Bridging *in-vitro*, *in-silico* and corporate realms for pharmaceutical drug discovery

Kumulative Dissertation

zur Erlangung des Doktorgrades (Dr. rer. nat.)

der Mathematisch-Naturwissenschaftlichen Fakultät

der Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von

Yojana Gadiya

aus Jodhpur, India

Bonn, 2025

Angefertigt mit Genehmigung
der Mathematisch-Naturwissenschaftlichen Fakultät
der Rheinischen Friedrich-Wilhelms-Universität Bonn

Gutachter/Betreuer: Univ. Prof. Dr. rer. nat. Martin Hofmann-Apitius
Gutachterin: Univ. Prof. Dr. rer. nat. Diana Imhof

Tag der Promotion: 9. September 2025
Erscheinungsjahr: 2025

Abstract

The onset of the COVID-19 pandemic in 2020 highlighted the urgent need for efficient navigation of big data. In response, numerous workflows and algorithms for data processing, aggregation, and analysis were developed and widely shared within the scientific community to address these emerging health challenges. However, these workflows have limitations, which can compromise their effectiveness and reliability. First, many workflows suffer from insufficient documentation and poor version control, making them difficult to reproduce, validate, and adapt for broader use. Second, they are often tailored to specific communities or domains, limiting their robustness and generalizability across diverse applications. Third, these workflows prioritize decisions on scientific data while overlooking critical aspects such as market applicability. Fourth, integration challenges stem from an overreliance on single-modality data analysis, neglecting the incorporation of heterogeneous data types. This limitation undermines the ability to make comprehensive go/no-go decisions based on a more holistic understanding of the problem. Finally, most existing workflows remain predominantly *in silico*, with minimal or no *in vitro* validation. This lack of experimental translation raises concerns about their applicability in real-world scenarios.

In our work, we developed reproducible and well-documented pipelines to integrate and consolidate knowledge across various domains to enhance pandemic preparedness. These pipelines contextualize knowledge through graphs constructed from data extracted via manual curation of scientific publications and experimental results. Designed with flexibility in mind, the pipelines are agnostic, allowing their application across multiple domains, including diverse therapeutic indication areas. Next, to expand the scope of the underlying data beyond research, we mined patent literature, a valuable resource capturing the marketing and commercial landscape of drug discovery. Using a tool we developed called PEMT, we identified patterns in compound-

and target-agnostic strategies employed in the commercial sector. Moreover, to address integration challenges between knowledge graphs and omics-based technologies, we merged knowledge graphs with transcriptomics data. This enabled us to decipher the mechanisms of action of key drug and drug-like candidates. Furthermore, we demonstrated the successful translation of *in silico* work into biological experiments. Specifically, we built a machine-learning model to predict the antibacterial activity of compounds and validated it by testing an external library for antibacterial activity in both *in silico* and *in vitro*. Our work highlights the importance of combining diverse data modalities with biological networks to gain profound insights into the mechanisms driving drug discovery. These workflows and approaches lay a strong foundation for identifying and prioritizing optimal drug candidates, facilitating their transition from preclinical studies to clinical trials and, ultimately, to the market.

Have courage to think differently, courage to invent, to travel the unexplored path, courage to discover the impossible and to conquer the problems and succeed.

Dr. APJ Abdul Kalam

Acknowledgments

I truly feel fortunate to have the opportunity to do my PhD remotely under the guidance of Prof. Dr. Martin Hofmann-Apitius. Thank you for all the support and flexibility you've given me over the past few years to achieve my career goals. I would also like to express my gratitude to Prof. Dr. Diana Imhof for agreeing to be an additional reviewer of my thesis and the members of my defence committee.

To Dr. Philip Gribbon, Andrea Zaliani and Dr. Carsten Claussen, my academic advisors and scientific mentors, I really can't seem to find the right words to say thank you. You have constantly pushed me to be a better version of myself and provided me with opportunities that in the past years have shaped me into who I am today. My friends and scientific coaches Dr. Daniel Domingo-Fernández and Dr. Sarah Mubeen, have always been there guiding me through this journey and providing me with emotional support whenever needed. They have ensured that the spirit of completing the work isn't lost through all the work I do. I owe you guys a lot and thank you for all the unasked support, constant paper reviews, encouragement, and guidance that you have given me on this journey.

Special thanks to all of my fellow PhD students at SCAI and colleagues at ITMP. Though the duration of my PhD was largely remote and our time together was ever so short (limited to the weekly syncs with Martin), I am so grateful for all the wonderful discussions and engagement that still took place offline and online. In particular, to Yasmin Salimi, Negin Babaiha and Giulia Panzarella, it was such a pleasure working with you! To the ITMP colleagues, specifically Gesa Witt and Dr. Reagon Karki, I am so glad to have had your support all these years since I first started working. Thank you for having confidence in me and checking in with me whenever I looked stressed about not being able to focus on my Ph.D. A heartfelt thank you to Alina Enns for her prompt coordination and “behind-the-scene” work over email. Warm regards to the Data Science team at ITMP! Also, special thanks to colleagues at Enveda, Dr. Viswa Colluru, Dr. David Healey, Dr. Shaurya Chanana, Dr. Efi Kontou, and Dr. António Preto. It was my good fortune to have had the constant support of such talented and kind colleagues like you.

Finally, to my family and friends who are never so far, cheering me on and helping me to keep things in perspective, I am so thankful. To my mom, dad and brother, my appreciation and gratitude towards you are boundless. To my boyfriend, thank you for listening to all my whining and providing me with constant support over the years both personally and professionally. You instilled in me the importance of integrity and a commitment to always seek knowledge, lessons that have stayed with me throughout my academic journey.

Disclaimer

The language in the thesis has been improved using AI tools such as ChatGPT and Grammarly.

Publications

Thesis Publications

1. Domingo-Fernández, D., Baksi, S., Schultz, B., **Gadiya, Y.**, Karki, R., Raschka, T., Ebeling, C., Hofmann-Apitius, M. and Kodamullil, A.T., 2021. COVID-19 Knowledge Graph: a computable, multi-modal, cause-and-effect knowledge model of COVID-19 pathophysiology. *Bioinformatics*, 37(9), pp.1332-1334.

   `https://doi.org/10.1093/bioinformatics/btaa834`

2. Karki, R., **Gadiya, Y.**, Zaliani, A. and Gribbon, P., 2023. Mpox Knowledge Graph: a comprehensive representation embedding chemical entities and associated biology of Mpox. *Bioinformatics advances*, 3(1), p.vbad045.

   `https://doi.org/10.1093/bioadv/vbad045`

3. **Gadiya, Y.**, Shetty, S., Hofmann-Apitius, M., Gribbon, P. and Zaliani, A., 2024. Exploring SureChEMBL from a drug discovery perspective. *Scientific data*, 11(1), p.507.

   `https://doi.org/10.1038/s41597-024-03371-4`

4. **Gadiya, Y.**, Zaliani, A., Gribbon, P. and Hofmann-Apitius, M., 2023. PEMT: a patent enrichment tool for drug discovery. *Bioinformatics*, 39(1), p.btac716.

   `https://doi.org/10.1093/bioinformatics/btac716`

5. **Gadiya, Y.**, Gribbon, P., Hofmann-Apitius, M. and Zaliani, A., 2023. Pharmaceutical patent landscaping: A novel approach to understand

patents from the drug discovery perspective. *Artificial Intelligence in the Life Sciences*, 3, p.100069.

`https://doi.org/10.1016/j.ailsci.2023.100069`

6. Domingo-Fernández, D., **Gadiya, Y.**, Patel, A., Mubeen, S., Rivas-Barragan, D., Diana, C.W., Misra, B.B., Healey, D., Rokicki, J. and Colluru, V., 2022. Causal reasoning over knowledge graphs leveraging drug-perturbed and disease-specific transcriptomic signatures for drug discovery. *PLoS computational biology*, 18(2), p.e1009909.

   `https://doi.org/10.1371/journal.pcbi.1009909`

7. **Gadiya, Y.**, Genilloud, O., Bilitewski, U., Brönstrup, M., von Berlin, L., Attwood, M., Gribbon, P. and Zaliani, A., 2024. Predicting antimicrobial class specificity of small molecules using machine learning. *Journal of Chemical Information and Modeling.* 65 (5), pp.2416-2431.

   `https://doi.org/10.1021/acs.jcim.4c02347`

Other Publications

† Equal authorship

8. Hopp, M. T., Domingo-Fernández D., **Gadiya, Y.**, Detzel, M. S., Graf, R., Schmalohr, B. F., Kodamullil, T. A., Imhof, D., and Hofmann-Apitius, M. 2021. Linking COVID-19 and heme-driven pathophysiologies: A combined computational–experimental approach. *Biomolecules*, 11(5), 644.

   `https://doi.org/10.3390/biom11050644`

9. Karki, R., Madan, S., **Gadiya, Y.**, Domingo-Fernández D., and Hofmann-Apitius, M. 2020. Data-driven modeling of knowledge assemblies in understanding comorbidity between type 2 diabetes mellitus and Alzheimer's Disease. *Journal of Alzheimer's Disease*, 78(1), 87-95.

   `https://doi.org/10.3233/JAD-200752`

10. Schultz, B., Zaliani, A., Ebeling, C., Reinshagen, J., Bojkova, D., Lage-Rupprecht, V., Karki, R., Lukassen, S., **Gadiya, Y.**, ... and Hofmann-Apitius, M. 2021. A method for the rational selection of drug

repurposing candidates from multimodal knowledge harmonization. *Scientific reports*, 11(1), 11049.

`https://doi.org/10.1038/s41598-021-90296-2`

11. Mubeen, S., Bharadhwaj, V. S., **Gadiya, Y.**, Hofmann-Apitius, M., Kodamullil, A. T., and Domingo-Fernández D. 2021. DecoPath: a web application for decoding pathway enrichment analysis. *NAR Genomics and Bioinformatics*, 3(3), lqab087.

    `https://doi.org/10.1093/nargab/lqab087`

12. Khatami S. G., Sargsyan A., Russo M. F., Domingo-Fernández D., Zaliani, A., Kaladharan, A., Sethumadhavan, P., Mubeen, S., **Gadiya, Y.**, Karki, R., ... and Kodamullil, T. A. 2024. Curating, Collecting, and Cataloguing Global COVID-19 Datasets for the Aim of Predicting Personalized Risk. *Data*. 2024; 9(2):25.

    `https://doi.org/10.3390/data9020025`

13. Alawathurage, T. Bharadhwaj, V. S. and **Gadiya, Y.** 2023. Classification of Images from Biomedical Literature. *In Computational Life Sciences: Data Engineering and Data Mining for Life Sciences* (pp. 569-595). Cham: Springer International Publishing.

    `https://doi.org/10.1007/978-3-031-08411-9_21`

14. Rocca-Serra, P., Gu, W., Ioannidis, V., Abbassi-Daloii, T., Capella-Gutierrez, S., Chandramouliswaran, I., Splendiani, A., Burdett, T., Giessmann, R., Henderson, R., Batista, D., Emam, I., **Gadiya, Y.**, ... and Sansone, S. A. 2023. The FAIR Cookbook-the essential resource for and by FAIR doers. *Scientific data*, 10(1), 292.

    `https://doi.org/10.1038/s41597-023-02166-3`

15. Alharbi, E.[†], **Gadiya, Y.**[†], Henderson, D., Zaliani, A., Delfin-Rossaro, A., Cambon-Thomsen, A., ... and Gribbon, P., 2022. Selection of data sets for FAIRification in drug discovery and development: Which, why, and how?. *Drug discovery today*, 27(8), pp.2080-2085.

    `https://doi.org/10.1016/j.drudis.2022.05.010`

16. Welter, D., Juty, N., Rocca-Serra, P., Xu, F., Henderson, D., Gu, W., Strubel, J., Giessmann, R.T., Emam, I., **Gadiya, Y.**, Abbassi-Daloii, T. ... and Burdett, T., 2023. FAIR in action-a flexible framework to guide FAIRification. *Scientific data*, 10(1), p.291.

    `https://doi.org/10.1038/s41597-023-02167-2`

17. Khorchani, T., **Gadiya, Y.**, Witt, G., Lanzillotta, D., Claussen, C. and Zaliani, A., 2022. SASC: A simple approach to synthetic cohorts for generating longitudinal observational patient cohorts from COVID-19 clinical data. *Patterns*, 3(4).

    `https://doi.org/10.1016/j.patter.2022.100453`

18. Karki, R., **Gadiya, Y.**, Gribbon, P. and Zaliani, A., 2023. Pharmacophore-based machine learning model to predict ligand selectivity for E3 ligase binders. *ACS omega*, 8(33), pp.30177-30185.

    `https://doi.org/10.1021/acsomega.3c02803`

19. Berg, H., Wirtz Martin, M. A., Altincekic, N., Alshamleh, I., Kaur Bains, J., Blechar, J., Ceylan,B., Jesus, V., Dhamotharan, K., Fuks, C., Gande, S., Hargittay, B., Hohmann, K., Hutchison, M., Korn, S., Krishnathas, R., Kutz, F., Linhard, V., Matzel, T., Meiser, N., Niesteruk, A., Pyper, D., Schulte, L., Trucks, S., Azzaoui, K., Blommers, M., **Gadiya, Y.** ... and Schwalbe, H. 2022. Comprehensive fragment screening of the SARS-CoV-2 proteome explores novel chemical space for drug development. *Angewandte Chemie*, 134(46), e202205858.

    `https://doi.org/10.1002/anie.202205858`

20. Karki, R., **Gadiya, Y.**, Shetty, S., Gribbon, P. and Zaliani, A., 2024. Pharmacophore-based ML model to filter candidate E3 ligands and predict E3 Ligase binding probabilities. *Informatics in Medicine Unlocked*, 44, p.101424.

    `https://doi.org/10.1016/j.imu.2023.101424`

21. **Gadiya, Y.**[†], Ammar, A., Willighagen, E., Martinat, D., Sima, A. C., Balci, H., and Abbassi-Daloii, T.[†]. 2023. Extending interoperability of experimental data using modular queries across biomedical resources.*BioHackrXiv.*

    `https://doi.org/10.37044/osf.io/mhsqp`

22. Arend, D., Del Conte, A., Feser, M., **Gadiya, Y.**, Gaignard, A., Castro, L.J., Mičetić, I., Moretti, S., Neumann, S., Rayya, N. and Tsueng, G., 2024. Bioschemas Resource Index for Chem and Plants. *BioHackrXiv.*

    `https://doi.org/10.37044/osf.io/yxunp`

23. Hussein, R., Balaur, I., Burmann, A., Ćwiek-Kupczyńska, H., **Gadiya, Y.**, Ghosh, ... and Gribbon, P. (2024). Getting ready for the European Health Data Space (EHDS): IDERHA's plan to align with the latest

EHDS requirements for the secondary use of health data. *Open Research Europe*, 4, 160.

`https://doi.org/10.12688/openreseurope.18179.1`

24. Greco, A., Karki, R., **Gadiya, Y.**, Deecke, C., Zaliani, A. and Gul, S., 2024. Pharmacological profiles of neglected tropical disease drugs. *Artificial Intelligence in the Life Sciences*, 6, p.100116.

    `https://doi.org/10.1016/j.ailsci.2024.100116`

25. **Gadiya, Y.**, Karki, R., Gribbon, P., and Zaliani, A. 2025. Artificial intelligence-driven patent analysis in drug discovery. Comprehensive Medicinal Chemistry 4th Edition - In Silico Tools. *Book in press*

26. Tanoli, Z., Fernández-Torras, A., Özcan, U. O., Kushnir, A., Nader, K. M., **Gadiya, Y.**, ... and Aittokallio, T. (2025). Computational drug repurposing: approaches, evaluation of in silico resources and case studies. *Nature Reviews Drug Discovery*, 1-22.

    `https://doi.org/10.1038/s41573-025-01164-x`

27. Domingo-Fernández, D.†, **Gadiya, Y.**†, Mubeen, S., Healey, D., Norman, B.H. and Colluru, V., 2023. Exploring the known chemical space of the plant kingdom: insights into taxonomic patterns, knowledge gaps, and bioactive regions. *Journal of Cheminformatics*, 15(1), p.107.

    `https://doi.org/10.1186/s13321-023-00778-w`

28. Rivas-Barragan, D.†, Domingo-Fernández, D.†, **Gadiya, Y.**† and Healey, D., 2022. Ensembles of knowledge graph embedding models improve predictions for drug discovery. *Briefings in Bioinformatics*, 23(6), p.bbac481.

    `https://doi.org/10.1093/bib/bbac481`

29. Domingo-Fernández, D.†, **Gadiya, Y.**†, Mubeen, S., Bollerman, T.J., Healy, M.D., Chanana, S., Sadovsky, R.G., Healey, D. and Colluru, V., 2023. Modern drug discovery using ethnobotany: a large-scale cross-cultural analysis of traditional medicine reveals common therapeutic uses. *Iscience*, 26(9).

    `https://doi.org/10.1016/j.isci.2023.107729`

30. Domingo-Fernández D.,**Gadiya, Y.**, Preto, A.J., Krettler, C.A., Mubeen, S., Allen, A., Healey, D. and Colluru, V., 2024. Natural products have

increased rates of clinical trial success throughout the drug development process. *Journal of Natural Products*, 87(7), pp.1844-1851.

`https://doi.org/10.1021/acs.jnatprod.4c00581`

31. DeLong, L.N., **Gadiya, Y.**, Galdi, P., Fleuriot, J.D. and Domingo-Fernández, D., 2024. Mars: A neurosymbolic approach for interpretable drug discovery. arXiv preprint *arXiv*:2410.05289.

    `https://doi.org/10.48550/arXiv.2410.05289`

32. Hart, C.E., **Gadiya, Y.**, Kind, T., Krettler, C.A., Gaetz, M., Misra, B.B., Healey, D., Allen, A., Colluru, V. and Domingo-Fernandez, D., 2025. Defining the limits of plant chemical space: challenges and estimations. *bioRxiv*, pp.2025-01.

    `https://doi.org/10.1101/2025.01.08.631938`

Contents

CHAPTER 1

Introduction

Drug discovery and development represent a complex and multifaceted process essential for the advancement of new medications to address a broad spectrum of diseases, including those with both met and unmet needs. This endeavour involves the amalgamation of diverse scientific disciplines, such as chemistry, biology, pharmacology and informatics, among others [1–3]. Historically, drug discovery has relied on a combination of empirical observations, serendipitous discoveries and systematic screening of chemical libraries [4, 5]. However, with advances in technology and our understanding of molecular biology and disease etiology, the drug discovery process has evolved to become more rational and data-driven [6–9]. This evolution has eventually empowered academic institutions to forge partnerships with the pharmaceutical industry, fostering efficient collaboration to drive future drug development endeavours [10, 11].

This chapter provides an overview of key concepts central to drug discovery, focusing on the generation, utilization, storage, and analysis of data within this domain. A significant emphasis is placed on the representation of this data in the form of graphs, which serve as powerful tools for organizing and visualizing complex interactions. Additionally, special attention is given to the importance of patent documents in this context, highlighting their utility as valuable sources of information for researchers. The subsequent chapters delve deeper into the exploration of biological systems by bridging the *in vitro* experimentation and *in silico* modeling realms, offering insights into innovative approaches aimed at advancing drug discovery methodologies.

1.1 Resources leveraged in drug discovery

1.1.1 Fundamentals of drug discovery process

The discovery of drugs traces back to early human civilization, where it was often based on chance occurrences or natural observations. Human communities, dispersed across diverse geographical regions and cultures, passed down knowledge about known herbs and medicines from one generation to the next [12–16]. The digitalization of scientific literature has made vast amounts of information, previously confined to books and manuscripts, easily accessible online. In contemporary times, the drug discovery process has become more systematic and regulated. Researchers are now required to conduct thorough basic research across various biological modules before progressing to the development of a drug candidate.

The modern drug discovery process unfolds through five distinct stages: discovery, pre-clinical research, clinical research, approval and market distribution (Figure 1). Spanning approximately 15 years and requiring an investment of over a billion dollars, this process is a comprehensive journey marked by scientific investigation and regulatory scrutiny [17, 18]. The initial stage, known as the **Research or Discovery stage**, is characterized by the formulation of multiple scientific hypotheses regarding the therapeutic effects of targeting specific proteins or biomarkers within a given disease area. To achieve this, researchers exhaustively explore the therapeutic landscape, gathering evidence to support each hypothesis from publicly available documents, including research papers and experimental data. Central to this stage is the identification of biomarker proteins or genes associated with the target disease, as well as the discovery of potential small molecules or biologics capable of modulating the activity of these biomarkers [4, 19, 20].

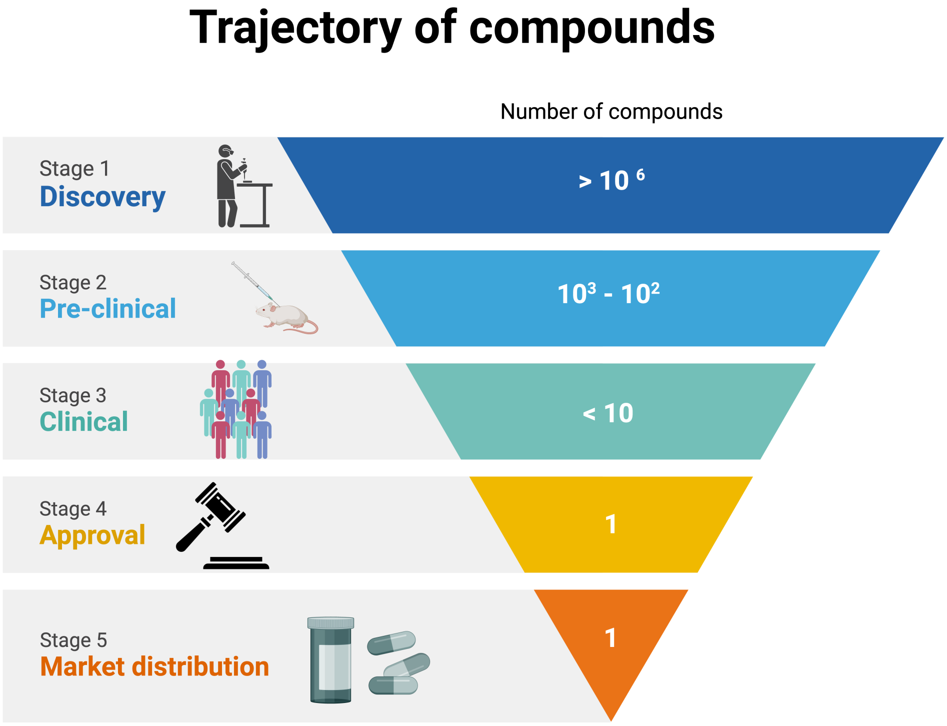


Figure 1: Drug discovery pipeline from compounds' perspective. The traditional drug discovery funnel involves a reduction in the number of identified compounds as it advances from the discovery phase toward clinical development. Figure inspired by [21].

Following the Research stage is the **Pre-clinical stage**, where experimental validation of the hypotheses occurs. Here, researchers conduct rigorous testing to identify "hit" compounds that exhibit activity against the desired target [22, 23]. These hit compounds then undergo a rigorous design and selection process aimed at optimizing their physicochemical properties. The primary objectives of this stage include enhancing compound potency and specificity, minimizing off-target interactions and mitigating potential toxicity [24]. By refining the aforementioned properties of candidate compounds, researchers aim to increase the likelihood of clinical success and pave the way for further development.

Now that we have validated our hypothesis in a laboratory setting (i.e., within a cellular environment), we progress towards validating the hypothesis at an organism level. This marks the stage of **Clinical research**, where additional parameters are assessed to understand the utility of the drug candidate. Clinical trials aim to test the intended use of the drug through safety and efficacy parameters [25, 26]. These clinical stages are divided into four subphases, Phase 1-3, each with different focuses [27].

- Phase 1 primarily focuses on evaluating the safety and dosage tolerance of the new drug.
- Phase 2 examines its efficacy and potential side effects in a small population cohort.
- Phase 3 involves monitoring the long-term adverse effects of the drug and is tested across a larger population cohort.

Following the successful completion of the three clinical phases, the drug candidate advances to the **Approval stage** or Phase 4, where the drug developer collaborates with the drug governing body to develop and refine the prescribing information for the new drug [28]. This information serves as the foundation for detailing the optimal usage of the new drug. Post-approval, the drug developer gains authorization to distribute the drug in the market. This step also establishes long-term safety monitoring protocols for the drug's efficacy, enabling collaborative adjustments to prescription and dosage information to address potential overuse measures [29]. Remarkably, only one out of a million compounds successfully advances through the drug discovery process to reach the consumer, as depicted in Figure 1.

In conclusion, while drug discovery presents numerous challenges and demands significant resources, it holds immense potential for addressing unmet medical needs. Despite the high attrition rates and complexities associated with bringing a drug to market, technological advancements, data analytics, and collaborative research efforts offer promising avenues for overcoming these hurdles [30, 31]. By leveraging cutting-edge tools and interdisciplinary collaboration, researchers are well-positioned to navigate the drug discovery process effectively and uncover new therapeutic opportunities for combating a wide range of diseases.

1.1.2 Exploring semantic rich data sources

The drug discovery process generates vast amounts of data across various stages, ranging from molecular properties of compounds to clinical trial results. This data plays a crucial role in identifying potential drug candidates, understanding their mechanisms of action, and assessing their safety and efficacy. Molecular data, such as chemical structures and properties, helps in screening and designing new compounds. Biological data, including gene expression profiles and protein interactions, provides insights into disease mechanisms and drug targets. Clinical data from trials informs on the drug's effectiveness and safety profile. Integration of these diverse datasets through advanced computational methods, such as machine learning and network analysis, enables researchers to identify promising drug candidates, predict their behaviour in biological systems and optimize their development. The utility of this data extends beyond individual drug discovery projects, contributing to the wider scientific understanding of diseases and drug responses.

Deciphering the drug discovery process requires a deep understanding of the complex interactions among genes, compounds, pathways and diseases within the human body. This understanding is essential for guiding decision-making throughout the drug development journey. As part of this process, a vast amount of data is generated, providing valuable insights into biological mechanisms and potential therapeutic targets. Collecting and standardizing this data is crucial for gaining a comprehensive understanding of biological diversity and facilitating effective selection of drug candidates. Below, we highlight some of the key experimentally derived sources of this invaluable information.

- **Bioactivity data.** Bioactivity data refers to experimental data generated from biological assays designed to measure the biological activity or potency of chemical compounds. These assays typically involve exposing biological samples, such as cells or tissues, to various concentrations (or doses) of the compound of interest and then measuring the resulting biological dose response. Bioassay data can provide valuable insights into the structure-activity relationship (SAR) and the pharmacological properties of compounds, including their efficacy, potency and potential toxicity [32]. This data, generated at the early stages of the drug discovery process, is widely used to assess the activity of candidate compounds, prioritize leads and optimize drug candidates for further development. In the past decades, a number of databases have been established to collect and structure the bioactivity data generated across the globe. These resources include PubChem [33], ChEMBL [34], BindingDB [35] among others [36]. In all these resources, there is a disproportionate representation of active and inactive compounds, with active compounds being significantly over-represented.

- ***Omics* data.** Omics data includes data generated from high-throughput assays the enable profiling and measuring comprehensively and simultaneously molecules of the same type from a biological sample. This allows for a holistic view of the biological system under study. Three major branches of omics data include metabolomics, proteomics, and genomics.

- **Metabolomics data.** Metabolomics data involves the comprehensive analysis of small molecules, referred to as metabolites, found in biological samples such as cells, tissues, or biofluids. These metabolites serve as the end products of cellular processes and participate in diverse biochemical pathways within an organism. By studying and comparing the metabolite profiles between healthy and diseased states, researchers can gain valuable insights into the underlying physiological and pathological processes. Techniques like mass spectrometry and nuclear magnetic resonance spectroscopy are frequently utilized to detect and quantify metabolites in biological samples, enabling researchers to uncover key metabolic signatures associated with various conditions. Publicly accessible databases housing such data include the Human Metabolome Database (HMDB) [37], MetaboLights [38] and Golm Metabolome Database (GMD) [39] and more [40].
- **Proteomics and Genomics.** Genomics and proteomics data play a role in deciphering the molecular mechanisms underlying diseases and drug responses at the molecular level. Genomics focuses on the comprehensive study of an organism's entire genome, including genes, their variations and their interactions. This data is instrumental in identifying disease-causing genetic mutations and uncovering potential drug targets as disease biomarkers. Proteomics, on the other hand, involves the large-scale study of proteins, their structures, functions, and interactions. Proteomic data provides crucial information on protein expression levels, post-translational modifications and protein-protein interactions, which are essential for elucidating disease pathways and identifying druggable targets. Together, genomics and proteomics data enable researchers to identify novel drug target interactions, optimize drug efficacy and personalize treatments based on an individual's genetic makeup and protein profiles. Numerous databases and resources offer access to genomic and proteomics data. For instance, the Library of Integrated Network-based Cellular Signatures (LINCS) [41] provides comprehensive gene and protein expression data related to drug or disease perturbations. Additionally, Bgee [42] offers valu-

able gene expression information at both the tissue and organism levels, facilitating cross-species comparisons and enhancing our understanding of gene expression patterns across different species.

- **Cell imaging data.** Cell imaging data encompasses the acquisition and analysis of images obtained from various cellular assays, revealing information on the morphology, behavior and function of cells in response to different experimental conditions, including drug treatments. High-throughput imaging technologies enable the simultaneous examination of thousands of cells, allowing researchers to screen large compound libraries for potential drug candidates. Alongside these, automated image analysis algorithms extract quantitative features from these images, facilitating the identification of cellular phenotypes associated with drug response or disease progression. Specifically, cell imaging data plays a crucial role in target identification, lead optimization and toxicity screening during the drug discovery process. Various experiments collect such imaging data, and corresponding databases like the Electron Microscopy Data Bank (EMDB) [43], Cell Image Library (CIL) [44], and Image Data Resource (IDR) [45] are available based on instrument type and experimental endpoints.

Throughout the drug discovery process, a wealth of data is generated to validate and provide valuable insights into biological mechanisms and potential therapeutic targets for novel drug candidates. Standardizing and collecting this data is essential for gaining a comprehensive understanding of biological diversity and enabling the effective selection of promising drug candidates. Moreover, the creation of a sustainable resource, such as those highlighted in this section, which systematically collects, stores and enables querying of this data, holds promise for reducing the overall cost of pre-clinical drug discovery. For example, by avoiding the repetition of experiments that have yielded the same conclusions in the past or by utilizing this data to train intelligent machine learning algorithms for the efficient selection of potential drug candidates.

1.1.3 Exploring non-semantic rich data sources

In addition to the data generated through laboratory experiments, a vast reservoir of information is gleaned from various scientific and non-scientific documents available on the internet [46, 47]. These documents include a wide range of sources, including research articles, books, clinical trial reports, conference proceedings, and regulatory documents. Through careful curation and analysis, valuable insights can be extracted from this wealth of textual information, contributing significantly to our understanding of biology. Furthermore, advancements in text mining, natural language processing and machine learning techniques have facilitated the systematic extraction and integration of knowledge from these diverse sources, enabling researchers to uncover hidden associations and generate novel hypotheses.

Books and scientific articles. Books and scientific articles are the fundamental resources of biomedical research. Scientific articles published in peer-reviewed journals provide detailed accounts of experimental findings, methodological approaches and theoretical frameworks relevant to drug development. These articles cover a wide range of topics, including the identification of disease targets, mechanism of action of drugs, pharmacokinetics, pharmacodynamics, and clinical trial results. Moreover, textbooks and reference books offer in-depth discussions on fundamental principles, techniques, and concepts in pharmacology, biochemistry, molecular biology, and other relevant disciplines. These resources provide researchers with a solid foundation of knowledge and serve as essential reference materials for understanding the underlying biology of diseases and the mechanisms of drug action [48–50]. By synthesizing information from books and scientific articles, researchers can gain critical insights, formulate hypotheses, and design experiments to advance drug discovery efforts.

Biological databases. Biological databases have revolutionized drug discovery by providing structured repositories of biological data, which were previously scattered across scientific literature and experimental sources in unstructured formats [51, 52]. To convert these unstructured resources into structured databases, manual, semi-automated or automated curation efforts through sophisticated text mining systems were performed. These databases offer centralized platforms where diverse types of biological data, including chemicals, proteins and diseases, are curated, standardized and made readily accessible. For instance, the STRING database offers comprehensive insights into protein-protein interactions gleaned from both literature and experimental sources [53]. Similarly, Open Targets serves as a platform aggregating metadata pertaining to targets or proteins and interactions from multiple independent resources [54]. These initiatives play a crucial role in facilitating research and fostering collaboration within the scientific community [55].

Patient-level data. In addition to the traditional research and discovery resources discussed earlier, a wealth of data is accumulated within clinical and hospital settings. This encompasses various measurements such as inflammatory signals in the blood, blood pressure readings, and data collected from smart devices like smartwatches, which monitor parameters such as heart rate and physical activity. Furthermore, advanced patient data includes radiological images obtained from techniques like computed tomography (CT) or magnetic resonance imaging (MRI), which offer insights into pathological abnormalities through visual analysis. These diverse data sources are invaluable for training machine learning (ML) and artificial intelligence (AI) models, enabling the tracking of disease progression and facilitating early detection of menacing illnesses. This approach, known as precision medicine, tailors treatment strategies to individual patients based on their unique characteristics and disease profiles [56, 57]. By leveraging patient-centric data in this manner, we can revolutionize healthcare delivery, providing personalized and proactive interventions to improve patient outcomes and quality of life.

