
II. Articles published before 1 January 2019 in journals included in the Ministry list dated 25 January 2017, provided that the journal received at least 10 points.				
	Total:	-	-	-
III. Other articles				
	Total:	-	-	-
	Total (part I- III):	**43,600**	**940**	-
IV. Scientific monographs/chapters in monographs published by publishers included in the list of the Ministry of Science and Higher Education[3] or by organizational units of entities whose publishers are included in this list				
-				
V. Other monographs or book chapters				
-				
VI. Patents				
-				

DYREKTOR
Biblioteki Uczelnianej
Warszawskiego Uniwersytetu Medycznego
mgr Agnieszka Czarnecka

[3] The list has been prepared in accordance with the regulations issued under Article 267(2)(2)(a) of the Act of 20 July 2018 – Law on Higher Education and Science (Journal of Laws 2022, item 574, as amended). The list of publishers of peer-reviewed scholarly monographs was announced by the Minister of Education and Science on 22 July 2021

ul. Żwirki i Wigury 63
02-091 Warszawa
www.biblioteka.wum.edu.pl
tel.: +48 22 116 60 11
biblioteka@wum.edu.pl

4. Scientific internships

One of the most important milestones in my scientific development was a research stay completed under the Erasmus+ Programme between February and May 2025. The visit took place at the Milano Chemometrics and QSAR Research Group, Department of Earth and Environmental Sciences, University of Milano-Bicocca (Italy), under the supervision of Prof. Davide Ballabio, an internationally recognized expert in chemometrics, QSAR modelling, and machine learning applications in chemistry.

The research stay provided an opportunity for intensive collaboration on advanced computational approaches to molecular property prediction, quantitative structure–activity relationship (QSAR) modelling, and the development of benchmarking strategies for predictive models. The work also addressed key methodological challenges in computational drug discovery, including model validation, applicability domain assessment, explainable artificial intelligence, and the characterization of local structure–activity relationships across chemical space.

Beyond the scientific outcomes, the mobility significantly strengthened my experience in conducting research within an international and interdisciplinary environment, fostering collaborations with leading experts in computational chemistry. The knowledge and skills acquired during the stay have been directly incorporated into my doctoral research and have led to an ongoing scientific collaboration with Prof. Ballabio's research group, resulting in joint publications and the development of new research initiatives in artificial intelligence-driven drug discovery.

5. Other significant information

In parallel with my doctoral research, I have been working in the Real-World Evidence (RWE) department at AstraZeneca, where I contribute to the planning, execution, and dissemination of observational studies. My responsibilities include study design, protocol development, statistical planning, scientific communication, and the preparation of peer-reviewed publications across therapeutic areas, particularly oncology and respiratory medicine. This experience has strengthened my ability to bridge academic research with industry-driven evidence generation and has provided valuable insight into the translation of clinical data into real-world healthcare outcomes.

Alongside my research activities, I actively disseminated the results of my work at national scientific meetings. During my doctoral studies, I delivered presentations at the XXXIX and XL Scientific Conferences of the Polish Medicinal Chemistry Society, held in Lublin (2023) and Gdańsk (2024), respectively. These conferences provided an opportunity to present my research to the medicinal chemistry community, receive valuable scientific feedback, and establish new academic collaborations.

My scientific profile has also attracted the interest of international innovation partners. Following discussions initiated through my research activities, I have been invited to collaborate with I&I Prague, a biotechnology incubator and innovation hub, where we are currently exploring the establishment of a joint incubation initiative aimed at translating my research into innovative drug discovery technologies. These ongoing negotiations represent an important opportunity to transform academic research into practical applications with commercial and societal impact.

Demonstrating my growing research independence, I also prepared and submitted my first competitive research grant proposal as Principal Investigator to the National Science Centre (NCN) PRELUDIUM 25 call in 2026. The proposal focuses on the development of novel methodologies for toxicity prediction.

6. Summary of Scientific Achievements

Category	Scientific achievement
Scientific publications	8 peer-reviewed publications
Total Impact Factor	43.6
Total MNiSW score	940 points
Doctoral thesis publication series	4 publications: 2 original research articles and 2 review articles
Leading authorship	4 first-author publications, including 2 as corresponding author
International research experience	Erasmus+ research internship, University of Milano-Bicocca, Italy (2025)
Research collaboration	INTEGRA project – Medical University of Warsaw and Warsaw University of Technology
Research funding	2 Young Researcher (“Młody Badacz”) grants, Medical University of Warsaw (2023, 2026)
Competitive grant applications	NCN PRELUDIUM 25 proposal submitted as Principal Investigator (2026)
Scientific conference activity	Presentations at the XXXIX and XL Scientific Conferences of the Polish Medicinal Chemistry Society (2023, 2024)
Pharmaceutical industry experience	Real-World Evidence, AstraZeneca
Innovation and technology transfer	Collaboration with I&I Prague toward the development of a potential incubation initiative

7. Abstracts

7.1. Abstract

The discovery of new small-molecule therapeutics remains one of the major challenges in modern medicine. Although advances in medicinal chemistry have expanded the number of biologically active compounds, only a small fraction of accessible chemical space has been explored. Traditional lead optimization is highly effective within individual analogue series but provides only limited insight into the broader structure–activity landscape. As a result, many promising regions of chemical space remain unexplored.

Among privileged heterocyclic scaffolds, 4-thiazolidinones have attracted considerable attention because of their broad biological activity and exceptional structural versatility. Hundreds of derivatives have been reported with anticancer, antibacterial and other pharmacological properties. However, knowledge of their structure–activity relationship remains fragmented across independent studies. This limits the identification of general design principles and hinders the rational exploration of new regions of thiazolidinone chemical space.

The aim of this doctoral thesis was to develop an integrated framework for the discovery and optimization of biologically active 4-thiazolidinone derivatives by combining medicinal chemistry with modern computational approaches. Local structure-activity relationship discontinuities and activity cliffs were treated as sources of design information and ultimately incorporated into *de novo* molecular generation.

The work combined compound design, organic synthesis, biological evaluation and computational modelling. Novel 4-thiazolidinone derivatives were synthesized and evaluated for anticancer and antibacterial activity, leading to the identification of promising lead compounds. Molecular docking and molecular dynamics simulations were used to investigate ligand–target interactions and support the interpretation of experimental results. Quantitative structure–activity relationship (QSAR) models were developed to characterize structure–activity relationships and predict biological activity across thiazolidinone chemical space.

The thesis then extended beyond conventional SAR analysis. Activity cliffs were investigated as local regions of the structure–activity landscape where small structural modifications produce large changes in biological activity. These regions represent highly informative molecular events but also remain one of the greatest challenges for modern machine learning

models. Rather than treating activity cliffs as prediction errors, this work explored their potential as sources of medicinal chemistry knowledge. Their role was analyzed across the AI pipeline, from molecular representation and model evaluation to generative molecular design.

Building on these findings, an activity cliff-informed *de novo* design strategy was developed for the rational generation of novel 4-thiazolidinone derivatives. The proposed framework integrates experimental data, predictive modelling and local SAR knowledge to prioritize compounds with improved biological potential. This work demonstrates that combining medicinal chemistry with artificial intelligence enables a more efficient exploration of thiazolidinone chemical space than conventional analogue-based optimization alone. More broadly, it highlights activity cliffs as actionable design principles for next generation drug discovery.

Keywords: 4-thiazolidinones; Medicinal Chemistry; Activity Cliffs; Artificial Intelligence

7.2. Abstract in Polish

Odkrywanie nowych leków małocząsteczkowych pozostaje jednym z największych wyzwań współczesnej medycyny. Pomimo ogromnego postępu chemii medycznej liczba związków przebadanych eksperymentalnie stanowi jedynie niewielki fragment dostępnej przestrzeni chemicznej. Klasyczna optymalizacja związków wiodących pozwala skutecznie rozwijać pojedyncze serie analogów, jednak dostarcza jedynie lokalnej wiedzy o zależnościach struktura–aktywność. W konsekwencji wiele potencjalnie wartościowych obszarów przestrzeni chemicznej pozostaje nadal niewykorzystanych.

Wśród uprzywilejowanych układów heterocyklicznych szczególne miejsce zajmują 4-tiazolidynony, które od wielu lat stanowią przedmiot intensywnych badań ze względu na swoją różnorodność strukturalną oraz szerokie spektrum aktywności biologicznej. Opisano setki pochodnych wykazujących między innymi aktywność przeciwnowotworową i przeciwbakteryjną. Dotychczasowa wiedza dotycząca zależności struktura–aktywność tych związków pozostaje jednak rozproszona pomiędzy licznymi, niezależnymi seriami analogów. Utrudnia to identyfikację uniwersalnych zasad projektowania nowych cząsteczek oraz racjonalną eksplorację słabo poznanych obszarów przestrzeni chemicznej 4-tiazolidynonów.

Celem rozprawy było opracowanie zintegrowanego podejścia do odkrywania i optymalizacji biologicznie aktywnych pochodnych 4-tiazolidynonu poprzez połączenie klasycznych metod chemii medycznej z nowoczesnymi metodami sztucznej inteligencji. Szczególną uwagę poświęcono lokalnym zależnościom struktura–aktywność, klifom aktywności oraz projektowaniu cząsteczek *de novo*.

Praca obejmowała projektowanie związków, ich syntezę, ocenę aktywności biologicznej oraz analizy obliczeniowe. Otrzymano nowe pochodne 4-tiazolidynonu, które oceniono pod względem aktywności przeciwnowotworowej i przeciwbakteryjnej, identyfikując najbardziej obiecujące struktury wiodące. W celu wyjaśnienia molekularnych podstaw obserwowanej aktywności zastosowano dokowanie molekularne oraz symulacje dynamiki molekularnej. Zależności struktura–aktywność analizowano również z wykorzystaniem modeli QSAR, które umożliwiły identyfikację cech strukturalnych odpowiedzialnych za aktywność biologiczną oraz predykcję właściwości nowych pochodnych.

Rozprawa wykracza jednak poza klasyczną analizę SAR. Szczególną uwagę poświęcono klifom aktywności, czyli lokalnym obszarom krajobrazu struktura–aktywność, w których

niewielkie zmiany struktury prowadzą do dużych zmian aktywności biologicznej. Regiony te stanowią jedno z największych wyzwań dla współczesnych modeli uczenia maszynowego, ale jednocześnie zawierają wyjątkowo cenną wiedzę z punktu widzenia chemii medycznej. W niniejszej pracy klify aktywności potraktowano nie jako źródło błędów predykcyjnych, lecz jako nośniki informacji wspierających racjonalne projektowanie nowych cząsteczek. Przeanalizowano ich znaczenie na kolejnych etapach wykorzystania sztucznej inteligencji w odkrywaniu leków – od reprezentacji molekularnych i oceny modeli predykcyjnych po generatywne projektowanie związków.

Na podstawie uzyskanych wyników opracowano strategię projektowania cząsteczek *de novo* uwzględniającą informacje pochodzące z klifów aktywności. Zaproponowane podejście integruje dane eksperymentalne, modelowanie predykcyjne oraz lokalną wiedzę o zależnościach struktura–aktywność w celu identyfikacji nowych pochodnych 4-tiazolidynonu o zwiększonym potencjale biologicznym. Uzyskane wyniki pokazują, że połączenie chemii medycznej ze sztuczną inteligencją umożliwia bardziej efektywną eksplorację przestrzeni chemicznej niż klasyczna optymalizacja oparta wyłącznie na seriach analogów. Jednocześnie wskazują, że klify aktywności mogą stanowić praktyczne narzędzie wspierające projektowanie nowych związków oraz rozwój kolejnej generacji metod wykorzystywanych w odkrywaniu leków.

Słowa kluczowe: 4-Thiazolidinony; Chemia Medyczna; Klify Aktywności; Sztuczna Inteligencja

8. Introduction with justification of the research topic

8.1. Drug discovery

The discovery and development of new therapeutic agents remain one of the greatest scientific and technological challenges in modern medicine.[1,2] Despite remarkable advances in molecular biology, medicinal chemistry, and pharmacology, the process of bringing a new drug from the initial concept to clinical application remains exceptionally time-consuming, expensive, and associated with a high attrition rate.[3,4] It is estimated that only a small fraction of compounds entering preclinical development ultimately reach the pharmaceutical market, while the overall cost of developing a single drug frequently exceeds one billion US dollars.[5] Consequently, improving the efficiency of early-stage drug discovery has become one of the major priorities of contemporary pharmaceutical research.[6]

The need for innovative therapeutic strategies is particularly evident in oncology and infectious diseases.[7,8] Cancer remains one of the leading causes of mortality worldwide, with approximately 20 million new cases and 9.7 million deaths reported in 2022.[9] According to projections of the World Health Organization (WHO), the annual number of newly diagnosed cancer cases is expected to exceed 35 million by 2050, primarily due to population growth and ageing.[10] Despite the introduction of targeted therapies and immunotherapy, many malignancies eventually develop resistance to treatment, creating a continuous demand for new small-molecule drugs with improved efficacy and selectivity.[11] Similar challenges are observed in antibacterial therapy, where the rapid emergence of antimicrobial resistance has substantially reduced the effectiveness of many currently available antibiotics, prompting the WHO to recognize antimicrobial resistance as one of the greatest global public health threats of the twenty-first century.[12,13]

Following the growing demand for new therapeutic agents, medicinal chemistry has increasingly focused on the identification of molecular scaffolds capable of interacting with diverse biological targets while offering broad opportunities for structural optimization.[14,15] Such privileged scaffolds constitute an important starting point for the development of novel bioactive compounds, as subtle structural modifications often enable substantial changes in biological activity, selectivity, and pharmacokinetic properties.[16] Consequently, the search for versatile heterocyclic systems remains one of the major directions of modern small-molecule drug discovery.[17]

8.2. Thiazolidinones as privileged scaffolds in medicinal chemistry

The 4-thiazolidinone core is a five-membered heterocyclic system containing sulfur at position 1, nitrogen at position 3, and a carbonyl group at position 4, and it has been investigated in medicinal chemistry for several decades because of its considerable synthetic and pharmacological versatility.[18] The scaffold can be diversified primarily through substitution at the C2, N3, and C5 positions, enabling the introduction of aromatic, heteroaromatic, aliphatic, and pharmacophoric fragments without eliminating the characteristic heterocyclic framework.[19] In particular, modification at C5 frequently leads to 5-ene-4-thiazolidinone derivatives, which constitute one of the most extensively studied subclasses because the exocyclic double bond provides additional possibilities for conjugation, molecular recognition, and structure–activity relationship (SAR) optimization.[20] This structural adaptability has supported the development of compounds with anticancer, antibacterial, antifungal, antiviral, antiparasitic, anti-inflammatory, and metabolic activities.[21]

The medicinal relevance of this chemical family is also demonstrated by the presence of closely related thiazolidinone-derived motifs in clinically used drugs. Epalrestat, an aldose-reductase inhibitor approved in Japan for diabetic peripheral neuropathy, contains a 2-thioxo-4-thiazolidinone, or rhodanine, moiety and represents one of the clearest examples of the therapeutic applicability of a thiazolidinone skeleton structure.[22] By contrast, the clinically established antidiabetic agents pioglitazone and rosiglitazone belong to the 2,4-thiazolidinedione class and therefore contain an additional carbonyl group at C2.[23] Their clinical success provides broader evidence that thiazolidinone-based heterocycles can combine biological activity with physicochemical and pharmacological properties compatible with therapeutic use.[24]

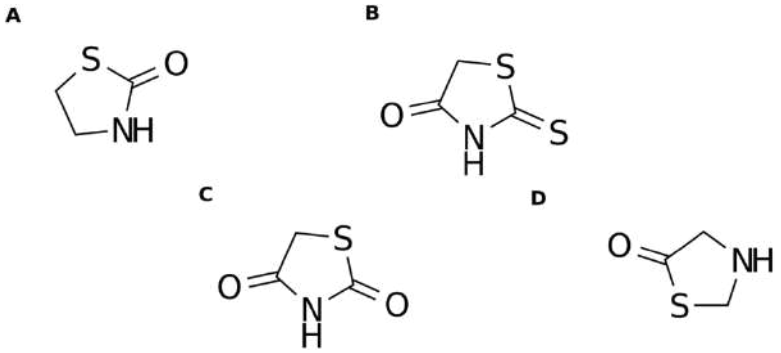


Figure 1. Representative structural variants of the thiazolidinone scaffold: (**A**) 1,3-thiazolidin-2-one, (**B**) 2-thioxo-1,3-thiazolidin-4-one, (**C**) 1,3-thiazolidine-2,4-dione, and (**D**) 1,3-thiazolidin-4-one.

Anticancer research represents one of the principal areas in which 4-thiazolidinone derivatives have been explored, with reported compounds affecting protein kinases, apoptotic pathways, oxidative balance, cell-cycle progression, and cytoskeletal organization.[25] The versatility of the scaffold has also encouraged the construction of hybrid molecules in which the 4-thiazolidinone ring is connected to another pharmacologically relevant fragment to improve potency or enable simultaneous modulation of multiple cancer-associated targets.[26]

One important design direction has involved the development of kinase inhibitors targeting pathways associated with cancer cell proliferation and survival. In a series of thiazolidin-4-one derivatives designed as dual EGFR/CDK2 inhibitors, compounds 5d, 5e, and 5f were identified as the most active antiproliferative members and displayed substantial inhibitory activity against both enzymes.[27] More recent structure-guided investigations have likewise shown that appropriately substituted thiazolidin-4-ones may inhibit wild-type EGFR at submicromolar concentrations, with reported IC_{50} values of 0.043-0.200 μM.[28]

Recent studies continue to demonstrate the remarkable anticancer potential of 4-thiazolidinone derivatives. For example, El-Zend et al. (2026) reported a series of multitarget 2,3-disubstituted 1,3-thiazolidin-4-one derivatives, among which the lead compound exhibited potent antiproliferative activity against MDA-MB-231 and MCF-7 breast cancer cells (IC_{50} = 2.25 and 6.70 μM, respectively). Mechanistic studies revealed G0/G1 cell-cycle arrest, caspase-3/7-mediated apoptosis, and simultaneous modulation of multiple cancer-related pathways, highlighting the potential of 4-thiazolidinones as promising multitarget anticancer agents.[29]

4-Thiazolidinone derivatives have also been extensively studied as potential antibacterial agents, including compounds active against both Gram-positive and Gram-negative microorganisms.[30] Their proposed antibacterial mechanisms include interference with bacterial cell-wall biosynthesis and inhibition of enzymes essential for DNA replication or other cellular processes.[31] In particular, inhibition of UDP-*N*-acetylenolpyruvylglucosamine reductase, commonly referred to as MurB, has been described as a potential mechanism because this enzyme participates in an essential stage of peptidoglycan biosynthesis and has no direct human counterpart.[32]

The activity of this chemical class is strongly dependent on substitution pattern and bacterial species. In an experimental study of twenty-one 2,3-diaryl-thiazolidin-4-ones, all investigated compounds showed antibacterial activity against the tested Gram-positive and Gram-negative strains, with Minimal Inhibitory Concentration (MIC) values ranging from 0.008 to 0.24 mg/mL and MBC values ranging from 0.0016 to 0.48 mg/mL.[33] Other studies combining microwave-assisted synthesis, microbiological evaluation, and molecular docking have identified 4-thiazolidinone derivatives with substantial antimicrobial effects and have suggested that appropriately positioned aromatic and halogen-containing substituents may enhance interactions with bacterial molecular targets.[34]

8.3. Navigating the underexplored regions of Structure-Activity Landscape

Despite the considerable number of thiazolidinone derivatives reported in the literature, the available compounds represent a collection of relatively narrow and largely disconnected chemical series rather than a systematically explored SAR landscape. Most medicinal chemistry studies focus on a limited number of closely related analogues designed around a specific substitution pattern and evaluated against a selected biological target, cancer cell line, or microbial strain.[35] Consequently, the conclusions derived from an individual series usually describe only a local region of the chemical space and cannot be transferred to structurally distant thiazolidinones or to compounds investigated using different biological endpoints. This fragmentation is also reflected in previous Quantitative Structure-Activity Relationship (QSAR) analyses of the scaffold, which have generally been developed for specific compound classes and individual activities rather than from a unified representation of the broader thiazolidinone family.[20]

The challenge is therefore no longer limited to generating additional derivatives but increasingly concerns the ability to determine which regions of the accessible structural space should be explored. Diversification at several positions of the thiazolidinone core, together with variation of peripheral aromatic, heteroaromatic, aliphatic, and hybrid pharmacophoric fragments, produces a rapidly expanding set of theoretically possible analogues. Even within much more restrictive molecular universes, systematic enumeration produces extraordinarily large numbers of structures; for example, GDB-17 contains 166.4 billion constitutionally distinct molecules composed of no more than 17 heavy atoms of selected elements.[36,37] The scale of such enumerated spaces illustrates why experimental synthesis and testing can cover only a minute and highly non-random fraction of potentially accessible molecular structures.

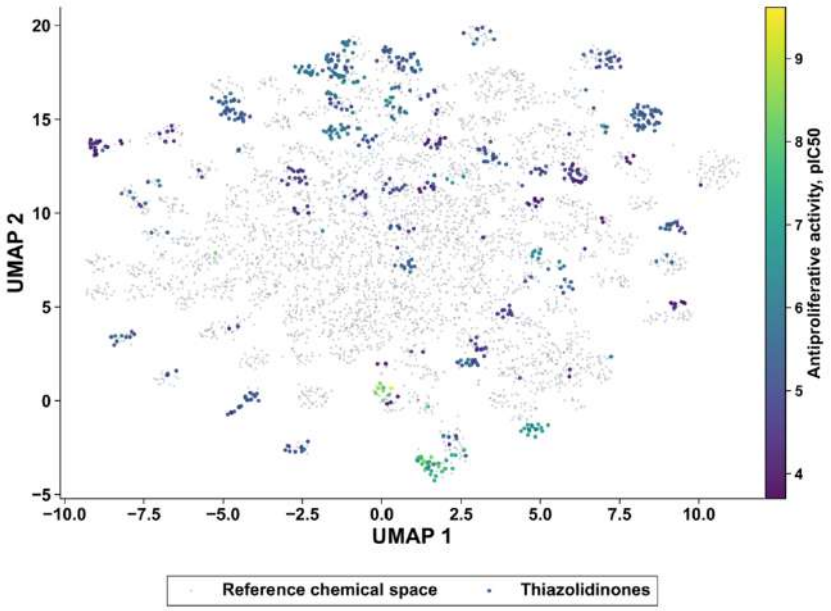


Figure 2. Two-dimensional representation of the chemical space occupied by compounds with experimentally determined antiproliferative activity against the A549 human lung adenocarcinoma cell line. Molecular structures were standardized, canonicalized and represented using ECFP4 fingerprints (Morgan radius = 2, 2048 bits) generated with RDKit. Dimensionality reduction was performed using Uniform Manifold Approximation and Projection (UMAP) to preserve local neighborhood relationships within the high-dimensional fingerprint space. Gray points represent all compounds included in the A549 dataset, whereas thiazolidinone derivatives are highlighted in color, illustrating their distribution within the broader SAR landscape associated with this biological endpoint.

The availability of several diversification sites means that the number of possible analogues increases multiplicatively with the number of admissible substituents. As a simple illustration, considering only 100 synthetically feasible alternatives at each of three independent diversification sites would generate 100^3 combinations before accounting for additional peripheral modifications, positional isomers, stereochemistry or different connection patterns. This calculation demonstrates why exhaustive synthesis and biological testing are impractical even when the range of structural modifications is deliberately restricted. More generally, both systematic enumeration and stochastic sampling studies have shown that chemically meaningful molecular spaces rapidly become too large to explore through experimental synthesis alone.[38]

Traditional optimization of thiazolidinone derivatives has largely relied on iterative design–make–test–analyse cycles, in which new analogues are generated through incremental structural modifications of previously active compounds. This strategy remains fundamental to medicinal chemistry because it integrates experimental SAR with considerations of synthetic accessibility, physicochemical properties and target engagement, enabling the efficient refinement of lead series toward improved biological profiles. However, its inherently local nature means that exploration is typically restricted to the structural neighborhood of known chemotypes. As a result, medicinal chemistry campaigns often generate detailed knowledge of individual analogue series while providing only limited insight into the broader organization of chemical space occupied by the scaffold. Consequently, structurally distant regions with biological potential may remain unsampled, despite representing plausible directions for lead optimization or scaffold diversification. Addressing this limitation requires moving beyond the analysis of isolated analogue series towards a broader understanding of how compounds are distributed across chemical space and how structural variation shapes biological activity at both local and global scales.[39]

8.4. Artificial Intelligence in modern drug discovery

The growing gap between the vastness of chemical space and the limited number of compounds that can be synthesized and experimentally evaluated has positioned artificial intelligence (AI) as one of the principal strategies for improving the efficiency of early-stage drug discovery.[40] AI enables a more rational allocation of experimental effort by identifying promising regions of chemical space, prioritizing compounds for synthesis and predicting biological properties before laboratory validation. As a result, computational modelling has evolved from a supportive tool for post-hoc interpretation into an integral component of prospective molecular design.[41–44]

This transformation has been driven by rapid advances in algorithms together with the growing availability of high-quality bioactivity databases such as ChEMBL and PubChem.[45–47] Modern computational workflows integrate diverse modelling strategies, including classical machine learning QSAR models, transformer-based molecular representations and deep learning models capable of proposing entirely new chemical structures optimized with respect to biological activity, physicochemical properties and synthetic accessibility.[48,49] In contrast to conventional virtual screening, where candidate molecules are selected from predefined libraries, generative AI enables the *de novo* design of compounds that may occupy previously unexplored regions of chemical space while remaining consistent with learned SAR.[50–52]

The success of these approaches depends not only on advances in algorithms but also on the quality of the information extracted from experimental data. AI models are ultimately limited by the knowledge contained within the datasets on which they are trained.[53] Most AI models learn dominant trends across large datasets and therefore perform best in regions where biological activity changes gradually with molecular structure. However, the most informative medicinal chemistry knowledge is often encoded in local regions of the SAR landscape.[54,55] Among these, activity cliffs— pairs of highly similar compounds with unexpectedly large differences in biological activity—identify structural modifications that produce disproportionate changes in potency.[56] As such, they encode critical knowledge on molecular features that drive biological activity and are often overlooked by conventional global SAR analyses.

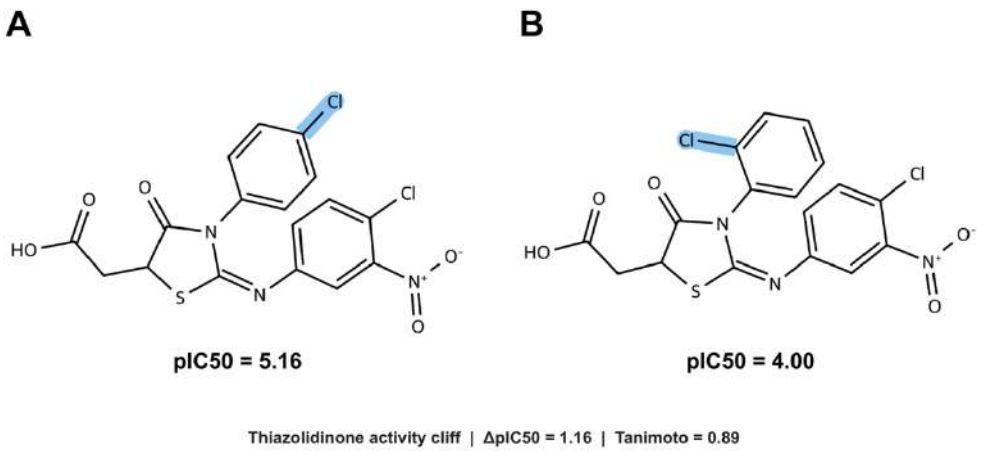


Figure 3. Representative example of an activity cliff identified within the 4-thiazolidinone dataset. Activity cliffs were operationally defined as pairs of compounds exhibiting high structural similarity (Tanimoto coefficient ≥ 0.80) calculated from ECFP4 fingerprints (Morgan radius = 2, 2048 bits, RDKit) together with a large difference in biological activity ($\Delta pIC_{50} \geq 1.0$), corresponding to at least a tenfold difference in potency. Structural similarity was calculated using the Tanimoto coefficient, whereas biological activity was expressed as pIC_{50} values derived from experimentally determined IC_{50} measurements. Only compounds evaluated using the same biological endpoint (A549 human lung adenocarcinoma cell line) were considered to ensure direct comparability of biological activity. The example illustrates how a minor structural modification can produce a pronounced change in biological activity despite high overall molecular similarity.

This observation has important implications for AI-driven molecular design. Activity cliffs consistently expose the limitations of conventional QSAR and remain one of the most demanding benchmarks for property prediction.[54] At the same time, they represent precisely those structural changes that are most valuable for lead optimization. Rather than treating activity cliffs as prediction failures, recent studies increasingly exploit them as sources of medicinal chemistry knowledge. Incorporating activity cliff information into molecular representations, model evaluation and generative workflows allows AI models to focus on structural transformations with the greatest impact on biological activity.[57]

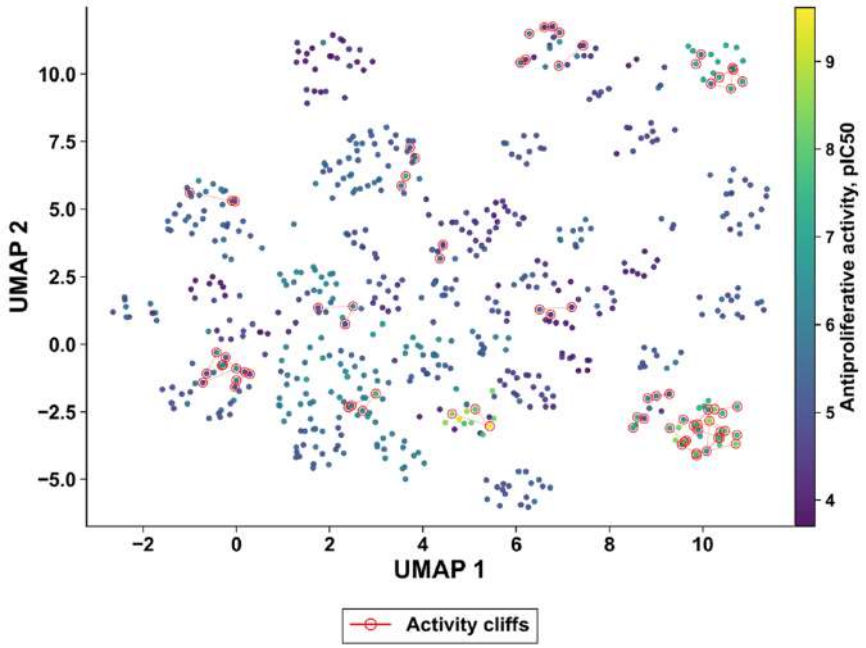


Figure 4. Two-dimensional representation of the chemical space occupied by 4-thiazolidinone derivatives evaluated against the A549 human lung adenocarcinoma cell line. Each point represents one unique compound and is colored according to biological activity expressed as pIC_{50}. Molecular structures were standardized, canonicalized and encoded as ECFP4 fingerprints (Morgan radius = 2, 2048 bits), followed by dimensionality reduction using Uniform Manifold Approximation and Projection (UMAP). Activity cliffs were operationally defined as pairs of compounds with high structural similarity (Tanimoto coefficient ≥ 0.90, calculated using ECFP4 fingerprints) and a difference in biological activity of at least one logarithmic unit ($\Delta pIC_{50} \geq 1.0$), corresponding to a minimum tenfold difference in potency. Red line segments connect compounds forming activity cliffs, highlighting local discontinuities within the structure–activity landscape

This concept is particularly relevant for scaffolds such as thiazolidinones, where multiple diversification sites generate large and only partially explored SAR landscapes. Integrating activity cliff knowledge with generative AI offers an opportunity to move beyond conventional analogue optimization towards *de novo* molecular design guided by experimentally validated local SAR. In this framework, activity cliffs become more than a challenge for predictive modelling; they become design principles that help identify biologically meaningful regions of chemical space and support the generation of novel thiazolidinone derivatives with improved therapeutic potential.

9. Aim of the thesis

The overall aim of this doctoral thesis was to develop and validate an integrated medicinal chemistry framework for the discovery and optimization of biologically active 4-thiazolidinone derivatives by combining classical medicinal chemistry approaches with modern computational methods. Special attention was given to understanding local SAR, exploiting activity cliffs as knowledge-rich molecular events, and applying this information to *de novo* molecular design.

The specific objectives were as follows:

- **Design and synthesis of novel 4-thiazolidinone derivatives** using rational medicinal chemistry approaches and structural optimization of the thiazolidinone scaffold.
- **Biological evaluation of the synthesized compounds**, including the assessment of anticancer and antibacterial activity and the identification of the most promising lead structures.
- **Investigation of the molecular mechanisms underlying biological activity** using computational approaches, including molecular docking, molecular dynamics simulations and SAR analysis.
- **Development and validation of predictive QSAR models** to identify molecular features associated with biological activity.
- **Characterization of local structure–activity discontinuities (activity cliffs)** and evaluation of their impact on molecular property prediction and medicinal chemistry optimization.
- **Development of an activity cliff-informed workflow** for *de novo* molecular design, integrating predictive modelling, molecular generation and knowledge extracted from local SAR to prioritize novel thiazolidinone derivatives with improved biological potential.
- **Evaluation of the applicability of artificial intelligence as a knowledge-driven strategy for exploring regions of thiazolidinone chemical space**, with the aim of improving molecular design beyond conventional analogue-based optimization.

10. Realization of the scientific objectives

The scientific objectives of this doctoral dissertation were realized through four complementary publications published between 2024 and 2026. The studies covered literature analysis, synthesis and biological evaluation of novel 4-thiazolidinone derivatives, investigation of activity cliffs, and development of an activity cliff-informed computational framework for *de novo* molecular design.

P.I focused on a structurally homogeneous group of (2-imino-4-oxo-1,3-thiazolidin-5-yl)acetic acid derivatives. The available literature was analyzed with respect to synthetic approaches, biological activities, molecular targets, and reported SAR. The study provided the scientific background for the subsequent experimental work.

P.II described the design, synthesis, structural characterization, and biological evaluation of fifteen novel 4-thiazolidinone derivatives. Their anticancer activity was assessed *in vitro*, and the most promising compounds were subjected to mechanistic studies and computational analyses.

P.III addressed the role of activity cliffs in modern small-molecule drug discovery. The study reviewed their impact on molecular representation, predictive modelling, model evaluation, and generative design, and proposed a three-layer activity cliff-aware framework integrating these stages.

P.IV translated this concept into a computational workflow for the *de novo* design of 4-thiazolidinone derivatives with potential activity against non-small cell lung cancer.

The experimental studies were conducted primarily at the Chair and Department of Biochemistry, Medical University of Warsaw. The computational component was further developed during a research stay at the Milano Chemometrics and QSAR Research Group, University of Milano-Bicocca, where work focused on QSAR modelling, model validation, and analysis of local SAR.

The research forming the basis of this dissertation was supported by a Researcher (“Młody Badacz”) grant awarded by the Medical University of Warsaw in 2023.

11. Materials and Methods

11.1. Literature analysis of 4-thiazolidinone derivatives

In **P.I**, a structurally defined group of (2-imino-4-oxo-1,3-thiazolidin-5-yl)acetic acid derivatives was systematically analyzed. Compounds reported between 1984 and the first quarter of 2024 were identified using Reaxys, PubMed, and Espacenet. The collected literature was evaluated with respect to synthetic approaches, structural modifications, biological activities, molecular targets, and proposed mechanisms of action.

The analysis was restricted to derivatives sharing the same central scaffold to enable more meaningful comparison of SAR within a chemically coherent group. Regioselectivity and stereochemistry of the investigated derivatives were also considered, as these structural features may influence both compound characterization and the interpretation of biological activity.

11.2. Chemistry, biological evaluation, and molecular modelling

In **P.II**, fifteen novel (2-imino-4-oxo-1,3-thiazolidin-5-yl)acetic acid derivatives were designed and synthesized. The compounds were divided into three structural series according to the heterocyclic substituent introduced into the thiazolidinone scaffold: tryptamine-, 1,2,4-triazole, and thiazole. A common synthetic route based on the reaction of appropriately substituted thiourea derivatives with maleic anhydride was applied, leading to cyclization and formation of the 4-thiazolidinone ring. Acetone was used as the reaction medium to favor regioselective product formation. The obtained derivatives were characterized by ^{1}H and ^{13}C Nuclear Magnetic Resonance (NMR) spectroscopy. Single-crystal X-ray diffraction of compound 2 was additionally performed to determine its three-dimensional structure and verify the regioselectivity and stereochemical outcome of the synthetic procedure.

The biological evaluation followed a stepwise screening strategy. Initial anticancer activity was investigated against human MDA-MB-231 breast adenocarcinoma and SW620 metastatic colorectal carcinoma cells, while immortalized human HaCaT keratinocytes were used as a non-cancerous reference. Cells were exposed to increasing concentrations of the synthesized derivatives for 72 h, and viability was determined using the 3-(4,5-Dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide (MTT) assay. IC_{50} values were calculated from concentration–response data, and compounds that did not reduce cell viability by at least 50% at the highest tested concentration were reported as IC_{50} >100 μM. Cytotoxic selectivity was assessed using the Selectivity Index (SI) calculated from IC_{50} values obtained in HaCaT and cancer cells.

Cell-cycle progression was evaluated in corresponding cell lines after exposure to concentrations corresponding to the individual IC_{50} values. DNA content was assessed by propidium iodide staining followed by flow cytometry, enabling quantification of cells in the principal phases of the cell cycle and identification of changes associated with cell-cycle arrest.

The pro-apoptotic activity was evaluated using Annexin V-FITC and propidium iodide staining followed by flow cytometry. This approach enabled differentiation between viable, early apoptotic, and late apoptotic/necrotic cell populations. Evaluation of the Interleukin-6 (IL-6) secretion was conducted via Enzyme-Linked Immunosorbent Assay (ELISA). Culture supernatants were collected and compared with untreated controls to determine changes in cytokine secretion.

The experimental studies were complemented by structure-based computational analyses aimed at investigating Epidermal Growth Factor Receptor (EGFR) as a potential molecular target. All fifteen synthesized compounds were docked into five crystallographic structures of EGFR: 1M17, 2ITO, 4I23, 4LQM, and 7AEM. The receptor structures were selected to provide diversity in binding-pocket geometry and native-ligand characteristics. Molecular docking was performed using Smina, with docking regions defined from the coordinates of the corresponding native ligands and extended by a 4 Å buffer. Predicted binding scores and ligand poses were compared with the native reference ligands.

Compounds selected based on the combined biological and docking results were further investigated by molecular dynamics simulations. Simulations were performed in OpenMM after receptor preparation and structural repair. The Amber ff14SB force field was applied to proteins and General AMBER Force Field (GAFF) to ligands. Protein–ligand complexes were solvated in explicit Transferable Intermolecular Potential with 3 Points (TIP3P) water with a 10 Å padding and 150 mM NaCl. Following energy minimization, gradual heating to 300 K, restraint release, and Constant Number of Particles, Volume, and Temperature ensemble (NVT) and Constant Number of Particles, Pressure, and Temperature ensemble (NPT) equilibration, production simulations were performed for 100 ns. The resulting trajectories were analyzed at the level of the whole protein, binding pocket, and ligand. Root Mean Square Deviation (RMSD) and Root Mean Square Fluctuation (RMSF) were used to characterize structural stability and molecular flexibility, while Molecular Mechanics/Poisson–Boltzmann Surface Area (MM/PBSA) and Molecular Mechanics/Generalized Born Surface Area (MM/GBSA) calculations were performed to provide an additional energetic comparison of the investigated

complexes. Computational assessment of selected drug-like and Absorption, Distribution, Metabolism, Excretion, and Toxicity (ADMET)-related properties complemented the experimental and molecular modelling analyses.

Statistical analyses of the experimental data were performed in GraphPad Prism. Results were expressed as mean ± standard deviation, and differences between experimental groups were evaluated using one-way Analysis of Variance (ANOVA) followed by appropriate post hoc comparisons, with $p < 0.05$ considered statistically significant.

11.3. Activity cliff-aware approaches in small-molecule drug discovery

In **P.III**, activity cliffs were investigated as local discontinuities in SAR and as potential sources of information for AI-assisted drug discovery. The methodological literature was analyzed across three interconnected stages: molecular representation and embedding, predictive modelling and model evaluation, and *de novo* molecular generation.

Different approaches to molecular similarity and activity cliff identification were compared. Special attention was given to data harmonization, molecular and bioactivity data curation, data-splitting strategies, applicability domains, and activity cliff-specific evaluation.

As an illustrative analysis, compounds with IC_{50} data against the A549 lung adenocarcinoma cell line were retrieved from ChEMBL, standardized, and analyzed using ECFP4-based similarity. Principal Component Analysis (PCA), t-Distributed Stochastic Neighbor Embedding (t-SNE), and structure–activity similarity (SAS) maps were used to visualize local regions in which structurally similar compounds exhibited large differences in potency.

11.4. Activity cliff-informed QSAR and *de novo* molecular design

In **P.IV**, thiazolidinone derivatives with experimentally determined activity against the A549 human lung adenocarcinoma cell line were retrieved from ChEMBL and supplemented with 21 compounds previously synthesized by our research group**.** To improve endpoint consistency, only IC_{50} values obtained using the MTT assay were retained. Measurements were standardized to nM, transformed to pIC_{50}, and subjected to structural curation including duplicate removal, tautomer handling, salt and solvate removal, and exclusion of structurally unsuitable records.

Activity cliffs were operationally defined as molecular pairs with Tanimoto similarity ≥0.80 computed on Extended Connectivity Fingerprints (ECFP) Morgan fingerprint and potency difference of ΔpIC_{50} ≥1.0 (10x -fold difference in IC_{50}). The steepness of local SAR was additionally quantified using the Structure–Activity Landscape Index (SALI). SAS maps, t-SNE, and multidimensional scaling were used to characterize the organization of the thiazolidinone structure–activity landscape.

Two complementary QSAR workflows were developed. In reverse QSAR, compounds were decomposed into chemically meaningful fragments using RDKit Breaking of Retrosynthetically Interesting Chemical Substructures (BRICS), and a binary fragment-incidence matrix was used as the molecular representation. Multiple classical machine-learning algorithms were compared over ten independent train–test splits. Model selection was based primarily on Area Under the Precision–Recall Curve (AUPRC) for the active class and was complemented by Area Under the Receiver Operating Characteristic Curve (AUROC), accuracy, precision, recall, F1-score, and Matthews Correlation Coefficient (MCC). Activity cliff pairs were maintained within the training data to prevent pairwise leakage and enable extraction of local SAR information.

The newly developed Cliff Active Fragment Extraction algorithm (CAFE) compared the fragment composition of the more and less active members of individual activity cliff pairs. Fragment differences were directionally weighted using SALI, allowing fragments associated with favorable potency changes to receive positive enrichment and those associated with unfavorable changes to receive negative enrichment. These contributions were accumulated across activity cliff pairs to generate an activity cliff-informed fragment library.

A separate QSAR benchmark was trained to predict binary label active/non-active on dataset encoded to ECFP with radius 2 at 1024 and 2048 bits, Molecular ACCess System (MACCS),

and 31 physicochemical descriptors selected using Boruta from approximately 1500 calculated features. Models were evaluated over repeated group-aware splits in which activity cliff components could occur in the test set. Selection prioritized predictive performance specifically within steep SAR regions using $RMSE_{cliff}$ (RMSE computed on activity cliff pairs, which reveals models' performance on difficult cases) and its stability across repeated evaluations.

De novo molecular generation was performed using an island-based fragment assembly framework built around distinct thiazolidinone cores. Generated compounds were subjected to chemical validity filters and ranked using predicted activity, Quantitative Estimate of Drug-likeness (QED), and synthetic accessibility. The CAFE LATE extension additionally incorporated positive fragment-level information obtained from CAFE analysis into candidate ranking. CAFE LATE was benchmarked against an otherwise identical baseline QSAR-guided workflow. Final candidate selection incorporated Pareto optimization, molecular diversity, scaffold representation, and structural de-duplication.

Generated molecules were subsequently evaluated by molecular docking against thirteen cancer-relevant protein structures. Native-ligand redocking was used to validate individual receptor models, with RMSD $\leq$2.0 Å defining successful reproduction of the crystallographic pose. Compounds achieving more favorable scores than the respective native ligands were advanced to molecular dynamics simulations and assessed using ligand and pocket RMSD/RMSF together with binding-energy calculations. These parameters were integrated into a multi-metric candidate prioritization strategy.

12. Results

12.1. 4-thiazolidinone chemical space

Literature analysis revealed broad structural diversity of 4-thiazolidinones, particularly at the substituents attached to the imino- and N-substituted regions of the scaffold. Heterocyclic substituents included triazole, benzothiazole, imidazole, pyrazole and coumarin systems, whereas simpler methyl, ethyl, allyl, cyclohexyl, and phenyl groups were frequent at other substitution sites.

Despite this structural diversity, most individual studies contained only a limited number of compounds sharing the investigated core, and only a few derivatives from individual series were commonly advanced to additional biological evaluation. No structural motif consistently associated with increased biological potency across the analyzed derivatives could therefore be identified. Derivatives containing the acetic acid substituent at C5 were also relatively rarely selected as lead compounds for further characterization.

Stereochemistry represented an additional unresolved characteristic of the available chemical space. The majority of reported derivatives were not evaluated as stereochemically defined compounds despite the presence of a stereogenic center at C5. Consequently, many biological results available for this structural class were obtained for racemic mixtures or compounds without explicitly defined stereochemistry.

12.2. Novel 4-thiazolidinone derivatives

Fifteen novel (2-imino-4-oxo-1,3-thiazolidin-5-yl)acetic acid derivatives were synthesized. Preliminary antibacterial screening did not reveal sufficient activity to support further optimization, and subsequent experimental work therefore focused on their antiproliferative potential.

The anticancer screening identified marked differences between individual compounds. Most derivatives showed IC_{50} values of approximately 100 μM or higher in both cancer cell lines. Compounds 5, 6, and 7 showed the most favorable combination of cancer cell cytotoxicity and selectivity toward HaCaT cells.

Compound 5 was the most active derivative against MDA-MB-231 cells, with an IC_{50} of 21.16 ± 0.36 μM and a SI of 4.73. Compound 7 showed activity against both cancer models, with IC_{50}

values of 36.82 ± 2.78 μM in MDA-MB-231 and 24.10 ± 3.98 μM in SW620 cells. The corresponding SIs were 2.71 and 4.16, respectively, with an IC_{50} >100 μM in HaCaT cells. Compound 6 was preferentially active against SW620 cells, with an IC_{50} of 41.31 ± 1.98 μM and a SI of 2.42, while its IC_{50} against MDA-MB-231 and HaCaT cells was ≥100 μM. (**Figure 5.A**)

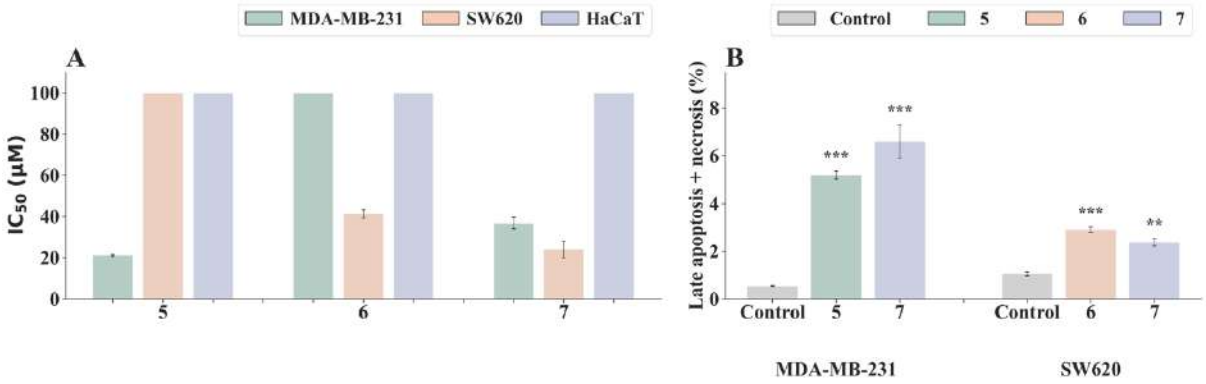


Figure 5. Biological activity of selected 4-thiazolidinone derivatives. (**A**) Cytotoxic activity of compounds 1–15 against MDA-MB-231 breast adenocarcinoma, SW620 metastatic colorectal carcinoma, and non-cancerous HaCaT keratinocyte cells, expressed as IC_{50} values after 72 h exposure. (**B**) Percentage of cells undergoing late apoptosis or necrosis after treatment with selected compounds 5, 6, and 7 at their respective IC_{50} concentrations for 72 h in MDA-MB-231 and SW620 cells. Data are presented as mean±SD. Statistical significance was assessed relative to the corresponding untreated control: $^{**}p \leq 0.001$, $^{***}p \leq 0.0001$.

Compounds 5, 6, and 7 were subsequently subjected to mechanistic cellular analyses. All three altered the distribution of apoptotic cell populations (**Figure 5.B**). Compound 7 produced the largest increase in late apoptosis or necrosis in MDA-MB-231 cells, reaching 6.60±0.69% compared with 0.15±0.029% in the corresponding control group ($p \leq 0.0001$). In SW620 cells, the late apoptotic/necrotic population reached 2.38±0.15% after treatment with compound 7 compared with 1.05±0.07% in the control ($p \leq 0.001$). Compound 6 increased the late apoptotic/necrotic SW620 population to 2.91±0.12% ($p \leq 0.0001$). Compound 5 increased this population in MDA-MB-231 cells to 5.19±0.17% compared with 0.93±0.015% in the corresponding control ($p \leq 0.0001$).

Changes were also observed in cell-cycle distribution after 24 h exposure. In SW620 cells, compound 7 increased the S- and G2/M-phase populations and was accompanied by a reduction in the G0/G1 population, reported at 48.89% ($p \leq 0.001$). Compound 6 produced a 1.7-fold increase in the S-phase population ($p \leq 0.05$) and a 2.77-fold increase in the sub-G1 fraction

compared with untreated cells. Compound-dependent alterations were also observed in MDA-MB-231 cells, with compounds 5 and 7 affecting progression through different phases of the cell cycle.

IL-6 secretion differed after treatment with the selected derivatives. In MDA-MB-231 cells, compound 5 reduced IL-6 secretion 4.1-fold relative to untreated cells ($p \leq 0.001$), corresponding to a 76% reduction. Following treatment with compound 7, IL-6 concentration was 6.63 ± 1.76 pg/mL compared with 8.67 ± 1.54 pg/mL in untreated MDA-MB-231 cells ($p \leq 0.01$). In SW620 cultures, compounds 6 and 7 reduced IL-6 secretion by 1.38-fold and 2.21-fold, respectively.

Molecular docking of compounds 5-7 against five EGFR structures produced more favorable minimum docking scores than the respective native ligands across the investigated structures. For 1M17, the native ligand score was −6.74 kcal/mol, compared with minimum scores of −8.7, −7.6, and −8.4 kcal/mol for compounds 5, 6, and 7, respectively. For 2ITO, the corresponding values were −7.34 kcal/mol for the native ligand and −8.8, −7.6, and −8.1 kcal/mol for compounds 5-7. For 4I23, the native score was −7.83 kcal/mol compared with −8.4, −8.6, and −8.7 kcal/mol. For 4LQM, the values were −1.51 kcal/mol for the native ligand and −8.7, −8.4, and −8.4 kcal/mol for compounds 5-7. The most favorable docking scores were obtained against 7AEM: −9.4, −8.7, and −9.2 kcal/mol for compounds 5, 6, and 7, respectively, compared with −7.71 kcal/mol for the native ligand.

Four 100 ns molecular dynamics simulations were subsequently performed for 7AEM: the native ligand and compounds 5, 6, and 7. Mean ligand RMSD values were 1.29 ± 0.27 Å for the native ligand, 1.22 ± 0.28 Å for compound 5, 0.73 ± 0.18 Å for compound 6, and 0.74 ± 0.31 Å for compound 7. Mean protein RMSD values were 8.41 ± 3.14 Å, 5.45 ± 1.30 Å, 6.47 ± 1.70 Å, and 6.13 ± 2.06 Å, respectively. Pocket RMSD values were 0.94 ± 0.15 Å for the native complex, 1.72 ± 0.24 Å for compound 5, 1.32 ± 0.24 Å for compound 6, and 0.70 ± 0.11 Å for compound 7.

12.3. Activity cliff-aware strategies for improving molecular modelling

Activity cliffs were identified as information-rich local SAR events that can be operationalized during molecular modelling. The review (**P.III**) integrated current approaches into a three-stage, closed-loop framework covering predictive modelling, *de novo* design, synthesis and experimental validation, and a return of obtained real-world data into a model, starting a next

research interval. Results highlighted strategies that specifically improve predictive and generative performance in steep SAR regions.

The accompanying A549 case study identified 574 activity cliff pairs among 32,598 standardized compounds using ECFP4 similarity ≥0.90 and $\Delta pIC_{50} \geq 1.0$. PCA, t-SNE, and SAS analyses showed that these discontinuities were embedded within structurally coherent regions rather than isolated from the surrounding chemical space.

This analysis provided the conceptual basis for the **P.IV** study, where activity cliff information was incorporated directly into QSAR-guided molecular generation.

12.4. *De novo* generation of the new thiazolidinone derivatives

In **P.IV**, the final curated thiazolidinone dataset contained 356 compounds. Of these, 117 compounds were classified as active (32.9%) and 239 as inactive (67.1%).

Pairwise analysis identified 47 activity cliff pairs involving 54 unique molecules. Thus, 15.2% of all compounds participated in at least one activity cliff. Among high-similarity molecular pairs, 349 pairs had Tanimoto similarity ≥0.80, of which 13.5% fulfilled the activity cliff criterion. At Tanimoto ≥0.85, 119 pairs were identified, of which 10.1% were activity cliffs. At Tanimoto ≥0.90, 18 pairs remained, of which 5.6% were classified as activity cliffs.

SALI analysis identified 12 pairs with SALI ≥10. Seven of these simultaneously fulfilled the predefined activity cliff criterion of Tanimoto ≥0.80 and $\Delta pIC_{50} \geq 1.0$. The remaining five represented highly similar pairs with steep SALI values but potency differences below the binary ΔpIC_{50} threshold.

Nine machine learning architectures were compared in the reverse-QSAR stage. Logistic regression achieved the highest mean AUPRC for the active class, reaching 0.889 (95% CI, 0.856–0.917), and was selected for subsequent fragment analysis. Eight compounds were misclassified: four false negatives and four false positives (**Figure 6. A and B**).

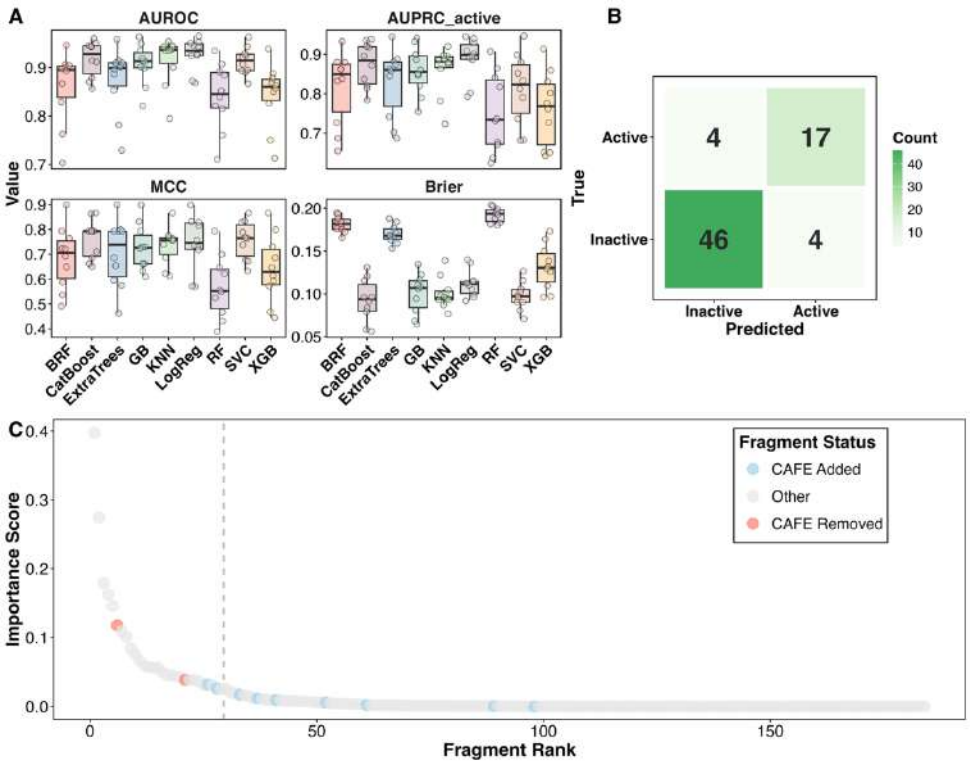


Figure 6. The scatter plot with error bars of reverse QSAR performance trained on nine ML engines (**A**). The confusion matrix of the best reverse QSAR model- logistic regression (**B**). The fragments ranking by the models' importance score (**C**).

Prediction-QSAR modelling showed consistently higher prediction errors for compounds participating in activity cliffs than for the complete test sets. Across most evaluated configurations and repeated splits, Expected Calibration Error (ECE) restricted to activity cliff compounds was also higher than the corresponding global ECE.

CAFE analysis identified activity-driving fragments that were missed by the baseline model, while also filtering out fragments that had been falsely identified as associated with activity (**Figure 6.C**). By aggregating directional, SALI-weighted structural differences across all activity cliff pairs, CAFE generated a refined fragment library enriched for locally relevant SAR information. This fragment-level knowledge was subsequently transferred to the generative stage through CAFE LATE, which applied a post hoc correction to candidate prioritization based on activity cliff-derived evidence.

Six final top 100 candidate sets were produced: one from the baseline workflow (classical QSAR models without CAFE/CAFE LATE interference) and five from CAFE LATE runs using different activity cliff weights. Increasing the CAFE contribution progressively changed

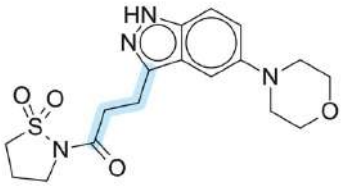
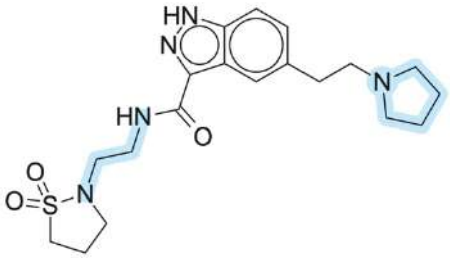
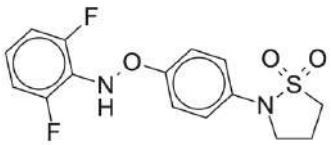
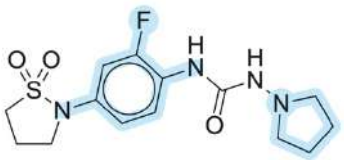
the composition of the highest-ranked molecular sets. At stronger CAFE LATE settings, overlap with the baseline top-ranked molecules decreased to 5%, then 1%, and finally 0%. Across all runs, 282 unique generated compounds were retained for structure-based evaluation.

All thirteen receptor models passed native-ligand redocking with RMSD ≤2.0 Å. In head-to-head docking comparisons, CAFE LATE candidates generally achieved more favorable docking scores than the baseline model, with the strongest improvements observed among the highest-ranked compounds. For selected targets, the mean improvement in the top-ranked pairs reached up to -2.5 kcal/mol.

Across 282 unique molecules and thirteen protein structures, 241 generated protein–ligand complexes achieved docking scores more favorable than the corresponding native reference. Of these better-than-native complexes, 63.9% originated from CAFE LATE generation. The three most favorable docking results within this subset were generated by CAFE LATE candidates, with minimum predicted binding energy versus native scores of −11.0 versus −9.7 kcal/mol, −10.9 versus −8.2 kcal/mol, and −10.8 versus −8.8 kcal/mol. Five generated molecules produced better-than-native docking scores against more than one protein target; four of these five originated from the CAFE LATE workflow. Molecular dynamics (MD) simulations were then performed for the better-than-native complexes. For the three targets represented by both generation strategies, 91 baseline-generated and 148 CAFE LATE-generated ligands were analyzed. The largest dataset was obtained for 1M17 and comprised 203 systems: 73 baseline compounds, 129 CAFE LATE compounds, and the native ligand. Across the complete comparison, the mean number of simultaneously improved MD metrics was 2.20 for CAFE LATE and 2.07 for the baseline workflow.

The final multi-metric ranking identified five ligands that outperformed the corresponding native reference across all five evaluated parameters: pocket RMSD, pocket RMSF, ligand RMSD, ligand RMSF, and binding energy. Four of these five compounds originated from CAFE LATE. Compound 239 ranked first, with a calculated binding energy of −7.78 kcal/mol compared with −4.09 kcal/mol for the native ligand. Compound 343 reached −7.70 kcal/mol. The remaining top-ranked compounds were molecule 86 from the baseline workflow at −7.31 kcal/mol, molecule 265 from CAFE LATE at −7.12 kcal/mol, and molecule 271 from CAFE LATE at −6.84 kcal/mol. Their corresponding QED/SA scores were 0.85/0.77, 0.72/0.81, 0.87/0.81, 0.80/0.78, and 0.87/0.82, respectively (**Table 1**).

Table 1. Top-performing ligands across molecular dynamics benchmark identified against 1M17 protein target. Fragments extracted by activity cliff-aware method (CAFE) are highlighted in color blue. All molecules show 5/5 improved metrics compared to native ligand results.

Ligand ID	Ligand Structure	Method	Binding Energy [kcal/mol]	QED | SA
239		CAFE LATE	-7.78	0.85 | 0.77
343		CAFE LATE	-7.70	0.72 | 0.81
86		Baseline	-7.31	0.87 | 0.81
265		CAFE LATE	-7.12	0.80 | 0.78
271		CAFE LATE	-6.84	0.87 | 0.82

13. Discussion

The four studies forming this dissertation were planned as successive elements of an iterative medicinal chemistry workflow. In **P.III**, this process was formalized as a three-layer, self-improving loop in which activity predictions inform generation, generated candidates return to synthesis and biological testing, and each new experimental round updates the knowledge available to subsequent design. The resulting cycle — *represent, evaluate, generate, test, represent* — places experimental evidence at both the beginning and the end of computational design. Prospective generative studies increasingly follow the same principle, with computationally designed compounds progressing to synthesis and experimental characterization before the next design iteration.[57–59]

The present thesis developed this loop around 4-thiazolidinones. **P.I** characterized the chemical space of thiazolidinones and defined the chemical problem, **P.II** supplied experimentally characterized molecules and biological data, **P.III** established the computational strategy suited to this data regime, and **P.IV** used this knowledge to design the next generation of molecules. Together, these studies trace a continuous path from chemical space characterization and experimental SAR to computational learning and prospective molecular design.

13.1. Fragmented thiazolidinones SAR defined the problem

The starting point emerged from **P.I**. Although 4-thiazolidinones have been explored across many biological applications, the structurally homogeneous subclass examined in the review was represented by relatively small series, uneven substitution patterns and limited stereochemical characterization. This produced a sparse local SAR landscape in which individual active compounds were easier to identify than transferable structural rules. The broader thiazolidinone literature shows the same tendency toward numerous scaffold variants developed in relatively focused chemical series, with biological activity strongly dependent on peripheral substitution.[60]

This setting is particularly demanding for rational molecular design and for data-driven modelling. Small medicinal chemistry datasets provide fewer labelled examples, while strong local activity contrasts make individual structural transitions disproportionately informative. Chemical language models can operate in low-data regimes, but their performance depends strongly on how effectively prior chemical knowledge and local task-specific information are transferred into the model.[61,62] Recent low-data benchmarks further show that predictive

performance becomes strongly dependent on the amount, quality and chemical context of the available data.[63] Activity cliffs intensify this problem because conventional models tend to smooth local potency discontinuities even when global prediction remains acceptable. **P.III** explicitly identified this behavior as a central limitation of classical QSAR modelling.[54]

The computational framework developed later in the thesis therefore emerged directly from the chemical and biological constraints identified in **P.I**. A small, chemically focused dataset with pronounced local SAR contrasts required a modelling strategy able to retain information carried by close molecular neighbors. **P.II** supplied this framework with a controlled experimental foundation: fifteen new derivatives expanded a defined region of thiazolidinone space and generated internally consistent biological data under common assay conditions. These data provided the experimental basis for a data-efficient strategy focused on the steepest regions of thiazolidinone SAR.

Anticancer activity provided the strongest experimental signal and supported deeper biological characterization. Preliminary antibacterial screening yielded insufficient activity for continued optimization, directing subsequent experimental development toward the anticancer phenotype. This early biological triage reduced the search space and concentrated future modelling on the endpoint supported by the strongest experimental signal.

13.2. Structural certainty is essential

A second issue identified in **P.I** concerned structural assignment. Reactions leading to substituted 4-thiazolidinones can generate regioisomeric or stereochemically distinct products, yet the literature often provides limited information on the structural origin of the isolated form.[64] For medicinal chemistry, such ambiguity directly weakens SAR interpretation because biological activity can only be assigned confidently when the tested molecular structure is defined.[65]

This issue was addressed experimentally in **P. II.** Considerable effort was therefore devoted to establishing the structure of the newly synthesized series, including crystallization of compound 2 and X-ray analysis. Single-crystal diffraction remains one of the strongest methods for resolving ambiguous regio- and stereochemical assignments, and it is routinely used when spectroscopic data alone allow more than one structural interpretation.[66,67] The crystal structure supported the assigned regioisomer and provided direct evidence for the regioselective course of the cyclization used for the series. Resolving this ambiguity before biological interpretation

was particularly important in a scaffold where subtle structural differences later proved sufficient to produce large changes in activity.

13.3. Experiments shaped future modelling

Experimental data generated in **P.II** showed that close structural similarity did not translate into continuous biological responses. Differences in cytotoxicity were accompanied by compound-specific effects on apoptosis, cell-cycle distribution and IL-6 secretion, producing a multidimensional phenotype rather than a single activity ranking. The study therefore yielded chemically defined and biologically informative compounds, although their anticancer potency still offered considerable scope for optimization. Compounds 5 and 7 represented the strongest profiles within the synthesized series, but the magnitude of their activity did not justify continuing analogue expansion by classical approaches alone.[68]

This result became a turning point in the design strategy. The experimental series showed that relatively small structural changes within the same thiazolidinone framework could produce pronounced differences in biological response. The same pattern was evident in the broader chemical space characterized in **P.I** and was subsequently confirmed in publicly available bioactivity data. In the A549 dataset analyzed in **P.III**, 574 activity cliff pairs were identified among 32,598 standardized compounds, forming coherent high-gradient regions rather than isolated outliers. Independent benchmarking has shown that such regions remain particularly difficult for molecular machine learning models: activity cliff compounds are predicted less reliably than the surrounding chemical space, and conventional QSAR performance can remain strong globally while failing around close analogues with big potency differences.[53,69] Large-scale molecular property benchmarks similarly show that activity cliffs expose prediction failures that remain obscured by aggregate test-set performance.[70]

The thiazolidinone-specific dataset used in **P.IV** reproduced this behavior. 47 activity cliff pairs involving 54 unique molecules were identified, corresponding to 15.2% of the curated dataset, and twelve pairs reached SALI values ≥10, defining compact regions of particularly disrupted SAR signal.

This combination of limited data volume, close structural neighborhoods and strong local activity contrasts created a specific modelling requirement. Data efficiency became more important than model complexity, and preservation of local SAR became more important than optimization of global prediction alone. This is consistent with studies showing that generative

and predictive molecular models can operate effectively in low-data regimes when chemical and biological information are incorporated in a task-aware manner.[71,72] More recent work also indicates that chemical diversity and information content can matter as much as raw dataset size for downstream molecular learning.[73]

13.4. Turning limitations into advantage

P.IV provided the practical response to these constraints by converting local SAR discontinuities into a design signal. Two newly developed algorithms, CAFE and CAFE LATE, were introduced to extract activity cliff-informed fragment knowledge and propagate it into the prioritization of newly generated molecules. This strategy is consistent with the emerging view that molecular generation benefits from task-specific constraints that preserve biologically relevant information during optimization.[59,74] Such constraints are particularly important in multi-objective generative design, where unconstrained optimization can exploit weaknesses of predictive models and produce apparently optimal but poorly supported candidates.[75]

The advantage became evident in the final candidate selection. The five highest-ranked complexes improved simultaneously over the native reference across all five MD criteria, and four of these five molecules were generated by CAFE LATE. The two leading activity cliff-informed candidates, 239 and 343, reached calculated binding energies of −7.78 and −7.70 kcal/mol against 1M17, while preserving improvement across the complete stability benchmark. At the population level, CAFE LATE also shifted ligand and pocket stability toward more favorable values relative to the baseline model, although the energetic advantage remained target dependent. Concordance across structural stability and energetic readouts provided a more stringent basis for candidate prioritization than optimization against a single computational score. Physics-guided generative frameworks increasingly adopt this strategy by coupling generation with molecular dynamics or free-energy calculations before experimental prioritization.[59] Related active-learning workflows likewise use computationally expensive physics-based calculations selectively, concentrating them on molecules carrying the highest expected design value.[76]

These findings place the present results within a broader challenge in AI-driven drug design: moving beyond computational optimization toward the generation of candidates with genuine experimental potential. A central question is therefore whether generative models merely produce chemically plausible structures or meaningfully enrich the subset worth advancing

experimentally. Prospective studies indicate that achieving this requires progressive integration of biological relevance, structural plausibility, and synthetic feasibility across the design cycle.[77] Recent molecular-design studies further demonstrate that learned SAR can be transferred into the generation of analogues with improved target activity, providing an experimental benchmark for this concept.[78] In this context, the value of CAFE LATE lies in carrying experimentally anchored local SAR into lead prioritization, so that information concentrated in the steepest regions of the activity landscape shapes the next molecular generation.

Across the four studies, SAR discontinuity therefore changed role: it emerged first as an experimental constraint, then as a modelling challenge, and ultimately as information that could be exploited for molecular design. This progression closes the computational part of the loop established at the beginning of the thesis and defines the generated candidates as the starting point for its next experimental iteration.

13.5. Study limitations

Several limitations should be considered when interpreting the present findings. The antibacterial part of the study was discontinued after preliminary screening showed insufficient activity to justify further optimization, and the biological development of the series therefore focused predominantly on anticancer effects.

The experimental evaluation was performed in a limited panel of *in vitro* models, and the molecular mechanisms proposed for the active compounds were supported mainly by biological assays and computational analyses. Direct biochemical confirmation of the predicted molecular targets would provide an additional level of mechanistic validation.

The computational framework was developed within a focused thiazolidinone chemical space and for the A549 activity endpoint. Its broader applicability to other scaffolds, biological systems and endpoints remains to be established.

Finally, the compounds generated by CAFE LATE have so far been evaluated computationally. Their prospective synthesis and biological testing will be required to determine whether the observed enrichment translates into improved experimental activity and to complete the *design–make–test–design* cycle.

14. Abbreviations

ADMET — Absorption, Distribution, Metabolism, Excretion, and Toxicity

AI — Artificial Intelligence

ANOVA — Analysis of Variance

AUPRC — Area Under the Precision–Recall Curve

AUROC — Area Under the Receiver Operating Characteristic Curve

BRICS — Breaking of Retrosynthetically Interesting Chemical Substructures

CAFE — Cliff Active Fragment Extraction

ECE — Expected Calibration Error

ECFP — Extended Connectivity Fingerprints

EGFR — Epidermal Growth Factor Receptor

ELISA — Enzyme-Linked Immunosorbent Assay

GAFF — General AMBER Force Field

IL-6 — Interleukin-6

MACCS — Molecular ACCess System

MCC — Matthews Correlation Coefficient

MD — Molecular Dynamics

MIC — Minimal Inhibitory Concentration

MM/GBSA — Molecular Mechanics/Generalized Born Surface Area

MM/PBSA — Molecular Mechanics/Poisson–Boltzmann Surface Area

MTT — 3-(4,5-Dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide

NMR — Nuclear Magnetic Resonance

NPT — Constant Number of Particles, Pressure, and Temperature ensemble

NVT — Constant Number of Particles, Volume, and Temperature ensemble

PCA — Principal Component Analysis

QED — Quantitative Estimate of Drug-likeness

QSAR — Quantitative Structure–Activity Relationship

RMSD — Root Mean Square Deviation

RMSF — Root Mean Square Fluctuation

SALI — Structure–Activity Landscape Index

SAR — Structure–Activity Relationship

SAS — Structure–Activity Similarity

SI — Selectivity Index

TIP3P — Transferable Intermolecular Potential with 3 Points

t-SNE — t-Distributed Stochastic Neighbor Embedding

UMAP — Uniform Manifold Approximation and Projection

WHO — World Health Organization

15. References

1. Tautermann, C. S. Current and Future Challenges in Modern Drug Discovery. *Methods in Molecular Biology* **2114**, 1–17 (2020).
2. Khanna, I. Drug discovery in pharmaceutical industry: productivity challenges and trends. *Drug Discov. Today* **17**, 1088–1102 (2012).
3. Kenakin, T. Know your molecule: pharmacological characterization of drug candidates to enhance efficacy and reduce late-stage attrition. *Nature Reviews Drug Discovery 2024 23:8* **23**, 626–644 (2024).
4. Waring, M. J. *et al.* An analysis of the attrition of drug candidates from four major pharmaceutical companies. *Nature Reviews Drug Discovery 2015 14:7* **14**, 475–486 (2015).
5. Paul, S. M. *et al.* How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nature Reviews Drug Discovery 2010 9:3* **9**, 203–214 (2010).
6. Roland, A., Fox, W. & Baker, A. Efficiency, effectiveness and productivity in pharmaceutical R&D. *Nat. Rev. Drug Discov.* **23**, 656–657 (2024).
7. Vasan, N., Baselga, J. & Hyman, D. M. A view on drug resistance in cancer. *Nature 2019 575:7782* **575**, 299–309 (2019).
8. Livermore, D. M. The need for new antibiotics. *Clinical Microbiology and Infection* **10**, 1–9 (2004).
9. Bray, F. *et al.* Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **74**, 229–263 (2024).
10. Global cancer burden growing, amidst mounting need for services. https://www.who.int/news/item/01-02-2024-global-cancer-burden-growing--amidst-mounting-need-for-services.
11. Holohan, C., Van Schaeybroeck, S., Longley, D. B. & Johnston, P. G. Cancer drug resistance: an evolving paradigm. *Nature Reviews Cancer 2013 13:10* **13**, 714–726 (2013).
12. Murray, C. J. *et al.* Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *The Lancet* **399**, 629–655 (2022).
13. Antimicrobial resistance. https://www.who.int/news-room/fact-sheets/detail/antimicrobial-resistance.

14. Brown, D. G. & Boström, J. Analysis of Past and Present Synthetic Methodologies on Medicinal Chemistry: Where Have All the New Reactions Gone?Miniperspective. *J. Med. Chem.* **59**, 4443–4458 (2016).
15. Vitaku, E., Smith, D. T. & Njardarson, J. T. Analysis of the Structural Diversity, Substitution Patterns, and Frequency of Nitrogen Heterocycles among U.S. FDA Approved PharmaceuticalsMiniperspective. *J. Med. Chem.* **57**, 10257–10274 (2014).
16. Karageorgis, G., Warriner, S. & Nelson, A. Efficient discovery of bioactive scaffolds by activity-directed synthesis. *Nature Chemistry 2014 6:10* **6**, 872–876 (2014).
17. Kabir, E. & Uzzaman, M. A review on biological and medicinal impact of heterocyclic compounds. *Results Chem.* **4**, 100606 (2022).
18. Brown, F. C. 4-Thiazolidinones. *Chem. Rev.* **61**, 463–521 (1961).
19. Tripathi, A. C. *et al.* 4-Thiazolidinones: The advances continue…. *Eur. J. Med. Chem.* **72**, 52–77 (2014).
20. Sharma, A. *et al.* Recent advances in synthetic strategies and SAR of thiazolidin-4-one containing molecules in cancer therapeutics. *Cancer and Metastasis Reviews 2023 42:3* **42**, 847–889 (2023).
21. Mech, D., Kurowska, A. & Trotsko, N. The Bioactivity of Thiazolidin-4-Ones: A Short Review of the Most Recent Studies. *International Journal of Molecular Sciences 2021, Vol. 22, Page 11533* **22**, 11533 (2021).
22. Ramirez, M. A. & Borja, N. L. Epalrestat: an aldose reductase inhibitor for the treatment of diabetic neuropathy. *Pharmacotherapy* **28**, 646–655 (2008).
23. Long, N., Le Gresley, A. & Wren, S. P. Thiazolidinediones: An In–Depth Study of Their Synthesis and Application to Medicinal Chemistry in the Treatment of Diabetes Mellitus. *ChemMedChem* **16**, 1717 (2021).
24. Nanjan, M. J., Mohammed, M., Prashantha Kumar, B. R. & Chandrasekar, M. J. N. Thiazolidinediones as antidiabetic agents: A critical review. *Bioorg. Chem.* **77**, 548–567 (2018).
25. Roszczenko, P. *et al.* 4-Thiazolidinone-Bearing Hybrid Molecules in Anticancer Drug Design. *Int. J. Mol. Sci.* **23**, 13135 (2022).
26. Finiuk, N. *et al.* Novel hybrid pyrrolidinedione-thiazolidinones as potential anticancer agents: Synthesis and biological evaluation. *Eur. J. Med. Chem.* **238**, 114422 (2022).
27. Tawfeek, H. N. *et al.* Design, synthesis, crystal structures and biological evaluation of some 1,3-thiazolidin-4-ones as dual CDK2/EGFR potent inhibitors with potential apoptotic antiproliferative effects. *Arabian Journal of Chemistry* **15**, 104280 (2022).

28. Elsisi, W. I., George, R. F., Syam, Y. M., Abd-Ellatef, G. E. F. & Abd El-Karim, S. S. Targeting EGFR with thiazolidin-4-ones: Structure-guided design, synthesis, and in silico profiling for anticancer drug discovery. *Eur. J. Med. Chem.* **300**, 118142 (2025).
29. El-Zend, M. A. *et al.* Design, synthesis, and multi-target anticancer evaluation of 1,3-thiazolodin-4-one analogues against breast cancer: mechanistic insights into estrogen metabolism, inflammation, angiogenesis, and oxidative stress. *RSC Adv.* **16**, 2528–2554 (2026).
30. Gupta, A., Singh, R., Sonar, P. K. & Saraf, S. K. Novel 4-Thiazolidinone Derivatives as Anti-Infective Agents: Synthesis, Characterization, and Antimicrobial Evaluation. *Biochem. Res. Int.* **2016**, 8086762 (2016).
31. Küçükgüzel, Ş. G., Oruç, E. E., Rollas, S., Şahin, F. & Özbek, A. Synthesis, characterisation and biological activity of novel 4-thiazolidinones, 1,3,4-oxadiazoles and some related compounds. *Eur. J. Med. Chem.* **37**, 197–206 (2002).
32. Küçükgüzel, G., Kocatepe, A., De Clercq, E., Şahin, F. & Güllüce, M. Synthesis and biological activity of 4-thiazolidinones, thiosemicarbazides derived from diflunisal hydrazide. *Eur. J. Med. Chem.* **41**, 353–359 (2006).
33. Tratrat, C. *et al.* Thiazolidin-4-Ones as Potential Antimicrobial Agents: Experimental and In Silico Evaluation. *Molecules* **27**, 1930 (2022).
34. Zangade, S. B., Shinde, A., Nalwar, Y. & Vibhute, Y. Microwave Assisted Synthesis of Some Novel Series of 4-Thiazolidinone Derivatives as Potent Antimicrobial Analogs. *Orbital - The Electronic Journal of Chemistry* **8**, (2016).
35. Devinyak, O., Zimenkovsky, B. & Lesyk, R. Biologically active 4-thiazolidinones: a review of QSAR studies and QSAR modeling of antitumor activity. *Curr. Top. Med. Chem.* **12**, 2763–2784 (2012).
36. Ruddigkeit, L., Van Deursen, R., Blum, L. C. & Reymond, J. L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.* **52**, 2864–2875 (2012).
37. Reymond, J. L. The Chemical Space Project. *Acc. Chem. Res.* **48**, 722–730 (2015).
38. Virshup, A. M., Contreras-García, J., Wipf, P., Yang, W. & Beratan, D. N. Stochastic Voyages into Uncharted Chemical Space Produce a Representative Library of All Possible Drug-Like Compounds. *J. Am. Chem. Soc.* **135**, 7296–7303 (2013).
39. Warr, W. A., Nicklaus, M. C., Nicolaou, C. A. & Rarey, M. Exploration of Ultralarge Compound Collections for Drug Discovery. *J. Chem. Inf. Model.* **62**, 2021–2034 (2022).

40. Qureshi, R. *et al.* AI in drug discovery and its clinical relevance. *Heliyon* **9**, e17575 (2023).
41. Walters, W. P. & Murcko, M. Assessing the impact of generative AI on medicinal chemistry. *Nature Biotechnology 2020 38:2* **38**, 143–145 (2020).
42. Bender, A. & Cortes-Ciriano, I. Artificial intelligence in drug discovery: what is realistic, what are illusions? Part 2: a discussion of chemical and biological data. *Drug Discov. Today* **26**, 1040–1052 (2021).
43. Vamathevan, J. *et al.* Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery 2019 18:6* **18**, 463–477 (2019).
44. Schneider, G. Automating drug discovery. *Nat. Rev. Drug Discov.* **17**, 97–113 (2018).
45. Zdrazil, B. *et al.* The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Res.* **52**, D1180–D1192 (2024).
46. PubChem. https://pubchem.ncbi.nlm.nih.gov/.
47. Selvaraj, C., Elakkiya, E., Prabhu, P., Velmurugan, D. & Singh, S. K. Advances in QSAR through artificial intelligence and machine learning methods. *QSAR in Safety Evaluation and Risk Assessment* 101–116 (2024) doi:10.1016/B978-0-443-15339-6.00033-3.
48. Elton, D. C., Boukouvalas, Z., Fuge, M. D. & Chung, P. W. Deep learning for molecular design—a review of the state of the art. *Mol. Syst. Des. Eng.* **4**, 828–849 (2019).
49. Sanchez-Lengeling, B. & Aspuru-Guzik, A. Inverse molecular design using machine learning: Generative models for matter engineering. *Science (1979).* **361**, 360–365 (2018).
50. Tong, X. *et al.* Generative Models for De Novo Drug Design. *J. Med. Chem.* **64**, 14011–14027 (2021).
51. Tang, X. *et al.* A survey of generative AI for de novo drug design: new frontiers in molecule and protein generation. *Brief. Bioinform.* **25**, (2024).
52. Crucitti, D., Pérez Míguez, C., Díaz Arias, J. Á., Fernandez Prada, D. B. & Mosquera Orgueira, A. De novo drug design through artificial intelligence: an introduction. *Frontiers in Hematology* **3**, 1305741 (2024).
53. Maggiora, G. M. On Outliers and Activity CliffsWhy QSAR Often Disappoints. *J. Chem. Inf. Model.* **46**, 1535–1535 (2006).
54. Van Tilborg, D., Alenicheva, A. & Grisoni, F. Exposing the Limitations of Molecular Machine Learning with Activity Cliffs. *J. Chem. Inf. Model.* **62**, 5938–5951 (2022).

55. Stumpfe, D. & Bajorath, J. Exploring Activity Cliffs in Medicinal ChemistryMiniperspective. *J. Med. Chem.* **55**, 2932–2942 (2012).
56. Cruz-Monteagudo, M. *et al.* Activity cliffs in drug discovery: Dr Jekyll or Mr Hyde? *Drug Discov. Today* **19**, 1069–1080 (2014).
57. Szostek, T. & Szulczyk, D. From obstacle to design advantage: activity cliff aware modeling for small-molecule drug discovery. *Drug Discov. Today* **31**, (2026).
58. Atz, K. *et al.* Prospective de novo drug design with deep interactome learning. *Nature Communications 2024 15:1* **15**, 3408- (2024).
59. Filella-Merce, I. *et al.* Optimizing drug design by merging generative AI with a physics-based active learning framework. *Communications Chemistry 2025 8:1* **8**, 238- (2025).
60. Kaminskyy, D., Kryshchyshyn, A. & Lesyk, R. 5-Ene-4-thiazolidinones – An efficient tool in medicinal chemistry. *Eur. J. Med. Chem.* **140**, 542 (2017).
61. Moret, M. *et al.* Leveraging molecular structure and bioactivity with chemical language models for de novo drug design. *Nature Communications 2023 14:1* **14**, 114- (2023).
62. Isigkeit, L. *et al.* Automated design of multi-target ligands by generative deep learning. *Nature Communications 2024 15:1* **15**, 7946- (2024).
63. Eraqi, B. A., Khizbullin, D., Nagaraja, S. S. & Sarathy, S. M. Molecular property prediction in the ultra-low data regime. *Communications Chemistry 2025 8:1* **8**, 201- (2025).
64. Mukhrish, Y. A Comprehensive Review on Thiazolidinones and their 5-ene Derivatives: Synthetic Methodologies and Pharmacological Profiles. *Mini. Rev. Org. Chem.* **23**, 32–72 (2026).
65. Klebe, G. Optical Activity and Biological Effect. *Drug Design* 67–82 (2024) doi:10.1007/978-3-662-68998-1_5.
66. Wang, Y. *et al.* Diversity-Oriented Synthesis as a Strategy for Fragment Evolution against GSK3β. *ACS Med. Chem. Lett.* **7**, 852–856 (2016).
67. Hiesinger, K. & Knapp, S. Accelerate Your Science: Direct-to-Biology Strategies in Medicinal Chemistry. *J. Med. Chem.* https://doi.org/10.1021/ACS.JMEDCHEM.6C00996 (2026) doi:10.1021/ACS.JMEDCHEM.6C00996.
68. Mas, P., Filoche-Rommé, B., Bianciotto, M. & Vuilleumier, R. Back to the future of lead optimization: benchmarking compound prioritization strategies. *Digital Discovery* **5**, 2983–2997 (2026).

69. Dablander, M., Hanser, T., Lambiotte, R. & Morris, G. M. Exploring QSAR Models for Activity-Cliff Prediction. *J. Cheminform.* **15**, (2023).
70. Deng, J. *et al.* A systematic study of key elements underlying molecular property prediction. *Nature Communications 2023 14:1* **14**, 6395- (2023).
71. Naser, M. Z. A review of machine learning with small and limited data. *Journal of Big Data 2026 13:1* **13**, 18- (2026).
72. Moret, M., Friedrich, L., Grisoni, F., Merk, D. & Schneider, G. Generative molecular design in low data regimes. *Nature Machine Intelligence 2020 2:3* **2**, 171–180 (2020).
73. Medina, B., Tibo, A., He, J., Janet, J. P. & Österbacka, N. Diversity Beats Size Scaling for Chemical Language Models. *npj Drug Discovery 2026 3:1* **3**, 26- (2026).
74. Loeffler, H. H., Wan, S., Klähn, M., Bhati, A. P. & Coveney, P. V. Optimal Molecular Design: Generative Active Learning Combining REINVENT with Precise Binding Free Energy Ranking Simulations. *J. Chem. Theory Comput.* **20**, 8308–8328 (2024).
75. Yoshizawa, T. *et al.* A data-driven generative strategy to avoid reward hacking in multi-objective molecular design. *Nature Communications 2025 16:1* **16**, 2409- (2025).
76. Gusev, F., Gutkin, E., Kurnikova, M. G. & Isayev, O. Active Learning Guided Drug Design Lead Optimization Based on Relative Binding Free Energy Modeling. *J. Chem. Inf. Model.* **63**, 583–594 (2023).
77. Bender, A. *et al.* Artificial intelligence in drug discovery — what it is, where we stand and the path forward. *Nature Reviews Drug Discovery 2026* 1–22 (2026) doi:10.1038/s41573-026-01496-2.
78. Hörmann, T. *et al.* Structural optimization of drug molecules with incrementally trained language models. *Nature Communications 2026 17:1* **17**, 3456- (2026).

16. Publication I

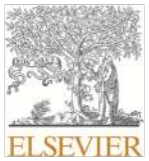
Exploring the bioactive potential of (2-imino-4-oxo-1,3-thiazolidin-5-yl)acetic acid derivatives: A comprehensive review.

Szostek, T., Otto-Ślusarczyk, D., Roszkowski, P., Struga, M., & Szulczyk, D. *Results in Chemistry*, *11*, 101828. (2024).

DOI: https://doi.org/10.1016/J.RECHEM.2024.101828

IF= 4.2, **MNiSW=** 20

Results in Chemistry 11 (2024) 101828



Contents lists available at ScienceDirect

Results in Chemistry

journal homepage: www.sciencedirect.com/journal/results-in-chemistry





Exploring the bioactive potential of (2-imino-4-oxo-1,3-thiazolidin-5-yl) acetic acid derivatives: A comprehensive review

Tomasz Szostek [a], Dagmara Otto-Ślusarczyk [a], Piotr Roszkowski [b], Marta Struga [a], Daniel Szulczyk [a,*]

[a] Chair and Department of Biochemistry, Medical University of Warsaw, 02-097 Warsaw, Poland
[b] Faculty of Chemistry, University of Warsaw, Pasteura 1, 02-093 Warszawa, Poland

ARTICLE INFO

Keywords:
Thiazolidinone
Anticancer
Antimicrobial
Antiprotozoal
Anti-inflammatory

ABSTRACT

Thiazolidinone is a scaffold known for its diverse biological activities, drawing the interest of researchers seeking to explore its potential. Numerous libraries of molecules with diverse structures and properties can be created based on this scaffold. We propose the extraction of compounds with a defined central skeleton, resulting from a repeatable methodology of synthesis, to analyze the biological properties of a heterogeneous set of molecules. This review offers a comprehensive overview of recent research, focusing on a homogeneous group of (2-imino-4-oxo-1,3-thiazolidin-5-yl)acetic acid derivatives. It highlights the most promising compounds that have emerged from recent studies, their molecular targets, and the underlying mechanisms contributing to their biological activity.

1. Introduction

The pursuit of novel therapeutic agents remains a cornerstone of medicinal chemistry, continually driving innovation and the exploration of diverse chemical scaffolds. Among these, thiazolidinone scaffolds have emerged as a versatile and potent heterocyclic core, demonstrating a broad spectrum of biological activities and holding significant promise in drug discovery and development. The broad therapeutic applicability of these compounds is due to their interactions with several biological targets, such as enzymes, receptors, and ion channels, highlighting the importance of understanding their mode of action at the molecular level [1,2].

However, analyzing an overly extensive group of compounds based on a single heterocyclic scaffold can lead to misleading conclusions. Within the analyzed heterogeneous collection of compounds, smaller homogeneous subsets can occasionally be identified (Fig. 1). These subsets are characterized by recurring substituents or specific structural modifications that occur exclusively at particular sites within the core structure.

Analyzing homogeneous groups of compounds offers several scientific advantages that can significantly enhance the quality and interpretability of research findings in medicinal chemistry and drug discovery. The preference for studying homogeneous groups stems from the need for precision, specificity, and clarity in understanding the structure–activity relationships (SAR) and pharmacological profiles of derivatives. Increased accuracy in SAR studies is possible because the physicochemical properties of the derivatives are similar, making it easier to attribute observed biological effects to specific structural modifications. Furthermore, compounds in a homogeneous group are more likely to interact with biological targets in a similar manner, simplifying the process of identifying the mode of action. Analysis in subsets reduces variability in the data, making statistical analysis more straightforward and reliable. This can lead to more robust conclusions and a better understanding of the compound's properties. Homogeneous groups of compounds tend to exhibit similar absorption, distribution, metabolism, excretion, and toxicity (ADMET) profiles. This consistency simplifies the prediction of these crucial properties, which are elementary in evaluating the potential of compounds as drug candidates. Focusing on such groups allows researchers to streamline the synthesis and optimization of compounds. Similar chemical structures often require similar synthetic pathways and conditions, saving time and resources in the drug development process.

In conclusion, analyzing homogeneous groups of compounds provides a focused, accurate, and efficient approach to medicinal chemistry and drug development. It enhances the understanding of compound-specific characteristics, optimizes resource utilization, and ultimately

* Corresponding author.
E-mail address: daniel.szulczyk@wum.edu.pl (D. Szulczyk).

https://doi.org/10.1016/j.rechem.2024.101828
Received 4 June 2024; Accepted 26 September 2024
Available online 27 September 2024
2211-7156/© 2024 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

contributes to the making of more effective therapeutic agents with fewer adverse events. This review provides a comprehensive overview of the current research on a homogeneous group of derivatives of (2-imino-4-oxo-1,3-thiazolidin-5-yl)acetic acid, focusing on the most promising compounds that have emerged from recent studies, their molecular targets, and the underlying mechanisms contributing to their biological activity. Our team analyzed compounds published between 1984 and the first quarter of 2024 using online databases such as Reaxys, PubMed, and Espacenet, based on the structural framework depicted in Fig. 1. The goal was to identify structures with diverse biological activities and to share the most noteworthy findings with the scientific community.

2. Preparation method of (2-imino-4-oxo-1,3-thiazolidin-5-yl) acetic acid derivatives

Numerous methods for synthesizing thiazolidinones have been extensively described in the literature [3]. However, when considering a uniform set of derivatives of (2-imino-4-oxo-1,3-thiazolidin-5-yl)acetic acid, these methods must be narrowed down. The simplest approach is to identify the most commonly occurring substrates in the reactions used to obtain these compounds. It turns out that derivatives of thiourea and thiosemicarbazide have long been used to prepare derivatives of (2-imino-4-oxo-1,3-thiazolidin-5-yl)acetic acid (Scheme 1). The first structures obtained using thiourea derivatives were reported as early as 1894 [4], followed by a series of compounds obtained in 1939 using thiosemicarbazones [5].

These procedures are frequently engaged because of the wide accessibility of substrates and the straightforward, one-step introduction of the selected heterocyclic ring. Maleic anhydride is used in both cases. The reaction can be conducted in both polar and nonpolar solvents. The reaction time depends on the specific reagents and solvent used but does not exceed 24 h.

It is worth noting that for most of the published derivatives presented later in the article, the *E/Z* stereochemistry of the compounds under study was not specified. In the case of symmetrically substituted thioureas or thiosemicarbazones, well-defined products were formed. Adding unsymmetric thiourea derivatives to maleimides or maleic anhydride could lead to two regioisomeric products, depending on the structure of unsymmetrical substrates and the specific solvent. This isomerism occurs due to the existing double bond between the C2 and N6 atoms. Furthermore, the regioselective formation of the designed product may also occur when unsymmetrical substrates are used and a specific solvent is chosen [6]. Based on the suggested mechanism of this transformation, the regioisomers show *E* configuration. Additionally, the obtained products have a stereogenic center at the C5 position, so the compounds exist as a racemic mixture. Individual enantiomers may therefore have different biological activities, and the results obtained for the racemate may not accurately reflect the properties of these compounds. Despite this, the stereochemistry of the molecules was not addressed by the authors in their research in this area.

3. Biological activities of derivatives of (2-imino-4-oxo-1,3-thiazolidin-5-yl)acetic acid

Many derivatives of (2-imino-4-oxo-1,3-thiazolidin-5-yl)acetic acid with diverse biological activities have been obtained. It is important to emphasize that the molecules were designed with an approach that (2-imino-4-oxo-1,3-thiazolidin-5-yl)acetic acid constituted a distinct central motif of the structures. Structural modifications were associated with the introduction of various substituents on the available nitrogen atoms. This versatility has inspired researchers to synthesize various compounds featuring this structural motif. Numerous studies in the scientific literature detail the various biological activities of these derivatives, many of which are discussed in this review.

3.1. Antiprotozoal activity

Toxoplasma gondii, a ubiquitous parasite, is responsible for causing toxoplasmosis in various species of mammals and birds. It's estimated that up to one-third of the global human population carries this parasite. While *Toxoplasma gondii* infections often remain asymptomatic, they can manifest clinically, particularly when affecting the uterus or individuals with compromised immune systems. Early research indicates that these infections may lead to symptoms such as lymphadenopathy, fever, weakness, debilitation, and even ophthalmic complications in immunocompetent individuals. Hence, there is a great requirement to discover safer and more efficient parasite-specific drugs for combating *T. gondii* infections and treating toxoplasmosis [7,8]. Interesting results in this field are presented for structures based on the title synthetic framework.

Trotsko *et al.* obtained three groups of thiazolidinone-derived compounds using thiosemicarbazones as starting materials, of which

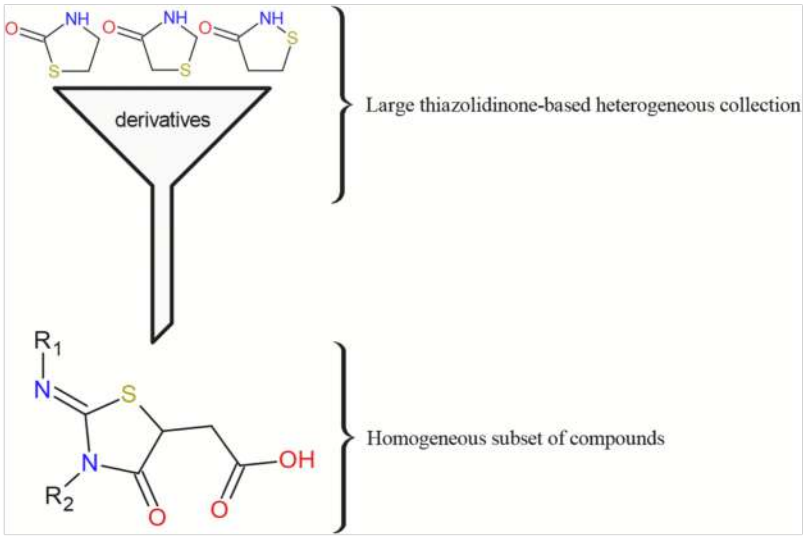
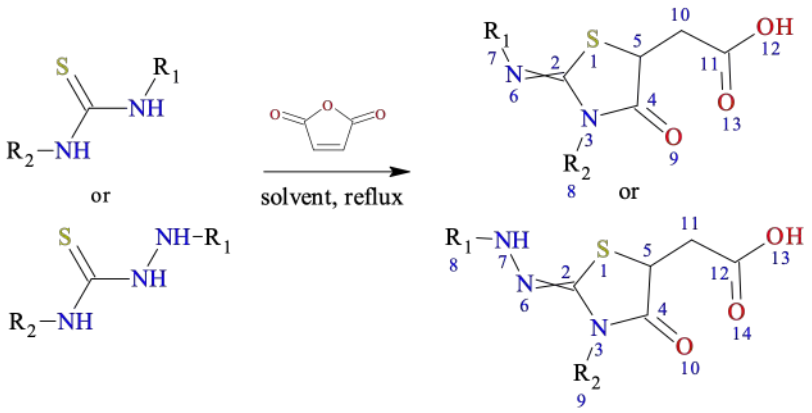


Fig. 1. Subset built on derivatives of (2-imino-4-oxo-1,3-thiazolidin-5-yl)acetic acid.



Scheme 1. A simplified synthesis scheme of derivatives of (2-imino-4-oxo-1,3-thiazolidin-5-yl)acetic acid.

fourteen were built on the (2-imino-4-oxo-1,3-thiazolidin-5-yl)acetic acid scaffold. Cytotoxicity was assessed for all derivatives, after which a group of non-cytotoxic compounds was tested for their ability to inhibit *T. gondii* growth *in vitro*. Six (4-oxothiazolidin-5-yl)acetic acid derivatives showed higher potency when compared to the reference, demonstrating inhibition of parasite proliferation at lower concentrations than sulfadiazine or its synergistic combination with trimethoprim.

It needs to be underlined that two of the investigated derivatives (compounds **1** and **2**, Fig. 2.) were at 90-fold lower IC_{50} values than sulfadiazine and more than 4-fold than sulfadiazine with trimethoprim [9].

Trypanosoma cruzi is a parasite that causes Chagas disease. Current chemotherapy for this disease is disappointing due to limited efficacy, particularly during the chronic phase, and is often linked with frequent side effects that lead to treatment discontinuation. Additionally, some strains of *T. cruzi* have developed resistance to gold-standard drugs such

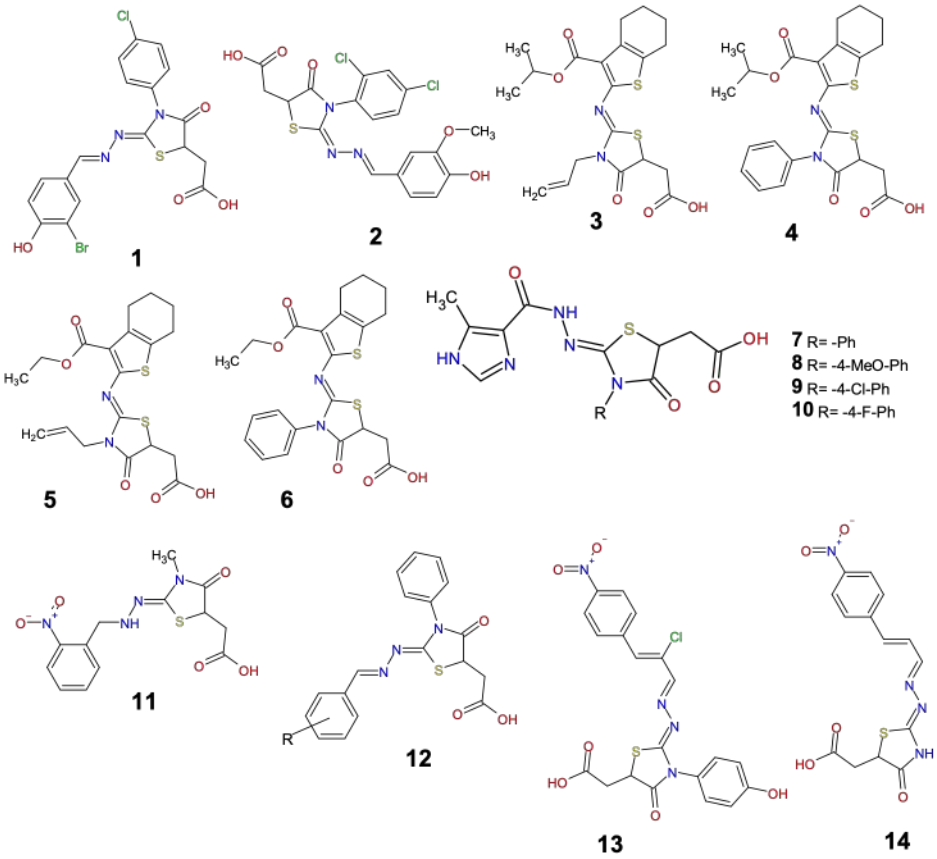


Fig. 2. Example derivatives of (2-imino-4-oxo-1,3-thiazolidin-5-yl)acetic acid possessing antiprotozoal activity.

as nifurtimox and benznidazole [10]. Researchers studied the potential trypanocidal activity against *T. cruzi* through cruzain inhibition by thiophen-2-thiourea and thiophen-2-iminothiazolidine derivatives [11]. Among the broader group, four (4-oxothiazolidin-5-yl)acetic acid derivatives (**3**–**6**, Fig. 2) were evaluated. Only three (**3**, **5**, **6**) were active towards the cruzain enzyme but at a moderate level. The level of cruzain inhibition did not significantly vary across various Triton concentrations, indicating that enzyme inhibition is not attributable to derivative accumulation. Furthermore, docking results indicated that the evaluated three compounds did not yield satisfactory Eb values and exhibited limited interaction with the Gly66 amino acid.

Another series of compounds with antiprotozoal and antimicrobial activity, synthesized from ethyl(5-methyl-1*H*-imidazole-4-carboxylate), was described based on their physical, analytical, and spectral data. The bigger part of the studied derivatives showed excellent anti-*T. gondii* activity compared to hydroxyurea and sulfadiazine. Additionally, it was shown that the majority of the derivatives in this study performed better towards intracellular tachyzoites. However, the tested compounds (**7**–**10**, Fig. 2) towards *T. gondii*-infected Vero cells maintained an effective result only at a 1 mM concentration. The 4-thiazolidinone derivatives were found to be less promising [12].

Tenório *et al.* presented results for two series of derivatives: thiosemicarbazides and corresponding 4-thiazolidinones with common (*ortho*- or *meta*- or *para*-)nitrophenyl structural arrangements (**11**, Fig. 2) [13]. Initial *in vitro* tests showed a marked reduction in the percentage of *T. gondii*-infected Vero cells and the mean number of intracellular parasites. Hydroxyurea and sulfadiazine were used as references. The authors claimed that the cytotoxicity of derivatives was lower than that of the references. The relative number of intracellular parasites in infected host cells after treatment with the compounds differed in the synthesized series, showing less promising results for synthesized ten derivatives with the (2-imino-4-oxo-1,3-thiazolidin-5-yl)acetic acid scaffold.

A similar study and outcomes were presented by another research team. A new series of 2-[(phenylmethylene)hydrazono]-4-oxo-3-phenyl-5-thiazolidineacetic acids (**12**, Fig. 2) were obtained from synthesized substituted benzaldehyde 4-phenyl-3-thiosemicarbazones. Two groups, each with sixteen derivatives, were analyzed. For anti-*Toxoplasma gondii* activity, most derivatives reduced the percentage of infected cells, contributing to parasite elimination. This effect on intracellular parasites also resulted in a lower mean number of tachyzoites. Furthermore, the majority of the 4-thiazolidinones exhibited potent toxicity against intracellular parasites, with IC_{50} values ranging from 0.05 to 1 mM [14]. Hydroxyurea and sulfadiazine were used as references in all studies against parasites. The authors did not identify leading structures but emphasized the need for structural modifications to achieve more definitive results.

Over a hundred compounds were investigated, demonstrating excellent antitrypanosomal activity against *Trypanosoma brucei brucei* and *Trypanosoma brucei gambiense*. The study identified compounds with submicromolar IC_{50} levels, high selectivity indexes, relatively low cytotoxicity in human primary fibroblasts, and low acute toxicity. [15]. Two compounds (**13**–**14**, Fig. 2) possessing the (2-imino-4-oxo-1,3-thiazolidin-5-yl)acetic acid scaffold were included in the study. Both derivatives did not reach high inhibition rates (1 mg/mL concentration) in the initial antitrypanosomal activity study towards *Trypanosoma brucei brucei* and were not included in further tests.

3.2. Anti-inflammatory activity

Inflammation is a complex and multifaceted biological response of body tissues to harmful stimuli, such as pathogens, damaged cells, or irritants. This process is a fundamental aspect of the body's defense mechanism, playing a critical role in healing. However, when inflammation becomes chronic, it is correlated with a large number of pathophysiological disorders, involving autoimmune diseases, heart disease, and various forms of cancer. Upon the occurrence of an inflammatory stimulus, the body's immune cells, particularly macrophages, are activated. These macrophages are pivotal in the immune response as they release a variety of pro-inflammatory molecules that serve to eliminate harmful stimuli and initiate healing. Among these molecules, the short-lived free radical nitric oxide (NO) is significant. Nitric oxide plays a crucial role in the inflammatory process due to its ability to mediate vasodilation, thereby increasing blood flow to the affected area. This increase in blood flow helps deliver essential nutrients and immune cells to the site of inflammation, facilitating the removal of pathogens and the repair of damaged tissue [16]. However, the production of nitric oxide and other pro-inflammatory mediators needs to be tightly regulated. Excessive or prolonged release of these molecules can lead to tissue damage and contribute to the development of chronic inflammatory conditions. In such scenarios, inflammation no longer serves as a protective response but instead perpetuates tissue injury and disease progression. Moreover, the inflammatory process is induced by several other origins, involving inherited tendencies, lifestyle preferences such as diet and exercise, and environmental exposures. For example, certain foods can exacerbate or mitigate inflammation, and regular physical activity has been shown to have an anti-inflammatory effect.

Understanding the complex mechanisms underlying inflammation and its causes is crucial for developing effective treatments for a wide array of inflammatory conditions. By modulating the inflammatory response, it may be possible to alleviate indicators, stop disease progress, and improve the quality of life for individuals suffering from chronic inflammatory diseases. In the search for new molecules with anti-inflammatory properties, researchers have turned their attention to structures based on the (2-imino-4-oxo-1,3-thiazolidin-5-yl)acetic acid scaffold.

Twenty-four newly synthesized derivatives of 1,2,3,4-tetrahydronaphthalen-6-yl-thiazole were assessed for their anti-inflammatory and analgesic activity (Haiba *et al.*) [17]. A group of 4-thiazolidinone-acetic acid derivatives was also included in the biological assessment. Anti-inflammatory activity was established by applying a carrageenan-induced paw edema bioassay in rats, using indomethacin as a reference standard. It was observed that some of the evaluated compounds, including one of the 4-thiazolidinone-acetic acid derivatives (**15**, Fig. 3.), demonstrated early action, showing percentages of edema inhibition higher than or equal to the reference.

This compound was found to have noteworthy edema inhibition at the first-hour post-administration (23.25 %). In this case, equipotency to the reference drug indomethacin was declared.

In a comparative study, the effects of the tested compounds on carrageenan-induced rat paw edema, their percentage of protection, activity in relation to indomethacin, and ulcerogenic effects were evaluated. The authors synthesized 3-acetylcoumarin thiosemicarbazones and then their cyclic analogues [18]. Most of the tested derivatives were more active than the reference compound in the first study. However, the introduction of a carboxylic acid moiety (**16**–**17**, Fig. 3.) led to less anti-inflammatory activity. Both compounds were not transferred for further testing.

A similar study was conducted for a new series of benzothiazole-thiazolidinone hybrids substituted with halogen, methyl, or methoxy groups on the aromatic rings (**18**, Fig. 3). The 4-thiazolidinone-acetic acid derivatives were obtained through the condensation reaction between diverse thioureas and maleic anhydride. In the carrageenan-induced rat paw edema model, the percentage decrease in inflammation in the test compound was measured three hours after carrageenan administration. This decrease was then compared to the reduction observed in animals treated with carrageenan and the reference standard diclofenac sodium. The evaluation revealed that certain compounds exhibited moderate anti-inflammatory activity [19].

Lesyk *et al.* prepared and studied the anti-inflammatory activity of novel 3-(2,3-dimethyl-1-phenyl-4-pyrazolon-5-yl)-4-thiazolidones [20]. Again, the paw edema inhibition test was used to evaluate the anti-inflammatory properties of selected derivatives. Paw volume levels

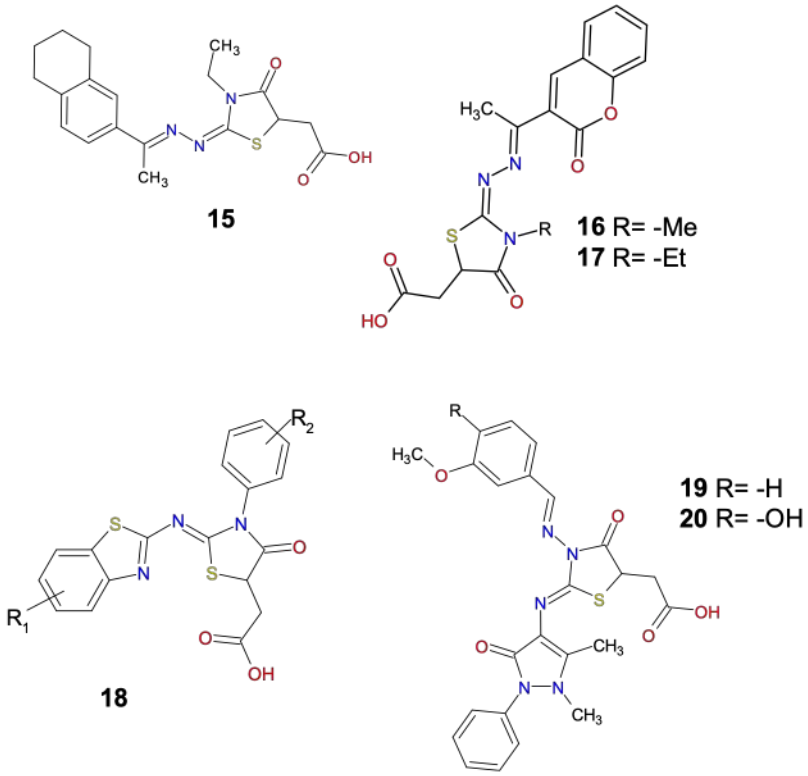


Fig. 3. Example derivatives of (2-imino-4-oxo-1,3-thiazolidin-5-yl)acetic acid possessing anti-inflammatory activity.

were measured, and percentages of inhibition were calculated at the beginning of the test, then after 3 h and 5 h following the formalin injection. Only two (**19**–**20**, Fig. 3.) derivatives reached the levels of the reference drugs. In this case, the two most potent compounds held the (2-imino-4-oxo-1–3-thiazolidin-5-yl)acetic acid central core.

3.3. Anticonvulsant activity

Effective management of seizures is attainable, but often at the cost of substantial toxic side effects. These adverse effects can vary widely in severity, ranging from mild cognitive impairment to life-threatening conditions such as aplastic anemia or hepatic failure [21]. This situation underscores a critical demand for the development of new anticonvulsant agents that not only offer enhanced efficacy in seizure control but also exhibit a markedly improved safety profile. This pursuit is driven by the need to fill the therapeutic gap left by current treatments, providing effective seizure management without compromising patient safety and well-being.

Several studies have been conducted to find new anticonvulsant drug candidates. Ragab *et al.* provided research results on the anticonvulsant

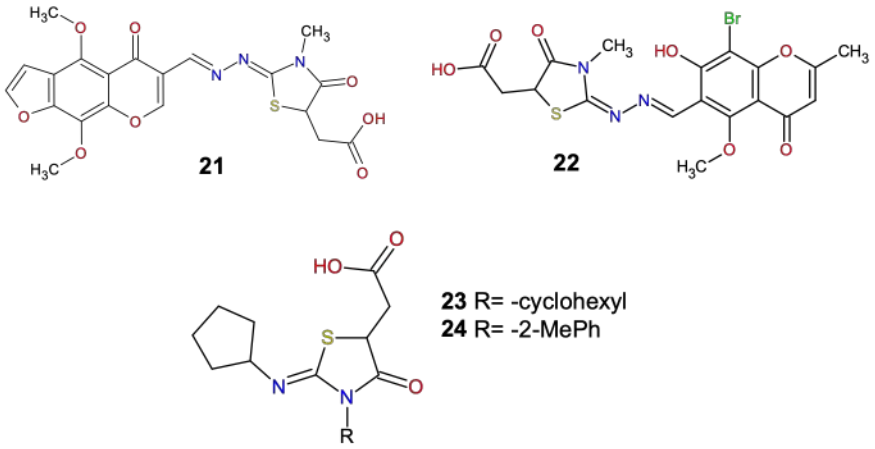


Fig. 4. Example derivatives of (2-imino-4-oxo-1,3-thiazolidin-5-yl)acetic acid possessing anticonvulsant activity.

activity of certain substituted furochromone, benzofuran, and flavone derivatives [22]. The authors prepared thiosemicarbazide and cyclic thiazolidinone derivatives from starting heterocyclic building blocks. All obtained compounds were examined for their potential anticonvulsant activity using subcutaneous pentylenetetrazole-induced seizures (scPTZ). The maximal electric shock-induced seizures (MES) method was also applied to all synthesized novel prepared derivatives. It was stated that the group of linear thiosemicarbazides showed 100 % protection in most cases at a dose of 150 mg/kg body weight in albino mice. However, when an acetic acid moiety was introduced at the 5-position of the thiazolidine-4-one nucleus (**21**, Fig. 4.), its protection percentage decreased from 100 % to 67 %.

Other synthesized (thiazolidin-5-yl)acetic acids showed comparable activity decreases or lacked protection from seizures. This phenomenon was explained by an imbalance in the lipophilicity/hydrophilicity ratio. Researchers presented similar bioactivity outcomes in another publication for comparable (thiazolidin-5-yl)acetic acid derivatives (**22**, Fig. 4.) [23]. Products were designed following a similar pattern, but 8-bromo-7-hydroxy-5-methoxy-2-methyl-4-oxo-4*H*-1-benzopyran-6-carbaldehyde and its 4-thio analog were used as starting materials.

A large group of 42 derivatives was synthesized, and their anticonvulsant and cardiovascular effects were investigated [24]. In the first step, 1,3-disubstituted thiosemicarbamides were obtained and then transformed into the corresponding 2-substituted-imino-3-substituted-5-carboxymethyl-4-thiazolidinones. Both groups were built using cyclopentyl or cyclohexyl substituents. Only thiazolidinones were transferred to biological tests. Studies on gross behavior, anticonvulsant activity, and effects on blood pressure were conducted in mice. Lethal dosages for synthesized compounds were established (LD_{50} mg/kg). The anticonvulsant activity was determined against pentylenetetrazol-induced seizures using a dose equivalent to 1/5 of the established LD_{50}. Although the study focused more on cardiovascular effects, it is worth noting that two compounds (**23**–**24**, Fig. 4.) reached 100 % protection levels. In parallel, both compounds caused no mortality after 24 h. Nine of the twenty evaluated derivatives showed no protection from seizures and a high percentage of 24-hour mortality. It is important to note that no reference substances were used in the studies.

3.3. Antimicrobial and antiviral activity

The exploration of innovative antimicrobial and antiviral agents continues to be a vital and challenging task for medicinal chemists due to several pressing factors. Firstly, the rapid evolution of drug-resistant strains of bacteria, viruses, and other pathogens poses a significant threat to global public health. These pathogens can quickly develop resistance to existing treatments, making it increasingly difficult to manage infections and disease outbreaks. As a result, there is a pressing need to develop novel antimicrobial and antiviral compounds that can effectively combat these resistant strains. Additionally, the emergence of new infectious diseases, often with pandemic potential, underscores the importance of ongoing research in this area. The COVID-19 pandemic is a stark reminder of infectious diseases' global impact, highlighting the critical need for the proactive development of new therapeutic agents that can be rapidly deployed against emerging threats.

Some papers report interesting results in the search for new antimicrobial or antiviral activity of derivatives of (2-imino-4-oxo-1,3-thiazolidin-5-yl)acetic acid [25]. In many research efforts, these compounds have been the primary focus of investigation. A series of 2-(arylimino)thiazolidin-4-ones and 2-(benzylidenehydrazono)-3-arylthiazolidin-4-ones was synthesized and investigated [26]. All the tested compounds have been evaluated for their *in vitro* antibacterial activity against selected clinically important pathogenic Gram-positive and Gram-negative bacterial strains by measuring the minimum inhibitory concentration (MIC) and the minimum bactericidal concentration (MBC). Gentamycin and tetracycline were used as reference substances. Incorporating an acetic acid side chain at the 5th position of the thiazolidinone ring, as seen in the example compounds **25** and **26** (Fig. 5.), led to analogs that exhibited neither antibacterial activity nor the ability to inhibit biofilm formation.

One of the compounds disclosed earlier [11] for antiprotozoal activity should also be mentioned for its interesting antimicrobial properties. It was found that 3-phenyl-2-[(5-methyl-1*H*-imidazole-4-yl) carbonyl]hydrazono-4-oxo-13-thiazolidin-5-yl-acetic acid (7) showed significant mean zone inhibition (MZI) against *Staphylococcus aureus* (19 mm) and *Bacillus subtilis* (27 mm). Unfortunately, the values of minimum inhibitory concentrations (MIC) were rather poor when compared with the standard drugs (chloramphenicol and rifampicin). In another study mentioned earlier [13], the authors presented preliminary results of antimicrobial activity toward selected bacterial strains. However, the values of MIC and MBC were rather high. In general, it can be concluded that synthesized derivatives were not active against the bacterial strains used. Only one compound (**27**, Fig. 5.) showed higher potency than reference chloramphenicol towards the *M. luteus* strain.

There are also examples of the search for antiviral compounds among derivatives of (2-imino-4-oxo-1,3-thiazolidin-5-yl)acetic acid. In a reported investigation, a series of thiosemicarbazones and thiazolidine analogs were synthesized and tested against a panel of viruses [27]. The inhibitory effect on viral growth was measured, and the prevention of cytopathogenic effects was observed using human embryonic kidney cell monolayers. Some of the derivatives were active against *herpes simplex virus strain Z* and *poliovirus no. 1* (Brunhilde) in direct contact or at time intervals after infection. Similar results were reported for *poliovirus no. 2* (Saukett) and vaccinia virus. Compounds were inactive against the measles virus. Examples of active derivatives are depicted below (**28** and **29**, Fig. 5.).

Research presented in a paper by Kobe *et al.* is interesting mostly because of the designed compounds [28]. The authors obtained ribofuranosyl/arabinofuranosyl conjugates with thiazolidinone derivatives (**30**, Fig. 5.). After an unsuccessful first attempt, two alternative synthetic routes were applied, involving glycosidation and amminolysis procedures. Several conjugates were obtained and investigated for antiviral activity. The biological experiments described in this paper utilized monolayer cultures of HeLa cells. These cells were grown and propagated in Minimum Essential Medium (MEM) enriched with 5–10 % fetal bovine serum. The cultures were maintained in a humidified atmosphere containing CO_2. While earlier studies found that beta-D-ribofuranosyl derivatives exhibited moderate antiviral activity, it was observed that both alpha- and beta-D-arabinofuranosyl derivatives showed no activity against *herpes simplex virus* and *poliovirus*.

In another study, a preliminary evaluation of antiviral properties was conducted for synthesized novel 1,2,4-triazolylthiazolidones [29]. The study concentrated on the synthesis and spectroscopic studies of six 1,3-thiazolidin-5-yl acetic acid-based compounds (**31**, Fig. 5.). The authors simply stated that in the evaluation against *herpes simplex virus*, the tested derivatives were inactive or toxic.

3.4. Anticancer activity

Research indicates that compounds featuring the thiazolidinone moiety demonstrate a broad spectrum of anticancer properties [30]. These effects are not only multifaceted but also target-specific, highlighting the system's capability to interact with a variety of anticancer biotargets. Such interactions suggest a mechanistic approach whereby thiazolidinones exert their antitumor effects by binding to and modulating the function of key proteins involved in cancer progression and survival [31–33]. Among these are phosphatases that play essential roles in cancer cell signaling pathways, such as the phosphatase of a regenerating liver (PRL-3) and the nonmembrane protein tyrosine phosphatase (SHP-2) [34–36]. These enzymes are pivotal in regulating pathways that control cell growth, migration, and survival, making them attractive targets for anticancer therapy. Additionally, thiazolidinones have been shown to interact with JNK-stimulating phosphatase-1 (JSP-1), a

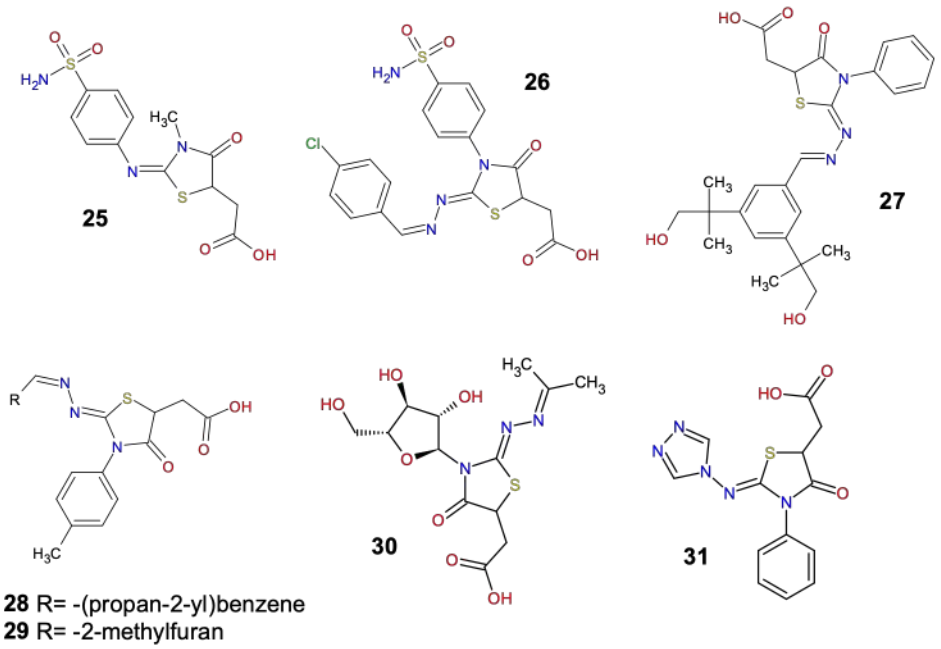


Fig. 5. Example derivatives of (2-imino-4-oxo-1,3-thiazolidin-5-yl)acetic acid possessing antimicrobial and antiviral activity.

component involved in stress response pathways that can influence cell death and survival [37,38]. Another significant aspect of the anticancer action of thiazolidinone compounds involves their ability to modulate the activity of tumor necrosis factor TNFα, a cytokine that plays a dual role in cancer by promoting inflammation but also aiding in cell death under certain conditions [39]. Furthermore, these compounds target the anti-apoptotic biocomplex Bcl-XL-BH3, crucial for the regulation of apoptosis, thus tilting the balance towards cell death in cancer cells [40]. Integrin αvβ3, a cell surface receptor involved in cell adhesion and migration, is yet another critical target of thiazolidinones [41]. By inhibiting this integrin, these compounds can potentially reduce the metastatic spread of cancer cells. The recent identification of thiazolidinones as inhibitors of necroptosis, a form of programmed cell death different from apoptosis, further broadens their therapeutic spectrum [42]. This mechanism offers a novel approach to killing cancer cells that have become resistant to traditional therapies. The ability of these compounds to interact with multiple key proteins and pathways underscores their potential as effective tools in cancer therapy. The continued exploration of thiazolidinones may provide invaluable insights into their mechanistic roles and pave the way for the development of novel anticancer treatments.