Integrating both biomedical and non-biomedical resources forms a comprehensive repository of knowledge and information on human biology. With the continuous advancement of technologies like laboratory tool robotics and large language models, an exponential increase in the quantity and quality of data generated and analyzed in this domain. Consequently, there is a growing need for frameworks that can efficiently gather, curate and represent this wealth of information in a meaningful manner. In the next section, we look into graph databases, one of the most engaging methodologies for structurally representing this data.

1.2 Graphs as a resource for data analysis

1.2.1 What are knowledge graphs?

The concept of graphs traces its roots back to 1736 when Swiss mathematician Leonhard Euler encountered the Seven Bridges of Königsberg problem [58]. This puzzle revolved around the city of Königsberg, which featured four landmasses connected by seven bridges over the Pregel River (Figure 2). The task at hand was to devise a walking route that would cross each bridge exactly once, culminating in a return to the initial starting point. Euler tackled this challenge in a novel way by abstracting the landmasses and bridges into what we now recognize as a graph — a structure comprising vertices (representing points) and edges (representing connections between these points). By transforming the problem into a graph, Euler revealed a fundamental insight: the impossibility of finding a solution due to the disproportionate ratio of bridges and landmasses. This groundbreaking work not only resolved the Seven Bridges problem but also laid the cornerstone for the emergence of graph theory — a mathematical discipline devoted to exploring the properties and applications of graphs in diverse fields.

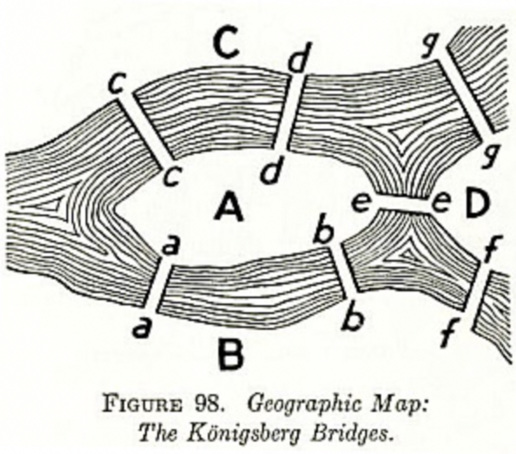


Figure 2: Depiction of the Königsberg problem. Figure taken from [59].

Commending on the evolution of graph theory as a fundamental tool, various fields, including computer science and biology, have adopted it for solving problems and scenarios. Especially in the drug discovery field, graphs have emerged as a key technology for storing and querying data, enabling applications in areas such as knowledge graph construction. Fundamentally, knowledge graphs (KGs) are structured representations of knowledge that capture relationships between entities in a domain [60]. They consist of nodes representing various entities, such as genes, proteins, diseases, drugs and pathways, with edges denoting the relationships between these entities, including activation, inhibition, and more. By connecting these entities and their relationships in a unified framework, KGs enable researchers to explore the intricate interactions, thus decoding biology. In contrast, data graphs represent the underlying data itself in a graph format. Unlike KGs, which focus on semantic relationships and domain-specific insights, data graphs primarily serve as mechanisms for organizing data within database systems. This structure allows for efficient information extraction, facilitating access and analysis of the stored data.

In the field of drug discovery, numerous biomedical KGs have been generated, drawing information from a variety of sources, including experimental and scientific evidence, as discussed in Sections 1.1.2 and 1.1.3. These KGs play a crucial role as foundational datasets for various graph-based tasks, such as link prediction [61–63], drug repurposing [64–66] and toxicity prediction [67–69]. Simultaneously, a number of benchmark KGs have been boosted in the past years to allow for modularizing the data aggregation and the machine learning task, allowing experts in both domains to contribute equally to this growing field. A critical distinction among these benchmark models is the quality of the underlying data and its downstream applicability. One notable example is OpenBioLink[1], one of the first biomedical benchmark datasets optimized for the task of link prediction [70]. This dataset is split into test and train sets, representing seven biomedical entities, namely gene, drug, disease, pathway, anatomy, phenotype, and GO term, systematically connected through knowledge gathered from public resources. Additionally, each edge in the graph has a confidence score associated with the occurrence of the relation in different resources and publications, creating a quality filter on top of the graph. The structured approach of OpenBioLink ensures that the data is both reliable and applicable for various biomedical research and machine learning tasks. On the other hand, to cope with the exponentially growing volume of research and data, information retrieval and automated knowledge discovery methods through AI are increasingly employed for the construction or updation of existing graphs. Examples of such efforts include Biomedical Knowledge Graph (BioKG) developed by Zhang *et. al.* [71] and the approach outlined by Babaiha *et al.* for updating existing pathophysiology mechanism graphs [72]. These AI-driven methods significantly enhance the efficiency of integrating new findings into comprehensive KGs, ensuring that they remain current and relevant. Moreover, questions regarding the final quality of the data ingested into these comprehensive graphs still remain [73].

1.2.2 Frameworks, architecture and tools for graph representation

As discussed in the previous section (Section 1.2.1), KGs play a pivotal role in efficiently representing complex data relationships. Thus, it is essential to delve into the different ways KGs are represented and developed by the

[1] `https://github.com/OpenBioLink/OpenBioLink`

community. Over the past years, KGs have gained recognition as powerful frameworks for pattern identification, leading to an increase in the diversity of representation formats and the development of sophisticated tools for generating and visualizing KGs.

KGs can be represented using different formats, each offering unique advantages for specific applications. Common representation formats include:

- **Resource Description Framework (RDF)**: RDF is a standard framework developed by the World Wide Web Consortium (W3C) for data interchange on the web[2] [74, 75]. It is structured using the RDF Schema (RDFS[3]), which provides the formal specification of concepts and relationships, ensuring semantic consistency and interoperability in a RDF graph. These definitions are often enhanced through use of Uniform Resource Identifiers (URIs) and, ontologies and controlled vocabularies, such as those provided by the OBO Foundry (`https://obofoundry.org/`). An RDF graph is classically composed of RDF triples, where each triple comprises of a subject, a predicate, and an object, collectively representing data in a structured and machine-readable format[4] (Figure 3). The resultant graph is stored in a triple-store and queried using SPARQL, a specialized query language.

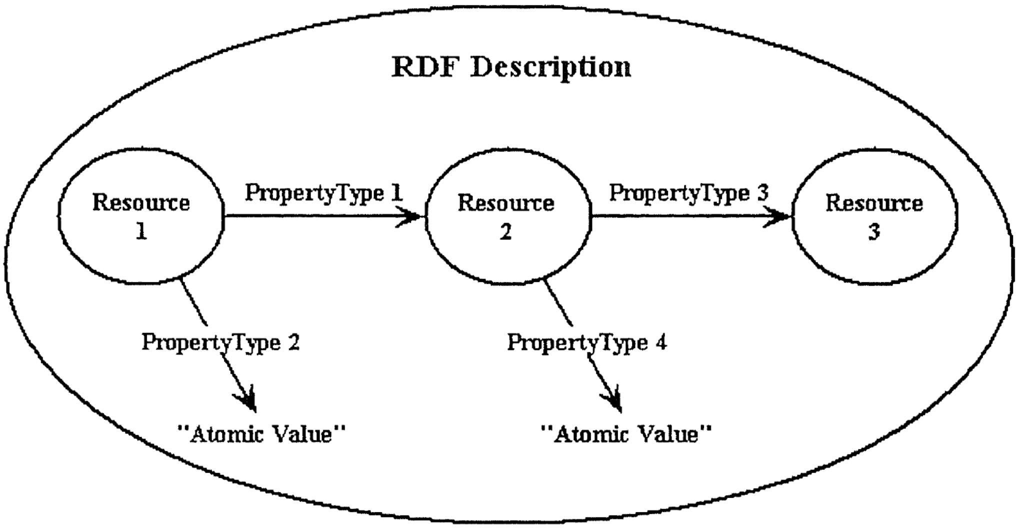


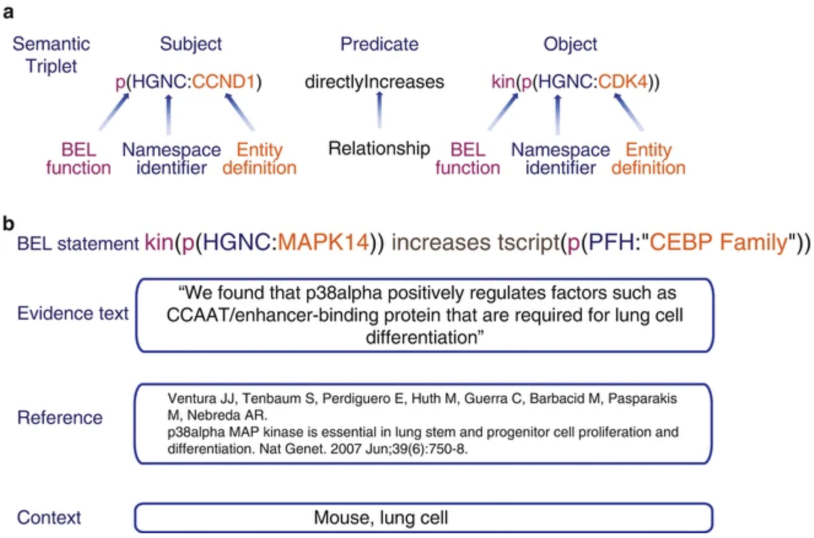
Figure 3: Depiction of an RDF Description. Figure taken from [74].

[2] `https://www.w3.org/TR/rdf-concepts/`

[3] `https://www.w3.org/TR/rdf-schema/`

[4] `https://www.w3.org/TR/rdf11-concepts/`

- **Biological Expression Language (BEL)**: BEL, created more than a decade ago by Selventa, is a specialized language developed to represent complex biological relationships in a structured and interpretable (both human and machine) format [76]. Fundamentally, it is designed to capture and convey knowledge about biological entities, such as genes, proteins, and small molecules, and their interactions with causal relationships. Similar to RDF, it encodes data in a triple-based structure where terms are grouped into subject–predicate–object BEL statements, with each statement describing a scientific finding (Figure 4).



"We found that p38alpha positively regulates factors such as CCAAT/enhancer-binding protein that are required for lung cell differentiation"

Ventura JJ, Tenbaum S, Perdiguero E, Huth M, Guerra C, Barbacid M, Pasparakis M, Nebreda AR.
p38alpha MAP kinase is essential in lung stem and progenitor cell proliferation and differentiation. Nat Genet. 2007 Jun;39(6):750-8.

Mouse, lung cell

Figure 4: An example of a BEL statement. (a) It shows the semantic dissection of a BEL triple into BEL-relevant entities (i.e. function, namespace, and entity). (b) It shows the translation of an "evidence" statement into the BEL statement with cross-reference. Figure taken from [76]

- **Property graph**: Recently democratized by commercial solutions like Neo4J (`https://neo4j.com/`), property graphs, also known as labeled property graphs, have emerged as powerful and flexible frameworks for representing and managing complex networks of interconnected data [77]. These graphs support the representation of both nodes (entities) and edges (relationships) with associated properties stored as a key-value pair (Figure 5). Additionally, nodes can be assigned one or more labels, allowing developers to tag or group them effectively. Unlike RDFs, property graphs offer greater structural flexibility, enabling rich, multi-dimensional data representation without a strict focus on semantic interoperability. This makes them particularly well-suited for scenarios involving intricate interrelations, such as social networks, recommendation systems, or datasets with entity- and relation-specific metadata, such as experimental protocols. Their adaptability, combined with an intuitive and expressive nature, facilitates efficient querying and analysis of complex datasets. Similar to RDFs, property graphs employ a specialized query language, CYPHER, to navigate and manipulate the data effectively.

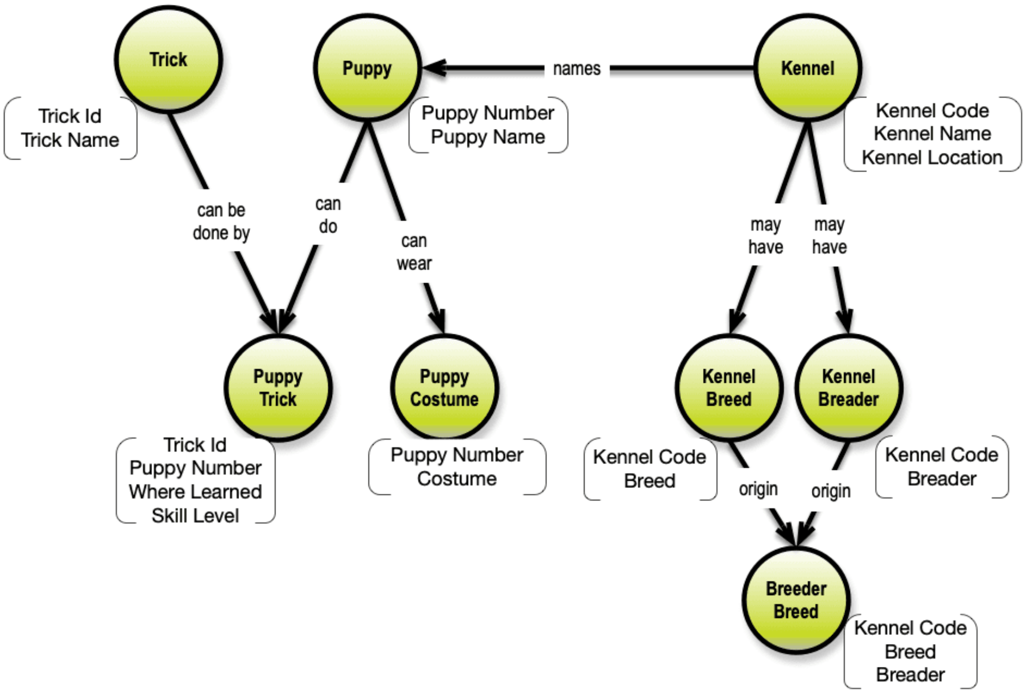


Figure 5: An example of a property graph model. As show in the figure, each node (in circle) has at least two or more properties associates with it. Figure taken from [78].

The representation formats mentioned thus far are only a subset from a larger list[5]. In addition to these, various tools have been developed to generate and visualize KGs. These tools facilitate the creation, management, and exploration of KGs, making it easier for researchers and data scientists to work with complex datasets. One such tool is Neo4j, a graph database that uses the property graph model to store and manage data. Neo4j offers a robust query language (called CYPHER) and a range of visualization options to explore graph data interactively. Its user-friendly interface allows for quick onboarding of both programmers and non-programmers, making

[5] `https://w3id.org/faircookbook/FCB070`

it a successful resource for future multi-stakeholder graph-based projects. Another notable tool is Ontotext's GraphDB (`https://www.ontotext.com/products/graphdb/`), which supports efficient storage, querying (through SPARQL), and visualization of highly scalable RDF databases. GraphDB integrates with various visualization tools and offers advanced reasoning capabilities. Alongside these software solutions, numerous libraries such as D3.js (`https://d3js.org/`), vis.js (`https://visjs.org/`), and Stanford Network Analysis Platform (SNAP) (`https://snap.stanford.edu/snap/`) provide extensive options for customization and visualization of large-scale, typically characterized with more than million nodes, graphs. These tools collectively enhance the usability, reachability, and applicability of KGs, making them indispensable in handling and interpreting complex, interconnected data across diverse stakeholder groups.

1.3 Patent documents as novel source of data

This section presents text from the following book chapter:

> **Yojana Gadiya**, Reagon Karki, Philip Gribbon, and Andrea Zaliani. 2025. Artificial intelligence-driven patent analysis in drug discovery. *Comprehensive Medicinal Chemistry 4th Edition - In Silico Tools. In press*

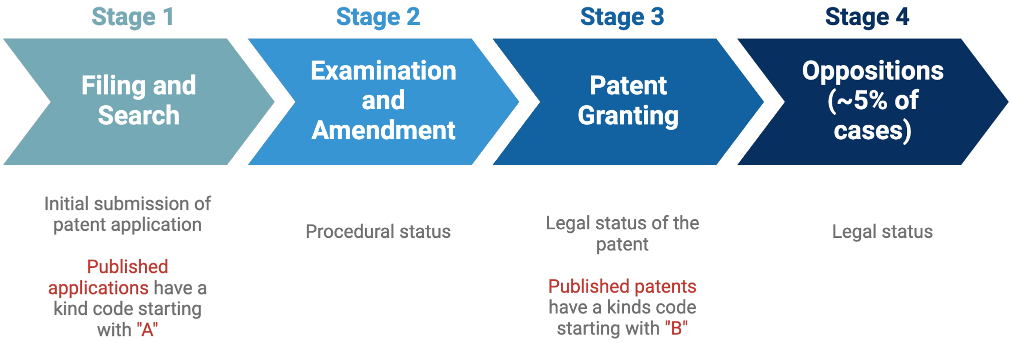
1.3.1 Patent filling process

As defined by the World Intellectual Property Organization (WIPO), a patent is an exclusive right granted for an invention[6]. These are written in a legal framework and provide rights to the inventor to make, use, and sell the invention for a definite period. Within drug development, these documents play a critical role in assessing the competitive landscape surrounding a lead candidate. In contrast, within drug discovery, they remain largely untapped resources.

Each patent document undergoes a rigorous approval and granting process (Figure 6). This process begins with the filing of the patent to a jurisdiction

[6] `https://www.wipo.int/patents/en/`

such as the United States (US), Europe (EU), or Japan (JP), among others. Upon filing, the patent document is assigned a kind code, typically starting with “Ax”. The kind code is an alphanumeric code used to tag and distinguish different types of utility patent documents based on their publication stage. Along with the kind code, each registered patent is given a unique identifier consisting of a two-letter country code and a patent assignment number. In recent years, in addition to local jurisdiction filings, super-jurisdictional filings have become a standard practice in pharmaceutical discovery (for e.g. application under the Patent Cooperation Treaty (PCT) allowing for IP across a large number of countries) [79]. This approach helps to streamline the process, reducing the costs and bureaucratic complexity associated with multiple filings across different jurisdictions. Once identified, the document undergoes a thorough examination and review by patent lawyers or officers who assess the scope of the patent and identify relevant prior art with similar scopes. Lawyers meticulously ensure that the novelty claimed in the patent document is not found in any public resources to date, including websites, books, publications, patents, and blogs. This examination process involves a cycle of discussions between the patent applicant (the owner or submitter of the patent document) and the patent office to address any open questions and refine or amend the patent as needed. When the patent office is satisfied with the description and claims of the patent document, its legal status changes from filed (with kind code Ax) to granted (with kind code Bx). This status change signifies that the patent is now officially published, and the technology described is trademarked by the patent owner. The whole process, from filing to approval of a patent, can take 18 months. This comprehensive process ensures that the patent is thoroughly vetted and its novelty and originality are established, thereby providing robust protection for the inventor’s intellectual property.



Stage 1
Filing and Search
Stage 2
Examination and Amendment
Stage 3
Patent Granting
Stage 4
Oppositions (~5% of cases)
Initial submission of patent application
Published applications have a kind code starting with "A"
Procedural status
Legal status of the patent
Published patents have a kinds code starting with "B"
Legal status

Figure 6: Patent granting process. The figure was taken from the Espacenet documentation on patent filing (`https://e-courses.epo.org/mod/streaming/view.php?id=9277`).

1.3.2 Patent document applications in drug discovery

As mentioned in the prior section, each patent document is scrutinized by an expert panel of lawyers, ensuring its transition from a patent application (with filing status) to a patent (with granted status). During the process, the documents are also assigned International Patent Classification (IPC) codes, which aid in categorizing them into relevant technology areas[7]. However, not all IPC codes are applicable to pharmaceuticals. Only a subset of these IPCs are relevant for pharmaceutical usage. Among them, the A61 (referring to "MEDICAL OR VETERINARY SCIENCE; HYGIENE") is the most recognized and widely used classification in the pharmaceutical field. This categorization enables efficient identification and analysis of patents related to drug development, medical treatments, and healthcare innovations, supporting better patent landscape analysis and decision-making in pharmaceutical research.

In the pharmaceutical industry, patent documents are leveraged primarily for patenting a drug or new chemical entity (NCE), showing promise for desired activity [80]. These patents not only protect the investment in research and development (R&D) but also grant the patent holder exclusive rights to manufacture, use, and sell the compound for a specified period. This exclusivity incentivizes innovation by ensuring that companies can recoup their R&D investments. Additionally, patents can include extensive information

[7] `https://www.wipo.int/en/web/classification-ipc`

on the synthesis process of the compound(s), their biological activity, their formulation, and their potential therapeutic uses. These pieces of information can be found distributed throughout the introduction, description, and claims sections of a patent document. This documentation serves as a comprehensive reference that can aid in further research and development, regulatory approval processes, and strategic planning for market entry. Furthermore, patents often highlight the unique structural and functional aspects of the compound, distinguishing it from prior art and emphasizing its novelty and utility. By leveraging patent data, companies can also track competitor activity, identify potential collaboration opportunities, and navigate the complex landscape of pharmaceutical innovation. Alongside compounds, patent documents offer insights into the evolution of research on specific gene targets. They provide a historical perspective, helping prioritize gene targets by revealing which ones have been the focus of sustained research and development efforts [81]. Additionally, patents contain claims on drug discovery tools, such as cDNA fragments, crystallized receptors, and protein probes. Understanding these claims is essential for navigating the intellectual property (IP) landscape and ensuring that new pharmaceutical targets are developed within a legally sound framework. Overall, the strategic analysis of patent documents is a powerful tool in drug discovery. It enables researchers to gather intelligence, generate hypotheses, prioritize targets, and understand the IP landscape. By integrating this approach into the research process, scientists can enhance their ability to innovate and develop effective new drugs.

1.3.3 Tools and resources for patent documents

Given the significance of patent documents, several open-source and commercial tools have been developed to capture and analyze patent literature. Below, we highlight a few popular tools and explore their capabilities (Table 1).

- **Espacenet**[8]: Dating back to patents from 1782, Espacenet offers open access to over 150 million patents worldwide. Developed by the European Patent Office (EPO), Espacenet is designed to store, retrieve, search, and query patent documents efficiently. This comprehensive database enables researchers, inventors, and professionals to explore a vast array of patented inventions and technological advancements

[8] https://worldwide.espacenet.com/patent/

across 35 technology fields. These fields include electrical engineering, chemistry, furniture, and food. Alongside searching for existing patents, Espacenet provides translation of non-English patents to English, thus serving as an advanced platform for patent landscaping. Patent offices make use of free text and wildcards to search the resources for chemical-relevant patents and claims [82].

- **PubChem with Google patents**[9]: PubChem, funded by the NIH and maintained by the National Center for Biotechnology Information (NCBI), stands as one of the largest and most comprehensive open chemistry databases globally [33]. It offers a vast repository of data encompassing both small molecules, such as fragments and drugs, and large molecules, including peptides and lipids. With data aggregated from over 1,000 diverse sources, PubChem provides an extensive array of detailed information on chemical compounds. This includes their structures, properties, biological activities, patents, safety, and toxicity profiles. A notable feature of PubChem is its collaboration with Google Patents (`https://patents.google.com`) to integrate compound-patent metadata. This partnership significantly enhances the knowledge base by linking chemical compounds to a wealth of patent information. However, it also introduces potential challenges, such as the incorporation of noise due to the vast volume of data [83]. The database is remarkably expansive, containing billions of references related to chemical compounds, diseases, proteins, genes, and more. These references are meticulously extracted from a variety of sources, including full-text documents, images, PDFs, and translations of approximately 130 million patents.

- **SureChEMBL**[10]: SureChEMBL, managed by the European Bioinformatics Institute (EMBL-EBI), is a comprehensive patent database that excels in extracting and indexing biomedical entities such as genes, diseases, organisms, drugs, and chemicals cited in patent documents [84]. The database integrates and harmonizes data from the IFI CLAIMS (`https://www.ificlaims.com/start.htm`) patent platform, enabling users to delve into the chemical landscape within patents, monitor the development of new compounds, and identify potential areas for innovation. What sets SureChEMBL apart is its unique capability to run a named entity recognition (NER) model that can dynamically annotate biomedical entities in any patent document of interest. Thus,

[9] `https://pubchem.ncbi.nlm.nih.gov/`

[10] `https://www.surechembl.org/`

the resource offers deep insights into the intersection of biomedical data and patent literature.

- **Chemical Abstract Service**[11]: Supported by the American Chemical Society (ACS), the Chemical Abstracts Service (CAS) is a premium database and registry for chemical compounds. CAS excels in curating data through both manual and automated methods, ensuring comprehensive and accurate chemical information [85]. This approach has enabled CAS to launch several solutions catered to the needs of drug discovery and other scientific research fields. One of CAS's flagship offerings is the CAS Registry (`https://www.cas.org/cas-data/cas-registry`), which collects extensive details on chemical substances, including chemical names, structures, and properties. This registry is one of the most authoritative sources of chemical information available, providing researchers with reliable data for their studies. CAS Patents (`https://www.cas.org/cas-data/cas-patents`) is another critical service offered by CAS, indexing patent documents from over 100 patent offices worldwide. This service provides structured information, including chemical structures and patent-related metadata.

- **Derwent**[12]: Derwent's patent database provides a competitive edge in its integration of the Derwent World Patents Index with the Derwent Patent Citation Index. This combination enables users to perform both patent and citation searches simultaneously, providing a more comprehensive view of the patent landscape. The database is particularly useful for reducing duplication in R&D efforts, tracking competitors' activities, avoiding patent infringement, and identifying potential gaps in the marketplace or licensing opportunities.

  We have listed only the most widely used tools for patent landscape analysis. However, the number of available tools (particularly commercial ones) is steadily increasing each month as the demand for comprehensive intellectual property analysis grows.

[11] `https://www.cas.org/`

[12] `https://clarivate.com/products/ip-intelligence/ip-data-and-apis/derwent-world-patents-index/`

Name	**Updates**	**Coverage**	**Type**
Espacenet	Daily	> 150 million patent documents	Public
PubChem	Annually	> 21 million patent documents	Public
SureChEMBL	Daily	24 million patent documents	Public
CAS	Daily	> 109 patent authorities	Private
Derwent	Daily	20 million patents	Private

Table 1: Summary of the patent resources and their data coverage.

1.4 Organization and aims of this thesis

The abundance of biomedical data available is invaluable for gaining insights into the understanding of biological mechanisms within the human body. The crucial task of ingesting, interpreting and honing its utility (by identifying meaningful patterns within this data) can be effectively addressed through the modelling of relationships between biological entities and their concepts in the form of networks. This dissertation primarily focuses on constructing graphical networks for the interpretation and contextualization of high-dimensional biological data. Additionally, it acknowledges the limitations inherent in the data present in biomedical networks and proposes remedies to address some of these. The dissertation concludes by showcasing the translation of computation analysis into real-world scenarios. Thus, the ensuing chapters of this thesis delve into three main objectives, presenting specific contributions to the field:

i. Develop mechanisms for the *generation of knowledge graphs* through the ingestion of multi-modular data resources (Chapter 2).

ii. Recognize the requirement for *incorporating patent literature* absent in current biomedical graph generation framework (Chapter 3).

iii. Demonstrate scenarios of *digital-to-laboratory validation* for empowering utility of *in-silico* approaches for drug discovery (Chapter 4).

In Chapter 2, we present two contrasting approaches that leverage existing data resources for the generation of knowledge graphs. Specifically, we introduce the ability to aggregate experimental and *in silico* data through established resources in a meaningful and efficient way (as per Section 2.2). In this work, we built a reproducible pipeline for future infectious diseases where sparse literature research exists, signifying preparedness for pandemic situations. Contrary to this, in Section 2.1, we built the graph centric around scientific literature. In both of these approaches, we demonstrated the usage of the graph to make deductions on potential drug candidates, especially in the context of drug repurposing as therapeutics for the indicated area.

Then, in Chapter 3 we delve into the limitations of current biomedical knowledge graphs. Documents such as patents are omitted from these resources due to a number of reasons. This chapter emphasizes the importance of considering patent documents as a resource for ingestion by demonstrating the vast amount of knowledge present in them (as outlined in Section 3.1). Furthermore, we developed a tool called PEMT (as indicated in Section 3.2) that can strategically integrate with current knowledge graph approaches to collect and expand the graph around pharmaceutical patent literature.

Finally, in Chapter 4, we conclude with three publications that adopt methods from graph theory for drug discovery-based applications. Specifically, these include Patent landscaping (as per Section 4.1), Antimicrobial Modeling (as per Section 4.3) and RPath (as per Section 4.2). While patent landscaping focuses on utilizing patent information extracted from PEMT for target prioritization and competitive landscaping, the Antimicrobial model demonstrates the capability of training *in-vitro*-validated machine learning models for the future. Finally, in concluding this chapter, we introduce RPath, which focuses on overlaying graph networks with transcriptomic signals (e.g., drug and disease). RPath's main functionality is to deconvolve a compound's potential mechanism of action.

Subsequently, these chapters are followed by a discussion of the themes presented, challenges faced, and possible future directions, which serves as the general conclusion of this thesis.

CHAPTER 2

Data warehousing through knowledge graphs

The advancements in data generation technologies, such as automated screening facilities and *omics* technologies, have triggered an explosion of high-throughput biological data, consequently leading to the development of numerous biological databases. However, no single database is comprehensive, necessitating efforts to improve data interoperability, reproducibility, and the integration of knowledge derived from these technological advancements. One systematic approach to addressing this challenge is through the use of knowledge graphs (KGs), which collate and represent databases graphically by connecting various entities (nodes) through syntactical relationships (edges). Additionally, a key challenge with these databases is the need for regular updates to maintain their relevance and accuracy. Therefore, efficient workflows must be established to facilitate the continuous updating of KGs with the latest information from all relevant databases. In this chapter, we discuss a couple of methodologies we have developed for generating and maintaining knowledge graphs, ensuring their scalability, adaptability and long-term utility in biological data integration.

2.1 COVID-19 Knowledge Graph: a multimodal, computable, cause-and-effect knowledge model of COVID-19 pathophysiology

This section presents the following publication (**see Appendix A.1**):

> Daniel Domingo-Fernández, Shounak Baksi, Bruce Schultz, **Yojana Gadiya**, Reagon Karki, Tamara Raschka, Christian Ebeling, Martin Hofmann-Apitius and Alpha Tom Kodamullil. 2021. COVID-19 Knowledge Graph: a computable, multi-modal, cause-and-effect knowledge model of COVID-19 pathophysiology. *Bioinformatics* 37, no. 9: 1332-1334.
>
> `https://doi.org/10.1093/bioinformatics/btaa834`

Authors' contributions

Daniel Domingo-Fernández: Conceptualization, Supervision, Writing - Original draft preparation and Visualization. *Shounak Baksi*: Methodology and Writing - Original draft preparation. *Bruce Schultz*: Software and Visualization. *Yojana Gadiya*: Methodology, and Writing - Reviewing and Editing. *Reagon Karki*: Methodology, and Writing - Reviewing and Editing. *Tamara Raschka*: Methodology, and Writing - Reviewing and Editing. *Christian Ebeling*: Software, and Writing - Reviewing and Editing. *Martin Hofmann-Apitius*: Conceptualization, and Writing - Reviewing and Editing. *Alpha Tom Kodamullil*: Conceptualization, Supervision and Writing - Original draft preparation

Summary

The COVID-19 pandemic, caused by the novel coronavirus *SARS-CoV-2*, emerged as a global health crisis in late 2019, profoundly impacting societies worldwide. In response to the rapid spread of the virus, there was an urgent need to develop novel antiviral therapeutics. Addressing this challenge necessitated a profound understanding of the intricate mechanisms underlying viral replication, infection and transmission. Numerous researchers mobilized efforts in this respect, leading to an exponential surge in scientific publications on the pathophysiology of the virus. However, the sheer volume of information generated posed a potential obstacle. Without coordinated efforts to systematically organize this knowledge, it risked becoming fragmented within individual research groups, leading to data and information silos.

One potential approach to tackle this situation would be through the generation of knowledge graphs (KGs) (introduced in Section 1.2.1). In the publication, *COVID-19 Knowledge Graph: a computable, multi-modal, cause-and-effect knowledge model of COVID-19 pathophysiology*, we systematically identified and manually curated over 150 full-text scientific publications from PubMed focused on the novel coronavirus. Our objective was to construct a comprehensive cause-and-effect network called COVID-19 Pharmacome to mitigate the existence of information silos. To achieve this goal, we curated a literature corpus, specifically incorporating publications discussing potential drug targets for COVID-19, elucidating biological pathways affected by the virus during replication within its human host, and providing insights into the downstream biological functions of viral proteins in the host.

To ensure interoperability across the different publications, each article was manually encoded, curated and transformed into a BEL complaint triple format featuring a structured source-relation-target framework (as outlined in Section 1.2.2). The COVID-19 Pharmacome comprises 4,016 nodes spanning over ten biological entity types, including chemicals and genes, and 10,232 relationships, culminating in the creation of a coherent and expansive KG. Additionally, we embedded the KG into a Biological Knowledge Miner[1] with an OrientDB database server, facilitating the effortless exploration of molecular interactions within BEL-driven networks.