In the development of aminoquinazoline derivatives as potential multitarget anticancer agents with apoptotic and anti-proliferative effects, interesting results were found [43]. Newly designed 4-aminoquinazoline-based scaffolds were combined with semi/thiosemicarbazide, acetohydrazide, or Schiff's base moieties in one structure. Some compounds in that group were transformed into cyclic analogs (**32**, Fig. 6.). These substances were tested against 60 human tumor cell lines derived from nine cell lines (leukemia, non-small cell lung cancer, colon, CNS, ovarian, renal, melanoma, prostate, and breast cancer) at a single dose of 10 μM. Their growth inhibition (GI %) was determined.

Cyclization of the thiosemicarbazide derivative resulted in its thiazolidinone analog (**32**, Fig. 6.) that showed comparable antitumor activities against renal UO-31 (GI % = 11.9) and breast MCF7 (GI % = 12.9) cancer cell lines. Cooperatively, the screening results for thiazolidinone derivatives were found to be superior to those for thiosemicarbazide precursors. Compounds were not transferred to further studies, but more active derivatives were synthesized for this study.

Sahran *et al.* reported the synthesis of 6-bromo-coumarin-ethylidene-hydrazonyl-thiazolyl and 6-bromo-coumarin-thiazolyl-based derivatives. Among those obtained, two 4-oxo-3-phenylthiazolidin-5-yl-acetic acids (**33** and **34**, Fig. 6.) were included in the evaluation of anticancer properties. The cytotoxic effects were measured using three selected cell lines: MCF-7 (representing breast cancer), A-549 (human alveolar basal epithelial cells), and CHO-K1 (an ovarian cell line). Across all these cell lines, the synthesized compounds demonstrated a moderate level of cytotoxic activity. One lead non-cyclic compound possessing an N-phenyl thiourea moiety was identified and transferred to further testing. Both thiazolidine-based derivatives didn't reach the reference values of IC_{50} (Flavopiridol MCF-7 = 0.00445 μM, A-549 = 0.00114 μM, CHO-K1 = 0.00241 μM) [44].

Pyrazolyl-substituted 2(3*H*)-furanone was allowed to react with different nitrogen nucleophiles such as hydrazine hydrate; then the obtained hydrazides were converted to new forms such as thiazolidinone (**35**, Fig. 6.) and oxadiazole pyridazinone derivatives. The authors then transferred derivatives to biological activity evaluation [45]. Selected compounds were evaluated as anticancer agents against three types of carcinoma cell lines: human liver carcinoma (HepG2), human colon carcinoma (HCT116), and human prostate cancer (PC3) using doxorubicin as a reference drug. Unfortunately, the depicted derivative (**35**, Fig. 6.) did not reach the desired levels of cytotoxic activity to be included in additional studies.

In another paper, novel compounds were prepared *via* condensation of 1-(3-chlorophenyl)-3-(4-methoxyphenyl)-1*H*-pyrazole-4-carboxaldehyde with substituted thiosemicarbazide, followed by cyclization reactions. Four of the obtained compounds (**36**–**39**, Fig. 6.), possessing the 4-oxo-3-phenylthiazolidin-5-yl-acetic acid scaffold, presented interesting biological activities [46]. Sixteen target compounds were preliminarily screened for their *in vitro* cytotoxic activity against HEPG-2, PC-3, and A-549 cell lines at a concentration of 100 μg/mL. Derivatives **37** and **39** (Fig. 6.) displayed potent activity of 97 ± 1.10 %

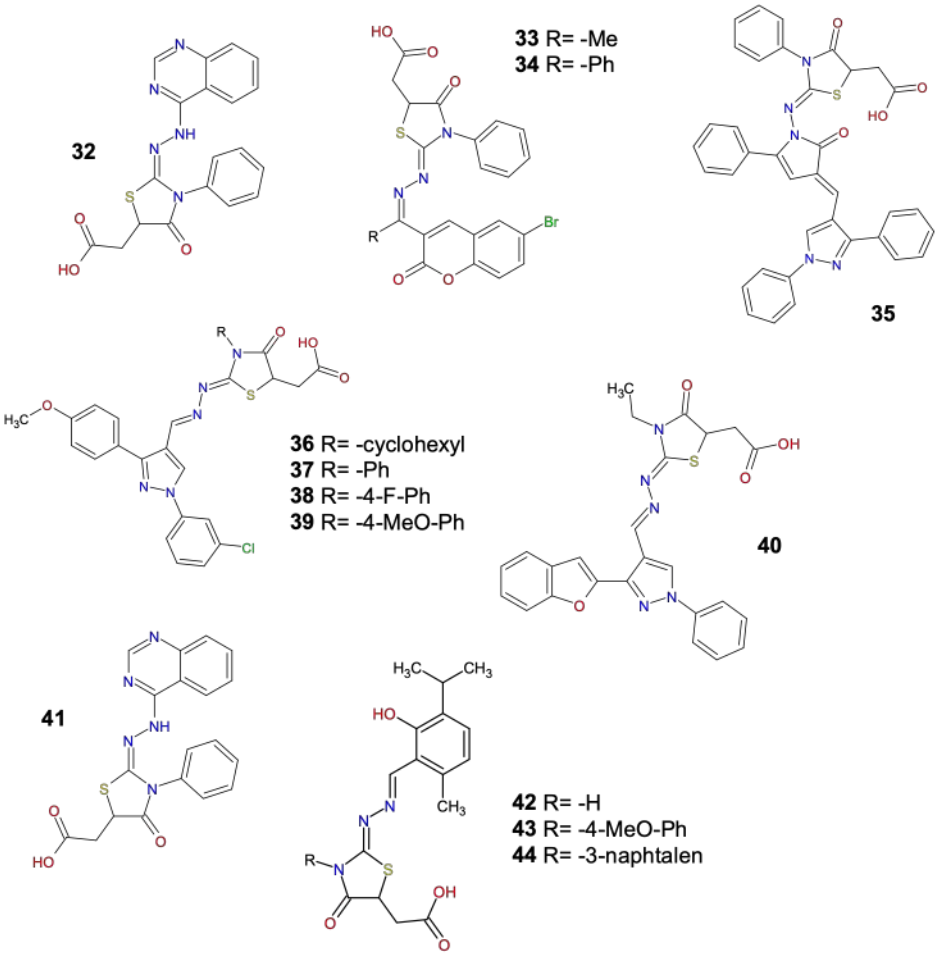


Fig. 6. Example derivatives of (2-imino-4-oxo-1,3-thiazolidin-5-yl)acetic acid possessing anticancer activity.

and 92.9 ± 1.62 % against HEPG-2, 97.7 ± 1.45 % and 96.9 ± 3.25 % against PC-3, and 79.6 ± 1.18 % and 92.2 ± 1.00 % against A-549, respectively. One of the compounds was not active, and another (**36**) gave potent cytotoxic activity against two cell lines only, HEPG-2 and PC-3 (98.4 ± 1.30 % and 98.5 ± 2.60 %, respectively). Doxorubicin was used as a reference drug. Interestingly, in the case of the HEPG-2 cancer cell line, the active derivative showed higher potency than the reference. In the glutathione S-transferase pi gene (GSTP1) inhibition assay, three derivatives (**36**–**38**, Fig. 6.) revealed potent inhibitory activity of 98.4 ± 0.03 %, 98.5 ± 0.05 %, and 98.8 ± 0.03 %, with IC_{50} values of 4.3 ± 0.29, 3.43 ± 0.32 μM, and 1.5 ± 0.18 μM, respectively. The most potent compounds were transferred to a molecular docking study. Results showed that the thiazole-pyrazole backbone of thiazolidinones is crucial for cytotoxic activity against the three types of carcinoma (HEPG-2, PC-3, and A-549) cell lines and inhibitory effect against GSTP1 enzyme. This may be due to the inclusion of a $–CH_2COOH$ moiety at the thiazolidinonyl scaffold, responsible for forming an H-bond between the oxygen of the C = O group and Gln39. In contrast, modifying compound **38** (Fig. 6.) by adding an electron-withdrawing group to the para position on the phenyl ring resulted in reduced affinity for GSTP1, leading to diminished anticancer efficacy across three carcinoma cell lines. On the other hand, other derivatives have shown antiproliferative effects in liver and prostate carcinoma evaluations. These findings highlight the potential of compounds as a novel class of GSTP1 inhibitors, meriting further investigation.

In another study, a series of pyrazole, thiazole, and thiazolidine derivatives incorporated into benzofuran were synthesized using 3-(benzofuran-2-yl)-1-phenyl-1*H*-pyrazol-4-carboxaldehyde as a starting material [47]. Among 41 compounds, one possessed the thiazolidin-5-yl-acetic acid scaffold (**40**, Fig. 6.). The primary objective of this study was to evaluate and compare the efficacy and toxicity of newly synthesized compounds against human liver carcinoma (HepG2) and human cervix carcinoma (HELA) cell lines relative to the established chemotherapeutic agent 5-fluorouracil (5-FU). This comparison aimed to identify agents that were not only more effective in inhibiting tumor growth but also exhibited reduced toxicity, potentially offering better therapeutic profiles for cancer treatment. It is unclear why only 7 from a group of 41 derivatives were evaluated in preliminary biological tests. The method of selecting molecules and the substantive justification were not presented. Unfortunately, compound **40** (Fig. 6.) was not included in the study.

Amin *et al.* presented the results of evaluating new derivatives for various cancer cell lines [48]. The newly synthesized 4-aminoquinazoline derivatives were assessed for their comprehensive anticancer capabilities, specifically focusing on their ability to induce apoptosis and inhibit cell proliferation. One of the designed compounds possessed the thiazolidin-5-yl-acetic acid scaffold (**41**, Fig. 6.). It was obtained through the cyclization of the thiosemicarbazide starting material.

Following this initial evaluation, a detailed *in vitro* investigation was conducted to measure the cytotoxic effects of these compounds utilizing the National Cancer Institute's (NCI) panel of 60 human cancer cell lines, encompassing nine distinct subpanels representing various cancer types. Derivative **41** (Fig. 6.) showed comparable anticancer activities against renal UO-31 (GI % = 11.9) and breast MCF7 (GI % = 12.9) cell lines. Overall, the screening results for the thiazolidinone derivatives were found to be superior to those of the thiosemicarbazide derivatives. Three of the most active compounds, exhibiting broad-spectrum activity against selected cancer cell lines, were selected for *in vitro* screening of their inhibitory activities against EGFR and CDK2 enzymes, excluding derivative **41**.

In a comprehensive paper on recent developments for cancer treatment targeting proviral integration sites for Moloney-murine leukemia virus (PIM) kinases, interesting thiazolidinones were included and discussed for their potency. Novel thymol 4-thiazolidinone derivatives were synthesized by connecting different compound fragments (**42**–**44**, Fig. 6.). It was found that these compounds showed enzyme inhibition effects involved in cancer and inflammation, such as COX-2, 5-LOX, and PIM. Focusing on anticancer properties, it was found that the compounds demonstrated a significantly higher inhibitory effect on PIM-2 relative to PIM-1. This enhanced specificity was elucidated through molecular docking studies, which identified a crucial hydrogen bond interaction between the hydroxyl group of the carboxyl (–COOH) functional group and the amino acid residue Ala122. This interaction is key to the compounds' increased affinity and selectivity for PIM-2. General observations suggest that the substituents present on the thiazolidinone ring might potentially decrease the enzymatic inhibition effectiveness against PIM-1 [49].

3.5. Patented derivatives of (2-imino-4-oxo-1,3-thiazolidin-5-yl)acetic acid possessing biological activity

Derivatives possessing the thiazolidin-5-yl-acetic acid scaffold have also become subjects of legal protection. Below are examples of structures that have been patented along with a brief description of the legal claims.

The patent EP0985670A1 focuses on novel compounds and methods for controlling undesirable plant species, particularly weeds, which are a significant agricultural problem globally. The invention introduces a series of novel compounds categorized as 1-(3-heterocyclylphenyl)isothiourea, −isourea, −guanidine, and −amidine. These compounds are primarily designed to function as herbicidal agents [50,51]. Depicted below are example structures **45**–**47** (Fig. 7.) that were part of the claim.

The next example patent concerns medicinally active compounds that inhibit peptidases, enzymes that cleave peptides at positions where the amino acid alanine is linked to p-nitroanilide. These substances are used for the prevention and treatment of a wide range of diseases characterized by excessive immune responses. This includes autoimmune diseases (like lupus and rheumatoid arthritis), allergies, and transplant rejections. The compounds are also applicable in the treatment of other chronic inflammatory diseases, which may include conditions such as inflammatory bowel disease and chronic obstructive pulmonary disease. The patent covers not only the individual medicinally active substances such as derivative **48** (Fig. 7.) but also pharmaceutical or cosmetic compositions that incorporate these substances [52,53].

Another invention claims that these new thiazolidine derivatives offer beneficial properties in antiviral medications, potentially contributing to more effective treatments against various viral diseases. The inclusion of these compounds (**49**–**50**, Fig. 7.) in different types of pharmaceutical formulations enhances their applicability in clinical settings, offering flexible treatment options [54].

The next patent describes novel sirtuin-modulating compounds (**51**, Fig. 7.) and their methods of use. Sirtuins are a family of proteins known to regulate important biological pathways in aging, stress responses, and metabolism. The invention provides significant therapeutic potential by modulating sirtuin activity, which is central to many physiological and pathological processes. The ability of these compounds to interact with sirtuin pathways allows for broad-spectrum applications in medicine, offering a foundational approach to developing treatments for diseases associated with aging and cellular stress [55].

4. Conclusions

The scientific community continuously follows recent advances in thiazolidinone-based compounds with various medicinal targets. However, the diversity of obtained compounds seems overwhelming in terms of structure–activity relationship studies due to the large heterogeneity of derivatives. In many cases, a single compound possesses thiazolidinone scaffolds as one of the few heterocyclic moieties. Our team analyzed a homogeneous group of derivatives of (2-imino-4-oxo-1,3-thiazolidin-5-yl)acetic acid. Presented derivatives are showing interesting results in antiprotozoal, anti-inflammatory, anticancer, antiviral, and antimicrobial studies. Some compounds targeting specific molecular targets were patented for their significant therapeutic potential. This

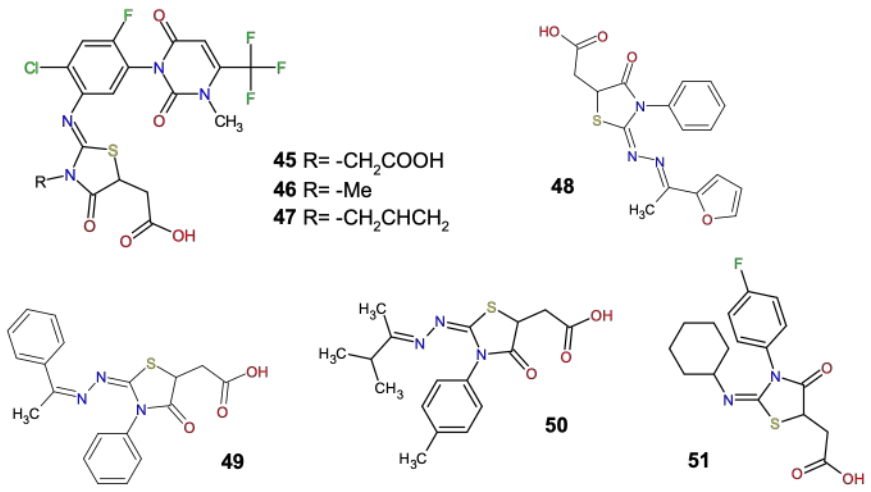


Fig. 7. Example patented derivatives of (2-imino-4-oxo-1,3-thiazolidin-5-yl)acetic acid possessing biological activity.

broadening of applications is attributed to the molecular adaptability of thiazolidinones, allowing for extensive structural modifications that enable the targeting of different biological pathways.

For most of the disclosed compounds, certain structural tendencies are apparent, regardless of their molecular targets. A proper structure–activity relationship (SAR) analysis should be conducted on a larger group of derivatives and focused on a single biological activity. Nonetheless, it is evident that, in most structures, the R_2 position (Scheme 1) is occupied by primary groups such as methyl, ethyl, allyl, cyclohexyl, and phenyl, which are occasionally substituted with hydroxy, methoxy, or halogen groups in the para position. In contrast, more varied moieties are observed at the N6 position (Scheme 1), predominantly heterocycles such as triazole, benzothiazole, imidazole, pyrazole, coumarin, and chromene. There are also substitutions with phenyl groups, but no aliphatic moieties. In the analyzed homogeneous subset, we observe a trend where compounds possessing an acetic acid substituent are rarely selected for further research. Most often, the lead compounds are those possessing different substituents on the thiazolidinone ring. However, it cannot be definitively stated that the cause of the poorer biological activity (regardless of the molecular target) is the substitution at the C5 position of the ring with acetic acid. A thorough study should be conducted to examine this substituent's impact on the compounds' physicochemical properties, such as solubility or pH, and to analyze its influence on the desired biological activity. In most published studies, only one or a few compounds containing the (2-imino-4-oxo-1,3-thiazolidin-5-yl)acetic acid core structure have progressed to further testing as "lead" derivatives. As a result, there are no consistent motifs associated with higher biological potency. The diversity of substituents should be enhanced by synthesizing a larger number of compounds in future studies. Given the limited data available, it is too early to draw definitive conclusions regarding the SAR.

Finally, it should be underlined that the stereochemistry of the majority of compounds under study was not specified. Additionally, the obtained products have a stereogenic center at the C5 (Scheme 1) position, so the compounds exist as a racemic mixture. The value of the presented conclusions regarding the biological activity of the studied compounds may be questioned due to the lack of stereochemical purity. Certainly, the results for mixtures of regioisomers or racemates can be treated as preliminary. However, to understand the true pharmacological potential of a molecule, it must be studied with defined stereochemistry.

CRediT authorship contribution statement

Tomasz Szostek: Writing – review & editing, Validation, Methodology, Investigation, Conceptualization. **Dagmara Otto-Ślusarczyk:** Validation, Methodology, Investigation. **Piotr Roszkowski:** Methodology, Formal analysis, Data curation. **Marta Struga:** Resources. **Daniel Szulczyk:** Writing – original draft, Supervision, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

References

[1] F.C. Brown, 4-Thiazolidinones, Chem. Rev. 61 (1961) 463–521, https://doi.org/10.1021/cr60213a002.

[2] D. Kaminskyy, A. Kryshchyshyn, R. Lesyk, 5-Ene-4-thiazolidinones - An efficient tool in medicinal chemistry, Eur. J. Med. Chem. 140 (2017) 542–594, https://doi.org/10.1016/j.ejmech.2017.09.031.

[3] A.K. Jain, A. Vaidya, V. Ravichandran, S.K. Kashaw, R.K. Agrawal, Recent developments and biological activities of thiazolidinone derivatives: A review, Bioorg. Med. Chem. 20 (2012) 3378–3395, https://doi.org/10.1016/j.bmc.2012.03.069.

[4] R. Tambach, Ueber Thiohydantoïnessigsäure und Diphenylthiohydantoïnessigsäure, Justus Liebigs Annalen Der Chemie 280 (1894) 233–246.

[5] J. M'Lean, F.J. Wilson, 225. The reaction between thiosemicarbazones and maleic anhydride, J. Chem. Soc. (1939) 1048–1050, https://doi.org/10.1039/JR9390001048.

[6] A.S. Pankova, M.A. Samartsev, I.A. Shulgin, P.R. Golubev, M.S. Avdontceva, M. A. Kuznetsov, Synthesis of thiazolidines via regioselective addition of unsymmetric thioureas to maleic acid derivatives, RSC Adv. 4 (2014) 51780–51786, https://doi.org/10.1039/C4RA07840C.

[7] M.M. McAllister, A decade of discoveries in veterinary protozoology changes our concept of "subclinical" toxoplasmosis, Vet. Parasitol. 132 (2005) 241–247, https://doi.org/10.1016/j.vetpar.2005.07.003.

[8] A.M. Tenter, A.R. Heckeroth, L.M. Weiss, *Toxoplasma gondii*: from animals to humans, Int. J. Parasitol. 30 (2000) 1217–1258, https://doi.org/10.1016/S0020-7519(00)00124-7.

[9] N. Trotsko, A. Bekier, A. Paneth, M. Wujec, K. Dzitko, Synthesis and In Vitro Anti-Toxoplasma gondii Activity of Novel Thiazolidin-4-one Derivatives, Molecules 24 (2019) 3029, https://doi.org/10.3390/molecules24173029.

[10] M.V. de O. Cardoso, L.R.P. de Siqueira, E.B. da Silva, L.B. Costa, M.Z. Hernandes, M.M. Rabello, R.S. Ferreira, L.F. da Cruz, D.R. Magalhães Moreira, V.R.A. Pereira, M.C.A.B. de Castro, P.V. Bernhardt, A.C.L. Leite, 2-Pyridyl thiazoles as novel anti-*Trypanosoma cruzi* agents: Structural design, synthesis and pharmacological evaluation, Eur. J. Med. Chem. 86 (2014) 48–59. https://doi.org/10.1016/j.ejmech.2014.08.012.

[11] Design, synthesis, molecular docking and biological evaluation of thiophen-2-iminothiazolidine derivatives for use against Trypanosoma cruzi, Bioorg. Med. Chem. 24 (2016) 4228–4240. https://doi.org/10.1016/j.bmc.2016.07.013.

[12] A.P. Liesen, T.M. de Aquino, C.S. Carvalho, V.T. Lima, J.M. de Araújo, J.G. de Lima, A.R. de Faria, E.J.T. de Melo, A.J. Alves, E.W. Alves, A.Q. Alves, A.J.S. Góes, Synthesis and evaluation of anti-Toxoplasma gondii and antimicrobial activities of thiosemicarbazides, 4-thiazolidinones and 1,3,4-thiadiazoles, Eur. J. Med. Chem. 45 (2010) 3685–3691, https://doi.org/10.1016/j.ejmech.2010.05.017.

[13] R.P. Tenório, C.S. Carvalho, C.S. Pessanha, J.G. de Lima, A.R. de Faria, A.J. Alves, E.J.T. de Melo, A.J.S. Góes, Synthesis of thiosemicarbazone and 4-thiazolidinone derivatives and their in vitro anti-*Toxoplasma gondii* activity, Bioorg. Med. Chem. Lett. 15 (2005) 2575–2578, https://doi.org/10.1016/j.bmcl.2005.03.048.

[14] T.M. de Aquino, A.P. Liesen, R.E.A. da Silva, V.T. Lima, C.S. Carvalho, A.R. de Faria, J.M. de Araújo, J.G. de Lima, A.J. Alves, E.J.T. de Melo, A.J.S. Góes, Synthesis, anti-*Toxoplasma gondii* and antimicrobial activities of benzaldehyde 4-phenyl-3-thiosemicarbazones and 2-[(phenylmethylene)hydrazono]-4-oxo-3-phenyl-5-thiazolidineacetic acids, Bioorg. Med. Chem. 16 (2008) 446–456, https://doi.org/10.1016/j.bmc.2007.09.025.

[15] A. Kryshchyshyn, D. Kaminskyy, O. Karpenko, A. Gzella, P. Grellier, R. Lesyk, Thiazolidinone/thiazole based hybrids – New class of antitrypanosomal agents, Eur. J. Med. Chem. 174 (2019) 292–308, https://doi.org/10.1016/j.ejmech.2019.04.052.

[16] S. Hosseininezhad, A. Ramazani, Thiazole ring- the antimicrobial, anti-inflammatory, and anticancer active scaffold, Arab. J. Chem. 16 (2023) 105234, https://doi.org/10.1016/j.arabjc.2023.105234.

[17] M.E. Haiba, S.S. Abd El-Karim, R.S. Gouhar, M.I. El-Zahar, S.A. El-Awdan, Synthesis and evaluation of anti-inflammatory and analgesic activity of some substituted thiazolyl and thiazolidinonyl tetrahydronapthalene derivatives, Med. Chem. Res. 23 (2014) 3418–3435, https://doi.org/10.1007/s00044-014-0926-z.

[18] D.H. Dawood, R.Z. Batran, T.A. Farghaly, M.A. Khedr, M.M. Abdulla, New Coumarin Derivatives as Potent Selective COX-2 Inhibitors: Synthesis, Anti-Inflammatory, QSAR, and 18. Molecular Modeling Studies, Arch. Pharm. (Weinheim) 348 (2015) 875–888, https://doi.org/10.1002/ardp.201500274.

[19] S.S. Bahekar, D.B. Shinde, Synthesis and Anti-inflammatory Activity of [2-(Benzothiazol-2-ylimino)-4-oxo-3-phenylthiazolidin-5-yl]-acetic Acid Derivatives, J. Kor. Chem. Soc. 47 (2003) 237–240, https://doi.org/10.5012/jkcs.2003.47.3.237.

[20] R. Lesyk, O. Vladzimirska, B. Zimenkovsky, S. Golota, I. Nektgayev, O. Cherpak, M. Leb'yak, O. Kozak, Synthesis and antiinflammatory activity of novel 3-(2,3-dimethyl-1-phenyl-4-pyrazolon-5-yl)-4-thiazolidones, Boll. Chim. Farm. 141 (2002) 197–201.

[21] P. Kwan, M.J. Brodie, Early Identification of Refractory Epilepsy, N. Engl. J. Med. 342 (2000) 314–319, https://doi.org/10.1056/NEJM200002033420503.

[22] F.-A.-F. Ragab, N.A.M. El-Sayed, A.A.H.M. Eissa, A.M.E. Kerdawy, Synthesis and Anticonvulsant Activity of Certain Substituted Furochromone, Benzofuran and Flavone Derivatives, Chem. Pharm. Bull. 58 (2010) 1148–1156, https://doi.org/10.1248/cpb.58.1148.

[23] F.-A.-E.-F. Ragab, G.S. Hassan, H.-A.-S. Abu Yossef, T.A.A. Yahya, H.A. Abd El-Latif, Synthesis and anticonvulsant activity of 4-oxo and 4-thioxo-8-bromobenzo-pyran derivatives, Arzneimittelforschung 60 (2010) 171–176.

[24] S. Nagar, H.H. Singh, J.N. Sinha, S.S. Parmar, Anticonvulsant and cardiovascular effects of substituted thiazolidones, J. Med. Chem. 16 (1973) 178–180, https://doi.org/10.1021/jm00260a027.

[25] S. Supajaruwong, S. Porahong, A. Wibowo, Y.-S. Yu, M.J. Khan, P. Pongchaikul, P. Posoknistakul, N. Laosiripojana, K.-C.-W. Wu, C. Sakdaronnarong, Scaling-up of carbon dots hydrothermal synthesis from sugars in a continuous flow microreactor system for biomedical application as in vitro antimicrobial drug nanocarrier, Sci.

Technol. Adv. Mat. 24 (2023) 2260298, https://doi.org/10.1080/14686996.2023.2260298.
[26] E.M. El-Hossary, Y.M. Nissan, K. Edkins, H. Bruhn, Synthesis and antibacterial activity of novel 2-(arylimino)thiazolidin-4-one and 2-(benzylidenehydrazono)-3-arylthiazolidin-4-one derivatives, J. App. Pharm. Sci. 6 (2016) 007–017, https://doi.org/10.7324/JAPS.2016.60502.
[27] A. Krbavcic, M. Plut, A. Pollak, M. Tisler, M. Likar, P. Schauer, Derivatives of 5-carboxymethylthiazolidine-2,4-dione, a new group of antiviral compounds, J. Med. Chem. 9 (1966) 430–431, https://doi.org/10.1021/jm00321a047.
[28] B. Rusjakovski, J. Kobe, B. Brdar, M. Japelj, U. Valcavi, Synthesis and antiviral activities of 2-substituted azino-3-(D-arabinofuranosyl)-5-carboxymethylene-thiazolidine-4-ones, Farmaco. Sci. 37 (1982) 764–780.
[29] M. Tisler, A. Andolsek, B. Stanovnik, M. Likar, P. Schauer, Synthesis and activity of some 1,2,4-triazolylthiazolidones, J. Med. Chem. 14 (1971) 53–54, https://doi.org/10.1021/jm00283a014.
[30] R.B. Lesyk, B.S. Zimenkovsky, 4-Thiazolidones: Centenarian History, Current Status and Perspectives for Modern Organic and Medicinal Chemistry, Curr. Org. Chem. 8 (n.d.) 1547–1577.
[31] D. Havrylyuk, B. Zimenkovsky, R. Lesyk, Synthesis and Anticancer Activity of Novel Nonfused Bicyclic Thiazolidinone Derivatives, Phosphorus Sulfur Silicon Relat. Elem. 184 (2009) 638–650, https://doi.org/10.1080/10426500802247563.
[32] D. Kaminskyy, B. Zimenkovsky, R. Lesyk, Synthesis and in vitro anticancer activity of 2,4-azolidinedione-acetic acids derivatives, Eur. J. Med. Chem. 44 (2009) 3627–3636, https://doi.org/10.1016/j.ejmech.2009.02.023.
[33] D. Havrylyuk, B. Zimenkovsky, O. Vasylenko, L. Zaprutko, A. Gzella, R. Lesyk, Synthesis of novel thiazolone-based compounds containing pyrazoline moiety and evaluation of their anticancer activity, Eur. J. Med. Chem. 44 (2009) 1396–1404, https://doi.org/10.1016/j.ejmech.2008.09.032.
[34] J.H. Ahn, S.J. Kim, W.S. Park, S.Y. Cho, J.D. Ha, S.S. Kim, S.K. Kang, D.G. Jeong, S.-K. Jung, S.-H. Lee, H.M. Kim, S.K. Park, K.H. Lee, C.W. Lee, S.E. Ryu, J.-K. Choi, Synthesis and biological evaluation of rhodanine derivatives as PRL-3 inhibitors, Bioorg. Med. Chem. Lett. 16 (2006) 2996–2999, https://doi.org/10.1016/j.bmcl.2006.02.060.
[35] H. Park, S.-K. Jung, D.G. Jeong, S.E. Ryu, S.J. Kim, Discovery of novel PRL-3 inhibitors based on the structure-based virtual screening, Bioorg. Med. Chem. Lett. 18 (2008) 2250–2255, https://doi.org/10.1016/j.bmcl.2008.03.013.
[36] A. Geronikaki, P. Eleftheriou, P. Vicini, I. Alam, A. Dixit, A.K. Saxena, 2-Thiazolylimino/heteroarylimino-5-arylidene-4-thiazolidinones as new agents with SHP-2 inhibitory action, J. Med. Chem. 51 (2008) 5221–5228, https://doi.org/10.1021/jm8004306.
[37] N.S. Cutshall, C. O'Day, M. Prezhdo, Rhodanine derivatives as inhibitors of JSP-1, Bioorg. Med. Chem. Lett. 15 (2005) 3374–3379, https://doi.org/10.1016/j.bmcl.2005.05.034.
[38] S. Kamila, H. Ankati, E.R. Biehl, An efficient microwave assisted synthesis of novel class of Rhodanine derivatives as potential HIV-1 and JSP-1 inhibitors, Tetrahedron Lett. 52 (2011) 4375–4377, https://doi.org/10.1016/j.tetlet.2011.05.114.
[39] P.H. Carter, P.A. Scherle, J.K. Muckelbauer, M.E. Voss, R.Q. Liu, L.A. Thompson, A. J. Tebben, K.A. Solomon, Y.C. Lo, Z. Li, P. Strzemienski, G. Yang, N. Falahatpisheh, M. Xu, Z. Wu, N.A. Farrow, K. Ramnarayan, J. Wang, D. Rideout, V. Yalamoori, P. Domaille, D.J. Underwood, J.M. Trzaskos, S.M. Friedman, R.C. Newton, C. P. Decicco, Photochemically enhanced binding of small molecules to the tumor necrosis factor receptor-1 inhibits the binding of TNF-alpha, Proc. Natl. Acad. Sci. USA 98 (2001) 11879–11884, https://doi.org/10.1073/pnas.211178398.
[40] A. Degterev, A. Lugovskoy, M. Cardone, B. Mulley, G. Wagner, T. Mitchison, J. Yuan, Identification of small-molecule inhibitors of interaction between the BH3 domain and Bcl-xL, Nat. Cell. Biol. 3 (2001) 173–182, https://doi.org/10.1038/35055085.
[41] R. Dayam, F. Aiello, J. Deng, Y. Wu, A. Garofalo, X. Chen, N. Neamati, Discovery of small molecule integrin alphavbeta3 antagonists as novel anticancer agents, J. Med. Chem. 49 (2006) 4526–4534, https://doi.org/10.1021/jm051296s.
[42] W. Zheng, A. Degterev, E. Hsu, J. Yuan, C. Yuan, Structure-activity relationship study of a novel necroptosis inhibitor, necrostatin-7, Bioorg. Med. Chem. Lett. 18 (2008) 4932–4935, https://doi.org/10.1016/j.bmcl.2008.08.058.
[43] N.H. Amin, M.T. El-Saadi, M.M. Abdel-Fattah, A.A. Mohammed, E.G. Said, Development of certain aminoquinazoline scaffolds as potential multitarget anticancer agents with apoptotic and anti-proliferative effects: Design, synthesis and biological evaluation, Bioorg. Chem. 135 (2023) 106496, https://doi.org/10.1016/j.bioorg.2023.106496.
[44] M.O. Sarhan, S.S. Abd El-Karim, M.M. Anwar, R.H. Gouda, W.A. Zaghary, M. A. Khedr, Discovery of New Coumarin-Based Lead with Potential Anticancer, CDK4 Inhibition and Selective Radiotheranostic Effect: Synthesis, 2D & 3D QSAR, Molecular Dynamics, In Vitro Cytotoxicity, Radioiodination, and Biodistribution Studies, Molecules 26 (2021) 2273, https://doi.org/10.3390/molecules26082273.
[45] W.S.I. Abou-Elmagd, A.K. EL-Ziaty, M.I. Elzahar, S.K. Ramadan, A.I. Hashem, Synthesis and antitumor activity evaluation of some N-heterocycles derived from pyrazolyl-substituted 2(3H)-furanone, Synth. Comm. 46 (2016) 1197–1208, https://doi.org/10.1080/00397911.2016.1193755.
[46] E.S. Nossier, S.M. El-hallouty, E.R. Zaki, Synthesis, anticancer evaluation and molecular modeling of some substituted thiazolidinonyl and thiazolyl pyrazole derivatives, Int. J. Pharm. Pharm. Sci. (2015) 353–359.
[47] M.I. El-Zahar, S.A. El-Karim, M.M. Anwar, Synthesis and Cytotoxicity Screening of Some Novel Benzofuranoyl-pyrazole Derivatives against Liver and Cervix Carcinoma Cell Lines, S. Afr. J. Chem. 62 (2009) 189–199.
[48] N.H. Amin, M.T. El-Saadi, M.M. Abdel-Fattah, A.A. Mohammed, E.G. Said, Development of certain aminoquinazoline scaffolds as potential multitarget anticancer agents with apoptotic and anti-proliferative effects: Design, synthesis and biological evaluation, Bioorg. Chem. 135 (2023) 106496, https://doi.org/10.1016/j.bioorg.2023.106496.
[49] L. Chen, W. Mao, C. Ren, J. Li, J. Zhang, Comprehensive Insights that Targeting PIM for Cancer Therapy: Prospects and Obstacles, J. Med. Chem. 67 (2024) 38–64, https://doi.org/10.1021/acs.jmedchem.3c01802.
[50] G.M. Karp, 1-(3-Heterocyclylphenyl)isothiourea, -isourea, -guanidine and -amidine compounds as herbicides, EP0985670A1 (2000).
[51] G.M. Karp, 1-(3-heterocyclylphenyl)isothiourea, -isourea, -guanidine and -amidine herbicidal agents, US2002137929A1 (2002).
[52] S. Ansorge, U. Bank, K. Nordhoff, M. Taeger, F. Striggow, Novel Alanyl-Amino Peptidase Inhibitors for Functionally Influencing Different Cells and Treating Immunological, Inflammatory, Neuronal, and Other Diseases, WO2005037257A2, 2005.
[53] S. Ansorge, U. Bank, K. Nordhoff, M. Taeger, F. Striggow, Novel alanyl-amino peptidase inhibitors for functionally influencing different cells and treating immunological, inflammatory, neuronal, and other diseases, US2007037752A1, 2007.
[54] Espacenet – search results, (n.d.). https://worldwide.espacenet.com/patent/search/family/025559692/publication/GB1258681A?q=pn%3DGB1258681A (accessed April 23, 2024).
[55] M. Milburn, J. Milne, N,n'-Dicyclic Isothiourea Sirtuin Modulators, WO2006094233A1, 2006.

17. Publication II

Synthesis, Cytotoxicity Evaluation, and Molecular Dynamics of 4-Thiazolidinone Derivatives: A Potential Path to EGFR-Targeted Anticancer Therapy.

Szostek, T., Otto-Ślusarczyk, D., Dobrowolski, M. A., Wiśniewski, M., Roszkowski, P., & Szulczyk, D.

Current Medicinal Chemistry, *33*, 1–29. (2025).

DOI: https://doi.org/10.2174/0109298673400295250913044755

IF= 3.2, **MNiSW=** 100

Send Orders for Reprints to reprints@benthamscience.net

Current Medicinal Chemistry, XXXX, *XX*, 1-00 1

RESEARCH ARTICLE

Synthesis, Cytotoxicity Evaluation, and Molecular Dynamics of 4-Thiazolidinone Derivatives: A Potential Path to EGFR-Targeted Anticancer Therapy

Tomasz Szostek[1], Dagmara Otto-Ślusarczyk[1], Michał A. Dobrowolski[2], Maciej Wiśniewski[3,4], Piotr Roszkowski[2] and Daniel Szulczyk[1,*]

[1]*Chair and Department of Biochemistry, Medical University of Warsaw, 02-097 Warsaw, Poland;* [2]*Faculty of Chemistry, University of Warsaw, Pasteura 1, 02-093 Warsaw, Poland;* [3]*Laboratory of Bioinformatics and Computational Genomics, Faculty of Mathematics and Information Science, Warsaw, Poland;* [4]*Laboratory of Functional and Structural Genomics, Center of New Technologies, University of Warsaw, Warsaw, Poland*

© 2025 The Author(s). Published by Bentham Science Publisher. This is an open access article published under CC BY 4.0 https://creativecommons.org/licenses/by/4.0/legalcode

ARTICLE HISTORY

Received: April 16, 2025
Revised: June 10, 2025
Accepted: June 20, 2025

DOI:
10.2174/0109298673400295250913044755

Abstract: ***Introduction*:** Heterocyclic compounds remain cornerstones of contemporary drug discovery because their ring-embedded heteroatoms confer adaptable electronics, conformational flexibility, and a broad spectrum of biological activities. The skeleton structure of 4-thiazolidinone is present in many cytotoxically active compounds and is often used in the design of new antitumor agents. This study aimed to synthesize, characterize, and evaluate the anticancer potential of fifteen new (2-imino-4-oxo-1,3-thiazolidin-5-yl)acetic acid derivatives.

***Methods*:** Compounds were synthesized using a consistent synthetic route involving a reaction between a thiourea derivative and maleic anhydride, which formed the thiazolidin-4-one ring through cyclization. The compounds were then categorized into three sets based on the attached heterocyclic rings (tryptamine, thiazole, and 1,2,4-triazole). The NMR and X-ray analysis followed the synthesis. Apoptotic effects, cell cycle arrest, IL-6 suppression, docking, and dynamics simulations were conducted. Preliminary cytotoxic activity was tested on metastatic colorectal cancer (SW620) and human breast adenocarcinoma (MDA-MB-231) cell lines using the MTT assay.

***Results*:** Compounds **5**, **6**, and **7** demonstrated notable selectivity indexes (4.73, 2.42, 4.16, respectively) and were further investigated for their mechanisms of action, revealing pro-apoptotic properties and the ability to induce cell cycle arrest. Additionally, compound **5** inhibited IL-6 secretion by 76%. *In silico* studies revealed the formation of an energetically stable complex between compound **5** and the EGFR crystal structure (min/-max binding affinities of -9.4|-8.0 kcal/mol, compared to the -7.71 kcal/mol for the native ligand).

***Discussion*:** This preliminary study provides compelling data on synthesized derivatives, but more advanced testing is needed to assess their therapeutic value fully. Compared with earlier reports on related thiazolidinone scaffolds, the present derivatives exhibit improved potency, clearer selectivity, and mechanistic features consistent with EGFR inhibition and cytokine modulation.

***Conclusion*:** These findings validate (2-imino-4-oxo-1,3-thiazolidin-5-yl)acetic acid as a privileged core for cytotoxic lead generation and indicate that strategic substitution with either a tryptamine moiety (compound **5**) or a 1,2,4-triazole ring (compound **7**) is particularly advantageous. These compounds are promising EGFR-targeting anticancer candidates, warranting further investigation.

Keywords: 4-thiazolidinone, anticancer, cytotoxicity, selectivity index, breast adenocarcinoma, molecular docking.

*Address correspondence to this author at the Chair and Department of Biochemistry, Medical University of Warsaw, 02-097 Warsaw, Poland; E-mail: daniel.szulczyk@wum.edu.pl

0929-8673/XX XXXX Bentham Science Publishers

1. INTRODUCTION

Regardless of many recent discoveries in the field of cancer treatment, its global epidemic remains one of the biggest challenges to the healthcare system. Despite promising data, as reported by the American Cancer Society, the total mortality rate associated with cancer in the United States decreased by 27% from 1991 to 2016 [1]. However, this phenomenon concerns only highly developed countries. According to Mattiuzzi and Lippi, global cancer frequency and mortality increase every year and are expected to maintain an incremental trend [2]. Moreover, currently used cancer treatments and newly discovered candidates for potential replacement are often highly toxic to normal human cells. This indicates that some cancer therapies, even if effective, cause dangerous adverse effects and reduce patients' quality of life [3]. Therefore, the development of new anticancer compounds is of critical importance, particularly those distinguished by high selectivity.

The group of 4-Thiazolidinones has been proven to be particularly biologically active scaffolds. Their wide spectrum of action includes anticancer, antiprotozoal, anti-inflammatory, anticonvulsant, and antimicrobial activity [4]. Currently, a few medicines containing a 4-thiazolidinone structure are available on the worldwide pharmaceutical market with particular emphasis on their use in diabetes treatment. Glitazones, including Pioglitazone (Fig. **1**) [5] and Rosiglitazone [6], are considered type 2 diabetes medicines, and Epalrestat was approved in Japan for the treatment of diabetic peripheral neuropathy [7]. Other therapeutic areas using medicines containing 4-thiazolidinones derivatives as active substances include Diuretics (Etozoline) [8], anticonvulsants (Ralitoline) [9], and, most recently, from 2021, Ponesimod (Ponvory), which was approved by the FDA for the treatment of multiple sclerosis and psoriasis [10].

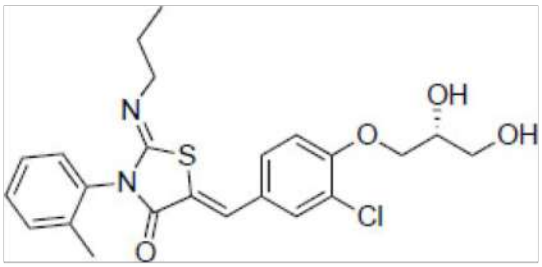


Fig. (1). Chemical structure of Ponesimod containing 4-thiazolidinone ring.

The skeleton structure of 4-thiazolidinone is present in many cytotoxically active compounds and is often used in the design of new antitumor agents. Türe *et al*. synthesized a series of new 4-thiazolidinone derivatives hybridized with Imatinib and investigated their cytotoxic properties. Three of the obtained hybrids demonstrated significant antiproliferative activity, further confirmed through apoptosis and cell cycle arrest tests. The most potent compound showed an IC_{50} of 4.86 ± 0.73 μM and 3.52 ± 0.91 μM for 24 and 48 hours, respectively, against K562 cells [11]. Szychowski *et al.* reported stimulation of ROS production in the squamous carcinoma cell line (SCC-15) after treatment with new 4-thiazolidinone derivatives (Les-2194, Les-3377, Les-3640) [12].

Particularly interesting are derivatives of (2-Imino-4-oxo-1,3-thiazolidin-5-yl)acetic acid, which were reported as potential antitumor agents by Amin *et al.* (2023). The compound was tested for antiproliferative activity with different scaffolds, such as thiosemicarbazides, and showed results of GI% = 11.9 and GI% = 12.9 against renal (UO-31) and breast (MCF7) cancer cell lines, respectively. Researchers concluded that preliminary anticancer studies suggested a better potential for (2-imino-4-oxo-1,3-thiazolidin-5-yl)acetic acid over thiosemicarbazides in this study [13]. Nossier *et al*. synthesized four new derivatives of 4-oxo-3-phenylthiazolidin-5-yl-acetic acid and evaluated their *in vitro* cytotoxic activity. In the cell viability test, the HEPG-2 (hepatoblastoma), PC-3 (prostatic adenocarcinoma), and A549 (non-small cell lung cancer) cell lines were treated with the compounds at a concentration of 100 μg/ml. Two synthesized compounds exhibited significant cytotoxic activity compared to the reference drug, doxorubicin. For the HEPG-2 cell line, the compounds showed viability percentages of 97 ± 1.10% and 92.9 ± 1.62%, with IC_{50} values of 3.43 ± 0.32 μM and 1.5 ± 0.18 μM, respectively. These compounds were subsequently subjected to molecular docking studies, which revealed an inhibitory effect against the GSTP1 enzyme [14].

The most recent research investigating the antitumor activity of novel (2-imino-4-oxo-1,3-thiazolidin-5-yl)acetic acid derivatives comes from 2024. They were hybridized with thymol and targeted against proviral integration sites for Moloney-murine leukemia (PIM) virus kinases. Researchers found that compounds exhibited inhibitory properties toward enzymes important for cancer metabolism and inflammation, including 5-LOX, COX-2, and PIM [15].

Moreover, 4-thiazolidinones were chosen because, unlike many frontline cytostatics that are virtually insoluble in water (*e.g.*, paclitaxel <0.01 mg/ml and < 60% oral bioavailability) [16], recent profiling shows

most thiazolidinone-bearing leads dissolve in phosphate-buffered saline at ≥0.2 mg/ml and satisfy Lipinski/Veber rules, giving them a markedly better predicted biopharmaceutical profile [17].

The aim of the study was to obtain new biologically active compounds. To maximize the probability of achieving the desired effect, a strategic approach was necessary during the design phase of the synthesis. As a solution, characteristic heterocyclic structures were selected for inclusion in the new (2-imino-4-oxo-1,3-thiazolidin-5-yl)acetic acid derivatives scaffolds, giving small molecule hybrids. Hybridization of molecules with well-documented activity has been proven to increase the probability of obtaining more cytotoxic, selective, multitargeting agents [18]. The choice of specific heterocycles in our research was based on recent scientific reports showing their potential for improving cytotoxic activity.

Heterocyclic compounds remain cornerstones of contemporary drug discovery because their ring-embedded heteroatoms confer adaptable electronics, conformational flexibility, and a broad spectrum of biological activities. Recent examples illustrate this versatility: bis-Schiff bases bearing a thiobarbiturate nucleus have emerged as potent α-glucosidase inhibitors, validated by docking and DFT analyses [19], and closely related thiobarbiturate derivatives display dual urease-inhibitory and antioxidant properties [20]. Benzimidazole frameworks continue to yield high-value leads, ranging from submicromolar prolyl-oligopeptidase antagonists [21] to neuroprotective analogues that target enzymes implicated in neurodegeneration [22]. Hydrazone-Schiff benzimidazoles likewise inhibit urease and perform robustly in molecular dynamics simulations [23], while oxadiazole-flurbiprofen hybrids extend anti-inflammatory chemistry into the anticancer arena [24]. Quantum-chemical studies further clarify how heteroatom substitution modulates charge distribution and ligand-target recognition in these motifs [25]. Collectively, such findings highlight the strategic value of heterocycles and rationalize our focus on 4-thiazolidinone hybrids as a fertile platform for anticancer lead optimization.

Tryptamine derivatives are considered mainly as 5-hydroxytryptamine receptor agonists. However, recent studies report that they can also possess cytotoxic activity. Salikov *et al.*, screened newly synthesized tryptamine analogs for antiproliferative activity against HEK293 (embryonic kidney), HEPG-2 (hepatoblastoma), and Jurkat (T lymphocyte) cell lines. The best candidate demonstrated an IC_{50} value of 4.3 ± 0.05 μM against the HEPG-2 cell line. Further confirmation of cytotoxic activity was obtained through Annexin V/PI staining, which indicated the presence of cells in apoptotic and necrotic states [26]. In another similar study, tryptamine derivatives exhibited even lower IC_{50} values. For example, the most active compound showed an IC_{50} value of 0.0015 μM against IGROV1 (human ovarian carcinoma) cells [27]. 1,2,4-Triazole and its derivatives possess a broad spectrum of biological activity including antifungal, antibacterial, and anticancer effects [28]. One of the most successful compounds containing 1,2,4-triazole, letrozole, has been approved for clinical tumor treatment [29]. *Guan et al.* reported a series of 1,2,4-triazoles hybridized with other N-heterocycles that showed promising cytotoxic activity through mechanisms of targeting VEGFR, TDO inhibition, microtubule inhibition, and others. Notably, the study mentions a hybrid of 1,2,3-triazole with 2,4-thiazolidinedione that exhibited higher cytotoxic activity against MCF-7 (human breast cancer) cells compared to the reference drug cisplatin [30]. The cytotoxic activity of thiazole derivatives is also well-known to scientists, and they are often used in designing new scaffolds with the potential of having antitumor activity [31]. Recent studies on thiazoles linked with other N-heterocycles, specifically thiosemicarbazones, have reported obtaining potentially cytotoxic active hybrids. After preliminary tests against HL-60 (human leukemia) cells, the best result of IC_{50} value was 3.67 ± 0.60 μM [32].

Thus, based on the extensive experience of our research team, as evidenced by the successful discovery of new biologically active compounds in previous studies [33-36], we propose a well-justified strategy for designing novel small-molecule hybrids. This approach focuses on aligning the highly active central structure of (2-imino-4-oxo-1,3-thiazolidin-5-yl)acetic acid with tryptamine, 1,2,4-triazole, and thiazole rings as active substituents. This strategy holds significant promise for the development of new potential antitumor agents, highlighting the critical importance of these molecular frameworks in enhancing both cytotoxic efficacy and selectivity.

2. MATERIALS AND METHODS

2.1. Chemistry

2.1.1. Chemicals and Reagents

Chloroform, methanol, and acetone were obtained from Sigma Aldrich (Saint Louis, MO, USA). Tryptamine (98%) was sourced from Acros Organics (Geel, Belgium), while 2-aminothiazole (≥99%) and 3-amino-1,2,4-triazole (≥98.5%) were purchased from Sig-

ma Aldrich (Saint Louis, MO, USA). All other reagents used were of analytical grade and applied without additional purification. Nuclear magnetic resonance (NMR) spectra were acquired using a Bruker AVANCE spectrometer (Bruker, Karlsruhe, Germany) operating at 300 or 500 MHz for ^{1}H NMR, and at 75 or 125 MHz for ^{13}C NMR. The spectra were recorded in $CDCl_3$ or a $CDCl_3/CD_3OD$ mixture (9:1), and chemical shifts (δ) are reported in ppm relative to tetramethylsilane (TMS). Electrospray ionization (ESI) mass spectra were measured on a high-resolution LCT Micromass TOF instrument (Micromass UK Limited, Manchester, UK). Melting points were determined using a KSP1D Melting Point Meter (A. Krüss Optronic, Hamburg, Germany) and are uncorrected. Thin-layer chromatography (TLC) was performed on silica gel plates (Merck Kiesegel GF254, Merck, Darmstadt, Germany) and visualized under UV light or by exposure to iodine vapors. Column chromatography was carried out at atmospheric pressure using Silica Gel 60 (230-400 mesh, Merck, Darmstadt, Germany), with a chloroform/methanol mixture (0-10%) as the mobile phase.

2.1.2. General Procedure for the Synthesis of the Compounds

A solution containing the thiourea derivative (146 mg, 1 mmol) and maleic anhydride (98 mg, 1 mmol) in acetone (15 mL) was stirred and heated under reflux for 15 hours. Upon completion, the solvent was removed under reduced pressure. The resulting residue was then refluxed with a small volume of a benzene/water mixture to eliminate any residual anhydride. The organic phase was separated, washed with water (30 mL), and dried over anhydrous sodium sulfate (Na_2SO_4). Following solvent evaporation under reduced pressure, the crude product was purified *via* column chromatography on silica gel using a chloroform/methanol gradient (0-10% MeOH) as the eluent. The purified compound was subsequently recrystallized from a 1:1 mixture of water and absolute ethanol.

2.1.2.1. [(2Z,5R)-2-(3-chloro-4-fluorophenyl)imino-3-(2-(1H-indol-3-yl)ethyl)-4-oxo-1,3-thiazolidin-5-yl]acetic acid (1)

White solid, Mp 122-124°C, yield: 76%. ^{1}H NMR (300 MHz, DMSO-d_6): δ 2.73-2.82 (m, 1H), 3.00-3.09 (m, 3H), 3.97-4.02 (m, 2H), 4.49 (dd, J = 3.9 Hz, 6.3 Hz, 1H), 6.82-6.87 (m, 1H), 6.90-6.93 (m, 1H), 6.96-7.01 (m, 1H), 7.05-7.10 (m, 1H), 7.21 (d, J = 2.4 Hz, 1H), 7.34-7.38 (m, 2H), 7.59 (d, J = 7.8 Hz, 1H), 10.86 (d, J = 2.4 Hz, 1H, NH), 12.77 (1H, COOH). ^{13}C NMR (75 MHz, DMSO-d_6): 22.4, 36.7, 43.0, 43.6, 110.6, 111.4, 117.4 (d, J_2 = 21.8 Hz), 118.2, 118.4, 119.7(d, J_2 = 18.0 Hz), 121.0, 121.4 (d, J_3 = 7.5 Hz), 122.7, 123.3, 127.2, 136.3, 145.5 (d, J_4 = 3.0 Hz), 154.1 (d, J_1 = 241.5 Hz), 156.3 (2xC), 171.7, 173.3. HRMS (ESI) m/z 446.9008 calc. for $C_{21}H_{17}ClFN_3O_3S$ [M+H]+; found 446.9011.

2.1.2.2. {(2Z,5R)-2-ethylimino-3-[2-(1H-indol-3-yl)ethyl]-4-oxo-1,3-thiazolidin-5-yl}acetic acid (2)

White solid, Mp 118-119°C, yield: 82%. ^{1}H NMR (300 MHz, DMSO-d6): δ 1.21 (t, J = 7.2 Hz, 3H, CH_3), 2.69-2.78 (m, 1H), 2.92-3.05 (m, 3H), 3.28 (q, J = 7.2 Hz, 3H, CH_2), 3.83-3.88 (m, 2H), 4.42 (dd, J = 3.9 Hz, 6.3 Hz, 1H), 6.96-7.01 (m, 1H), 7.04-7.07 (m, 1H), 7.17 (d, J = 2.4 Hz, 1H), 7.34 (d, J = 8.1 Hz, 1H), 7.62 (d, J = 7.5 Hz, 1H), 10.83 (d, J = 2.1 Hz, 1H, NH), 12.72 (1H, COOH). ^{13}C NMR (75 MHz, DMSO-d6): 15.9, 22.5, 37.1, 42.7, 43.1, 46.0, 110.8, 111.4, 118.3 (2xC), 121.0, 122.9, 127.2, 136.3, 151.0, 171.7, 172.9. HRMS (ESI) m/z 346.4244 calc. for $C_{17}H_{20}N_3O_3S$ [M+H]+; found 346.4246.

2.1.2.3. {(2Z,5R)-3-[2-(1H-indol-3-yl)ethyl]-4-oxo-2-phenylimino-1,3-thiazolidin-5-yl}acetic acid (3)

White solid, Mp 133-135°C, yield: 79%. ^{1}H NMR (300 MHz, DMSO-d_6): δ 2.70-2.79 (m, 1H), 2.98-3.11 (m, 3H), 3.98-4.03 (m, 2H), 4.45 (dd, J = 3.6 Hz, 9.3 Hz, 1H), 6.88-6.91 (m, 2H), 6.96-7.01 (m, 1H), 7.05-7.16 (m, 2H), 7.23 (d, J = 2.1 Hz, 1H), 7.34-7.39 (m, 3H), 7.62 (d, J = 7.8 Hz, 1H), 10.87 (d, J = 2.4 Hz, 1H, NH), 12.75 (1H, COOH). ^{13}C NMR (75 MHz, DMSO-d_6): 22.4, 36.8, 43.0, 43.3, 110.6, 111.5, 118.1, 118.4, 120.9 (2C), 123.1, 124.2, 127.3, 129.3 (2C), 136.3, 148.3, 154.4, 171.7, 173.3. HRMS (ESI) m/z 394.4684 calc. for $C_{21}H_{20}N_3O_3S$ [M+H]+; found 394.4687.

2.1.2.4. {(2Z,5R)-2-cyclohexylimino-3-[2-(1H-indol-3-yl)ethyl]-4-oxo-1,3-thiazolidin-5-yl}acetic acid (4)

White solid, Mp 120-121°C, yield: 83%. ^{1}H NMR (300 MHz, DMSO-d_6): δ 1.07-1.41 (m, 5H), 1.49-1.77 (m, 5H), 2.18-2.5 (m, 1H), 2.64-2.70 (m, 1H*), 2.73-2.81 (m, 1H), 2.92-3.03 (m, 3H), 3.45 (t, J = 7.2 Hz, 2H), 3.45 (t, J = 7.5 Hz, 2H*), 4.34 (dd, J = 3.6 Hz, 8.7 Hz, 1H), 4.40 (dd, J = 3.6 Hz, 9.3 Hz, 1H*), 6.94-7.00 (m, 1H), 7.03-7.09 (m, 1H), 7.16 (d, J = 2.1 Hz, 1H), 7.30 (d, J = 8.1 Hz, 1H), 7.55 (d, J = 7.8 Hz, 1H), 7.63 (d, J = 7.8 Hz, 1H*), 10.81 (s, 1H, NH), 12.75 (1H, COOH). ^{13}C NMR (75 MHz, DMSO-d_6):

22.4, 25.7, 26.3 (C*), 27.6, 27.7 (C*), 33.3, 33.3 (C*), 37.3, 42.3, 42.7 (C*), 43.0, 53.1, 54.4 (C*), 110.7, 111.3, 111.4 (C*), 118.1, 118.2 (C*), 118.3 (C*), 118.4, 120.8, 121.0 (C*), 122.8, 122.9 (C*), 127.3 (C*), 127.4, 136.1, 136.2 (C*), 149.1 (C*), 151.3, 171.5, 171.7 (C*), 172.9 (C*), 173.2. HRMS (ESI) m/z 400.5164 calc. for $C_{21}H_{26}N_3O_3S$ [M+H]+; found 400.5167.

2.1.2.5. {(2Z,5R)-2-(2-chloro-5-methylphenyl)imino-3-[2-(1H-indol-3-yl)ethyl]-4-oxo-1,3-thiazolidin-5-yl}acetic acid (5)

White solid, Mp 129-131°C, yield: 87%. ^{1}H NMR (300 MHz, DMSO-d_6): δ 2.82-2.88 (m, 1H), 3.00-3.13 (m, 3H), 3.18 (s, 3H), 4.00-4.05 (m, 2H), 4.50 (dd, J = 3.6 Hz, 8.4 Hz, 1H), 4.57 (dd, J = 4.2 Hz, 7.8 Hz, 1H*), 6.88-6.76 (d, J = 2.4 Hz, 1H), 6.94-7.01 (m, 1H), 7.05-7.11 (m, 2H), 7.23 -7.25 (m, 1H), 7.35-7.38 (m, 1H), 7.42-7.48 (m, 1H), 7.60 (d, J = 7.8 Hz, 1H), 10.78 (s, 1H*, NH), 10.88 (s, 1H, NH), 12.87 (1H, COOH). ^{13}C NMR (75 MHz, DMSO-d_6): 16.9, 22.5, 22.6 (C*), 36.2 (C*), 36.6, 43.1, 43.3 (C*), 43.5, 110.5 (C*), 110.6, 111.3 (C*), 111.5, 118.1, 118.2 (C*), 118.3 (C*), 118.4, 119.6, 120.9 (C*), 121.1, 122.8 (C*), 123.2, 123.9 (C*), 123.9, 127.2 (C*), 127.2, 128.4 (C*), 128.4, 130.4, 132.0, 136.1 (C*), 136.3, 148.4 (C*), 148.4, 155.1 (C*), 155.4, 170.5 (C*), 171.6, 173.2 (C*), 173.4. HRMS (ESI) m/z 442.9374 calc. for $C_{22}H_{21}ClN_3O_3S$ [M+H]+; found 442.9377.

2.1.2.6. [(2Z,5R)-4-oxo-2-phenylimino-3-(1H-1,2,4--triazol-5-yl)-1,3-thiazolidin-5-yl]acetic acid (6)

White solid, Mp 154-156°C, yield: 34%. ^{1}H NMR (300 MHz, DMSO-d_6): δ 3.12-3.35 (m, 2H, CH_2), 4.59 (s, 1H, CH), 6.84-7.60 (m, 5H), 8.58 (s, 1H, CH) 11.18 (s, 1H, NH), 11.80 (s, 1H, COOH). ^{13}C NMR (75 MHz, DMSO-d_6): 43.52, 51.73, 119.52, 120.60, 124.33 (2C), 125.50, 128.52 (2C), 128.81, 138.76, 166.88, 171,76. HRMS (ESI) m/z 318.3304 calc. for $C_{13}H_{12}N_5O_3S$ [M+H]+; found 318.3307.

2.1.2.7. [(2Z,5R)-2-(4-methylphenyl)imino-4-oxo-3-(1H-1,2,4-triazol-5-yl)-1,3-thiazolidin-5-yl]acetic acid (7)

White solid, Mp 157-159°C, yield: 29%. ^{1}H NMR (300 MHz, DMSO-d_6): δ 2.29 (s, 3H, CH_3), 3.05-3.08 (m, 2H, CH_2), 4.60 (s, 1H, CH), 7.16-7.32 (m, 3H), 7.46-7.48 (d, J = 8.4 Hz, 2H), 8.56 (s, 1H, CH) 11.13 (s, 1H, NH), 11.72 (s, 1H, COOH). ^{13}C NMR (75 MHz, DMSO-d_6): 20.55, 43.26, 51.71, 120.56 (2C), 124.36 (2C), 133.10 (2C), 134.76, 136.19, 145.69, 157.46, 166.77. HRMS (ESI) m/z 332.3574 calc. for $C_{14}H_{14}N_5O_3S$ [M+H]+; found 332.3577.

2.1.2.8. [(2Z,5R)-4-oxo-2-phenylimino-3-(1,3-thiazol-2-yl)-1,3-thiazolidin-5-yl]acetic acid (8)

White solid, Mp 144-146°C, yield: 72%. ^{1}H NMR (300 MHz, DMSO-d_6): δ 3.14-3.17 (m, 2H, CH_2), 4.60-4.64 (m, 1H, CH), 7.40-7.42 (m, 3H), 7.43-7.48 (m, 3H), 7.60-7.61 (d, J = 2.1 Hz, 1H), 7.60-7.61 (d, J = 1.5 Hz, 1H), 12.82 (s, 1H, COOH). ^{13}C NMR (75 MHz, DMSO-d_6): 43.38, 60.83, 117.26, 120.71, 124.18, 129.21 (2C), 135.40, 137.71, 139.86, 148.21, 155.39, 161.45, 169.01. HRMS (ESI) m/z 334.3874 calc. for $C_{14}H_{12}N_3O_3S_2$ [M+H]+; found 334.3876.