In addition to its search utility, the COVID-19 Pharmacome has proven valuable in specific applications related to heme biology and drug repurposing

[1] For more information see `https://www.covid19-knowledgespace.de/`

(Figure 7). Hopp *et al.* (2021) conducted an analysis by superimposing the Heme KG [86] with the COVID-19 Pharmacome, identifying four biomarker proteins (i.e. two host cell proteins namely ACE2 and TMPRSS2, and two viral proteins namely 7a and S protein) associated with COVID-19 infection [87]. This collaborative effort revealed shared inflammatory pathways among these biomarkers, uncovering potential targets for therapeutic development for COVID-19. Concurrently, Schultz *et al.* (2021) extended the COVID-19 Pharmacome by integrating ten distinct disease maps and multiple experimental data resources related to *SARS-CoV-2* pathophysiology [88]. This extension enabled the identification of novel drug combinations such as Remdesivir-Thioguanosine and Nelfinavir-Raloxifene. Experimental validation further confirmed the efficacy and synergistic effects of these repurposed combinations. Overall, these use cases exemplify the translational potential of the COVID-19 Pharmacome in advancing our understanding and therapeutic strategies for combating COVID-19.

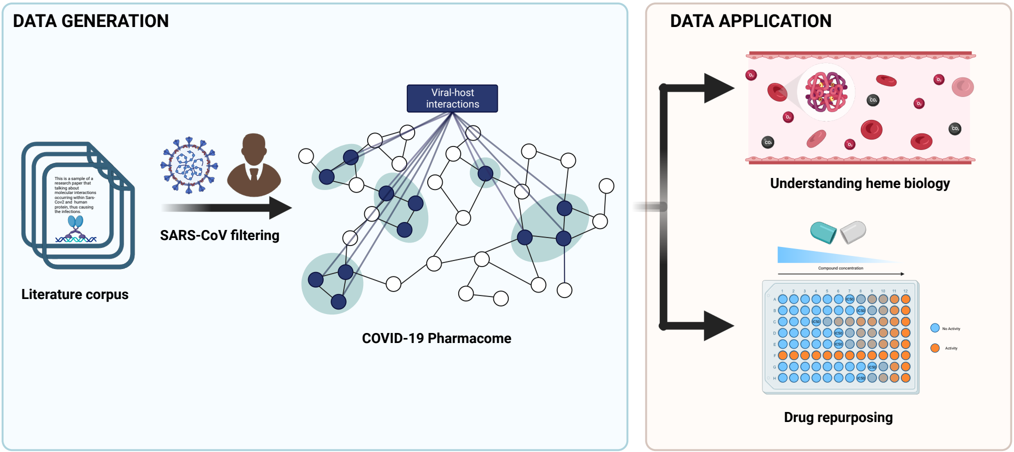


Figure 7: Graphical summary of paper. Created with BioRender.com

2.2 Mpox Knowledge Graph: a comprehensive representation embedding chemical entities and associated biology of Mpox

This section presents the following publication **(see Appendix A.2)**:

> Reagon Karki, **Yojana Gadiya**, Andrea Zaliani and Philip Gribbon. 2023. Mpox Knowledge Graph: a comprehensive representation embedding chemical entities and associated biology of Mpox. *Bioinformatics Advances* 3, no. 1: vbad045.
>
> `https://doi.org/10.1093/bioadv/vbad045`.

Authors' contributions

Reagon Karki: Conceptualization, Methodology, Software, Supervision, Writing - Original draft preparation, Validation and Visualization. *Yojana Gadiya*: Writing - Original draft preparation, and Writing - Reviewing and Editing. *Andrea Zaliani*: Writing - Original draft preparation, and Writing - Reviewing and Editing. *Philip Gribbon*: Writing - Reviewing and Editing.

Summary

In May 2022, the outbreak of *Monkeypox virus* (Mpox) prompted the World Health Organization to declare it a global health emergency.[2]. This development raised concerns globally as Mpox cases emerged across numerous countries, highlighting the persistent threat and the potential for a novel pandemic. This situation underscores the importance of a swift response, particularly considering the scientific unpreparedness witnessed during the COVID-19 pandemic. With the number of Mpox cases on the rise, an urgent and well-informed approach is paramount.

Addressing this situation could involve a potential strategy of generating biomedical networks from available literature resources, employing a combination of manual curation and text-mining-based tools. Given the scarcity of literature resources on Mpox, it is essential to devise strategies to overcome such limitations. In the publication, *Mpox Knowledge Graph: a comprehensive representation embedding chemical entities and associated biology of Mpox*, we present a methodology for developing a KG in cases where limited or no information on the infectious disease can be found in the literature. In order to do so, we built a BEL-compliant KG to depict biomedical entities (e.g. proteins and chemicals) and the fundamental mechanisms connecting these entities in the context of Mpox. We identified and collected chemicals from chemical data resources such as PubChem [33] and ChEMBL [34], and protein and disease interactions from UniProt [89] and DISEASES [90] respectively (Figure 8). We maximized the information utility from UniProt and ChEMBL by extending the KG with associated assays, biological processes, molecular functions, and pathways. This was one of the first attempts to combine experimental, curated and text-mining information for scientific endeavours in Mpox.

As mentioned previously, we retrieved chemical and bioassay data relevant to Mpox from PubChem and ChEMBL. To enhance the specificity of bioassays, we filtered them based on two criteria: a) chemicals demonstrating activity in the submicromolar range and b) assays capturing direct chemical interactions, focusing on binding or functional assays. Ultimately, the Mpox KG consists of 9,117 nodes and 44,516 relationships capturing associations within the domain of Mpox. Notably, an instance of the graph in the CytoScape compliant format was hosted on the NDExbio platform[3].

[2] `https://www.ecdc.europa.eu/en/monkeypox-outbreak`

[3] Network available at `https://doi.org/10.18119/N9SG7D`

Moreover, information on *in-silico* viral-host interactions was limited as the research on Mpox was still in its early stages. To address this, orthologs for the viral proteins in the graph were identified through BLAST sequencing, a method previously validated by Zhou *et al.*, 2014 [91]. This analysis revealed a close protein ortholog of Mpox in *Vaccinia virus* (OPG148), which is crucial for viral DNA replication through interaction with uracil, shedding light on the Mpox viral replication process. From our KG, approved compounds like Nevirapine and Zalcitabine were found to share substructure similarities with uracil. The hypothesis is that their larger chemical structure (about 100 Dalton difference) could act as a potential competitive inhibitor for uracil DNA glycosylase, inhibiting the functioning of the protein and thereby suppressing viral replication. Furthermore, a similar substructure was observed in Tecovirimat, currently used as a therapeutic for Mpox[4]. Using the approach developed in this study, we aim to identify novel therapeutics for future infectious diseases characterized by limited research availability. This proactive strategy positions us to swiftly deploy experiments, creating a pathway for developing effective countermeasures.

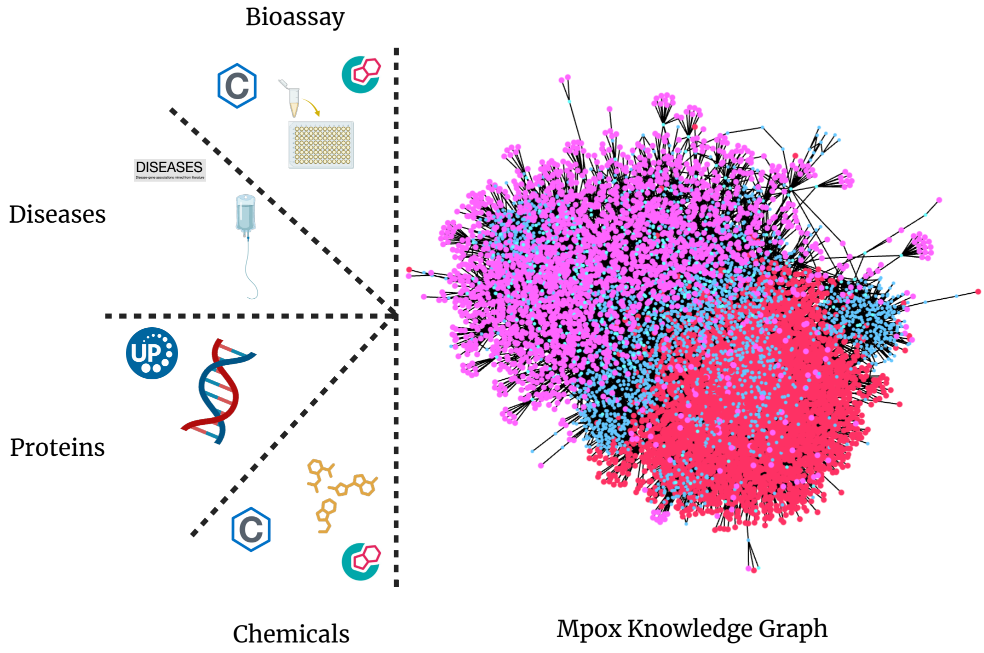


Figure 8: Graphical summary of paper. Created with BioRender.com

[4] https://www.cdc.gov/mpox/hcp/clinical-care/tecovirimat.html

CHAPTER 3

Unlocking patent documents as a unique data resource

Traditionally, biomedical research has relied on information extracted from scientific literature and experimental data. While these resources are invaluable for drug discovery and lead identification, an often underutilized yet highly valuable source is patent literature. Patent documents provide unique insights into emerging research areas, particularly from market and commercial perspectives, helping to identify potential growth opportunities in the pharmaceutical and biotechnology sectors. In this chapter, we explore the impact of integrating patent literature into knowledge graphs through two key studies. These studies collectively investigate the potential of patent documents as a data source for constructing knowledge graphs and propose a pipeline for incorporating extracted information into existing knowledge frameworks. Our approach began with an assessment of the availability and extent of chemobiology data in the open-access patent database SureChEMBL[1]. Building on this, we developed a tool that enables systematic searching and querying of pharmaceutical patent documents within this resource. By leveraging these insights, we highlight the potential of patent literature as a valuable complement to traditional data sources in drug discovery and knowledge integration.

[1] https://www.surechembl.org/

3.1 Exploring SureChEMBL from a drug discovery perspective

This section presents the following publication **(see Appendix A.3)**:

> **Yojana Gadiya**, Simran Shetty, Martin Hofmann-Apitius, Philip Gribbon and Andrea Zaliani. 2024. Exploring SureChEMBL from a drug discovery perspective *Scientific Data*, 11, 507.
>
> `https://doi.org/10.1038/s41597-024-03371-4`

Authors' contributions

Yojana Gadiya: Conceptualization, Methodology, Software, Supervision, Writing - Original draft preparation, Validation, and Visualization. *Simran Shetty*: Resources, Methodology and Software. *Andrea Zaliani*: Writing - Original draft preparation, Methodology and Validation. *Philip Gribbon*: Writing - Reviewing and Editing, and Validation. *Martin Hofmann-Apitius*: Writing - Reviewing and Editing

Summary

Patent documents have played significant roles in drug discovery, ranging from competitive and academic intelligence to research and innovation index (see Section 1.3.2). Acknowledging the importance of these legal documents, several manual and automated workflows have been developed to annotate biomedical-related entities such as protein, compounds, and gene sequences accurately [35, 92, 93]. Of these, the annotation of chemical compounds from patent documents has played a pivotal role for medicinal chemists to understand the underlying patenting landscape for potential drug lead candidates. Consequently, the establishment of a public patent database like SureChEMBL has allowed academic researchers to leverage patent document annotation tools that would otherwise be inaccessible due to their commercial nature (see Section 1.3.3) [84]. In this section and the one that follows, we introduce you to data found in a public patent data resource namely, SureChEMBL and its applicability in drug discovery pipelines.

While several patent intelligence surveys have explored the information within patent literature, the choice of datasets used for these analyses have often been constrained by factors such as time limitations (e.g., until 2017), the type of patent (e.g., IPC code A61K), or the specific country in which the patent document is filed and/or granted (typically the United States). In the publication titled *Exploring SureChEMBL from a drug discovery perspective*, we adopt a pharmacological perspective to scrutinize and evaluate the data within SureChEMBL. The dataset considered in this study surpasses previous limitations encompassing an exhaustive time period (from 2015 onward), all types of life science patent documents and spanning across the globe (i.e., the United States, Europe and Japan). The primary goal of the paper was to understand the quality of data annotated by SureChEMBL's open-source and proprietary tools, thereby shedding light on the significance of patent data. Consequently, we conduct an evaluation of the database from both a research and development (R&D) and medicinal chemistry standpoint to assess the characteristics of the chemical space found in patent documents for drug discovery.

We designed a set of chemoinformatics experiments to investigate: a) the discoverability of the compounds within SureChEMBL in publicly accessible chemical databases such as PubChem [33], ChEMBL [34] and DrugBank [94], b) the chemical space of compounds concerning their drug-like properties, and c) the documentation of the transition of compounds from pre-clinical

to clinical stages based on data available in established repositories such as DrugBank (Figure 9).

From the aforementioned experiment, we uncovered four key findings. Firstly, an impressive over 90% of compounds successfully linked to public resources like PubChem. This was attributed to their automated integration and annotation processes, particularly from Google patents. Secondly, we observed that a mere 0.02% of the extensive 10 million compounds in SureChEMBL have transitioned to the market, underscoring a predominance in the research phase. Furthermore, the analysis of the molecular properties of the preclinical patent compounds revealed a surge in patent development around macrocyclic compounds, like cyclic peptides. This trend was discerned through meticulous assessments of physicochemical properties. Notably, a fraction of data within SureChEMBL encapsulates non-drug candidates like assay-interfering compounds. Conclusively, delving into the scaffold space, as defined by Bemis-Murcko, debunked a rich chemical space of approximately 3 million distinct scaffolds within this comprehensive literature source.

Through this work, we outlined the extensive chemobiology information found within patent documents collected in SureChEMBL. This was achieved by emphasizing three key outcomes: i) recognition of the need for a stronger correlation between SureChEMBL and prominent compound public databases such as PubChem and ChEMBL, ii) the realization that reusing data from SureChEMBL required a pre-filtering to accurately select lead-like compounds with potential bioactivity when modelling their corresponding activity landscape, and iii) by incorporating these compounds into the existing chemical space, we expanded our access to a broader and more diverse range of chemical entities. Thus, this work guides researchers in maximizing the potential utility of SureChEMBL for drug discovery perspectives.

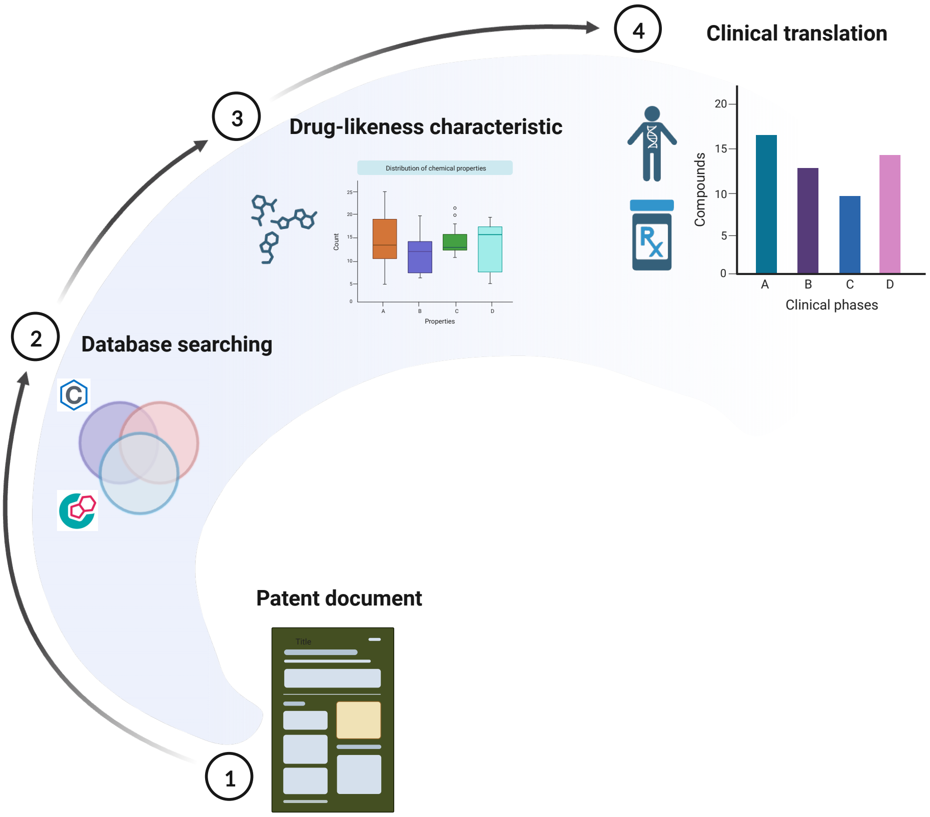


Figure 9: Graphical summary of paper. Created with BioRender.com

3.2 PEMT: a patent enrichment tool for drug discovery

This section presents the following publication **(see Appendix A.4)**:

> **Yojana Gadiya**, Andrea Zaliani, Philip Gribbon and Martin Hofmann-Apitius. 2023. PEMT: a patent enrichment tool for drug discovery. *Bioinformatics*, 39(1), btac716.
>
> `https://doi.org/10.1093/bioinformatics/btac716`

Authors' contributions

Yojana Gadiya: Conceptualization, Methodology, Software, Writing - Original draft preparation, Validation and Visualization. *Andrea Zaliani*: Conceptualization, Writing - Original draft preparation, Methodology, Supervision and Validation. *Philip Gribbon*: Writing - Reviewing and Editing. *Martin Hofmann-Apitius*: Writing - Reviewing and Editing

Summary

Drug discovery research has relied in recent years on hypotheses generated from biomedical knowledge graphs like the ones presented in the preceding chapter (Chapter 2). These knowledge graphs are usually assembled from data available in scientific publications and experimental data deposited on various open-source, and sometimes proprietary, database platforms. However, despite the wealth of information available, one valuable resource has remained largely underutilized: patent literature. This is primarily due to the complexity of legal language, which makes it challenging to mine and integrate patent data into existing knowledge frameworks.

In the previous section, we provided a brief overview of SureChEBML, a public patent database, and understood the significance of extracting valuable information from its chemical universe. Despite serving as an extensive collection of life science patent literature, SureChEMBL relies on manual search and retrieval methods for patent documents related to a specific compound. This reliance on manual efforts may be attributed to lower level of user engagement, possibly influenced by the availability of competitive commercial patent vendors such as CAS SciFinder[2], iamIP[3], Derwent[4] and others (see Section 1.3.3). In the publication, *PEMT: a patent enrichment tool for drug discovery*, we developed the Patent EnrichMent Tool (PEMT), an open-source Python software designed to assist researchers in drug discovery. This tool facilitates the extraction of pharmaceutically relevant patent literature related to genes or proteins of interest, utilizing compound and/or biological modulators. The pharmaceutical significance of the patent documents is determined by the International Patent Classification (IPC) classification (see Section 1.3.2).

For a given indication, PEMT can extract patent documents in either a target-agnostic or a compound-agnostic approach (Figure 10). In the target-agnostic approach, where genes associated with the specific indication are pre-known, PEMT assists in the extraction of *in vivo* validated compounds and biological modulators that directly regulate these genes. Subsequently, the tool establishes connections between the modulators and patent documents by systematically querying SureChEMBL. During this stage, pre-

[2] `https://www.cas.org/solutions/cas-scifinder-discovery-platform/cas-scifinder-n`

[3] `https://iamip.com/`

[4] `https://clarivate.com/products/ip-intelligence/ip-data-and-apis/derwent-world-patents-index/`

filtering criteria are applied to narrow down “active” patent documents falling under 34 IPC classes covering aspects such as compound synthesis, formulation, compound-gene modulation, compound-gene-disease modulations and gene-disease modulations. In the compound-agnostic approach, where the interaction between the compound and the disease is known, PEMT facilitates the extraction of patent documents related to these compounds, following the same rigorous methodology outlined earlier. This dual approach provides researchers with a comprehensive and flexible tool for patent literature extraction in the context of drug discovery.

To illustrate the utility of PEMT, we extracted patent documents focused on rare diseases. Among the 56 genes selected in Orphanet database, we found 133 patent documents, of which 85 belonged to pharmaceutical industries, 23 to academic institutes and 25 patent documents to individuals. Beyond merely collecting a patent corpus, we used PEMT to explore the patent landscape associated with specific genes. We discovered instances where the patent attention span of a gene correlated with the indication areas where the gene was identified as a biomarker. By applying the same methodology to the remaining targets, our study offers insights into gene target prioritization, showcasing its evolution throughout the research period through a thorough analysis of the associated patent landscape for each gene.

With the development and open-sourcing of the PEMT tool, we introduce automation in patent collection, streamlining a traditionally manual and time-intensive process. As the first of its kind, this Python-based tool aids experts in efficiently searching for compounds of interest, overcoming the limitations of conventional manual searches. Such tools are crucial for assembling patent cohorts, which play a key role in downstream research and decision-making. Historically, commercial platforms such as Thomson Reuters’ Merged Markush Service (MMS), CAS’s MARPAT, and Questel’s MMS have dominated the industry by providing robust compound structure-based patent searches. However, with the shift toward open science, tools like PEMT and Chemicalstripes are gaining traction. Aurich *et al.* introduced Chemicalstripes[5], an R-based package that facilitates the exploration of compound patents sourced from PubChem [95]. This package takes PubChem Compound Identifiers (CIDs) and a specified patent date range as inputs, generating visual representations of the compound patent landscape. The visualization employs colored stripes, where the colors correspond to patent counts [96]. Using this tool, the authors analyzed the emergence of chemicals in the

[5] `https://gitlab.lcsb.uni.lu/eci/chemicalstripes`

environmental sector, with a particular focus on pollutants and agrochemicals. Until now, patent analysis tools have been either commercially restrictive or reliant on labor-intensive manual efforts. The advent of automated and semi-automated solutions like PEMT marks a significant shift, enabling faster, more efficient data aggregation and improving decision-making processes in patent research.

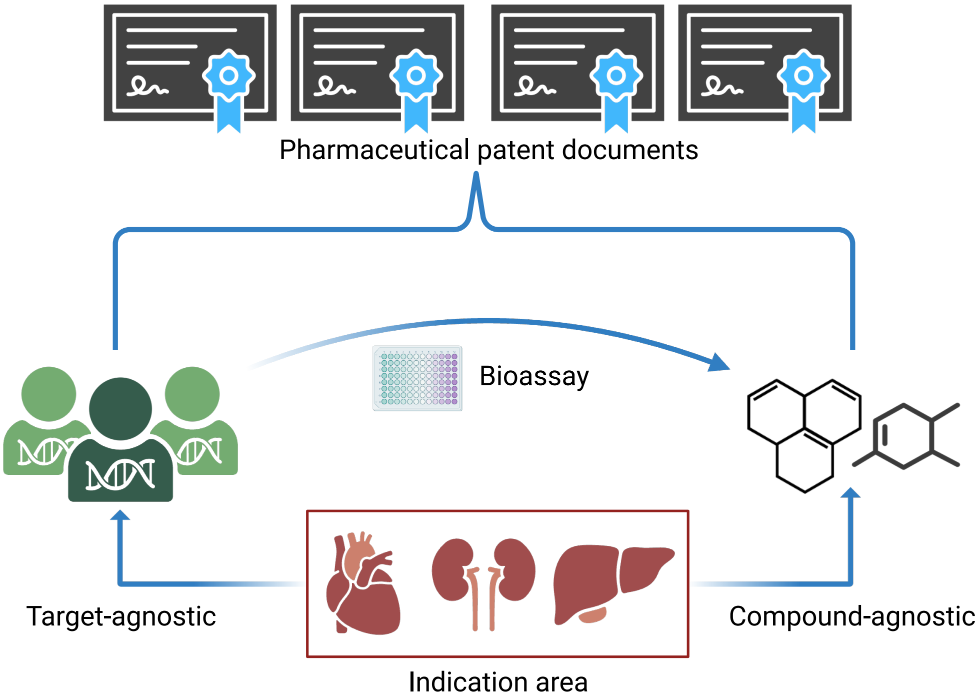


Figure 10: Graphical summary of paper. Created with BioRender.com

CHAPTER 4

Leveraging knowledge graphs for decision making

In the previous chapters, we introduced data warehousing approaches that consolidate data into structured information and knowledge, creating a centralized database. However, an equally important aspect to explore is the practical application and impact of this aggregation in day-to-day research. Just as a tool can serve multiple purposes, the methodologies discussed earlier can be utilized in various ways to enhance decision-making in drug discovery and development. This chapter delves into alternative approaches for interrogating these aggregated datasets, demonstrating how they can be leveraged to generate meaningful hypotheses for drug discovery. Additionally, we discuss the experimental validation of *in-silico* generated hypotheses in an *in-vitro* setting. We begin by illustrating the value of integrating biological patent data into KGs for comprehensive landscape analysis, particularly from an industrial perspective. Next, we examine how the contextual depth within KGs can facilitate the identification and validation of promising drug candidates in early-stage discovery. Finally, we present a predictive algorithm designed to infer cellular-level modulations using transcriptomic data integrated into the KG, aiding in the identification of Mechanisms of Action (MoA).

4.1 Pharmaceutical patent landscaping: A novel approach to understand patents from the drug discovery perspective.

This section presents the following publication **(see Appendix A.5)**:

> **Yojana Gadiya**, Philip Gribbon, Martin Hofmann-Apitius and Andrea Zaliani. 2023. Pharmaceutical patent landscaping: A novel approach to understand patents from the drug discovery perspective. *Artificial Intelligence in the Life Sciences*, 3, 100069.
>
> `https://doi.org/10.1016/j.ailsci.2023.100069`

Authors' contributions

Yojana Gadiya: Project administration, Formal analysis and Writing – original draft. *Philip Gribbon*: Writing – review and editing. *Martin Hofmann-Apitius*: Writing – review and editing. *Andrea Zaliani*: Project administration, Formal analysis, Writing – review and editing, and Writing – original draft.

Summary

Patent literature is an underutilized resource in drug discovery, often seen primarily as a means to access or reinforce drug monopolies. Within this scenario, a prominent approach employed by pharmaceutical companies is known as patent landscaping, which involves a thorough examination of patent documents to discern underlying data patterns. Landscaping plays a non-secondary role in understanding competitor strategy in pharma discovery. This landscaping approach serves a pivotal role in deciphering competitor strategies within the pharmaceutical discovery landscape. Moreover, patent literature serves as a rich source of scientific insights that may not always be captured in traditional scientific publications. Consequently, there is a lack of open-source tools available to academics and researchers for analyzing patent documents. To overcome this limitation, in our previous endeavours, we delved into the public patent data repository (Section 3.1) and developed a tool called PEMT (Section 3.2) to facilitate the extraction of patent-relevant data within specific contexts. In this section, we leverage this tool to investigate pharmaceutical patent landscaping techniques for an indication area.

Pharmaceutical patenting activity, which primarily covers claims related to therapeutic design, chemical synthesis, formulation, and other key aspects, offers valuable insights into drug development and prescribing trends [71, 97, 98]. The patent documents corresponding to these activities cover two fundamental components: the compound itself and its application. The compound is identified by its structure image, depicted in the form of a Markush scaffold, or through its name which could be either its trade or generic name. Meanwhile, the application field is usually described in the text found within the claims or description sections of the patent document. Both these elements in combination play a crucial role in defining the scope of the underlying patent, making it a subject of study for patent lawyers and pharmaceutical research scientists who aim to comprehend the freedom-to-operate (FTO) space associated with the patent documents.

In the publication, *Pharmaceutical patent landscaping: A novel approach to understand patent data from the drug discovery perspective*, we use our open-source patent extractor PEMT and systematically analyze patent documents to identify patterns within patent owners, targets and small molecules around which the patent documents are formulated (Figure 11). We focus our applications on two indication areas, rare diseases and Alzheimer's disease, by leveraging pre-existing mechanistic information aggregated from publications.

From the genes that are involved in these diseases, we formulate a patent corpus that serves as the basis for landscaping. Throughout the paper, we identified and assessed case scenarios that allow in understanding of the patterns within patent documents.

Commencing with an exploration of genes associated with Alzheimer's and Rare diseases using publicly available knowledge graphs Neurommsig[1] and OrphaNet[2], respectively, we compiled a repository of compounds and patent documents. This effort yielded a corpus of 23,000 compounds sourced from 14,000 patents. Analysis of these patent documents unveiled recurring patterns, such as the prevalence of patent applications across specific compound classes and a notable discrepancy between filed and granted patent applications (for definition, refer to Section 1.3.1). Moreover, a temporal examination of patent data provided insights into the indication landscape over the past decade, offering a glimpse into the target interests of major pharmaceutical players and their patenting activities in these areas. Beyond competitive analysis, these insights can inform target prioritization strategies in research endeavors. For instance, targets demonstrating sustained interest in the patent landscape could be prioritized, whereas those exhibiting waning interest may warrant reevaluation due to potential toxicity concerns evidenced by higher clinical trial failure rates. Ultimately, this study underscores the utility of PEMT in informing scientific decisions within the realm of early drug discovery.

Patent landscaping, when applied using tools and resources like the one mentioned above, can benefit diverse stakeholders, from attorneys and portfolio managers to researchers and stakeholders. Large organizations either have dedicated in-house asset management teams (comprising legal, technical, business, and marketing experts) that conduct patent landscaping for new developments or assets or outsource to professionals. The outcomes of such landscaping can reveal trends that indicate the future potential of technologies or fields, aiding in strategic planning and informed R&D decisions. For instance, Liu *et al.*, among other researchers, analyzed the patent landscape across multiple coronavirus species to assess technological investments and advancements, ultimately guiding future efforts in combating pandemics by leveraging insights from past research and development [99, 100]. In the field of drug repurposing, which investigates existing compounds for new therapeutic applications, Murke analyzed cancer-related patents from 2014 to 2019. His study identified Hippel-Lindau syndrome, a rare cancer, as the only

[1] `https://neurommsig.scai.fraunhofer.de/`

[2] `https://www.orpha.net/`

rare disease with potential repurposing opportunities based solely on patent data [101]. Furthermore, his findings indicated that neuropsychiatry could be a promising area for future drug repurposing efforts. Patent landscaping thus play a critical role in tracking the growth and impact of various research fields while providing insights into future directions.

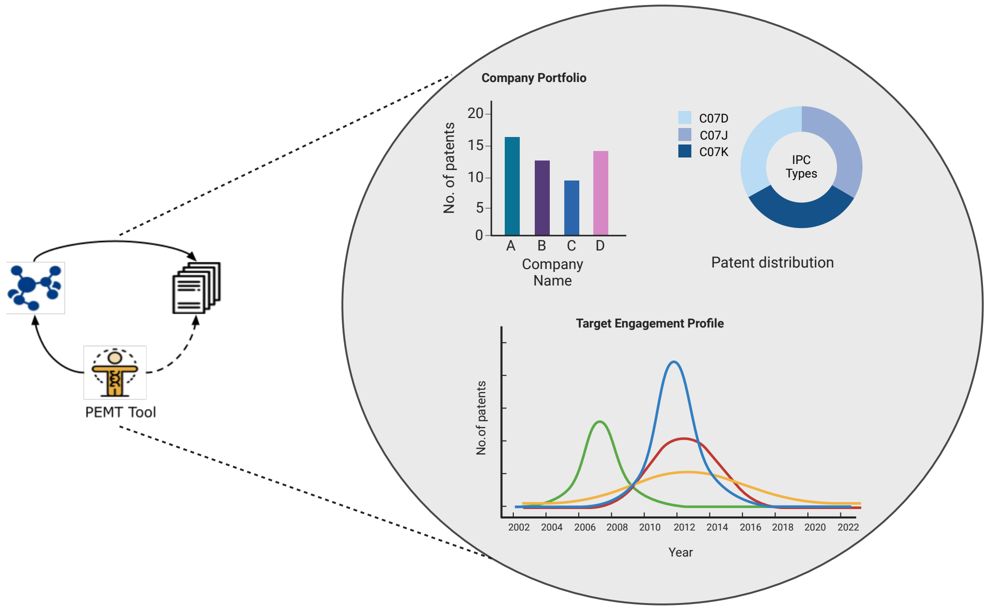


Figure 11: Graphical summary of paper. Created with BioRender.com

4.2 Causal reasoning over knowledge graphs leveraging drug-perturbed and disease-specific transcriptomic signatures for drug discovery

This section presents the following publication **(see Appendix A.6)**:

> Daniel Domingo-Fernández, **Yojana Gadiya**, Abhishek Patel, Sarah Mubeen, Daniel Rivas-Barragan, Chris W. Diana, Biswapriya B. Misra, David Healey, Joe Rokicki and Viswa Colluru. 2022. Causal reasoning over knowledge graphs leveraging drug-perturbed and disease-specific transcriptomic signatures for drug discovery. *PLoS computational biology*, 18(2), e1009909.
>
> `https://doi.org/10.1371/journal.pcbi.1009909`

Authors' contributions

Daniel Domingo-Fernández: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft and Writing – review and editing. *Yojana Gadiya*: Formal analysis, Investigation, Methodology, Software, Validation, and Visualization. *Abhishek Patel*: Formal analysis and Investigation. *Sarah Mubeen*: Data curation, Formal analysis, Writing – original draft and Writing – review and editing. *Daniel Rivas-Barragan*: Data curation, Software and Writing – review and editing. *Chris W. Diana*: Project administration, Resources, Supervision and Visualization. *Biswapriya B. Misra*: Investigation, Project administration, Validation, Writing – original draft and Writing – review and editing. *David Healey*: Funding acquisition, Project administration, Supervision, Writing – original draft, and Writing – review and editing. *Joe Rokicki*: Funding acquisition, Methodology, Project administration and Supervision. *Viswa Colluru*: Funding acquisition, Project administration, Resources, Supervision and Writing – review and editing.