2.1.2.9. [(2Z,5R)-2-(4-methoxyphenyl)imino-4-oxo-3-(1,3-thiazol-2-yl)-1,3-thiazolidin-5-yl]acetic acid (9)

White solid, Mp 139-141°C, yield: 80%. ^{1}H NMR (300 MHz, DMSO-d_6): δ 3.12-3.14 (m, 2H, CH_2), 3.81 (s, 3H, OCH_3), 4.55-4.59 (m, 1H, CH), 7.03-7.10 (m, 3H), 7.23-7.31 (m, 3H), 7.39-7.40 (d, J = 3.6 Hz, 1H), 7.59-7.60 (d, J = 3.6 Hz, 1H), 12.78 (s, 1H, COOH). ^{13}C NMR (75 MHz, DMSO-d_6): 43.48, 55.35, 114.27 (2C), 117.17, 121.76, 127.80, 129.48 (2C), 139.83, 141.26, 159.20, 161.68, 171.76, 173.46. HRMS (ESI) m/z 364.4134 calc. for $C_{15}H_{14}N_3O_3S_2$ [M+H]+; found 364.4136.

2.1.2.10. [(2Z,5R)-2-(4-iodophenyl)imino-4-oxo-3-(1,3-thiazol-2-yl)-1,3-thiazolidin-5-yl]acetic acid (10)

White solid, Mp 162-165°C, yield: 83%. ^{1}H NMR (300 MHz, DMSO-d_6): δ 3.13-3.17 (m, 2H, CH_2), 4.57-4.65 (m, 1H, CH), 7.20-7.23 (m, 2H), 7.23-7.31 (m, 3H), 7.40-7.41 (d, J = 2.4 Hz, 1H), 7.60-7.61 (d, J = 2.1 Hz, 1H), 12.75 (s, 1H, COOH). ^{13}C NMR (75 MHz, DMSO-d_6): 43.75, 60.91, 115.35, 117.38, 123.34 (2C), 130.71 (2C), 138.04 (2C), 147.81, 156.85, 161.20, 168.88. HRMS (ESI) m/z 460.2839 calc. for $C_{14}H_{11}IN_3O_3S_2$ [M+H]+; found 460.2842.

2.1.2.11. [(2Z,5R)-2-(2-chlorophenyl)imino-4-oxo-3-(1,3-thiazol-2-yl)-1,3-thiazolidin-5-yl]acetic acid (11)

White solid, Mp 147-148°C, yield: 72%. ^{1}H NMR (300 MHz, DMSO-d_6): δ 3.27-3.34 (m, 2H, CH_2), 5.73-5.75 (m, 1H, CH), 7.41-7.65 (m, 5H), 7.67-7.70 (t, J = 3.6 Hz, 1H), 7.82-7.88 (m, 1H), 12.84 (s, 1H, COOH). ^{13}C NMR (75 MHz, DMSO-d_6): 34.04, 60.76, 115.55, 123.03, 129.11, 130.61, 131.88, 132.66,

133.29, 137.82, 156.49, 170.86 (2C), 177.63. HRMS (ESI) m/z 368.8294 calc. for $C_{14}H_{11}ClN_3O_3S_2$ [M+H]+; found 368.8297.

2.1.2.12. [(2Z,5R)-2-(4-bromophenyl)imino-4-oxo-3-(1,3-thiazol-2-yl)-1,3-thiazolidin-5-yl]acetic acid (12)

White solid, Mp 150-152°C, yield: 69%. ^{1}H NMR (300 MHz, DMSO-d_6): δ 3.13-3.15 (m, 2H, CH_2), 4.56-4.60 (m, 1H, CH), 7.24-7.27 (m, 2H), 7.40-7.53 (m, 3H), 7.60-7.62 (d, J = 3.6 Hz, 1H), 7.67-7.68 (d, J = 3.6 Hz, 1H), 12.86 (s, 1H, COOH). ^{13}C NMR (75 MHz, DMSO-d_6): 36.25, 43.70, 117.35, 127.36, 128.63, 131.54, 133.11, 134.10, 136.45, 139.89, 161.29, 168.88, 171.84. 173.87. HRMS (ESI) m/z 413.2834 calc. for $C_{14}H_{11}BrN_3O_3S_2$ [M+H]+; found 413.2837.

2.1.2.13. [(2Z,5R)-2-(3,4-dichlorophenyl)imino-4-oxo-3-(1,3-thiazol-2-yl)-1,3-thiazolidin-5-yl]acetic acid (13)

White solid, Mp 122-125°C, yield: 71%. ^{1}H NMR (300 MHz, DMSO-d_6): δ 3.66 (s, 2H, CH_2), 4.56-4.61 (t, J = 6.0 Hz, 1H, CH), 7.36-7.44 (m, 2H), 7.40-7.53 (m, 2H), 7.61-7.63 (d, J = 3.6 Hz, 1H), 7.72-7.73 (d, J = 2.1 Hz, 1H), 12.93 (s, 1H, COOH). ^{13}C NMR (75 MHz, DMSO-d_6): 36.28, 43.83, 117.51, 129.18, 130.55, 131.15, 131.29, 131.79, 135.03, 139.92, 161.02, 166.82, 168.70, 170.91. HRMS (ESI) m/z 403.2714 calc. for $C_{14}H_{10}Cl_2N_3O_3S_2$ [M+H]+; found 403.2717.

2.1.2.14. [(2Z,5R)-2-(3-chloro-4-methylphenyl)imino-4-oxo-3-(1,3-thiazol-2-yl)-1,3-thiazolidin-5-yl]acetic acid (14)

White solid, Mp 146-148°C, yield: 77%. ^{1}H NMR (300 MHz, DMSO-d_6): δ 3.57 (s, 3H, CH_3), 4.42-4.43 (m, 2H, CH_2), 5.68-5.70 (d, J = 7.2 Hz, 1H, CH), 7.34-7.38 (m, 3H), 7.57-7.58 (d, J = 2.1 Hz, 1H), 7.70-7.72 (d, J = 3.9 Hz, 2H), 12.90 (s, 1H, COOH). ^{13}C NMR (75 MHz, DMSO-d_6): 45.98, 50.95, 116.96, 120.83, 121.79, 123.21, 130.92, 131.52, 131.98, 134.81, 139.84, 162.74, 166.88, 169.06, 169.31. HRMS (ESI) m/z 382.8564 calc. for $C_{15}H_{13}ClN_3O_3S_2$ [M+H]+; found 382.8567.

2.1.2.15. [(2Z,5R)-2-(3-chloro-4-fluorophenyl)imino-4-oxo-3-(1,3-thiazol-2-yl)-1,3-thiazolidin-5-yl]acetic acid (15).

White solid, Mp 147-149°C, yield: 82%. ^{1}H NMR (300 MHz, DMSO-d_6): δ 3.34-3.35 (m, 2H, CH_2), 4.56-4.60 (t, J = 3.0 Hz, 1H, CH), 7.41-7.61 (m, 3H), 7.61-7.64 (m, 1H), 7.66-7.70 (m, 2H), 12.85 (s, 1H, COOH). ^{13}C NMR (75 MHz, DMSO-d_6): 43.76, 48.60, 117.32, 117.46, 117.62, 119.60, 119.86, 129.68, 129.78, 130.74, 132.09, 132.14, 139.91, 161.18. HRMS (ESI) m/z 386.8198 calc. for $C_{14}H_{10}ClFN_3O_3S_2$ [M+H]+; found 386.8201.

The example NMR spectra are given in the Supplementary file (see Figs. **S1**-**S30**).

2.2. X-Ray Studies

The experiment was performed on a SuperNova, single-source at offset/far, Eos diffractometer. The crystal was kept at 100.00(10) K during data collection. Using Olex2 [37], the structure was solved with the XS structure solution program using Direct Methods and refined with the XL refinement package using Least Squares minimization.

2.3. Biological Assay

2.3.1. Cell Line and Culture

The human cell lines utilized in this study included MDA-MB-231 (breast adenocarcinoma), SW620 (metastatic colorectal carcinoma), and HaCaT (immortalized human keratinocytes), all obtained from the American Type Culture Collection (ATCC, Rockville, USA). Cell culturing was conducted under specific conditions: HCT116 and SW620 cells were maintained in Minimum Essential Medium (MEM; Thermo Fisher Scientific, Waltham, MA, USA), whereas MDA-MB-231 and HaCaT cells were grown in Dulbecco's Modified Eagle Medium (DMEM) with high glucose content. The culture media were supplemented with 10% fetal bovine serum (FBS; Sigma-Aldrich, St. Louis, MO, USA), 20 mM HEPES buffer (Biowest, Nuaillé, France), and antibiotics (100 U/mL penicillin and 100 µg/mL streptomycin; Gibco, Grand Island, NY, USA). All cell lines were incubated at 37 °C in a humidified atmosphere containing 5% CO_2 until reaching approximately 80-90% confluence [36].

2.3.2. MTT Assay

To assess cytotoxic properties, the MTT assay was employed across three distinct cell lines, with results expressed as IC_{50} values. The selected cancer models included human metastatic colon carcinoma (SW620) and human breast adenocarcinoma (MDA-MB-231). As a reference for selectivity, a non-cancerous human keratinocyte line derived from adult skin (HaCaT) was used, and results were further evaluated using the Selectivity Index (SI).

The MTT assay is based on the enzymatic reduction of the yellow tetrazolium salt, 3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide (MTT), into purple formazan crystals by mitochondrial dehydrogenases in metabolically active cells.

Cells were seeded into 96-well plates at a density of 1×10^4 cells per well and allowed to adhere during a 24-hour incubation period. Subsequently, they were treated with varying concentrations (20-100 µM) of 1,3-thiazolidin-5-yl derivatives, while the control group remained untreated. After a 72-hour incubation at 37 °C in a humidified atmosphere with 5% CO_2, MTT solution (0.5 mg/mL) was added to each well, followed by an additional 4-hour incubation under the same conditions. To solubilize the resulting formazan crystals, a 1:1 (v/v) mixture of DMSO and isopropanol was used.

Absorbance was measured at 570 nm using a MultiscanGo microplate reader (Thermo Scientific). Cell viability was calculated as the percentage of MTT reduction in treated cells relative to the untreated control, using the formula: ([A]/[B]) × 100%, where [A] represents the absorbance of treated cells and [B] denotes the absorbance of the control group. Lower MTT values indicate reduced viability. IC_{50} values, defined as the concentration of compound that inhibits 50% of cell viability, were determined using GraphPad Prism 8.0.1 software [36]. The MTT assay was conducted across a concentration range of 1 to 100 µM. For compounds that failed to induce at least 50% inhibition of cell viability at 100 µM, IC_{50} values were recorded as ">100 µM" to indicate the absence of significant cytotoxicity within the tested range.

2.3.3. Cell Cycle and Apoptosis

2.3.3.1. Cell Cycle Analysis

SW620 and MDA-MB-231 cells were seeded in six-well plates at a density of 1×10^5 cells per well and incubated for 24 hours to facilitate attachment. Following this period, compounds 5, 6, and 7 were introduced at concentrations equivalent to their respective IC_{50} values, excluding the control wells, and cells were incubated for an additional 24 hours. Subsequently, both adherent and non-adherent cells were harvested and centrifuged at 400 × g for 5 minutes at 4 °C. The cell pellets were washed twice with 0.9% NaCl and fixed overnight in 500 µL of pre-cooled 70% ethanol.

Prior to flow cytometric analysis, samples were centrifuged at 850 × g for 5 minutes at 4 °C and rinsed with phosphate-buffered saline (PBS). Cells were then incubated at 37 °C for 30 minutes in the dark with 50 µL of RNase (100 µg/mL) and 200 µL of propidium iodide (PI; 50 µg/mL). Before measurement, each sample was diluted with an additional 100 µL of PBS. Cell cycle distribution was analyzed using a FACSVerse flow cytometer (Becton Dickinson, Franklin Lakes, NJ, USA). All experiments were carried out in quadruplicate [34].

2.3.3.2. Apoptosis

SW620 and MDA-MB-231 cells were seeded into 6-well plates at a density of 1×10^5 cells per well and incubated for 24 hours to promote cell attachment. Following this, compounds 5, 6, and 7 were administered at concentrations corresponding to their respective IC_{50} values, while control wells remained untreated. Cells were then incubated for an additional 72 hours.

Apoptosis was assessed using the commercially available FITC Annexin V Apoptosis Detection Kit I (Biosciences Pharmingen, San Jose, CA, USA). After treatment, cells were harvested, washed with phosphate-buffered saline (PBS), and stained with Annexin V-FITC and propidium iodide (PI), following the manufacturer's instructions (Becton Dickinson). Flow cytometry analysis was performed using a FACSVerse instrument (Becton Dickinson, Franklin Lakes, NJ, USA) to evaluate apoptotic status. Based on staining, cells were categorized as early apoptotic (Annexin V+/PI−) or late apoptotic/necrotic (Annexin V+/PI+) [34].

2.3.4. Human Interleukin 6

To assess interleukin-6 (IL-6) production, a human IL-6 ELISA kit (Diaclone SAS, Besançon Cedex, France) was used. The SW620 and MDA-MB-231 cell lines were treated with compounds 5, 6, and 7 at concentrations equivalent to their IC_{50} values and incubated for 72 hours. Non-treated cells were included as a control group. Following incubation, culture supernatants were collected and analyzed for IL-6 content using the enzyme-linked immunosorbent assay method, following the supplier's instructions. Each experimental condition was performed in triplicate.

2.3.5. In Silico Studies

2.3.5.1. Molecular Docking

To investigate interactions with the Epidermal Growth Factor Receptor, we performed pocket-based molecular docking of 15 studied molecules against 5 different crystal structures of the receptor, each featuring a defined active site and a confirmed antagonist of the receptor under study (1M17, 2ITO, 4I23, 4LQM,

7AEM). In selecting the crystal structures, we aimed for the greatest possible diversity in the Shape of the Binding Pocket (pocket lDDT metric) and Tanimoto Ligand Similarity while maintaining the appropriate antagonist properties of the native ligand.

Molecular docking was performed using the classical software AutoDock Smina 1.1.2 [38], targeting the native ligand pocket with an additional amount of buffer space measuring 4 Å. Subsequently, the docking scoring function results were compared with the results for native ligands. Visualizations were presented using PyMOL software.

2.3.5.2. Molecular Dynamics Simulation

Molecular dynamics simulations for selected complexes were performed using OpenMM 8.2.0 [39]. Prior to this, the protein was properly repaired by adding missing residues and amino acids, as well as standardizing non-standard residues. This process was carried out using MODELLER 10.6 [40] and the PDBFixer package.

For the dynamics simulations, the Amber ff14SB force field [41] was applied for the protein, while the gaff-2.11 force field was used for the ligand. The entire complex was placed in a solvent using the TIP3P water model with a padding of 10.0 Å. The ion concentration was set to 150 mM NaCl.

The prepared system underwent a multi-step preliminary dynamics process, which included the following stages: energy minimization, heating from 50 to 300 K, release of backbone restraints, and equilibration under NVT conditions, followed by NPT conditions. After these preparations, a full molecular dynamics simulation of the complex was conducted for 100 ns with a 1ps timestep and 100ps write interval.

The detailed parameters for the dynamics and each stage of the preparation process are presented in Table **1**.

Subsequently, the trajectories were analyzed using the PyTraj 2.0.2 python package [42], and free energy calculations were performed *via* gmx_MMPBSA 1.6.4 software [43]. MM/PBSA [44] and MM/GBSA [45] were calculated for every 100 simulation snapshots.

When analyzing molecular dynamics by the protein trajectory, we refer to the trajectories of all C-alpha atoms in the protein. By the ligand trajectory, we consider all atoms of the ligand molecule except for hydrogen. Finally, the pocket trajectory encompasses the amino acids located within a 6 Å distance from the native ligand in the first frame of its simulation.

2.3.5.3. ADMET Prediction

To perform the prediction of key-druglike parameters, the open source ADMET-AI python package [46] was used in our local batch environment. The ottained-data file of predictions was given to standardization using MaxAbsScaller, which is an estimator for scaling each feature by its maximum absolute value, ensuring that all values lie between -1 and 1 while preserving the data's sparsity. Then, a standardized data file was provided to hierarchical clustering using the Ward method, which groups compounds based on their similarities across selected molecular properties. The ward method minimizes the variance within each cluster, creating a more distinct clustering pattern. To create a visualization as a heatmap with a dendrogram, the cluster map function from Seaborn was used, which automatically performs hierarchical clustering and generates a heatmap with a dendrogram. To create a radial chart, standard python coding with the NumPy library was used. To display the plots with matplotlib.pyplot library was used.

2.3.6. Statistical Analysis

Statistical evaluations were performed using GraphPad Prism 9 (San Diego, CA, USA). Results are expressed as mean ± standard deviation (SD) based on a minimum of three independent experiments. Differences between groups were analyzed using one-way ANOVA, followed by Dunnett's post hoc test for multiple comparisons. A p-value of less than 0.05 was considered statistically significant [34].

3. RESULTS AND DISCUSSION

3.1. Chemistry

3.1.1. Synthesis

In this paper, we disclose fifteen novel (2-imino-4-oxo-1,3-thiazolidin-5-yl)acetic acid derivatives. The synthetic strategy involved obtaining five compounds in each group based on the heterocyclic substituent used (Table **2**). We aimed to test the effectiveness of the reaction in terms of yields, universality, and biological activity outcomes. The obtained derivatives were categorized into three sets based on the heterocyclic rings attached to the core 4-thiazolidinone motif: tryptamine, thiazole, and 1,2,4-triazole. The synthetic route was consistent across all derivatives, involving a reaction between a thiourea derivative and maleic anhydride, resulting in the formation of thiazolidin-4-one ring analogs through cyclization (Fig. **2**).

Table 1. Summary of parameters used in molecular dynamics simulations.

Properties	Molecular Dynamics Simulation Step					
	Minimization	**Heating**	**Backbone Restrains Removal**	**NVT Equilibration**	**NPT Equilibration**	**Production Run**
Time [ns]	~ until 10*kj/mole	0.025	0.5	0.5	0.5	100
Temperature	50	from 50 to 300	300	300	300	300
Integrator	Langevin Integrator					
Thermostat	Andersen Thermostat					
Barostat	MonteCarlo Barostat					
Timestep Value [ps]	1.0					
Friction Value [1/ps]	1.0					
NonBonded Method	Periodic Mesh Ewald with 10.0Å cutoff					

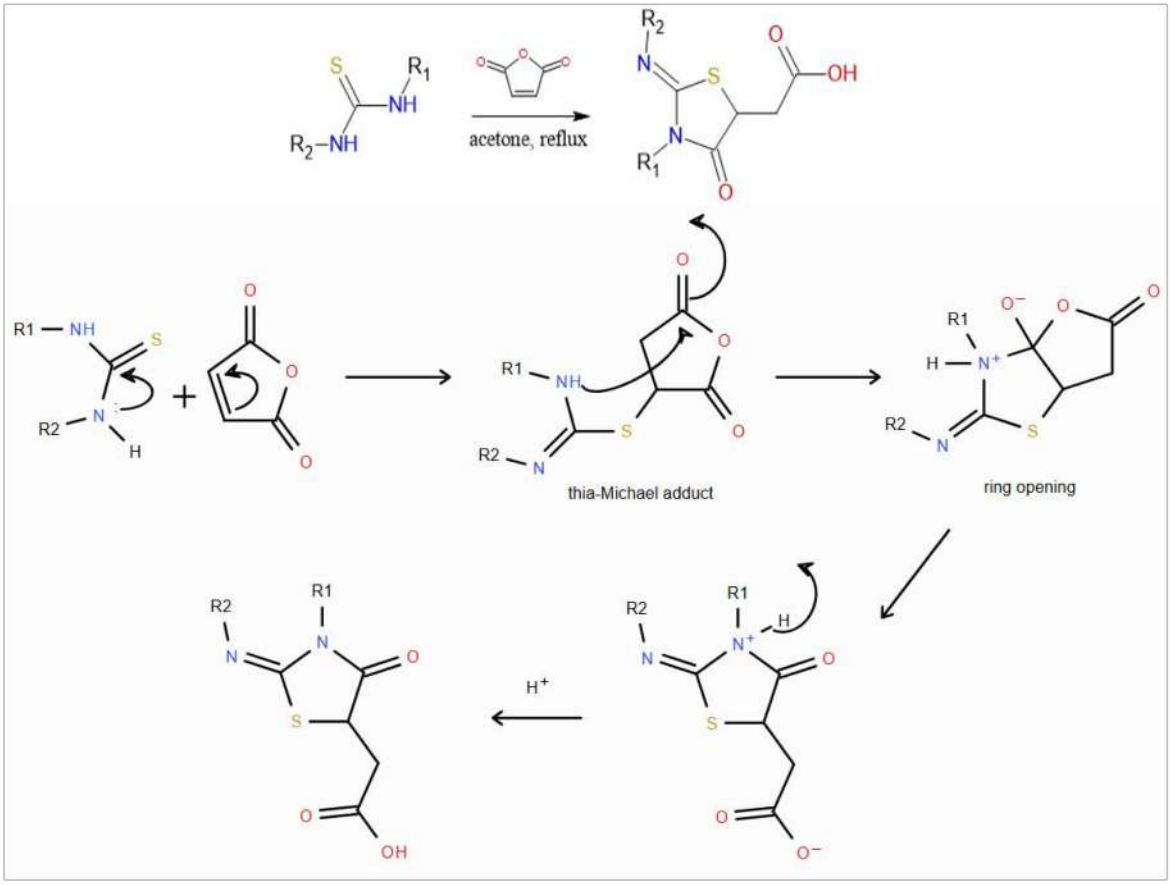


Fig. (2). The scheme of the synthetic pathway of new (2-imino-4-oxo-1,3-thiazolidin-5-yl)acetic acid derivatives (compounds **1-15**) with proposed general mechanism of product formation.

Table 2. Summary of main heterocyclic substituent (R_1) and additional substituents (R_2) applied for the synthesis of new (2-imino-4-oxo-1,3-thiazolidin-5-yl)acetic acid derivatives (compounds 1-15).

R_1	R_2	
Tryptamine	1(-3Cl,4F-Ph) 2 (-Et), 3 (-Ph)	4 (-cyclohexyl) 5 (-2Cl,4Me-Ph)
1,2,4-Triazole	6 (-Ph) 7 (-4Me-Ph)	
Thiazole	8 (-Ph) 9 (-OMe-Ph) 10 (-4I-Ph) 11 (-2Cl-Ph)	12 (-4Br-Ph) 13 (-3,4Cl-Ph) 14 (-3Me,4Cl-Ph) 15 (-3F,4Cl-Ph)

Based on previous reports, acetone [47] was selected as the solvent for this reaction to predefine regioselectivity, as the nucleophilic attack by the nonequivalent nitrogen atoms of the thiourea could result in varying ratios of regioisomers. We can confirm that the use of acetone promoted the formation of a single regioisomer, as reported in the previous studies [48, 49].

However, the synthesis of 4-thiazolidinones combined with a 1,2,4-triazole ring resulted in unsatisfactory yields due to the low solubility of the triazole substrate under the reaction conditions. This limitation led to the formation of only two derivatives in this group (compounds **6** and **7**). Despite this, the promising preliminary anticancer screening results for these compounds motivated us to include them in further tests, which are detailed in this article. Additionally, we plan to perform separate studies on a larger group of 4-thiazolidinone and triazole hybrids, applying alternative synthetic pathways to improve yields, solubility, and stereochemical purity.

The structures of all obtained products were confirmed using NMR spectroscopic analysis, including ^{1}H and ^{13}C NMR, revealing characteristic chemical shifts consistent with the proposed molecular frameworks. In the case of ^{1}H NMR, the most identifiable were: the methine proton (adjacent to carbonyl, and attached to the acetic acid group), the methylene protons of the CH_2COOH side chain, and the COOH proton. For ^{13}C NMR: the C=O (lactam carbon), the C=N (imino group), and the COOH carbon (carboxylic acid).

It should be noted that for (2-imino-4-oxo-1,3-thiazolidin-5-yl)acetic acid derivatives, tautomerism and H-bonding may appear and influence the NH, COOH, or other group signals not only due to the central 4-thiazolidinone skeleton, but also due to the chosen heterocycles. The Supplementary file includes the specific data. In addition, confirmation of stereochemistry was achieved *via* X-ray analysis of the monocrystal of compound **2** (see section 2.1.2). Such a study is widely considered the gold standard or final proof for structure confirmation because it provides direct, three-dimensional atomic-level details of a molecule.

3.1.2. X-ray Diffraction Studies

After synthesis, compound **2** was successfully submitted for crystallization using a 1:1 mixture of water and absolute ethanol, resulting in the isolation of a monocrystal structure. This allowed us to perform X-ray diffraction studies, which confirmed the regioselective nature of the presented synthetic route under the given conditions and previously described formation of a single regioisomer. Crystal data and structure refinement are summarized in Table **3**. The Cambridge Crystallographic Data Centre (CCDC 2456121) contains the supplementary crystallographic data for this paper. These data can be obtained free of charge *via*https://www.ccdc.cam.ac.uk/structures/. To our knowledge, this is the first crystal structure of (2-imino-4-oxo-1,3-thiazolidin-5-yl)acetic acid derivative reported to date.

Compound **2** crystallizes in the monoclinic crystal system, belonging to the space group $P2_1/c$. The asymmetric unit contains one molecule of {(2Z,5R)-2-ethylimino-3-[2-(1*H*-indol-3-yl)ethyl]-4-oxo-1,3-thiazolidin-5-yl}acetic acid, along with two water molecules (Fig. **3**). The crystal packing reveals several intermolecular N-H···O hydrogen bonds, including N3-H3···O20 (2.794 Å) and N22-H22···O21 (2.745 Å), as illustrated in Fig. (**4**). Moreover, the overall structure is further stabilized by O-H···O hydrogen bonding interactions involving the water molecules, with bond lengths ranging from 2.766 to 3.007 Å.

Table 3. Crystallographic parameters and structure refinement details for compound 2.

Properties	Compound 2
Empirical formula	$C_{17}H_{19}N_3O_3S_1,2(H_2O)$
Formula weight	381.44
Temperature [K]	100.00(10)
Space group	$P2_1/c$
Unit cell dimensions	
a [Å]	7.4109(2)
b [Å]	12.5547(4)
c [Å]	19.1743(7)
a [°]	90.00
β [°]	100.129(3)
γ [°]	90.00
Volume *V* [$Å^3$]	1756.21(10)
Z [molecules/cell]	4
$D_{calculated}$ [g cm^{-3}]	1.443
Absorption coefficient μ/mm^{-1}	0.219
2*θ* range for data collection [°]	3.9-52.72
Limiting indices	-6 <= *h* => 9
	-12 <= *k* => 15
	-20 <= *l* => 23
Reflections collected/unique	6615/3594
Data/parameters	3594/254
Goodness of Fit	1.041
Final *R* index ($I > 2\sigma$)	0.0543
wR^2	0.1334
Largest diff. Peak and hole [$Å^{-3}$]	0.90 and -0.60

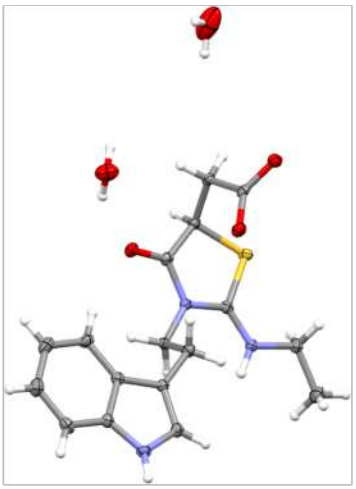


Fig. (3). The molecular structure of compound **2** with two water molecules showing displacement ellipsoids at the 50% probability level.

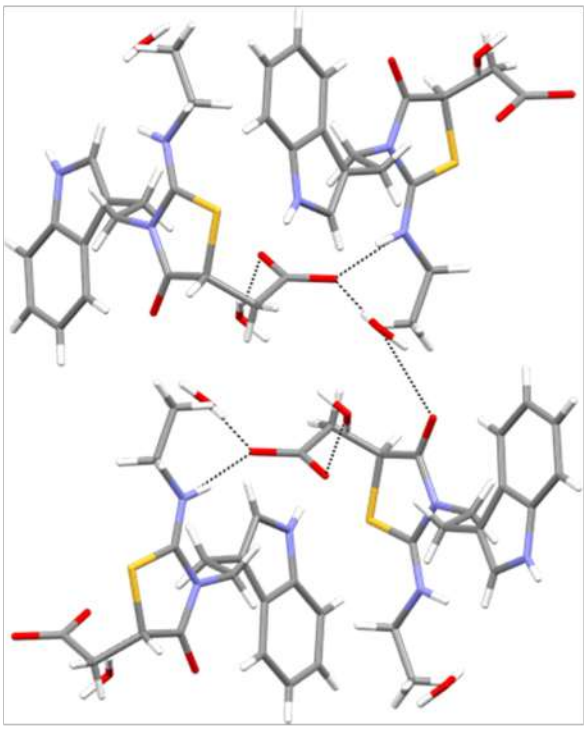


Fig. (4). Packing arrangement and key intermolecular interactions within the crystal lattice of compound 2, viewed along the [100] crystallographic axis.

This unequivocal structural confirmation not only validates the regio- and stereochemical integrity of our synthetic route but also provides a solid foundation for the subsequent biological evaluation of these thiazolidinone derivatives in cytotoxicity assays in the next steps of this research.

3.2. Cytotoxic Activity

A preliminary assessment of cytotoxic activity was performed for all newly obtained compounds (Table **4**). It was expressed in IC_{50} values and established using the MTT method. The tumor cells used in the study were derived from human metastatic colon cancer (SW620) and human breast adenocarcinoma (MDA-MB-231) cell lines. The results of IC_{50} values against cancer cell lines were compared to those obtained from tests with the human immortal keratinocyte cell line (HaCaT), representing normal cells, and expressed as selectivity index (SI).

Most of the new 4-thiazolidinone derivatives from the first set, containing a tryptamine ring, were found not to be active toward the studied cancer cell lines. However, compound **5** bearing a 2-chloro-4-methylophenyl substituent exhibited significant cytotoxic activity at a level of IC_{50} =21.16 ± 0.36 μM, registered against the MDA-MB-231 cancer cell line. It is worth emphasizing that HaCaT cells were not affected by compound **5,** which resulted in a high selectivity index of 4.73. In the group of 4-thiazolidinones containing a thiazole ring, most of the compounds did not show antiproliferative activity. The best result in this group was observed for compound **15**, presenting IC_{50}= 53.24 ± 2.80 μM against SW620 cell line.

In general, the thiazole-substituted derivatives exhibited the weakest cytotoxic activity among the tested compounds, with several showing IC_{50} values above 100 μM. This decreased potency may be attributed to the electronic and steric characteristics of the thiazole moiety, which introduces planarity and rigidity that may impair optimal receptor binding. In contrast to the more flexible triazole and tryptamine derivatives, the thiazole ring may reduce compound permeability or hinder critical interactions within the active site of molecular targets.

Table 4. Assessment of cell viability reduction by the studied compounds *via* MTT[a] assay.

Compound	Cancer Cell Lines				Normal Cells
	MDA[f]		SW 620[d]		HaCaT[e]
	IC_{50}[b]	SI[c]	IC_{50}[b]	SI[c]	IC_{50}[b]
1	100	1	100	1	>100
2	100	1	100	1	>100
3	100	1	100	1	>100
4	100	1	100	1	>100
5	21.16 ± 0.36	4.73	100	1	>100
6	100	1	41.31 ± 1.98	2.42	>100
7	36.82 ± 2.78	2.71	24.10 ± 3.98	4.16	>100
8	100	1	100	1	>100
9	100	1	100	1	>100
10	77.18 ± 3.68	1.30	100	1	>100
11	100	1	100	1	>100
12	100	1	100	1	>100
13	100	0.24	75.71 ± 8.06	0.31	23.48 ± 7.45
14	100	1	100	1	>100
15	100	1	53.24 ± 2.80	1.88	>100
Ref[g]	1.65±0.23	0.28	1.83 ± 0.10	0.16	0.46 ± 0.03

[a] The MTT assay is a commonly applied colorimetric technique to evaluate cell viability based on metabolic activity. It functions through NADPH-dependent oxidoreductase enzymes that convert the MTT tetrazolium salt into an insoluble, purple-colored formazan product. [b] Data are reported as mean ± SD. The IC_{50} (μM) value indicates the compound concentration that reduces cell growth by 50% in comparison to untreated control cells, following 72 hours of exposure. [c] The Selectivity Index (SI) was determined using the equation: SI = IC_{50} for non-cancerous cells / IC_{50} for cancerous cells. [d] Human metastatic colon cancer cell line (SW620). [e] Human keratinocyte cell line derived from adult skin (HaCaT). [f] Human breast adenocarcinoma cells (MDA).
[g] Doxorubicin was used as a standard reference drug due to its established role in anticancer therapy.

The 1,2,4-triazole derivatives (compounds **6** and **7)** contained phenyl and 4-methylophenyl substituents, respectively, and exhibited significant cytotoxic activity. Compound **7** showed IC_{50}= 36.82 ± 2.78 μM and IC_{50}= 24.10 ± 3.98 μM, against MDA-MB-231 and SW620 cancer cells, respectively. Compound **6** had slightly lower activity, with an IC_{50}= 41.31 ± 1.98 μM (SW620). Neither of the compounds exhibited cytotoxic properties toward the normal human cell line (HaCaT), giving IC_{50} values greater than 100 μM. This indicated high selectivity indexes for compounds **6** and **7** at 2.42 and 4.16, respectively.

Based on their favorable antiproliferative profiles observed in the MTT assay, particularly in the MDA-MB-231 cell line and extraordinary selectivity, compounds **5**, **6**, and **7** were selected for further investigation. These derivatives exhibited the most pronounced cytotoxic activity within the tested series. Their enhanced performance suggested a promising structure-activity relationship, warranting more detailed mechanistic and molecular docking studies to elucidate their potential mode of action and interaction with oncogenic targets.

3.3. Investigation of Anticancer Mechanism

3.3.1. Apoptotic Activity

Due to the results from preliminary cytotoxic activity evaluations against human metastatic colon cancer (SW620) and human breast adenocarcinoma (MDA-MB-231) cell lines, compounds **5**, **6**, and **7** were selected for further testing to investigate their mechanism of action. To establish the apoptotic effects of the novel (2-imino-4-oxo-1,3-thiazolidin-5-yl)acetic acid derivatives on these cancer cell lines, flow cytometry followed by annexin V-FITC and PI staining was used. As shown in Figs. (**5** and **6**), we observed significant differences in the number of cells in the apoptotic state in samples treated with the studied compounds compared to the untreated control group, suggesting that these compounds may possess proapoptotic properties. Overall, MDA-MB-231 cells demonstrated greater sensitivity to the studied compounds in inducing apoptosis compared to SW620 cells.

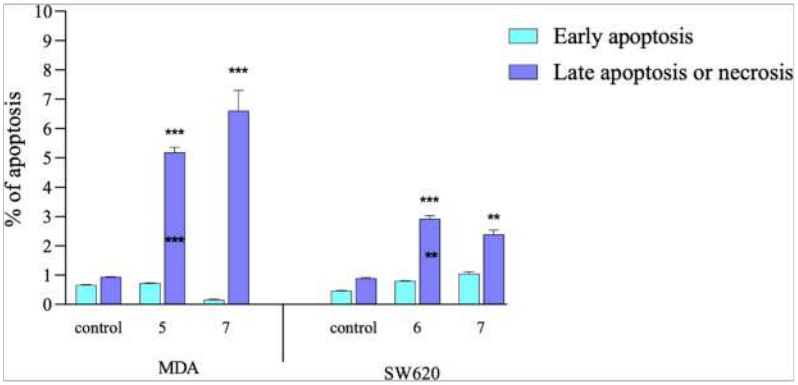


Fig. (5). Effect of novel N-derivatives of (2-imino-4-oxo-1,3-thiazolidin-5-yl)acetic acid **5**, **6,** and **7** on late apoptosis or necrosis and early apoptosis in human metastatic colon cancer (SW620) and human breast adenocarcinoma (MDA-MB-231) cell lines. Cells were treated with compounds at their IC_{50} values and incubated for 72h, followed by staining with annexin V-FITC and propidium iodide (PI). The assay was performed using flow cytometry. The outcomes are expressed as the proportion of cells undergoing early apoptosis and those in late apoptosis or necrotic stages. *** $p \leq 0.0001$, ** $p \leq 0.001$, in comparison to the control group. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

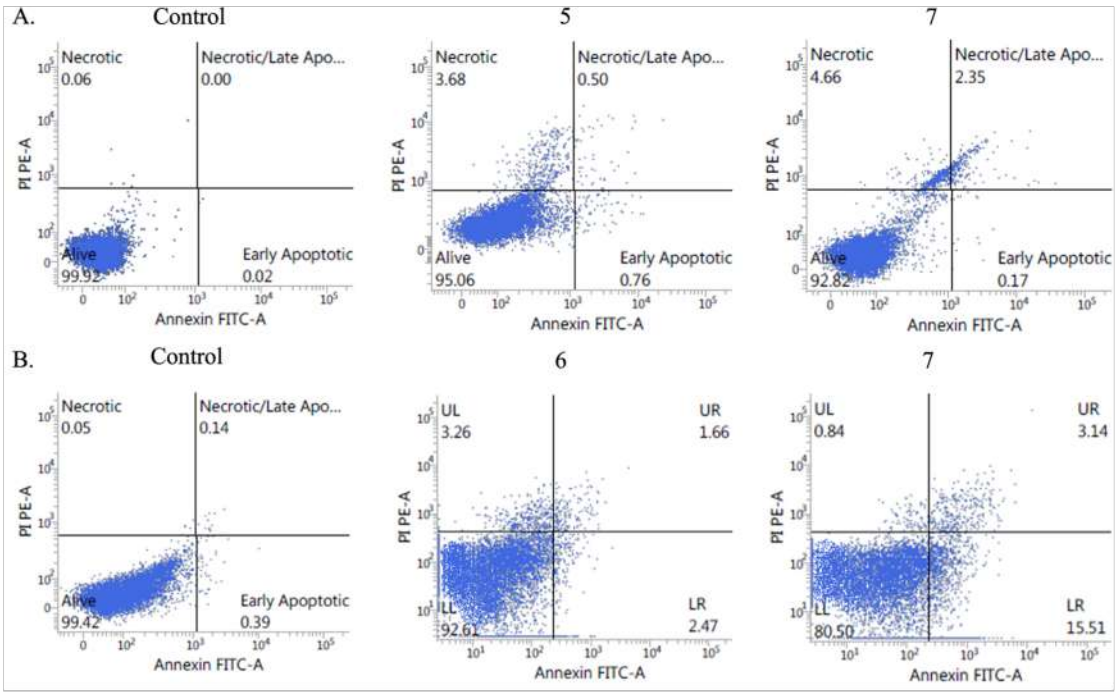


Fig. (6). Representative percentages presented as dot plots illustrate the apoptosis assessment of (**A**) MDA-MB-231, (**B**) SW620, and other cancer cell lines treated with compounds **5**, **6**, and **7**. The analysis was carried out *via* flow cytometry using Annexin V-FITC/PI double staining, distinguishing between Q1-viable cells, Q2-early apoptotic, Q3-late apoptotic/necrotic, and Q4-necrotic populations. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

Compound **7** (Figs. **5** and **6**) had the strongest proapoptotic effect, leading 6.60 ± 0.69% ($p \leq 0.0001$) and 2.38 ± 0.15% ($p \leq 0.001$) of MDA-MB-231 and SW620 cells, respectively, to late apoptosis or necrosis, compared to 0.15 ± 0.029% for MDA-MB-231 and 1.05 ± 0.07% for SW620 in the control groups. Compound **6** acted similarly towards SW620 cells, resulting in 2.91 ± 0.12% ($p \leq 0.0001$) of cells being registered in the stage of late apoptosis or necrosis (Figs. **5** and **6B**). Furthermore, derivative **5** led a significant number of MDA-MB-231 cells to enter the state of late apoptosis or necrosis (Figs. **5** and **6A**) with a percentage of 5.19 ± 0.17% ($p \leq 0.0001$), compared to 0.93 ± 0.015% in the control group.

The results obtained from the apoptosis assay confirmed the observations from the MTT experiment, revealing the properties of compounds **5**, **6,** and **7** in inducing a moderate number of cells to late apoptosis and necrosis. Additionally, MDA-MB-231 cancer cells showed slightly higher sensitivity to this effect. This suggests that the antiproliferative properties of the studied compounds may be related to their potential to induce proapoptotic effects in cancer cells.

Since apoptosis often arises from earlier disruptions of cell-cycle control, we asked whether the same compounds that triggered late-stage cell death might also halt proliferation at specific checkpoints. Therefore, after confirming that derivatives **5**, **6**, and **7** significantly increased the fraction of cells in late apoptosis or necrosis, we analyzed their impact on cell-cycle distribution in both MDA-MB-231 and SW620 cell lines. The next section details how these compounds influence the G1, S, and G2/M phase populations, shedding light on whether cell-cycle arrest precedes and possibly drives the proapoptotic effects observed above.

3.3.2. Induction of Cell Cycle Arrest

The cell cycle is a series of events, including DNA duplication, protein translation, and gene transcription, which take place in a single cell before it divides into two new cells [50]. Dysregulation of crucial cell cycle proteins and the cell cycle itself is a hallmark of cancer cells, enabling their uncontrolled proliferation and growth. Thus, therapies targeting dysregulated cell cycle processes in cancer cells can potentially hinder their development [51-53].

To evaluate the effect of compounds **5**, **6**, and **7** on cell cycle progression in MDA-MB-231 and SW620 cancer cells, the cells were exposed to IC_{50}-equivalent concentrations of the compounds and incubated for 24 hours. Cell cycle distribution was then analyzed using flow cytometry.

Both compounds **5** and **7** influenced cell cycle progression in MDA-MB-231 cells, although their effects differed (Figs. **7** and **8A**). Compound **7** caused cell cycle arrest in the G0/G1 phase, as indicated by a higher percentage of cells in this phase compared to untreated controls. A reduction in the proportion of cells in the S and G2/M phases accompanied this. These findings imply that compound **7** may exhibit antiproliferative activity, potentially promoting apoptosis, which is further supported by a modest increase in the sub-G1 fraction, representing apoptotic or dead cells. In contrast, compound **5** led to an accumulation of cells in the G2/M phase, with a concurrent decline in the S phase population, suggesting a different mechanism of cell cycle disruption.

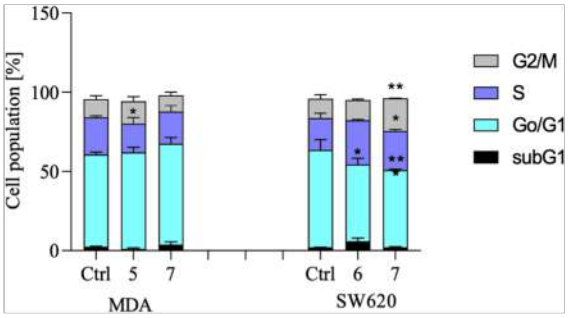


Fig. (7). Cell cycle analysis of MDA-MB-231 and SW620 cells was performed after 24-hour treatment with compounds 5, 6, and 7 at concentrations equivalent to their respective IC_{50} values. The SubG1 population corresponds to apoptotic and non-viable cells. Data are shown as mean ± standard deviation (SD) from three independent experiments. ** $p \leq 0.001$, * $p \leq 0.05$, compared to the untreated control. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

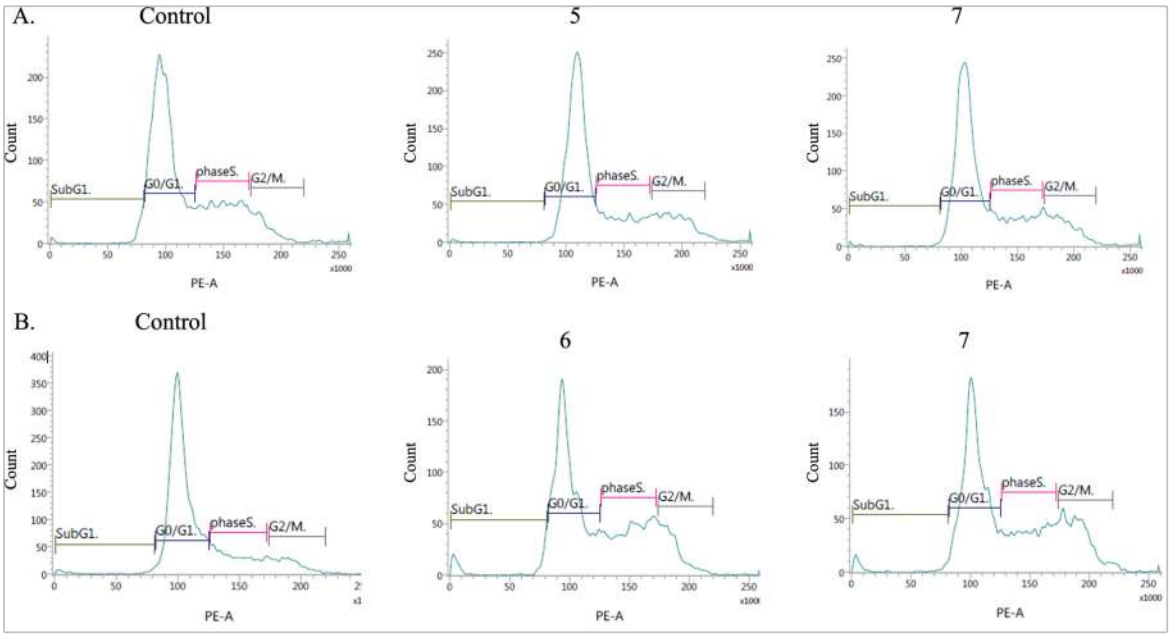


Fig. (8). Flow cytometric evaluation of cell cycle distribution was conducted following 24-hour exposure of (**A**) MDA and (**B**) SW620 cancer cells to compounds 5, 6, and 7 at concentrations corresponding to their IC_{50} values. The SubG1 phase indicates the proportion of apoptotic or non-viable cells. Data are expressed as mean ± standard deviation (SD) from three independent experiments. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

In SW620 colorectal cancer cells, compound **7** demonstrated the most pronounced effect, significantly increasing the proportion of cells in the G2/M phase, alongside a rise in the S phase population, relative to the control group (Figs. **7** and **8B**). This was coupled with a marked reduction in G0/G1 phase cells (48.89%, $p \leq 0.001$), suggesting cell cycle arrest at a later checkpoint. Treatment with compound **6** resulted in a 1.7-fold increase in the S phase cell population ($p \leq 0.05$), along with a decrease in cells residing in the G0/G1 phase. Furthermore, exposure to compound **6** led to a 2.77-fold increase in the sub-G1 fraction compared to the control, indicating elevated levels of apoptosis or cell death.

The results obtained from the cell cycle investigation are consistent with those from previous analyses. The induction of cell cycle arrest in the G0/G1 phase and an increased population of cells in the sub-G1 phase, representing dead cells, confirmed the proapoptotic potential of compounds **6** and **7**. Compound **5** demonstrated antiproliferative potential through inhibition of cell growth in the S phase, preventing cells from progressing to mitosis, which occurs in the G2/M phase.

Because checkpoint blockade and apoptosis are often modulated by inflammatory cues—especially the IL-6/JAK-STAT3 axis, which can either push tumor cells into G0/G1 arrest or, conversely, promote survival depending on context [54, 55] —we next examined whether our thiazolidinone derivatives alter IL-6 signaling. The following section therefore, reports how compounds **5**, **6,** and **7** affect IL-6 secretion and pathway activation in MDA-MB-231 and SW620 cultures, clarifying whether modulation of this cytokine accompanies the cell cycle and apoptotic changes described above.

3.3.3. Human Interleukin 6

To better understand the cytotoxic mechanisms of action, (2-imino-4-oxo-1,3-thiazolidin-5-yl)acetic acid derivatives (compounds **5**, **6**, and **7**) were subjected to a human interleukin-6 (IL-6) assay (Fig. **9**). IL-6 is a prototypical, pro-inflammatory cytokine involved in maintaining homeostasis [56] and is overexpressed in almost every type of cancer. A higher concentration of IL-6 has been shown to promote tumorigenesis by regulating crucial metabolic and signaling pathways. There-

fore, compounds that reduce IL-6 secretion in cancer cells can be considered potential antitumor agents [57].

To conduct a Human IL-6 test, previously used cancer cells were treated with the compounds at concentrations corresponding to their IC_{50} values. Among the human cancer cell lines used in the study, MDA-MB-231 was identified as more sensitive to the inhibition of IL-6 release compared to SW620. The strongest effect on MDA-MB-231 cells was observed with compound **5**, which inhibited IL-6 secretion by a factor of 4.1, compared to the control group ($p \leq 0.001$).

However, for compound **7**, the registered concentration of IL-6 was 6.63±1.76 pg/mL, compared to 8.67±1.54 pg/mL in the control group, indicating its moderate inhibitory influence on the same cancer cell line ($p \leq 0.01$). For SW620 cells, inhibition of IL-6 secretion was observed at levels of 1.38-fold and 2.21--fold compared to the control group, for compounds **6** and **7**, respectively.

The reduction of IL-6 concentration in cells caused by the studied compounds complements the previous results from cytotoxic evaluation tests. MDA-MB-231 cells were shown to be more sensitive to exposure to the studied compounds compared to the SW620 cancer cell line. The strongest effect was registered for compound **5**, which inhibited IL-6 secretion by 76%, compared to the control group.

Having shown that compounds **5**, **6**, and **7** stall the cell cycle and drive apoptosis, we next asked how these molecules might engage their protein targets at the atomic level. Therefore, the following section presents our molecular docking studies, in which each compound was fitted into the active sites of key cancer-related kinases. These *in silico* results help explain the distinct cell cycle checkpoints affected by each derivative and provide a structural basis for the antiproliferative trends observed *in vitro*.

3.4. *In Silico* Studies

3.4.1. Molecular Docking and Molecular Dynamics Simulation

Molecular docking is a bioinformatics technique used to predict the conformation of a ligand within the active site of a macromolecule and to estimate the binding affinity of the resulting protein-ligand complex. This prediction is guided by strictly defined scoring functions. It is important to emphasize that molecular docking primarily serves as a preliminary screening tool, paving the way for more complex and computationally intensive molecular dynamics simulations of the complex.

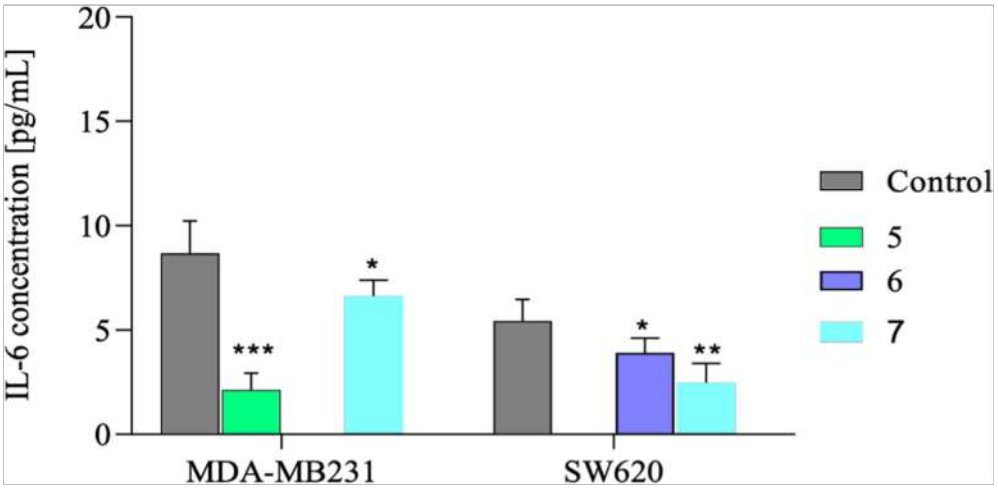


Fig. (9). The influence of novel (2-imino-4-oxo-1,3-thiazolidin-5-yl)acetic acid derivatives 5, 6, and 7 on IL-6 secretion was evaluated using ELISA in supernatants collected from human metastatic colon cancer (SW620) and human breast adenocarcinoma (MDA-MB-231) cell cultures. Results are presented as mean ± standard deviation (SD) from three independent experiments, each performed in triplicate. *** $p \leq 0.001$, ** $p \leq 0.01$, * $p \leq 0.05$, relative to the control group. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

The MDA-MB-231 cancer cell line demonstrated greater sensitivity to the studied compounds during preliminary evaluations of cytotoxicity, apoptosis induction, and cell cycle arrest. Scientific reports have shown that various types of human breast cancer, including human breast adenocarcinoma, often exhibit overexpression of the epidermal growth factor receptor (EGFR) [58]. Moreover, combined therapy of the well--known antitumor agent doxorubicin with EGFR inhibitors improved the efficacy of cytotoxic action and was identified as a synergistic interaction [59].

To determine whether the cytotoxic mechanism of the candidate compounds (**5**, **6,** and **7**) identified in the *in vitro* phase is associated with potential interference with EGFR receptor, we conducted docking experiments using five crystal structures of EGFR (1M17, 2ITO, 4I23, 4LQM, and 7AEM) [60-64], each featuring defined active pockets. These structures correspond to conformations bound with five different receptor antagonists: erlotinib, gefitinib, dacomitinib, PD-168393, and brigatinib (Fig. **10**). We evaluated the three studied molecules (compounds **5**, **6**, and **7**) against these active pockets.

This preliminary study aimed to evaluate whether compounds **5**, **6**, and **7**, later discussed as ligands, possess the biophysical potential to act as antagonists for EGFR. Such findings could provide valuable insights into their mechanism of action, which, despite existing hypotheses based on known pharmacodynamic effects, remains unconfirmed. Notably, the scoring function results for these compounds significantly exceeded those of the native ligands, as shown in Table **5**.

Through the docking results obtained from the static crystal structures, we derived not only a numerical evaluation of binding affinity but also a spatial representation of the molecular conformation. This spatial arrangement served as the initial positioning for subsequent molecular dynamics simulations. In total, we performed four simulations: three involving the target molecules within the most promising binding pocket (7AEM), and one with the same pocket bound to the native ligand.

The best-docked poses from each complex, along with the native pose of the original ligand, were used as starting points for our molecular dynamics simulations. From the generated trajectories, we analyzed and compared changes in the RMSD (root mean square deviation) of the protein's C-alpha atoms, the active site, and the entire ligand. Additionally, we evaluated the RMSF (root mean square fluctuation) of protein residues, with particular emphasis on those in the active site and the ligand atoms.

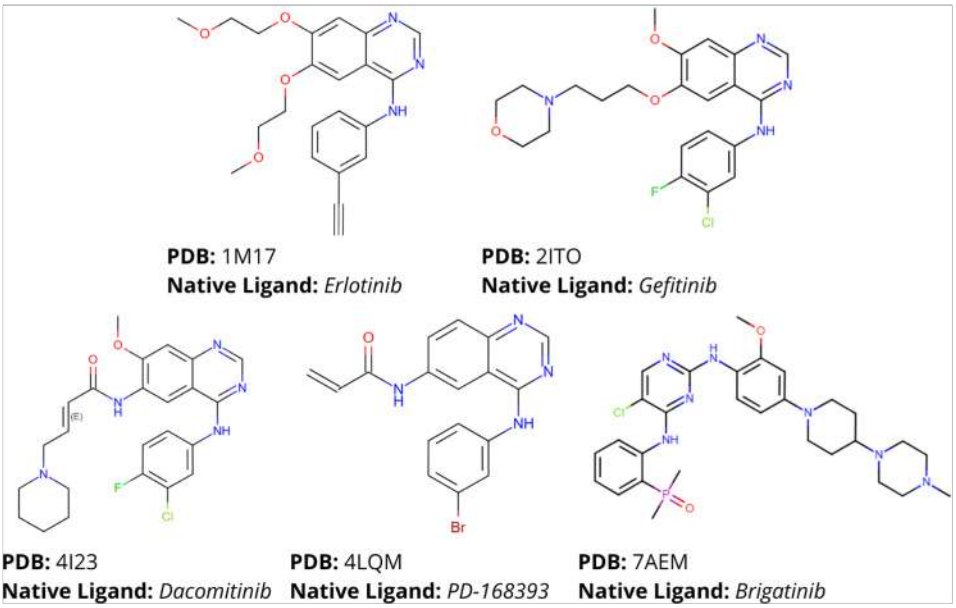


Fig. (10). Native ligands from RCSB protein data bank EGFR complexes.

Table 5. Comparison of average AutoDock SMINA results for native and studied molecules.

Compound	Crystal ID				
	1M17	**2ITO**	**4I23**	**4LQM**	**7AEM**
	Predicted Binding Score [kcal/mol] (↓)				
Native Molecule	-6.74	-7.34	-7.83	-1.51	-7.71
	10 Poses Predicted Binding Score (min|max) [kcal/mol]				
5	-8.7|-7.5	-8.8|-8.1	-8.4|-7.6	-8.7|-7.9	**-9.4**|-8.0
6	-7.6|-6.9	-7.6|-7.0	-8.6|-7.3	-8.4|-7.5	-8.7|-7.5
7	-8.4|-7.8	-8.1|-7.4	-8.7|-7.2	-8.4|-7.8	**-9.2**|-8.0

Table 6. Root mean square deviation (RMSD) of ligand, protein, and pocket in angstrom.

			Compound	
	native	**5**	**6**	**7**
Ligand RMSD (mean|std) [Å]	1.29 | 0.27	1.22 | 0.28	0.73 | 0.18	0.74 | 0.31
Protein RMSD (mean|std) [Å]	8.41 | 3.14	5.45 | 1.3	6.47 | 1.70	6.13 | 2.06
Pocket RMSD (mean|std) [Å]	0.94 | 0.15	1.72 | 0.24	1.32 | 0.24	0.70 | 0.11

The first stage of analysis involved calculating the superimposed root mean square deviation (RMSD) of atomic positions, using the first simulation frame as the reference. As shown in Fig. (**11** and **12**) and Table **6**, the RMSD of protein trajectories revealed stable mean values (~5-6 Å) with low deviations in the presence of compounds **5-7**, indicating improved global structural stability over time. In contrast, the RMSD of ligands **6** and **7** showed similar values, suggesting comparable dynamic behavior, while the native ligand and compound **5** exhibited lower average RMSD values (~1.2 Å), which may reflect difficulties in achieving a thermodynamically stable conformation. A similar trend was observed in the active site residues. Except in the presence of compound **5**, the active site RMSD values remained relatively constant, averaging below 1.5 Å.

In contrast, the RMSF analysis indicated greater flexibility in the heavy atoms of the native ligand compared to the studied molecules. This observation suggests that the tested compounds may align more effectively with the active site. The increased flexibility of the native ligand could be attributed to its larger size and the spatial constraints of the active site pocket. Consequently, larger ligands, such as the native ligand, may exhibit less favorable interactions compared to smaller molecules like compounds **5-7**.

Surprisingly, the analysis of free binding energy using two methods (Table **7**) revealed that, despite the significant mobility of the native ligand, it exhibited the most stable affinity towards the protein. In contrast, molecule **6** displayed a different performance due to its relatively constant intermolecular van der Waals interactions (Fig. **11** and **12**), which led to a significant and unfavorable increase in binding energy. A detailed analysis of the trajectory movements of the native ligand and molecule **6** complexes, visualized in Fig. (**13**), highlights that the stability of the native ligand arises from its incomplete embedding within the active pocket. For the native ligand, only a portion of its atoms interacts with the pocket, while the remaining atoms, responsible for high fluctuations, rotate freely at a greater distance from the protein surface.

Table 7. ΔMM/PBSA and ΔMM/GBSA total energy comparison between native and studied molecules.

			Compound	
	Native	**5**	**6**	**7**
ΔMM/PBSA	-34.42 | 0.82	94.22 | 3.40	1584.83 | 149.11	562.43 | 84.10
ΔMM/GBSA	-41.84 | 0.38	72.15 | 3.20	1554.31 | 149.25	536.40 | 84.36

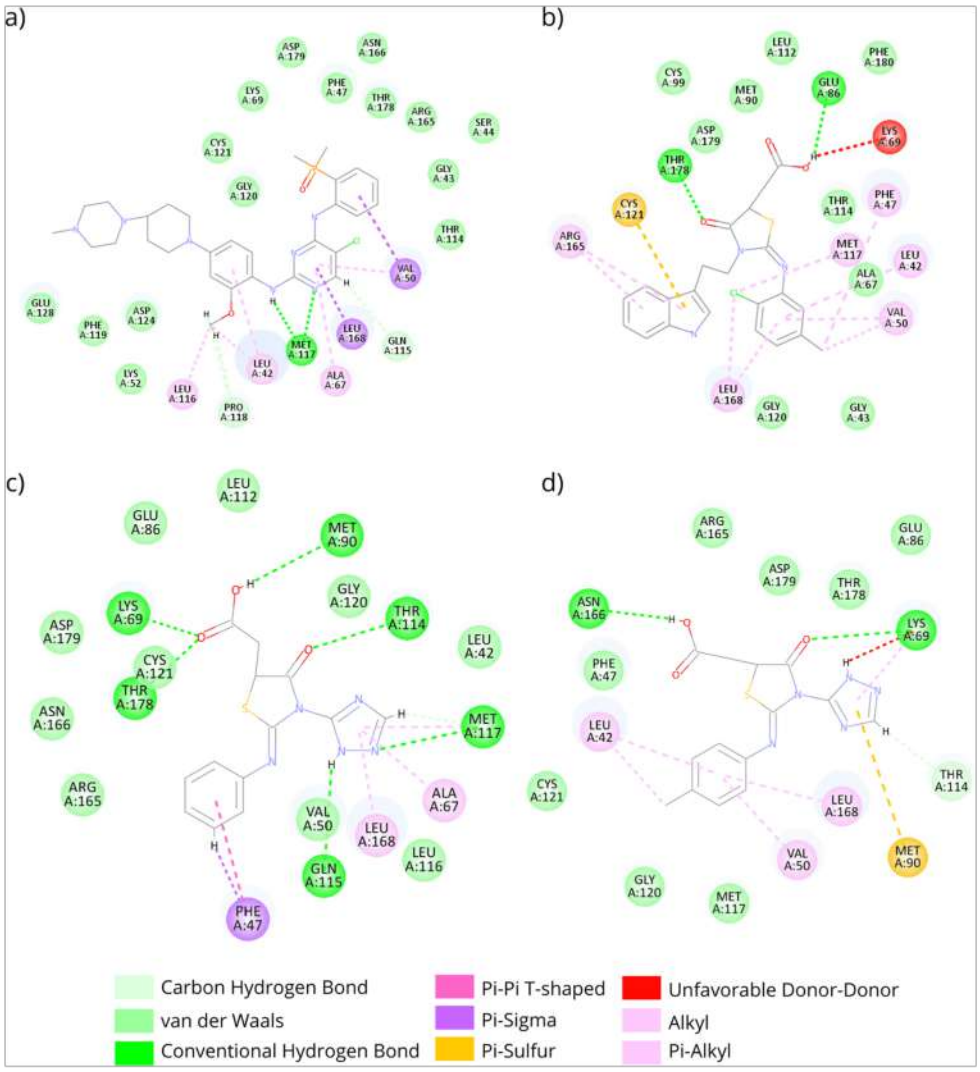


Fig. (11). A visualization of intermolecular interactions for EGFR protein complexes with: (**a**) the native molecule, (**b**) compound **5**, (**c**) compound **6**, and (**d**) compound **7**.