Summary

As outlined in an earlier chapter, biological knowledge graphs provide a powerful framework for integrating and deciphering causal relationships among biological entities. However, since these graphs are primarily constructed from scientific literature, they are inherently susceptible to research bias, which can impact their interpretation. For instance, well-studied biomarker proteins or genes may appear extensively linked to diseases, whereas lesser-studied genes may be significantly underrepresented. This bias presents challenges, particularly when using knowledge graphs for target prioritization or drug mechanism of action (MoA) elucidation, as they often lack context-specific validation. Additionally, given the complexity of biological systems, critical insights may be missing altogether from these resources. To address these limitations, a promising approach is to complement literature-derived knowledge graphs with experimentally validated data, particularly through the integration of *omics* datasets.

In the publication, *Causal reasoning over knowledge graphs leveraging drug-perturbed and disease-specific transcriptomic signatures for drug discovery*, we introduce RPath, an algorithm designed to integrate transcriptomic data with existing knowledge graphs in order to identify novel drug-target mechanisms for an indication area (Figure 12). While several algorithms exist for integrating networks with *omics* data [102–104], none of them have been extensively applied in the field of drug discovery until now. Our framework utilizes the RPath algorithm, which reasons over paths in a graph to identify all possible effects of a given drug on a disease, taking into account both causal relations between the drug and disease and their associated transcriptomic signature. Through RPath, we can explore, predict and recommend clinically relevant drugs for targeted diseases. Given the effectiveness of this approach, this section explores the prospect of leveraging RPath for elucidating the MoA of a drug and predicting novel drug-disease pairs.

The RPath algorithm is initiated by identifying all feasible causal paths connecting a drug to the disease via a protein in the graph. For each of these paths, it integrates drug-induced transcriptomic data with the primary objective to identify metapaths in the graph that align with the experimental findings. This entails ensuring that the effect of a causal edge (e.g., inhibition or activation) between a drug and its target is consistent with the transcriptomic profile of the target influenced by the drug. Subsequently, for these “concordant” metapaths, RPath validates the impact of the target

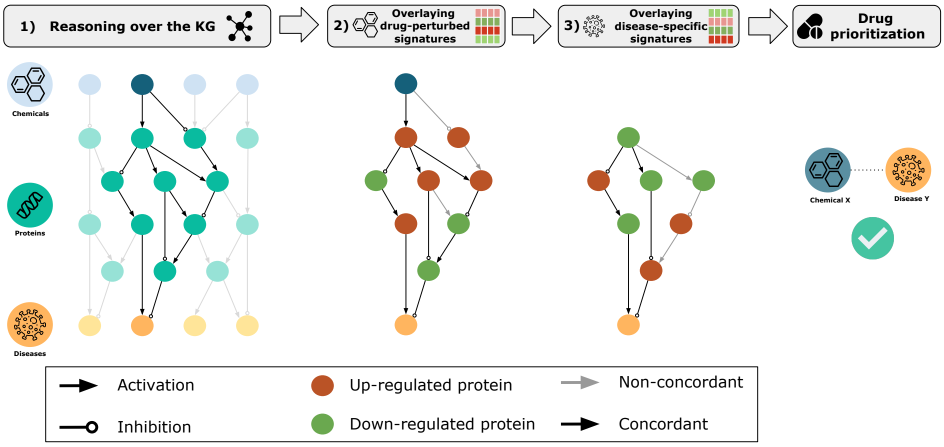


Figure 12: Graphical summary of paper.

on the disease by leveraging disease-perturbed transcriptomic data. In this context, rather than seeking concordant patterns, we seek "non-concordant" ones. This approach ensures that the proteins are targeted by the drug and that the genes play a role in disease dysregulation.

To demonstrate the utility of RPath, we leveraged open-source benchmark knowledge graphs such as OpenBioLink, along with transcriptomic data from the L1000[3] and GEO[4] dataset. Using these resources, we conducted a benchmarking study comparing RPath against 11 classical methods commonly used for drug-disease prediction. RPath consistently outperformed these approaches, most of which rely solely on network proximity information, such as nearest neighbors and the underlying graph structure. Furthermore, we showcased RPath's ability to prioritize drugs based on biologically relevant MoA, using the cancer drug candidates ponatinib and bicalutamide as case studies. By integrating biological context into drug prioritization, RPath enhances the reliability of insights, mitigates biases, and enables a more data-driven understanding of disease mechanisms and therapeutic targets.

[3] https://clue.io/
[4] https://www.ncbi.nlm.nih.gov/geo/

4.3 Predicting antimicrobial class specificity of small molecules using machine learning

This section presents the following publication **(see Appendix A.7)**:

> **Yojana Gadiya**, Olga Genilloud, Ursula Bilitewski, Mark Brönstrup, Lenoie von Berlin, Marie Attwood, Philip Gribbon and Andrea Zaliani. 2024. Predicting antimicrobial class specificity of small molecules using machine learning. *Journal of Chemical Information and Modeling.* 65 (5), pp.2416-2431
>
> `https://doi.org/10.1021/acs.jcim.4c02347`

Authors' contributions

Yojana Gadiya: Conceptualization, Methodology, Writing - Original draft preparation and Visualization. *Andrea Zaliani*: Conceptualization, Supervision, Methodology, Writing - Original draft preparation and Visualization. *Lenoie von Berlin*: Visualization. *Olga Genilloud*: Investigation. *Ursula Bilitewski*: Investigation. *Mark Brönstrup*: Investigation. *Marie Attwood*: Writing - Reviewing and Editing. Philip Gribbon: Writing - Reviewing and Editing.

Summary

Up to this point, we have explored strategies for data aggregation (Chapters 2 and 3), enabling on-the-fly collection of structured, machine-ready data. Additionally, we have discussed methods to complement data analysis with transcriptomic insights (Section 4.2). However, these efforts have primarily focused on *in silico* tools and computational approaches, with limited transition toward real-world laboratory validation. Now, we bridge this gap. In this section, we showcase a groundbreaking example of how *in silico* findings can translate into experimental settings, demonstrating their real-world applicability. This final publication is a pivotal step—moving beyond theoretical modeling to tangible scientific outcomes—where computational predictions meet laboratory validation, reinforcing the impact of data-driven discoveries in a biological context.

Machine learning (ML) has emerged as a powerful tool for accelerating and optimizing drug discovery, particularly in addressing critical challenges such as drug-resistant pathogens [105–107]. While thousands of open-source ML models, including widely recognized ones like ChatGPT[5] and DeepSeek[6], have been developed and democratized, their application in drug discovery remains a focal point of research. The advancement of explainable AI (XAI) has further facilitated the integration of ML models into decision-making processes, ensuring transparency and interoperability. Here we highlight the transformative role of ML in combating antibiotic resistance (AMR), showcasing how these models can be strategically applied to accelerate the discovery of novel antibiotics. By harnessing data-driven insights, ML-driven approaches offer a promising avenue for tackling one of the most urgent global health threats, paving the way for innovative and effective therapeutic solutions.

In the publication, *Predicting Antimicrobial Class Specificity of Small Molecules Using Machine Learning*, we systematically aggregated a comprehensive dataset of *in vitro* tested compounds across various microbial strains (Figure 13). These microbial strains were categorized into four major classes: Gram-positive (GP) bacteria, Gram-negative (GN) bacteria, *Mycobacterium tuberculosis* (Acid-Fast bacteria), and fungi. This effort led to the development of the Antimicrobial-KG (`https://antimicrobial-kg.serve.scilifelab.se/`), a knowledge graph that integrates data from three bioassay-specific

[5] `https://chatgpt.com/`

[6] `https://www.deepseek.com/`

repositories - ChEMBL, CO-ADD, and SPARK. A key consideration in this aggregation process was ensuring compliance with the FAIR principles: Findability (F), Accessibility (A), Interoperability (I) and Reusability (R), which define the quality and usability of scientific data [108]. In summary, the Antimicrobial-KG collected information on 81,490 chemicals tested across 1,373 bacterial species.

Next, we focused on developing XAI models using the data collected in the Antimicrobial-KG. Our goal was to predict the activity of compounds across the four bacterial classes mentioned earlier. For model development, we employed a two-step funnel approach. First, we trained a cohort of ML models (Random Forest, XGBoost, Linear Regression, Naïve Bayes, Gradient Boosting Tree, and Decision Tree) to identify the best-performing model. Once identified, we fine-tuned this model through hyperparameter optimization. Alongside model selection, we evaluated the predictive performance across multiple chemical representations of the compounds in the form of molecular fingerprints. These fingerprints capture key structural features, helping to identify molecular properties that contribute to antimicrobial activity. Our analysis identified Random Forest with the MHFP6 fingerprint as the optimal model, achieving an accuracy of 75.9% and a Cohen's Kappa score of 0.68. Despite the inherent explainability of XAI models, molecular fingerprints, being bit-vector representations of chemical structures, can sometimes obscure the interpretability of results. To enhance transparency and facilitate future chemical optimization, we incorporated physicochemical properties (such as LogP, polar surface area, hydrogen bond acceptor, hydrogen bond donor etc.) as an additional fingerprinting method. This integration provided a more intuitive understanding of compound activity, complementing the predictive capabilities of the XAI model.

We applied our model to two compound libraries, EU-OPENSCREEN[7] (EU-OS) and the Enamine Antibacterial Library[8], to assess its effectiveness in predicting active compounds against GP bacteria, GN bacteria, and fungal pathogens. For the Enamine Antibacterial Library, which contains 32,000 compounds, our model predicted only 10% to be active for antimicrobial purposes. Among these, 69% were predicted to be active against Gram-positive bacteria, 22% against Acid-Fast bacteria, 6% against fungi, and 3% against Gram-negative bacteria. With the EU-OS library, we adopted a different approach wherein we evaluated the impact of ML-based predictions.

[7] `https://www.eu-openscreen.eu/`

[8] `https://enamine.net/compound-libraries/targeted-libraries/antibacterial-library`

We screened the entire library of approximately 100,000 compounds *in vitro* against seven microbial strains: *Candida auris* DSM21092, *Staphylococcus aureus* ATCC 29213, *Pseudomonas aeruginosa*, *Candida albicans* ATCC 64124, *Enterococcus faecalis* ATCC 29212, *Aspergillus fumigatus* ATCC 46645, and *Escherichia coli* ATCC 25922. Simultaneously, we applied our model to predict compound activity. Notably, our model successfully predicted over 30% of the active compounds that were experimentally validated through full-library screenings. Additionally, it identified microbial strains, such as *Pseudomonas aeruginosa* and *Escherichia coli*, where no activity was observed, demonstrating its potential to significantly reduce the financial costs associated with high-throughput screening (HTS). These findings emphasize the real-world applicability of our models, not only enabling *in-silico* predictions but also validating their effectiveness through *in-vitro* efficacy testing.

With the rise of GPT-based tools like BioGPT[9], the research community is expected to see an influx of machine-generated scientific hypotheses. However, validating these hypotheses in experimental settings will be crucial to differentiate genuine scientific insights from AI-generated hallucinations. By developing the Antimicrobial-KG and ML model, we have demonstrated a solution that integrates both *in-silico* predictions and *in-vitro* validation to guide future hypothesis generation. Beyond experimental validation, our work has also identified key physicochemical properties relevant to antimicrobial activity for different microbial classes (i.e. Gram-positive, Gram-negative, Acid-fast and Fungi). These chemical characteristics would guide drug optimization in antimicrobial drug discovery (ADD) projects. In conclusion, our findings present a promising approach for filtering out non-antimicrobial scaffolds from public and commercial screening collections, thus providing an efficient and cost-effective strategy to support AMR research and foster collaboration in drug discovery.

[9] `https://biologpt.com/`

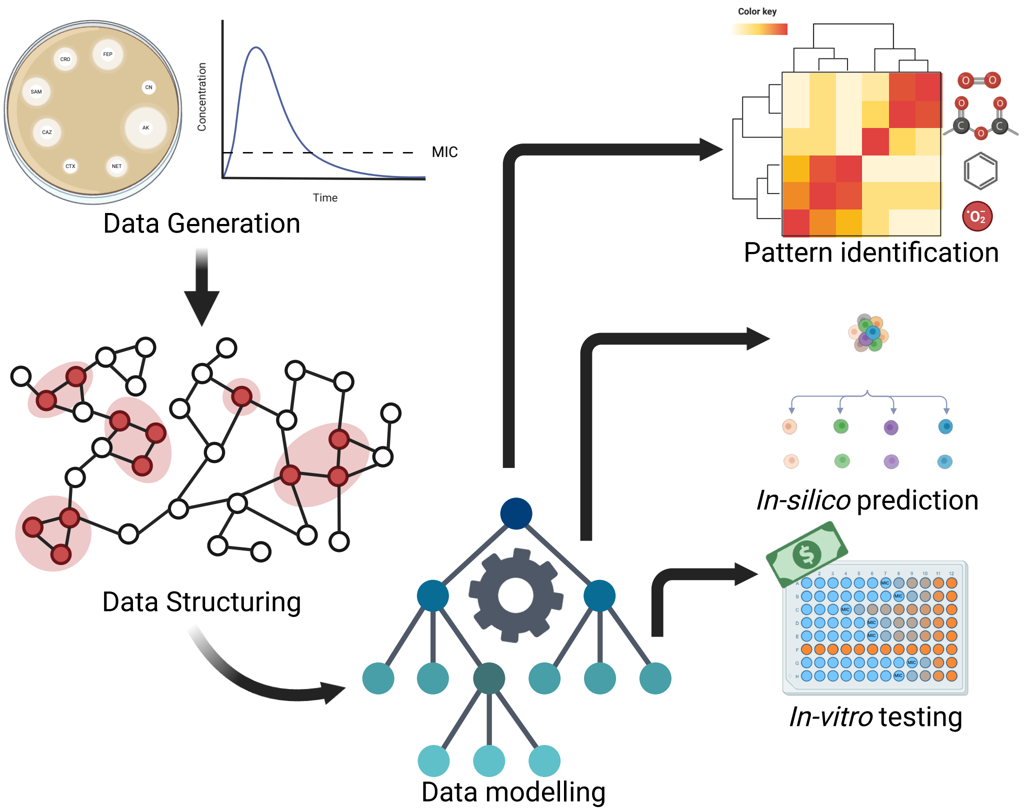


Figure 13: Graphical summary of paper. Created in BioRender.

CHAPTER 5

Conclusion and outlook

In recent years, rapid technological advancements have led to an unprecedented surge in biological and chemical data within drug discovery. While this wealth of information holds immense potential, the challenge lies in effectively harnessing its insights. Extracting meaningful knowledge requires sophisticated analytical tools capable of integrating and interpreting data from diverse sources. However, beyond technological capabilities, a comprehensive understanding of the multidimensional aspects of the research question spanning biology, chemistry, and related disciplines is essential.

To address this, there is a growing need for develop novel approaches that enable the intelligent and holistic interpretation of complex datasets. These methodologies must not only extract accurate insights but also ensure their relevance and applicability to real-world challenges. Moreover, the slow pace of updates to existing resources has highlighted the opportunity for dynamic, on-the-fly data graphs and knowledge generation. By embracing a systems-level perspective and leveraging cutting-edge analytical techniques, we can unlock the full potential of these vast datasets, accelerating drug discovery and ultimately advancing healthcare outcomes.

This dissertation has aspired to construct a comprehensive framework for identifying potential drug candidates during the early stages of drug discovery (Figure 14). Our exploration commenced by elucidating classical methodologies for systematically gathering and organizing drug discovery-related data into KGs, spanning both literature and experimental domains (Chapter 2). Subsequently, we underscored the significance of incorporating a third dimension, the industrial landscape delineated by patent documents, which holds considerable sway in the selection of drug candidates (Chapter

3). Lastly, we transcended conventional representations of the compound mechanism of action (MoA) by employing graph-based and ML approaches to nominate candidate compounds, culminating in their experimental validation and thus forging a vital connection between the realms of *in-silico* and *in-vitro* experimentation (Chapter 4).

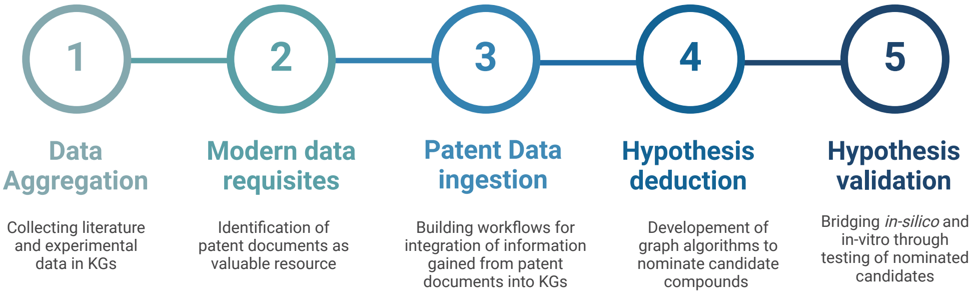

1
2
3
4
5
Data Aggregation
Collecting literature and experimental data in KGs
Modern data requisites
Identification of patent documents as valuable resource
Patent Data ingestion
Building workflows for integration of information gained from patent documents into KGs
Hypothesis deduction
Developement of graph algorithms to nominate candidate compounds
Hypothesis validation
Bridging *in-silico* and in-vitro through testing of nominated candidates

Figure 14: Graphical overview of thesis. Created with BioRender.com

This dissertation introduces a range of knowledge- and data-driven computational methodologies tailored to support early-stage drug discovery endeavours. This research sheds light on the valuable insights derived from patent documents, a resource often overlooked until the later stages of drug development. By doing so, it raises awareness of their potential significance in informing candidate selection. Moreover, the thesis showcases advanced methodologies that demonstrate how both knowledge- and data-driven approaches can synergistically complement one another, thereby streamlining the drug discovery process. Beyond methodological discussions, the integrative efforts undertaken underscore the tangible impact of these resources in hypothesis generation and validation. In summary, the mechanistic insights unearthed during early drug discovery research have the potential to expand our understanding of the underlying pathophysiology of human conditions. By revealing novel interactions and crosstalks, these insights pave the way for a more informed and strategic selection of promising drug candidates.

In Chapter 2, we presented two publications that demonstrate the creation of KGs through two distinct methodologies: literature mining and experimental studies. The literature mining (presented in Section 2.1) entails the extraction and synthesis of information from a vast array of scientific publications, enabling the construction of graphs that represent the collective knowledge and insights documented in scientific literature. On the other hand, the experimental approach (as showcased in Section 2.2) involves harnessing data derived directly from publically disseminated experimental studies, such as *in-vitro* assays or *in-vivo* experiments, to construct graphs that encapsulate empirical findings and relationships between various entities. By exploring these two complementary approaches, we aim to showcase the versatility and efficacy of leveraging existing data resources for the generation of comprehensive and informative KGs. Through the integration of experimental data and insights from literature, these knowledge graphs serve as powerful tools for advancing research, facilitating data-driven discoveries, and enhancing our understanding of complex biological systems and phenomena.

Then, in Chapter 3, we explore a crucial component that has been missing from existing KGs but holds immense potential to enhance their context and content, particularly within the domain of drug discovery: patent documents. These documents serve as a rich source of information, encapsulating valuable insights into novel compounds, formulations, therapeutic targets and innovative approaches in drug development. By integrating patent data into KGs, we aim to bridge the gap between scientific knowledge and real-world applications, providing a comprehensive understanding of the drug discovery landscape. Through our examination of patent documents, we seek to uncover hidden connections, identify emerging trends and/or gaps, and unlock new opportunities for innovation and collaboration within the pharmaceutical industry. By addressing this missing piece, we strive to enrich KGs with invaluable insights from the realm of intellectual property, empowering researchers and stakeholders with a holistic view of the drug discovery market.

In Section 3.1, we focus on a public patent database SureChEMBL, which aggregates and annotated biomedical entities such as genes, compounds and diseases found in patent documents. Here, we explored the utility of patent data available in SureChEMBL for drug discovery based on four key criteria: *i)* presence of patent compounds in public compound databases like ChEMBL, *ii)* the coverage assessment of the drug-like space among patent compounds through criteria like Lipinski's Rule of five (Ro5), *iii)* the evaluation of chemical diversity represented by Murcko scaffolds found in patent compounds and *iv)* the prospective tracing of patent compounds through clinical trials to market approval as drugs. Recognizing the value of patent data within SureChEMBL for advancing drug discovery, we developed the Patent EnrichMent Tool (PEMT) in Section 3.2. This tool enables systematic querying and extraction of patent documents related to target genes and compounds, thereby providing researchers with valuable insights to inform their drug discovery efforts. Through these initiatives, we aim to empower the scientific community with enhanced access to patent-derived knowledge, facilitating innovation and progress in the field of drug discovery.

Finally, in Chapter 4, we culminate our exploration by delving into three publications that discuss potential mechanisms for downstream utility of graphs. In Section 4.1, we demonstrated the effectiveness of analyzing patent documents as a means to gain insights into the prevailing market interest in drug discovery. Through landscaping analysis, we identified target and disease portfolios of organizations, shedding light on the competitive landscape within the industry. Additionally, we unraveled the evolving importance of targets and diseases over different years, providing valuable temporal perspectives on research trends and priorities. Next, in Section 4.2, we presented RPath, a novel graph algorithm which integrates transcriptomic data with KGs for the purpose of selecting compound-disease pairs with therapeutic potential. Our hypothesis centered on the effectiveness of a network-based strategy, which amalgamates paths linking drug and disease-related nodes within a KG with gene expression data, surpassing the capabilities of conventional graph-based methods relying on network proximity measures, such as shortest paths. By validating our approach with drug-disease pairs from clinical trials, we confirmed the efficacy of RPath in prioritizing clinically investigated pairs. Notably, our analysis revealed a significant proportion of clinically relevant pairs among those highlighted by RPath. Furthermore, our findings extended into the field of oncology, where RPath successfully identified novel drug-disease pairs. By integrating transcriptomic data with KGs, RPath enhances the prospect of identifying promising therapeutic avenues, ultimately contributing to the advancement of precision medicine and the development of novel treatments.

Lastly, in Section 4.3, we introduced an ML model built on Antimicrobial-KG, a KG aggregated from experimental *in-vitro* screening assays, to predict the antimicrobial activity of compounds. The experimental validations of *in-silico* compound activity conducted in this publication revealed the ability to transpire results from *in-silico* to *in-vitro* or *in-vivo* level, in addition to showcasing the impact and applicability of the model antimicrobial drug discovery world. By leveraging the ML model to analyze structural and functional relationships among compounds and their activity, this XAI approach streamlines and assist the compound selection process, leading to more efficient and targeted drug discovery efforts. In summary, the introduction of the ML model to the antimicrobial community represents a significant advancement in computational drug discovery methodologies, offering a more financially informed approach to AMR.

This thesis contributes to the field of drug discovery in numerous ways. First, by the establishment of “reusable” workflows for KG generation, the work has demonstrated the capability of building graphs on-the-fly. Second, the development of one of the first open-source patent mining tool, PEMT, designed to seamlessly integrate with existing KG workflows,has facilitates patent data enrichment with ease. Third, with development of algorithms like RPath, the work provides basis for enabling data driven hypothesis generation. And finally, the thesis has successfully demonstrated the bridging of *in-silico* to *in-vitro* worlds through ML-based predictions and subsequent experimental validation showcasing the importance of cross-talk between both the worlds in drug discovery. Moreover, a key emphasis of this thesis is on adhering to the principles of Findability, Accessibility, Interoperability and Reusability (FAIR). All resources, including software, analytical scripts, and pipelines, have been meticulously packaged and tested in accordance with established software development standards endorsed by the scientific community. These resources are made publicly available through open-source platforms such as GitHub, Zenodo, or PyPi, ensuring widespread accessibility and usability.

Future outlook

Recent advances in the drug discovery field are progressively showing a significant surge in the interest in the development of KG across various domains, including scientific and non-scientific fields. This growing interest has led to the creation of numerous automated workflows designed to scale up the generation of KG, moving beyond the traditional methods that relied heavily on manual curation and/or semi-automated techniques. These automated workflows have the potential to vastly increase the efficiency and scope of KG creation, enabling the handling of larger datasets and more complex relationships. However, a major challenge in automated methods for KG generation is the often low emphasis on ensuring the quality of data within these graphs. Automated systems can rapidly process and integrate vast amounts of information, but they may also propagate errors, inconsistencies, and biases present in the source data. For instance, several cases of AI-based hallucination, where AI generates plausible but nonsensical statements, have been reported [109–111]. These inaccuracies can reduce the reliability of the knowledge graph built from such GPT-based models, posing significant challenges for researchers. Such gaps in knowledge hold ramifications for our ability to piece together the mechanisms governing the biological processes we seek to investigate.

Within the context of our work, incorporating pharmaceutical patent-based information, a pioneering effort, can enhance our understanding of the commercial and research landscape of compounds and genes. Such interlinked knowledge graphs can delineate "freedom-to-operate" areas for drug repositioning or the development of novel drugs within a unexplored chemical space. This approach ensures that researchers can navigate intellectual property landscapes more effectively, identifying opportunities and potential barriers. Furthermore, integrating experimental results with literature-based information can provide a more comprehensive understanding of data provenance. This synergy allows for the creation of KGs that reflect real-world scenarios and the complex biology of organisms, thereby improving the reliability and applicability of the findings. We anticipate that future work will build upon these foundations by developing automated AI-based methods that contextualize and summarize patent, experimental, and research-based data together. These advancements will streamline the synthesis of vast amounts of information, facilitating more informed decision-making in drug discovery and development. Additionally, the ability to quantify and distinguish ground truth from AI-generated hallucinations will be crucial in the future of ML/AI.

In drug discovery, the successful transition of *in-silico* predictions to *in-vitro* and ultimately *in-vivo* experiments could mark a significant breakthrough, accelerating the path from computational insights to tangible therapeutic advancements.

In summary, while the advancement of automated workflows for knowledge graph generation marks a significant step forward in various fields, it is imperative to address the data quality challenges that come with these automated methods. Ensuring the accuracy and reliability of the information in knowledge graphs will be key to their successful application and the realization of their full potential in driving innovation and discovery.

References

1. Berdigaliyev, N. & Aljofan, M. An overview of drug discovery and development. *Future medicinal chemistry* **12,** 939–947 (2020).
2. Sadybekov, A. V. & Katritch, V. Computational approaches streamlining drug discovery. *Nature* **616,** 673–685 (2023).
3. Drews, J. Drug discovery: a historical perspective. *science* **287,** 1960–1964 (2000).
4. Hughes, J. P., Rees, S., Kalindjian, S. B. & Philpott, K. L. Principles of early drug discovery. *British journal of pharmacology* **162,** 1239–1249 (2011).
5. Grabley, S. & Thiericke, R. *Drug discovery from nature* (Springer Science & Business Media, 1998).
6. Al-Ali, H. The evolution of drug discovery: from phenotypes to targets, and back. *MedChemComm* **7,** 788–798 (2016).
7. Sliwoski, G., Kothiwale, S., Meiler, J. & Lowe, E. W. Computational methods in drug discovery. *Pharmacological reviews* **66,** 334–395 (2014).
8. Pasrija, P., Jha, P., Upadhyaya, P., Khan, M., Chopra, M., *et al.* Machine learning and artificial intelligence: a paradigm shift in big data-driven drug design and discovery. *Current Topics in Medicinal Chemistry* **22,** 1692–1727 (2022).
9. Minnich, A. J. *et al.* AMPL: a data-driven modeling pipeline for drug discovery. *Journal of chemical information and modeling* **60,** 1955–1968 (2020).
10. Takebe, T., Imai, R. & Ono, S. The current status of drug discovery and development as originated in United States academia: the influence of industrial and academic collaboration on drug discovery and development. *Clinical and translational science* **11,** 597–606 (2018).

11. Canady, M. From outsourced to open: the continuing evolution of the drug discovery business model. *Drug Discov.,* 9 (2012).

12. Patwardhan, B., Vaidya, A. D. & Chorghade, M. Ayurveda and natural products drug discovery. *Current science,* 789–799 (2004).

13. Li, J. W.-H. & Vederas, J. C. Drug discovery and natural products: end of an era or an endless frontier? *Science* **325,** 161–165 (2009).

14. Newman, D. J., Cragg, G. M. & Snader, K. M. The influence of natural products upon drug discovery. *Natural product reports* **17,** 215–234 (2000).

15. Butler, M. S. The role of natural product chemistry in drug discovery. *Journal of natural products* **67,** 2141–2153 (2004).

16. Newman, D. J. & Cragg, G. M. Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019. *Journal of natural products* **83,** 770–803 (2020).

17. Singh, N. *et al.* Drug discovery and development: introduction to the general public and patient groups. *Frontiers in Drug Discovery* **3,** 1201419 (2023).

18. Fougner, C., Cannon, J., Smith, J., Leclerc, O., *et al.* Herding in the drug development pipeline. *Nature reviews. Drug Discovery* (2023).

19. Hefti, F. F. Requirements for a lead compound to become a clinical candidate. *BMC neuroscience* **9,** S7 (2008).

20. Gashaw, I., Ellinghaus, P., Sommer, A. & Asadullah, K. What makes a good drug target? *Drug discovery today* **16,** 1037–1043 (2011).

21. Masarone, S. *et al.* Advancing predictive toxicology: overcoming hurdles and shaping the future. *Digital Discovery.* `https://doi.org/10.1039/D4DD00257A` (2025).

22. Martis, E., Radhakrishnan, R. & Badve, R. High-throughput screening: the hits and leads of drug discovery-an overview. *Journal of Applied Pharmaceutical Science,* 02–10 (2011).

23. Frearson, J. A. & Collie, I. T. HTS and hit finding in academia–from chemical genomics to drug discovery. *Drug discovery today* **14,** 1150–1158 (2009).

24. Singh, G. in *Pharmaceutical Medicine and Translational Clinical Research* 47–63 (Elsevier, 2018).

25. Kandi, V. & Vadakedath, S. Clinical trials and clinical research: a comprehensive review. *Cureus* **15** (2023).

26. Piantadosi, S. *Clinical trials: a methodologic perspective* (John Wiley & Sons, 2024).

27. Tamimi, N. A. & Ellis, P. Drug development: from concept to marketing! *Nephron Clinical Practice* **113,** c125–c131 (2009).

28. Lipsky, M. S. & Sharp, L. K. From idea to market: the drug approval process. *The Journal of the American Board of Family Practice* **14,** 362–367 (2001).

29. *Step 5: FDA Post-Market Drug Safety Monitoring — fda.gov* `https://www.fda.gov/patients/drug-development-process/step-5-fda-post-market-drug-safety-monitoring`. [Accessed 01-05-2024]. 2018.

30. Roberts, R. A. *et al.* Reducing attrition in drug development: smart loading preclinical safety assessment. *Drug discovery today* **19,** 341–347 (2014).

31. Morgan, P. *et al.* Impact of a five-dimensional framework on R&D productivity at AstraZeneca. *Nature reviews Drug discovery* **17,** 167–181 (2018).

32. Humbeck, L. & Koch, O. What can we learn from bioactivity data? Chemoinformatics tools and applications in chemical biology research. *ACS chemical biology* **12,** 23–35 (2017).

33. Li, Q., Cheng, T., Wang, Y. & Bryant, S. H. PubChem as a public resource for drug discovery. *Drug discovery today* **15,** 1052–1057 (2010).

34. Gaulton, A. *et al.* ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic acids research* **40,** D1100–D1107 (2012).

35. Liu, T., Lin, Y., Wen, X., Jorissen, R. N. & Gilson, M. K. BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic acids research* **35,** D198–D201 (2007).

36. Tiikkainen, P. & Franke, L. Analysis of commercial and public bioactivity databases. *Journal of chemical information and modeling* **52,** 319–326 (2012).

37. Wishart, D. S. *et al.* HMDB 5.0: the human metabolome database for 2022. *Nucleic acids research* **50,** D622–D631 (2022).

38. Haug, K. *et al.* MetaboLights: a resource evolving in response to the needs of its scientific community. *Nucleic acids research* **48,** D440–D444 (2020).

39. Kopka, J. *et al.* GMD@ CSB. DB: the Golm metabolome database. *Bioinformatics* **21,** 1635–1638 (2005).

40. Go, E. P. Database resources in metabolomics: an overview. *Journal of Neuroimmune Pharmacology* **5,** 18–30 (2010).

41. Koleti, A. *et al.* Data Portal for the Library of Integrated Network-based Cellular Signatures (LINCS) program: integrated access to diverse large-scale cellular perturbation response data. *Nucleic acids research* **46,** D558–D566 (2018).

42. Bastian, F. B. *et al.* The Bgee suite: integrated curated expression atlas and comparative transcriptomics in animals. *Nucleic acids research* **49,** D831–D847 (2021).

43. EMDB—the Electron Microscopy Data Bank. *Nucleic Acids Research* **52,** D456–D465 (2024).

44. Bray, M.-A. *et al.* A dataset of images and morphological profiles of 30 000 small-molecule treatments using the Cell Painting assay. *Gigascience* **6,** giw014 (2017).

45. Williams, E. *et al.* Image Data Resource: a bioimage data integration and publication platform. *Nature methods* **14,** 775–781 (2017).

46. Brown, N. *et al.* Big data in drug discovery. *Progress in medicinal chemistry* **57,** 277–356 (2018).

47. Samuelsson, G., Bohlin, L., *et al. Drugs of natural origin: a treatise of pharmacognosy.* **Ed. 7** (CRC Press Inc., 2017).

48. Sang, S. *et al.* SemaTyP: a knowledge graph based literature mining method for drug discovery. *BMC bioinformatics* **19,** 1–11 (2018).

49. Rossanez, A., Dos Reis, J. C., Torres, R. d. S. & de Ribaupierre, H. KGen: a knowledge graph generator from biomedical scientific literature. *BMC medical informatics and decision making* **20,** 1–24 (2020).

50. Xu, J. *et al.* Building a PubMed knowledge graph. *Scientific data* **7,** 205 (2020).

51. Robbins, R. J. Biological databases: A new scientific literature. *Publishing Research Quarterly* **10,** 3–27 (1994).

52. Landsman, D., Gentleman, R., Kelso, J. & Francis Ouellette, B. *DATABASE: a new forum for biological databases and curation* 2009.

53. Szklarczyk, D. *et al.* The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic acids research* **51,** D638–D646 (2023).

54. Ochoa, D. *et al.* Open Targets Platform: supporting systematic drug–target identification and prioritisation. *Nucleic acids research* **49,** D1302–D1310 (2021).

55. Ma, L. *et al.* Database commons: a catalog of worldwide biological databases. *Genomics, Proteomics & Bioinformatics* **21,** 1054–1058 (2023).

56. Duffy, D. J. Problems, challenges and promises: perspectives on precision medicine. *Briefings in bioinformatics* **17,** 494–504 (2016).

57. Dugger, S. A., Platt, A. & Goldstein, D. B. Drug development in the era of precision medicine. *Nature reviews Drug discovery* **17,** 183–196 (2018).

58. Adams, C. Leonhard Euler and the Seven Bridges of Königsberg. *The Mathematical Intelligencer* **33,** 18–20 (2011).

59. Kadesch, R. *Problem Solving Across the Disciplines* ISBN: 9780136541875. `https://books.google.de/books?id=EiUQAAAACAAJ` (Prentice Hall, 1997).

60. Ehrlinger, L. & Wöß, W. Towards a definition of knowledge graphs. *SEMANTiCS (Posters, Demos, SuCCESS)* **48,** 2 (2016).

61. Biswas, R. *Embedding based link prediction for knowledge graph completion* in *Proceedings of the 29th ACM international conference on information & knowledge management* (2020), 3221–3224.

62. Ali, M., Hoyt, C. T., Domingo-Fernández, D. & Lehmann, J. *Predicting Missing Links Using PyKEEN.* in *ISWC (Satellites)* (2019), 245–248.

63. Shu, D. *et al.* Knowledge Graph Large Language Model (KG-LLM) for Link Prediction. *arXiv preprint arXiv:2403.07311* (2024).

64. Ghorbanali, Z., Zare-Mirakabad, F., Salehi, N., Akbari, M. & Masoudi-Nejad, A. DrugRep-HeSiaGraph: when heterogenous siamese neural network meets knowledge graphs for drug repurposing. *BMC bioinformatics* **24,** 374 (2023).

65. Gao, Z., Ding, P. & Xu, R. KG-Predict: A knowledge graph computational framework for drug repurposing. *Journal of biomedical informatics* **132,** 104133 (2022).

66. Zhu, C. *et al.* Rdkg-115: Assisting drug repurposing and discovery for rare diseases by trimodal knowledge graph embedding. *Computers in Biology and Medicine* **164,** 107262 (2023).

67. Myklebust, E. B., Jimenez-Ruiz, E., Chen, J., Wolf, R. & Tollefsen, K. E. *Knowledge graph embedding for ecotoxicological effect prediction* in *The Semantic Web–ISWC 2019: 18th International Semantic Web Conference, Auckland, New Zealand, October 26–30, 2019, Proceedings, Part II 18* (Springer, 2019), 490–506.

68. Hao, Y., Romano, J. D. & Moore, J. H. Knowledge graph aids comprehensive explanation of drug and chemical toxicity. *CPT: Pharmacometrics & Systems Pharmacology* **12,** 1072–1079 (2023).

69. Evangelista, J. E. *et al.* Toxicology knowledge graph for structural birth defects. *Communications Medicine* **3,** 98 (2023).

70. Breit, A., Ott, S., Agibetov, A. & Samwald, M. OpenBioLink: a benchmarking framework for large-scale biomedical link prediction. *Bioinformatics* **36,** 4097–4098 (2020).

71. Zhang, H.-L. & Li, Y. The Patent Landscape of BRAF Target and KRAS Target. *Recent Patents on Anti-Cancer Drug Discovery* **18,** 495–505 (2023).

72. Babaiha, N. S. *et al.* A natural language processing system for the efficient updating of highly curated pathophysiology mechanism knowledge graphs. *Artificial Intelligence in the Life Sciences* **4,** 100078 (2023).

73. Babaiha, N. S. *et al.* Rationalism in the face of GPT hypes: Benchmarking the output of large language models against human expert-curated biomedical knowledge graphs. *Artificial Intelligence in the Life Sciences* **5,** 100095 (2024).

74. Pan, J. Z. in *Handbook on ontologies* 71–90 (Springer, 2009).

75. Decker, S., Mitra, P. & Melnik, S. Framework for the semantic Web: an RDF tutorial. *IEEE Internet Computing* **4,** 68–73 (2000).

76. Boue, S. *et al.* Causal biological network database: a comprehensive platform of causal biological network models focused on the pulmonary and vascular systems. *Database* **2015,** bav030 (2015).

77. Angles, R. *The Property Graph Database Model.* in *AMW* (2018).

78. Frisendal, T. *The future history of time in data models* en. `https://www.dataversity.net/the-future-history-of-time-in-data-models/`. Accessed: 2024-6-12. Aug. 2019.

79. WIPO. `https://www.wipo.int/pct/en/users/summary.html`.

80. Simmons, E. S. Prior art searching in the preparation of pharmaceutical patent applications. *Drug discovery today* **3,** 52–60 (1998).

81. Senger, S. Assessment of the significance of patent-derived information for the early identification of compound–target interaction hypotheses. *Journal of Cheminformatics* **9,** 1–8 (2017).

82. Marttin, E. & Derrien, A.-C. How to apply examiner search strategies in Espacenet. A case study. *World Patent Information* **54,** S33–S43 (2018).

83. Ohms, J. Validity of PubChem compounds supplied by Patentscope or SureChEMBL. *World Patent Information* **70,** 102134 (2022).

84. Papadatos, G. *et al.* SureChEMBL: a large-scale, chemically annotated patent document database. *Nucleic acids research* **44,** D1220–D1228 (2016).

85. Tate, F. A. Chemical abstracts service. *American Journal of Health-System Pharmacy* **23,** 63–67 (1966).

86. Humayun, F. *et al.* A computational approach for mapping heme biology in the context of hemolytic disorders. *Frontiers in bioengineering and biotechnology* **8,** 74 (2020).

87. Hopp, M.-T. *et al.* Linking COVID-19 and heme-driven pathophysiologies: A combined computational–experimental approach. *Biomolecules* **11,** 644 (2021).

88. Schultz, B. *et al.* A method for the rational selection of drug repurposing candidates from multimodal knowledge harmonization. *Scientific reports* **11,** 11049 (2021).

89. UniProt: the universal protein knowledgebase in 2023. *Nucleic acids research* **51,** D523–D531 (2023).

90. Grissa, D., Junge, A., Oprea, T. I. & Jensen, L. J. Diseases 2.0: a weekly updated database of disease-gene associations from text mining and data integration. *Database* **2022,** baac019 (2022).

91. Zhou, H. *et al.* Stringent homology-based prediction of H. sapiens-M. tuberculosis H37Rv protein-protein interactions. *Biology direct* **9,** 1–30 (2014).

92. Magariños, M. P. *et al.* Illuminating the druggable genome through patent bioactivity data. *PeerJ* **11,** e15153 (2023).

93. Lee, B., Kim, T., Kim, S.-K., Lee, K. H. & Lee, D. Patome: a database server for biological sequence annotation and analysis in issued patents and published patent applications. *Nucleic acids research* **35,** D47–D50 (2007).

94. Wishart, D. S. *et al.* DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic acids research* **34,** D668–D672 (2006).

95. Aurich, D., Schymanski, E. L., de Jesus Matias, F., Thiessen, P. A. & Pang, J. Revealing Chemical Trends: Insights from Data-Driven Visualization and Patent Analysis in Exposomics Research. *Environmental Science & Technology Letters* **11,** 1046–1052 (2024).

96. Arp, H. P. H., Aurich, D., Schymanski, E. L., Sims, K. & Hale, S. E. Avoiding the next silent spring: our chemical past, present, and future. *Environmental Science & Technology* **57,** 6355–6359 (2023).

97. Kong, X. *et al.* STING as an emerging therapeutic target for drug discovery: Perspectives from the global patent landscape. *Journal of Advanced Research* **44,** 119–133 (2023).

98. Lahiry, S. & Rangarajan, K. Patent Landscape for Indian Biopharmaceutical Sector: A Strategic Insight. *Flexible Strategies in VUCA Markets,* 31–47 (2018).

99. Liu, K. *et al.* Global landscape of patents related to human coronaviruses. *International journal of biological sciences* **17,** 1588 (2021).

100. Sharma, R., Sharma, R. & Singla, R. K. Drug Discovery, Diagnostic, and therapeutic trends on Mpox: A patent landscape. *Current Research in Biotechnology* **7,** 100173 (2024).

101. Mucke, H. A. *What patents tell us about drug repurposing for cancer: A landscape analysis* in *Seminars in cancer biology* **68** (2021), 3–7.

102. Hill, S. M. *et al.* Context specificity in causal signaling networks revealed by phosphoprotein profiling. *Cell systems* **4,** 73–83 (2017).

103. Babur, Ö. *et al.* Causal interactions from proteomic profiles: Molecular data meet pathway knowledge. *Patterns* **2** (2021).

104. Catlett, N. L. *et al.* Reverse causal reasoning: applying qualitative causal knowledge to the interpretation of high-throughput data. *BMC bioinformatics* **14,** 1–14 (2013).

105. Peterson, E. & Kaur, P. Antibiotic resistance mechanisms in bacteria: relationships between resistance determinants of antibiotic producers, environmental bacteria, and clinical pathogens. *Frontiers in microbiology* **9,** 2928 (2018).

106. Naghavi, M. *et al.* Global burden of bacterial antimicrobial resistance 1990–2021: a systematic analysis with forecasts to 2050. *The Lancet* **404,** 1199–1226 (2024).

107. Essack, S. Y. & Lenglet, A. Bacterial antimicrobial resistance burden in Africa: accuracy, action, and alternatives. *The Lancet Global Health* **12,** e171–e172 (2024).

108. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data* **3,** 1–9 (2016).

109. Amer, P. *Is your AI hallucinating? New approach can tell when chatbots make things up* 2024.

110. Siontis, K. C., Attia, Z. I., Asirvatham, S. J. & Friedman, P. A. *ChatGPT hallucinating: can it get any more humanlike?* 2024.

111. Sovrano, F., Ashley, K. & Bacchelli, A. *Toward eliminating hallucinations: Gpt-based explanatory ai for intelligent textbooks and documentation* in *CEUR Workshop Proceedings* (2023), 54–65.

APPENDIX A

Appendix

A.1 COVID-19 Knowledge Graph: a computable, multi-modal, cause-and-effect knowledge model of COVID-19 pathophysiology

Reprinted with permission from “Domingo-Fernández D., Baksi, S., Schultz, B., Gadiya, Y., Karki, R., Raschka, T., Ebeling, C., Hofmann-Apitius, M. and Kodamullil, A. T. (2021). COVID-19 Knowledge Graph: a computable, multi-modal, cause-and-effect knowledge model of COVID-19 pathophysiology. *Bioinformatics*, 37(9), 1332-1334.”

Bioinformatics, 37(9), 2021, 1332–1334
doi: 10.1093/bioinformatics/btaa834
Advance Access Publication Date: 6 December 2020
Applications Note

Systems biology

COVID-19 Knowledge Graph: a computable, multi-modal, cause-and-effect knowledge model of COVID-19 pathophysiology

Daniel Domingo-Fernández [1,2,*], Shounak Baksi[3,*], Bruce Schultz [1,*], Yojana Gadiya [1,2], Reagon Karki[1,2], Tamara Raschka[1,2], Christian Ebeling[1], Martin Hofmann-Apitius [1,2] and Alpha Tom Kodamullil [1,2,*]

[1]Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), 53754 Sankt Augustin, Germany, [2]Bonn-Aachen International Center for Information Technology (B-IT), University of Bonn, 53113 Bonn, Germany and [3]Causality Biomodels, KINFRA Hi-Tech Park, Cochin, Kerala 683503, India

*To whom correspondence should be addressed.

Associate Editor: Lenore Cowen

Received on April 14, 2020; revised on July 26, 2020; editorial decision on September 9, 2020; accepted on September 11, 2020

Abstract

Summary: The COVID-19 crisis has elicited a global response by the scientific community that has led to a burst of publications on the pathophysiology of the virus. However, without coordinated efforts to organize this knowledge, it can remain hidden away from individual research groups. By extracting and formalizing this knowledge in a structured and computable form, as in the form of a knowledge graph, researchers can readily reason and analyze this information on a much larger scale. Here, we present the COVID-19 Knowledge Graph, an expansive cause-and-effect network constructed from scientific literature on the new coronavirus that aims to provide a comprehensive view of its pathophysiology. To make this resource available to the research community and facilitate its exploration and analysis, we also implemented a web application and released the KG in multiple standard formats.

Availability and implementation: The COVID-19 Knowledge Graph is publicly available under CC-0 license at https://github.com/covid19kg and https://bikmi.covid19-knowledgespace.de.

Contact: daniel.domingo.fernandez@scai.fraunhofer.de or shounak.baksi@causalitybiomodels.com or bruce.schultz@scai.fraunhofer.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The COVID-19 crisis has prompted a response of the scientific community that is unparalleled in history. Research organizations have dedicated their entire workforce to combat the pandemic. Tens of thousands of researchers in hundreds of universities, governmental laboratories and industrial research departments have entirely focused their efforts on understanding the virus pathophysiology, finding drugs that interfere with its life cycle and developing immunization strategies for future vaccines (Chahrour *et al.*, 2020).

While the steep increase in research activities in the COVID-19 context has led to an unprecedented increase of scientific publications, it becomes challenging to identify genuine novel findings and discern them from those that are already known. The process of discriminating ‘known knowns’ from ‘unknown knowns’ can be supported by knowledge graphs (KGs), as they provide a means to capture, represent and formalize structured information (Nelson *et al.*, 2019). Furthermore, although these KGs were originally developed to describe interactions between entities, they are complemented by a broad range of algorithms that have been proven to partially automate the process of knowledge discovery (Cowen *et al.*, 2017; Humayun *et al.*, 2020). Importantly, novel machine learning techniques can generate latent, low-dimensional representations of the KG which can then be utilized for downstream tasks such as clustering or classification (Hamilton *et al.*, 2017).

In this article, we present an approach to lay the foundation for a comprehensive KG in the context of COVID-19. Our work is complemented by a web application that enables users to comprehensively explore the information contained in the KG. To facilitate the ease of usage and interoperability of our KG, we have released its content in various standard formats to promote its adoption and enhancement by the scientific community.

2 Material and methods

In this section, we outline the methodology used to: (i) select the corpus, (ii) generate the COVID-19 KG and (iii) develop the web application for exploring the KG.

2.1 Selection of scientific literature

For the creation of the KG, scientific literature related to COVID-19 was retrieved from open access and freely available journals (see details in Supplementary Text). This corpus was then filtered based on available information about potential drug targets for COVID-19, biological pathways in which the virus interferes to replicate in its human host, and information on the various viral proteins along with their functions. Finally, the articles were prioritized based on the level of information that could be captured in the modeling language used to build the KG.

2.2 Constructing the COVID-19 Knowledge Graph

Evidence text from the prioritized corpus was manually encoded in Biological Expression Language (BEL) as a triple (i.e. source node—relation—target node) including metadata about the nodes and their relationships as well as corresponding provenance and contextual information. BEL scripts generated from this curation work are freely available at https://github.com/covid19kg along with their network representations in several other standard formats (e.g. SIF, GraphML and NDEx). By making this data available in multiple formats, we are seeking to facilitate the analysis of the KG with a broad range of methods/software as well as promote its integration into other biological databases and web services such as the one presented in the following section.

2.3 Web application

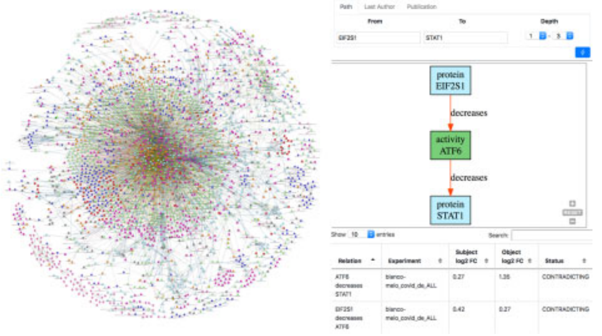
To better aid the exploration and usage of the generated COVID-19 Knowledge Graph, a web application was developed using Biological Knowledge Miner (BiKMi), an in-house software package designed for exploring pathways and molecular interactions within a BEL-derived network. The front-end of the application was constructed using the Python Django web framework, while the back-end of the software is implemented using OrientDB, a multi-model database management system that allows for both relational and graph queries to be made against a database via its API (Supplementary Text), which opens the avenue towards systematic comparison of different COVID models.

3 Results

We introduce a KG that comprises mechanistic information on COVID-19 published in 160 original research articles. In its current state, the COVID-19 KG incorporates 4016 nodes, covering 10 entity types (e.g. proteins, genes, chemicals and biological processes) and 10 232 relationships (e.g. increases, decreases and association), forming a seamless interaction network (Supplementary Text). Given the selected corpora, these cause-and-effect relations primarily denote host-pathogen interactions as well as comorbidities and symptoms associated with COVID-19. Furthermore, the KG contains molecular interactions related to host invasion (e.g. spike glycoprotein and its interaction with the host via receptor ACE2) and the effects of the downstream inflammatory, cell survival and apoptosis signaling pathways.

A key aspect of the COVID-19 KG is in its large coverage of drug–target interactions along with the biological processes, genes and proteins associated with the novel coronavirus. We have identified over 300 candidate drugs currently being investigated in the context of COVID-19 (Supplementary Text), including proposed repurposing candidates and drugs under clinical trial.

Along with the KG, we implemented a web application (https://bikmi.covid19-knowledgespace.de) for querying, browsing and navigating the KG (Fig. 1). The visualization enables users to explore and query the network (e.g. filtering nodes or edges, or calculating paths between nodes of interest). Additionally, it enables users to upload *omics* data and validate its signals against the knowledge contained in the network. To demonstrate this feature, the web application is loaded with the transcriptomics experiments conducted by Blanco-Melo *et al.* (2020).



From: EIF2S1
To: STAT1

Relation	Experiment	Subject log2 FC	Object log2 FC	Status
ATF6 decreases STAT1	blanco-melo_covid_de_ALL	0.27	1.38	CONTRADICTING
EIF2S1 decreases ATF6	blanco-melo_covid_de_ALL	0.42	0.27	CONTRADICTING

Fig. 1. (Left) Visualization of the COVID-19 KG in BiKMi. (Right) Querying paths between two nodes and verifying their consistency with transcriptomics data

4 Discussion

The novel coronavirus has motivated a profound response by the scientific community and has led to the rapid publishing of COVID-19 research. As an attempt to organize and formally represent the most current knowledge of the virus, we have introduced a KG comprising mechanistic information around COVID-19 biology and pathophysiology. The presented KG provides a comprehensive overview of relevant viral protein interactions and their downstream molecular mechanisms. Additionally, it also includes the vast majority of potential drug–targets as well as clinical manifestations associated with comorbidities and symptoms. Given the biological complexity and the sparse information we currently have on the pathophysiology of the virus, mechanistic knowledge contained in the KG could be promising for the discovery of yet hidden interactions. The COVID-19 KG presented here is part of a bigger ecosystem that integrates disease maps with three of the largest pathway databases (Domingo-Fernández *et al.*, 2019).

Not only do we provide a web application to make the content accessible to the research community, but we also have released the KG in a variety of standard formats. In doing so, we aim to foster an exchange of information across similar modeling approaches (Ostaszewski *et al.*, 2020) (Supplementary Text) as well as to facilitate its analytic use on both knowledge- and data-driven methods. Furthermore, the knowledge present in high-quality manually curated approaches can be combined with the information extracted by scalable text mining approaches such as Wang *et al.* (2020), which enable to systematically scan COVID-19 literature and construct KGs based on entity co-occurrence or relation extraction. However, combining both modeling approaches involves understanding their differences as well as relative strengths and weaknesses, some of which are discussed in Supplementary Text. Finally, we plan to make future releases of the KG to ensure the most up-to-date content as well as to benefit from its integration and crosstalk with other similar activities (i.e. #covidpathways).

Funding

This work was supported by the MAVO and ICON programs of the Fraunhofer Society.

Conflict of Interest: none declared.

References

Blanco-Melo,D. *et al.* (2020) Imbalanced host response to SARS-CoV-2 drives development of COVID-19. *Cell*, **181**, 1036–1045.

Chahrour,M. *et al.* (2020) A bibliometric analysis of COVID-19 research activity: a call for increased output. *Cureus*, **12**, e7357.

Cowen,L. *et al.* (2017) Network propagation: a universal amplifier of genetic associations. *Nat. Rev. Genet.*, **18**, 551–562.

Domingo-Fernández,D. *et al.* (2019) PathMe: merging and exploring mechanistic pathway knowledge. *BMC Bioinformatics*, **20**, 243.

Hamilton,W.L. *et al.* (2017) Representation learning on graphs: methods and applications. *IEEE Data Eng. Bull.*, **40**, 52–74

Humayun,F. *et al.* (2020) A computational approach for mapping heme biology in the context of hemolytic disorders. *Front. Bioeng. Biotechnol.*, **8**, 74.

Nelson,W. *et al.* (2019) To embed or not: network embedding as a paradigm in computational biology. *Front. Genet.*, **10**, 381.

Ostaszewski,M. *et al.* (2020) COVID-19 disease map, building a computational repository of SARS-CoV-2 virus-host interaction mechanisms. *Sci. Data*, **7**, 1–4.

Wang,Q. *et al.* (2020) COVID-19 Literature Knowledge Graph Construction and Drug Repurposing Report Generation. https://www.cell.com/cell/fulltext/S0092-8674(20)30489-X?_return.

A.2 Mpox Knowledge Graph: a comprehensive representation embedding chemical entities and associated biology of Mpox

Reprinted with permission from "Karki, R., Gadiya, Y., Zaliani, A., and Gribbon, P. (2023). Mpox Knowledge Graph: a comprehensive representation embedding chemical entities and associated biology of Mpox. *Bioinformatics Advances*, 3(1), vbad045.".

Bioinformatics Advances, 2023, vbad045
https://doi.org/10.1093/bioadv/vbad045
Advance Access Publication Date: 3 April 2023
Application Note

Systems biology

Mpox Knowledge Graph: a comprehensive representation embedding chemical entities and associated biology of Mpox

Reagon Karki [1,2,]*, Yojana Gadiya [1,2], Andrea Zaliani[1,2] and Philip Gribbon[1,2]

[1]Discovery Research ScreeningPort, Fraunhofer Institute for Translational Medicine and Pharmacology (ITMP), Schnackenburgallee 114, 22525 Hamburg, Germany and [2]Fraunhofer Cluster of Excellence for Immune-Mediated Diseases (CIMD), Theodor Stern Kai 7, 60590 Frankfurt, Germany

*To whom correspondence should be addressed.

Associate Editor: Nicola Mulder

Received on October 19, 2022; revised on March 20, 2023; editorial decision on March 26, 2023; accepted on March 31, 2023

Abstract

Summary: The outbreak of Mpox virus (MPXV) infection in May 2022 is declared a global health emergency by WHO. A total of 84 330 cases have been confirmed as of 5 January 2023 and the numbers are on the rise. The MPXV pathophysiology and its underlying mechanisms are unfortunately not yet understood. Likewise, the knowledge of biochemicals and drugs used against MPXV and their downstream effects is sparse. In this work, using Knowledge Graph (KG) representations we have depicted chemical and biological aspects of MPXV. To achieve this, we have collected and rationally assembled several biological study results, assays, drug candidates and pre-clinical evidence to form a dynamic and comprehensive network. The KG is compliant with FAIR annotations allowing seamless transformation and integration to/with other formats and infrastructures.

Availability and implementation: The programmatic scripts for Mpox KG are publicly available at https://github.com/Fraunhofer-ITMP/mpox-kg. It is hosted publicly at https://doi.org/10.18119/N9SG7D.

Contact: reagon.karki@itmp.fraunhofer.de

Supplementary information: Supplementary data are available at *Bioinformatics Advances* online.

1 Introduction

The recent coronavirus disease 2019 (COVID-19) pandemic has drastically changed the way research and scientific studies operate in areas of infectious and epidemic diseases. Although new discoveries and uncovering pathophysiology are the ultimate expectations, a new aspect that has been crucial to these is the response time (Khanna *et al.*, 2020). Despite the expertise and technologies of the highest levels in hospitals, pharmaceutical companies and research institutes, the response was not always timely. This clearly was the impact of lack of preparedness and since then we are determined to avoid experiencing the same chaos in future epidemics (Villa *et al.*, 2020). One of the setbacks in such a situation was the unavailability of enough research data with metadata compliant with Findable, Accessible, Interoperable and Reproducible (FAIR) data principles, consequently leading to mapping gaps between different domains of scientific studies. A number of efforts have emerged since then to harmonize sparse data and better understand the etiology of the disease (Harrison *et al.*, 2021; Schmidt *et al.*, 2021).

The ongoing multi-country outbreak of Mpox virus (MPXV) which started in May 2022 (https://www.ecdc.europa.eu/en/Mpox-outbreak) has been declared a global health emergency and stands as another potential threat of pandemic. Unfortunately, the etiology of MPXV is not known and therefore, there is an urgent need to decipher it. This involves identifying the involvement of viral and host proteins in the infection, their interactions, virus replication biology and potential drug candidates to perturb viral processes and mechanisms within the host. Additionally, from drug discovery and therapeutic perspective, it is important to know active molecules and their pharmacology either as a direct effect on viral–host interactions or as cellular toxicology effects. Understanding all the above-mentioned aspects will help accelerate drug repurposing and drug discovery processes. In this work, we have created a comprehensive Mpox Knowledge Graph (KG) that represents chemical-specific information such as chemicals and drugs active against MPXV along with their side effects, biological-specific information such as proteins, and their associated biological processes. The KG is represented with standard ontologies aligning it with the FAIR data principles. Furthermore, the KG is available in various graph formats to facilitate data handling and processing as required by the scientific community.

2 Materials and methods

The overall methodology used for the creation of the Mpox KG is divided into three main steps (i) biological/chemical resources identification, (ii) data harmonization and standardization and (iii) KG generation (Fig. 1).

2.1 Resource identification

The chemical compounds used or tested against MPXV were retrieved from public chemical data resources, i.e. PubChem (Kim *et al.*, 2021), ChEMBL (Davies *et al.*, 2015) databases (last accessed: 12 January 2022). We queried PubChem with NCBI Taxonomy Identifier (ID) of the virus (NCBITaxon: 10244) and the associated chemicals were listed under the table ‘Chemicals and Bioactivities’ and sub-table ‘Tested Compounds’. Since ChEMBL has its independent ontology for taxonomy, the database was queried using ChEMBL ID for MPXV (i.e. CHEMBL613120). Afterwards, we selected chemicals with pChEMBL value > 6 from either binding or functional assays since this condition ensures bioactivity of a given chemical. In general, this value is half-maximal response concentration/potency/affinity on a negative logarithmic scale. Next, using the taxonomy ID of MPXV, we collected reviewed protein entries (Swiss-Prot) from UniProt (Apweiler *et al.*, 2004). For human proteins, we queried DISEASES, a human disease database, with DOID: 3292 (Grissa *et al.*, 2022). Lastly, Open Targets Platform was used to fetch information about the ‘druggability’ of proteins reported from studies (Ochoa *et al.*, 2021).

2.2 Programmatic methods for data fetching and harmonization

The programmatic scripts and methods written were written in python (version 3.10) and are available at https://github.com/Fraunhofer-ITMP/mpox-kg. Firstly, we converted PubChem IDs to ChEMBL IDs because the information about chemical-associated proteins, assays, mechanism of actions, pathways and diseases can be fetched from the ChEMBL API using ChEMBL IDs. After combining the identified proteins with proteins from UniProt and DISEASES, we used the UniProt API to extend the information on molecular functions, biological processes, sequences, pathways and diseases. Furthermore, we identified additional ChEMBL compounds that target proteins collected in the workflow and repeated the aforementioned steps of using ChEMBL and UniProt API. Lastly, we mapped and harmonized chemical and protein names as they were collected from different resources and had different identifiers. For example, PubChem entry with compound ID 16124688 is registered in ChEMBL as CHEMBL1257073. We have used ChEMBL IDs for standard and uniform representation of chemicals. Similarly, UniProt IDs were converted to HUGO names as it helps researchers to readily identify a given protein.

2.3 Construction of Mpox KG

The Mpox KG is represented in the form of semantic triples using Biological Expression Language (BEL) with metadata annotation on nodes and relations using the PyBEL framework (Hoyt *et al.*, 2018). PyBEL is a software tool built to facilitate data parsing, semantics validation and visualization of data generated in BEL format. The framework provides a library of functions for exploring, querying and analyzing the KG. Moreover, the KG is exported to other formats such as json, csv, sql, graphml and Neo4j which enables systematic comparison or integration with other KGs. The KG is hosted publicly in NDExbio platform under the URL: https://doi.org/10.18119/N9SG7D (Pillich *et al.*, 2021).

3 Results

While literature and public data resources contain sparse MPXV-related information, our approach has reached out to various resources and built a bridge between the chemical and biological worlds, thus yielding a comprehensive KG. Starting from chemicals associated with MPXV, we were able to identify corresponding assays with bioactivity, target proteins and their biological processes. Moreover, with MPXV and human proteins, we not only summoned knowledge about the aforementioned aspects but also identified chemicals targeting these proteins.

Our query from PubChem retrieved 24 chemicals, all of which were successfully mapped to corresponding ChEMBL IDs. Similarly, eight chemicals were retrieved from ChEMBL, out of which six were identical with the PubChem chemicals. The search in UniProt fetched 11 MPXV proteins whereas DISEASES returned 19 human proteins. Using these results as the primer for the KG, we created a KG using ChEMBL and UniProt API. The KG comprised 9117 nodes and 44 516 relationships where we have identified 565 putative drugs targeting human and viral proteins. The full summary of KG statistics is available in Supplementary Figures S1 and S2. The proteins represented in the KG were further labeled with ‘druggability’ information using Open Targets (Supplementary Table S1). Additionally, we performed a sequence similarity search for the MPXV proteins and identified human homologs with sequence identity >35%. These results are provided in Supplementary Table S2.