Molecular docking and dynamics studies were complementary to previous results obtained from the *in vitro* phase of the research and revealed that compounds **5**, **6**, and **7** exhibited the strongest binding to the 7AEM structure of EGFR, which represents the active kinase domain in complex with brigatinib. As EGFR activation triggers both PI3K/AKT and RAS-MAPK signaling cascades, it is likely that our compounds may affect both pathways by interfering with ATP binding at the kinase site. This hypothesis aligns with known mechanisms of EGFR tyrosine kinase inhibitors and their impact on downstream signaling in EGFR-positive tumors [65]. EGFR is commonly overexpressed in many types of human breast cancer, which indicates that an inhibitory effect on this factor's expression and action is desirable and can explain the cytotoxic effect on MDA-MB-231 cancer cells [58]. Among all analyzed structures, molecule **5** was identified as the most promising candidate. It demonstrated the most similar binding affinity to the native li-

gand, which represents well-known antagonists of the EGFR. In contrast, the least energetically stable complex was observed for compound **6**, which did not exhibit cytotoxic activity against the MDA-MB-231 cancer cell line. This effect was instead noted in SW620 cancer cells, which have low EGFR expression, making therapies targeting this receptor largely ineffective [66].

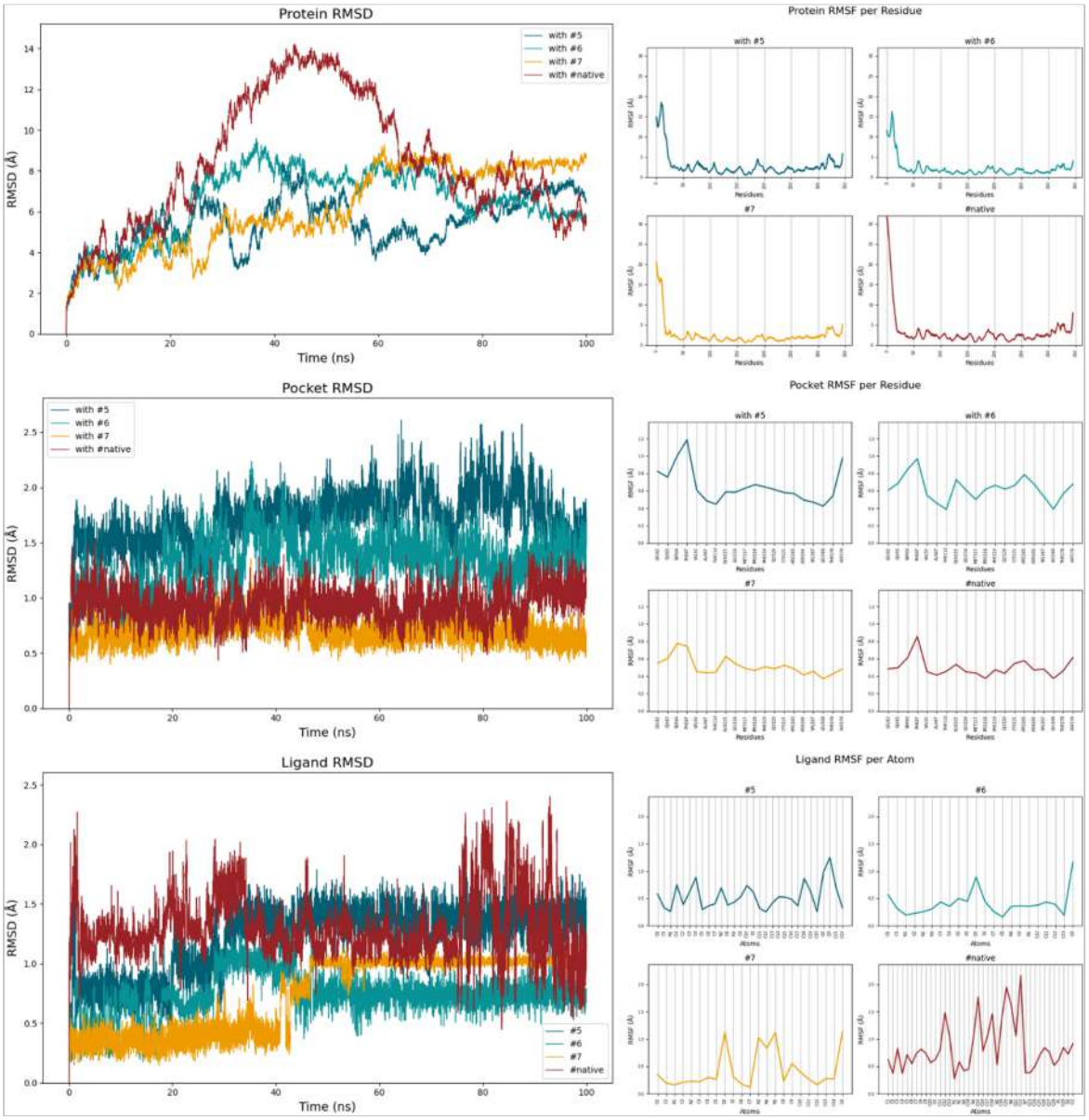


Fig. (12). On the left, from top to bottom, are comparisons of RMSD for proteins, pockets, and ligands derived from Molecular Dynamics simulations. On the right are RMSF plots, corresponding respectively to protein amino acids, pocket amino acids, and ligand atoms. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

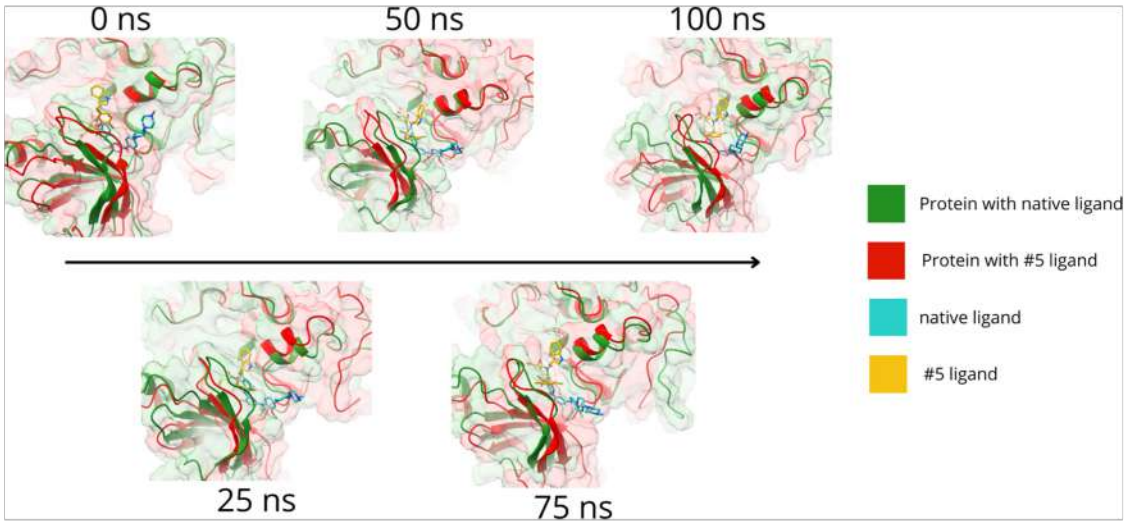


Fig. (13). A visualization comparing the shape of the protein-ligand complex between the native protein-molecule complex and the protein-molecule 5 complex at 0 ns, 25 ns, 50 ns, 75 ns, and 100ns. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

3.4.2. ADMET

Computer science and its various *in silico* models are increasingly used in the drug design process, assisting in the evaluation of newly discovered compounds. The integration of traditional scientific knowledge with the high computational power of artificial intelligence (AI) models helps to reduce costs and risks associated with preclinical and clinical trials, while also accelerating the drug discovery process [67, 68].

To assess the potential of the newly obtained (2-imino-4-oxo-1,3-thiazolidin-5-yl)acetic acid derivatives, we utilized the new ADMET-AI [46] model to predict key drug-like parameters and compared them to those of the reference antitumor drug etoposide [69], using standard statistical tools.

As key drug-like parameters, we selected LogP, bioavailability, quantitative estimate of drug-likeness (QED), and topological polar surface area (TPSA):

- LogP: A measure of a compound's hydrophobicity, reflecting its ability to cross cell membranes. It helps predict a drug's permeability, influencing its absorption and distribution [70].
- Bioavailability: Refers to the proportion of the drug that reaches systemic circulation in an active form after administration. It's critical for balancing efficacy, safety, and patient compliance [71].
- Quantitative Estimate of Drug-likeness (QED): This metric quantifies how similar a compound is to known drugs by considering properties such as molecular weight, solubility, and lipophilicity, helping prioritize candidates during drug development [72]
- Topological Polar Surface Area (TPSA): A molecular descriptor that estimates a compound's polarity, which in turn affects its absorption, distribution, and permeability through biological membranes [73].

After prediction, the obtained data were standardized for better visualization on the heatmap and submitted to hierarchical cluster analysis, displayed as a dendrogram. As depicted in Fig. (**14**), the studied compounds formed clusters based on the similarities in their predicted parameter values. The closer and lower the branches are in the hierarchical clustering, the greater the similarity between the compounds. Particularly interesting are the placements of compounds **6** and **7,** which tend to be closer to and connected with the reference drug etoposide, despite the higher level of the branch. The connection between compounds at higher levels of the branches in the cluster may suggest that they share similarities in major parameters, with some exceptions. Compounds **6** and **7** registered the same QED and TPSA scores, 0.95 and 1.0, respectively, compared to etoposide, with scores of 0.73 and 0.77. The LogP values predicted for derivatives **6** and

7 were at the levels of 0.31 and 0.37, respectively, compared to 0.54 for Etoposide (see Fig. **15**). The most significant difference between the studied compounds and the reference drug was observed in bioavailability, with a score difference of 0.59, which may account for the height of the cluster branch.

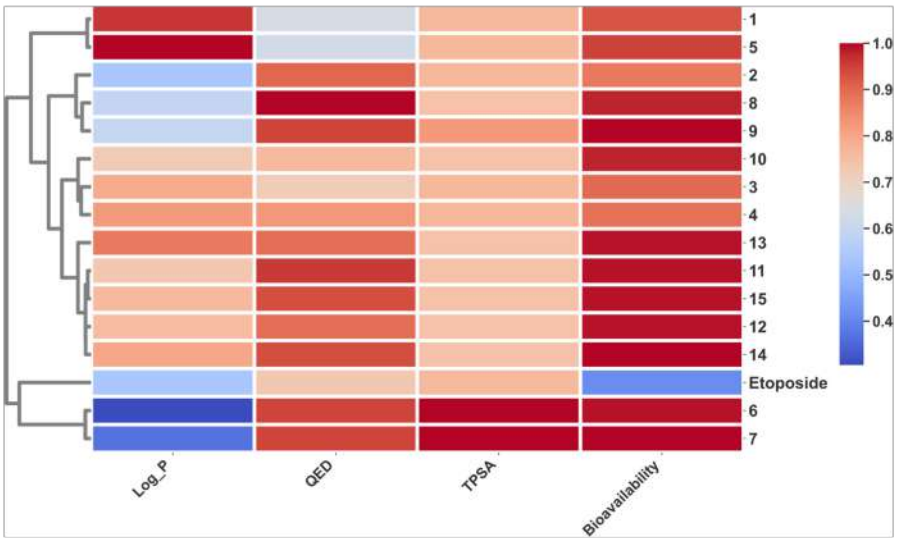


Fig. (14). A heatmap chart with hierarchical cluster analysis of predicted and standardized key drug-like parameters- logP, bioavailability, Quantitative Estimate of Drug-likeness (QED), and Topological Polar Surface Area (TPSA), for all novel (2-imino-4-oxo-1,3-thiazolidin-5-yl)acetic acid derivatives, compared to the reference drug Etoposide, and performed by ADMET-AI model. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

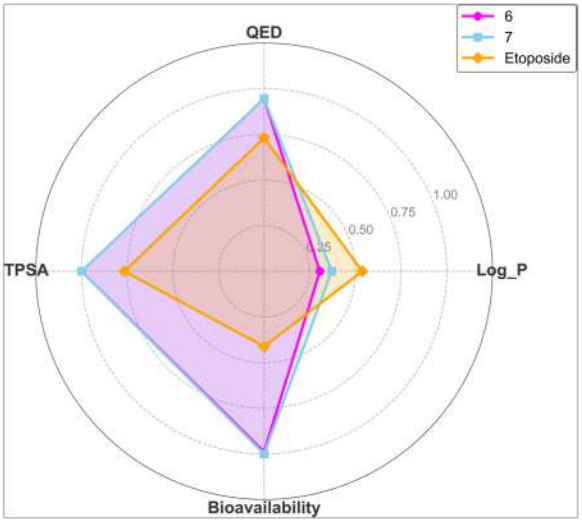


Fig. (15). A radar chart of compounds **6** and **7**, comparing predicted and standardized values of QED, logP, bioavailability, and TPSA to the reference drug Etoposide. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

The results of *in silico* studies suggest that compounds **6** and **7** may have potential drug-like properties, due to their similarities to the reference antitumor medicine, Etoposide, in key parameters crucial for drug design. Moreover, the high QED and TPSA scores further emphasize their drug-likeness, reflecting structural and chemical similarities to known medicines, along with potentially favorable surface polarity that could enhance solubility and permeability. The height of the cluster branch, representing dissimilarities between compounds and the reference drug, is primarily due to significant differences in bioavailability. Compounds **6** and **7** achieved the highest possible bioavailability scores, indicating potentially better administration and absorption in the human body, compared to etoposide.

CONCLUSION

The present work has established a coherent structure-activity narrative for a new family of fifteen (2-imino-4-oxo-1,3-thiazolidin-5-yl)acetic-acid derivatives and has identified two particularly promising molecules, compounds **5** and **7,** as credible leads for the development of anticancer drugs. Their superiority was first demonstrated in cytotoxic screening, where both showed low, micromolar IC_{50} values and the highest selectivity indices (2.42 - 4.73) in the panel, and was then reinforced by mechanistic assays. Flow-cytometry analysis revealed that compound **7** imposes a G0/G1 blockade in the highly EGFR-positive MDA-MB-231 line, whereas compound **5** accumulates cells in G2/M. In both cases, the arrest was accompanied by a significant rise in the sub-G1 fraction, confirming a shift toward apoptosis. Consistent with these findings, MDA-MB-231 cells proved more vulnerable than SW620 cells, and compound **7** produced the strongest pro-apoptotic effect, driving 6.60 ± 0.69% of the breast-cancer population into late apoptosis or necrosis.

Biochemical assays complemented the cell cycle data. Compound **5** suppressed IL-6 secretion in MDA-MB-231 cultures by 76%, linking its antiproliferative activity to the reduction of pro-inflammatory, tumor-promoting signaling. At the molecular level, docking experiments showed that this compound forms an energetically favorable complex with the EGFR kinase domain, surpassing the native ligand in calculated binding affinity. The 100 ns molecular dynamics simulations further demonstrated that the complex remains conformationally stable throughout the trajectory. Compound **7** distinguished itself *in silico* by combining equally strong docking scores with excellent bioavailability and drug-likeness parameters, its physicochemical profile closely matching that of the clinical reference etoposide.

Despite these encouraging *in vitro* and *in silico* results, several limitations should be acknowledged. This preliminary study provides compelling data on synthesized derivatives, but more advanced testing is needed to fully assess their therapeutic value. First, a broader panel of cancer cell lines will be included in future *in vitro* tests, involving the most potent compounds **5** and **7**. Second, only two triazole hybrids were evaluated due to solubility constraints and reaction conditions limitations. Alternative synthetic strategies will be employed to expand this series. Building on these encouraging results, future work will focus on broadening the scope of biological evaluation, including ADMET characterization, *in vivo* validation, and synthetic refinement to fully explore the clinical promise of this thiazolidinone series. Combination strategies involving the lead compounds may further enhance their therapeutic relevance.

Taken together, these findings validate (2-imino-4-oxo-1,3-thiazolidin-5-yl)acetic acid as a privileged core for cytotoxic lead generation and indicate that strategic substitution with either a tryptamine moiety (compound **5**) or a 1,2,4-triazole ring (compound **7**) is particularly advantageous. Compared with earlier reports on related thiazolidinone scaffolds [74], the present derivatives exhibit improved potency, clearer selectivity, and mechanistic features consistent with EGFR inhibition and cytokine modulation. On this foundation, future studies will focus on (i) Broadening the chemical space around the tryptamine and triazole vectors to generate a second-generation library, (ii) Refining synthetic protocols, especially for the hybrid 4-thiazolidinone and 1,2,4-triazole framework to improve yield and facilitate late-stage diversification, and (iii) Extending the evaluation to pharmacokinetic profiling and *in vivo* efficacy models. By coupling iterative scaffold optimization with targeted kinase-panel screening and metabolic stability assays, we aim to produce thiazolidinone-based candidates exhibiting nanomolar potency, robust selectivity, and a safety profile suitable for preclinical advancement.

AUTHORS' CONTRIBUTIONS

The authors confirm their contribution to the paper as follows: TS.; writing- original draft, formal analysis. D.O.-Ś; supervision, project administration. MW.; methodology, software. MAD.; validation. PR.; supervision, data curation. D.S.; writing- review and editing, supervision, and conceptualization. All authors reviewed the results and approved the final version of the manuscript.

LIST OF ABBREVIATIONS

ADME	= Absorption, Distribution, Metabolism, and Excretion
ADMET	= Absorption, Distribution, Metabolism, Excretion and Toxicity
AI	= Artificial Intelligence
ATP	= Adenosine-5′-triphosphate
COX-2	= Cyclo-oxygenase-2
DFT	= Density-functional Theory
DMSO	= Dimethyl Sulfoxide
EGFR	= Epidermal Growth-factor Receptor
ELISA	= Enzyme-linked Immunosorbent Assay
ESI	= Electrospray Ionisation
FACS	= Fluorescence-activated Cell Sorting
FITC	= Fluorescein Isothiocyanate
GBSA	= Generalized Born Surface-area (energy)
GI%	= Growth-inhibition Percentage
HRMS	= High-resolution Mass Spectrometry
IC_{50}	= Half-maximal Inhibitory Concentration
IL-6	= interleukin-6
LogP	= Logarithm of the n-octanol/Water Partition Coefficient
MAPK	= Mitogen-activated Protein Kinase
MD	= Molecular Dynamics
MM/GBSA	= Molecular-mechanics/Generalised Born Surface-area (energy)
MM/PBSA	= Molecular-mechanics/Poisson-Boltzmann Surface-area (energy)
MTT	= 3-(4,5-dimethylthiazol-2-yl)-2, 5-diphenyltetrazolium Bromide
NMR	= Nuclear-magnetic Resonance
PBS	= Phosphate-buffered Saline
PI	= Propidium Iodide
PIM	= Proviral-integration-site for Moloney-murine-leukemia Virus Kinase
PI3K	= Phosphatidylinositol-3-kinase
QED	= Quantitative Estimate of Drug-likeness
RAS	= Rat Sarcoma (GTPase)
RMSD	= Root-mean-square Deviation
RMSF	= Root-mean-square Fluctuation
ROS	= Reactive Oxygen Species
SI	= Selectivity Index
STAT3	= Signal Transducer and Activator of Transcription 3
TPSA	= Topological Polar Surface Area
VEGFR	= Vascular Endothelial Growth-factor receptor
5-LOX	= 5-lipoxygenase

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

HUMAN AND ANIMAL RIGHTS

Not applicable.

CONSENT FOR PUBLICATION

Not applicable.

AVAILABILITY OF DATA AND MATERIALS

Detailed experimental data supporting the findings of this study are available upon request.

FUNDING

This work was supported by the internal grant of the Medical University of Warsaw (1WK/2/M/M-B/N/23).

CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENTS

MW's work has been supported by the Polish National Science Centre (2020/37/B/NZ2/03757). Computations were performed thanks to the Laboratory of Bioinformatics and Computational Genomics, Faculty of Mathematics and Information Science, Warsaw University of Technology, using the Artificial Intelligence HPC platform financed by the Polish Ministry of Science and Higher Education (decision no. 7054/I-A/SP/2020 of 2020-08-28).

SUPPLEMENTARY MATERIAL

Supplementary material is available on the publisher's website along with the published article.

REFERENCES

[1] Siegel, R.L.; Miller, K.D.; Jemal, A. Cancer statistics, 2019. *CA Cancer J. Clin.,* **2019**, *69*(1), 7-34. http://dx.doi.org/10.3322/caac.21551 PMID: 30620402

[2] Mattiuzzi, C.; Lippi, G. Current Cancer Epidemiology. *J. Epidemiol. Glob. Health,* **2019**, *9*(4), 217-222. http://dx.doi.org/10.2991/jegh.k.191008.001 PMID: 31854162

[3] Basak, D.; Arrighi, S.; Darwiche, Y.; Deb, S. Comparison of Anticancer drug toxicities: paradigm shift in adverse effect profile. *Life,* **2021**, *12*(1), 48. http://dx.doi.org/10.3390/life12010048 PMID: 35054441

[4] Nirwan, S.; Chahal, V.; Kakkar, R. Thiazolidinones: Synthesis, reactivity, and their biological applications. *J. Heterocycl. Chem.,* **2019**, *56*(4), 1239-1253. http://dx.doi.org/10.1002/jhet.3514

[5] DeFronzo, R.A.; Inzucchi, S.; Abdul-Ghani, M.; Nissen, S.E. Pioglitazone: The forgotten, cost-effective cardioprotective drug for type 2 diabetes. *Diab. Vasc. Dis. Res.,* **2019**, *16*(2), 133-143. http://dx.doi.org/10.1177/1479164118825376 PMID: 30706731

[6] Wagstaff, A.J.; Goa, K.L. Spotlight on rosiglitazone in the management of type 2 diabetes mellitus. *Treat. Endocrinol.,* **2002**, *1*(6), 411-414. http://dx.doi.org/10.2165/00024677-200201060-00006 PMID: 15832493

[7] Ramirez, M.A.; Borja, N.L. Epalrestat: an aldose reductase inhibitor for the treatment of diabetic neuropathy. *Pharmacotherapy,* **2008**, *28*(5), 646-655. http://dx.doi.org/10.1592/phco.28.5.646 PMID: 18447661

[8] McLoughlin, C. Diuretic drugs. *Drugs,* **1986**, *31*(Suppl 4(2)), 50-51. http://dx.doi.org/10.2165/00003495-198600314-00006

[9] Fischer, W.; Bodewei, R.; Satzinger, G. Anticonvulsant and sodium channel blocking effects of ralitoline in different screening models. *Naunyn Schmiedebergs Arch. Pharmacol.,* **1992**, *346*(4), 442-452. http://dx.doi.org/10.1007/BF00171088 PMID: 1331817

[10] Markham, A. Ponesimod: First Approval. *Drugs,* **2021**, *81*(8), 957-962. http://dx.doi.org/10.1007/s40265-021-01523-z PMID: 33939119

[11] Türe, A.; Ergül, M.; Ergül, M.; Altun, A.; Küçükgüzel, İ. Design, synthesis, and anticancer activity of novel 4-thiazolidinone-phenylaminopyrimidine hybrids. *Mol. Divers.,* **2021**, *25*(2), 1025-1050. http://dx.doi.org/10.1007/s11030-020-10087-1 PMID: 32328961

[12] Szychowski, K.A.; Leja, M.L.; Kaminskyy, D.V.; Binduga, U.E.; Pinyazhko, O.R.; Lesyk, R.B.; Gmiński, J. Study of novel anticancer 4-thiazolidinone derivatives. *Chem. Biol. Interact.,* **2017**, *262*, 46-56. http://dx.doi.org/10.1016/j.cbi.2016.12.008 PMID: 27965178

[13] Amin, N.H.; El-Saadi, M.T.; Abdel-Fattah, M.M.; Mohammed, A.A.; Said, E.G. Development of certain aminoquinazoline scaffolds as potential multitarget anticancer agents with apoptotic and anti-proliferative effects: Design, synthesis and biological evaluation. *Bioorg. Chem.,* **2023**, *135*, 106496. http://dx.doi.org/10.1016/j.bioorg.2023.106496 PMID: 36989735

[14] Nossier, E.; El-Hallouty, S.; Zaki, E. Synthesis, anticancer evaluation and molecular modeling of some substituted thiazolidinonyl and thiazolyl pyrazole derivatives. *Int. J. Pharm. Pharm. Sci.,* **2015**.

[15] Chen, L.; Mao, W.; Ren, C.; Li, J.; Zhang, J. Comprehensive insights that targeting pim for cancer therapy: prospects and obstacles. *J. Med. Chem.,* **2024**, *67*(1), 38-64. http://dx.doi.org/10.1021/acs.jmedchem.3c01802 PMID: 38164076

[16] Konno, T.; Watanabe, J.; Ishihara, K. Enhanced solubility of paclitaxel using water-soluble and biocompatible 2-methacryloyloxyethyl phosphorylcholine polymers. *J. Biomed. Mater. Res. A,* **2003**, *65A*(2), 209-214. http://dx.doi.org/10.1002/jbm.a.10481 PMID: 12734814

[17] Yudaev, P.; Tupikov, A.; Chistyakov, E. Organocyclophosphazenes and materials based on them for pharmaceuticals and biomedicine. *Biomolecules,* **2025**, *15*(2), 262. http://dx.doi.org/10.3390/biom15020262 PMID: 40001565

[18] Abbot, V.; Sharma, P.; Dhiman, S.; Noolvi, M. N.; Patel, H. M.; Bhardwaj, V. Small hybrid heteroaromatics: resourceful biological tools in cancer research. *RSC Advances,* **2017**, *7*, 19857-19887. http://dx.doi.org/10.1039/c6ra24662a

[19] Gul, S.; Jan, F.; Alam, A.; Shakoor, A.; Khan, A.; AlAsmari, A.F.; Alasmari, F.; Khan, M.; Bo, L. Synthesis, molecular docking and DFT analysis of novel *bis*-Schiff base derivatives with thiobarbituric acid for α-glucosidase inhibition assessment. *Sci. Rep.,* **2024**, *14*(1), 3419. http://dx.doi.org/10.1038/s41598-024-54021-z PMID: 38341468

[20] Gul, S.; Maab, S.; Rafiq, H.; Alam, A.; Rehman, M.U.; Assad, M.; AlAsmari, A.F.; Alasmari, F.; Ibrahim, M.; Khan, M. Exploring bis-Schiff Bases with Thiobarbiturate Scaffold: *In Vitro* Urease Inhibition, Antioxidant Properties, and *In Silico* Studies. *Russ. J. Bioorganic Chem.,* **2024**, *50*(5), 1627-1638. http://dx.doi.org/10.1134/S1068162024050030

[21] Shakoor, A.; Alam, A.; Jan, F.; Khan, M.; Ali, M.; Ullah, S.; Khan, A.; AlAsmari, A.F.; Alasmari, F.; Al-Ghafri, A.; Al-Harrasi, A. Novel benzimidazole derivatives as effective inhibitors of prolyl oligopeptidase: synthesis, *in vitro and in silico* analysis. *Future Med. Chem.,* **2024**, *16*(1), 43-58. http://dx.doi.org/10.4155/fmc-2023-0267 PMID: 38054466

[22] Shakoor, A.; Alam, A.; Ali, M.; Ullah, S.; Halim, S.A.; Rehman, M.U.; AlAsmari, A.F.; Alasmari, F.; Khan, A.; Khan, M.; Al-Harrasi, A. Synthesis of novel benzimidazole analogs for neurodegenerative diseases by targeting prolyl oligopeptidase. *ChemistrySelect,* **2024**, *9*(34), e202401904. http://dx.doi.org/10.1002/slct.202401904

[23] Shakoor, A.; Jan, F.; Rahman, S.; Ali, M.; Ibrahim, M.; Khan, H.; Alam, A.; Khan, A.; Ali, A.; Al-Olayan, E.; Abukhadra, M.R.; Al-Harrasi, A.; Khan, M. Synthesis, urease inhibitory activity, molecular docking, dynamics, mmgbsa and dft studies of hydrazone-schiff bases bearing benzimidazole scaffold. *Chem. Biodivers.,* **2025**, *22*(3), e202402096.

http://dx.doi.org/10.1002/cbdv.202402096 PMID: 39530214

[24] Ahmad, S.; Khan, M.; Alam, A.; Ajmal, A.; Wadood, A.; Khan, A.; AlAsmari, A.F.; Alharbi, M.; Alshammari, A.; Shakoor, A. Novel flurbiprofen clubbed oxadiazole derivatives as potential urease inhibitors and their molecular docking study. *RSC Advances,* **2023**, *13*(37), 25717-25728. http://dx.doi.org/10.1039/D3RA03841F PMID: 37649663

[25] Al-Saleem, M.; Mohamed Ahmed, M.; Riyadh, S.; Alruwaili, A.; Zaki, M.; Gomha, S. Synthesis, molecular characterization, and antimicrobial evaluation of hydrazones derived from 2-hydrazinobenzimidazole. *Drug Des. Devel. Ther.,* **2025**, *19*, 4437-4456. http://dx.doi.org/10.2147/DDDT.S521541 PMID: 40453211

[26] Salikov, R.F.; Belyy, A.Yu.; Khusnutdinova, N.S.; Vakhitova, Y. Synthesis and cytotoxic properties of tryptamine derivatives. *Bioorganic & Medicinal Chemistry Letters,* **2015**, *25*(17), 3597-3600. http://dx.doi.org/10.1016/j.bmcl.2015.06.070

[27] Simonetti, G.; Boga, C.; Durante, J.; Micheletti, G.; Telese, D.; Caruana, P.; Ghelli Luserna di Rorà, A.; Mantellini, F.; Bruno, S.; Martinelli, G.; Calonghi, N. Synthesis of novel tryptamine derivatives and their biological activity as antitumor agents. *Molecules,* **2021**, *26*(3), 683. http://dx.doi.org/10.3390/molecules26030683 PMID: 33525621

[28] Aggarwal, R.; Sumran, G. An insight on medicinal attributes of 1,2,4-triazoles. *Eur. J. Med. Chem.,* **2020**, *205*, 112652.https://doi.org/https://doi.org/10.1016/j.ejmech.2020.112652 http://dx.doi.org/10.1016/j.ejmech.2020.112652 PMID: 32771798

[29] Cohen, M.H.; Johnson, J.R.; Justice, R.; Pazdur, R. Approval summary: letrozole (Femara® tablets) for adjuvant and extended adjuvant postmenopausal breast cancer treatment: conversion of accelerated to full approval. *Oncologist,* **2011**, *16*(12), 1762-1770. http://dx.doi.org/10.1634/theoncologist.2011-0287 PMID: 22089970

[30] Guan, Q.; Gao, Z.; Chen, Y.; Guo, C.; Chen, Y.; Sun, H. Structural modification strategies of triazoles in anticancer drug development. *Eur. J. Med. Chem.,* **2024**, *275*, 116578. http://dx.doi.org/10.1016/j.ejmech.2024.116578 PMID: 38889607

[31] Jain, S.; Pattnaik, S.; Pathak, K.; Kumar, S.; Pathak, D.; Jain, S.; Vaidya, A. Anticancer potential of thiazole derivatives: a retrospective review. *Mini Rev. Med. Chem.,* **2018**, *18*(8), 640-655. http://dx.doi.org/10.2174/1389557517666171123211321 PMID: 29173166

[32] de Santana, T.I.; Barbosa, M.O.; Gomes, P.A.T.M.; da Cruz, A.C.N.; da Silva, T.G.; Leite, A.C.L. Synthesis, anticancer activity and mechanism of action of new thiazole derivatives. *Eur. J. Med. Chem.,* **2018**, *144*, 874-886. http://dx.doi.org/10.1016/j.ejmech.2017.12.040 PMID: 29329071

[33] Jędrzejczyk, M.; Morabito, B.; Żyżyńska-Granica, B.; Struga, M.; Janczak, J.; Aminpour, M.; Tuszynski, J.A.; Huczyński, A. Novel combretastatin a-4 analogs—design, synthesis, and antiproliferative and anti-tubulin activity. *Molecules,* **2024**, *29*(10), 2200. http://dx.doi.org/10.3390/molecules29102200 PMID: 38792062

[34] Jeleń, M.; Otto-Ślusarczyk, D.; Morak-Młodawska, B.; Struga, M. Novel tetracyclic azaphenothiazines with the quinoline ring as new anticancer and antibacterial derivatives of chlorpromazine. *Int. J. Mol. Sci.,* **2024**, *25*(8), 4148. http://dx.doi.org/10.3390/ijms25084148 PMID: 38673734

[35] Szulczyk, D.; Bielenica, A.; Roszkowski, P.; Dobrowolski, M.A.; Olejarz, W.; Kmiecik, S.; Podsiad, M.; Struga, M. Synthetic transition from thiourea-based compounds to tetrazole derivatives: Structure and biological evaluation of synthesized new N-(furan-2-ylmethyl)-1*H*-tetrazol-5-amine derivatives. *Molecules,* **2021**, *26*(2), 323. http://dx.doi.org/10.3390/molecules26020323 PMID: 33435194

[36] Szostek, T.; Szulczyk, D.; Szymańska-Majchrzak, J.; Koliński, M.; Kmiecik, S.; Otto-Ślusarczyk, D.; Zawodnik, A.; Rajkowska, E.; Chaniewicz, K.; Struga, M.; Roszkowski, P. Design and synthesis of menthol and thymol derived ciprofloxacin: influence of structural modifications on the antibacterial activity and anticancer properties. *Int. J. Mol. Sci.,* **2022**, *23*(12), 6600. http://dx.doi.org/10.3390/ijms23126600 PMID: 35743043

[37] Dolomanov, O.V.; Bourhis, L.J.; Gildea, R.J.; Howard, J.A.K.; Puschmann, H. *OLEX2* : a complete structure solution, refinement and analysis program. *J. Appl. Cryst.,* **2009**, *42*(2), 339-341. http://dx.doi.org/10.1107/S0021889808042726

[38] Koes, D.R.; Baumgartner, M.P.; Camacho, C.J. Lessons learned in empirical scoring with smina from the CSAR 2011 benchmarking exercise. *J. Chem. Inf. Model.,* **2013**, *53*(8), 1893-1904. http://dx.doi.org/10.1021/ci300604z PMID: 23379370

[39] Eastman, P.; Swails, J.; Chodera, J.D.; McGibbon, R.T.; Zhao, Y.; Beauchamp, K.A.; Wang, L.P.; Simmonett, A.C.; Harrigan, M.P.; Stern, C.D.; Wiewiora, R.P.; Brooks, B.R.; Pande, V.S. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLOS Comput. Biol.,* **2017**, *13*(7), e1005659. http://dx.doi.org/10.1371/journal.pcbi.1005659 PMID: 28746339

[40] Eswar, N.; John, B.; Mirkovic, N.; Fiser, A.; Ilyin, V.A.; Pieper, U.; Stuart, A.C.; Marti-Renom, M.A.; Madhusudhan, M.S.; Yerkovich, B.; Sali, A. Tools for comparative protein structure modeling and analysis. *Nucleic Acids Res.,* **2003**, *31*(13), 3375-3380. http://dx.doi.org/10.1093/nar/gkg543 PMID: 12824331

[41] Maier, J.A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K.E.; Simmerling, C. Ff14SB: Improving the accuracy of protein side chain and backbone parameters from ff99SB. *J. Chem. Theory Comput.,* **2015**, *11*(8), 3696-3713. http://dx.doi.org/10.1021/acs.jctc.5b00255 PMID: 26574453

[42] Roe, D.R.; Cheatham, T.E., III PTRAJ and CPPTRAJ: Software for processing and analysis of molecular dynamics trajectory data. *J. Chem. Theory Comput.,* **2013**, *9*(7), 3084-3095. http://dx.doi.org/10.1021/ct400341p PMID: 26583988

[43] Valdés-Tresanco, M.S.; Valdés-Tresanco, M.E.; Valiente, P.A.; Moreno, E. Gmx_MMPBSA: A new tool to perform end-state free energy calculations with GROMACS. *J. Chem. Theory Comput.,* **2021**, *17*(10), 6281-6291. http://dx.doi.org/10.1021/acs.jctc.1c00645 PMID: 34586825

[44] Wang, J.; Cai, Q.; Xiang, Y.; Luo, R. Reducing grid-dependence in finite-difference Poisson-Boltzmann calculations. *J. Chem. Theory Comput.,* **2012**, *8*(8), 2741-2751. http://dx.doi.org/10.1021/ct300341d PMID: 23185142

[45] Onufriev, A.; Bashford, D.; Case, D.A. Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins,* **2004**, *55*(2), 383-394. http://dx.doi.org/10.1002/prot.20033 PMID: 15048829

[46] Swanson, K.; Walther, P.; Leitz, J.; Mukherjee, S.; Wu, J.C.; Shivnaraine, R.V.; Zou, J. ADMET-AI: a machine learning ADMET platform for evaluation of large-scale chemical libraries. *Bioinformatics,* **2024**, *40*(7), btae416. http://dx.doi.org/10.1093/bioinformatics/btae416 PMID: 38913862

[47] Pankova, A.S.; Samartsev, M.A.; Shulgin, I.A.; Golubev, P.R.; Avdontceva, M.S.; Kuznetsov, M.A. Synthesis of thiazolidines *via* regioselective addition of unsymmetric thioureas to maleic acid derivatives. *RSC Advances,* **2014**, *4*(93), 51780-51786. http://dx.doi.org/10.1039/C4RA07840C

[48] Pankova, A. S.; Golubev, P. R.; Khlebnikov, A. F.; Ivanov, A. Y.; Kuznetsov, M. A. Thiazol-4-one derivatives from the reaction of monosubstituted thioureas with maleimides: structures and factors determining the selectivity and tautomeric equilibrium in solution. *Beilstein J Org Chem,* **2016**, *12*, 2563-2569. http://dx.doi.org/10.3762/BJOC.12.251

[49] Achoui, N.; Zaioua, K.; Hammoutène, D.; Kolli-Nedjar, B.; Akacem, Y. Interaction of thiourea and urea with maleimide: comparative theoretical DFT study. *Heliyon,* **2019**, *5*(9), e02330. http://dx.doi.org/10.1016/j.heliyon.2019.e02330 PMID: 31687533

[50] Wang, Z. Cell cycle progression and synchronization: an overview. *Methods Mol. Biol.,* **2022**, *2579*, 3-23. http://dx.doi.org/10.1007/978-1-0716-2736-5_1 PMID: 36045194

[51] Jamasbi, E.; Hamelian, M.; Hossain, M.A.; Varmira, K. The cell cycle, cancer development and therapy. *Mol. Biol. Rep.,* **2022**, *49*(11), 10875-10883. http://dx.doi.org/10.1007/s11033-022-07788-1 PMID: 35931874

[52] Evan, G.I.; Vousden, K.H. Proliferation, cell cycle and apoptosis in cancer. *Nature,* **2001**, *411*(6835), 342-348. http://dx.doi.org/10.1038/35077213 PMID: 11357141

[53] Vakili-Samiani, S.; Khanghah, O.J.; Gholipour, E.; Najafi, F.; Zeinalzadeh, E.; Samadi, P.; Sarvarian, P.; Pourvahdani, S.; Kelaye, S.K.; Hamblin, M.R.; Feizi, A.A.H. Cell cycle involvement in cancer therapy; WEE1 kinase, a potential target as therapeutic strategy. *Mutat. Res.,* **2022**, *824*, 111776. http://dx.doi.org/10.1016/j.mrfmmm.2022.111776 PMID: 35247630

[54] Kang, S.; Narazaki, M.; Metwally, H.; Kishimoto, T. Historical overview of the interleukin-6 family cytokine. *J. Exp. Med.,* **2020**, *217*(5), e20190347. http://dx.doi.org/10.1084/jem.20190347

[55] Johnson, D.E.; O'Keefe, R.A.; Grandis, J.R. Targeting the IL-6/JAK/STAT3 signalling axis in cancer. *Nat. Rev. Clin. Oncol.,* **2018**, *15*(4), 234-248. http://dx.doi.org/10.1038/nrclinonc.2018.8 PMID: 29405201

[56] Moran, D.M.; Mattocks, M.A.; Cahill, P.A.; Koniaris, L.G.; McKillop, I.H. Interleukin-6 mediates G(0)/G(1) growth arrest in hepatocellular carcinoma through a STAT 3-dependent pathway. *J. Surg. Res.,* **2008**, *147*(1), 23-33. http://dx.doi.org/10.1016/j.jss.2007.04.022 PMID: 17574577

[57] Kumari, N.; Dwarakanath, B. S.; Das, A.; Bhatt, A. N. Role of interleukin-6 in cancer progression and therapeutic resistance. *Tumour Biol,* **2016**, *37*(9), 11553-11572. http://dx.doi.org/10.1007/S13277-016-5098-7

[58] Hossein-Nejad-Ariani, H.; Althagafi, E.; Kaur, K. Small peptide ligands for targeting egfr in triple negative breast cancer cells. *Sci. Rep.,* **2019**, *9*(1), 2723. http://dx.doi.org/10.1038/s41598-019-38574-y PMID: 30804365

[59] Abrahams, B.; Gerber, A.; Hiss, D.C. Combination Treatment with EGFR Inhibitor and Doxorubicin Synergistically Inhibits Proliferation of MCF-7 Cells and MDA-MB-231 Triple-Negative Breast Cancer Cells *In Vitro* . *Int. J. Mol. Sci.,* **2024**, *25*(5), 3066. http://dx.doi.org/10.3390/ijms25053066 PMID: 38474312

[60] Stamos, J.; Sliwkowski, M.X.; Eigenbrot, C. Structure of the epidermal growth factor receptor kinase domain alone and in complex with a 4-anilinoquinazoline inhibitor. *J Biol Chem,* **2002**, *277*, 46265-46272.

[61] Yun, C-H.; Boggon, T.J.; Li, Y.; Woo, S.; Greulich, H.; Meyerson, M.; Eck, M.J. Structures of lung cancer-derived EGFR mutants and inhibitor complexes: mechanism of activation and insights into differential inhibitor sensitivity. *Cancer Cell,* **2007**, *11*, 217-227. http://dx.doi.org/10.2210/pdb2ito/pdb

[62] Gajiwala, K.S.; Feng, J.; Ferre, R.; Ryan, K.; Brodsky, O.; Stewart, A. Crystal structure of the wild-type EGFR kinase domain in complex with dacomitinib (soaked). *Protein Data Bank,* **2013**. http://dx.doi.org/10.2210/pdb4I23/pdb

[63] Yun, C.H.; Eck, M.J. EGFR L858R/T790M in complex with PD168393. *Protein Data Bank,* **2013**. http://dx.doi.org/10.2210/pdb4LL0/pdb

[64] Hargreaves, D. Studies Towards a Reversible EGFR C797S Triple Mutant Inhibitor Series. *Protein Data Bank,* **2021**. http://dx.doi.org/10.2210/pdb7AEM/pdb

[65] Ugolini, A.; Kenigsberg, M.; Rak, A.; Vallée, F.; Houtmann, J.; Lowinski, M.; Capdevila, C.; Khider, J.; Albert, E.; Martinet, N.; Nemecek, C.; Grapinet, S.; Bacqué, E.; Roesner, M.; Delaisi, C.; Calvet, L.; Bonche, F.; Semiond, D.; Egile, C. Discovery and Pharmacokinetic and Pharmacological Properties of the Potent and Selective MET Kinase Inhibitor 1-{6-[6-(4-Fluorophenyl)-[1,2,4]triazolo[4,3-b]pyridazin-3-ylsulfanyl]benzothiazol-2-yl}-3-(2--morpholin-4-ylethyl)urea (SAR125844). *J. Med. Chem.,* **2016**, *59*(15), 7066-7074. http://dx.doi.org/10.1021/acs.jmedchem.6b00280 PMID: 27355974

[66] Balin-Gauthier, D.; Delord, J.P.; Rochaix, P.; Mallard, V.; Thomas, F.; Hennebelle, I.; Bugat, R.; Canal, P.; Allal, C. *In vivo* and *in vitro* antitumor activity of oxaliplatin in combination with cetuximab in human colorectal tumor cell lines expressing different level of EGFR. *Cancer Chemother. Pharmacol.,* **2006**, *57*(6), 709-718. http://dx.doi.org/10.1007/s00280-005-0123-3 PMID: 16320055

[67] Zhong, F.; Xing, J.; Li, X.; Liu, X.; Fu, Z.; Xiong, Z.; Lu, D.; Wu, X.; Zhao, J.; Tan, X.; Li, F.; Luo, X.; Li, Z.; Chen, K.; Zheng, M.; Jiang, H. Artificial intelligence in drug design. *Sci. China Life Sci.,* **2018**, *61*(10), 1191-1204.

http://dx.doi.org/10.1007/s11427-018-9342-2 PMID: 30054833

[68] Vemula, D.; Jayasurya, P.; Sushmitha, V.; Kumar, Y.N.; Bhandari, V. CADD, AI and ML in drug discovery: A comprehensive review. *Eur. J. Pharm. Sci.,* **2023**, *181*, 106324. http://dx.doi.org/10.1016/j.ejps.2022.106324 PMID: 36347444

[69] Meresse, P.; Dechaux, E.; Monneret, C.; Bertounesque, E. Etoposide: discovery and medicinal chemistry. *Curr. Med. Chem.,* **2004**, *11*(18), 2443-2466. http://dx.doi.org/10.2174/0929867043364531 PMID: 15379707

[70] Wildman, S.A.; Crippen, G.M. Prediction of physicochemical parameters by atomic contributions. *J. Chem. Inf. Comput. Sci.,* **1999**, *39*(5), 868-873. http://dx.doi.org/10.1021/ci990307l

[71] Ma, C.Y.; Yang, S.Y.; Zhang, H.; Xiang, M.L.; Huang, Q.; Wei, Y.Q. Prediction models of human plasma protein binding rate and oral bioavailability derived by using GA–CG–SVM method. *J. Pharm. Biomed. Anal.,* **2008**, *47*(4-5), 677-682. http://dx.doi.org/10.1016/j.jpba.2008.03.023 PMID: 18455346

[72] Bickerton, G. R.; Paolini, G. v.; Besnard, J.; Muresan, S.; Hopkins, A. L. Quantifying the chemical beauty of drugs. *Nat. Chem.,* **2012**, *4*(2), 90-98. http://dx.doi.org/10.1038/nchem.1243

[73] Estrada, E. Spectral Moments of the Edge-Adjacency Matrix of Molecular Graphs. 2. Molecules Containing Heteroatoms and QSAR Applications. *J. Chem. Inf. Comput. Sci.,* **1997**, *37*(2), 320-328. http://dx.doi.org/10.1021/ci960113v

[74] Szostek, T.; Otto-Ślusarczyk, D.; Roszkowski, P.; Struga, M.; Szulczyk, D. Exploring the bioactive potential of (2-imino-4-oxo-1,3-thiazolidin-5-yl)acetic acid derivatives: A comprehensive review. *Results Chem.,* **2024**, *11*, 101828.https://doi.org/https://-doi.org/10.1016/j.rechem.2024.101828 http://dx.doi.org/10.1016/j.rechem.2024.101828

DISCLAIMER: The above article has been published, as is, ahead-of-print, to provide early visibility but is not the final version. Major publication processes like copyediting, proofing, typesetting and further review are still to be done and may lead to changes in the final published version, if it is eventually published. All legal disclaimers that apply to the final published article also apply to this ahead-of-print version.

18. Publication III

From obstacle to design advantage: activity cliff aware modeling for small-molecule drug discovery.

Szostek, T., & Szulczyk, D.

Drug Discovery Today, *31*(1). (2026).

DOI: https://doi.org/10.1016/J.DRUDIS.2025.104589

IF= 8.7, **MNiSW=** 200

Drug Discovery Today • Volume 31, Number 1 • January 2026 REVIEWS

KEYNOTE (GREEN)

From obstacle to design advantage: activity cliff aware modeling for small-molecule drug discovery

Tomasz Szostek *, Daniel Szulczyk

Department of Biochemistry, Medical University of Warsaw 02-097 Warsaw, Poland

Activity cliffs (ACs), defined as near structural neighbors with large potency differences, provide high-value information on structure–activity relationship discontinuities that medicinal chemists have leveraged for years. Yet modern artificial intelligence-based discovery frameworks often smooth this signal, yielding confident but unrealistic proposals. We present a three-layer roadmap that treats ACs as a first-class design feature: representation is tuned to expose AC signal during pretraining and sampling; evaluation focuses on steep regions with early recognition, pair-level retrieval, and calibrated uncertainty under fair, series-aware splits; and *de novo* generation is steered toward the best candidates. Integrated as a closed, self-improving loop, this framework helps practitioners build AC-aware drug design tools.

Keywords: activity cliffs; structure–activity relationship; cheminformatics; artificial intelligence; drug design



Tomasz Szostek received his master of pharmacy degree and is currently a fourth-year doctoral candidate at the Medical University of Warsaw, Poland. His research focuses on drug discovery, with particular emphasis on the synthesis and biological evaluation of small-molecule compounds bearing the 4-thiazolidinone scaffold. He completed postgraduate studies in bioinformatics, which enables him to integrate modern *in silico* approaches, such as QSAR modeling, molecular docking, and cheminformatics-driven analysis, into his projects. His recent work combines computational and experimental strategies to identify novel anticancer agents, particularly those active against lung and breast cancer cell lines. His broader interests include artificial intelligence-driven drug design and predictive toxicology



Daniel Szulczyk is an associate professor at the Medical University of Warsaw, Poland. He earned his MSc in chemistry from the University of Warsaw (2008) and his PhD in pharmaceutical sciences from the Medical University of Warsaw (2013). He previously served as an assistant in the Department of Medical Chemistry (2008–13) and, since 2018, has been an adjunct in the Department of Biochemistry, Faculty of Medicine. His research focuses on drug discovery and medicinal chemistry, with particular emphasis on the design and synthesis of biologically active heterocycles such as oxadiazoles, thiadiazoles, and tetrazoles. He combines organic synthesis, activity testing, structure–activity relationship studies, lipophilicity profiling, and computational methods to guide the design of new molecules.

Introduction

Rapid advances in artificial intelligence (AI) applied to drug-design workflows promise substantial improvements across the target-specific small-molecule discovery process. Beyond pipeline acceleration, cost reduction,[p1],[p2],[p3] data-driven target identification, and potency prediction,[p4],[p5] modern models now enable *de novo* molecular design through powerful generative paradigms.[p6],[p7],[p8],[p9] Despite evolving solutions, the economics and risk profile of R&D in the industry remain stubborn. Recent audits have estimated single-drug development costs to be in the hundreds of millions to multi-billion USD range, while clinical success still hovers at around ~10–15%.[p10],[p11] At the same time, there are several flagship AI-enabled drug programs that have progressed to clinical development, including the allosteric TYK2 inhibitor TAK-279/NDI-034858 and DSP-0038 for Alzheimer's disease.[p12],[p13] Despite this momentum, cheminformatics still faces obstacles that, if addressed, would substantially improve the predictive accuracy and generative capability of current models.

Activity cliffs (ACs), pairs of closely related molecules displaying disproportionately large potency differences, are the most prominent example of these obstacles.[p14] Even single-atom sub-

* *Corresponding author. Szostek, T. (tomasz.szostek@wum.edu.pl)*

1359-6446/© 2025 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).
https://doi.org/10.1016/j.drudis.2025.104589

stitutions can induce potency changes of tenfold or more,[p15] leading to costly mis-prioritizations: models that interpolate well in smooth regions of chemical space often fail at the edges of the structure–activity relationship (SAR), where one motif change determines program success.[p16] Importantly, in medicinal chemistry, ACs have long been treated as informative signals of SAR determinants rather than 'noise' or outliers, and their utility is well documented. But AI practice – from classical quantitative structure–activity relationships (QSARs) through graph neural networks (GNNs) and transformers to diffusion-based 3D generators and reinforcement learning (RL) – tends to be optimized towards objectives that favor averaging, effectively smoothing the landscape.[p17] As a result, ACs become invisible to predictors and propagate into generators, degrading calibration and ranking power exactly where medicinal chemists make high-stakes decisions. Crucially, ACs are not noise: they encode high-information structural edits that, if recognized and exploited, can accelerate SAR navigation rather than blur it.[p18],[p19],[p20] Consistently across public studies, performance degrades on AC-enriched evaluations, even when average metrics remain strong.

In this article, we advocate for a modeling focus that treats ACs as a design-relevant signal rather than a liability, and we show how to make this focus operational in practice. We present a comprehensive, data-centric, layer-by-layer review of the cheminformatics workflow that covers molecular representation, model assessment and prediction, and *de novo* design, all integrated into a closed, AC-aware learning system. Our goal is to summarize the best practices in AC management that give measurable gains in calibration and ranking, exactly where medicinal chemists make high-stakes choices. To our knowledge, there is no comprehensive review that treats ACs as a useful signal and pulls together working solutions built on this idea. Prior articles either review ACs conceptually within medicinal chemistry or survey AI for small-molecule discovery with only brief mentions of ACs.[p21],[p22] The motivation is straightforward: most models still drop in performance on AC-rich data sets, and the first AC-aware predictors and generators are now appearing. This makes it the right moment to turn a long-standing obstacle into a repeatable advantage for design.

Activity cliffs: definitions, measures, and good practices

In cheminformatics, ACs are most often defined as pairs of structurally similar compounds that show a large potency difference against the same target (Figure 1). Such local discontinuities in the SAR can degrade model performance: They make it harder for models trained with smooth, average-loss objectives to interpolate and to generalize reliably. In practice, definitions vary because researchers choose different operational settings for the two core criteria of ACs: molecular similarity and potency difference (Δp).[p10],[p15],[p16],[p17],[p18],[p19],[p21],[p22] Throughout this section, we make these choices explicit, show how they change, show which pairs are labeled as ACs, and explain the implications for model training and evaluation.

Before we introduce pairwise definitions, we first detail how Δp is computed and which similarity metrics and pairing

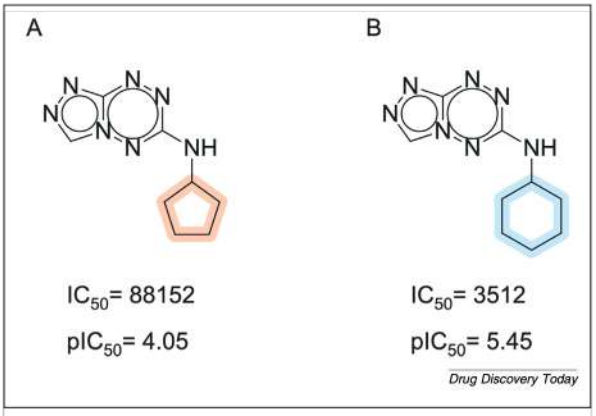


FIGURE 1

Example of an AC pair identified with 2048-bit ECFP4 fingerprints and the Tanimoto index (similarity $\tau \geq 0.90$) combined with a ≥tenfold potency difference ($\delta = 1.0$). The compounds were retrieved from ChEMBL and have bioassay data (IC_{50}) against the lung adenocarcinoma cell line (A549).

schemes are commonly used. Regarding similarity, a key design choice is the molecule representation. Two major families of embeddings are two-dimensional (2D; extracted from molecular graphs) and three-dimensional (3D; adding spatial orientation), each capturing a different aspect of structure. The most popular 2D (bit-based, numerical vector) representations are:

- Extended connectivity/Morgan fingerprint (ECFP/Morgan). Circular fingerprints in which each bit marks the presence of an atom-centered environment within a given radius. The bit vector length controls hashing granularity. It is widely used in high-throughput screening (HTS), QSAR analysis, and structure–activity similarity maps (SAS maps). Typical parameters: radius 2–3 (ECFP4/ECFP6) and 1024–2048 bits.[p23]
- MACCS keys. Faster but less expressive than ECFP. Useful for baselines and cross-checks.[p24]
- Atom pairs (APs). Encodes labeled atom–atom pairs and their shortest topological distances while capturing dispersed similarity patterns.[p25]
- Topological torsion (TT). Encodes labeled four-atom sequences (torsions). Often more selective than APs for local changes.[p26] To represent a level of similarity of two molecules, scientists use similarity coefficients, which on a 2D fingerprint is mostly computed using the Tanimoto index (Tanimoto coefficient, Jaccard index), measuring overlap relative to the union of bits. Popular alternatives are the Dice/Cosine/Soergel index and Tversky index (asymmetric), which generalizes Jaccard.[p27],[p28]

Recently, graph-based similarity methods have gained special attention, of which the two most common are:

- Maximum common substructure (MCS). Defines similarity by the size of the largest common subgraph (the 'core') between two molecules; similarity can be expressed as the fraction of

atoms/bonds in the MCS. The method is chemically intuitive, allows control over scaffold consistency, and is less sensitive to hashing artifacts.[p29]

- Graph-based similarity. Represents molecules as labeled graphs, where atoms are nodes and bonds are edges. Graph edit distance (GED) is the minimal cost of transforming one graph into the other. Weisfeiler–Lehman (WL) graph kernels are fast kernels that compare graphs via iterative node 'relabeling' (subtree patterns).[p30],[p31]

3D embeddings complement 2D topology by capturing shape complementarity and pharmacophoric arrangements that underlie molecular recognition. They can reveal similarities missed by 2D methods, but they rely on explicit spatial conformers.[p32],[p33],[p34] Analogous to 2D practice, 3D similarity metrics can be used in pairwise AC definitions by imposing a threshold on a shape metric (e.g., Shape Tanimoto or TanimotoCombo) and combining it with a Δp cutoff. We refer to such cases as 3D AC, highlighting discontinuities invisible to purely topological criteria:

- Rapid overlay of chemical structures (ROCS). Represents each molecule's volume as a sum of Gaussians and search orientations that maximize volumetric overlap (Shape Tanimoto). Pharmacophoric features ('color': H-bond donors/acceptors, aromatic, cation/anion) are overlaid in parallel. The standard score is Tanimoto Combo = Shape Tanimoto + Color Tanimoto (range ≈ 0–2). ROCS performs strongly in ligand-based virtual screening and scaffold hopping.[p35],[p36]
- Ultrafast shape recognition (USR) characterizes 3D shapes using moments of distance distributions from a set of reference points, yielding a pose-independent, alignment-free descriptor that is orders of magnitude faster than overlay methods. Ultrafast shape recognition with CREDO atom types (USRCAT) extends USR by adding pharmacophoric categories ('color') while retaining speed.[p37]

The most popular, abundant, and consistently curated end points in public resources such as ChEMBL are IC_{50} and K_i. They capture complementary biological information (functional effect versus equilibrium binding) with a well-defined theoretical link. However, mixing heterogeneous end points (e.g., EC_{50}/AC_{50} with K_i/IC_{50}) increases variance and can confound AC calls. Even within one end point class, cross-source variability warrants caution.[p38],[p39],[p40] IC_{50} depends on many conditions, such as assay type, substrate concentration, mechanism of action, and time of incubation, making it questionable, in some cases, to even compare only IC_{50} values (from different studies), whereas K_i is an equilibrium inhibition constant.[p41] Under competitive inhibition at equilibrium, the relationship between them is:

$$Ki = \frac{IC_{50}}{1 + \frac{[S]}{Km}} \quad (1)$$

where [S] is substrate concentration and K_m is the Michaelis constant. Equation (1) should be used only when its assumptions are justified. Otherwise, avoid converting IC_{50} to K_i or mixing end points in the same Δp calculation.[p42],[p43],[p44]

The Δp is defined as the difference of two given potency values on a $-\log_{10}$ scale (e.g., pK_i):

$$\Delta p = |p_1 - p_2| \quad (2)$$

Because the scale is logarithmic, Δp has a direct fold-change interpretation: Δp = 0.5, 1.0, or 2.0 corresponds approximately to 3×, 10×, or 100× differences, respectively. These thresholds are widely used as operational cutoffs in the AC literature, with $\Delta p \geq 1.0$ (≥10-fold) being the most common choice. Recent *Drug Discovery Today* work shows that dose response context can create potency discontinuities that mimic structural ACs. To separate structure from dose, we recommend reporting Hill slope, top and bottom constraints, and replicate quality for pairs used in AC calls. When dose-driven effects are suspected, perform a sensitivity check and only retain the pair as an AC if the discontinuity persists after harmonizing dose parameters.[p45]

In the context of Δp, two complementary views are used to define ACs. First, fixed global thresholds among similar pairs (e.g., ECFP4/Tanimoto ≥ 0.85 or Matched Molecular Pairs (MMP)-based matches), with fixed cutoffs at $\Delta p \geq 1.0$ and $\Delta p \geq 2.0$, which aids interpretability and benchmarking across methods. Second, continuous, threshold-free formulations, where the Structure Activity Landscape Index (SALI) ranks pairs by local steepness. Activity difference maps and consensus similarity landscapes likewise treat ACs as the extreme tail of a potency–similarity continuum rather than a binary label. In practice, the continuous view is useful for sensitivity analyses and for comparing representations without committing to a single cutoff.[p46]

Crucially, Δp cutoffs are not universal. Potency distributions differ by target and activity class, so a single global threshold yields different precision–recall trade-offs across classes. Recent large-scale analyses therefore define class-specific, statistically grounded thresholds from the empirical distribution of Δp within each target/activity class (under a fixed similarity regime), using simple rules such as class-wise quantiles (e.g., top 5–10% of Δp) or robust z-scores. This target-set dependent approach improves detection precision and interpretability and should be preferred when class sizes permit.[p47]

Most importantly, ACs do not only emerge from the data science point of view as a molecular representation and Δp definition. They are a property of the interaction between a ligand and its biological target. Different proteins exhibit different propensities to produce a steep SAR: rigid, sterically gated pockets tend to generate abrupt affinity drops upon small substituent changes, whereas flexible or solvent-exposed sites accommodate edits more smoothly. Structural analyses have repeatedly rationalized single-atom or single-motif ACs in co-crystal contexts (loss/gain of a key H-bond, displacement of conserved waters, emergent steric clash), underscoring that ACs are frequently target-dependent rather than only ligand artifacts.[p15],[p18],[p20],[p22] This is precisely why ACs should be diagnosed and explored as a first-class feature in the context of targets.

When high-quality spatial conformations of protein and ligand are available, ACs can be analyzed in the binding site to check shape/pharmacophore complementarity and water net-

works, and to attribute Δp to concrete interactions (π–cation, salt bridges, steric gating). Operationally, use pocket-aware 3D overlays and alignment/shape methods to expose ACs that 2D methods might miss. When reliable conformations are not available, a protein-aware but light in 3D path remains: build per-target models on harmonized K_i/K_d subsets to reduce assay condition variance, add target descriptors or protein embeddings to inject binding site context, and set class-specific Δp thresholds (quantile or robust z-score) so that AC calling reflects a target's intrinsic steepness rather than a global cutoff.