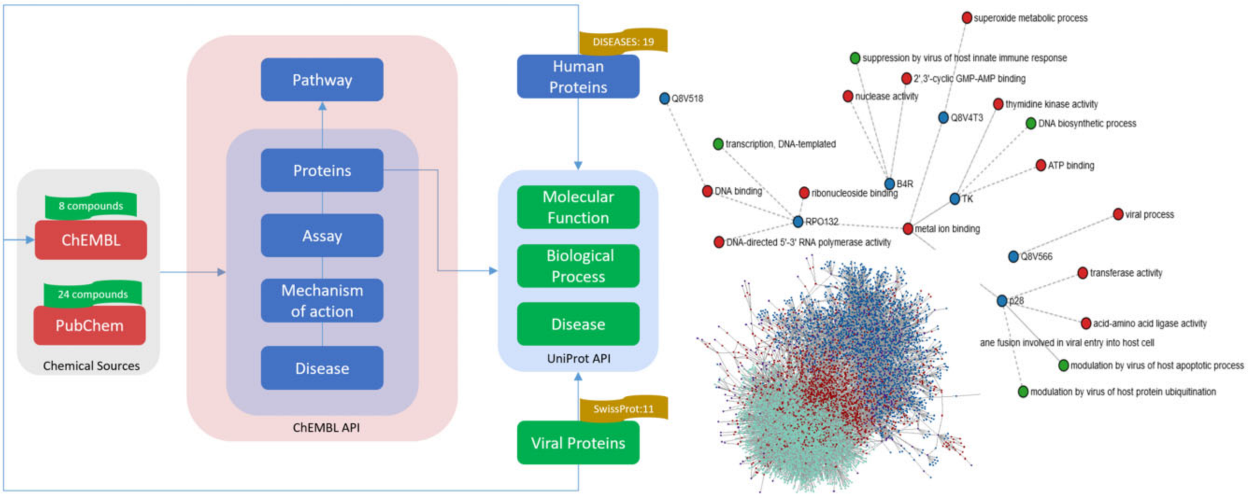


Fig. 1. A schematic representation of the KG workflow (left) and visualization of the KG (right)

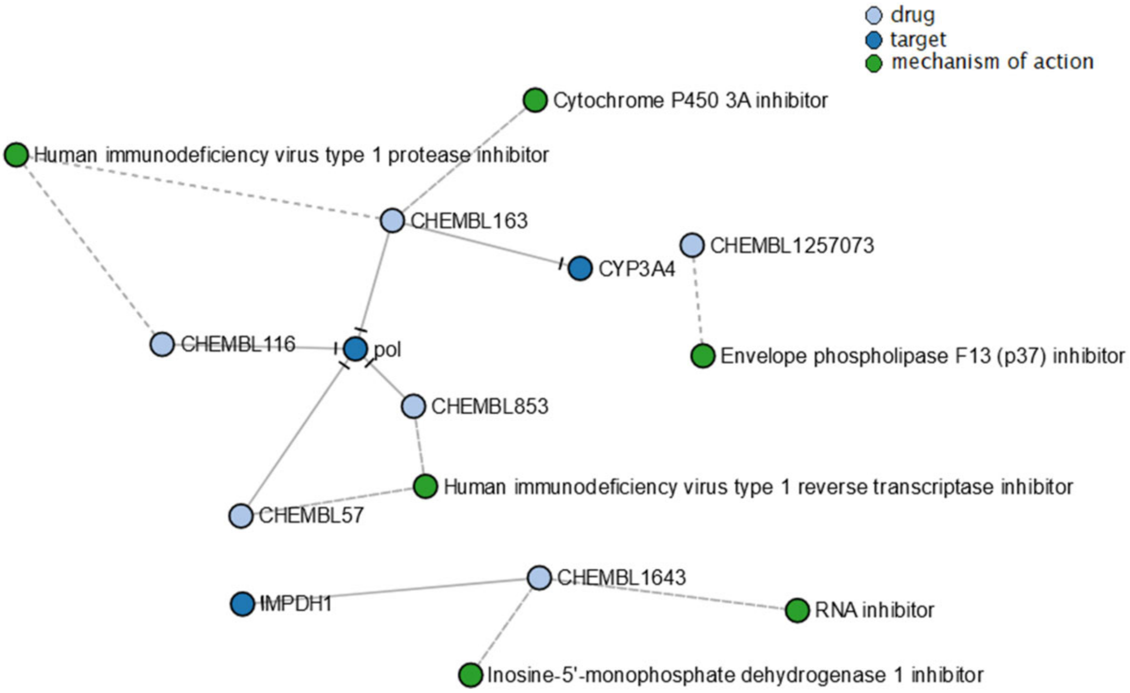

drug
target
mechanism of action
Cytochrome P450 3A inhibitor
Human immunodeficiency virus type 1 protease inhibitor
CHEMBL163
CYP3A4
CHEMBL1257073
CHEMBL116
pol
Envelope phospholipase F13 (p37) inhibitor
CHEMBL853
Human immunodeficiency virus type 1 reverse transcriptase inhibitor
CHEMBL57
CHEMBL1643
IMPDH1
RNA inhibitor
Inosine-5'-monophosphate dehydrogenase 1 inhibitor

Fig. 2. A section of the KG showing drugs, their target(s) and mechanism of action(s)

As there are no Mpox-specific drugs, we tried to identify drugs that were used against similar viruses. Our query to the KG identified 15 drugs that were in different phases of clinical trials. After filtering out for drugs that were in Phase IV, 12 drugs remained (Supplementary Table S3). To this, we filtered the list by retaining drugs with known mechanism of action(s), which resulted in 6 drugs (Fig. 2). Out of these, Ribavirin (CHEMBL1643) has been known to be a direct inhibitor of human IMPDH1, a protein involved in the regulation of cell growth (Te *et al.*, 2007). Likewise, Nevirapine (CHEMBL57) and Zalcitabine (CHEMBL853) have been shown to function as human immunodeficiency virus (HIV) type 1 reverse transcriptase inhibitors (pol) by inhibiting the corresponding protein (de Clercq, 2005; Ghosn *et al.*, 2009). Similarly, Amprenavir (CHEMBL116) and Ritonavir (CHEMBL163) are other known pol inhibitors, with the latter also functioning as human CYP3A4 inhibitor (Sadler *et al.*, 1999; Sevrioukova and Poulos, 2010).

In order to decipher any possible link between Mpox proteins and pol, we performed a BLAST analysis against Mpox proteins (Target database: UniProtKB Swiss-Prot and Mpox Taxonomy: 10244) to find their relatedness and discover the potential to repurpose its drugs against Mpox. The search identified 6 proteins with overlapping amino acids, all in the C-terminal of the protein. Among these, OPG148 (UniProt: A0A7H0DNC0) had the best overlap with 23 amino acids mapped across 97 amino acids of both proteins (Supplementary Fig. S4). Since, Mpox does not have any reported protein–ligand interaction so far, we extended the BLAST search with OPG148 [Target database: UniProtKB with 3D structure (PDB)] to identify Mpox-related proteins with ligand bound conformations. The analysis identified orthologous OPG148 (UniProt: P20995) in Vaccinia virus with 97.2% similarity (Supplementary Fig. S5). Afterwards, we explored its PDB structures and filtered for identifiers with ligands bound to the ortholog. We found that PDB structures 4YGM and 4YIG had Uracil (CHEMBL566) bound to the ortholog (Burley *et al.*, 2023). Next, we super-imposed Uracil onto Nevirapine and Zalcitabine to find out its structural resemblance. We found that Uracil-Zalcitabine had a complete substructure match while Nevirapine, although partial, had a good 3D super-imposition (Supplementary Fig. S6). Through this approach, we could swiftly identify candidate ligands for Mpox OPG148 using its ligand-complexed ortholog protein.

Lastly, we identified Tecovirimat (CHEMBL1257073) in the KG which has been previously reported to function as Envelope phospholipase F13 (p37) inhibitor (Yang *et al.*, 2005). Tecovirimat is one of the smallpox-specific drugs tested against Mpox and has been proven to induce anti-viral effects in Orthopoxvirus infections (Duraffour *et al.*, 2015). With these lines of evidence, we have shortlisted FDA-approved drugs which bear the potential to be repurposed against Mpox and therefore suggest the need of further research and investigation.

4 Discussion

The COVID-19 pandemic has alerted the scientific community at different levels such as identifying early predictors, understanding pathogen biology and consequent pathophysiology, selecting first line of treatments, identifying putative drugs, and enabling data availability which is crucial for all sorts of research activities (Bibi *et al.*, 2022; Domingo-Fernández *et al.*, 2021). The lesson learned from COVID-19 is to be prepared and act immediately to avoid the next unprepared ‘COVID-19-esque’ situation. In this regard, aligning with FAIR data principles, we have created a Mpox KG which is a comprehensive representation of biological and chemical entities associated with MPXV. Our analysis identified six drugs with known mechanism of actions and their target proteins which could be useful in upcoming Mpox studies. This highlights a straightforward application of the KG as it allows identification and selection of putative active Mpox-related chemicals. To our knowledge, ours is the first KG in MPXV research.

One significant strength of our KG is that it not only embeds, harmonizes and visualizes entities but also serves as a primer for downstream analyses. For example, a chemoinformatician can readily run similarity search analyses using the chemicals represented in the KG. Similarly, a biologist working with a certain protein and chemical can quickly find out other chemicals targeting the same protein. We know that this approach is not without limitations but embedding KGs in drug-discovery process allows faster and innovative hypotheses generation. Considering these, we aim to facilitate ongoing and upcoming MPXV studies by serving a useful resource to different research groups and therefore will continue to update the KG actively. One of our next updates will include annotation of proteins with MPXV-specific omics data. Moreover, we plan to explore functional interactions of orthologous proteins in other Orthopoxviruses. Finally, we plan to reach out to other public resources for enriching the knowledge in the KG.

Authors' contributions

R.K. ideated and implemented the KG. R.K. Y.G., A.Z. and P.G. wrote the article.

Funding

This work was supported by Horizon Europe's BY-COVID project [101046203] and EOSC Future Project [101017536].

Conflict of Interest: none declared.

References

Apweiler,R. *et al.* (2004) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, **32**, D115–D119.

Bibi,N. *et al.* (2022) Drug repositioning against COVID-19: a first line treatment. *J. Biomol. Struct. Dyn.*, **40**, 12812–12826.

Burley,S.K. *et al.* (2023) RCSB protein data bank (RCSB.org): delivery of experimentally-determined PDB structures alongside one million computed structure models of proteins from artificial intelligence/machine learning. *Nucleic Acids Res.*, **51**, D488–D508.

Davies,M. *et al.* (2015) ChEMBL web services: streamlining access to drug discovery data and utilities. *Nucleic Acids Res.*, **43**, W612–W620.

de Clercq,E. (2005) Emerging anti-HIV drugs. *Expert Opin. Emerg. Drugs*, **10**, 241–273.

Domingo-Fernández,D. *et al.* (2021) COVID-19 knowledge graph: a computable, multi-modal, cause-and-effect knowledge model of COVID-19 pathophysiology. *Bioinformatics*, **37**, 1332–1334.

Duraffour,S. *et al.* (2015) ST-246 is a key antiviral to inhibit the viral F13L phospholipase, one of the essential proteins for orthopoxvirus wrapping. *J. Antimicrob. Chemother.*, **70**, 1367–1380.

Ghosn,J. *et al.* (2009) HIV-1 resistance to first-and second-generation non-nucleoside reverse transcriptase inhibitors. *AIDS Rev.*, **11**, 165–173.

Grissa,D. *et al.* (2022) DISEASES 2.0: a weekly updated database of disease–-gene associations from text mining and data integration. *Database*, **2022**.

Harrison,P.W. *et al.* (2021) The COVID-19 data portal: accelerating SARS-CoV-2 and COVID-19 research through rapid open access data sharing. *Nucleic Acids Res.*, **49**, W619–W623.

Hoyt,C.T. *et al.* (2018) PyBEL: a computational framework for biological expression language. *Bioinformatics*, **34**, 703–704.

Khanna,R.C. *et al.* (2020) COVID-19 pandemic: lessons learned and future directions. *Indian J. Ophthalmol.*, **68**, 703–710.

Kim,S. *et al.* (2021) PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Res.*, **49**, D1388–D1395.

Ochoa,D. *et al.* (2021) Open targets platform: supporting systematic drug–target identification and prioritisation. *Nucleic Acids Res.*, **49**, D1302–D1310.

Pillich,R.T. *et al.* (2021) NDEx: accessing network models and streamlining network biology workflows. *Curr. Protoc.*, **1**, e258.

Sadler,B.M. *et al.* (1999) Safety and pharmacokinetics of amprenavir (141W94), a human immunodeficiency virus (HIV) type 1 protease inhibitor, following oral administration of single doses to HIV-infected adults. *Antimicrob. Agents Chemother.*, **43**, 1686–1692.

Schmidt,C.O. *et al.*; NFDI4Health Task Force Covid-19. (2021) Making COVID-19 research data more accessible-building a nationwide information infrastructure. *Bundesgesundheitsblatt, Gesundheitsforschung, Gesundheitsschutz*, **64**, 1084–1092.

Sevrioukova,I.F. and Poulos,T.L. (2010) Structure and mechanism of the complex between cytochrome P4503A4 and ritonavir. *Proc. Natl. Acad. Sci. USA*, **107**, 18422–18427.

Te,H.S. *et al.* (2007) Mechanism of action of ribavirin in the treatment of chronic hepatitis C. *Gastroenterol. Hepatol.*, **3**, 218.

Villa,S. *et al.* (2020) The COVID-19 pandemic preparedness or lack thereof: from China to Italy. *GHM*, **2**, 73–77.

Yang,G. *et al.* (2005) An orally bioavailable antipoxvirus compound (ST-246) inhibits extracellular virus formation and protects mice from lethal orthopoxvirus challenge. *J. Virol.*, **79**, 13139–13149.

A.3 Exploring SureChEMBL from a drug discovery perspective

Reprinted with permission from “Gadiya, Y., Shetty, S., Hofmann-Apitius, M., Gribbon, P., and Zaliani, A. (2024). Exploring SureChEMBL from a drug discovery perspective. *Scientific Data*, 11, 507.”

www.nature.com/scientificdata

scientific data

OPEN

ANALYSIS

Exploring SureChEMBL from a drug discovery perspective

Yojana Gadiya[1,2,3] ✉, Simran Shetty[1,2,4], Martin Hofmann-Apitius[3,5], Philip Gribbon[1,2] & Andrea Zaliani[1,2]

In the pharmaceutical industry, the patent protection of drugs and medicines is accorded importance because of the high costs involved in the development of novel drugs. Over the years, researchers have analyzed patent documents to identify freedom-to-operate spaces for novel drug candidates. To assist this, several well-established public patent document data repositories have enabled automated methodologies for extracting information on therapeutic agents. In this study, we delve into one such publicly available patent database, SureChEMBL, which catalogues patent documents related to life sciences. Our exploration begins by identifying patent compounds across public chemical data resources, followed by pinpointing sections in patent documents where the chemical annotations were found. Next, we exhibit the potential of compounds to serve as drug candidates by evaluating their conformity to drug-likeness criteria. Lastly, we examine the drug development stage reported for these compounds to understand their clinical success. In summary, our investigation aims at providing a comprehensive overview of the patent compounds catalogued in SureChEMBL, assessing their relevance to pharmaceutical drug discovery.

Introduction

Patent documents are legal documents that disclose an invention to the public (https://www.wipo.int/patents/en/). With this disclosure, the holder of a valid patent document generally has the exclusive right to make, use, and sell the invention for approximately 20 years in a given jurisdiction[1,2]. In drug discovery, researchers explore patent documents to identify competing interests associated with a drug candidate across various organizations, such as pharmaceutical companies, universities, or individuals[3]. Additionally, patent documents serve as a catalyst for medicinal chemists, empowering them to optimize their drug candidates strategically and ensure their alignment with freedom-to-operate (FTO) zones that may exist outside of the scope of the claimed patent coverage[4].

Pharmaceutical-based patenting activity, which mainly covers claims related to therapeutic design, synthesis, and formulation, among others claims, reveals critical information pertaining to the development and prescription of drugs and biologics. In doing so, it serves as a valuable resource for understanding the landscape and dynamics of the pharmaceutical industry[5–9]. Pharmaceutical patent documents cover two fundamental components: the compound itself and its application[10,11]. The compound is usually identified in its various forms, such as within a Markush structure, a trade/generic name, etc. Patent documents claim a compound by its structure or even claim a family of structures (based on a scaffold). The chemical structure information is the basis for conducting chemical patent searches by scientists and professionals and is leveraged by commercial vendors in the form of expert software tools and services (eg. CAS-SciFinder)[12]. The application field(s) of a patent document is usually found in the claims or description sections of the document in text format. A claim's text description is a legally focused document that often contains specialized terminology and jargon integral to the patent domain. This content plays a crucial role in defining the scope of the underlying patent document, making it a subject of study for patent lawyers and pharma R&D scientists who seek to comprehend the FTO space associated with the patent documents.

[1]Fraunhofer Institute for Translational Medicine and Pharmacology (ITMP), Schnackenburgallee 114, 22525, Hamburg, Germany. [2]Fraunhofer Cluster of Excellence for Immune-Mediated Diseases (CIMD), Theodor Stern Kai 7, 60590, Frankfurt, Germany. [3]Bonn-Aachen International Center for Information Technology (B-IT), University of Bonn, 53113, Bonn, Germany. [4]Hamburg University of Applied Sciences (HAW), 20099, Hamburg, Germany. [5]Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Schloss Birlinghoven, 53757, Sankt Augustin, Germany. ✉e-mail: Yojana.Gadiya@itmp.fraunhofer.de

Pharmaceutical patent analysis covers a wide range of topics, including patenting trends[13], tools for patent protection[14,15] and the identification of novel chemical entities for certain treatments or diseases[16,17]. In a study by Falaguera and Mestres[18,19], compounds mined from SureChEMBL, a public patent database, were found to be a collection of starting materials, intermediate products, or pharmacologically relevant compounds (i.e. compounds that target genes or diseases)[18]. In order to make these compound classifications, the authors employed chemoinformatic methods, such as the matched molecular pair (MMP) analysis[20] and the maximum common substructure (MCS) search[18], as alternatives to the generic Markush structural searches[21,22]. Additionally, these methods have allowed for the generation of metrics to assess the chemical novelty and patentability of new compounds[23]. Despite these efforts, the majority of the aforementioned analyses have been limited to patent documents filed and/or granted in the United States of America (USA). This can be attributed to three main reasons: (i) the United States is the world's largest pharmaceutical market[24], (ii) the presence of an easy-to-use US-centric public patent database, the United States Patent and Trademark Office (USPTO), which allows for bulk download of patent documents and their metadata[25], and (iii) the availability of resources, such as the FDA's Orange Book, which allows for the tracking of drug candidates and their corresponding patent documents through time[26]. Furthermore, these analyses restrict pharmaceutical patent documents to those tagged with International Patent Classification (IPC) code A61K, an IPC class that includes hygiene-related patent documents in addition to medicinal ones, potentially merging non-pharmaceutical annotations.

SureChEMBL (https://www.surechembl.org/) is an extensive publicly available patent compound data catalogue for the life sciences[27]. This database identifies compounds, along with other biomedical entities, such as genes and diseases, from patent documents through the use of automated text and image mining pipelines. Furthermore, SureChEMBL keeps an individual record of each extracted compound, associating it with structural information (i.e., SMILES and InChIKeys) and the section of the patent document (i.e. claims, title, description, etc.) where the compound was extracted from. In this study, we aim to investigate the relevance of compounds annotated by SureChEMBL's pipeline with respect to approved drugs in the pharmaceutical market. Specifically, we applied a medicinal chemistry lens on compounds in SureChEMBL to identify patterns within their molecular scaffolds, as well as the physiochemical properties of patented compounds. Moreover, we assessed the similarity of these compounds to drugs through drug-likeness traits defined by Lipinski (Rule of Five). Rather than limiting the investigation to patent documents found in the United States, as done in all previous methodologies, we broaden our scope to a diverse dataset of patent applications filed and/or granted globally. By doing so, we covered larger IPC patent classes, including information on the medicinal utility of compounds and their formulations. Furthermore, our exploration scrutinizes the availability of compounds described in patent documents and beyond, specifically those annotated by public compound databases. We conclude by delving into the clinical candidate space of the patent compounds in order to understand the success rate of progressing compounds from patent application to clinical practice.

Results

In the following subsections, we evaluate the chemical space of patent documents found in SureChEMBL, an open-access public patent database. First, we provide a brief statistical summary of the data present in SureChEMBL, with a focus on the country where the patent documents were first registered. Afterwards, we discuss the searchability (i.e. the ability to search for compounds in other databases) of the patent compounds within large chemical databases, namely PubChem, ChEMBL and DrugBank. Next, we review the findability (i.e. the ability to identify the section in patent documents through which the compounds were annotated) of the patent compounds. Following this, we explored the drug-likeness of patent compounds through rules like Ro5 and beyond, evaluated their structural diversity through the Murcko scaffold, and reviewed the presence of structural alerts like Pan-assay interference structures (PAINS) in these compounds. Finally, we briefly discuss the progression of a compound from patent documents to the market through clinical trials.

Quantitative overview of data in SureChEMBL. From the statistical side, our dataset included a collection of 10 million compounds found in over 1.5 million patent applications (including both granted and non-granted) between 2015 and 2022. Throughout this study, we used the term "patent compounds" to identify those compounds that were captured and annotated by SureChEMBL's internal pipeline to be associated with a patent application. The patent documents in SureChEMBL are captured across a number of patent offices, namely the USPTO, European Patent Office (EPO), Japan Patent Office (JPO) and World Intellectual Property Organization (WIPO). However, it is worth noting that WIPO usually consists of only filed patent documents and does not have the authority to grant any patent. Additionally, the patent documents in SureChEMBL cover a broad range of IPC classes, such as human necessities (A01, A23, A24, A61, A62B), chemistry and metallurgy-oriented (C05, C06, C07, C08, C09, C10, C11, C12, C13, C14) and physics (G01N), all of which are part of this study.

In SureChEMBL, a patent document is assigned a unique SureChEMBL patent number (SCPN) (for eg. US-1234567-A1) that consists of a country code (based on the country the patent document was first registered in), followed by a 7–11 digit number, and a patent kind code. In our study, we used the country code in the SCPN to understand the distribution of patent documents across different patent offices. This exploration revealed that patent documents were predominantly filed in the United States and Europe, with 57.3% (24% granted) and 26.6% (11% granted) patents, respectively. Moreover, SureChEMBL also aggregated compounds from Japan (JP), but we found the contribution of these patent documents to be less than 1%[27] (Fig. 1a).

As mentioned previously, each patent document is associated with a "patent kind code" by SureChEMBL. This code is a two-letter alphanumeric code that assists patent officers and reviewers in efficiently distinguishing different kinds of patent documents, such as utility, design, or plant patents. Utility patent documents involve the "discovery and invention of new and useful processes", design patent documents cover the invention of a "novel design for an article of manufacture", and plant patent documents cover the scope for "discovering an asexually

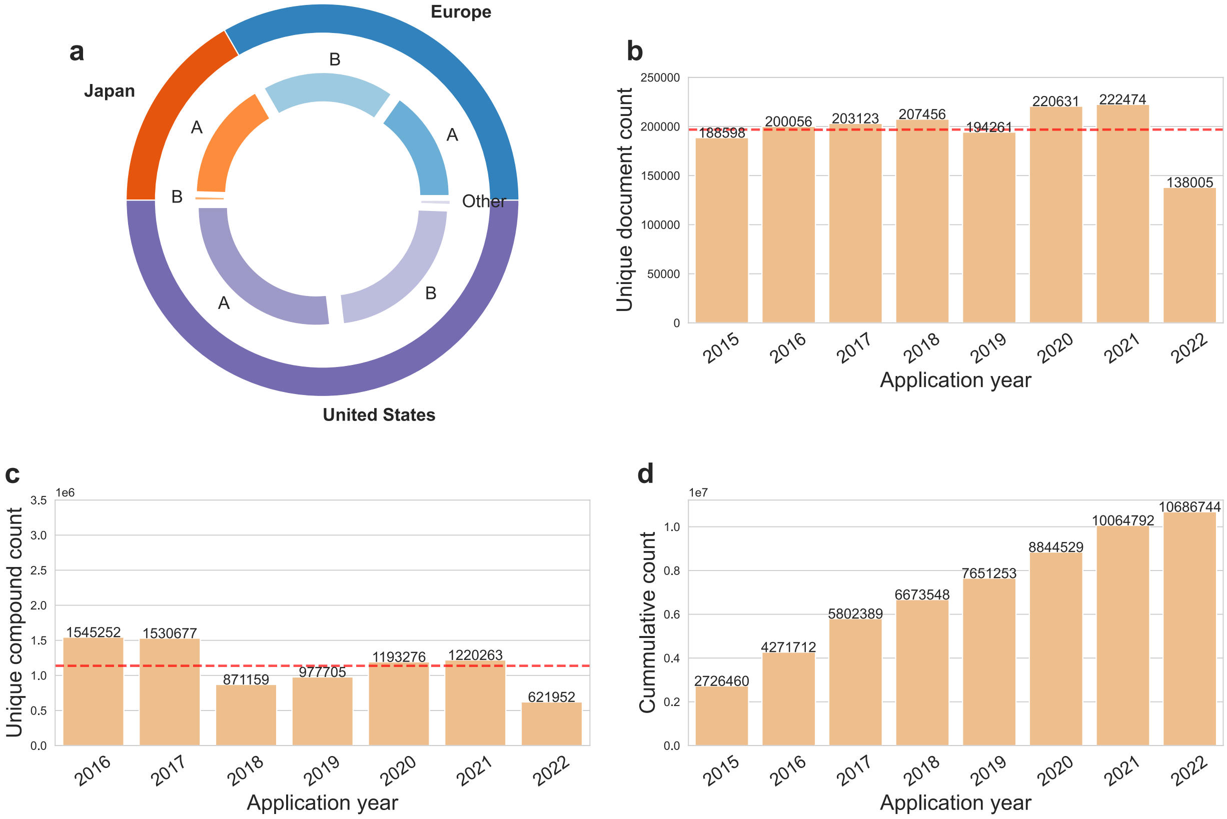


Fig. 1 (**a**) Distribution of patent document types by their patent kind codes across countries. (**b**) Distribution of patent documents filed over the application years. (**c**) Distribution of patent compounds over the application years. (**d**) Cumulative compound count over the years. The red line in subplots b and c indicates the average number of patents and compounds, respectively. Additionally, the counts displayed in subplots b and c are deduplicated counts for patents documents and compounds respectively.

Country	Patent kind code	Number of patent compounds
United States (US)	Filed (A)	6,111,699
United States (US)	Granted (B)	5,207,524
United States (US)	Design Patent (S)	2
United States (US)	Reissue Patent (E)	46,215
United States (US)	Plant Patent (P)	402
Europe (EP)	Filed (A)	2,648,617
Europe (EP)	Granted (B)	3,173,776
Japan (JP)	Filed (A)	254
Japan (JP)	Granted (B)	7

Table 1. Summary of the number of compounds found with respect to patent kind and country of filing. The compounds were counted based on their unique InChIKey representation in SureChEMBL.

reproducing variety of plant". To understand the proportion of these three different patent document types in SureChEMBL, we studied the patent kind codes for each of the registered patent documents in each jurisdiction individually. We identified large proportions of utility patent documents, including both filed (indicated by kind code *AX*) and granted patents (indicated by kind code *BX*) in each of the three countries (i.e. the United States, Europe and Japan), with the prior patent document type being predominant (Fig. 1a). Additionally, in the United States, we found the presence of a small proportion of "other" patent document classes, such as design patent documents (indicated by kind code *SX*), reissued patent documents (indicated by kind code *EX*) and plant patent documents (indicated by kind code *PX*) (summarized in Table 1).

Next, we investigated the distribution of patent documents and their compounds yearly. To do so, we collected all the patent documents and the patent compounds in SureChEMBL and distinguished them based on their application number and InChIKeys, respectively. On average, we found that 196,826 patent applications have been filed and patented each year (Fig. 1b,c). Moreover, an average of 6 compound references were

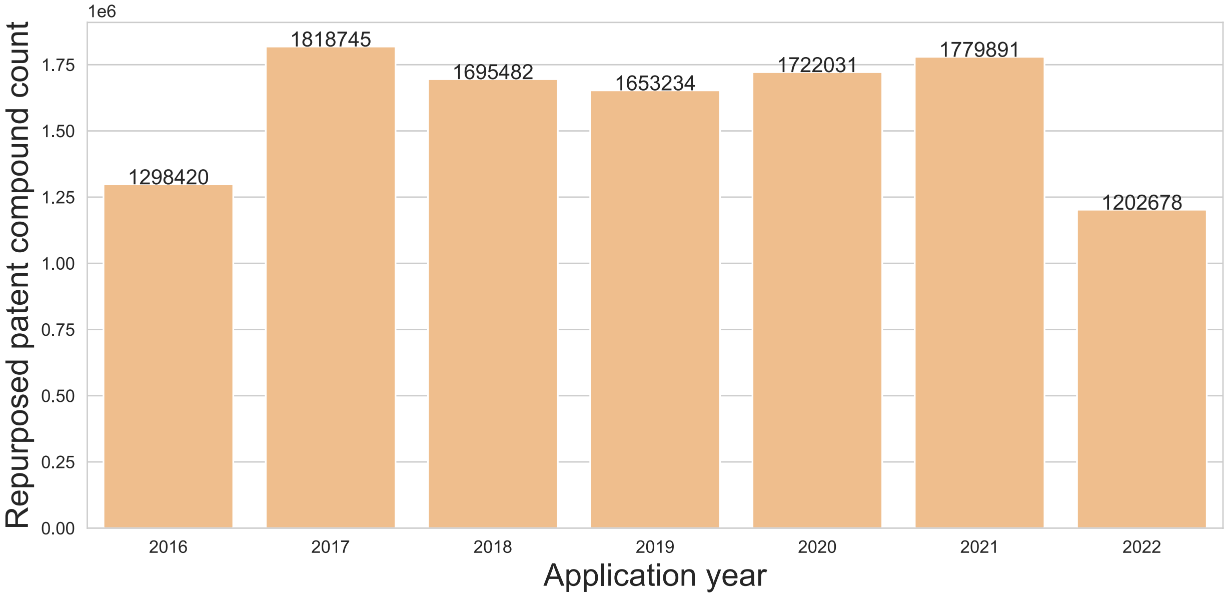


Fig. 2 Repurposed patent compound distribution over the application years. Distribution of patent compounds that have been found in patent documents from previous application years. It is to be noted that the repurposing scenario shown here is only considered from 2015 onwards.

identified per patent document in SureChEMBL. In addition to this, we examined the occurrence of patent compounds in more than one patent document. This analysis illustrated that, of the nearly 1 million patent compounds, 0.2% were associated with multiple patent documents (Fig. 2). A detailed investigation revealed that the majority (95%; 10,148,500 of 10,686,744) of these patent compounds appeared in fewer than 5 patent documents. On the other hand, 11,613 (of 10,686,744) patent compounds were found to be promiscuous across patent documents with their appearance in more than 1,000 documents each.

PubChem demonstrates highest coverage of patent compounds. A large number of compound-centric biological databases have been established in the past decades[28–30]. These databases have served various purposes in drug discovery, from identifying the bioactivity of unknown compounds[31,32] to the prediction of mechanisms of action[33,34], or simply for the virtual screening of drug candidates[35,36]. While previous research by Joerg Ohms (2022) explored the coverage of patent documents in relation to chemicals in two patent databases (SureChEMBL and Patentscope), the scope of this study was limited to manually comparing a set of chemicals between PubChem Substance and patent compounds[37]. Thus, to systematically understand the coverage of patent compounds in prominently used chemical databases, we analyzed the structural overlap between the compounds cited in patent documents and those found in three public chemical databases, namely PubChem, ChEMBL and DrugBank. The structural overlap was performed using the InChIKey representation of the compound in SureChEMBL against the three resources.

Upon identifying common compounds across these resources (Fig. 3a), two key findings were revealed. Firstly, only 0.02% (2,096 out of 10 million) of the patent compounds were eventually approved for one or more indication areas, according to data extracted from DrugBank, and secondly, PubChem retrieved compounds exhibited the highest overlap (91.5%) with the patent compounds from SureChEMBL. In contrast, ChEMBL demonstrated only a 0.1% overlap with patent compounds in SureChEMBL, an indicator that both resources occupy different chemical spaces. As illustrated in Fig. 3a, a small percentage (5.5%) of patent compounds were specific to the SureChEMBL database. Among these compounds, more than half were mined from US-based patent documents, while the remaining have been mined from EPO- or WIPO-based patent documents. Additionally, an examination of the annual count of SureChEMBL-specific compounds revealed a gradual decrease over time (Fig. 3b).

Images as the major source for compound annotation in patent documents. A patent document is a structured document containing sections including the title, abstract, description and claims[38]. Among these patent document sections, determining the location from which a compound was mined can provide insights into the correlation between the compound and the patent's applicability. For instance, if a compound was mentioned in the description section, it is likely to be associated with prior art (i.e. "referenced" compounds) relevant to the patent document. On the other hand, a compound mentioned in the claims section would likely pertain to the novel invention disclosed in the patent document.

In SureChEMBL, a patent document consists of four sections: title, abstract, description and claims. Along with these specific sections, chemical structure images and molfile (specifically restricted to patent documents collected from USPTO) serve as sources of compound annotation in SureChEMBL. Notably, these later sources (i.e. images and molfiles) were only annotated for patent applications after 2007[27]. Together, these six

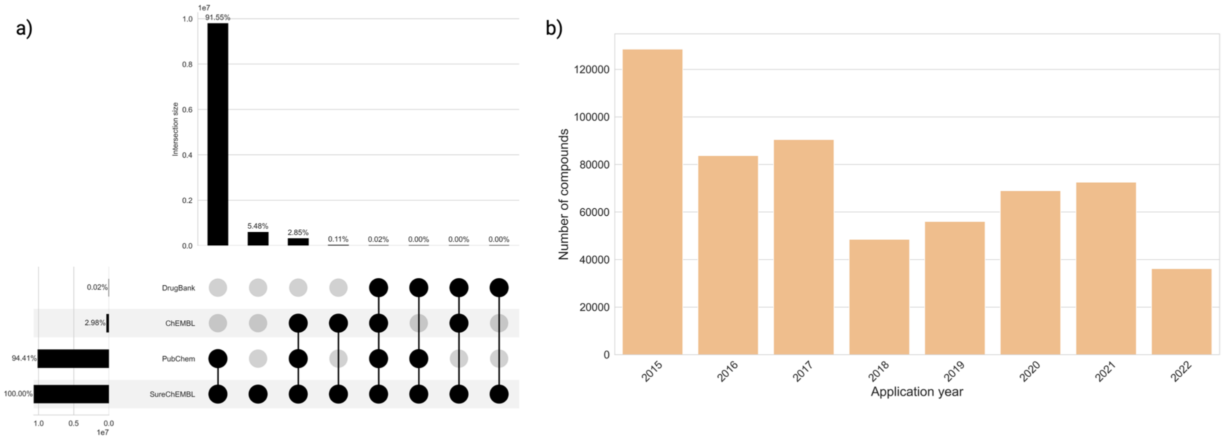


Fig. 3 (**a**) Distribution of patent compounds across four chemical resources, namely SureChEMBL, PubChem, ChEMBL and DrugBank. (**b**) Distribution of the proportion of patent compounds specific to SureChEMBL. The figure was formatted with BioRender.com.

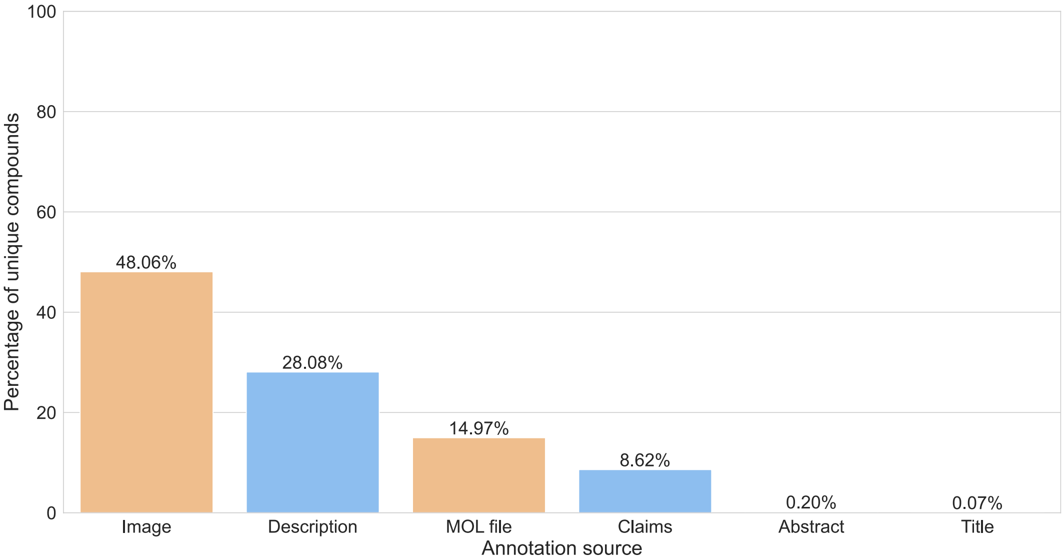


Fig. 4 Percentage of compounds annotated from the various patent document sources. Each bar in the figure corresponds to deduplicated compounds annotated specifically from the individual section of the patent document. The textual sections of a patent document (blue) are distinguished from the additional sources for annotation (orange) based on their colour.

sources of the patent document were utilized for biomedical entity annotations in SureChEMBL using numerous public and proprietary mining tools[27]. Hence, to provide an overview of the major sources surrounding the annotation of patent compounds, we investigated the sections frequently mined and annotated for compounds in SureChEMBL. We first calculated the average number of sources associated with patent documents in SureChEMBL. This analysis revealed that approximately 31.2% of patent compounds are found in more than one of the six sources. Next, we performed a thorough examination of the sources with regard to the compounds. We found that the description section (with ~28.08%) of the patent document was the major source of textual data involved in the extraction of patent compounds (Fig. 4). As illustrated in the figure, both the additional patent document sources (images with ~48% and molfiles with ~15%) were part of the top three sources for data annotations in SureChEMBL.

Over half of patent compounds show compliance with Ro5 framework.

To improve the efficiency of the drug discovery process, scientists have formulated guidelines or rules based on key determinant properties of compounds of drug likeness. Lipinski[39] and Veber *et al.*[40] provided the framework for the Rule of Five (Ro5), depending on physicochemical properties, to enhance the oral bioavailability of a compound[39–41]. Later, Doak *et al.* (2014), along with other researchers, extended the Ro5 for the oral bioavailability space of drugs, referring to it as the beyond Ro5 (bRo5) space[42–44]. The criteria of bRo5 supported the selection of cell-permeable

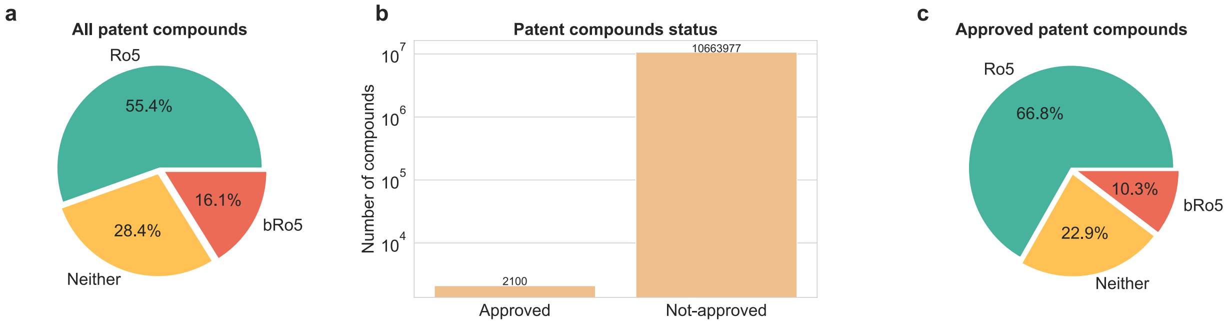


Fig. 5 Drug-like compliance of patent compounds. (**a**) Radial chart demonstrating the percentages of the Ro5 and bRo5 framework compliances of patent compounds. As shown, about 30% of patent compounds do not comply with either of the two frameworks. (**b**) Overview of the drug approval status of patent compounds. (**c**) Radial chart demonstrating the percentages of the Ro5 and bRo5 framework compliances of patent compounds that are approved drugs (i.e. 2,100 compounds).

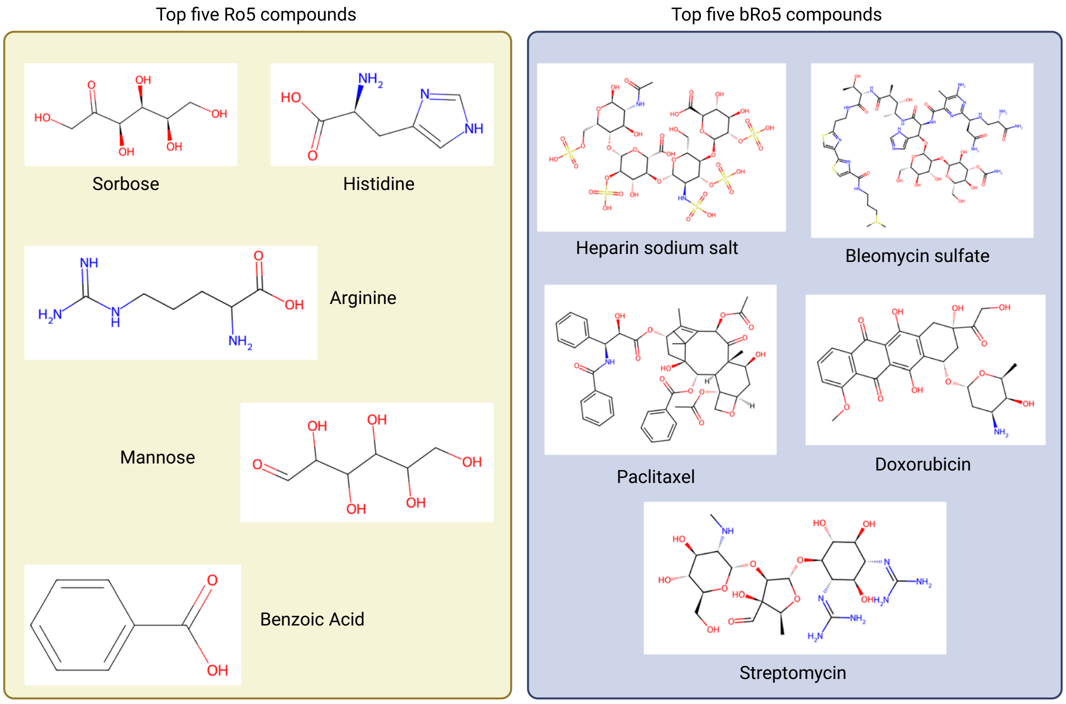


Fig. 6 Top five prevalent patent compounds in the Ro5 and bRo5 categories.

clinical candidates that demonstrated good pharmacokinetics (PK) and explored the "undruggable" targets, both of which could not have been possible previously with the Ro5 filtering.

To profile the drug-like space for the patent compounds, we explored the Ro5 and bRo5 space of these compounds. This analysis revealed that in the past decade, 55.46% of compounds complied with the Ro5 framework, and 16.11% compounds did so with the bRo5 (Fig. 5). The remaining compounds (28.43%) complied neither with the Ro5 nor bRo5 spaces. Next, we divided the compounds into two categories, approved and non-approved, based on the data in DrugBank. As mentioned in the previous sections, a very low number of patent compounds were approved. Moreover, a consistent trend was found, with more than half of the compounds complying with the Ro5 framework, among both approved and non-approved (Fig. 5).

Finally, we explored the most frequent patent compounds, focusing on the top five based on their prevalence in patent documents (Fig. 6). In the Ro5 class, traditional sugars such as Sorbose (SCHEMBL762) and Mannose (SCHEMBL1812) were found in over 200,000 and 150,000 patent documents, respectively. Essential amino acids like histidine (SCHEMBL3259) and arginine (SCHEMBL1790) were also prevalent, appearing in 150,000 to 200,000 patent documents. On the other hand, in the bRo5 class, we found heparin (SCHEMBL11557), a naturally occurring human metabolite, in more than 95,000 patent documents. Moreover, prominent drugs such as Paclitaxel (SCHEMBL3976), Bleomycin sulphate (SCHEMBL1599) and Doxorubicin (SCHEMBL3243), which are therapeutic drugs for treating cancer and antibacterial drug Streptomycin (SCHEMBL3276) were other patent compounds found in the bRo5 class with occurrence in about 90,000 patent documents. It is important

Year	Molecular weight	LogP	# HBA	# HBD	# RotB	# Rings	# Stereoisomers
2015	407.84	3.92	5.07	1.49	5.92	3.46	4.40
2016	437.60	4.29	5.36	1.55	6.31	3.84	5.05
2017	450.68	4.56	5.33	1.62	6.52	4.01	7.01
2018	451.07	4.61	5.35	1.53	6.52	4.02	6.57
2019	460.80	4.67	5.49	1.57	6.48	4.21	6.52
2020	471.94	4.90	5.56	1.55	6.52	4.42	6.47
2021	474.90	4.90	5.58	1.57	6.51	4.47	5.78
2022	498.41	5.43	5.70	1.50	6.71	4.94	6.37

Table 2. Physicochemical properties of patent compounds. For compounds found in each year, an average of the different molecular properties: (i) molecular weights, (ii) LogP, (iii) the number of hydrogen bond acceptors (# HBA), (iv) the number of hydrogen bond donors (# HBD), (v) the number of rotatable bonds (# RotB), (vi) the number of any ring (# Rings) and (vii) the number of stereoisomers were calculated.

to acknowledge that when dealing with thousands of patent documents associated with a compound, not all of them would be irrelevant. In fact, for a successfully approved active pharmaceutical ingredient (API) with potential, many follow-up patent applications may emerge. These could pertain to its synthesis, specific drug delivery system, novel indication area, or potential combination therapy with another ingredient. Moreover, APIs are frequently cited as prior art in patent documents, underscoring their significance in the pharmaceutical landscape.

Patent compounds show signs of enhanced chemical structural diversity.

Recently, PROteolysis-TArgeting Chimeras (PROTACS) have been identified as novel therapeutics with the potential to progress into clinics[45,46]. They achieve protein degradation by "hijacking" the cell's ubiquitin-proteasome system (UPS) and bringing together the target protein and an E3 ligase. Due to their non-adherent Ro5 characteristics, these molecules have not undergone "classical" prior optimization for oral bioavailability[47] and CNS penetration[48]. Additionally, in the past few years, interest has grown in the generation of macrocyclic compounds, those that retain the original scaffold or structure of existing compounds but contain additional functional groups or side chains, allowing for ring-shaped structures[49,50]. With the increasing interest in such compounds as potential clinical and drug candidates, we determined the physicochemical properties of patent compounds to check whether the chemical space expansion reflected PROTAC-like and macrocyclic compounds, among others, in recent years.

Table 2 summarizes the average physicochemical characteristics of patent compounds found annually. A gradual increase in characteristic molecular properties (i.e. molecular weight, the hydrogen bond donor and acceptors, LogP, and the number of rings) was observed. These properties, especially molecular weight, were nearing the upper limit of the Ro5 criteria. Specifically, the average molecular weight for patent compounds was between 400 and 500 Daltons, the average LogP was between 4 and 5, the average number of hydrogen bond acceptors and donors were below 6 and 2, respectively, and the average number of rotatable bonds was between 6 and 7. In addition, a consistent increase in the number of rings in patent compounds was also found in the past decade. Our analysis also uncovered some common findings with respect to molecular properties across potential clinical candidates. For example, we observed that over a million patent compounds exhibited with two or more chiral centers. This proportion supasses the number of compounds with only one chiral center. For either group, no enantiopurity is registered, so it is not possible to roughly assess how many enantioselective synthetic steps have been used. This finding aligns closely with recent research conducted by Scott *et al.*[51].

Additionally, we were interested in identifying bioactive compounds that could be covalent binders, showed existing polypharmacology, or were reactive in nature. To achieve this, one strategy involved examining the published biological activities and selectivities of the compounds. However, adopting this approach could lead to a very small subset of patent compounds (2,000–5,000), potentially yielding inconclusive results due to the existence of data silos surrounding the publication of biological data in patent documents. Thus, we used an alternative approach to understand the polypharmacological nature of patent compounds. This involved confirming the presence of Pan-Assay INterference Structures (PAINS), which are frequently used in drug discovery to flag and mark compounds that could cause interference in bioassays[44,52]. Consequently, such compounds that contain one or more PAINS alerts are usually removed during pre-clinical research due to their polypharmacological nature marking them as high risks for off-target effects.

In total, we identified 277 PAINS alerts in the patent compounds. This represents approximately 3.7% of all patent compounds in SureChEMBL that show the presence of at least one of these PAINS alerts (Fig. 7a). The most prominent of these PAINS alerts include the presence of azo compounds (Azo_a(324); 18.7%; 71,014 compounds), the presence of compounds involving one aniline and two alkyl groups with either an additional alkyl group (Anil_di_alk_a(478), 14%; 53,127 compounds) or an additional carbon (Anil_di_alk_c(246), 5.7%; 21,662 compounds), the presence of compound containing an indole, a phenyl and an alkyl group (Indol_3yl_alk(461), 8.5%; 32,290 compounds), and the presence of compounds containing catechol substructures (Catechol_a(92), 5.9%; 22,452 compounds). Moreover, aromatic PAINS like quinone also make it to the top of the list with about 5.2% (19,665) patent compounds. Interestingly, bioassay reagent materials like dyes (15,568 compounds) and Mannich bases (15,154 compounds) also appear in this list of PAINS alerts.

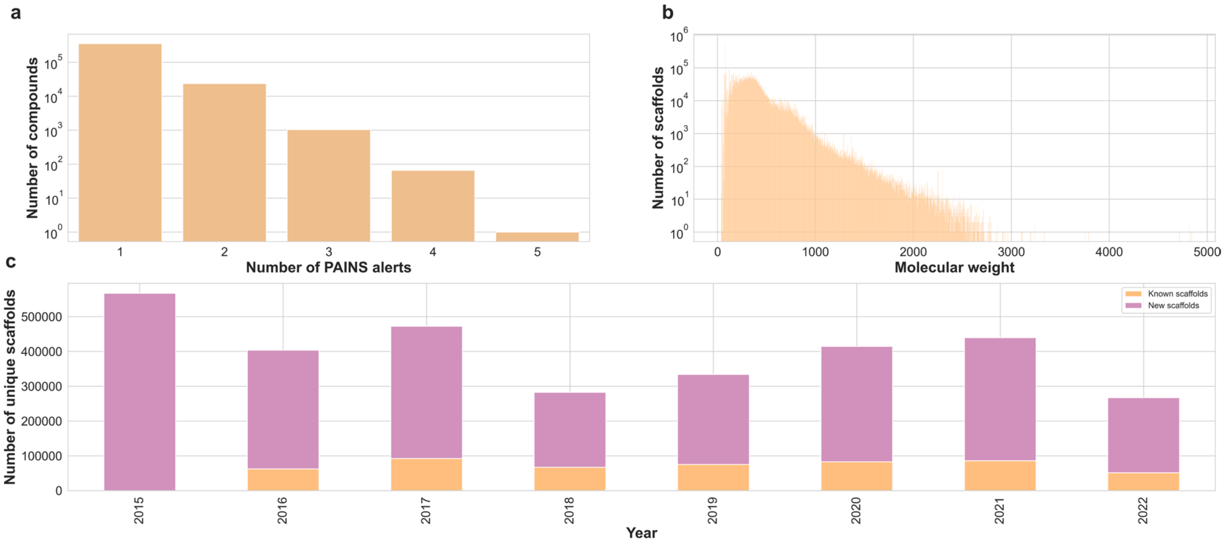


Fig. 7 (**a**) Count of PAINS alerts found to be associated with patent compounds. (**b**) Distribution of the molecular weight of the generic Murcko scaffold across the patent compounds. (**c**) Count of unique Murcko scaffolds found per year in patent documents. For each year, we distinguish the "known" scaffolds (orange) from "new" scaffolds (pink) based on the occurrence of the Murcko scaffold SMILES in previous years.

Patent Field	Number of scaffolds	Percentage of scaffold (%)
Abstract	8,120	0.19
Claims	315,413	7.73
Description	811,216	19.89
Image (after 2007)	2,261,009	55.44
MOL file (after 2007)	677,947	16.62
Title	4,494	0.11

Table 3. Summary of the number of unique scaffolds found in the individual patent document sections.

To conclude this exploration, we reduced the compounds to their generic Bemis-Murcko (BM) scaffold to quantify the scaffold diversity of patent compounds. The advantage of using a BM scaffold is two-fold: first, since these compounds are derived from patent documents, mapping them back to their original chemical definition would provide clues on how they were derived, and second, this representation retains the rings and side chains found in the compounds, unlike the graph framework that replaces all heteroatoms to carbon and collapses all bonds to single bonds notations[53]. This analysis revealed that 3 million distinct scaffolds encompass patent compounds in SureChEMBL. These scaffolds cover a broad range of molecular sizes, spanning from a compact molecule of 38.01 Daltons to a large molecule of 4841.19 Daltons (Fig. 7b). The year-wise comparison of the BM scaffold revealed that annually an average of 332,942 new generic scaffolds were patented (Fig. 7c). Moreover, trends showcasing a fluctuating number of scaffolds with a recent decline around the COVID-19 pandemic (2021–22) were seen. Tracing back the patent document source (i.e., the section or source from which the compound was annotated) revealed that more than half (55.44%) of the scaffolds were found to be associated with chemical images in patent documents, while only 19.89% were associated with the description section of patent document. Moreover, about 16.62% were found to be extracted from the claims section of the patent document (Table 3).

In this analysis, it is necessary to acknowledge that reducing the compounds to their generic BM scaffolds would result in the generation of promiscuous compounds like benzene or furan. These common scaffolds might not exactly be found in patent documents but would be part of a larger scaffold shown in these documents. Hence, our approach also unveiled a larger number of such common scaffolds. Figure 8 summarizes the top ten commonly identified scaffolds in patent compounds. As shown, the majority of these scaffolds have a single ring with heteroatoms causing it to weigh about a few hundred daltons. The most prominent ones include single cyclic scaffolds such as cyclohexane and tetrahydropyran or double-ring compounds like naphthalene and diphenylmethane.

Tracing a subset of approved drugs back to their patent documents. Pre-established regulations like the Patent Act of 1990 have aided drug proprietors in patenting novel pharmaceuticals or repurposing existing candidates to safeguard inventions under intellectual property laws before their integration into clinical practices[54,55]. Consequently, many disparities have arisen between the drug names present in patent documents

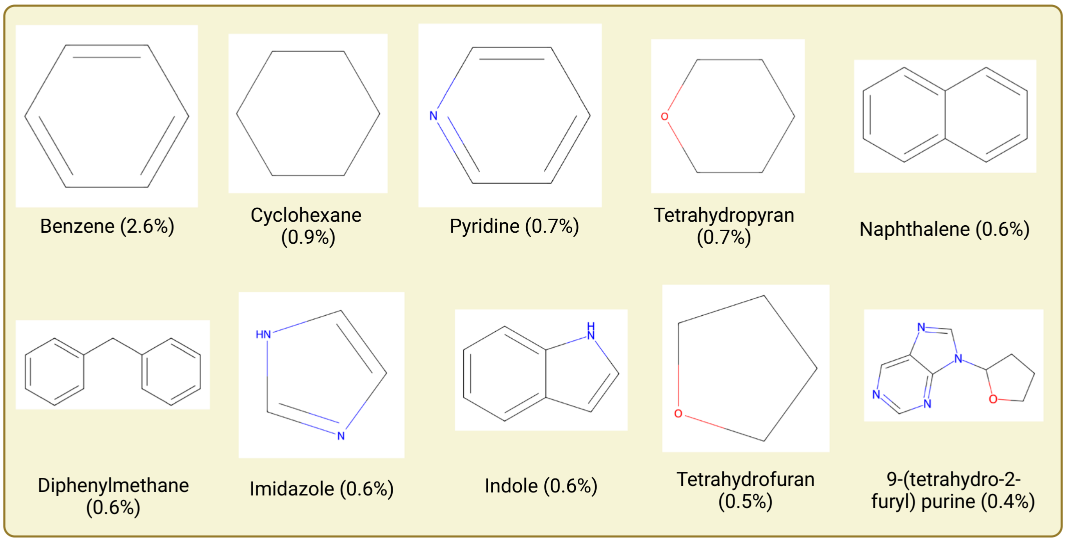


Fig. 8 The top Murcko scaffold of compounds and their respective frequency found in patent documents. The percentages only sum up to around 8.2% of the total patent compound scaffolds, a clear sign of the structural diversity within the patent chemical space.

and their corresponding brand name, posing a challenge in finding associated patent documents[56]. In the past, successful endeavours were made to link drugs to patent documents by the FDA's Orange Book and the World Intellectual Property Organization (WIPO) through their Pat-INFORMED tool[57]. Acknowledging the complexity of the drug nomenclature across the different stages of drug development, we leveraged the chemical representation (InChIKey) of patent compounds to generate an inventory of their corresponding clinical status in humans.

To do so, we started by looking into the intersection of the chemical space of patent compounds with investigational (i.e. drugs that have reached clinical trials) and withdrawn (i.e. drugs that have been discontinued) drugs in DrugBank. We found that only 3,235 of the ten million patent compounds have reached clinical trials, with a mere 0.0008% (85 compounds) falling under the withdrawn drugs category (Fig. 9a). In addition, databases such as ChEMBL enable identifying the drug research status (i.e., preclinical, clinical, and approved) of compounds, and hence we leverage this resource to identify the farthest research stage a patent compound traversed to in the drug discovery pipeline. Figure 9b depicts that compounds in patent documents are distributed across various research phases, ranging from preclinical to clinical, with the majority concentrated in the preclinical stage. Of these patent compounds, roughly 1% of the drugs were approved for one or more indication areas, according to ChEMBL. Furthermore, 1.6% of the patent compounds had no information (classified as "unknown" by ChEMBL) regarding their clinical stage and were likely to be lost during translation from patent documents to clinics or failed to be captured and annotated by the resource database. Furthermore, Phases 2 and 3 of clinical trials exhibited a greater proportion of patent compounds than Phase 1.

Discussion

Patent documents play an essential role in drug discovery and biological annotation pipelines, such as those that capture a molecule's image and convert it to SMILES, or those that highlight gene and disease names in the patent documents. This work focuses on patent compounds found in SureChEMBL, a patent database for life sciences, and assesses the annotation quality for drug-like molecules and drug discovery-related documents. To the best of our knowledge, this is the first systematic effort done in this direction with the entire database.

Initially, we started by inspecting the jurisdiction coverage of patent documents in SureChEMBL. As expected, countries such as the United States and Europe had the highest number of patent compounds. In contrast, a very low percentage of patent documents from Japan (through JPO) were present. This low number is confirmed by SureChEMBL, acknowledging their limited access to bibliographic information from Japanese patent documents (i.e. titles and abstracts) provided by the JPO[27,58]. Furthermore, challenges arise in converting Japanese text to English for the ingestion and storage of biomedical entities in SureChEMBL, thereby exacerbating the limitations. Previously, the absence of machine-translated full texts from which patent compounds could be extracted posed a bottleneck. However, recent advancements in integrated AI annotation within the resource offer potential mitigation for this issue in the future (https://www.ebi.ac.uk/about/news/technology-and-innovation/ai-annotations-increase-patent-data-in-surechembl/). Next, we looked into the quantitative aspect of data in SureChEMBL. A small fraction (0.2%) of compounds were identified in more than one patent document. This could entail the presence of either repurposing patent documents (i.e., a patent application on the reuse of known drugs for a different indication) or compounds annotated from utility patent documents (i.e., a patent application dealing with approved drugs in different pharmaceutical forms or route of administration for specific treatments).

In order to assess the ease with which patent compounds can be identified in the literature, we searched the compound structures across three major compound databases (i.e. PubChem, ChEMBL and DrugBank).

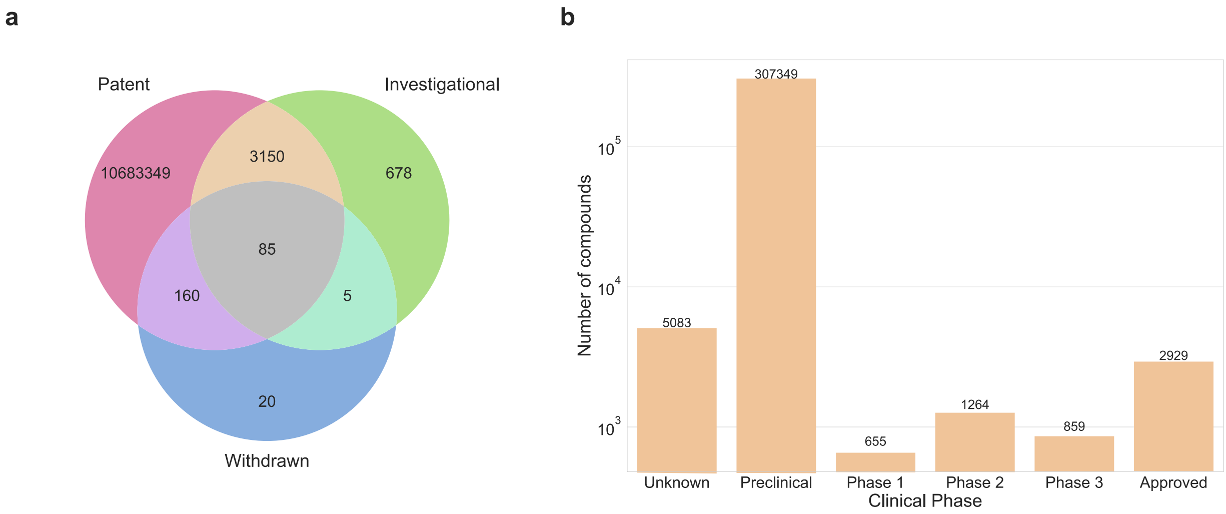


Fig. 9 (**a**) Euler figure of the chemical space across the patent compounds, investigational and withdrawn drugs found in DrugBank. (**b**) Distribution of patent compounds in the different clinical phases as per ChEMBL.

In this analysis, we found that a very low percentage of approved drugs (found in DrugBank) were a part of SureChEMBL. This low percentage could be attributed to three reasons: Firstly, patent documents are often crafted to encompass a broader chemical structure-activity landscape than the clinical candidates, thus safeguarding the candidates' secrecy[19,47]. This is typically achieved through the use of Markush structures in patent claims, allowing for coverage of a wide range of structural variations, including potential drug candidates that may still be unknown at the time of patent filing. Additionally, active pharmaceutical ingredients (APIs) are utilized to extend the applicability of the patent. Secondly, the changes in the nomenclature of the drug candidate as it progresses through the clinical pipeline, obscure its visibility. Despite SureChEMBL's cross-reference dictionary enabling the retrieval of patent compounds through multiple depictions (e.g., SMILES, IUPAC names, etc.), this limitation arises from inconsistencies in the names used by patent assignees or holders. Additionally, this could be due to the limited information provided by the patent assignees or holders in the patent document (a consequence of the former reason), thus hindering automated systems and pipelines (like SureChEMBL) from accurately recognizing the exact structure of the approved drug. Finally, the third limitation pertains to the use of DrugBank as a proxy for representing the approved drug space. DrugBank, being a commercial database, provides limited information for academic research. Moreover, our results also revealed an analogous chemical space being occupied between patent compounds in SureChEMBL and PubChem. This is unsurprising, given that PubChem leverages the SureChEMBL database to broaden its underlying chemical space. Furthermore, it's worth noting that in the near future, PubChem could accommodate additional patent compounds through its automated patent annotation pipeline. This pipeline establishes connections between compounds and relevant patent documents referenced in Google Patents[59]. Furthermore, a small proportion of compounds were not found in any public compound resources and were instead confined to SureChEMBL, indicating the presence of compounds from proprietary libraries used by patent assignees for drug discovery.

Following this, we investigated the major sources for compound annotation within SureChEMBL. These sources included known sections of patent documents (i.e. title, abstract, description and claims), images and MOL files. The description section was identified as the prominent textual source for compound annotation, highlighting that the annotated compound could be a part of the primary invention, whether it be related to its synthesis, formulation or application within a specific area. Moreover, it is essential to note that the text within the description section could also be too general, and include compounds such as assay buffers and reactants which are important for the compound stability or assay protocol but not necessarily the main scope of the patent document. Moreover, a small proportion (~15%) of compounds are also annotated from the MOL file, which is one of the basic files required for compound patent documents filed to the USPTO.

Additionally, we explored the chemical space of patent documents to delineate their structural diversity and drug-likeness space. Naturally, most of the compounds complied with the Ro5 framework for drug-likeness. This was further confirmed by looking at the underlying physicochemical distribution of patent compounds (as shown in Table 2). We also identified a small proportion of patent compounds outside the Ro5 space, falling into the bRo5 space; a similar trend was observed in the case of approved drugs in recent years[43]. Besides this, an increase in the physicochemical properties of the patent compounds was observed, which may have been due to multiple factors, including increasing interest in the development of PROTAC-like and signalling macrocyclic compounds (for e.g., cyclic peptides or cyclic kinase inhibitors)[60] or progress with regards to chemical synthesis capabilities[61,62]. Our exploration regarding the bioactivity of patent compounds led to the recognition of PAINS liable compounds in SureChEMBL. The presence of these assay-interfering compounds is not surprising given that a study by Capuzzi *et al.*[63] showed that FDA-approved drugs containing PAINS were more active than non-PAINS-containing drugs[63]. This has indeed led researchers like Senger *et al.*[64] to question the filtering of promiscuous compounds during the early drug discovery steps[64]. Nevertheless, the question of whether these

PAINS alert patent compounds are problematic (showing false positive results in assays) or innocuous (due to their possible polypharmacology activity) remains unsolved. From our perspective, this finding highlights the notion that the mere identification of compounds from patent documents is not sufficient to identify potential drug candidates. In the case of SureChEMBL, there is a need for performing a medicinal chemistry-oriented filtration to eliminate non-active or activity-interfering patent compounds prior to their downstream utility.

In a similar manner, we also explored the generic BM scaffolds of the patent compounds. This exploration showed a decline in the number of scaffolds generated in the past years. This could be attributed to several reasons, such as exceptional factors like the COVID-19-dependent blockade of some patenting activities in 2022, or structural reasons rooted within medicinal chemistry syntheses pipelines. Lastly, we concluded our analysis by addressing the drug discovery path of a patent drug by looking into its transition from a patent document to post-approval. Here we reported that of all the patent compounds found in ChEMBL, only 1% of the compounds were approved drugs. This is not surprising as an analysis by Brown[65] showed that hit compounds evolve through the drug discovery pipeline as they undergo structural modification that ensures their entry into clinics[65]. Moreover, a larger proportion of patent compounds were found in Phase 2 and Phase 3 than in Phase 1. This is attributed to the fact that these trial phases (i.e. Phases 2 and 3) typically yield a larger number of scientific publications, assuming the trials were successful[66]. In conclusion, our work provides a medicinal chemistry perspective on the chemical landscape formed by patent compounds, thus laying a foundation for the future utility of SureChEMBL. Furthermore, we believe that understanding the state of the art in terms of patent compounds and their scaffolds is crucial for enhancing innovation by exploring novel chemical space while minimizing the risks associated with inadvertently reusing chemical space for specific and undesired target indications.

We acknowledge certain limitations in our analysis that warrant attention. Firstly, our analysis relies on open-source data, which may introduce potential data quality issues. For instance, periodic updates of data sources could lead to temporary gaps, possibly resulting in inaccuracies in our analysis, particularly in areas such as the clinical status of compounds, as discussed in our study. Secondly, we assumed that all patent compounds in SureChEMBL are relevant to drug discovery, as they may pertain to either an indication area or drug formulation. However, this assumption may not always hold true, and it would be preferable to refine our analysis by focusing on a subset of patent IPC codes, as outlined by Gadiya *et al.*[67], to create a more drug discovery-specific patent document dataset. Also, the annotation source in SureChEMBL does not include the context from which the compounds were annotated in the patent document, at least in its data dump. This shortcoming makes it difficult to distinguish compounds that have been "referred" (i.e. prior art) in patent documents from those "claimed" (i.e. novelty). This has eventually led researchers to perform an additional layer of annotation on top of SureChEMBL-extracted patent documents[18,20]. Lastly, we briefly look into drug repurposing patent documents and recognize the potential value in identifying specific indication areas where drug repurposing patent documents are concentrated. This could offer insights into the similarity of MoA across different disease indications for certain compounds.

Discussions on the efficiency of patent documents for drug discovery have been raised numerous times, given their inability to disclose the drug candidate, thereby protecting the compounds' IP. This study aims to shed light on the utility of compound data in patent documents in the context of drug discovery. By leveraging SureChEMBL as the patent resource and untapping its chemical space, we open the avenue for the use of this resource for chemoinformatic-based models. For example, patent compounds could be used to extend and expand existing chemical datasets by enriching the structure-activity relationship landscape around the lead candidate. Such an approach could be a potential alternative to the generative AI-based approaches, provided that a patent document around the lead molecule exists and has been previously mapped. Hence, SureChEMBL has vast potential that has yet to be mined and leveraged for drug discovery purposes.