After establishing how potency differences are computed and how molecular similarity is represented, we now make the pairing step explicit. This step determines which molecular pairs are eligible to be called Acs, and it is therefore the main knob that controls both sensitivity (how many ACs did you find?) and specificity (how many are chemically credible?). Operationally, 'pairing' makes AC eligibility explicit. We first fix a similarity regime and an end point within a target, then enumerate candidate pairs and score potency differences on the p-scale. In practice, complementary families cover most use cases: ECFP/Tanimoto for close analogs in 2D topology, MMP/MMS to attribute changes to single edits or coherent R-group progressions, MCS to enforce a strict shared core, SALI/SAS-style landscape diagnostics to quantify local steepness, and a 3D track (shape alignment/docking) to reveal ACs that are invisible in 2D.[p15],[p27],[p48],[p49],[p50],[p51],[p52],[p53],[p54],[p55],[p56],[p57],[p58],[p59],[p60],[p61],[p62] Consistent assay metadata and transparent thresholds (τ, Δp) matter as much as algorithmic choice. Table S1 in the supplementary information consolidates concise definitions, recommended defaults, known failure modes (e.g., end point mixing, hubness, oversmoothing), and concrete mitigations for each pairing family. For the simplified pairing method selection, follow the decision diagram in Figure S1 in the supplementary information.

Large-scale surveys on ChEMBL have systematically quantified how often ACs occur and how many compounds participate, across multiple AC definitions (fingerprint-based ACs, MMP-ACs, and retrosynthetic MMP-ACs with class-dependent thresholds). In an early global analysis of high confidence potency data, ~22–34% of active compounds participated in at least one AC with ≥100-fold potency difference, whereas only ~4.7–5.7% of qualifying similar pairs formed ACs. Thus, ACs are rare at the pair level, but common at the compound level. Subsequent updates across ChEMBL releases reported remarkably stable proportions (~4.6–6.9% of qualifying pairs form ACs), with ~23–35% of compounds involved, and documented 16 096 class-dependent MMP ACs across 212 activity classes. These trends persisted over a decade, indicating steady growth in absolute AC counts while relative frequencies remain constant. Collectively, these studies demonstrated that ACs are not outliers, but a recurrent feature of medicinal chemistry data.[p17],[p63]

To summarize, building generative and predictive models in chemistry requires special caution and understanding of used chemical representations and biological information. The best solution for building effective models is a data-centric approach, where we focus on careful database building and its curation. In this approach, we can emphasize several common, universal good practices that, when implemented into a pipeline, can increase model efficacy, interpretability, and correct AC recognition. During the stage of data collecting, assay harmonization must be ensured. For reliable Δp, prefer a single-end-point subset with only K_i or $IC_{50}/EC_{50}/K_d$ results without mixing them. The results obtained should be consistent with bioassay type, bioassay conditions (e.g., incubation time, concentration range), organism type, and cell line. The results should be mapped to $-\log_{10}$ (pChEMBL) with attention to non-trivial variability across sources, avoiding non-specific results with qualifiers ('<', '>', '≥', '≤'), especially for regression models, and exclusion of non-reliable data sources.[p39],[p64] Before computing any form of fingerprints or descriptors, a correct chemical compounds library must be cleaned and normalized from obvious structural duplicates on canonicalized Simplified Molecular Input Line Entry System (SMILES) or InCHIKeys, as well as from more hidden duplicates such as tautomers, protonated compounds, stereoisomers, salt, and solvate forms. The presence of such duplicates exposes the model to data leakage, especially when a random training/test set split is applied, leading to falsely increased accuracy values [area under the curve (AUC)]. A good practice is to prefer time-based splits or scaffold and series/cluster splits. Moreover, the current bioinformatic tools allow for the detection and removal of pan-assay interference compounds (PAINS) from the data set.[p65],[p66] To avoid overfitting, for classification models, address class imbalance with stratification or class-weighted losses; for regression models, ensure broad response-range coverage and consider robust losses when noise is heterogeneous.[p67],[p68] Declare the molecular representation and its parameters and define the applicability domain (AD) matched to that representation, understood as a chemical space in which models' predictions are reliable. Finally, among the results, release a complete, reproducible protocol: data sets, thresholds, calculations, software, code, and all methods to regenerate splits and features.

Activity cliff aware modeling

Clean data provide models with a solid foundation, but in cheminformatics, the landscape of SAR rarely changes smoothly. The greatest changes often result from small structural edits that flip potency by orders of magnitude. Recognizing this phenomenon is the first step in moving from treating them as a malfunction or statistical outlier to using them as a concentrated, actionable source of SAR signal. If you want to use ACs as a signal, first choose a pairing strategy consciously by comparing alternative methods (see Table S1 and Figure S1 in the supplementary information) and run a sensitivity analysis across thresholds so that the calling rules match your end point.[p69] Then follow the good practices and methods presented in this section, which explain how to turn ACs from a liability into leverage.

Layer I. Embeddings and pretraining (AC-aware molecule representation)

In chemical ligand-based modeling, a representation maps a molecule to a vector of computed features (fingerprints, descriptors) or learned patterns (e.g., GNN). The embedding is the vector consumed by the model, such as the final GNN graph state or the fingerprint bit vector.[p70] Latent similarity denotes proximity in this embedding space, measured with cosine or Euclidean dis-

tance. It can align with structural similarity, such as ECFP Tanimoto, or with biological similarity, such as activity, but frequently it does not. The similar property principle remains a heuristic, and ACs are systematic exceptions. Benchmarks confirm this gap. MoleculeACE reports that widely used machine learning (ML) and deep learning (DL) methods perform well on global metrics, yet underperform on AC compounds. Two representation effects contribute to this behavior. Oversmoothing, common for GNNs, compresses neighborhoods and attenuates local AC SAR.[p71] Hubness in high-dimensional spaces produces universal neighbors and conceals discontinuities.[p72] Taken together, both effects put AC neighbors close to each other in the learned space, despite large potency gaps. t-distributed stochastic neighbor embedding (t-SNE) or uniform manifold approximation and projection (UAMP) plots can visualize this, but must be read cautiously: they preserve local structure only approximately and distort global geometry.[p73] The practical conclusion is that simple, latent proximity is not AC robustness, unless the representation is explicitly trained to keep AC neighbors apart. To illustrate this issue, we conducted a simple AC exercise using a comparison of principal component analysis (PCA), which showed global variance and t-SNE (which allows an insight into local neighbors differences and clusterization) on compounds with bioequivalent results (IC_{50} in nM) against the A549 cancer cell line (fetched from ChEMBL database).[p74]

From 32 598 cleaned and standardized compounds, we identified 574 AC pairs using ECPF4 with Tanimoto threshold $\tau \geq 0.9$ and $\Delta p \geq 1.0$ (×tenfold difference). Figure 2 integrates four views.

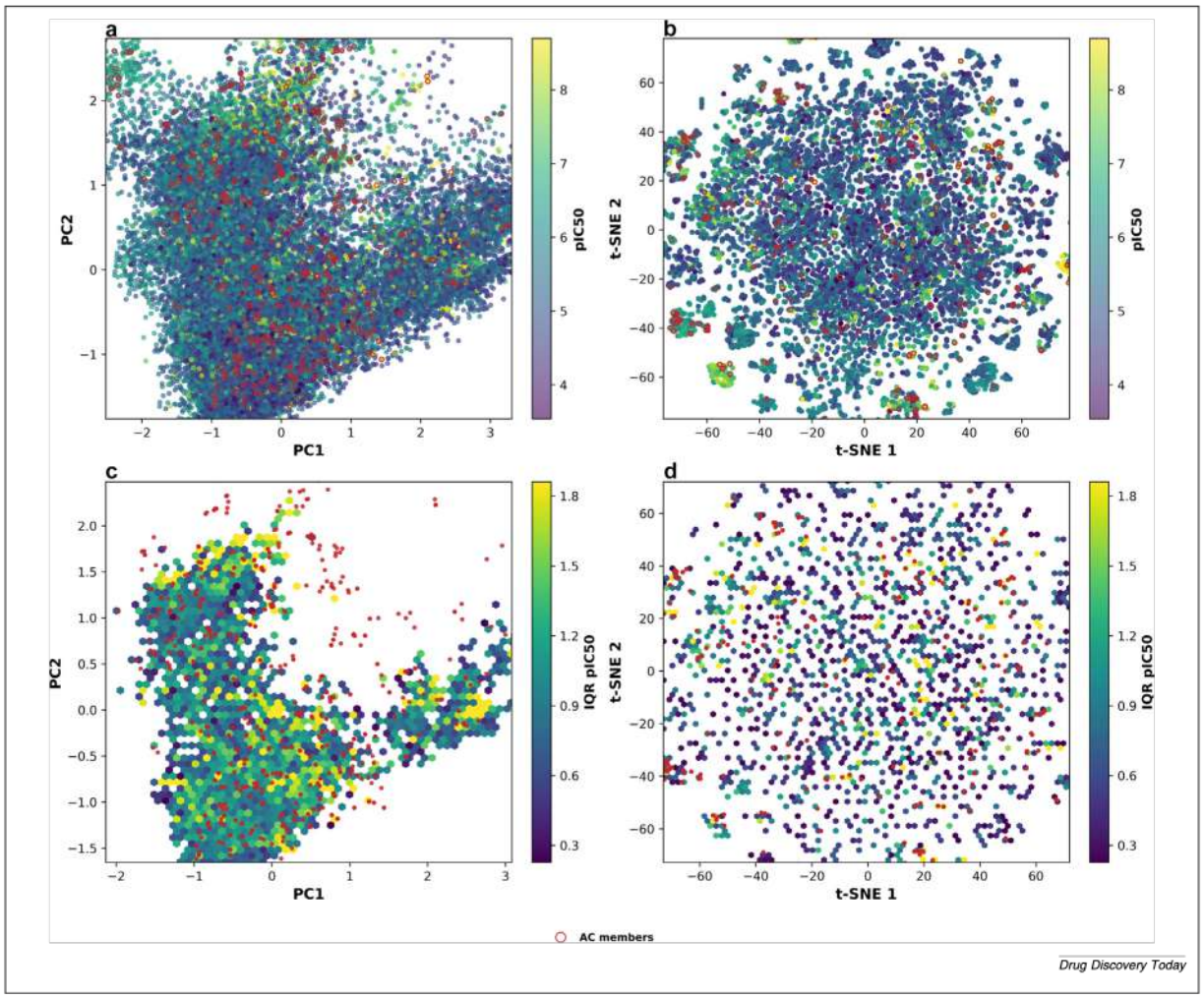


FIGURE 2

(a) PCA (PC1–PC2) and **(b)** t-SNE (2D) projections of ECFP4 fingerprints. Color encodes pIC_{50} (shared scale), red outlines highlight ACs. Both views reveal numerous ACs occurring in regions of close latent proximity. **(c, d)** Closer views of PCA (c) and t-SNE (d) plots. Color encodes the interquartile range of pIC_{50} (IQR pIC_{50}) within each hex (a robust measure of local variability), with a shared color scale. Red markers highlight ACs. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Figure 2a and Figure 2b show PCA and t-SNE colored by pIC_{50}, which locates potency gradients in chemical space, but does not quantify local steepness. Figure 2c and Figure 2d zoom into regions with high interquartile range (IQR) of pIC_{50}, marking zones where predictions are unstable, and ranking errors accumulate. These plots are descriptive: they reveal global trends and local variance, but not the true local slope of the activity landscape. Figure 3 shows an SAS map with ACs that occupy the high similarity and high Δp region, and appear as continuous, ridge-like corridors. These ridges are coherent features of the activity landscape rather than isolated anomalies. In PCA, variance compresses along these ridges. Together, Figures 2 and 3 connect AC with SAR discontinuity and show where standard averaging models are most likely to be misled. This motivates AC-aware pretraining and evaluation precisely in these high-gradient regions to improve ranking, calibration, and uncertainty.

Triplet-shaped latent geometry methods align the embedding space with potency contrasts by design. The family is represented by ACtriplet and ACANet/ACA.[p75],[p76] Both treat ACs as a metric learning signal: ACtriplet separates pretraining (triplet and bio-triplet loss functions with subsequent fine-tuning) from downstream supervision, whereas ACANet folds online triplet mining directly into the predictive objective. In practice, these approaches are strong defaults for moderately sized, assay-consistent data sets, dominated by close analogs. Both methods tend to improve the root mean squared error (RMSE) computed from AC compounds ($RMSE_{cliff}$) without sacrificing global fit, and they are comparatively straightforward to drop into existing GNN stacks. Triplet geometry is especially valuable when molecules that exhibit large Δp nonetheless occupy nearby regions of the activity landscape, because it enforces separation in latent space commensurate with potency contrast. These methods are

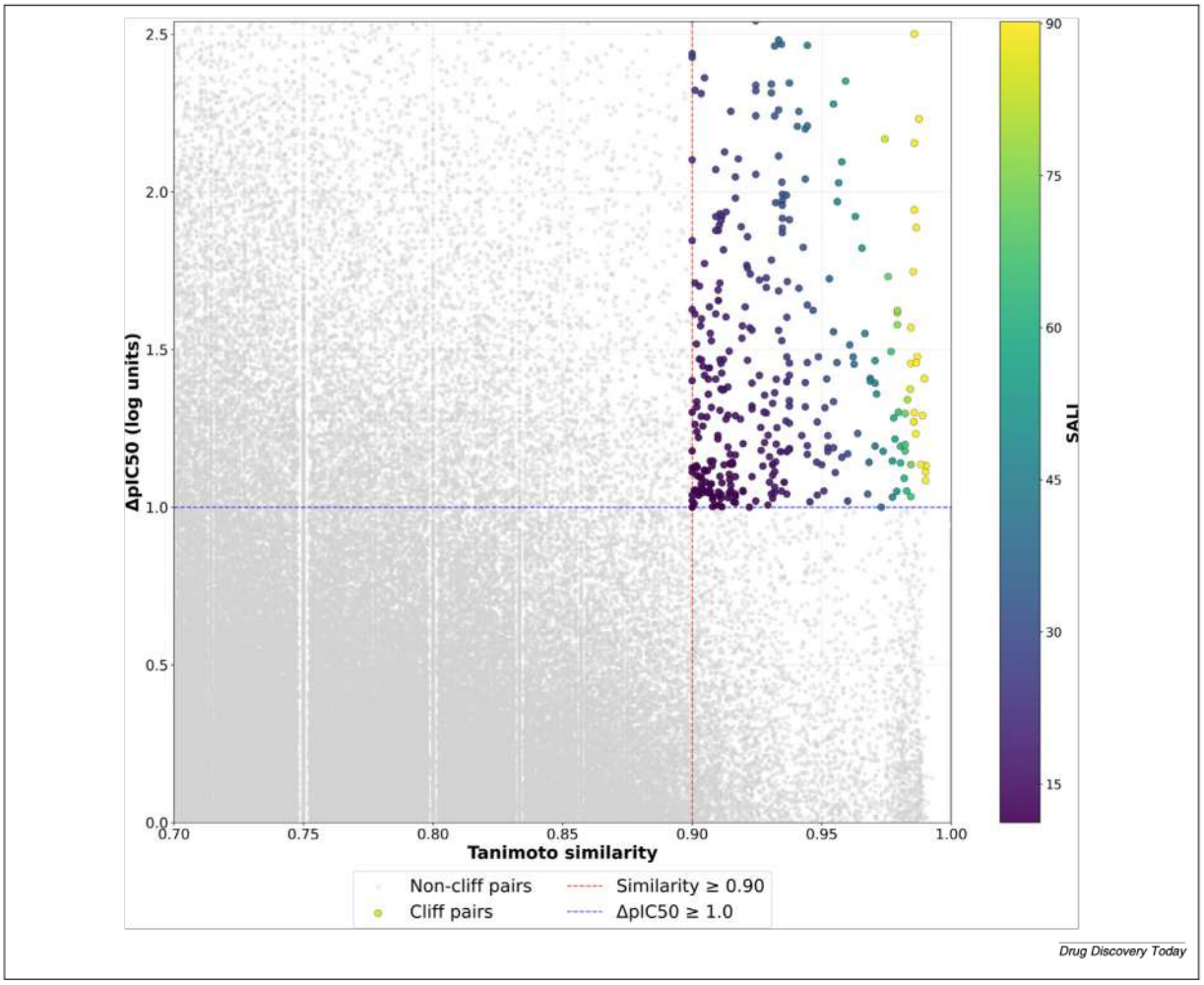


FIGURE 3
Structure–activity similarity (SAS) map. Each point is a compound pair built from ECFP4 neighbors. The x axis is Tanimoto similarity τ. The y axis is the potency difference Δp on the log scale. Color encodes SALI. Warm colors mark a steeper local slope.

most appropriate when 2D topology captures the relevant SAR, and they are not recommended when shape- or conformation-driven discontinuities are suspected to underlie the signal.

A second category, molecular pair anchored encoders, exploits MMPs as structured supervision to inject the local SAR directly into the embedding. SiamACLoss is the canonical representative: two weight-sharing encoder–decoders reconstruct fingerprints while pulling non-AC MMPs together and pushing AC MMPs apart by a margin, so latent distance scales with Δp.[p77] This methods family is attractive when your project workflow already mines MMPs and when you want edit-level interpretability that maps naturally to R-group reasoning. Because supervision comes from pairs, training is comparatively light. Moreover, fine-tuning the pretrained encoder into potency regression preserves the AC-aware geometry without requiring end-to-end architectural changes. Use this path when you need a transparent 'one-edit, one-effect' narrative and when you prefer to use 2D embeddings while correcting the usual collapse of near neighbors. Avoid it when the ACs in your data set are primarily 3D or when your series lacks reliable pair enumeration. A sibling on the predictive side is ACGCN, which ingests either the two full MMP molecules or the explicit core plus substituents, and classifies the pair as AC/non-AC.[p78] It is a pragmatic bridge between representation and Layer II evaluation, excelling when your goal is to triage candidate edits rather than to learn a universal molecule encoder.

A third category, image-centric fine-edit encoders, is represented by MaskMol and is motivated by empirical observation that graph encoders tend to blur micro-edits as molecular similarity increases, whereas chemically faithful 2D depictions preserve the very cues that flip potency.[p79] Masked image pretraining with chemically guided masking produces embeddings that keep AC neighbors apart and transfer downstream tasks. It is an attractive method for detecting subtle molecular features of ACs, such as local aromatic decoration or heteroatom positioning, using a standard input pipeline enriched by imagine AC logic without the costs of full 3D frameworks.

SemiMol addresses the low label regime head-on through curriculum and confidence.[p80] It is a supervised learning framework with an instructor that filters proxy-labeled examples, allowing the model to learn in smooth regions of SAR first. Then it graduates into AC, high-variance cases to add more nuance to the basic knowledge. This design is the right first move when you are early in a program, labels are sparse or heterogeneous, and you nevertheless want the encoder to experience enough AC-like perturbations to generalize.

Finally, physics- and reaction-informed pretraining, which injects structural context before task supervision, is exemplified by SCAGE and REMO.[p81],[p82] SCAGE injects 3D geometry and functional group semantics into a graph transformer, so attention becomes explicitly distance-aware. It excels when ACs are genuinely shape driven by steric gating, 3D pharmacophore alignment, or conformational locking, and when reliable conformers are available. REMO replaces single-molecule masking with reaction-center objectives learned from curated transformations, allocating capacity to the atoms and bonds that actually change, which is precisely where potency flips tend to occur. It is a sensible default when your series evolves through well-defined transformations and when you want the encoder to 'expect' minimal edits with maximal effects. Both methods typically require heavier pretraining budgets and cleaner inputs (reliable conformer and reaction data) but repay that investment with the most faithful treatment of discontinuities that emerge from spatial shape and chemistry, rather than 2D topology.

Crucially, the added complexity and computational cost of 3D transformers such as SCAGE is justified when the biological system is dominated by rigid or highly structured binding sites, when stereochemistry, atropoisomerism, ring puckering, halogen bonding, or metal coordination repeatedly coincide with SALI spikes around ACs, and when docking or structural assays indicate shape complementarity as a first-order driver for potency flips. Conversely, in flat landscapes with topology-driven SARs and limited labels, robust 2D baselines (ECFP4 plus regularized learners), MMP-anchored encoders, or triplet-shaped GNNs often match or exceed 3D methods at a fraction of the cost, and sample complexity (with simple ECFP4 combined with classical ML methods often winning over deeper architecture, especially in AC-rich benchmarks). In low-data regimes, complex encoders are prone to overfitting and unstable calibration, so we favor pair-anchored pretraining, conservative series-aware splits, and explicit early recognition objectives. Table 1 consolidates the head-to-head evidence (data sets, split protocols, AC-subset metrics, code availability, and compute footprint) so that a practitioner can choose a method under realistic constraints. Complete detailed methods descriptions and implementation notes are provided in the supplementary information online ('Layer I – embeddings and pretraining').

Layer II. Prediction under ACs: objectives, sampling, and validation

The problem of assessing both overall fit and the more trustworthy, AC-focused performance was clearly surfaced by the MoleculeACE benchmark.[p17] The authors evaluated predictive power on a broad suite of traditional ML and DL models across 30 molecular target data sets collected from the ChEMBL database. In general, every method struggled with AC performance, but surprisingly, less advanced ML models showed better results than deep architecture frameworks, proving their better application to small-molecule drug design tasks. The authors hypothesize that the difference in performance might come from suboptimal molecular representation (addressed in 'Layer I'). MoleculeACE is a state-of-the-art tool for AC-aware model assessment, enabling standardized identification of ACs with different methods, biological activity thresholds, and reporting models, enabling fair comparison of ML/DL methods in order to choose the one best suited for a particular data set. The benchmark introduced a shift in the model evaluation through introducing a new, currently widely used metric, $RMSE_{cliff}$ (RMSE computed on molecules being part of at least one AC pair), which together with the global metric, reveals models' predictive power around difficult SAR regions.

ACNet[p83] complements MoleculeACE by recasting ACs as a pairwise classification, rather than regression. It is a curated data set of 400 000 MMPs with biological annotations across 190 targets (ACs are identified based on ×100-fold Δp) extracted from the ChEMBL database and designed as a benchmark for the evaluation of the performance of AC-aware models. The authors

TABLE 1

Head-to-head comparison of AC-aware molecular representation and pretraining methods.

When to apply	Method	Data set (splits)	Primary metric/gain	Code availability	Compute footprint
SAR dominated by close analogs; 2D topology captures essential structural edits easy to use with; GNN architecture	ACtriplet	MoleculeACE (stratified random sampling on AC labels)	$RMSE_{cliff}$ = 0.603–0.995, consistently lower than non-pretrained DL baselines	https://github.com/githubXin123/ACtriplet.git	Moderate; 2D GNN pretraining + finetuning
	ACANet/ ACA	MoleculeACE (nested, series-aware)	$RMSE_{cliff}$ 0.711 vs 0.742; average improvement 6.59% across all data sets	https://github.com/shenwanxiang/ACANet.git	Moderate; 2D GNN, online triplet mining adds overhead
Edit-level accountability; fast triage of candidates R-group edits	SiamACLoss	MoleculeACE + 7 diagnostic subsets (per subset)	Overall improvements of $RMSE_{cliff}$ and $AUROC_{cliff}$	https://github.com/cmvcordova/acsuite.git	Light; 2D encoders with weight sharing
	ACGCN	ChEMBL v28 MMPs for three single-target sets (random split)	AC classification with AUROC reporting	https://github.com/chunkyun/ACGCN.git	Light; no pretraining, three-layered GCN
Subtle micro-edits drive flip potency; graph models blurring close neighbors' edits; AC uplift without conformers	MaskMol	MoleculeACE (scaffold split)	ACE average gain of 34.43%	https://github.com/ZhixiangCheng/MaskMol.git	Moderate; 2D only + pretraining
Sparse or noisy labels; heterogenous end points; prioritize robustness and calibration under limited supervision	SemiMol	MoleculeACE (stratified random sampling on AC labels)	$RMSE_{cliff}$ reduced on average by 26.5% vs baselines	https://github.com/smiles724/InstructMol.git	Moderate; 2D GNN with SSL instructor
Shape, pharmacophore, steric alignment drives ACs; stereochemistry is crucial; reliable conformers or reactions	SCAGE	MoleculeACE + 3D functional groups (per-target splits with 3D features)	Major $RMSE_{cliff}$ improvements for 23/30 targets	https://github.com/KazeDog/SCAGE.git	Heavy; conformer generation + 3D attention
	REMO	MoleculeACE and ACNet style (AC-aware splits)	$RMSE_{cliff}$ and Graphomer. First DL model to beat ECFP ML baseline on AC benchmark	–	Heavy; huge reaction-aware pretraining

release multiple target-specific subsets (small/medium/large 'tasks') to teach models which local structural edits drive large potency changes, and additionally they include a cross-target 'mixed' setting with target-wise splits to probe generalization. Trained and evaluated across these subsets, descriptor-based ML baselines with ECFP fingerprints consistently outperform deep encoders, attaining ROC–AUC (area under the receiver operating characteristic curve) up to 0.984 ± 0.003 on the large split. ACNet provides AUC/precision on AC pairs alongside AC-stratified regression metrics ($RMSE_{cliff}$) to avoid optimistic bias from non-AC baselines. Because AC pairs are a minority class, ROC–AUC can look deceptively strong under rarity. Accordingly, we treat the area under the precision–recall curve (AUCPR) with reported prevalence and early recognition of enrichment factor on AC (EF_{AC}) as the primary retrieval metrics for AC detection.

Precision on SALI-ranked pairs[p84] allows early retrieval of the steepest AC. After conducting SALI analysis, in which a large number indicates severe ACs, define an elite set (top k SALI pairs or top q%) with pairs ranked by models' AC score/probability and compute precision. This method directly tests whether the model helps triage candidates into high-information, AC-rich SAR regions. Because ACs are in general rare among data sets, we suggest treating their detection as an imbalanced retrieval task and report the AUCPR (and $AUCPR_{cliff}$ computed on AC pairs),[p85] which is more informative than ROC–AUC under class imbalance. The AC prevalence should always be reported, so AUCPR is interpretable against the random baseline. To quantify early-recognition performance, we suggest reporting the EF_{AC}.[p86] $EF_{AC\pi} = 1$ corresponds to random ranking, and values >1 indicate that the top α fraction of the list is proportionally richer in true ACs than expected by chance. Building on this early-recognition view, the subsequent analysis examines whether the enriched top-ranked ACs are identified with calibrated confidence via conformal prediction (CP) and calibration diagnostics, and how these calibrated signals can be operationalized in practice through AC-targeted active learning (AL).

Large predicted Δp between close structural neighbors is useful only when accompanied by calibrated uncertainty. A widely used framework is CP,[p87] which provides per-sample prediction sets/intervals with a finite-sample validity guarantee. In regression, inductive CP associates with each input x_i in an interval (Li, Ui). In classification models, CP returns a label set with error controlled at α. Because CP validity is marginal over the full distribution, good practice is to stratify uncertainty by landscape region, especially for the AC subset. Two transparent readouts are therefore reported: $coverage_{AC}$ and $width_{AC}$. These summaries are shown alongside the non-AC subset, confirming that intervals are wider (but still correctly covering) where SAR is discontinuous. Because CP validity is marginal over the full distribution, we explicitly stratify calibration by landscape

region. Concretely, compute $coverage_{AC}$ and $width_{AC}$ on the subset of molecules that participate in at least one AC pair (n_{AC}), and $coverage_{non\text{-}AC}$ and $width_{non\text{-}AC}$ on the complement. Claims of improved decision reliability near steep SAR should be supported by $coverage_{AC} \approx 1 - \alpha$, with $width_{AC}$ increasing only moderately relative to $width_{non\text{-}AC}$. In practice, we also recommend SALI-stratified calibration curves (coverage and median width across SALI quantiles) to localize over- or under-confidence along activity ridges. In parallel, report NLL (negative log-likelihood for probabilistic regressors) for probabilistic regressors and ECE (expected calibration error for classification models) for classifiers separately on AC and non-AC strata. Reductions in NLL_{AC}/ECE_{AC} are the relevant improvements under ACs. These AC-specific summaries make uncertainty statements auditable exactly where Δp is large and models are most likely to mislead. In cheminformatics pipelines, CPSign is commonly used for inductive/transductive CP.[p88] In deep models, uncertainty is often decomposed via ensembles or evidential heads on GNNs to separate epistemic models from aleatoric data/assay components. Therefore, CP is complemented by NLL, which evaluates sharpness and calibration of the predictive density, and ECE. Both metrics are reported separately on AC and non-AC subsets to detect overconfidence near ACs.

Crucially, all AC-aware metrics are definition sensitive, which bears directly on the fairness of comparisons and on the transferability of conclusions across projects; therefore, we treat 'What counts as an AC?' as a controlled factor rather than a constant and recommend exploring the same chemical space under multiple similarity kernels and potency thresholds while holding the split protocol fixed. In practice, this means repeating AC detection with alternative embeddings (e.g., ECFP4, ECFP6, MMP-based or pharmacophore/MCS similarity) at several similarity cutoffs (e.g., Tanimoto ≥ 0.8, ≥0.9, ≥0.95), potency thresholds (e.g., Δp ≥ 1.0/2.0/5.0), and engines (various ML versus DL models). Report the number of identified AC pairs for each recipe and threshold. Assess the best model architecture using $RMSE_{cliff}$ for regressive and $AUCPR_{cliff}$ for classification tasks. To characterize the SAR landscape, where decisions are taken, summarize early recognition with $EF_{AC\alpha}$ at small budget ($\alpha = 1 - 5\%$), and audit trustworthiness on the AC subset with $coverage_{AC}$ and $width_{AC}$ from CP. Quantify actionability by the percentage of high SALI pairs recovered in top k or q%, along with the observed fluctuation of false positives near assay-variability bands. Operational details, exact formulas, and a reporting checklist are compiled in Table S2 in the supplementary information so that practitioners can implement the discussed procedures without ambiguity.

With uncertainty quantified and calibrated, we close the prediction layer by introducing AL that targets AC neighborhoods.[p89],[p90] Recent studies in drug discovery have demonstrated AL's leverage for prioritizing syntheses/screens, offering concrete patterns for batch sizing, exploration–exploitation balancing, and recalibration after each assay cycle. These design maps can be applied to AC-aware acquisition by swapping target-agnostic heuristics for SALI-conditioned uncertainty querying. The algorithm prioritizes molecular pairs that are simultaneously informative (uncertain) and probably AC-bearing, as established by high SALI. A practical acquisition score for a candidate pair (i, j) is:

$$A(i,j) = \lambda \widetilde{SALI}(i,j) + (1-\lambda)\,\widetilde{U}(i,j), \tag{3}$$

where $\widetilde{SALI}$ is SALI standardized to [0,1], $\widetilde{U}$ is a standardized uncertainty proxy (e.g., CP interval width or ensemble variance) on the model's AC score, and $\lambda \in [0,1]$ trades off AC likelihood and epistemic uncertainty.

The model should not only predict two near-identical molecules that differ strongly in activity, but also indicate where in the structure this difference originates. Explanations (saliency/attribution maps) tell us which atoms/scaffold fragments the model believes are responsible for large Δp. That is the purpose of XAI (Explainable AI) under AC.[p91] Concretely, AC pairs are aligned using maximum common substr (MCS) with the edited fragment explicitly marked, then per-atom attribution maps are computed with a method appropriate to the model class [integrated gradients for neural networks; Grad-CAM-style methods for graph neural networks (GNNs); SHAP (SHapley Additive exPlanations), which quantify feature contributions based on cooperative game theory, for descriptor-based models] and normalized.[p92],[p93],[p94] We read the maps with a small set of auditable checks that test three things at once: Do attributions stay consistent on the shared scaffold; do they concentrate on the actual edit; and do they recover any known AC motif? This keeps the model from crediting the wrong atoms for the activity shift.

Beyond small molecules, AC-aware principles were extended to peptides via the AMPAC framework,[p95] which formalizes sequence-level ACs and supplies a curated benchmark, applying Layer II diagnostics to antimicrobial peptides (AMPs). AMPAC defines ACs at the sequence level using normalized BLOSUM62 similarity (≥ 0.9), where BLOSUM62 is a substitution matrix quantifying evolutionary similarity between amino-acid sequences, combined with a ≥2-fold change in minimal inhibitory concetration (MIC), and organizes paired AMPs into a curated benchmark for training and evaluation. The study applies identity-controlled splits and reports consistent readouts such as AUROC/AUPRC for AC detection and correlation-based metrics for MIC regression, across a broad model suite [various classical ML and pretrained protein language models (PLMs)].[p96] Results show one clear pattern – large PLMs such as ESM2 set the best baselines for AC detection and MIC, but their performance is still modest. A Spearman value of around 0.47 on MIC leaves clear headroom for improvement, and confirms that the analysis of discontinuous SARs is hard, even with strong encoders.

Layer III. AC-aware small molecule generation

Collectively, the AC metrics, calibrated uncertainty (CP/UQ), and AL indicate where the SAR breaks and how models behave around these breaks. The natural progression is to use these signals as a design advantage and as rewards that guide lead optimization. Modern *de novo* design should treat local chemical edits leading to extensive potency difference as first-class information for lead optimization. The most prominent example of such tools is ACARL[p97] (Activity Cliffs-Aware Reinforcement Learning).

ACARL integrates AC signals directly into a *de novo* design, injecting high-impact SARs from AC-rich regions and steering generation toward them, allowing drug candidate optimization. The framework consists of two crucial functions: activity cliff index (ACI) for AC detection, and contrastive loss in RL, a function that prioritizes learning from ACs. ACI is similar to the SALI form of molecular landscape visualization, but more generalized and applicable to various potency metrics, such as molecular docking. ACI is defined as:

$$ACI(x,y;f) := \frac{|f(x)-f(y)|}{d_T(x,y)}, \quad x,y \in \delta, \tag{4}$$

where d_T is the Tanimoto distance between molecules represented as ECFP. A pair of molecules is declared as an AC when two conditions are met:

$$f(y)| > \alpha_1 \text{ and } ACI(x, y; f) > \alpha_2 \tag{5}$$

Fine-tuning of the pipeline proceeds from a pretrained SMILES Transformer decoder, with a reward that preserves language-model knowledge while pushing the objective:

$$R(x) = \log - \text{likelihood}(x) + \sigma \cdot f(x) \tag{6}$$

To handle imbalance and highlight high-value SARs, ACARL uses a weighted replay across all past steps, emphasizing ACs:

$$L_\omega\left(\theta^{(k)}\right) = \frac{1}{\sum_{i=1}^{n}\sum_{j=1}^{k}\omega_{ij}} \sum_{i=1}^{n}\sum_{j=1}^{k}\omega_{ij} L(x_i^{(j)};\theta^{(k)}), \tag{7}$$

where $\omega_j \geq 0$ increases for samples linked to detected ACs, allocating more updates to high-impact regions. By leveraging dense SARs previously treated as outliers, ACARL consistently improves target metrics, indicating that gains come from explicit AC-aware prioritization rather than replay alone. The authors also note a key limitation: current pipeline evaluation leans on docking-based protocols, which might misalign with *in vivo* outcomes, leading to oversight of the most optimal structures. In practice, ACARL is complementary to synthesizability and multi-objective filters.

In published reports, ACARL is the only explicitly AC-aware generator that quantifies gains against strong RL baselines under a common docking-score objective. On three tested molecular targets, ACARL improves top one docking from −9.83/−11.05/−9.76 kcal/mol (generic RL) to −13.92/−12.15/−12.92 kcal/mol (Δ = 4.09/1.10/3.16) and strengthens top 100 means while maintaining or increasing internal diversity (IntDiv ≈ 0.86–0.88), indicating that replay weighting toward AC-rich edits yields harder-binding proposals without mode collapse.

Whereas ACARL encodes AC signals by design, the next tools are not explicitly AC-aware; however, they are presented here because they can be configured to operate under the same logic. REINVENT provides an RL framework that accepts AC-oriented scores, such as predicted |Δp| or ACI with CP and uncertainty gating and with XAI checks, and mmpdb (matched molecular pair engine for AC-aware generation) supplies a matched molecular pair single edit mechanism that keeps proposals local and chemically plausible. Together, they implement the AC logic through reward functions and constraints rather than through a new architecture, and they therefore offer a practical route to AC-aware generation within standard pipelines.

REINVENT 4[p98] is an advanced tool for new molecule generation, R-group replacement, scaffold hopping, library design, linker design, and lead optimization. Its RL objective enables the injection of AC signals as a reward or augmented likelihood. The pipeline built around RL allows various molecule optimization strategies, where some actions can be rewarded when beneficial for the output. The main logic behind that is driven by augmented log-likelihood:

$$R(x) = \log p_{prior}(x) + \sigma \cdot S(x) \tag{8}$$

where $\log p_{prior}(x)$ preserves the pretrained language model and S (x) ∈ [0,1] is a task score, where σ > 0 modulates task pressure (a similar structure to that of ACARL in Equation (6)). To apply AC logic into RL, define a pair relative score using a AC-calibrated predictor f and nearest neighbor $NN(x)$ in reference series S:

$$ACI_{pred}(x;f) = \frac{|f(x)-f(NN(x))|}{1-sim_T(x,NN(x))}, \quad sim_T = Tanimoto(ECFP) \tag{9}$$

This parallels ACI from the ACARL pipeline (Equation (4)), but at the same time uses generated molecule x against the closest analog in S. A gated task score then reads:

$$S_{AC}(x) = 1\{(U-L) \leq \tau\} \cdot 1\{|f(x)| > \alpha_1\} \cdot \min\left(1, \alpha_2 ACI_{pred}(x;f)\right), \tag{10}$$

where [L, U] is the conformal interval for $f(x)$ (uncertainty gate), τ controls admissible uncertainty, and α_1, α_2 are potency/ACI thresholds (like in Equation (5)). Plugging S_{AC} into Equation (8) yields an AC-aware REINVENT objective that rewards large predicted |Δp| under high similarity and rejects low-confidence cases.

Mmpdb[p99] complements AC-aware design by turning local, single-edit transformations into first-class actions for the generator, so that rewarded Δp reflects one auditable medicinal chemistry change rather than diffuse scaffold drift. Practically, mmpdb mines a rule set R of MMP transforms from a project library (Hussain–Rea cut and RDKit canonicalization; https://www.rdkit.org/docs/) and applies them as constrained edits during generation. Each rule $r \in R$ indicates a one-step proposal:

$$x\prime = Tr(x), \tag{11}$$

which is accepted only if locality/feasibility filters hold, for example:

$$sim_T(x,x') \geq s_{min}, \ |\Delta MW| \leq \delta_{MW}, \quad SA(x\prime) \leq SA_{max}, \tag{12}$$

with sim_T as Tanimoto similarity computed on ECFP, SA as a synthetic accessibility,[p100] and optional alerts such as PAINS applied as hard stops. AC-awareness is then layered in by scoring each accepted analog x' relative to a nearest neighbor $NN(x')$ drawn from a reference series S (seed set or growing memory). By using an AC-calibrated predictor f from Layer II, define the predicted AC strength as

$$ACI_{pred}(x\prime;f) = \frac{|f(x\prime)-f(NN(x\prime))|}{1-sim_T(x\prime,NN(x\prime))}, \tag{13}$$

and plug it into a gated AC score for RL in REINVENT:

$$S_{AC}(x) = 1\{(U-L) \leq \tau\} \cdot 1\{|f^*(x)| > \alpha_1\} \cdot \min\left(1, \frac{ACI_{pred}(x';f)}{\alpha_2}\right), \tag{14}$$

where [L, U] is the CP interval on uncertainty gate $f(x')$, f^* orients the objective (e.g., docking binding energy E_{dock} or pIC_{50}), and α_1, α_2 are potency/ACI thresholds as in the ACARL criteria. The augmented-likelihood objective in REINVENT keeps the language-model prior while rewarding AC-consistent moves:

$$R_{aug}(x') = \log p_{prior}(x') + \sigma S_{AC}(x\prime) \quad (15)$$

In this configuration, mmpdb guarantees that rewarded improvements come from single, interpretable edits within a series; nearest-neighbor anchoring enforces the local definition of an AC, and the CP gate suppresses noisy jumps.

Presented generative frameworks operationalize signal from ACs through RL architectures with scoring functions, weights, and rewards that steer AC-aware candidate design. However, the flywheel of the drug-design pipeline is the activity prediction of generated molecules, which can be misleading when the model used for this task is overconfident or overfitted to AC logic. This failure mode yields high-scoring models that become experts at maximizing predicted activity with unrealistic candidates exploiting the model's flaws, the opposite of making cheminformatics more laboratory-oriented. To prevent this, tune each generative model with chemical boundaries that limit generation by medicinal chemistry reality rules. Concretely, use a synthesizability score (e.g., RDKit's SA_score) to inject knowledge of synthetic accessibility and reduce the probability of proposing impossible-to-create candidates; apply general ADMET (absorption, distribution, metabolism, excretion, toxicity) filtering to constrain liabilities; and include QED (quantitative estimate of drug-likeness) scoring to balance potency with drug-likeness. These constraints should be enforced early (during training and sampling), not only as post-hoc filters, and can be complemented by simple novelty and diversity caps to avoid trivial exploitation.

Conclusions

ACs create dense, high-value regions of chemical space that are deliberately explored by medicinal chemists and can point directly toward new active chemotypes. To properly exploit the knowledge concentrated at the edges of SARs, *in silico* models must pay special attention to those regions, steering the modeling process. In practice, regularized 2D baselines (e.g., ECFP4 with SVM/GBM) often match or surpass more complex, deep architectures on AC-enriched tests under fair, series-aware splits and careful data curation. Deeper models earn their place when the task supplies the right inductive bias: AC-aware objectives or pretraining that shape latent distances around Δp, trustworthy 3D shape/conformer information where ACs are genuinely shape-driven, and multimodal target–ligand embeddings that capture protein context. This can be complemented by a structure-based design that proactively reads the AC signal in

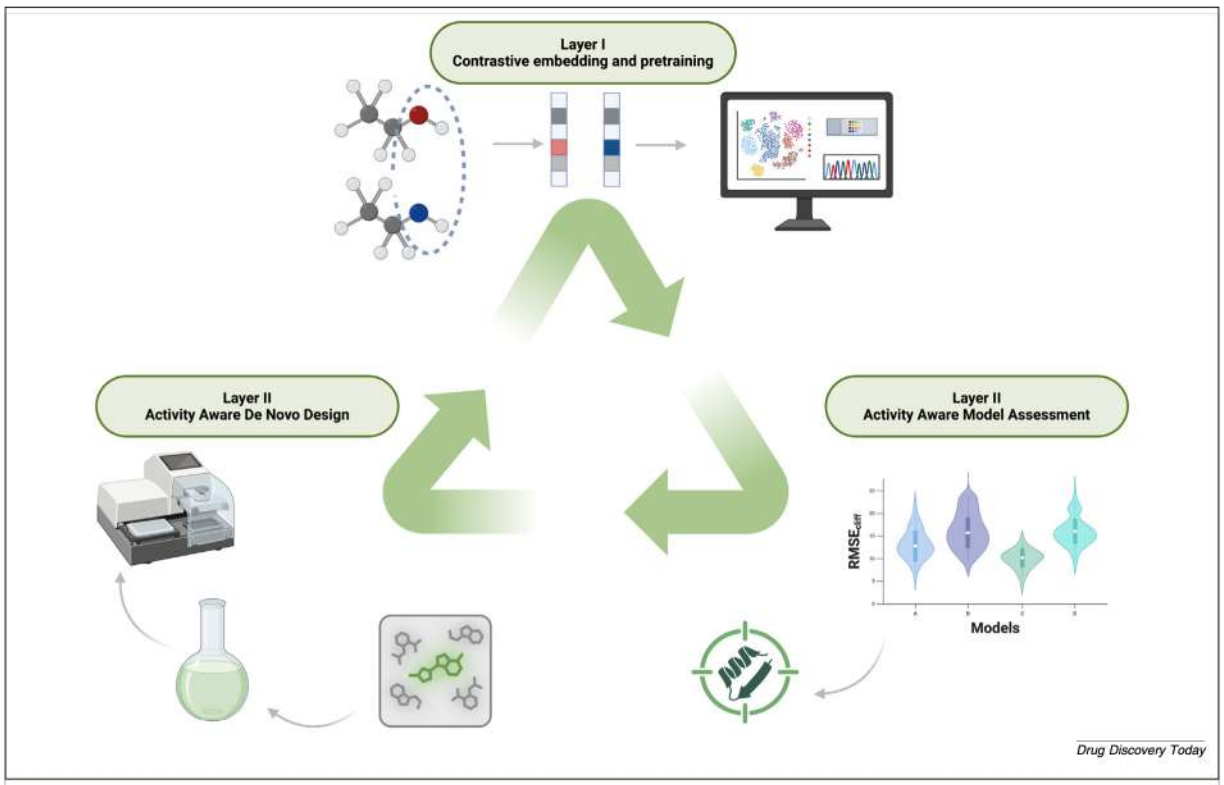


FIGURE 4
A simplified roadmap for AC-aware modeling, built on a self-improving, three-layered, closed loop.

three dimensions. Close analogs aligned in a common binding pose frequently expose the minimal interaction changes – such as the gain or loss of a hydrogen bond or salt bridge, a halogen-bond geometry shift, or a narrow steric gate – that convert a smooth SAR into discontinuity. Docking and virtual screening studies already show that 3D analyses can rationalize, and occasionally anticipate, AC behavior when shape complementarity and electrostatics dominate, whereas 2D AC mining alone can miss such cases. Looking ahead, progress will likely come from better embedding and pretraining schemes that bind ACs to structure, including physics-informed training signals and related formulations, rather than from capacity alone.[p101] Treating ACs as a signal, not a nuisance, reframes both benchmarking and design practice and can raise the quality of proposed molecules when applied with disciplined curation and fair evaluation.

To satisfy practitioners, we propose a three-layered dynamic loop: a guidance for cheminformaticians in creating effective drug design generators, built on the foundation of leveraging local SAR discontinuities, to outperform conventional AC-blind workflows. Layer I concentrates sampling and pretraining on steep regions so that molecular representations carry AC-relevant geometry into downstream tasks. Layer II then evaluates ACs using early recognition and pair-triage metrics that surface the most informative micro-edits responsible for large potency flips, while calibrated uncertainty keeps decisions honest in high-gradient neighborhoods. The resulting embeddings recalibrate the predictors, and generators in Layer III propose candidates that explicitly probe uncertain and AC-rich SAR regions flagged by Layer II predictors, rather than only optimizing the static scorer. In practice, this yields the cycle shown in Figure 4, Represent → Evaluate → Generate → Test → Represent, in which a newly identified AC can trigger contrastive re-weighting in Layer I, an uncertainty spike can redirect Layer III toward targeted exploration rather than exploitation, and each synthesis round both validates and reshapes the latent geometry used upstream. As a result, ACs become the signals that drive the loop, not just anomalies to be explained away.

Funding sources

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of interests

T.S. reports financial support was provided by Medical University of Warsaw. D.S. declares no competing financial interests.

Reproducibility package

All configuration files, environment specifications, fixed seeds, example ChEMBL query notes, and code reproducing the figures are available at: https://github.com/TomaszSzostek/AC_review.git.

CRediT authorship contribution statement

Tomasz Szostek: Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Daniel Szulczyk:** Writing – review & editing, Supervision.

Appendix A. Supplementary material

Supplementary material to this article can be found online at https://doi.org/10.1016/j.drudis.2025.104589.

Data availability

The data that has been used is confidential.

References

1. Qureshi R et al. AI in drug discovery and its clinical relevance. *Heliyon*. 2023;9, e17575. https://doi.org/10.1016/j.heliyon.2023.e17575.
2. Zhong F et al. Artificial intelligence in drug design. *Sci China Life Sci*. 2018;61:1191–1204. https://doi.org/10.1007/s11427-018-9342-2.
3. Selvaraj C, Chandra I, Singh SK. Artificial intelligence and machine learning approaches for drug design: challenges and opportunities for the pharmaceutical industries. *Mol Divers*. 2022;26:1893–1913. https://doi.org/10.1007/s11030-021-10326-z.
4. Sharma V et al. Role of artificial intelligence in drug discovery and target identification in cancer. *Curr Drug Deliv*. 2024;21:870–886. https://doi.org/10.2174/1567201821666230905090621.
5. Jiménez-Luna J, Grisoni F, Weskamp N, Schneider G. Artificial intelligence in drug discovery: recent advances and future perspectives. *Expert Opin Drug Discov*. 2021;16:949–959. https://doi.org/10.1080/17460441.2021.1909567.
6. Selvaraj C, Elakkiya E, Prabhu P, Velmurugan D, Singh SK. Advances in QSAR through artificial intelligence and machine learning methods. In: Hong H, ed. *QSAR in Safety Evaluation and Risk Assessment*. Academic Press; 2024:101–116. https://doi.org/10.1016/B978-0-443-15339-6.00033-3.
7. Tang X et al. A survey of generative AI for de novo drug design: new frontiers in molecule and protein generation. *Brief Bioinform*. 2024;25, bbae338. https://doi.org/10.1093/bib/bbae338.
8. Tong X et al. Generative models for de novo drug design. *J Med Chem*. 2021;64:14011–14027. https://doi.org/10.1021/acs.jmedchem.1c00927.
9. Crucitti D, Pérez Míguez C, Díaz Arias Jà, Fernandez Prada DB, Mosquera OA. De novo drug design through artificial intelligence: an introduction. *Front Hematol*. 2024;3, 1305741. https://doi.org/10.3389/frhem.2024.1305741.
10. Sertkaya A, Beleche T, Jessup A, Sommers BD. Costs of drug development and research and development intensity in the US, 2000–2018. *JAMA Netw Open*. 2024;7, e2415445. https://doi.org/10.1001/jamanetworkopen.2024.15445.
11. Sun D, Gao W, Hu H, Zhou S. Why 90% of clinical drug development fails and how to improve it? *Acta Pharm Sin B*. 2022;12:3049–3062. https://doi.org/10.1016/j.apsb.2022.02.002.
12. Wang J, Lomakin IB, Batista VS, Bunick CG. A triple-action inhibitory mechanism of allosteric TYK2-specific inhibitors. *J Invest Dermatol*. 2025;145:3158–3173. https://doi.org/10.1016/j.jid.2025.04.025.
13. Cheng F et al. Artificial intelligence and open science in discovery of disease-modifying medicines for Alzheimer's disease. *Cell Rep Med*. 2024;5, 101379. https://doi.org/10.1016/j.xcrm.2023.101379.
14. Maggiora GM. On outliers and activity cliffs: why QSAR often disappoints. *J Chem Inf Model*. 2006;46:1535. https://doi.org/10.1021/ci060117s.
15. Hu H, Bajorath J. Activity cliffs produced by single-atom modification of active compounds: systematic identification and rationalization based on X-ray structures. *Eur J Med Chem*. 2020;207, 112846. https://doi.org/10.1016/j.ejmech.2020.112846.
16. Dimova D, Bajorath J. Advances in activity cliff research. *Mol Inform*. 2016;35:181–191. https://doi.org/10.1002/minf.201600023.
17. Stumpfe D, Hu H, Bajorath J. Advances in exploring activity cliffs. *J Comput Aided Mol Des*. 2020;34:929–942. https://doi.org/10.1007/s10822-020-00315-z.
18. Cruz-Monteagudo M, Medina-Franco JL, Pérez-Castillo Y, Nicolotti O, Cordeiro F, Borges F. Activity cliffs in drug discovery: Dr Jekyll or Mr Hyde? *Drug Discov Today*. 2014;19:1069–1080. https://doi.org/10.1016/j.drudis.2014.02.003.

19. Vogt M, Huang Y, Bajorath J. From activity cliffs to activity ridges: informative data structures for SAR analysis. *J Chem Inf Model*. 2011;51:1848–1856. https://doi.org/10.1021/ci2002473.
20. Bajorath J. Duality of activity cliffs in drug discovery. *Expert Opin Drug Discov*. 2019;14:517–520. https://doi.org/10.1080/17460441.2019.1593371.
21. Stumpfe D, Bajorath J. Exploring activity cliffs in medicinal chemistry. *J Med Chem*. 2012;55:2932–2942. https://doi.org/10.1021/jm201706b.
22. Guha R. On exploring structure–activity relationships. In: Kortagere S, editor. *Silico Models for Drug Discovery (Methods in Molecular Biology)*, vol. 993. Totowa, NJ: Humana Press; 2013:81–94. https://doi.org/10.1007/978-1-62703-342-8_6.
23. Rogers D, Hahn M. Extended-connectivity fingerprints. *J Chem Inf Model*. 2010;50:742–754. https://doi.org/10.1021/ci100050t.
24. Durant JL, Leland BA, Henry DR, Nourse JG. Reoptimization of MDL keys for use in drug discovery. *J Chem Inf Comput Sci*. 2002;42:1273–1280. https://doi.org/10.1021/ci010132r.
25. Carhart RE, Smith DH, Venkataraghavan R. Atom pairs as molecular features in structure–activity studies: definition and applications. *J Chem Inf Comput Sci*. 1985;25:64–73. https://doi.org/10.1021/ci00046a002.
26. Nilakantan R, Bauman N, Dixon JS, Venkataraghavan R. Topological torsion: a new molecular descriptor for SAR applications. Comparison with other descriptors. *J Chem Inf Comput Sci*. 1987;27:82–85. https://doi.org/10.1021/ci00054a008.
27. Bajusz D, Rácz A, Héberger K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J Cheminform*. 2015;7:20. https://doi.org/10.1186/s13321-015-0069-3.
28. O'Hagan S, Kell DB. MetMaxStruct: a Tversky-similarity-based strategy for analysing the (sub)structural similarities of drugs and endogenous metabolites. *Front Pharmacol*. 2016;7:266. https://doi.org/10.3389/fphar.2016.00266.
29. Hariharan R et al. MultiMCS: a fast algorithm for the maximum common substructure problem on multiple molecules. *J Chem Inf Model*. 2011;51:788–806. https://doi.org/10.1021/ci100297y.
30. Gao X, Xiao B, Tao D, Li X. A survey of graph edit distance. *Pattern Anal Appl*. 2010;13:113–129. https://doi.org/10.1007/s10044-008-0141-y.
31. Togninalli M, Ghisu E, Llinares-López F, Rieck B, Borgwardt K. Wasserstein Weisfeiler–Lehman graph kernels. In: Wallach HM, Larochelle H, Beygelzimer F, d'Alché-Buc F, Fox EB, eds. *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Curran Associates; 2019:6439–6449.
32. Willett P. Similarity-based virtual screening using 2D fingerprints. *Drug Discov Today*. 2006;11:1046–1053. https://doi.org/10.1016/j.drudis.2006.10.005.
33. Matter H, Pötter T. Comparing 3D pharmacophore triplets and 2D fingerprints for selecting diverse compound subsets. *J Chem Inf Comput Sci*. 1999;39:1211–1225. https://doi.org/10.1021/ci980185h.
34. Nicholls A et al. Molecular shape and medicinal chemistry: a perspective. *J Med Chem*. 2010;53:3862–3886. https://doi.org/10.1021/jm900818s.
35. Khan SU, Ahemad N, Chuah L-H, Naidu R, Htar TT. Sequential ligand- and structure-based virtual screening for GPER-1 modulators. *RSC Adv*. 2019;9:2525–2538. https://doi.org/10.1039/C8RA09318K.
36. Hawkins PCD, Skillman AG, Nicholls A. Comparison of shape-matching and docking as virtual screening tools. *J Med Chem*. 2007;50:74–82. https://doi.org/10.1021/jm0603365.
37. Schreyer AM, Blundell T. USRCAT: real-time ultrafast shape recognition with pharmacophoric constraints. *J Cheminform*. 2012;4:27. https://doi.org/10.1186/1758-2946-4-27.
38. Zdrazil B et al. The ChEMBL database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Res*. 2024;52:D1180–D1192. https://doi.org/10.1093/nar/gkad1012.
39. Kalliokoski T, Kramer C, Vulpetti A, Gedeck P. Comparability of mixed IC_{50} data: a statistical analysis. *PLoS One*. 2013;8, e61007. https://doi.org/10.1371/journal.pone.0061007.
40. Landrum GA, Riniker S. Combining IC_{50} or K_i values from different sources is a source of significant noise. *J Chem Inf Model*. 2024;64:1560–1567. https://doi.org/10.1021/acs.jcim.4c00049.
41. Damiani E, Solorio JA, Doyle AP, Wallace HM. How reliable are *in vitro* IC_{50} values? *Toxicol Lett*. 2019;302:28–34. https://doi.org/10.1016/j.toxlet.2018.12.004.
42. Cheng Y, Prusoff WH. Relationship between the inhibition constant (K_I) and the concentration of inhibitor which causes 50 per cent inhibition (I_{50}) of an enzymatic reaction. *Biochem Pharmacol*. 1973;22:3099–3108. https://doi.org/10.1016/0006-2952(73)90196-2.
43. Burlingham BT, Widlanski TS. An intuitive look at the relationship of K_i and IC_{50}: a more general use for the Dixon plot. *J Chem Educ*. 2003;80:214. https://doi.org/10.1021/ed080p214.
44. Cheng HC. Determination of K_B or K_i from IC_{50}: a closer look at Cheng–Prusoff, Schild and related equations. *J Pharmacol Toxicol Methods*. 2001;46:61–71. https://doi.org/10.1016/S1056-8719(02)00166-1.
45. van Tilborg D, Alenicheva A, Grisoni F. Exposing the limitations of molecular machine learning with activity cliffs. *J Chem Inf Model*. 2022;62:5938–5951. https://doi.org/10.1021/acs.jcim.2c01073.
46. Ramírez-Palma D, Martinez-Mayorga K. Disruptions in bioactivity driven by dose: a challenge for drug discovery. *Drug Discov Today*. 2025;30, 104472. https://doi.org/10.1016/j.drudis.2025.104472.
47. Hu H, Stumpfe D, Bajorath J. Systematic identification of target set-dependent activity cliffs. *Future Sci OA*. 2019;5, FSO363. https://doi.org/10.4155/fsoa-2018-0089.
48. Rácz A, Bajusz D, Héberger K. Life beyond the Tanimoto coefficient: similarity measures for interaction fingerprints. *J Cheminform*. 2018;10:7. https://doi.org/10.1186/s13321-018-0302-y.
49. Dossetter AG, Griffen EJ, Leach AG. Matched molecular pair analysis in drug discovery. *Drug Discov Today*. 2013;18:724–731. https://doi.org/10.1016/j.drudis.2013.03.003.
50. 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. https://doi.org/10.1021/ci900450m.
51. Wawer M, Bajorath J. Local structural changes, global data views: graphical substructure–activity relationship trailing. *J Med Chem*. 2011;54:2944–2951. https://doi.org/10.1021/jm200026b.
52. Keefer CE, Chang G. The use of matched molecular series networks for cross-target SAR translation and potency prediction. *MedChemComm*. 2017;8:2067–2078. https://doi.org/10.1039/C7MD00465F.
53. Cao Y, Jiang T, Girke T. An MCS-based algorithm for searching and predicting drug-like compounds. *Bioinformatics*. 2008;24:i366–i374. https://doi.org/10.1093/bioinformatics/btn186.
54. Dimova D, Stumpfe D, Hu Y, Bajorath J. Analog series-based scaffolds: computational design and exploration. *Future Sci OA*. 2016;2, FSO139. https://doi.org/10.4155/fsoa-2016-0058.
55. Guha R, van Drie JH. Structure–activity landscape index: identifying and quantifying activity cliffs. *J Chem Inf Model*. 2008;48:646–658. https://doi.org/10.1021/ci7004093.
56. López-López E, Prieto-Martínez FD, Medina-Franco JL. Activity landscape and molecular modeling to explore dual epigenetic inhibitors. *Molecules*. 2018;23:3282. https://doi.org/10.3390/molecules23123282.
57. 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. https://doi.org/10.1021/jm401120g.
58. Poli G, Seidel T, Langer T. Conformational sampling of small molecules with iCon: performance vs OMEGA. *Front Chem*. 2018;6:229. https://doi.org/10.3389/fchem.2018.00229.
59. Irwin JJ et al. An aggregation advisor for ligand discovery. *J Med Chem*. 2015;58(17):7076–7087. https://doi.org/10.1021/acs.jmedchem.5b01105.
60. Garcia-Hernandez C, Fernández A, Serratosa F. Learning the edit costs of graph edit distance for ligand-based VS. *Curr Top Med Chem*. 2020;20:1582–1598. https://doi.org/10.2174/1568026620666200603122000.
61. Medina-Franco JL, Yongye AB, Pérez-Villanueva J, Houghten RA, Martínez-Mayorga K. Multitarget SARs via activity-difference maps and consensus similarity. *J Chem Inf Model*. 2011;51:2427–2439. https://doi.org/10.1021/ci200281v.
62. Yongye AB, Byler K, Santos R, Martínez-Mayorga K, Maggiora GM, Medina-Franco JL. Consensus models of activity landscapes with multiple representations. *J Chem Inf Model*. 2011;51:1259–1270. https://doi.org/10.1021/ci200081k.
63. Stumpfe D, Bajorath J. Frequency of occurrence and potency range distribution of activity cliffs in bioactive compounds. *J Chem Inf Model*. 2012;52:2348–2353. https://doi.org/10.1021/ci300288f.
64. Bento AP et al. The ChEMBL bioactivity database: an update. *Nucleic Acids Res*. 2014;42:D1083–D1090. https://doi.org/10.1093/nar/gkt1031.
65. Sheridan RP. Time-split cross-validation for estimating prospective prediction. *J Chem Inf Model*. 2013;53:783–790. https://doi.org/10.1021/ci400084k.
66. Baell JB, Holloway GA. New substructure filters for removal of PAINS. *J Med Chem*. 2010;53:2719–2740. https://doi.org/10.1021/jm901137j.
67. Fourches D, Muratov E, Tropsha A. Trust, but verify II: a practical guide to chemogenomics data curation. *J Chem Inf Model*. 2016;56:1243–1252. https://doi.org/10.1021/acs.jcim.6b00129.
68. Seal S, (Hartung T) Rabbit RR. Advice for bad computational toxicologists. *NAM J*. 2025;1, 100024. https://doi.org/10.1016/j.namjnl.2025.100024.

69. Hu Y, Stumpfe D, Bajorath J. Advancing the activity cliff concept. *F1000Research*. 2013;2:199. https://doi.org/10.12688/f1000research.2-199.v1.
70. Nguyen-Vo TH, Teesdale-Spittle P, Harvey JE, Nguyen BP. Molecular representations in bio-cheminformatics. *Memetic Comput*. 2024;16:519–536. https://doi.org/10.1007/s12293-024-00414-6.
71. Keriven N. Not too little, not too much: theoretical analysis of graph (over) smoothing. In: Koyejo S, Mohamed S, Agarwal A, Belgrave D, Cho K, Oh A, eds. *Proceedings of the 36th International Conference on Neural Information Processing Systems*. Curran Associates Inc.; 2022:2268–2281.
72. Radovanovic M, Nanopoulos A, Ivanovic M. Hubs in space: popular nearest neighbors in high-dimensional data. *J Mach Learn Res*. 2010;11:2487–2531.
73. Andronov M, Fedorov MV, Sosnin S. Exploring chemical reaction space with reaction difference fingerprints and parametric t-SNE. *ACS Omega*. 2021;6:30743–30751. https://doi.org/10.1021/acsomega.1c04778.
74. Akella LB, DeCaprio D. Cheminformatics approaches to analyse diversity in screening libraries. *Curr Opin Chem Biol*. 2010;14:325–330. https://doi.org/10.1016/j.cbpa.2010.03.017.
75. Yu X, Wang Y, Chen L, Li W, Tang Y, Liu G. ACtriplet: an improved deep learning model for activity cliff prediction by integrating triplet loss and pre-training. *J Pharm Anal*. 2025;15, 101317. https://doi.org/10.1016/j.jpha.2025.101317.
76. Shen W *et al*. Activity cliff-informed contrastive learning for molecular property prediction. *ChemRxiv*. Preprint posted online November 7, 2024. https://doi.org/10.26434/chemrxiv-2023-5cz7s-v2.
77. Valdez Córdova CM. Towards learning activity cliff-aware molecular representations [poster presentation]*International Conference on Machine Learning Conference: LatinX in AI (LXAI) Research Workshop 2024, Vienna, Austria*. https://doi.org/10.52591/lxai202407274.
78. Park J, Sung G, Lee S, Kang S, Park C. ACGCN: graph convolutional networks for activity cliff prediction between matched molecular pairs. *J Chem Inf Model*. 2022;62:2341–2351. https://doi.org/10.1021/acs.jcim.2c00327.
79. Cheng Z et al. MaskMol: knowledge-guided molecular image pre-training framework for activity cliffs with pixel masking. *BMC Biol*. 2025;23:279. https://doi.org/10.1186/s12915-025-02389-3.
80. Wu F. A semi-supervised molecular learning framework for activity cliff estimation. In: Larson K, ed. *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*:6080–6608.
81. Qiao J et al. Self-conformation-aware pre-training with substructure interpretability. *Nat Commun*. 2025;16:4382. https://doi.org/10.1038/s41467-025-59634-0.
82. Tang H *et al*. Contextual molecule representation learning from chemical reaction knowledge. *arXiv*. Preprint posted online February 21, 2024. https://doi.org/10.48550/arXiv.2402.13779.
83. Zhang Z, Zhao B, Xie A, Bian Y, Zhou S. Activity cliff prediction: dataset and benchmark. *arXiv*. Preprint posted online February 15, 2023. https://doi.org/10.48550/arXiv.2302.07541.
84. LeDonne NC, Rissolo K, Bulgarelli J, Tini L. Use of structure-activity landscape index curves and curve integrals to evaluate the performance of multiple machine learning prediction models. *J Cheminform*. 2011;3:7. https://doi.org/10.1186/1758-2946-3-7.
85. Saito T, Rehmsmeier M. The precision–recall plot is more informative than ROC on imbalanced data. *PLoS One*. 2015;10, e0118432. https://doi.org/10.1371/journal.pone.0118432.
86. Truchon J-F, Bayly CI. Evaluating virtual screening methods: good and bad metrics for early recognition. *J Chem Inf Model*. 2007;47:488–508. https://doi.org/10.1021/ci600426e.
87. Astigarraga M, Sánchez-Ruiz A, Colmenarejo G. Conformal prediction in cheminformatics: current applications and new challenges. *Artif Intell Life Sci*. 2025;7, 100127. https://doi.org/10.1016/j.ailsci.2025.100127.
88. Arvidsson McShane S, Norinder U, Alvarsson J, Ahlberg E, Carlsson L, Spjuth O. CPSign: conformal prediction for cheminformatics modeling. *J Cheminform*. 2024;16:75. https://doi.org/10.1186/s13321-024-00870-9.
89. Loeffler HH, Wan S, Klähn M, Bhati AP, Coveney PV. Optimal molecular design: generative active learning with binding free-energy ranking. *J Chem Theory Comput*. 2024;20:8308–8328. https://doi.org/10.1021/acs.jctc.4c00576.
90. Filella-Merce I et al. Optimizing drug design by merging generative AI with a physics-based active learning framework. *Commun Chem*. 2025;8:238. https://doi.org/10.1038/s42004-025-01635-7.
91. Jiménez-Luna J, Skalic M, Weskamp N. Benchmarking molecular feature attribution methods with activity cliffs. *J Chem Inf Model*. 2022;62:274–283. https://doi.org/10.1021/acs.jcim.1c01163.
92. Qi Z, Khorram S, Fuxin L. Visualizing deep networks by optimizing with integrated gradients. *arXiv*. Preprint posted online May 2, 2019. https://doi.org/10.48550/arXiv.1905.00954.
93. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization*2017 IEEE International Conference on Computer Vision (ICCV)*:618–626.
94. Van den Broeck G, Lykov A, Schleich M, Suciu D. On the tractability of SHAP explanations. *J Artif Intell Res*. 2022;74:851–886. https://doi.org/10.1613/jair.1.13412.
95. Li K et al. AMPCliff: quantitative definition and benchmarking of activity cliffs in antimicrobial peptides. *J Adv Res*. 2025. https://doi.org/10.1016/j.jare.2025.04.046. Epub ahead of print.
96. Weissenow K, Heinzinger M, Rost B. Protein language-model embeddings for fast, alignment-free protein structure prediction. *Structure*. 2022;30:1169–1177. https://doi.org/10.1016/j.str.2022.05.004.
97. Hu X, Liu G, Zhao Y, Zhang H. Activity cliff-aware reinforcement learning for de novo drug design. *J Cheminform*. 2025;17:54. https://doi.org/10.1186/s13321-025-01006-3.
98. Loeffler HH et al. Reinvent 4: modern AI-driven generative molecule design. *J Cheminform*. 2024;16:20. https://doi.org/10.1186/s13321-024-00812-5.
99. Dalke A, Hert J, Kramer C. mmpdb: an open-source matched molecular pair platform for large multiproperty data sets. *J Chem Inf Model*. 2018;58:902–910. https://doi.org/10.1021/acs.jcim.8b00173.
100. Fukunishi Y, Kurosawa T, Mikami Y, Nakamura H. Prediction of synthetic accessibility based on commercially available compound databases. *J Chem Inf Model*. 2014;54:3259–3267. https://doi.org/10.1021/ci500568d.
101. Lee M, Kim H, Joe H, Kim H-G. Multi-channel PINN: scalable and transferable neural networks for drug discovery. *J Cheminform*. 2019;11:46. https://doi.org/10.1186/s13321-019-0368-1.

19. Publication IV

De novo design of anticancer 4-thiazolidinone derivatives: a generative framework shaped by activity cliffs.

Szostek, T., Wiśniewski, M., Ballabio, D., Consonni, V., Plewczyński, D., & Szulczyk, D.

Journal of Cheminformatics, *18*(1), 88-. (2026).

DOI: https://doi.org/10.1186/S13321-026-01216-3/FIGURES/9

IF= 7.9, **MNiSW**=100

Szostek *et al. Journal of Cheminformatics* (2026) 18:88
https://doi.org/10.1186/s13321-026-01216-3

Journal of Cheminformatics

RESEARCH Open Access

De novo design of anticancer 4-thiazolidinone derivatives: a generative framework shaped by activity cliffs

Tomasz Szostek[1*], Maciej Wiśniewski[2], Davide Ballabio[3], Viviana Consonni[3], Dariusz Plewczyński[2,4] and Daniel Szulczyk[1]

Abstract

Activity cliffs (AC) correspond to large potency differences between highly similar compounds and pose a persistent challenge for both predictive modeling and de novo molecular design, particularly in small and underexplored areas of the chemical space. In this study, we introduce an AC-aware generative framework for the de novo design of anticancer thiazolidinone derivatives relevant to non-small cell lung cancer. We identify difficult regions of structure–activity landscape and extract chemical pattern signals hidden in its discontinuities directly from AC pairs using a dedicated algorithm. These signals are incorporated into a fragment-based generative pipeline to guide molecular construction and candidate selection. Compared with a standard QSAR-guided approach, the proposed framework yields a higher fraction of candidates with favorable docking profiles across multiple cancer-relevant targets and produces ligands with more stable binding modes in molecular dynamics simulations. Among the top five candidates selected by a molecular dynamics, multi-metric consensus combining docking affinity, binding stability (ligand and pocket RMSD/RMSF), and energetic criteria, four were generated using the AC-aware CAFE LATE strategy. The two highest-ranked hits exhibited binding energies of − 7.78 and − 7.70 kcal/mol, markedly more favorable than the native reference (− 4.09 kcal/mol), and consistently outperformed native ligands across all evaluated stability metrics. Those results suggest that explicit use of AC information can improve the quality of de novo generated small-molecule candidates in data-limited settings.

Scientific Contribution statement

This study introduces a new AC-informed small-molecule design framework that explicitly exposes crucial structure–activity information hidden in steep regions of the SAR landscape and uses it as an operationalized design signal. Our algorithm converts AC knowledge into reverse QSAR and generation leverage, enabling exploration of unknown areas of the chemical space, and discovery of the best performing candidates supported in molecular dynamics simulations. To the best of our knowledge, this is the first reverse QSAR and fragment-based generative framework to operationalize AC information at the fragment level within a unified machine learning workflow.

Keywords Activity cliffs, Drug discovery, 4-thiazolidinones, QSAR, Machine learning, Anticancer

*Correspondence:
Tomasz Szostek
tomasz.szostek@wum.edu.pl
Full list of author information is available at the end of the article

© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Introduction

Lung cancer remains the most diagnosed cancer and the leading cause of cancer death worldwide. In 2022 it accounted for ~ 2.48 million new cases (12.4% of all cancers) and ~ 1.82 million deaths (18.7%) [1]. Despite advances in screening and therapeutics, the absolute global burden continues to rise, driven by population aging and growth alongside persistent exposures to major risk factors. The World Health Organization projects more than 35 million new cancer cases per year by 2050 [2]. Consequently, the clinical management of the predominant histologic category of non-small cell lung cancer (NSCLC) centers on comprehensive biomarker testing to guide targeted and individualized therapies and immunotherapies, as reflected in contemporary international guidelines [3]. Nevertheless, durable disease control in advanced NSCLC remains uncommon because relapse and progression are frequent, reflecting tumor heterogeneity and acquired resistance (e.g. secondary epidermal growth factor receptor mutation [4]), bypass signaling [5], and lineage transformation [6]. Recent randomized trials have moved targeted and immune therapies into earlier stages, for example perioperative durvalumab improved event-free survival and pathologic complete response [7], and adjuvant alectinib prolonged disease-free survival after resection of anaplastic lymphoma kinase positive disease (ALINA) [8], but they have not eliminated relapse. Given the clinical reality and therapeutic pressure, there is a profound need for new small-molecules discovery, which could expand possibilities of NSCLC selective treatment.

Thiazolidinone is a widely investigated, pharmacologically versatile scaffold and remains a promising core chemotype for anticancer lead discovery. As summarized in prior reviews, thiazolidinones have attracted sustained attention from medicinal chemists owing to their broad spectrum of biological activities, including anticancer effects, and their synthetically accessible substitution patterns [9, 10]. Recent studies further underscore the anticancer potential of this pharmacophore. Podolak et al. (2025) reported a new series of highly active derivatives with lead achieving $IC_{50} = 0.85$ μM in MCF-7 human breast adenocarcinoma cells and 0.28 μM in Jurkat human T-cell leukemia cells, while exhibiting markedly weaker effects in normal human dermal fibroblasts (HDF $IC_{50} = 16.9$ μM), indicating useful selectivity [11]. In a complementary series, Shawky et al. (2023) disclosed 2,3-diaryl-1,3-thiazolidin-4-ones with IC_{50} values ranging from 0.02 to 17.02 μM across MCF-7 human breast adenocarcinoma cells, HT-29 human colorectal adenocarcinoma cells, and A2780 human ovarian carcinoma cells, along with in vivo tumor growth suppression in an Ehrlich solid carcinoma model [12]. Most importantly, several studies show that thiazolidinone derivatives exhibit measurable cytotoxicity against NSCLC models. For example, novel 4-thiazolidinone-based compounds were screened for in vitro anticancer activity and demonstrated inhibitory effects on the A-549 human cell line in MTT assays, with IC_{50} values in the low-to-mid micromolar range [13]. Moreover, related thiazolidinone hybrids have shown apoptosis induction and cell cycle arrest in A-549 cells, reinforcing the relevance of this scaffold to NSCLC drug discovery [14].

Yet the thiazolidinone chemical space understood as the conceptual space of compounds with common chemical motif, remains underexplored, particularly along substituent vectors relevant to potency, selectivity, and drug-like properties. This motivates a systematic, structure-guided expansion in search of the new effective anticancer agents.

Modern in silico methods, including Artificial Intelligence (AI) powered frameworks, have become a standard tool in medicinal chemistry for de novo small-molecules generation, giving a promise of effective digital data processing and analysis. These approaches typically rely on predictive models trained on large chemical databases, which learn structure–property and structure–activity relationships and subsequently generate or prioritize new molecular structures with desired biological profiles [15]. Statistically, discovery phase of active molecule lasts from 3 to 6 years, but applying AI driven computational aided drug design (CADD) can accelerate the process by even 2 years. Moreover, it can shorten the clinical trial duration by ~ 15—30% and significantly rise drug candidate's early-stage success (Phase I 80–90% vs historical 40–65%) [16, 17], as evidenced by small-molecule TNIK inhibitor INS018_055 which was designed using AI powered methodology and has successfully reached 2a phase of clinical trial targeting fibrosis [18]. Despite undeniable progress, several assessments caution that the real-world impact of AI on drug discovery remains uneven indicating necessary improvements in data quality management and models architecture [19].

From computational chemistry perspective, building AI models that capture a continuous structure–activity relationship (SAR) and reliable performance is primarily disrupted by pairs of very similar compounds with large potency differences (Δp)- Activity Cliffs (AC) [20, 21]. These discontinuities function as outliers that break the apparent correlation between shared chemical motifs and biological effects, where even single-atom edits can yield ≥ tenfold potency shifts [22]. As a result, generative and predictive models with tendency to generalization often produce overly optimistic estimates and inflated accuracy metrics, wasting time and resources, and risking overlooking of relevant candidates [23]. Importantly,

medicinal chemists view AC as a dense, high-gradient, information rich regions of SAR landscape, which can be operationalized in AI pipelines using dedicated techniques enhancing the models' performance [24]. At the molecular representation level, methods such as ACtriplet [25], ACANT [26], MaskMol [27] and SiamACLoss [28] inject pairwise/contrastive AC information during model training, explicitly separating AC neighbors in latent space and improving sample efficiency and reliability. For evaluation, MoleculeACE formalized AC-aware benchmarking by introducing $RMSE_{cliff}$ (Root Mean Square Error calculated on AC pairs) and subset metrics [23], while ACNet [29] complements this by casting AC detection as a pairwise classification problem. Finally, de novo generative modeling was reshaped to explore difficult SAR region with Activity Cliffs Aware Reinforcement Learning (ACARL) [30] via contrastive objectives.

Quantitative Structure Activity Relationship (QSAR) models learn a statistical relationship between molecular structure and biological activity [31]. Once trained, they can predict unknown activity for new compounds within a defined Applicability Domain (AD), understood as a pre-specified region of chemical space where the model's assumptions and uncertainty bounds justify using the prediction. Reverse QSAR (r-QSAR, often called inverse-QSAR) turns the problem around. Starting from desired property values or activity profiles, it searches the chemical space for structures likely to satisfy those targets, typically by coupling prediction model (QSAR) with a structure generator/decoder or chemically constrained mutation engine [32].

Rapid progress of AI and its implementation to QSAR modeling extended representation learning, and brought a great improvement to data interpretability, accelerating hypothesis testing in drug discovery pipelines. At the same time, extensive comparative studies in small-molecule frameworks, comparing Machine Learning (ML) and Deep Learning (DL) methods, consistently reveal superiority of simpler ML approaches. Rigorous evaluation under AC-aware conditions related to strict assessment metrics such as $RMSE_{cliff}$ further highlights this advantage [23]. As evidenced in recent analysis conducted on 62,820 trained models, classical molecular representation aligned with ML methods can be the best choice, especially for smaller datasets [33].

In this study, we continue to expand thiazolidinones chemical space to identify new candidates against NSCLC [34]. To increase the likelihood of discovering highly active analogues in a limited and underexplored dataset, we introduce an AC-informed generative framework built around r-QSAR and fragment-based design. In contrast to previously reported AC-aware approaches, which primarily incorporate cliff information at the level of model training, latent-space shaping, loss design, or reinforcement-learning control, our strategy operates at the level of chemically interpretable building blocks. Specifically, it extracts directional SAR information from AC pairs, resolves this signal into fragment-level enrichment rules, and reinjects the corrected fragment knowledge into r-QSAR-guided molecular generation. To the best of our knowledge, this is the first study to combine r-QSAR, fragment-level AC knowledge extraction, and de novo design within one coherent small-molecule discovery workflow. This new approach turns AC from sources of model error into actionable signal, focusing search on high-value regions of chemical space and yielding a small library of thiazolidinone derivatives with higher probability of desirable potency gains against NSCLC.

Methods

Data acquisition and standardization

To establish a robust modeling foundation and define actionable chemical space for downstream analysis, we enforced strict data management rules. We retrieved all compounds, bearing the thiazolidinone scaffold with reported activity against A549 adenocarcinoma cell line, commonly used as a model of NSCLC from ChEMBL database (CHEMBL392), using the platform's curated bioactivity records, and added 21 compounds previously synthesized by our team [34–36]. For endpoint consistency, we selected a single readout of IC_{50} values, derived exclusively from MTT colorimetric assay, and converted all measurements to pIC_{50}, in common nM unit for concentration [38]:

$$pIC_{50} = -log_{10}(IC_{50}[M]) = 9 - log_{10}(IC_{50}[nM]) \quad (1)$$

This choice aligns with best practice for harmonizing potency scales and leverages ChEMBL's conventions for comparability across studies. The dataset encoded as canonical SMILES was subjected to a standard curation workflow to minimize artifacts and leakage: removal of exact duplicates, resolution of tautomeric variants, stripping salts/solvates, and exclusion of structurally implausible or ill-specified records ("odd compounds"), though molecular weight (MolWt < 900 Da) and LogP < 8 filters. This follows established recommendations that rigorous, transparent curation is a precondition for reliable QSAR/ML modeling [39]. For subsequent classification tasks, each chemical of the cleaned set was binarized to an activity label y_i using the common cheminformatics threshold $IC_{50} \leq 10$ μM (= active, otherwise inactive), a cutoff frequently adopted in benchmarking and target-agnostic screens [40]:

$$y_i = \begin{cases} 1 & if\, IC_{50_i} \leq 10000nM \\ 0 & else \end{cases} \quad (2)$$

Chemical space and structure activity landscape

To align with community standards, we define AC as a pair of compounds with high 2D structural similarity computed on Morgan fingerprints ECFP4 (radius = 2, length = 2048) and Jaccard Tanimoto similarity index $\tau \geq 0.8$, combined with a $\Delta pIC_{50} \geq 1.0$ (10- fold difference in IC_{50}) using RDKit [23, 41, 42]. For a dataset of n valid molecules, all unordered pairs (i, j) with $i < j$ are enumerated. For each pair, we compute Tanimoto similarity s_{ij} and potency difference ΔpIC_{50}. A pair is retained as an AC if it meets both thresholds: high 2D similarity ($s_{ij} \geq \tau$, default $\tau = 0.80$) and large potency contrast ($\Delta pIC_{50} \geq \delta$, default $\delta = 1.0$ on IC_{50}). For every retained pair we quantify AC severity with Structure Activity Landscape Index *SALI* defined as:

$$SALI_{ij} = \frac{\Delta pIC_{50}}{1 - s_{ij} + \varepsilon} \quad (3)$$

where s_{ij} is Tanimoto similarity and $\varepsilon = 1 \times 10^{-6}$ is added in the denominator to guard against division by zero for near-identical pairs. *SALI* is used strictly as a diagnostic in this step and later as a weight in fragment enrichment in r-QSAR.

Two complementary visual diagnostics are produced directly from the fingerprints and potency values (these maps are only chemical space visualizations and are not used by any predictive model):

- Structure Activity Similarity (SAS) map: *SALI* scatter plots ΔpIC_{50} versus Tanimoto for all pairs, using a continuous colormap for *SALI* values on the subset that meets AC criteria. The similarity and potency thresholds (τ, δ) are drawn as dashed reference lines to delimit the AC region [43].
- t-SNE embedding with AC overlay: projects molecule level fingerprints to two dimensions using t-SNE [44]. Points are colored by pIC_{50}, and molecules participating in at least one AC (present in mol_i or mol_j) are overlaid as highlighted markers.
- Multidimensional Scaling (MDS)**:** projects molecular fingerprints into two dimensions using classical metric multidimensional scaling based on pairwise dissimilarities derived from ECFP4 fingerprints. The dissimilarity matrix is computed as 1—Tanimoto similarity, and molecules are embedded such that inter-point distances best preserve the original dissimilarities in the low-dimensional space. Points are colored by pIC_{50} values, while molecules participating in at least one AC pair are marked as triangles. This representation provides a global, distance-preserving baseline view of the chemical space, complementary to the nonlinear t-SNE projection [45].

Machine learning models

Both the r-QSAR and prediction-QSAR model were trained on 9 classical ML engines using scikit learn and imbalanced libraries [45]: logistic regression (LogReg), linear support vector machine (Linear SVM), radial basis function support vector machine (RBF-SVM; SVC), k-nearest neighbors (k-NN), random forest (RF), balanced random forest (BRF), ExtraTrees, gradient boosting (GB), XGBoost, and CatBoost.

Reverse QSAR

We trained a r-QSAR classifier across nine distinct ML engines to identify the best one suited for the task: mining chemical patterns mostly accountable for anticancer activity against A549 cancer cell line. All compounds in the final thiazolidinone dataset were defragmented to chemically meaningful fragments using RDKit BRICS [42]. Those fragments were normalized to customized SMILES in which broken bonds are marked with attachment points [*]. We then build a binary fragment incidence matrix:

$$X \in \{0,1\}^{N \times F}, X_{i,f} = 1, iffispresentinmoleculei \quad (4)$$

where N is the number of molecules in training set, and $\mathcal{L} = \{f_1, ., f_F\}$ is the vocabulary of normalized fragments (each f corresponds to one column of X). This representation is analogous to MACCS keys [47], and enables the classifier to learn SAR as fragment presence/absence and to rank fragments by learned importance [48]. This preserves the highest learned importance per normalized fragment. As a sanity filter, any fragment observed only in the inactive class has its importance set to 0 before selection to prevent one-class artifacts.

To avoid indirect data leakage from splitting AC pairs across folds, we generated ten independent repeats of a stratified, group-aware train/test split. For each repeat:

- We constructed the set of molecules that participate in at least one AC pair (from the *SALI* analysis table).
- All molecules in this set were forced into train (AC pairs and their neighbors remain together, preserving the AC logic).
- The remaining non-AC molecules were split by class-stratified sampling to form test at the target proportion (default 20%) the rest joined train.

- Class labels remain balanced across repeats by stratification. Random seeds were varied per repeat to ensure independence.

This policy keeps AC relationships intact and prevents the model from "seeing one side of an AC" in train and the other in test set, which would inflate AC related performance. The operating point (classification threshold) used for thresholded metrics within each repeat was chosen by scanning probability scores $\rho \in [0; 1]$ and selecting a single operating threshold that maximized the designated objective (F1-score by default) on that repeat's held-out set.

We summarize r-QSAR performance by the mean over the 10 repeats of the following metrics: $AUPRC_{active}$ emphasizing performance on the positive class under imbalance (chosen as a primary metric), AUROC, Accuracy, Precision, Recall, F1- score, Matthews Correlation Coefficient (MCC). For each engine we report the mean of these metrics across the 10 repeats the best model is selected by maximizing mean $AUPRC_{active}$. Tie-breakers, when needed, follow (in order) higher AUROC, higher MCC, and higher F1 at the chosen operating point.

The best r-QSAR model builds the fragment library from the ranked fragment-importance profile. Let $I(j)$ denote the learned importance of fragment f, and let $I(j)$ denote the corresponding importance value after sorting in descending order. The cumulative importance curve is then defined as:

$$g(k) = \frac{\sum_{j=1}^{k} I_{(j)}}{\sum_{j=1}^{F} I_{(j)}}, k = 1, .., F \tag{5}$$

where k is the number of top-ranked fragments included in the cumulative sum. Among all importance mass, $g(k)$ shows what fraction is already covered by the top-k fragments. The final number of important fragments is chosen by Kneedle detector [49], which searches for natural bend in importance curve ("the elbow"), and keep the compact, high-signal features.

This split strategy minimizes pairwise leakage but also renders the AC-free, and therefore easier test set, so the reported global metrics are overoptimistic. The performance is expected to degrade on real-world scenarios. To recover SAR signal hidden inside AC and correct r-QSAR fragments choices, we introduce a new logic Cliff Active Fragment Extraction (CAFE): a lightweight, training-only postprocessor. CAFE requires exactly three artifacts:

- A binary fragment incidence matrix X from Eq. (4)
- SMILES list of the same N molecules, to map AC pairs to row indices in X.
- Table of listed AC pairs (i, j) with ΔpIC_{50} values, Tanimoto similarity s_{ij}, potency gap ΔpIC_{50}, and $SALI$.

CAFE processes each AC pair in turn. It first checks that both molecules are present in the X and that ΔpIC_{50}, values are available. The partner with the highest ΔpIC_{50} is treated as the more active molecule (i^+), the partner with lower ΔpIC_{50} as inactive (i^-), and the pair is oriented:

$$i^+ = argmax\{pIC_{50}\}, i^- = argmin\{pIC_{50}\} \tag{6}$$

Each molecule is represented as a binary vector over the same fragment set, where 1 means the fragment is present and 0 means it is absent. Let x_{i^+} and x_{i^-} denote the binary fragment-incidence vectors of the more active and less active analogues, respectively, corresponding to their rows in matrix X. CAFE performs element-wise subtraction to obtain a signed delta vector:

$$d_ij = x_{(i^+)} - x_{(i^-)}\{-1, 0, +1\}^F \tag{7}$$

which encodes fragment by fragment, whether a feature helps explain the potency flip for this pair. Each entry of d_{ij} takes one of three values $\{-1, 0, +1\}$. Positive one means the fragment is present only in the more active partner. Negative one means the fragment is present only in the less active partner. Zero means the fragment is present in both partners or in neither partner, nor carries information about the potency flip:

$$d_{ij,}(f) = \begin{cases} +1, & if x_{i^+(f)} = 1 and x_{i^-(f)} = 0 \\ 0, & if x_{i^+(f)} = x_{i^-(f)} \\ -1, & if x_{i^+(f)} = 0 and x_{i^-(f)} = 1 \end{cases} \tag{8}$$

where f indexes the customized fragment vocabulary $\mathcal{L} = \{f_1, \ldots, f_F\}$. Each f corresponds to one column of the binary matrix X (one feature). For each pair, a fragment that occurs only in the more active partner receives a positive contribution equal to the $SALI$ value of that pair. A fragment that occurs only in the less active partner receives a negative contribution of the same magnitude. Fragments with delta equal to zero are ignored. $SALI$ values are clipped to 99th percentile q to prevent single extreme pairs from dominating the sum:

$$\omega_{ij} = min(SALI_{ij}, c), c = quantile_q(\{SALI_{ij} : (i, j) \in Train\}) \tag{9}$$

In the present dataset, this cap corresponded to a $SALI$ value of 14.92 and affected only 1 of 47 AC pairs (2.13%). Comparison of fragment rankings obtained with and without clipping showed complete preservation of the main enrichment signal (top 10 overlap = 10/10; top 20 overlap = 20/20; Spearman correlation = 1.000),

indicating that clipping acted purely as a stabilization step preventing disproportionate influence from a rare extreme pair, rather than materially altering fragment prioritization.

The per-pair contribution is then:

$$e_{ij}(f) = \omega_{ij} d_{ij}(f) \tag{10}$$

This comparison is repeated across all valid AC pairs in the training data, and the contributions are accumulated per fragment. A fragment that is repeatedly unique to the more active side builds a large positive Enrichment Factor $E_{(f)}$) and becomes a candidate to add. A fragment that is repeatedly unique to the less active side builds a large negative score and becomes a candidate to remove:

$$E_{(f)} = \sum_{(i,j)\in Train} e_{ij}(f) = \sum_{(i,j)\in Train} \omega_{ij} d_{ij}(f) \tag{11}$$

The output is a structured table that maps each fragment to its total AC enrichment score and lists the supporting AC pairs with their row indices, so every recommendation has traceable provenance. Figure 1 illustrates the basic scoring on a toy example. Consider AC pair consisting of more potent analogue i^+ (active) and less potent analogue i^- (inactive), represented by binary fragment presence vectors x_{i+}, and x_{i-} respectively, over three fragments $f_1, f_2 and f_3$ with a *SALI* value of 3.2. Subtracting the binary vectors yields a directed d_{ij} with entries -1, 0, or +1. A value of +1 indicates that the fragment is present only in the more active analogue and therefore contributes +3.2 to its enrichment score. A value of -1 indicates that the fragment is present only in the less active analogue and therefore contributes -3.2. A value of 0 means that the fragment is shared by both analogues or absent from both and thus does not contribute to the potency flip. When such contributions are accumulated across many AC pairs, local SAR discontinuities are transformed into a stable fragment-level enrichment signal. The final fragment library is then corrected by CAFE through promotion of fragments repeatedly associated with the more active side of AC pairs and demotion of fragments repeatedly associated with the less active side.

Prediction QSAR

We trained a binary classification QSAR across the same nine ML engines using four molecular representations: circular Morgan/ECFP fingerprints with radius 2 at 1024 and 2048 bits (hash-based, neighborhood-expansion), MACCS structural keys (classical 166-bit MDL keyset), and 31 physicochemical descriptors selected by Boruta feature selection algorithm from a pool of ≈1500 computed features (all-relevant wrapper over Random Forest) [50]. Continuous features are z-standardized, binary fingerprints are used as is. Performance is estimated over ten independent repeats. For each repeat, we build AC groups from the *SALI* table and perform a stratified, group-aware split that keeps each AC component in the

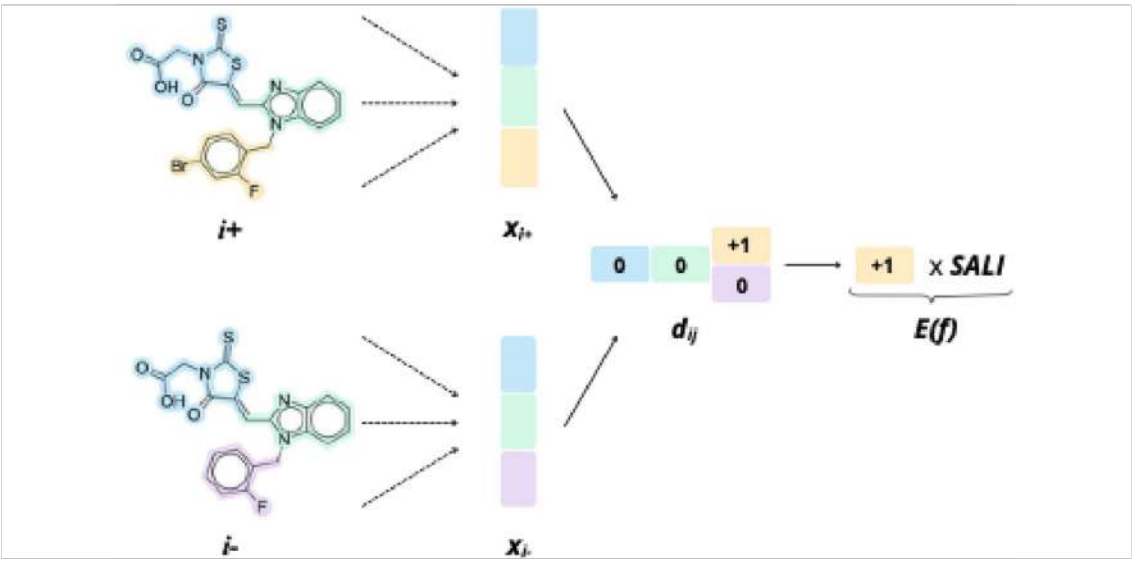


Fig. 1 Operational principle of CAFE on a toy example. Fragment-level differences between the more and less potent analogues are encoded by the signed vector $d_{ij} = x_{i+} - x_{i-}$, allowing AC pairs to generate positive, negative, or null fragment-enrichment contributions

same fold (train or test), with test ≈ 20% and class proportions preserved [51]. Unlike r-QSAR (where test was AC free), here AC can appear in test, enabling AC-aware error estimates. After fitting on train, each model produces probabilities $\rho \in [0,1]$. We select operating threshold $t*$ per repeat by grid search on a held-out validation split of train:

$$t* = argmax MCC_{(t)} \tag{12}$$

where $MCC_{(t)}$ denotes the Matthews Correlation Coefficient evaluated at decision threshold t on the validation split.

Model selection prioritized AC-aware accuracy: we chose the configuration with the lowest and most stable $RMSE_{cliff}$ across repeats, the smallest mean and the narrowest 95% CI on AC compounds, as recommended by MoleculeACE benchmark [23]. Because our prediction task is binary (active vs inactive), we adapt this AC-aware evaluation principle and assess models on the AC subset using the squared error between predicted probabilities and binary labels- the Brier score [52], reported as its square root ($RMSE$). For each repeat r, let $C^{(r)}$ be the test molecules that belong to at least one AC pair, then:

$$RMSE_{cliff}^{(r)} = \sqrt{\frac{1}{|C^{(r)}|}\sum_{i \in C^{(r)}} (\hat{p}_i - y_i)^2} \tag{13}$$

This metric is a proper scoring rule for Bernoulli outcomes (minimized at the true class probability), hence it jointly rewards calibration and accuracy on the steep SAR region [53]. For aggregation across repeats, let $(r = 1, \ldots, R)$, and $c_{(r)} = RMSE_{cliff}^{(r)}$, then:

$$\tilde{c} = \frac{1}{R}\sum\nolimits_{r=1}^{R} c_{(r)} \tag{14}$$

$$CI95\% = [Q_{2.5}, Q_{97.5}] \tag{15}$$

$$\mu = Q_{97.5}(c_{(r)}) - Q_{2.5}(c_{(r)}) \tag{16}$$

The best model is chosen based on the lowest final score (Sc) from aggregated mean $RMSE_{cliff}$ from ten repeats, combined with confidence interval width correction (μ):

$$Sc = \tilde{c} + \mu \tag{17}$$

This strategy allows to choose model with the lowest and most stable predictions on AC pairs.

De novo *design*

To expand the thiazolidinone chemical space we use an island framework [54]. Each island is anchored on a distinct thiazolidinone core extracted from the final fragment library, and generation proceeds independently on each island, so coverage is spread across cores rather than collapsing onto a single neighborhood. Fragments and cores expose attachment points as [*] in SMILES. The generator aligns a library fragment to a compatible attachment on a core and builds a candidate molecule until no [*] attachment is left. Fragment sampling is two-phase. In the coverage phase, we iterate across the deduplicated fragment set on every core, so each core–fragment combination is tried at least once, fragment order is skewed to under-used pieces to prevent early collapse. After roughly 30% of the target is reached, we switch to a bandit phase: each fragment carries a probability that is updated from its recent post-generation performance. A small, fixed exploration mass is retained, so learning never shuts off diverse fragments. Island scheduling mirrors this logic. During coverage, islands are visited round-robin to ensure balanced growth. In bandit mode, island selection becomes performance-aware so that productive cores get proportionally more attempts without starving others.

Every candidate must pass basic medicinal chemistry gates before scoring. We require successful sanitization in RDKit terms (valence, aromaticity, formal charge consistency), full resolution of all placeholder atoms, a non-trivial size window (atom count 2–200), valid SMILES, and a molecular weight between 150 and 800 Da. Only molecules that satisfy these rules are forwarded to evaluation. After passing basic gates, candidate is transferred to multi-objective scoring consisting of the best selected QSAR classifier, QED [55] (Quantitative Estimate of Drug-Likeliness), RDKit implementation of Ertl–Schuffenhauer SA score (measuring synthetic accessibility), which was invert-normalized:

$$SA\prime = 1 - min(\frac{SA - 1}{9}, 1) \tag{18}$$

where lower raw SA scores indicate easier synthesis, and the transformed $SA\prime$ rescales this quantity so that higher values correspond to more favorable synthetic accessibility.

The aggregate score for molecule (m) used for ranking the best candidates is a convex combination:

$$\begin{aligned}\varphi(m) = {} & \beta_{QSAR}\hat{p}(m) + \beta_{SA}SA\prime(m) \\ & + \beta_{QED}QED(m) + \lambda_{CAFE}B(m)\end{aligned} \tag{19}$$

where $\hat{p}(m)$ is the QSAR predicted probability, $B(m)$ is the CAFE LATE bonus defined below, and β $(\beta_{QSAR} = 0.4, \beta_{SA} = 0.4, \beta_{QED} = 0.2)$ are the tunable weights for each of the scorers. Then λ_{CAFE} controls the contribution of the AC-aware CAFE LATE bonus to the

final ranking score. However, the baseline predictor can still underestimate analogues that carry favorable AC-enriched fragments, particularly near steep SAR regions. To address this, we introduce CAFE LATE, a lightweight post hoc selection-stage extension of CAFE. Whereas CAFE operates during fragment prioritization by extracting enrichment signal from AC pairs, CAFE LATE reuses this supervised fragment knowledge during candidate ranking by assigning an additional bonus to newly generated molecules containing positively enriched AC fragments.

For each newly synthesized molecule m, it checks whether each AC fragment f occurs in m using RDKit substructure match of the normalized fragment pattern (same key used by CAFE). Presence is encoded by the indicator:

$$\delta_f(m) = 1\{f \subseteq m\} \tag{20}$$

Multiple occurrences of the same fragment f in molecule m are counted once (prevents length bias from repeated substructures). CAFE LATE converts fragment level evidence into a single signed bonus by summing the previously computed $E(f)$ over fragments found in m:

$$B(m) = \sum_{f \in f_{AC}} \delta_f(m) E(f) \tag{21}$$

AC positive motifs ($E(f) > 0$) increase the ranking score through trough $B(m)$, and molecules containing no positively enriched CAFE fragments have $B(m) = 0$ and therefore remain unchanged by CAFE LATE.

Final selection proceeds in two parallel arms. In the vanilla arm, used here to denote the baseline workflow without AC-derived post hoc boosting, we set $\lambda = 0$ and select the best 100 compounds by ranked aggregated score $\varphi(m)$. In the CAFE arm we run the same pipeline repeatedly over a grid of $\lambda_{CAFE} = \{0.4; 0.6; 0.8; 1.0; 1.5\}$ values, each time independently producing the best 100 hits ranked by $\varphi(m)$. Final selection follows a multi-stage funnel that guarantees quality and diversity while always filling to the target size. We first pre-rank candidates by $\varphi(m)$ and retain a working pool. We then compute a Pareto front on the three base objectives QSAR, SA′ and QED to keep only non-dominated solutions. The front is re-ranked by $\varphi(m)$, after which we apply diversity selection with ECFP4 (radius 2, 2048 bits) and Tanimoto similarity using a greedy best-first procedure that starts from the highest-scoring molecule and iteratively admits the next most dissimilar candidate. If diversity constraints make the set too small the similarity threshold is relaxed stepwise until the target count is reached. A minimum representation per core scaffold is enforced, canonical SMILES de-duplication is applied, and ties are broken by $\varphi(m)$. Each ablation arm yields exactly 100 molecules. For discovery-oriented analyses we additionally build a union of all unique candidates from CAFE boosted runs, which is used downstream to surface "all-star" candidates irrespective of the specific weight that promoted them. The simplified scheme of new molecules generator with enhanced AC-aware methods CAFE and CAFE LATE is presented on Fig. 2.

Molecular docking

Protein structures were retrieved from the Protein Data Bank (PDB) [56], identified by their PDB IDs (1M17, 4I23, 8GUA, 4EKL, 5OQ4, 4LMN, 5HD4, 6SCM, 5EHR, 4L7B, 5L2S, 5L2I, 5L2T), stored locally and preprocessed to remove alternate conformations, water molecules, ions, and crystallization additives. Hydrogen atoms were added, and protonation states were assigned using pdbfixer [57]. Cleaned receptor structures were converted to PDBQT format using AutoDockTools [58]. Ligand structures, provided as SMILES representations, were transformed into 3D conformations using Open Babel [59] and converted to PDBQT format. Ligand protonation states were assigned during 3D preparation using the standard Open Babel workflow under consistent, near-physiological conditions across the full ligand set.

The docking grids were automatically generated based on native ligand coordinates. The grid box was defined using the ligand center of mass with a padding of 4 Å to ensure the binding site was fully encompassed [60]. Docking simulations were conducted using Smina, an optimized fork of AutoDock Vina [61], using a fixed protocol across all receptors and ligands. Parameters were set to default values, including exhaustiveness of 24 and 10 docking poses per ligand-receptor pair. Native ligand re-docking served as validation, with the root mean square deviation (RMSD) calculated to assess docking accuracy. An RMSD $\leq$ 2.0 Å was considered a successful redocking [62].

Docking poses were evaluated based on calculated binding affinities (kcal/mol). For downstream comparative analyses, the retained pose for each ligand–receptor pair was defined as the lowest-energy valid pose returned under the same docking setup. Docking scores were used as a comparative prioritization criterion within this fixed protocol rather than as absolute estimates of binding affinity.

Ligands with docking scores more favorable than those of the corresponding native ligands under the same protocol were selected as preliminary hits for further refinement. Interaction profiles of the best docking poses were analyzed using the Protein–Ligand Interaction Profiler (PLIP) [63], generating detailed reports of interaction types such as hydrogen bonds, hydrophobic interactions,

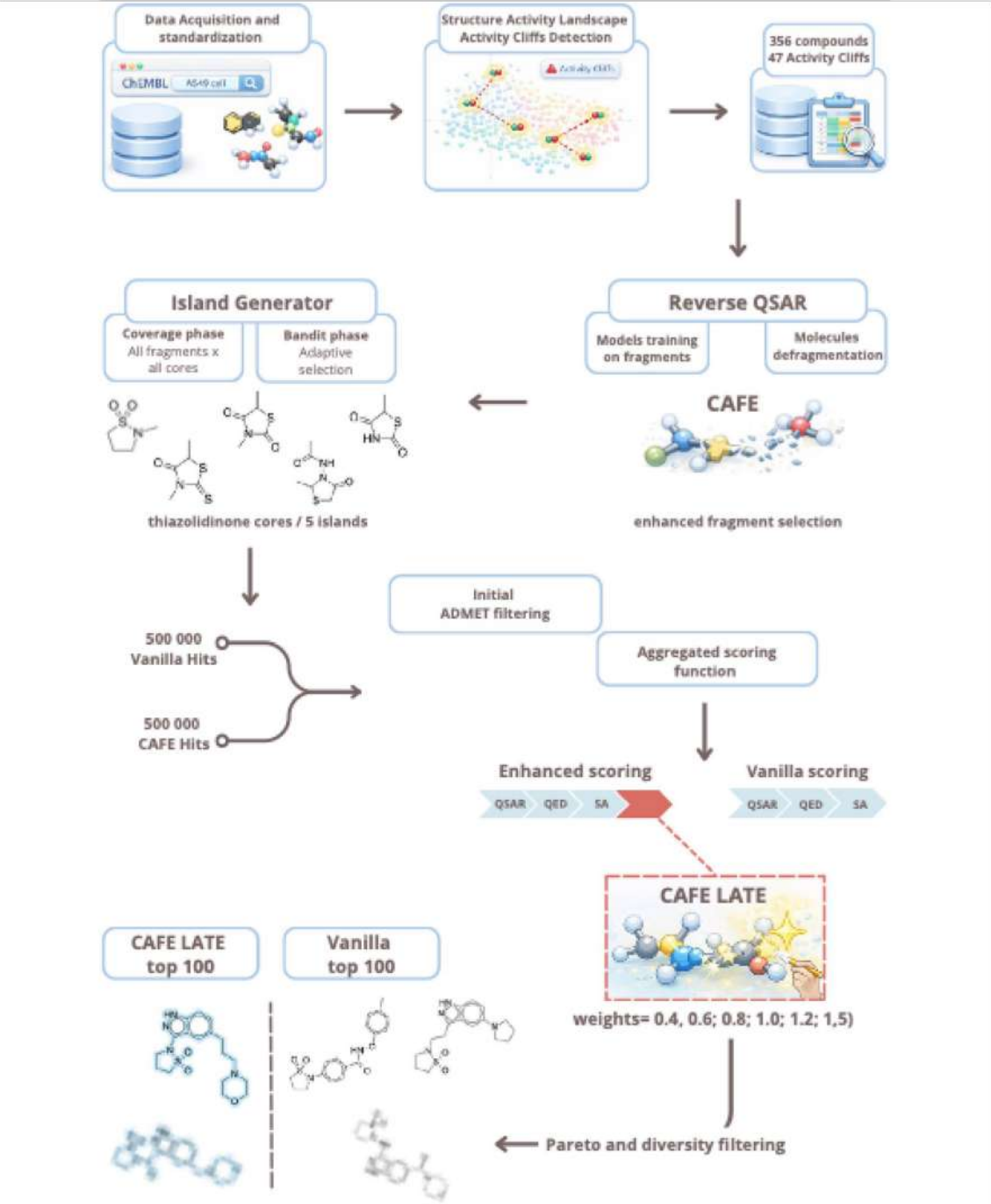


Fig. 2 Overview of the AC-aware de novo design workflow, with simplified depiction of CAFE and CAFE LATE intervention

and π-stacking. Complex structures (receptor-ligand) of the top hits were visualized using py3Dmol [64] within Jupyter Notebooks and 2D interaction diagrams were created using python scripts. Visualizations included highlighting binding sites, ligand conformations, and receptor-ligand interactions.

Ablation analysis

To quantify the contribution of CAFE/CAFE LATE beyond baseline QSAR scoring, we dock, per receptor, the top 100 compounds from vanilla (baseline) arm and each CAFE top 100 (for every λ_{CAFE}) under the same protocol, filter invalid poses (non-negative scores) and native ligands, and then form one-to-one matched pairs within each target by rank with a molecular-weight tolerance of ≤25 Da (unmatched ranks are discarded to preserve comparability). For every retained pair we compute:

$$\Delta E = E_{CAFE} - E_{VAN} \tag{20}$$

on minimal binding affinities, so that $\Delta E < 0$ denotes a CAFE LATE candidate win (lower binding affinity), and we summarize per-target and pooled effects using the paired Wilcoxon signed-rank test on ΔE with two-sided p-values and the matched-pairs rank-biserial correlation as effect size, alongside outcome-focused statistics that reflect medicinal-chemistry priorities: namely the mean ΔE, the 5th-percentile improvement (Q5) capturing uplift among the very best candidates, and the 25th-percentile (Q25) for the top quartile. Finally, across CAFE LATE weights we identify the "best" configuration by the most negative mean Q5 averaged over receptors (ties broken by mean ΔE), use this single configuration for head-to-head top-k visualizations against vanilla (ranks 1–6 per receptor), and report the unified CAFE union only for discovery summaries

Molecular dynamics

Based on the molecular docking results, both native ligand complexes and ligands exhibiting superior predicted binding affinities compared to their native counterparts (*better-than-native* hits) were selected for further dynamic evaluation. In total, molecular dynamics (MD) simulations were performed for approximately 245 protein–ligand complexes, enabling a systematic assessment of binding stability and interaction persistence beyond static docking poses. As a result, four protein structures were used in the MD campaign and downstream analysis: 1M17, 4I23, 5L2S, 5OQ4.

All MD simulations were carried out using OpenMM 8.2.0 [65]. Prior to simulation, receptor structures were carefully repaired to ensure structural completeness and consistency. Missing residues and atoms were reconstructed, non-standard residues were standardized, and protonation states were preserved, using PDBFixer. Ligand topologies and parameters were generated separately to ensure compatibility with the selected force fields.

Protein atoms were described using the Amber ff14SB force field [66], while ligands were parameterized with GAFF-2.11 [67]. Each protein–ligand complex was solvated in an explicit TIP3P water box with a minimum padding of 10.0 Å from the solute to the box edge. Physiological conditions were mimicked by adding Na^+ and Cl^- ions to achieve a final salt concentration of 150 mM.

A multistage equilibration protocol was applied to all systems prior to production simulations. This protocol consisted of energy minimization to remove steric clashes, gradual heating from 50 to 300 K, progressive release of backbone restraints, and equilibration under constant volume (NVT), followed by constant pressure (NPT) conditions.

Temperature and pressure were controlled using Langevin dynamics and a Monte Carlo barostat, respectively, while long-range electrostatics were treated with the Particle Mesh Ewald (PME) method. Following equilibration, each complex was subjected to a 100 ns production MD simulation at 300 K. A timestep of 1 fs was employed, and trajectory frames were saved every 100 fs for subsequent analysis. This extensive MD campaign enabled comparative evaluation of native versus better-than-native ligands in terms of structural stability, interaction longevity, and dynamic behavior within the binding site.

To quantify stability and pose retention, we computed Ligand RMSD (only heavy atoms), after fitting to the protein backbone; reported as time series, Pocket RMSD (residues within a defined radius around the ligand within 6 Å) to capture local receptor stability. Moreover, RMSF per residue, with emphasis on binding-site residues (e.g., residues within 6 Å of the ligand in the starting structure) to measure binding-site flexibility and Ligand–pocket COM distance and a bound-state fraction, defined by a geometric criterion, to detect partial unbinding or large rearrangements.

Results and discussion

Chemical space and structure activity landscape

After standardization, de-duplication, and scope filtering, the operational thiazolidinone dataset consisted of 356 compounds (Additional file 1), with a workable potency class balance: 117 actives (32.9%) and 239 inactives (67.1%). We removed a small set of "odd compounds" bearing cholesterol derived motifs (Additional file 2), because they lie outside thiazolidinone area of the chemical space and would inject nontransferable signal

for small-molecule optimization. Using ECFP4/Tanimoto similarity τ and a log-scale potency difference (ΔpIC_{50}), we identified 47 AC pairs (Additional file 3), covering 54 unique molecules, meaning 15.2% of the dataset participate in at least one AC pair. Within all high-similarity neighborhoods we observe: 349 pairs with $\tau \geq 0.80$ (of which 13.5% are AC), 119 pairs with $\tau \geq 0.85$ (10.1% AC), and 18 pairs with $\tau \geq 0.90$ (5.6% AC), indicating that while most close neighbors behave smoothly, a non-trivial fraction encodes steep SAR.

To visualize thiazolidinones chemical space, we conducted a t-SNE projection of ECFP4 space (Fig. 3A), which reveals molecules gathered in dense scaffold consistent clusters with broadly coherent potency, concentrating easy to learn SAR signal. Critically, AC pairs are mostly located on the edges of those smooth regions, where minor structural edits yield disproportionate activity changes. This is precisely where disrupted signal miscalibrates model, leading to overgeneralization and bad predictions. Those observations are only unpowered by MDS projection showing structurally similar compounds aggregated together while AC are randomly scattered among whole chemical space leading to SAR discontinuation (see Figure S1).

To quantify those discontinuities, hidden in clusters, we computed *SALI* which captures how different a pair of seemingly similar compounds is and set it against structural similarity to produce a SAS map. As a result, we detected twelve AC pairs with $SALI \geq 10$, forming compact "hot zones" of very steep SAR. Of these, 7 pairs also meet our AC rule ($\Delta pIC_{50} \geq 1.0$ and $\tau \geq 0.8$) (Fig. 3B). The remaining 5 are *SALI* only steep pairs where extreme similarity magnifies moderate ΔpIC_{50}, still mechanistically informative, and easy to miss with a binary AC definition. Together, t-SNE, MDS and the SAS map provide complementary views of the landscape: rule-based AC emphasize large ΔpIC_{50} among close neighbors, while *SALI* also surfaces micro-edits that create sharp local gradient. Practically, the dataset offers favorable mix for model guided design: cluster structured regions support generalization, and exposed AC carry valuable insights about crucial structural edits driving activity. In the next section we convert these high-gradient signals into supervised fragment knowledge for r-QSAR and generation, so that the library prioritizes thiazolidinone variants carrying the key modifications linked to activity gains.

Fragments library

Across trained ML engines, using the area under the precision–recall curve for the active class ($AUPRC_{active}$) as the primary criterion, logistic regression delivered the best mean performance of 0.889 (95% CI 0.856–0.917), and was therefore selected as the r-QSAR architecture (Fig. 3A, detailed metrics in Table S.1). The model misclassified 8 compounds (4 false negatives and 4 false positives), as shown on confusion matrix (Fig. 4B) these errors concentrate near high-gradient regions where small structural edits produce large potency flips, outside those zones, interpolation is stable and the classifier behaves as expected. This pattern justifies our choice of $AUPRC_{active}$ (robust to prevalence and aligned with the practical objective of recovering actives) and motivates a fragment level audit of failures. To identify and extract fragments responsible for those failures, we implemented a new logic CAFE, which corrects model choices by injecting important SAR signal hidden in AC pairs. Consistent with that, vanilla r-QSAR (without

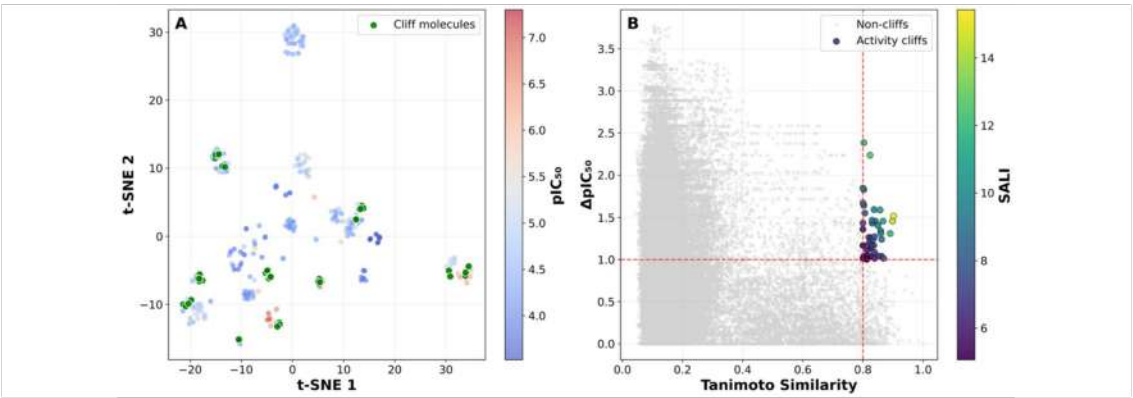


Fig. 3 **A** t-SNE projection of the thiazolidinone chemical space, computed on ECFP/Tanimoto, colored by pIC_{50} values. Green scatters are AC molecules. **B** Distribution of *SALI* values across molecule pairs in thiazolidinone dataset

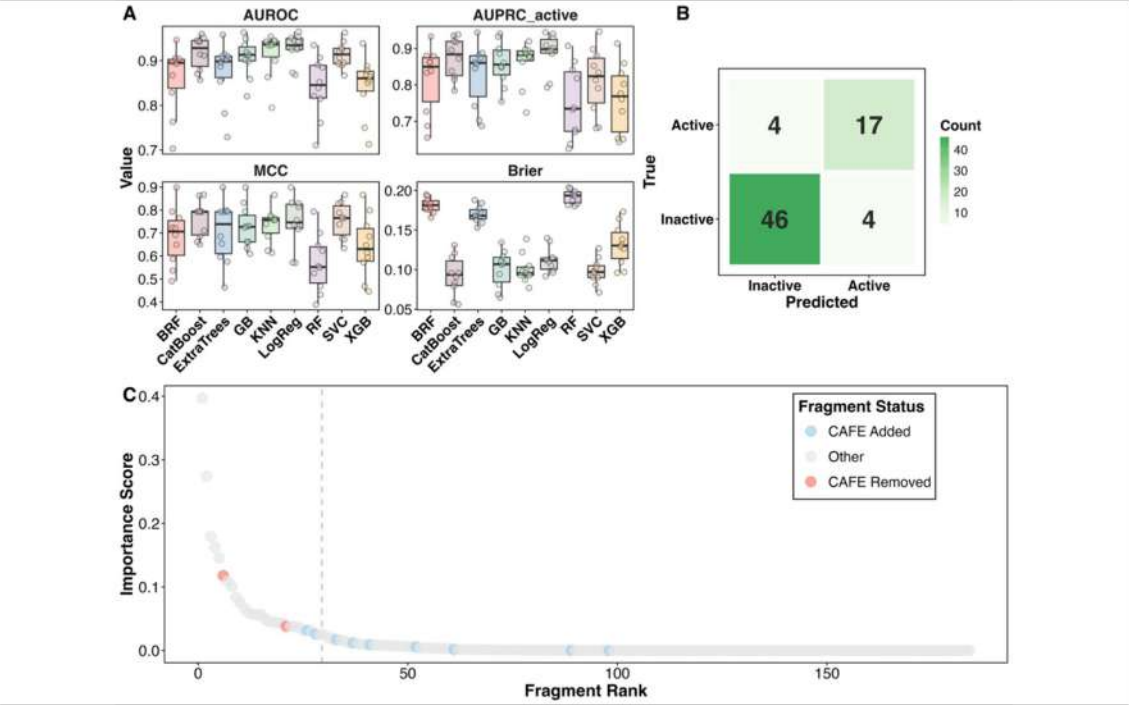


Fig. 4 The scatter plot with error bars of r-QSAR performance trained on nine ML engines **A**. The confusion matrix of the best r-QSAR model- logistic regression **B**. The fragments ranking by the models' importance score (**C**)

CAFE interference) identified 25 fragments with importance score cutoff as the most influential on compound's activity. However, as depicted on Fig. 4C, CAFE analysis revealed that r-QSAR overlooked nine important activity driving fragments, and critically included two fragments that are likely to occur on the inactive side of AC. Those falsely recognized motifs explain models' failures and metrics dropouts. If used as building blocks, they would bias generation toward suboptimal proposals. The intervention is surgical and keeps the library consistent with the measured landscape: smooth SAR regions still support generalization, while high-gradient edits are now explicitly represented as constructive design rules rather than noise. The final fragments library modified by CAFE (Table S.2), gives promise for generation of more reliable proposals.

Hit discovery

De novo generated thiazolidinone candidates must be preliminary assessed for predicted activity against the A549 lung cancer cell line, to prioritize productive design and by filtering out suboptimal building blocks combinations. For this task, we follow a MoleculeACE guideline and select the best prediction model based on AC-aware metric ($RMSE_{cliff}$), instead of global ones that could show overoptimistic results when tested on non-AC molecules. We used our version of that metric tuned for the classification task specificity, which identified BRF trained on MACCS as the most AC-aware model with the most robust results of $RMSE_{cliff}$ = 0.421, 95% CI 0.324–0.487 (see Fig. 5A and Table S3 for detailed performance metrics).

This model choice is further substantiated by the broader error and calibration patterns observed across QSAR settings. Global RMSE is frequently lower than $RMSE_{cliff}$, indicating that performance estimated over the random test set is often more optimistic than performance measured in steep SAR regions. In Fig. 5B, many model configurations lie above the dashed identity line, confirming that prediction error on AC compounds is systematically higher than the corresponding global error. This pattern highlights the extent to which AC distort the apparent predictive performance of QSAR models and demonstrates that global error metrics alone can underestimate the true difficulty of the task. Importantly, the pink points representing mean global RMSE calculated on the AC-free test sets used in the r-QSAR setting are, in most cases, shifted toward even lower values, further

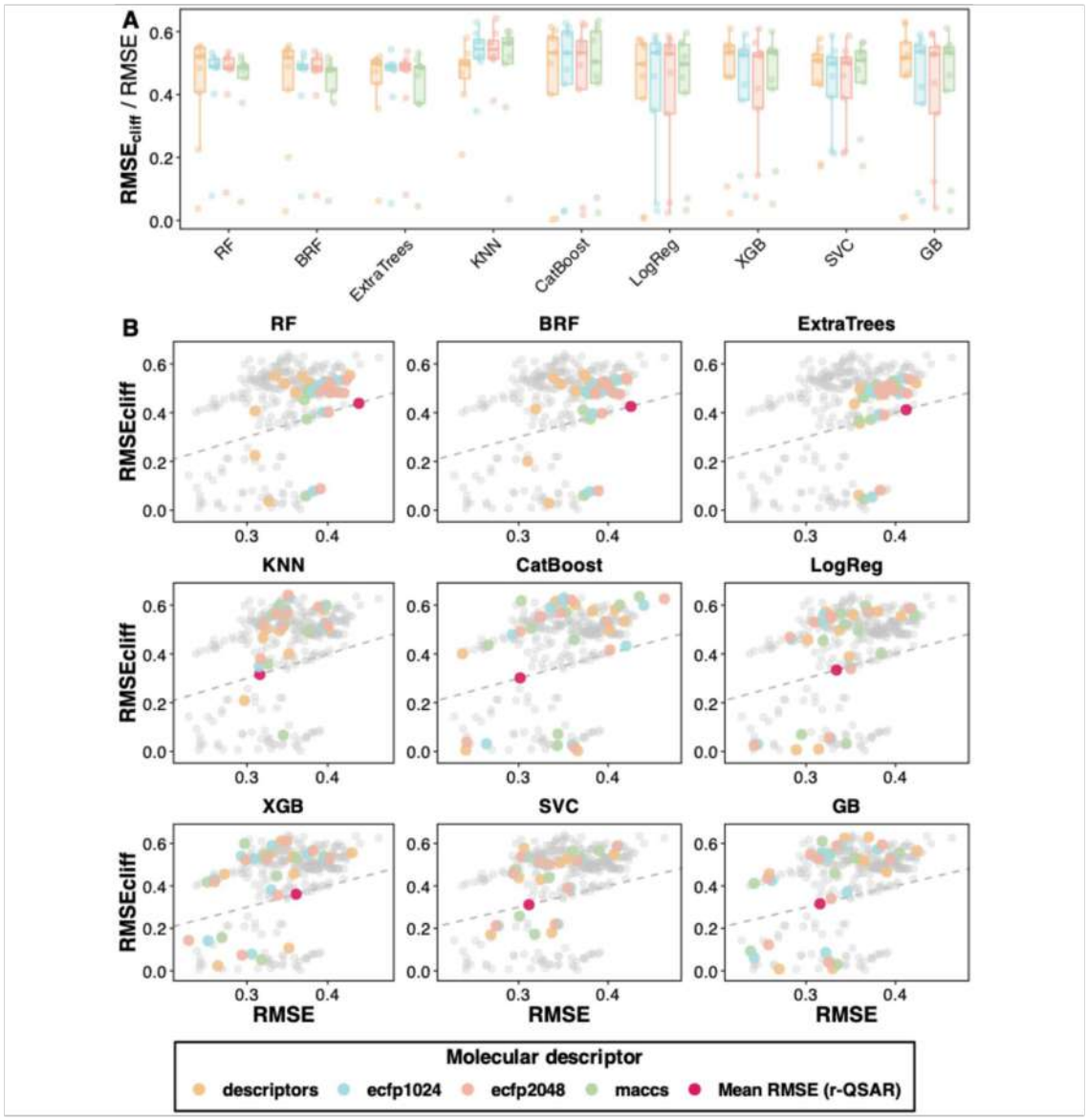


Fig. 5 **A** The $RMSE_{cliff}$ values of QSAR models and **B** comparison of $RMSE_{cliff}$ values against global RMSE computed on several molecular representations from ten independent training splits

illustrating how removal of AC from the evaluation set creates an artificially easier and more favorable assessment scenario. The same trend is reflected in the calibration analysis (Fig. 6). For most QSAR settings and across the majority of split realizations, AC restricted expected calibration error (ECE_{cliff}) exceeds the corresponding global ECE, again with many points located above the dashed line. Thus, activity cliffs impair not only predictive accuracy, but also the reliability of probabilistic outputs. Taken together, these results show that AC define a region of chemical space in which both error magnitude and probability calibration deteriorate, making global metrics insufficiently informative for model selection. In this context, the use of $RMSE_{cliff}$ as the primary criterion

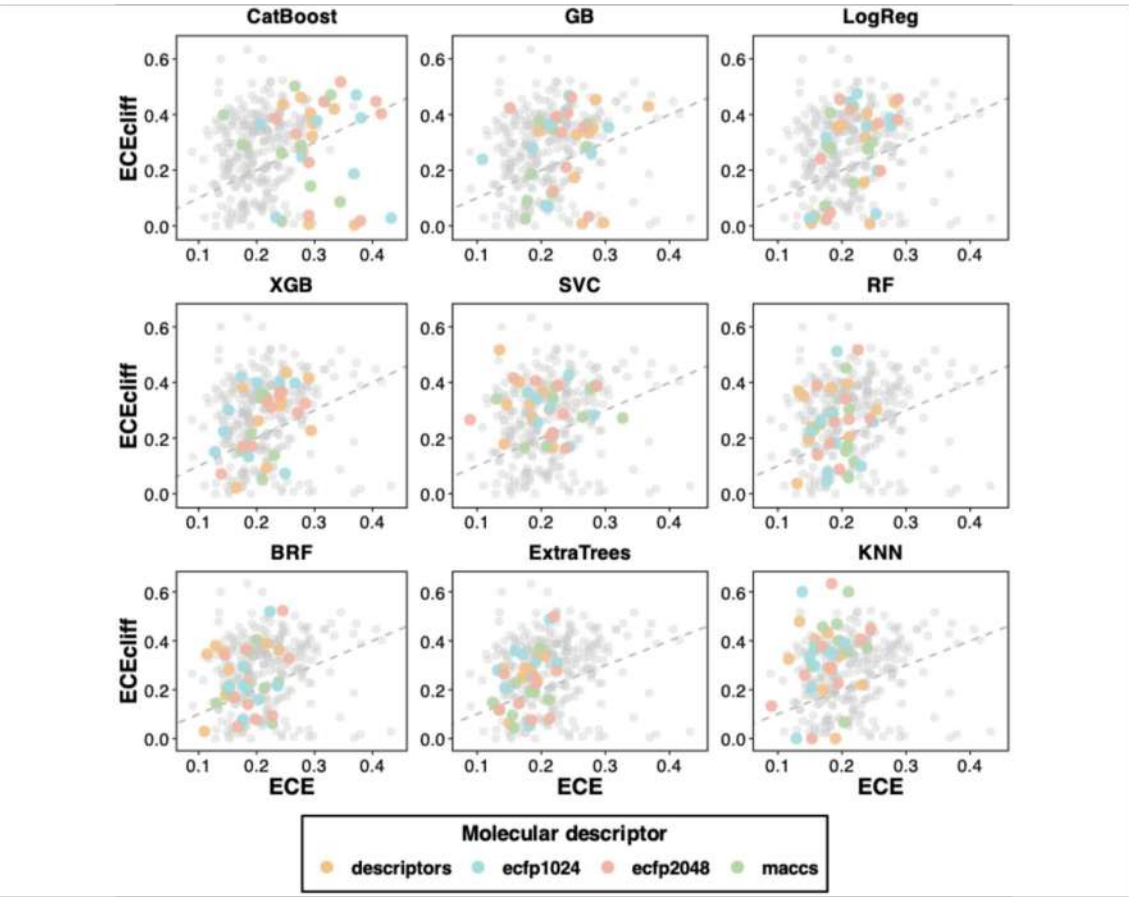


Fig. 6 Model calibration across evaluated QSAR configurations, assessed using global Expected Calibration Error (ECE) and AC restricted ECE (ECE_{cliff})

is not only methodologically justified, but necessary to identify a model that remains robust in the most challenging and decision-relevant part of the SAR landscape.

Having established a CAFE-corrected fragment library and selected the best-performing QSAR scorer, we next deployed both components in the de novo design stage. At the same time, we remained aware that even the most AC-robust available classifier, selected here through $RMSE_{cliff}$ cannot fully resolve the intrinsic predictive difficulty associated with steep SAR regions. Because learning activity directly from chemical structure becomes particularly challenging in the vicinity of AC, we introduced CAFE LATE as a supervised post hoc correction layer that adjusts the QSAR-based ranking of newly generated candidates carrying positively enriched CAFE fragments.

Generation was therefore conducted in two parallel arms. The first was the vanilla branch, corresponding to the baseline workflow without AC-derived post hoc correction. The second was the AC-aware branch, in which the corrected fragment library was combined with CAFE LATE supervision during candidate ranking. To determine the most effective contribution of the AC-aware bonus while preserving balanced exploration of chemical space, we evaluated a grid of post hoc boosting weights, $\lambda_{CAFE} = \{0.4; 0.6; 0.8; 1.0; 1.5\}$, applied independently during candidate scoring.

Because the final fragment library contained five distinct thiazolidinone ring variants, we adopted an island-style genetic generator to distribute exploration more evenly across the scaffold space. In both branches, 500,000 candidates were generated and subsequently standardized and evaluated using the aggregated score

$\varphi(m)$, composed of the QSAR-predicted probability, synthetic accessibility, and drug-likeness scores. In the AC-aware branch, this baseline score was further corrected by the CAFE LATE bonus under the tested λ_{CAFE} settings. As a result, we obtained one top100 set from the vanilla branch and one top100 set from each AC-aware weight setting, giving six top100 datasets in total. The overlap between the vanilla and CAFE LATE top-ranked sets decreased progressively with increasing λ_{CAFE}, reaching 5%, 1%, and finally 0% for the strongest boosting settings. In total, 282 unique candidates were identified and forwarded to molecular modeling studies (see Additional file 4).

Mechanism supported lead selection

To confirm results from preliminary assessment and choose the best candidates we implemented mechanistic prediction studies and compared the docking scores of all unique 282 candidates against those of the corresponding native ligands across well-characterized protein targets relevant to A549 cancer cell line. Those included:

- 1M17, Epidermal Growth Factor Receptor (EGFR) wild kinase domain co-crystallized with gefitinib a reversible, ATP-site, type-I binder that blocks receptor autophosphorylation. In A549, upstream receptor inhibition may be partially bypassed by constitutively active KRAS signaling [68].
- 4I23, EGFR wild kinase domain co-crystallized with dacomitinib, clinically used inhibitor [69].
- 8GUA, phosphoinositide 3-kinase alpha (PI3Kα, E542K mutant) with alpelisib (BYL-719). Direct structural reference for the class-I PI3Kα pocket engaged by an approved α-selective inhibitor [70].
- 4EKL, AKT1 co-crystallized with the clinical ATP-site inhibitor ipatasertib (GDC-0068). Supports docking and pharmacophore transfer for the central PI3k and AKT node [71].
- 5OQ4, class I PI3K/mTOR dual inhibitor (PQR309/paxalisib) in complex with PI3K. Relevant for KRAS mutant cancers [72].
- 4LMN, MEK1 bound to cobimetinib (allosteric, type-III site). Defines the MEK1 allosteric pocket and activation-loop geometry; essential for modeling the RAS → RAF → MEK → ERK cascade that sits immediately downstream of KRAS in A549 [73].
- 5HD4, ERK2 co-crystallized with SCH772984 (inhibitor induced pocket). Shows the distinctive inhibitor induced binding mode in ERK. Valuable when probing terminal MAPK pathway blockade after KRAS/MEK [74].
- 6SCM, SOS1 in complex with BI-3406 (protein–protein interaction blocker). Orthosteric disruption of the SOS1–KRAS nucleotide exchange interface. Structurally and mechanistically aligned to dampening KRAS activation irrespective of the specific codon-12 substitution in A549 [75].
- 5EHR, SHP2 (PTPN11) co-crystallized with the allosteric inhibitor SHP099. Stabilizes SHP2 in an autoinhibited conformation at the interface of its SH2 and PTP domains. Structurally underpins strategies that blunt receptor-tyrosine-kinase-RAS signaling input feeding the KRAS pathway[74].
- 4L7B, KEAP1 Kelch domain bound to a small-molecule antagonist. High-resolution Kelch-propeller pocket usable for modeling disruption of KEAP1-NRF2. Relevant because A549 carries KEAP1 loss-of-function and exhibits NRF2 pathway dysregulation [76].
- L2S, CDK6 co-crystallized with abemaciclib, 5L2I with palbociclib, and 5L2T with ribociclib. Three orthogonal, drug-resolved views of the CDK6 ATP site that support modeling cytostatic control over G1/S progression. Useful phenotypic leverage even when signaling is KRAS dominant [77].

All native ligands were successfully redocked into their cognate receptors, and the resulting poses matched the crystallographic references with RMSD ≤ 2.0 Å (Table S4), qualifying the receptor ensemble and docking setup. Preliminary analysis revealed influence of CAFE LATE on designing candidates with more favorable docking scores against many of studied molecular targets. The largest mean paired improvement was observed for the 5EHR receptor in the λ_{CAFE}= 0.6 dataset ($\Delta = -0.139$ kcal/mol across 95 matched pairs; Wilcoxon signed-rank test, $p < 0.0001$). Although differences of this magnitude should not be overinterpreted in absolute scoring terms, their directionality was consistent under the same docking protocol and therefore remains informative as a comparative signal. Importantly, the effect of AC-aware boosting became more pronounced at the top end of the ranking, which is the practically relevant region for hit prioritization. For instance, the mean improvements in binding affinities of top 5 matched pairs of $\lambda_{CAFE} = 0.8$ and vanilla dataset for 5EHR, 4I23, 4L7B were -2.5 kcal/mol, -0.980 kcal/mol and -0.7 kcal/mol, respectively. As depicted on Fig. 7 in head-to-head comparison of top 6 docking results from $\lambda_{CAFE} = 0.8$ dataset to vanilla, hits generated with AC boosting achieved better predicted binding affinities for most targets (points below dashed line symbolize CAFE LATE candidate with more favorable binding affinity colored by protein target).

Across the 282 unique combined CAFE LATE and vanilla hit molecules docked against all thirteen validated protein structures, 241 protein–ligand complexes

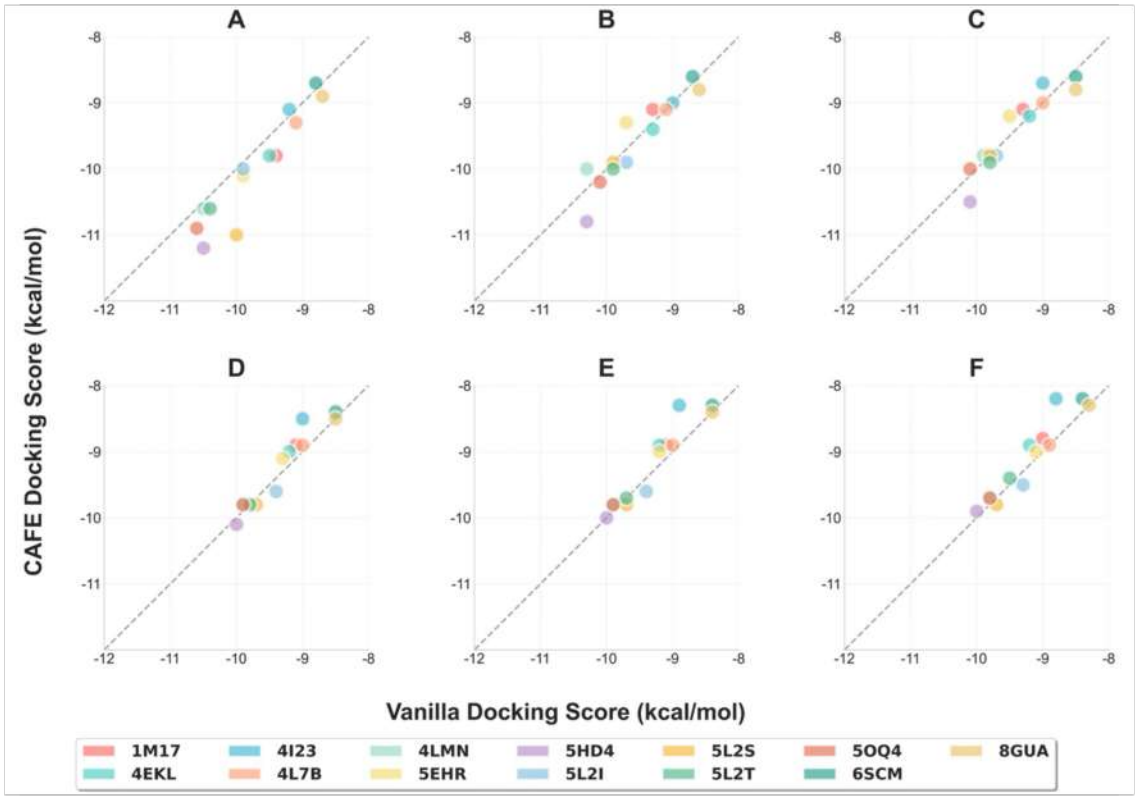


Fig. 7 Comparison of top 6 ranked molecules between CAFE and vanilla approaches across 13 molecular targets (**A**–**F** corresponding to 1–6 ranked hits, respectively)

achieved docking scores more favorable than the corresponding native ligand (see Additional file 5). Among those, 63.9% originated from CAFE LATE arm, indicating that vanilla generation model would miss more than a half ligand–protein complexes beating referent results, highly decreasing a chance for in vitro success. The best three docking scores in better-than-native dataset were formed by CAFE LATE candidates (−11.0 | −9.7; −10.9 | −8.2; −10.8 | −8.8 kcal/mol). Most importantly, five candidates showed better results than native ligands for more than one molecular target suggesting a potential multitargeted activity, and four from them were designed with AC-aware modeling strategy. For illustration, molecule **314** (from dataset with $\lambda_{CAFE} = 0.6$) formed a particularly strong complexes with 1M17: min|max scores of −9.8 | −7.4 kcal/mol, 5OQ4: −9.8 | −7.3 kcal/mol, and 5L2S: −11.0 | −9.7 kcal/mol, compared to the native ligands results of −7.1 | −6.4; −9.1 | −8.1; −10.2 | −8.9, respectively.

All complexes from better-than-native dataset were transferred to structure-based comparative refinement: short replicate molecular dynamics (MD) simulations to quantify pose stability and interaction persistence, followed by MM-GBSA postprocessing to estimate relative binding free energies from MD snapshots. To reduce receptor conformation bias, per-ligand results were aggregated across the validated receptor ensemble (robust statistics over the best conformations), and a consensus ranking integrated post-MD MM-PBSA, pose stability, and contact occupancy.

We evaluated the proposed CAFE LATE molecule design approach against a vanilla baseline using MD simulations across four protein–ligand complexes from better-than-native molecular docking analysis: 1M17 (n = 203 total systems), 5OQ4 (n = 23), 4I23 (n = 16), and 5L2S (n = 3). For each complex, the native ligand was re-docked and simulated under the same protocol and used as a per-target reference. For each designed ligand we computed: Pocket RMSD, Pocket RMSF, Ligand RMSD, Ligand RMSF, and Binding Energy. The whole distribution of results is presented on Figure S2.

In the cross-section by targets where we have both methods (1M17, 5OQ4, and 4I23) we run 91 vanilla ligands and 148 generated by the proposed algorithm. CAFE LATE shifts the distributions towards better structural stability more often than the baseline vanilla approach, which is best quantified as the percentage of candidates better than native one on each metric. Pooled across proteins, CAFE LATE molecules win more often on Pocket RMSD, Pocket RMSF, Ligand RMSD and Ligand RMSF, as shown in Table 1.

In MD MM-PBSA alone, the picture is more protein-dependent and does not show a unidirectional advantage of CAFE LATE across the entire set, suggesting that the main strength of CAFE LATE is revealed in stability metrics, while binding energy is more sensitive to the nature of the pocket and the underlying target difficulty. The largest and most representative set is 1M17 (203 systems in total: 73 vanilla, 129 CAFE LATE, and native), so we treat it as the main test. In 1M17, CAFE LATE consistently increases the percentage of candidates more stable than native in both ligand and pocket metrics, especially for Ligand RMSD and Ligand RMSF, as well as others. For MM-PBSA in 1M17, both methods achieve a comparable, moderate rate of outperforming native one. The key point, however, is that despite the moderate "average" advantage on MM-PBSA, CAFE LATE delivers strong individual hits that simultaneously satisfy all stability and energy criteria. In 1M17, CAFE LATE candidates such as **239** or **343** improved on all metrics relative to native, while in the vanilla set, similar cases are rarer.

To show whether the method produces individual hot candidates or rather a general shift in the distribution we present Fig. 8A, where the mean of the number of metrics improved simultaneously is higher for CAFE LATE (2.20 vs 2.07), reflecting a subtle but systematic shift towards multimetric improvement.

In 5OQ4 (23 complexes: 8 vanilla, 14 CAFE LATE, and native one), the advantage of AC-aware candidates is more pronounced and covers both stability and energy. CAFE LATE molecules more often improve Pocket RMSF (42.9% vs. 12.5%), Ligand RMSD (57.1% vs. 37.5%), Ligand RMSF (78.6% vs. 50.0%), and Binding Energy (42.9% vs. 25.0%), which indicates that the CAFE LATE algorithm better targets stable poses while maintaining or enhancing energetic interactions. The stability vs energy tradeoff is particularly well illustrated by the Pareto analysis shown in Fig. 8B. In this representation. The quadrant corresponding to simultaneously improved stability and binding energy (lower RMSD and Binding Energy) provides an immediate visual summary of method performance. In this region, ligands generated by CAFE are markedly more abundant than those obtained with the vanilla approach. Notably, ligand **241** exemplifies this behavior by outperforming the native reference across all five evaluated metrics, including a substantially improved Binding Energy value of − 5.951. The whole calculation metrics are available in Additional file 6.

To identify the most promising candidates generated in this study, we ranked all non-native ligands across protein targets according to their multi-metric improvement relative to the native redocked reference. The five highest-ranked ligands' complexes all outperform the native reference simultaneously across all five evaluated metrics, representing the strongest overall candidates identified in this benchmark shown in Table 2. Notably, 4 out of the 5 top ligands were generated by the CAFE LATE method, despite the presence of strong vanilla candidates, indicating that our algorithm not only increases the overall hit rate but also enriches the good quality region of chemical space. As depicted on Fig. 9, the best overall results were observed for hit 239 with computed binding energy of −7.78 kcal/mol formed with 1M17 target (compared to native ligand result of −4.09 kcal/mol), what complemented stability of favorable protein ligand interactions previously detected in molecular docking simulations (see Figure S3).

To place these hits in the context of the modeled chemical space, we additionally identified their nearest analogs in the training set using the same ECFP4/Tanimoto definition applied throughout the study. The resulting similarities were moderate (0.338–0.537; Additional file 7), indicating that the top-ranked compounds remain chemically anchored in the thiazolidinone design space without being trivial replicas of the training molecules.

Table 1 Summary of molecular dynamics simulation metrics for studied protein—ligand complexes

Metric	CAFE LATE molecules			Vanilla molecules		
	Better	Worse	%	Better	Worse	%
Pocket RMSD	95	54	**63.76**	55	37	59.78
Pocket RMSF	41	108	**27.52**	21	71	22.83
Ligand RMSD	70	79	**46.98**	37	55	40.22
Ligand RMSF	80	69	**53.69**	43	49	46.74
Binding Energy	55	94	36.91	41	51	**44.57**

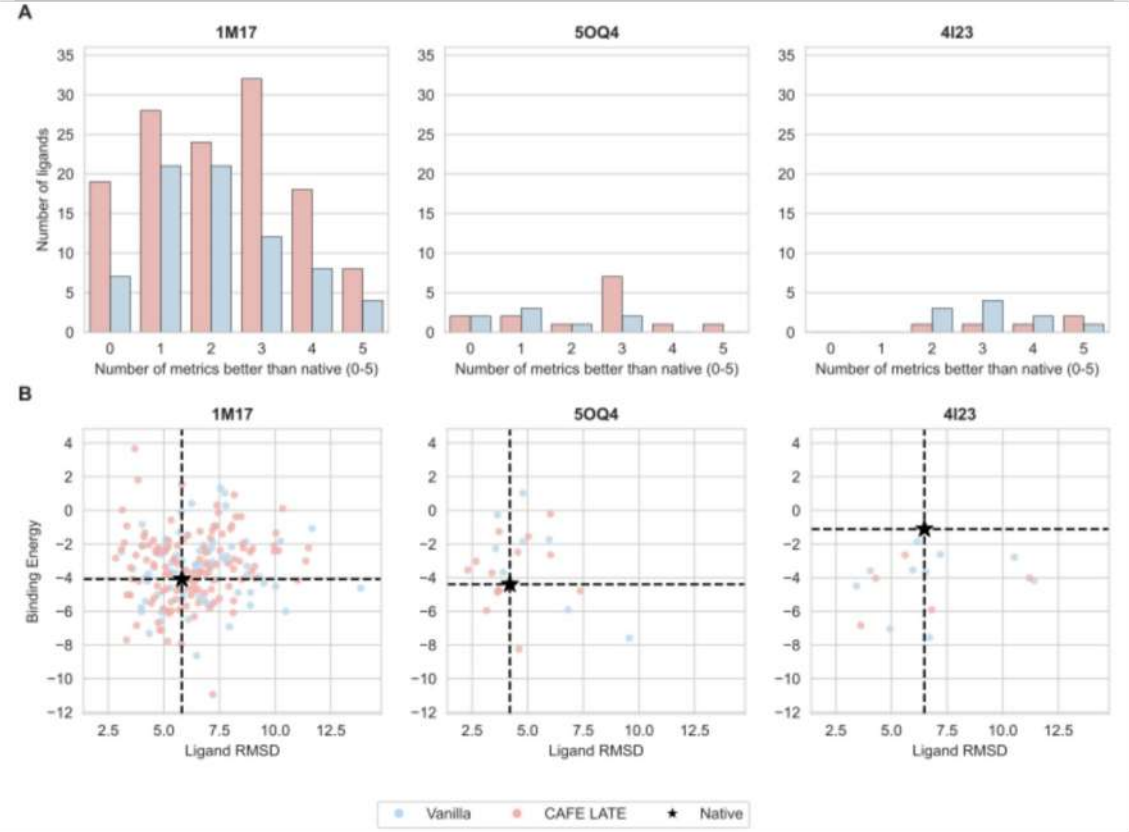


Fig. 8 **A** Multi-metric improvement profile relative to the native ligand. For each protein, we show the distribution of the number of molecular dynamics metrics for which a given ligand outperforms the native reference for each protein target. **B** Stability and energy tradeoff for ligand binding across protein complexes. Scatter plots show Ligand RMSD versus Binding Energy for ligands generated by the vanilla approach and by the CAFE method for each protein complex

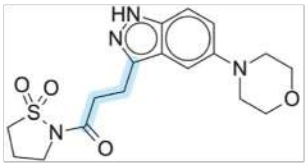
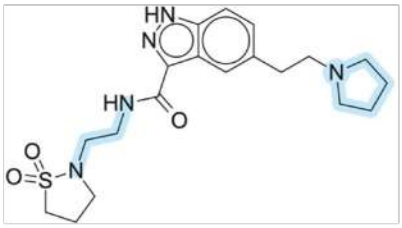
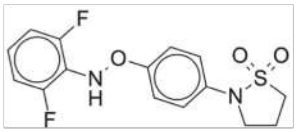
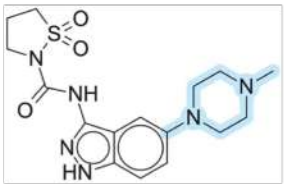
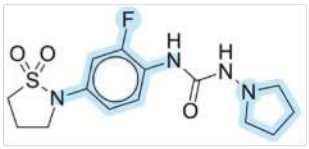
Conclusions

As development of in silico approaches continue to expand across cheminformatics and drug design, the present results have not reached the promised level yet, indicating the need for paradigm shift in building reliable predictive and generative models. In this work, we address this limitation by enriching classical, well-performing ML frameworks with an AC-aware fragment-based strategy, implemented as the CAFE and CAFE LATE algorithms. Unlike previously reported AC-aware methods, such as SiamACLoss and ACTriplet, which increase AC sensitivity during model training, or MoleculeACE, which formalizes cliff-centered benchmarking and evaluation, our framework aims the same problem from the perspective of drug design logic.

The proposed approach explicitly detects AC pairs, mines fragments hidden within steep SAR regions, removes fragments falsely identified as activity-driving, and recovers overlooked fragments that may promote activity, feeding this corrected fragment knowledge back into de novo molecular generation. Using this AC-aware design logic, we generated and evaluated thiazolidinone derivatives in a focused chemical space relevant to NSCLC. Among 282 unique molecules docked against multiple molecular targets, 241 protein–ligand complexes achieved predicted binding affinities superior to their native ligands, indicating favorable intermolecular interactions. These candidates were further subjected to molecular dynamics simulations to assess the stability of ligand binding within the active site.

Across docking and molecular dynamics benchmarks, CAFE LATE consistently produced a higher proportion of candidates with stable binding modes than the vanilla QSAR-guided baseline. This effect was particularly evident in structural stability metrics, including pocket RMSD and RMSF, where AC-aware candidates more

Table 2 Top-performing ligands across molecular dynamics benchmark identified against 1M17 protein target

Ligand ID	Ligand structure	Method	Binding energy [kcal/mol]	QED | SA
239		CAFE LATE	−7.78	0.85 | 0.77
343		CAFE LATE	−7.70	0.72 | 0.81
86		Vanilla baseline	−7.31	0.87 | 0.81
265		CAFE LATE	−7.12	0.80 | 0.78
271		CAFE LATE	−6.84	0.87 | 0.82

Fragments extracted by AC-aware method CAFE are highlighted in color blue
All molecules show 5/5 improved metrics compared to native ligand results

frequently outperformed native references. More importantly, the AC-aware framework generated the strongest individual hits, characterized by simultaneous improvements across multiple metrics: pocket RMSD, pocket RMSF, ligand RMSD, ligand RMSF, and binding energy. Compounds 239 and 343 emerged as the most promising examples, exhibiting both favorable binding energies (−7.78 and −7.70 kcal/mol, respectively, compared to native ligand -4.09 kcal/mol) and stable protein–ligand interactions. Importantly, these values are interpreted here in a comparative, protocol-dependent manner and are meant to indicate relative trends rather than absolute binding affinities of studied molecules.

These results suggest that incorporating AC-aware fragment logic into small-molecule generative frameworks can enhance exploration of unknown areas of the chemical space, reduce systematic design errors caused by misleading SAR generalization, and enrich the region of highest-quality candidates with increased likelihood of experimental success. At the same time, the present study is limited by its fully computational nature, reliance on docking and force-field based MD approximations, and restriction to a single chemical scaffold and a predefined set of protein targets. While all evaluations reported here were conducted following current community best practices for data curation and benchmarking, future work

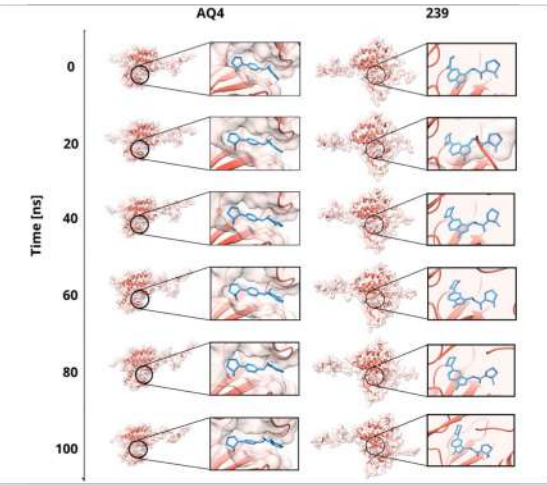


Fig. 9 Visualization of head-to-head comparison of protein–ligand complexes (1M17 with native vs the best generated compound 239) using MD simulation in 100 ns

will focus on experimental validation and on extending the proposed framework to additional scaffolds and biological targets to further test its generality and practical utility.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13321-026-01216-3.

Supplementary Material 1.
Supplementary Material 2.
Supplementary Material 3.
Supplementary Material 4.
Supplementary Material 5.
Supplementary Material 6.
Supplementary Material 7.
Supplementary Material 8.

Acknowledgements
DP and MW research was funded by Warsaw University of Technology within the Excellence Initiative: Research University (IDUB) programme and co-supported by Polish National Science Centre (2020/37/B/NZ2/03757). Computations were performed thanks to the Laboratory of Bioinformatics and Computational Genomics, Faculty of Mathematics and Information Science, Warsaw University of Technology using Artificial Intelligence HPC platform financed by Polish Ministry of Science and Higher Education (decision no. 7054/IA/SP/2020 of 2020-08-28).

Author contributions
TS– Conceptualization, formal analysis, data curation, software, validation, writing- original draft. MW– Formal analysis, data curation, validation, software, writing. DB– Supervision, resources, writing- review & editing VC– Writing-Review & editing DP – Supervision DS – Supervision, Project administration.

Funding
This research received no external funding.

Data availability
De novo thiazolidinones generative pipeline and CAFE/CAFE LATE architecture: (https://github.com/TomaszSzostek/AC_aware_modeling.git) Molecular docking: (https://github.com/TomaszSzostek/Molecular_docking.git) Molecular dynamics and final leads selection: (https://github.com/maciejwisniewski-drugdiscovery/MolecularDynamicsPipeline/tree/main) The datasets supporting the conclusions of this article are included within the article (and its additional files).

Declarations

Competing interests
The authors declare no competing interests.

Author details
[1]Chair and Department of Biochemistry, Medical University of Warsaw, 02-097 Warsaw, Poland. [2]Laboratory of Bioinformatics and Computational Genomics, Faculty of Mathematics and Information Science, Warsaw University of Technology, Koszykowa 75, 00-662 Warsaw, Poland. [3]Milano Chemometrics and QSAR Research Group, Department of Earth and Environmental Sciences, University of Milano Bicocca, Piazza della Scienza 1, 20126 Milano, Italy. [4]Laboratory of Functional and Structural Genomics, Centre of New Technologies, University of Warsaw, Banacha 2C, 02-097 Warsaw, Poland.

Received: 13 January 2026 Accepted: 1 May 2026
Published online: 11 May 2026

References

1. Bray F, Laversanne M, Sung H et al (2024) Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 74(3):229–263. https://doi.org/10.3322/caac.21834
2. World Health Organization. Global cancer burden growing, amidst mounting need for services. https://www.who.int/news/item/01-02-2024-global-cancer-burden-growing-amidst-mounting-need-for-services. Accessed 2 November 2025.
3. Zer A, Ahn MJ, Barlesi F et al (2025) Early and locally advanced non-small-cell lung cancer: ESMO clinical practice guideline for diagnosis, treatment and follow-up. Ann Oncol 36(11):1245–1262. https://doi.org/10.1016/j.annonc.2025.08.003
4. Anastasios D, Aaron M (2025) EGFR-mutated lung cancer—letting the butterfly out of the cocoon. N Engl J Med 393(17):1743–1745. https://doi.org/10.1056/NEJMe2509843
5. Xiang Y, Liu X, Wang Y et al (2024) Mechanisms of resistance to targeted therapy and immunotherapy in non-small cell lung cancer: promising strategies to overcoming challenges. Front Immunol 15:1366260. https://doi.org/10.3389/fimmu.2024.1366260
6. Yang Y, Fan S (2024) Small cell lung cancer transformations from non-small cell lung cancer: biological mechanism and clinical relevance. Chin Med J Pulm Crit Care Med 2(1):42–47. https://doi.org/10.1016/j.pccm.2023.10.005
7. Herbst RS, Heymach JV, Mok T et al (2023) Perioperative durvalumab for resectable non-small-cell lung cancer. N Engl J Med 389(18):1672–1684. https://doi.org/10.1056/NEJMoa2304875
8. Wu YL, Dziadziuszko R, Ahn JS et al (2024) Alectinib in resected ALK-positive non-small-cell lung cancer. N Engl J Med 390(14):1265–1276. https://doi.org/10.1056/NEJMoa2310532
9. Szostek T, Otto-Ślusarczyk D, Roszkowski P, Struga M, Szulczyk D (2024) Exploring the bioactive potential of (2-imino-4-oxo-1,3-thiazolidin-5-yl) acetic acid derivatives: a comprehensive review. Results Chem 11:101828. https://doi.org/10.1016/j.rechem.2024.101828
10. Sharma A, Sharma D, Saini N et al (2023) Recent advances in synthetic strategies and SAR of thiazolidin-4-one containing molecules in cancer therapeutics. Cancer Metastasis Rev 42:847–889. https://doi.org/10.1007/s10555-023-10106-1

11. Podolak M, Horishny V, Dudchak R, Gornowicz A, Czarnomysy R, Mural D, Holota S, Bielawski K, Lesyk R, Bielawska A (2025) Synthesis and anticancer activity evaluation of new 5-({5-Nitrofuran-2-yl}allylidene)-2-thioxo-4-thiazolidinones. Pharmaceuticals 18(11):1598. https://doi.org/10.3390/ph18111598
12. Shawky AM, Almalki FA, Abdalla AN et al (2023) Discovery and optimization of 2,3-diaryl-1,3-thiazolidin-4-one-based derivatives as potent and selective cytotoxic agents with anti-inflammatory activity. Eur J Med Chem 259:115712. https://doi.org/10.1016/j.ejmech.2023.115712
13. Abkar Aras M, Jafari N, Moshtaghi Zonouz A, Hamishehkar H (2025) An efficient green synthesis and in vitro evaluation of novel adamantane-containing 4-thiazolidinone derivatives as anticancer agents. J Mol Struct 1321:139708. https://doi.org/10.1016/j.molstruc.2024.139708
14. Roszczenko P, Holota S, Szewczyk OK et al (2022) 4-Thiazolidinone-Bearing hybrid molecules in anticancer drug design. Int J Mol Sci 23(21):13135. https://doi.org/10.3390/ijms232113135
15. Wu Z, Chen S, Wang Y et al (2024) Current perspectives and trend of computer-aided drug design: a review and bibliometric analysis. Int J Surg 110(6):3848–3878. https://doi.org/10.1097/JS9.0000000000001289
16. Serrano DR, Luciano FC, Anaya BJ, Ongoren B, Kara A, Molina G, Ramirez BI, Sánchez-Guirales SA, Simon JA, Tomietto G et al (2024) Artificial intelligence (AI) applications in drug discovery and drug delivery: revolutionizing personalized medicine. Pharmaceutics 16(10):1328. https://doi.org/10.3390/pharmaceutics16101328
17. Gangwal A, Ansari A, Ahmad I et al (2024) Generative artificial intelligence in drug discovery: basic framework, recent advances, challenges, and opportunities. Front Pharmacol 15:1331062. https://doi.org/10.3389/fphar.2024.1331062
18. Ren F, Aliper A, Chen J et al (2025) A small-molecule TNIK inhibitor targets fibrosis in preclinical and clinical models. Nat Biotechnol 43(1):63–75. https://doi.org/10.1038/s41587-024-02143-0
19. Hasselgren C, Oprea TI (2024) Artificial intelligence for drug discovery: are we there yet? Annu Rev Pharmacol Toxicol 64:527–550. https://doi.org/10.1146/annurev-pharmtox-040323-040828
20. Maggiora GM (2006) On outliers and activity cliffsWhy QSAR often disappoints. J Chem Inf Model 46(4):1535. https://doi.org/10.1021/ci060117s
21. Szostek T, Szulczyk D (2025) From obstacle to design advantage: activity cliff aware modeling for small-molecule drug discovery. Drug Discov Today. https://doi.org/10.1016/j.drudis.2025.104589
22. Hu H, Bajorath J (2020) Activity cliffs produced by single-atom modification of active compounds: systematic identification and rationalization based on X-ray structures. Eur J Med Chem 207:112846. https://doi.org/10.1016/j.ejmech.2020.112846
23. van Tilborg D, Antichav A, Grisoni F (2022) Exposing the limitations of molecular machine learning with activity cliffs. J Chem Inf Model 62(23):5938–5951. https://doi.org/10.1021/acs.jcim.2c01073
24. Bajorath J (2019) Duality of activity cliffs in drug discovery. Expert Opin Drug Discov 14(6):517–520. https://doi.org/10.1080/17460441.2019.1593371
25. Yu X, Wang Y, Chen L et al (2025) ACtriplet: an improved deep learning model for activity cliffs prediction by integrating triplet loss and pre-training. J Pharm Anal 15(8):101317. https://doi.org/10.1016/j.jpha.2025.101317
26. Shen W, Cui C, Su X, Zhang Z, Velez-Arce A, Wang J et al (2024) Activity cliff-informed contrastive learning for molecular property prediction. ChemRxiv. https://doi.org/10.26434/chemrxiv-2023-5cz7s-v2
27. Cheng Z, Xiang H, Ma P, et al. MaskMol: knowledge-guided molecular image pre-training framework for activity cliffs. arXiv. 2024.2409.12926.
28. Wu F. A semi-supervised molecular learning framework for activity cliff estimation. In: Proceedings of the 33rd International Joint Conference on Artificial Intelligence. 2024. 6080–6088.
29. Zhang Z, Zhao B, Xie A, Bian Y, Zhou S (2023) Activity cliff prediction: dataset and benchmark. arxiv. https://doi.org/10.48550/arXiv.2302.07541
30. Hu X, Liu G, Zhao Y, Zhang H (2025) Activity cliff-aware reinforcement learning for de novo drug design. J Cheminform 17(1):54. https://doi.org/10.1186/s13321-025-01006-3
31. Banerjee A, Roy K, Gramatica P (2025) A bibliometric analysis of the cheminformatics/QSAR literature (2000–2023) for predictive modeling in data science using the SCOPUS database. Mol Divers 29:3703–3715. https://doi.org/10.1007/s11030-024-11056-8
32. Gantzer P, Creton B, Nieto-Draghi C (2020) Inverse-QSPR for de novo design: a review. Mol Inform 39(4):1900087. https://doi.org/10.1002/minf.201900087
33. Deng J, Yang Z, Wang H, Ojima I, Samaras D, Wang F (2023) A systematic study of key elements underlying molecular property prediction. Nat Commun 14(1):6395. https://doi.org/10.1038/s41467-023-41948-6
34. Szostek T, Otto-Ślusarczyk D, Dobrowolski MA, Wiśniewski M, Roszkowski P, Szulczyk D (2025) Synthesis, cytotoxicity evaluation, and molecular dynamics of 4-thiazolidinone derivatives: a potential path to EGFR-targeted anticancer therapy. Curr Med Chem. https://doi.org/10.2174/0109298673400295250913044755
35. Bento AP, Gaulton A, Hersey A et al (2014) The ChEMBL bioactivity database: an update. Nucleic Acids Res. https://doi.org/10.1093/nar/gkt1031
36. Zdrazil B, Felix E, Hunter F et al (2024) The ChEMBL database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. Nucleic Acids Res 52(D1):D1180–D1192. https://doi.org/10.1093/nar/gkad1004
37. Nacht M, Dracheva T, Gao Y et al (2001) Molecular characteristics of non-small cell lung cancer. Proc Natl Acad Sci U S A 98(26):15203–15208. https://doi.org/10.1073/pnas.261414598
38. Kalliokoski T, Kramer C, Vulpetti A, Gedeck P (2013) Comparability of mixed IC_{50} data - a statistical analysis. PLoS ONE 8(4):e61007. https://doi.org/10.1371/journal.pone.0061007
39. Fourches D, Muratov E, Tropsha A (2010) Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. J Chem Inf Model 50(7):1189–1204. https://doi.org/10.1021/ci100176x
40. Brocidiacono M, Popov KI, Tropsha A. An improved metric and benchmark for assessing the performance of virtual screening models. Preprint. ArXiv. 2024.
41. Stumpfe D, Hu H, Bajorath J (2020) Advances in exploring activity cliffs. J Comput Aided Mol Des 34(9):929–942. https://doi.org/10.1007/s10822-020-00315-z
42. RDKit: Open-source cheminformatics. Zenodo. 10.5281/zenodo.591637. Accessed 11 Nov 2025.
43. Guha R, Van Drie JH (2008) Structure--activity landscape index: identifying and quantifying activity cliffs. J Chem Inf Model 48(3):646–658. https://doi.org/10.1021/ci7004093
44. Pedregosa F, Varoquaux G, Gramfort A et al (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825–2830
45. Kruskal JB (1964) Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika 29:1–27. https://doi.org/10.1007/BF02289565
46. Lemaître G, Nogueira F, Aridas CK (2017) Imbalanced-learn: a Python toolbox to tackle the curse of imbalanced datasets in machine learning. J Mach Learn Res 18(17):1–5
47. Durant JL, Leland BA, Henry DR, Nourse JG (2002) Reoptimization of MDL keys for use in drug discovery. J Chem Inf Comput Sci 42(6):1273–1280. https://doi.org/10.1021/ci010132r
48. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Guyon I, von Luxburg U, Bengio S, et al., eds. Advances in Neural Information Processing Systems. Curran Associates; 2017.
49. Satopaa V, Albrecht J, Irwin D, Raghavan B. Finding a "Kneedle" in a haystack: detecting knee points in system behavior. In: 2011 31st International Conference on Distributed Computing Systems Workshops. 2011. https://doi.org/10.1109/ICDCSW.2011.20
50. Kursa MB, Jankowski A, Rudnicki WR (2010) Boruta – A System for Feature Selection. Fund Inform 101(4):271–285. https://doi.org/10.3233/FI-2010-288
51. Wu Z, Ramsundar B, Feinberg EN et al (2018) MoleculeNet: a benchmark for molecular machine learning. Chem Sci 9(2):513–530. https://doi.org/10.1039/C7SC02664A
52. Rufibach K (2010) Use of Brier score to assess binary predictions. J Clin Epidemiol 63(8):938–939. https://doi.org/10.1016/j.jclinepi.2009.11.009
53. Gneiting T, Raftery AE (2007) Strictly proper scoring rules, prediction, and estimation. J Am Stat Assoc 102(477):359–378. https://doi.org/10.1198/016214506000001437
54. Whitley D, Rana S, Heckendorn RB (1999) The island model genetic algorithm: on separability, population size and convergence. J Comput Inf Technol 7(1):33–47

55. Bickerton G, Paolini G, Besnard J et al (2012) Quantifying the chemical beauty of drugs. Nat Chem 4:90–98. https://doi.org/10.1038/nchem.1243
56. Berman HM, Westbrook J, Feng Z et al (2000) The protein data bank. Nucleic Acids Res 28(1):235–242. https://doi.org/10.1093/nar/28.1.235
57. Eastman P, Friedrichs MS, Chodera JD et al (2013) OpenMM 4: a reusable, extensible, hardware independent library for high performance molecular simulation. J Chem Theory Comput 9(1):461–469. https://doi.org/10.1021/ct300857j
58. Morris GM, Huey R, Lindstrom W et al (2009) AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. J Comput Chem 30(16):2785–2791. https://doi.org/10.1002/jcc.21256
59. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR (2011) Open Babel: an open chemical toolbox. J Cheminform 3:33. https://doi.org/10.1186/1758-2946-3-33
60. Trott O, Olson AJ (2010) AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem 31(2):455–461. https://doi.org/10.1002/jcc.21334
61. Koes DR, Baumgartner MP, Camacho CJ (2013) Lessons learned in empirical scoring with smina from the CSAR 2011 benchmarking exercise. J Chem Inf Model 53(8):1893–1904. https://doi.org/10.1021/ci300604z
62. Pagadala NS, Syed K, Tuszynski J (2017) Software for molecular docking: a review. Biophys Rev 9(2):91–102. https://doi.org/10.1007/s12551-016-0247-1
63. Salentin S, Schreiber S, Haupt VJ, Adasme MF, Schroeder M (2015) PLIP: fully automated protein-ligand interaction profiler. Nucleic Acids Res 43(W1):W443–W447. https://doi.org/10.1093/nar/gkv315
64. Rego N, Koes D (2015) 3Dmol.js: molecular visualization with WebGL. Bioinformatics 31(8):1322–1324. https://doi.org/10.1093/bioinformatics/btu829
65. Eastman P, Galvelis R, Peláez RP et al (2024) OpenMM 8: molecular dynamics simulation with machine learning potentials. J Phys Chem B 128(1):109–116. https://doi.org/10.1021/acs.jpcb.3c06662
66. Tian C, Kasavajhala K, Belfon KAA et al (2019) ff19SB: amino-acid-specific protein backbone parameters trained against quantum mechanics energy surfaces in solution. J Chem Theory Comput 16(1):528–552. https://doi.org/10.1021/acs.jctc.9b00591
67. He X, Man VH, Yang W, Lee TS, Wang J (2020) A fast and high-quality charge model for the next generation general AMBER force field. J Chem Phys. https://doi.org/10.1063/5.0019056
68. Stamos J, Sliwkowski MX, Eigenbrot C (2002) Structure of the epidermal growth factor receptor kinase domain alone and in complex with a 4-anilinoquinazoline inhibitor. J Biol Chem 277(48):46265–46272. https://doi.org/10.1074/jbc.M207135200
69. Gajiwala KS, Feng J, Ferre R et al (2013) Insights into the aberrant activity of mutant EGFR kinase domain and drug recognition. Structure 21(2):209–219. https://doi.org/10.1016/j.str.2012.11.014
70. Liu X, Zhou Q, Hart JR et al (2022) Cryo-EM structures of cancer-specific helical and kinase domain mutations of PI3Kα. Proc Natl Acad Sci U S A 119(46):e2215621119. https://doi.org/10.1073/pnas.2215621119
71. Lin K, Lin J, Wu WI et al (2012) An ATP-site on-off switch that restricts phosphatase accessibility of Akt. Sci Signal 5:ra37. https://doi.org/10.1126/scisignal.2002618
72. Beaufils F, Cmiljanovic N, Cmiljanovic V, Bohnacker T, Melone A, Marone R, Jackson E, Zhang X, Sele A, Borsari C, Mestan J, Hebeisen P, Hillmann P, Giese B, Zvelebil M, Fabbro D, Williams RL, Rageot D, Wymann MP (2017) 5-(4,6-Dimorpholino-1,3,5-triazin-2-yl)-4-(trifluoromethyl)pyridin-2-amine (PQR309), a potent, brain-penetrant, orally bioavailable, pan-Class I PI3K/mTOR inhibitor as clinical candidate in oncology. J Med Chem 60(17):7524–7538. https://doi.org/10.1021/acs.jmedchem.7b00930
73. Hatzivassiliou G, Haling J, Chen H et al (2013) Mechanism of MEK inhibition determines efficacy in mutant KRAS- versus BRAF-driven cancers. Nature 501:232–236. https://doi.org/10.1038/nature12441
74. Jha S, Morris EJ, Hruza A, Mansueto MS, Schroeder GK, Arbanas J, McMasters D, Restaino CR, Dayananth P, Black S, Elsen NL, Mannarino A, Cooper A, Fawell S, Zawel L, Jayaraman L, Samatar AA (2016) Dissecting therapeutic resistance to ERK inhibition. Mol Cancer Ther 15(4):548–559. https://doi.org/10.1158/1535-7163.MCT-15-0172
75. Hofmann MH, Gmachl M, Ramharter J, Savarese F, Gerlach D, Marszalek JR, Sanderson MP, Kessler D, Trapani F, Arnhof H, Rumpel K, Botesteanu D-A, Ettmayer P, Gerstberger T, Kofink C, Wunberg T, Zoephel A, Fu S-C, Teh JL, Böttcher J, Pototschnig N, Schachinger F, Schipany K, Lieb S, Vellano CP, O'Connell JC, Mendes RL, Moll J, Petronczki M, Heffernan TP, Pearson M, McConnell DB, Kraut N (2021) BI-3406, a potent and selective SOS1–KRAS interaction inhibitor, is effective in KRAS-driven cancers through combined MEK inhibition. Cancer Discov 11(1):142–157. https://doi.org/10.1158/2159-8290.CD-20-0142
76. Jnoff E, Albrecht C, Barker JJ et al (2014) Binding mode and structure-activity relationships around direct inhibitors of the Nrf2-Keap1 complex. ChemMedChem 9:699–705. https://doi.org/10.1002/cmdc.201300525
77. Chen P, Lee NV, Hu W et al (2016) Spectrum and degree of CDK drug interactions predicts clinical performance. Mol Cancer Ther 15(10):2273–2281. https://doi.org/10.1158/1535-7163.MCT-16-0300

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

20. Authors' contribution statements

Słupno, 17.07.2026.
(miejscowość, data)

Daniel Szulczyk
.....
(imię i nazwisko)

OŚWIADCZENIE

Jako współautor pracy pt. „*Exploring the bioactive potential of (2-imino-4-oxo-1,3-thiazolidin-5-yl)acetic acid derivatives: A comprehensive review.*" oświadczam, iż mój własny wkład merytoryczny w przygotowanie, przeprowadzenie i opracowanie badań oraz przedstawienie pracy w formie publikacji stanowi: 5%.

Wkład Tomasza Szostka w powstawanie publikacji obejmował: zbieranie oraz analizę literatury naukowej, pisanie publikacji naukowej.

Jednocześnie wyrażam zgodę na wykorzystanie w/w pracy jako część rozprawy doktorskiej mgr. Tomasza Szostka

Daniel Szulczyk
.....
(podpis oświadczającego)

Warszawski Uniwersytet Medyczny
Katedra i Zakład Biochemii

Dr hab. n. med. i n. o zdr. Daniel Szulczyk

...Słupno, 17.07.2026.
(miejscowość, data)

Daniel Szulczyk
.....
(imię i nazwisko)

OŚWIADCZENIE

Jako współautor pracy pt. „*Synthesis, Cytotoxicity Evaluation, and Molecular Dynamics of 4-Thiazolidinone Derivatives: A Potential Path to EGFR-Targeted Anticancer Therapy*." oświadczam, iż mój własny wkład merytoryczny w przygotowanie, przeprowadzenie i opracowanie badań oraz przedstawienie pracy w formie publikacji stanowi: 5%.

Wkład Tomasza Szostka w powstawanie publikacji obejmował: Syntezę nowych związków chemicznych, przeprowadzenie badań biologicznych, zbieranie, analizowanie i opracowywanie wyników, przygotowanie publikacji naukowej.

Jednocześnie wyrażam zgodę na wykorzystanie w/w pracy jako część rozprawy doktorskiej mgr. Tomasza Szostka

Daniel Szulczyk
.....
(podpis oświadczającego)

Warszawski Uniwersytet Med...
Katedra i Zakład Biochem...
Dr hab. n. med. i n. o zdr. Daniel Sz...

...Słupno, 17.07.2026.
(miejscowość, data)

Daniel Szulczyk
.....
(imię i nazwisko)

OŚWIADCZENIE

Jako współautor pracy pt. „*From obstacle to design advantage: activity cliff aware modeling for small-molecule drug discovery*." oświadczam, iż mój własny wkład merytoryczny w przygotowanie, przeprowadzenie i opracowanie badań oraz przedstawienie pracy w formie publikacji stanowi: 5%.

Wkład Tomasza Szostka w powstawanie publikacji obejmował: Zbieranie i analizę literatury naukowej oraz przygotowanie publikacji naukowej.

Jednocześnie wyrażam zgodę na wykorzystanie w/w pracy jako część rozprawy doktorskiej mgr. Tomasza Szostka

Daniel Szulczyk
.....
(podpis oświadczającego)

Warszawski Uniwersytet Medyczny
Katedra i Zakład Biochemii
Dr hab. n. med. i n. o zdr. Daniel Szulczyk

.....Słupno, 17.07.2026..
(miejscowość, data)

Daniel Szulczyk
.....
(imię i nazwisko)

OŚWIADCZENIE

Jako współautor pracy pt. „*De novo design of anticancer 4-thiazolidinone derivatives: a generative framework shaped by activity cliffs*." oświadczam, iż mój własny wkład merytoryczny w przygotowanie, przeprowadzenie i opracowanie badań oraz przedstawienie pracy w formie publikacji stanowi: 3%.

Wkład Tomasza Szostka w powstawanie publikacji obejmował: Syntezę nowych związków chemicznych, badania cheminformatyczne, przeprowadzenie badań biologicznych, zbieranie, analizowanie i opracowywanie wyników, przygotowanie publikacji naukowej.

Jednocześnie wyrażam zgodę na wykorzystanie w/w pracy jako część rozprawy doktorskiej mgr. Tomasza Szostka

Daniel Szulczyk
.....
(podpis oświadczającego)

Katedra i Zakład Biochemii
Dr hab. n. med. i n. o zdr. Daniel Szulczyk

16.07.2026, Warszawa
(miejscowość, data)

Dagmara Otto-Ślusarczyk
(imię i nazwisko)

OŚWIADCZENIE

Jako współautor pracy pt. „*Exploring the bioactive potential of (2-imino-4-oxo-1,3-thiazolidin-5-yl)acetic acid derivatives: A comprehensive review*." oświadczam, iż mój własny wkład merytoryczny w przygotowanie, przeprowadzenie i opracowanie badań oraz przedstawienie pracy w formie publikacji stanowi:

Mój udział procentowy w publikacji określam jako 4%

Wkład Tomasza Szostka w powstawanie publikacji obejmował: zbieranie oraz analizę literatury naukowej, pisanie publikacji naukowej.

Jednocześnie wyrażam zgodę na wykorzystanie w/w pracy jako część rozprawy doktorskiej mgr. Tomasza Szostka

.....
(podpis oświadczającego)

16.07.2026, Warszawa
(miejscowość, data)

Dagmara Otto-Ślusarczyk
(imię i nazwisko)

OŚWIADCZENIE

Jako współautor pracy pt. „*Synthesis, Cytotoxicity Evaluation, and Molecular Dynamics of 4-Thiazolidinone Derivatives: A Potential Path to EGFR-Targeted Anticancer Therapy.*" oświadczam, iż mój własny wkład merytoryczny w przygotowanie, przeprowadzenie i opracowanie badań oraz przedstawienie pracy w formie publikacji stanowi:
Mój udział procentowy w przygotowaniu publikacji określam jako 5%
Wkład Tomasza Szostka w powstawanie publikacji obejmował: Syntezę nowych związków chemicznych, przeprowadzenie badań biologicznych, zbieranie, analizowanie i opracowywanie wyników, przygotowanie publikacji naukowej.

Jednocześnie wyrażam zgodę na wykorzystanie w/w pracy jako część rozprawy doktorskiej mgr. Tomasza Szostka

.....
(podpis oświadczającego)

Warszawa, 20.07.2026
(miejscowość, data)

Piotr Roszkowski
(imię i nazwisko)

OŚWIADCZENIE

Jako współautor pracy pt. „*Exploring the bioactive potential of (2-imino-4-oxo-1,3-thiazolidin-5-yl)acetic acid derivatives: A comprehensive review*." oświadczam, iż mój własny wkład merytoryczny w przygotowanie, przeprowadzenie i opracowanie badań oraz przedstawienie pracy w formie publikacji stanowi:
5%.

Wkład Tomasza Szostka w powstawanie publikacji obejmował: zbieranie oraz analizę literatury naukowej, pisanie publikacji naukowej.

Jednocześnie wyrażam zgodę na wykorzystanie w/w pracy jako część rozprawy doktorskiej mgr. Tomasza Szostka

Roszkowski P.

(podpis oświadczającego)

Warszawa, 20.06.2026
(miejscowość, data)

Piotr Roszkowski
(imię i nazwisko)

OŚWIADCZENIE

Jako współautor pracy pt. „*Synthesis, Cytotoxicity Evaluation, and Molecular Dynamics of 4-Thiazolidinone Derivatives: A Potential Path to EGFR-Targeted Anticancer Therapy*." oświadczam, iż mój własny wkład merytoryczny w przygotowanie, przeprowadzenie i opracowanie badań oraz przedstawienie pracy w formie publikacji stanowi:
4%.

Wkład Tomasza Szostka w powstawanie publikacji obejmował: Syntezę nowych związków chemicznych, przeprowadzenie badań biologicznych, zbieranie, analizowanie i opracowywanie wyników, przygotowanie publikacji naukowej.

Jednocześnie wyrażam zgodę na wykorzystanie w/w pracy jako część rozprawy doktorskiej mgr. Tomasza Szostka

Roszkowski P.

(podpis oświadczającego)

Warszawa 12.08. 2026

prof. Marta Struga

Katedra i Zakład Biochemii WUM

OŚWIADCZENIE

Jako współautor pracy pt. „*Exploring the bioactive potential of (2-imino-4-oxo-1,3-thiazolidin-5-yl)acetic acid derivatives: A comprehensive review*." oświadczam, iż mój własny wkład merytoryczny w przygotowanie, przeprowadzenie i opracowanie badań oraz przedstawienie pracy w formie publikacji stanowi:
Pomoc w budowaniu koncepcji pracy, przeglądu literatury oraz przygotowania manuskryptu, który oceniam na 5%.
Wkład Tomasza Szostka w powstawanie publikacji obejmował: zbieranie oraz analizę literatury naukowej, pisanie publikacji naukowej.

Jednocześnie wyrażam zgodę na wykorzystanie w/w pracy jako część rozprawy doktorskiej mgr. Tomasza Szostka

Marta Katarzyna Struga

Elektronicznie podpisany przez Marta Katarzyna Struga
Data: 2026.08.12 10:36:40 +02'00'

(podpis oświadczającego)

Warszawa, 21.07.2026

Michał Dobrowolski

OŚWIADCZENIE

Jako współautor pracy pt. „*Synthesis, Cytotoxicity Evaluation, and Molecular Dynamics of 4-Thiazolidinone Derivatives: A Potential Path to EGFR-Targeted Anticancer Therapy.*" oświadczam, iż mój własny wkład merytoryczny w przygotowanie, przeprowadzenie i opracowanie badań oraz przedstawienie pracy w formie publikacji stanowi:

Uzyskanie monokryształu związku numer 2, przeprowadzenie pomiaru i analizy jego struktury krystalicznej, opracowanie oraz depozyt danych w bazie CSD, przygotowanie publikacji naukowej. Co szacuję na 5%.

Wkład Tomasza Szostka w powstawanie publikacji obejmował: Syntezę nowych związków chemicznych, przeprowadzenie badań biologicznych, zbieranie, analizowanie i opracowywanie wyników, przygotowanie publikacji naukowej.

Jednocześnie wyrażam zgodę na wykorzystanie w/w pracy jako część rozprawy doktorskiej mgr. Tomasza Szostka

.....

(podpis oświadczającego)

Warszawa, 24.07.2026
(miejscowość, data)

Maciej Wiśniewski
(imię i nazwisko)

OŚWIADCZENIE

Jako współautor pracy pt. „*Synthesis, Cytotoxicity Evaluation, and Molecular Dynamics of 4-Thiazolidinone Derivatives: A Potential Path to EGFR-Targeted Anticancer Therapy*." oświadczam, iż mój własny wkład merytoryczny w przygotowanie, przeprowadzenie i opracowanie badań oraz przedstawienie pracy w formie publikacji stanowi:

5% i obejmował analizę symulacji dynamiki molekularnej

Wkład Tomasza Szostka w powstawanie publikacji obejmował: Syntezę nowych związków chemicznych, przeprowadzenie badań biologicznych, zbieranie, analizowanie i opracowywanie wyników, przygotowanie publikacji naukowej.

Jednocześnie wyrażam zgodę na wykorzystanie w/w pracy jako część rozprawy doktorskiej mgr. Tomasza Szostka

M. Wiśniewski
(podpis oświadczającego)

Maciej Wiśniewski
02.08.2026, Warszawa
(miejscowość, data)

Maciej Wiśniewski
(imię i nazwisko)

OŚWIADCZENIE

Jako współautor pracy pt. „De novo design of anticancer 4-thiazolidinone derivatives: a generative framework shaped by activity cliffs" oświadczam, iż mój własny wkład merytoryczny w przygotowanie, przeprowadzenie i opracowanie badań oraz przedstawienie pracy w formie publikacji stanowi:

5% i obejmowała analizę danych z symulacji dynamiki molekularnej

Wkład Tomasza Szostka w powstawanie publikacji obejmował: Konceptualizację, metodologię, eksperymenty badawcze, zbieranie i analizę wyników oraz przygotowanie publikacji naukowej.

Jednocześnie wyrażam zgodę na wykorzystanie w/w pracy jako część rozprawy doktorskiej mgr. Tomasza Szostka

Maciej Wiśniewski
(podpis oświadczającego)

Milan, 15/7/2026

Davide Ballabio

DECLARATION

As a co-author of the publication entitled, De *novo design of anticancer 4-thiazolidinone derivatives: a generative framework shaped by activity cliffs*"
I hereby declare that my own substantive contribution to the preparation, conduct and elaboration of the research, as well as to the presentation of the work in the form of a publication, was as follows: supervision, resources, writing- review & editing (total percentage contribution 10%).

The contribution of Tomasz Szostek to the preparation of the publication included: conceptualization, methodology, experiments conduct, data analysis, and preparation of the publication.
At the same time, I consent to the use of the above-mentioned publication as part of the doctoral dissertation of MPharm Tomasz Szostek.



Milano, 16/07/2026

Viviana Consonni

DECLARATION

As a co-author of the publication entitled *De novo design of anticancer 4-thiazolidinone derivatives: a generative framework shaped by activity cliffs*"
I hereby declare that my own substantive contribution to the preparation, conduct and elaboration of the research, as well as to the presentation of the work in the form of a publication, was as follows: writing-review and editing (total percentage contribution 5%).

The contribution of Tomasz Szostek to the preparation of the publication included: conceptualization, methodology, experiments conduct, data analysis, and preparation of the publication.
At the same time, I consent to the use of the above-mentioned publication as part of the doctoral dissertation of MPharm Tomasz Szostek.



Seattle, USA
(miejscowość, data)

Dariusz Plewczyński
(imię i nazwisko)

7/10/26

OŚWIADCZENIE

Jako współautor pracy pt. „De novo design of anticancer 4-thiazolidinone derivatives: a generative framework shaped by activity cliffs" oświadczam, iż mój własny wkład merytoryczny w przygotowanie, przeprowadzenie i opracowanie badań oraz przedstawienie pracy w formie publikacji stanowi:

5% i obejmował konceptualizację, analizę i przygotowanie publikacji.

Wkład Tomasza Szostka w powstawanie publikacji obejmował: Konceptualizację, metodologię, eksperymenty badawcze, zbieranie i analizę wyników oraz przygotowanie publikacji naukowej.

Jednocześnie wyrażam zgodę na wykorzystanie w/w pracy jako część rozprawy doktorskiej mgr. Tomasza Szostka

(podpis oświadczającego)