Methods

Collecting compound data from patent literature. We used SureChEMBL (https://www.surechembl.org/), an extensive publicly available patent compound data catalogue, as a source for patent documents and metadata[27]. We obtained all the tab-separated data files (.txt) from the FTP server of the resource (ftp://ftp.ebi.ac.uk/pub/databases/chembl/SureChEMBL/data/map/). The legacy data from 1994–2014 (identified by the file name SureChEMBL_map_20141231.txt.gz) had a different data format, lacking patent information, which complicated the transition from compounds to corresponding patent documents. Consequently, the legacy data was excluded from the analysis presented in the study. Ultimately, the compounds from 2015–2022 were utilized as the patent compound collection for this work.

Compound databases utilized for compound metadata annotation. To identify compounds annotated within SureChEMBL, we mapped them to three large independent compound data resources, namely PubChem (v.2023)[59], ChEMBL (v.32)[68], and DrugBank (v.5.1.10)[69]. The data from these resources was obtained either by querying the REST API service (as in the case of PubChem through PubchemPy), or by downloading a local data dump of the resources (as in the cases of DrugBank, via an academic licence, and ChEMBL, via the SQL database from their FTP server). A compound is said to be the same across two resources provided that an exact match of the InChIKey is present.

Moreover, we leveraged the clinical stage annotation of compounds in ChEMBL ("max phase") to annotate the clinical phase of a corresponding compound found in SureChEMBL. Additionally, DrugBank was used to validate the compounds that have been approved and distinguish them from those that had been withdrawn. It is worth noting that the commercial nature of DrugBank, while ensuring faster updates than public databases, might limit certain information to premium users.

Chemical space exploration using physicochemical properties of compounds. The chemical space of compounds was explored using two well-defined and established rules:

a) Ro5 - Extending Lipinski's Rule of Five (Ro5)[39,40], Veber *et al.*[40] added rotatable bonds (NRotB) and the topological polar surface area (TPSA) as additional features, proposing that compounds should have a TPSA < 140 and NRotB < 12 to enhance the probability of sufficient oral bioavailability[42].
b) beyond Ro5/bRo5 - According to Doak *et al.* (2016), compounds with $500 < \text{MW} < 3000$ daltons with at least one property beyond the *extended* Ro5 (i.e., LogP > 7.5 or LogP < 0, hydrogen bond donors (HBD) > 5, hydrogen bond acceptors (HBA) > 10, polar surface (TPSA) > 200, and rotatable bonds (NRotB) > 20 fall in this category[43].

A desalting step using RDKit was performed for the patent compound. In addition to these two rules, we also assessed the chemical space underlying patent documents by conducting a scaffold diversity assessment on compounds derived from patent documents. To accomplish this, we simplified the desalted compounds into their Murcko scaffolds, preserving the generic forms of ring components, linkers and side chains[44]. These Murcko scaffolds were then used to examine the occurrence of newly registered scaffolds on an annual basis. The generic Murcko scaffold for the patent compounds was generated using RDKit's "Scaffolds.MurckoScaffold.MurckoScaffoldSmiles()" function. We also identified PAINS alerts within the patent compounds. This was performed using the RDKit library using the codebase from the TeachOpenCADD tutorials[70].

Data availability

The data used in this study can be accessed on Zenodo[71]. The "Figures" directory consists of all the figures shown in this study. The "Mappings" directory consists of JSON serialized data dumps corresponding to the physiochemical properties, PAINS alerts and Murcko scaffolds for compounds found in SureChEMBL. The "Processed" directory involved the successful mapping of SureChEMBL compounds to external public databases like PubChem and ChEMBL. Finally, the "Raw" directory includes the combined original data dump of SureChEMBL from its FTP server (ftp://ftp.ebi.ac.uk/pub/databases/chembl/SureChEMBL/data/map/).

Code availability

The Python scripts and Jupyter notebooks supporting the conclusions of this study can be accessed and downloaded via GitHub (https://github.com/Fraunhofer-ITMP/patent-clinical-candidate-characteristics). The repository is structured into "Data" and "Notebook" sections. The "Data" section is the exact replica of the Zenodo dump mentioned previously. The "Notebook" section includes all the analyses presented in this study organized based on the sections within the results.

Received: 25 January 2024; Accepted: 13 May 2024;
Published online: 16 May 2024

References

1. Grabowski, H. G., DiMasi, J. A. & Long, G. The roles of patents and research and development incentives in biopharmaceutical innovation. *Health Affairs* **34**, 302–310 (2015).
2. Kesselheim, A. S., Sinha, M. S. & Avorn, J. Determinants of market exclusivity for prescription drugs in the United States. *JAMA Internal Medicine* **177**, 1658 (2017).
3. Dunn, M. K. Timing of patent filing and market exclusivity. *Nature Reviews. Drug Discover/Nature Reviews. Drug Discovery* **10**, 487–488 (2011).
4. Sayle, R. A., Petrov, P., Winter, J. & Mureșan, S. Improved chemical text mining of patents using infinite dictionaries, translation and automatic spelling correction. *Journal of Cheminformatics* **3** (2011).
5. Gadiya, Y., Gribbon, P., Hofmann-Apitius, M. & Zaliani, A. Pharmaceutical patent landscaping: A novel approach to understand patents from the drug discovery perspective. *Artificial Intelligence in the Life Sciences* **3**, 100069 (2023).
6. Kong, X. *et al.* STING as an emerging therapeutic target for drug discovery: Perspectives from the global patent landscape. *Journal of Advanced Research* **44**, 119–133 (2023).
7. Zhang, H. & Li, Y. The patent landscape of BRAF Target and KRAS Target. *Recent Patents on Anti-cancer Drug Discovery* **18**, 495–505 (2023).
8. Song, C. H., Han, J., Jeong, B. & Yoon, J. Mapping the patent landscape in the field of personalized medicine. *Journal of Pharmaceutical Innovation* **12**, 238–248 (2017).
9. Lahiry, S. R. & Rangarajan, K. Patent landscape for Indian biopharmaceutical sector: A Strategic insight. in *Flexible systems management* 31–47, https://doi.org/10.1007/978-981-10-8926-8_3 (2018).
10. Mucke, H. A. Intellectual property considerations. in *The Royal Society of Chemistry eBooks* 264–279, https://doi.org/10.1039/9781839163401-00264 (2022).
11. Strittmatter, S. M. Overcoming Drug Development Bottlenecks With Repurposing: Old drugs learn new tricks. *Nature Medicine* **20**, 590–591 (2014).
12. Senger, S. Assessment of the significance of patent-derived information for the early identification of compound–target interaction hypotheses. *Journal of Cheminformatics* **9** (2017).
13. Colen, L., Belderbos, R., Kelchtermans, S. & Leten, B. Many are called, few are chosen: the role of science in drug development decisions. *The Journal of Technology Transfer* https://doi.org/10.1007/s10961-022-09982-6 (2023).
14. Schmitt, V. J., Walter, L. & Schnittker, F. C. Assessment of patentability by means of semantic patent analysis – A mathematical-logical approach. *World Patent Information* **73**, 102182 (2023).
15. Fabry, B., Ernst, H., Langholz, J. & Koster, M. P. Patent portfolio analysis as a useful tool for identifying R&D and business opportunities—an empirical application in the nutrition and health industry. *World Patent Information* **28**, 215–225 (2006).
16. Grego, T., Pęzik, P., Couto, F. M. & Rebholz-Schuhmann, D. Identification of chemical entities in patent documents. in *Lecture notes in computer science* 942–949, https://doi.org/10.1007/978-3-642-02481-8_144 (2009).
17. Farre-Mensa, J., Hegde, D. & Ljungqvist, A. What Is a Patent Worth? Evidence from the U.S. Patent "Lottery". *The Journal of Finance* **75**, 639–682 (2019).
18. Falaguera, M. J. & Mestres, J. Identification of the core chemical structure in SUReCHEMBL patents. *Journal of Chemical Information and Modeling* **61**, 2241–2247 (2021).

19. Falaguera, M. J. & Mestres, J. Congenericity of claimed compounds in patent applications. *Molecules* **26**, 5253 (2021).
20. Kunimoto, R. & Bajorath, J. Exploring sets of molecules from patents and relationships to other active compounds in chemical space networks. *Journal of Computer-aided Molecular Design* **31**, 779–788 (2017).
21. Wagner, Ş., Sternitzke, C. & Walter, S. G. Mapping Markush. *Research Policy* **51**, 104597 (2022).
22. Deng, W., Berthel, S. J. & So, W. V. Intuitive patent Markush Structure Visualization tool for medicinal chemists. *Journal of Chemical Information and Modeling* **51**, 511–520 (2011).
23. Wills, T. J. & Lipkus, A. H. Structural approach to assessing the innovativeness of new drugs finds accelerating rate of innovation. *ACS Medicinal Chemistry Letters* **11**, 2114–2119 (2020).
24. Kim, J. & Lee, S. Patent databases for innovation studies: A comparative analysis of USPTO, EPO, JPO and KIPO. *Technological Forecasting & Social Change* **92**, 332–345 (2015).
25. Marco, A. C., Graham, S. & Apple, K. The USPTO Patent Assignment Dataset: Descriptions and Analysis. *Social Science Research Network* https://doi.org/10.2139/ssrn.2849634 (2015).
26. Hill, L. L. The Orange Book. *Nature Reviews. Drug Discovery* **4**, 621 (2005).
27. Papadatos, G. *et al.* SureChEMBL: a large-scale, chemically annotated patent document database. *Nucleic Acids Research* **44**, D1220–D1228 (2015).
28. Ferrence, G. M. *et al.* CSD Communications of the Cambridge Structural Database. *IUCrJ* **10**, 6–15 (2023).
29. Southan, C., Sitzmann, M. & Mureşan, S. Comparing the chemical structure and protein content of CHEMBL, DrugBank, Human Metabolome Database and the Therapeutic Target database. *Molecular Informatics* **32**, 881–897 (2013).
30. Ghani, S. S. A comprehensive review of database resources in chemistry. *Eclética Química* **45**, 57–68 (2020).
31. Tamura, S., Miyao, T. & Bajorath, J. Large-scale prediction of activity cliffs using machine and deep learning methods of increasing complexity. *Journal of Cheminformatics* **15** (2023).
32. Van Tran, T. T., Wibowo, A., Tayara, H. & Chong, K. T. Artificial intelligence in Drug toxicity Prediction: Recent advances, challenges, and future perspectives. *Journal of Chemical Information and Modeling* **63**, 2628–2643 (2023).
33. Lagunin, A. *et al.* CLC-Pred 2.0: a freely available web application for in silico prediction of human cell line cytotoxicity and molecular mechanisms of action for druglike compounds. *International Journal of Molecular Sciences* **24**, 1689 (2023).
34. Chen, W., Liu, X., Zhang, S. & Chen, S. Artificial intelligence for drug discovery: Resources, methods, and applications. *Molecular Therapy. Nucleic Acids* **31**, 691–702 (2023).
35. Bhattacharjee, A. K. Pharmacophore-based virtual screening of large compound databases can aid "big data" problems in drug discovery. in *Elsevier eBooks* 231–246, https://doi.org/10.1016/b978-0-323-85713-0.00014-1 (2023).
36. Almansour, N. M., Allemailem, K. S., Aty, A. A. A. E., Ismail, E. I. F. & Ibrahim, M. A. A. In Silico Mining of Natural Products Atlas (NPATLAS) database for identifying effective BCL-2 inhibitors: molecular docking, molecular dynamics, and pharmacokinetics characteristics. *Molecules* **28**, 783 (2023).
37. Ohms, J. Validity of PubChem compounds supplied by Patentscope or SureChEMBL. *World Patent Information* **70**, 102134 (2022).
38. Jessop, D., Adams, S. & Murray-Rust, P. Mining chemical information from open patents. *Journal of Cheminformatics* **3** (2011).
39. Lipinski, C. A. Drug-like properties and the causes of poor solubility and poor permeability. *Journal of Pharmacological and Toxicological Methods* **44**, 235–249 (2000).
40. Veber, D. F. *et al.* Molecular properties that influence the oral bioavailability of drug candidates. *Journal of Medicinal Chemistry* **45**, 2615–2623 (2002).
41. Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews* **64**, 4–17 (2012).
42. Doak, B. C., Zheng, J., Dobritzsch, D. & Kihlberg, J. How beyond rule of 5 drugs and clinical candidates bind to their targets. *Journal of Medicinal Chemistry* **59**, 2312–2327 (2015).
43. Bemis, G. W. & Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *Journal of Medicinal Chemistry* **39**, 2887–2893 (1996).
44. Baell, J. B. & Walters, M. A. Chemistry: Chemical con artists foil drug discovery. *Nature* **513**, 481–483 (2014).
45. Ermondi, G., Jiménez, D. G. & Sebastiano, M. R. Rational control of molecular properties is mandatory to exploit the potential of PROTACs as oral drugs. *ACS Medicinal Chemistry Letters* **12**, 1056–1060 (2021).
46. Jiménez, D. G. *et al.* Designing Soluble PROTACs: Strategies and preliminary guidelines. *Journal of Medicinal Chemistry* **65**, 12639–12649 (2022).
47. Ermondi, G., Jiménez, D. G. & Caron, G. PROTACs and building blocks: the 2D chemical space in very early drug discovery. *Molecules* **26**, 672 (2021).
48. Tashima, T. Proteolysis-Targeting Chimera (PROTAC) Delivery into the Brain across the Blood-Brain Barrier. *Antibodies* **12**, 43 (2023).
49. Xie, J. & Bogliotti, N. Synthesis and applications of Carbohydrate-Derived Macrocyclic Compounds. *Chemical Reviews* **114**, 7678–7739 (2014).
50. Zhao, Z. & Bourne, P. E. Rigid scaffolds are promising for designing macrocyclic kinase inhibitors. *ACS Pharmacology & Translational Science* **6**, 1182–1191 (2023).
51. Scott, K. A. *et al.* Stereochemical diversity as a source of discovery in chemical biology. *Current Research in Chemical Biology* **2**, 100028 (2022).
52. Chakravorty, S. J. *et al.* Nuisance compounds, PAINS filters, and dark chemical matter in the GSK HTS collection. *SLAS Discovery* **23**, 532–544 (2018).
53. Langdon, S. R., Brown, N. & Blagg, J. Scaffold diversity of exemplified medicinal chemistry space. *Journal of Chemical Information and Modeling* **51**, 2174–2185 (2011).
54. Malbon, J., Lawson, C. & Davison, M. The WTO Agreement on Trade-Related Aspects of Intellectual Property Rights: A Commentary. (Edward Elgar Publishing, 2014).
55. Motari, M. *et al.* The role of intellectual property rights on access to medicines in the WHO African region: 25 years after the TRIPS agreement. *BMC Public Health* **21** (2021).
56. Thakkar, K. & Billa, G. The concept of: Generic drugs and patented drugs vs. brand name drugs and non-proprietary (generic) name drugs. *Frontiers in Pharmacology* **4** (2013).
57. SCHULTZ, M. Pat-INFORMED: A new tool for drug procurement. *WIPO MAGAZINE* 30–36 (2018).
58. Senger, S., Bartek, L., Papadatos, G. & Gaulton, A. Managing expectations: assessment of chemistry databases generated by automated extraction of chemical structures from patents. *Journal of Cheminformatics* **7** (2015).
59. Kim, S. *et al.* PubChem 2023 update. *Nucleic Acids Research* **51**, D1373–D1380 (2022).
60. Guo, Y. *et al.* An Integrated Strategy for Assessing the Metabolic Stability and Biotransformation of Macrocyclic Peptides in Drug Discovery toward Oral Delivery. *Analytical Chemistry* **94**, 2032–2041 (2022).
61. Münzfeld, L. *et al.* Synthesis and properties of cyclic sandwich compounds. *Nature* **620**, 92–96 (2023).
62. Gao, X. *et al.* Enantioselective Synthesis of Chiral Medium-Sized Cyclic Compounds via tandem Cycloaddition/Cope Rearrangement Strategy. *ACS Catalysis* **9**, 1645–1654 (2019).
63. Capuzzi, S. J., Muratov, E. & Tropsha, A. Phantom PAINS: Problems with the Utility of Alerts for Pan-Assay INterference CompoundS. *Journal of Chemical Information and Modeling* **57**, 417–427 (2017).

64. Senger, M. R., Fraga, C. A. M., Dantas, R. F. & Silva, F. P. Filtering promiscuous compounds in early drug discovery: is it a good idea? *Drug Discovery Today* **21**, 868–872 (2016).
65. Brown, D. G. An analysis of successful Hit-to-Clinical Candidate pairs. *Journal of Medicinal Chemistry* **66**, 7101–7139 (2023).
66. Cuschieri, S. Clinical trial publications. *Saudi Journal of Anaesthesia* **13**, 42 (2019).
67. Gadiya, Y., Zaliani, A., Gribbon, P. & Hofmann-Apitius, M. PEMT: a patent enrichment tool for drug discovery. *Bioinformatics* **39** (2022).
68. Gaulton, A. *et al*. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research* **40**, D1100–D1107 (2011).
69. Wishart, D. S. *et al*. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research* **46**, D1074–D1082 (2017).
70. Sydow, D., Morger, A., Driller, M. & Volkamer, A. TeachOpenCADD: a teaching platform for computer-aided drug design using open source packages and data. *Journal of Cheminformatics* **11** (2019).
71. Gadiya, Y. Dataset for manuscript titled "Exploring SureChEMBL from a drug discovery perspective". *Zenodo (CERN European Organization for Nuclear Research)* https://doi.org/10.5281/zenodo.10210061 (2023).

Acknowledgements

We would like to thank the authors of the public resources used in our work for making their datasets available to the scientific community. The REMEDi4ALL project has received funding from the European Union's Horizon Europe Research & Innovation programme under grant agreement No. 101057442. This work reflects only the authors' view, and the EC is not responsible for any use that may be made of the information it contains. We would also like to thank Dr. Sarah Mubeen for helping us improve the language and content of the manuscript.

Author contributions

Y.G. conceived the work. Y.G. and A.Z. contributed to the ideation. Y.G. and S.S. performed the analysis. M.H.A., P.G., Y.G. and A.Z. have written the manuscript. All the authors have reviewed and approved the final manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Y.G.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

A.4 PEMT: a patent enrichment tool for drug discovery

Reprinted with permission from "Gadiya, Y., Zaliani, A., Gribbon, P., and Hofmann-Apitius, M. (2023). PEMT: a patent enrichment tool for drug discovery. *Bioinformatics*, 39(1), btac716.".

Bioinformatics, 39(1), 2023, btac716
https://doi.org/10.1093/bioinformatics/btac716
Advance Access Publication Date: 2 November 2022
Applications Note

Data and text mining

PEMT: a patent enrichment tool for drug discovery

Yojana Gadiya [1,2,*], **Andrea Zaliani**[1,2], **Philip Gribbon**[1,2] and **Martin Hofmann-Apitius** [3,4]

[1]Fraunhofer Institute for Translational Medicine and Pharmacology (ITMP), Hamburg 22525, Germany, [2]Fraunhofer Cluster of Excellence for Immune-Mediated Diseases (CIMD), Frankfurt 60590, Germany, [3]Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Sankt Augustin 53754, Germany and [4]Bonn-Aachen International Center for Information Technology (B-IT), University of Bonn, Bonn 53113, Germany

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on July 26, 2022; revised on October 10, 2022; editorial decision on October 31, 2022; accepted on November 1, 2022

Abstract

Motivation: Drug discovery practitioners in industry and academia use semantic tools to extract information from online scientific literature to generate new insights into targets, therapeutics and diseases. However, due to complexities in access and analysis, patent-based literature is often overlooked as a source of information. As drug discovery is a highly competitive field, naturally, tools that tap into patent literature can provide any actor in the field an advantage in terms of better informed decision-making. Hence, we aim to facilitate access to patent literature through the creation of an automatic tool for extracting information from patents described in existing public resources.

Results: Here, we present PEMT, a novel patent enrichment tool, that takes advantage of public databases like ChEMBL and SureChEMBL to extract relevant patent information linked to chemical structures and/or gene names described through FAIR principles and metadata annotations. PEMT aims at supporting drug discovery and research by establishing a patent landscape around genes of interest. The pharmaceutical focus of the tool is mainly due to the subselection of International Patent Classification codes, but in principle, it can be used for other patent fields, provided that a link between a concept and chemical structure is investigated. Finally, we demonstrate a use-case in rare diseases by generating a gene-patent list based on the epidemiological prevalence of these diseases and exploring their underlying patent landscapes.

Availability and implementation: PEMT is an open-source Python tool and its source code and PyPi package are available at https://github.com/Fraunhofer-ITMP/PEMT and https://pypi.org/project/PEMT/, respectively.

Contact: yojana.gadiya@itmp.fraunhofer.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Patents are an untapped source of scientific information which nevertheless play a vital role in reflecting the progress of organizations involved in drug discovery. Scientific content contained within patent documents can be overlooked due to the lengthy process involved in making the patent public, when compared with scientific publications which can be readily accessible to the community via pre-print servers. Moreover, the legal jargon used to describe the claims of the patent applications also introduces additional complexities in data analysis when compared with classical scientific publications. However, analysis of patents can be key in assessing and reviewing, for instance, a companies' disease strategy (Roskams-Edris *et al.*, 2019) or making well-informed target selection and prioritization decisions (Jin and Wong, 2014).

Patent literature provides a different view of drug discovery by focusing on industry-specific genes or targets rather than scientific literature (Mucke, 2021). Databases such as PATENTSCOPE (https://www.wipo.int/patentscope/en/) and Espacenet (https://worldwide.espacenet.com/) specifically provide search and retrieval capabilities that cater to the specialized structured format of patent applications (Donald *et al.*, 2018). Despite the intrinsic value in analysing all relevant patent data in a project, there is often a limit on the number of patents individual users can retrieve within any period of time. In addition, many existing resources have associated access fees, which may not be affordable for academia and small-scale companies. This situation gives rise to a need for open-source resources for analysing patent content.

Here, we present patent enrichment tool (PEMT), an automated patent extraction tool that takes advantage of information from patent databases and connects genes or chemicals to this information retrospectively. While patent officers are specifically trained in legal aspects of patent analysis, drug researchers have few open-source tools that enable them to perform a qualitative landscape evaluation on genes linked to certain diseases. PEMT is designed to provide them with such a tool. Furthermore, we demonstrate the applicability of this tool in the rare disease domain and show how the tool can assist the scientific community in exploring this untapped resource.

2 Materials and methods

2.1 Implementation details

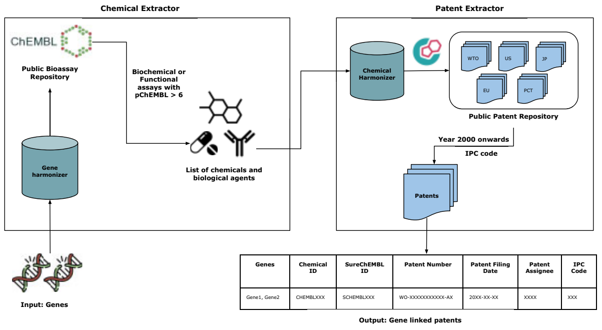
PEMT is written in version-controlled software with Python 3.8 and can be found at PyPi (https://pypi.org/project/PEMT/) and on GitHub (https://github.com/Fraunhofer-ITMP/PEMT).

2.2 Data harmonization

PEMT deals with three types of entities: patents, genes and modulators, such as chemicals or biological agents. Each of these entities are harmonized by mapping them to one or more well-known identifiers using cross-references available from existing biological databases. For genes, we leveraged the HUGO Gene Nomenclature Committee (Povey *et al.*, 2001) database that enabled easy conversion of gene symbols or names to UniProt identifiers (UniProt Consortium, 2015). Similarly, to map gene identifiers to ChEMBL identifiers, and ChEMBL modulators to SureChEMBL identifiers (Papadatos *et al.*, 2016), we made use of ChEMBL's cross-reference system (Gaulton *et al.*, 2012). Overall, genes were represented with UniProt and ChEMBL identifiers, modulators with ChEMBL and SureChEMBL identifiers and patents with application numbers. The reasoning behind the selection of the abovementioned resources is discussed in Supplementary Text S1. This harmonization step served two main purposes: first, it increased the alignment of the underlying data resources with FAIR principles and second, it allowed for data extraction from multiple resources in an efficient manner.

2.3 Design architecture

PEMT takes a two-step approach to collect relevant patent documents (Fig. 1). In the first step, chemical and biological modulators that directly regulate (i.e. activation or inhibition) specified genes of interest are extracted. For each gene of interest, a gene harmonization step described in the previous section is performed. Once we have the corresponding ChEMBL IDs for the gene, we query ChEMBL to identify experimentally validated compounds for the gene. Since ChEMBL has multiple approaches for flagging experimental data, we restricted the identified compounds to have binding or functional activity on the gene (Supplementary Text S2). Thus, the chemical extractor stage generated a pre-filtered and *in vivo* validated list of chemical and biological modulators for the gene.



Genes	Chemical ID	SureChEMBL ID	Patent Number	Patent Filing Date	Patent Assignee	IPC Code
Gene1, Gene2	CHEMBLXXX	SCHEMBLXXX	WO-XXXXXXXXXXX-AX	20XX-XX-XX	XXXX	XXX

Fig. 1. Overview of the framework of the PEMT

In the second stage, interlinking of identified modulators to patent documents is done by systematically querying SureChEMBL, a patent database. In order to extract relevant patents, first the chemical and biological modulators from the previous stage are harmonized to SureChEMBL IDs with the chemical harmonization. This is followed by querying the SureChEMBL database to extract all corresponding patent documents that reference the modulator. Simultaneously, a patent class-based filtering was performed to retrieve patent documents playing a role in essential chemical or biological roles (Supplementary Text S3). PEMT recognizes each patent document as unique based on its patent number which consists of a country code (e.g. WO/US/EP), a 7–11 digit number and the patent grant status number (e.g. A1 and B2) each separated by a dash (-). Together, these two steps facilitate the linking of patent documents to genes of interest via chemicals or biological modulators.

3 Case scenario

3.1 Data generation

Orphanet (www.orpha.net) is a service catalog of rare disease data which include the Orphanet Ontology of Rare Diseases, as well as information on the epidemiological occurrence and the biological mechanisms (disease–gene interactions) involved in rare diseases (Weinreich *et al.*, 2008). We made use of two data categories from Orphanet, that is, epidemiological data and biological mechanisms, to demonstrate the applicability of PEMT.

As a starting point, we selected a subset of genes based on their corresponding disease prevalence from the current list of 3886 diseases and 2637 genes available in Orphanet. We also filtered out the diseases with an incidence rate lower than 10. This selection criterion resulted in 59 diseases and 56 genes, representing the most common of the rare diseases, where significant patenting activity would be reasonably anticipated to have occurred in the past. These 56 genes were then provided as inputs for the PEMT tool.

3.2 Data analysis

PEMT generated connections between genes to patents via modulators. Based on publicly available bioassays, only 12 genes were associated with modulators (Supplementary Fig. S1). The number of agents ranged between 115 (for fibroblast growth factor receptor 3 (FGFR3)) and 1 (for methyl-CpG binding protein 2 (MECP2)) and a total of 469 were extracted from this step. A further reduction in the chemical space (of the 469 chemicals) was performed, based on their patentability, resulting in 143 modulators. From these 143 modulators, only 97 were identified to be linked to pharmaceutically relevant patents (Supplementary Table S2). Ultimately, 135 unique patents belonging to 10 genes (2 did not have any patent information connected to them), linked via 97 modulators, were extracted using the PEMT tool (Supplementary Fig. S2).

Patents covered two geographical jurisdictions: 83 from the USA (United States Patent and Trademark Office (USPTO)) and 19 from the European Patent Office (EPO). Furthermore, 106 patents are not yet granted while 29 patents have already been granted (Supplementary Table S3). Of the 106 non-granted patents, we found that 33 patents were from the World Intellectual Property Organization which includes patent applications but does not grant them, unlike EPO or USPTO who have patent granting authority. We also found that these patents cover a range of different cooperative patent classification codes: C07D (82 patents), C07K (18 patents), C12Q (14 patents), C07H (4 patents), A61P (9 patents), C12P (3 patents), C07F (5 patents) and C07C (1 patent) and are distributed over the years 2000–2021 (Supplementary Fig. S3). Moreover, upon assessing the patent assignees, we found 85 patents belonged to pharmaceutical industries, 23 patents to academic institutes and 25 patents to individuals.

Lastly, we looked at the historical timeline for each patented gene to understand the genes' importance over the patent landscape (Supplementary Fig. S4). FGFR3, one of the genes that received major attention between 2002 and 2008, was found to be associated with multiple diseases. In the beginning of this time period, FGFR3 was found to be associated with Muenke syndrome which had no interventional clinical studies, leading to a peak in patenting activity. Despite these efforts, only a single completed clinical trial in 2005 (NCT00106977) was performed that aimed at better understanding disease aetiology. Later, in 2004, it was found that FGFR3 played a role in carcinomas, such as multiple myeloma and bladder carcinoma (Grand *et al.*, 2004) and interest within the industry to understand the effect of this target in carcinoma remained. Nonetheless, only a marginal efficacy was demonstrated in this case as well (Chae *et al.*, 2017) and in the following years, interest in the target began to wane. Conversely, other targets, including SLC2A3 began surfacing in patents in 2015 and contributed to the highest number of patents, primarily due to their association with Huntington's disease, a prominent neurodegenerative disease.

4 Discussion and future work

The PEMT is an open-source and easy-to-use tool providing the scientific community with the opportunity to extract pharmaceutically relevant patent documents for genes of interest. This efficient linking of genes to patents could potentially open a link to enrich existing biomedical knowledge graphs with unmined patent literature. Moreover, PEMT can help in providing an overview of the target prioritization shift over time. Hence, with the PEMT, we hope to enhance the accessibility and applicability of patent literature within the scientific community.

Acknowledgements

The authors thank Daniel Domingo-Fernández, Sarah Mubeen and the anonymous reviewers for their comments and suggestions on the manuscript.

Funding

This work was supported by the German Federal Ministry of Education and Research (BMBF) [01GM2003A] (project 'TreatKCNQ').

Conflict of Interest: The authors have declared that no conflict of interests exists.

References

Chae,Y.K. *et al.* (2017) Inhibition of the fibroblast growth factor receptor (FGFR) pathway: the current landscape and barriers to clinical application. *Oncotarget*, **8**, 16052–16074.

Donald,K.E. *et al.* (2018) Tips for reading patents: a concise introduction for scientists. *Exp. Opin. Ther. Pat.*, **28**, 277–280.

Gaulton,A. *et al.* (2012) ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.*, **40**, D1100–D1107.

Grand,E.K. *et al.* (2004) Targeting FGFR3 in multiple myeloma: inhibition of t (4; 14)-positive cells by SU5402 and PD173074. *Leukemia*, **18**, 962–966.

Jin,G. and Wong,S.T. (2014) Toward better drug repositioning: prioritizing and integrating existing methods into efficient pipelines. *Drug Discov. Today*, **19**, 637–644.

Mucke,H.A. (2021) What patents tell us about drug repurposing for cancer: a landscape analysis. *Semin. Cancer Biol.*, **68**, 3–7.

Papadatos,G. *et al.* (2016) SureChEMBL: a large-scale, chemically annotated patent document database. *Nucleic Acids Res.*, **44**, D1220–D1228.

Povey,S. *et al.* (2001) The HUGO gene nomenclature committee (HGNC). *Hum. Genet.*, **109**, 678–680.

Roskams-Edris,D. *et al.* (2019) Medical methods patents in neuromodulation. *Neuromodulation*, **22**, 398–402.

UniProt Consortium. (2015) UniProt: a hub for protein information. *Nucleic Acids Res.*, **43**, D204–D212.

Weinreich,S.S. *et al.* (2008) Orphanet: a European database for rare diseases. *Ned. Tijdschr. Geneeskd.*, **152**, 518–519.

A.5 Pharmaceutical patent landscaping: A novel approach to understand patents from the drug discovery perspective

Reprinted with permission from “Gadiya, Y., Gribbon, P., Hofmann-Apitius, M., and Zaliani, A. (2023). Pharmaceutical patent landscaping: A novel approach to understand patents from the drug discovery perspective. *Artificial Intelligence in the Life Sciences*, 3, 100069.”.

Artificial Intelligence in the Life Sciences 3 (2023) 100069

Contents lists available at ScienceDirect

Artificial Intelligence in the Life Sciences

journal homepage: www.elsevier.com/locate/ailsci

Pharmaceutical patent landscaping: A novel approach to understand patents from the drug discovery perspective

Yojana Gadiya[a,b,c,*], Philip Gribbon[a,b], Martin Hofmann-Apitius[c,d], Andrea Zaliani[a,b]

[a] Fraunhofer Institute for Translational Medicine and Pharmacology (ITMP), Schnackenburgallee 114, Hamburg 22525, Germany
[b] Fraunhofer Cluster of Excellence for Immune-Mediated Diseases (CIMD), Theodor Stern Kai 7, Frankfurt 60590, Germany
[c] Bonn-Aachen International Center for Information Technology (B-IT), University of Bonn, Bonn 53113, Germany
[d] Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Schloss Birlinghoven, Sankt Augustin 53757, Germany

ARTICLE INFO

Keywords:
Patent analysis
Drug discovery
Patent landscape
Rare disease
Alzheimer's disease
Drug repurposing
Target importance

ABSTRACT

Patents play a crucial role in the drug discovery process by providing legal protection for discoveries and incentivising investments in research and development. By identifying patterns within patent data resources, researchers can gain insight into the market trends and priorities of the pharmaceutical and biotechnology industries, as well as provide additional perspectives on more fundamental aspects such as the emergence of potential new drug targets. In this paper, we used the patent enrichment tool, PEMT, to extract, integrate, and analyse patent literature for rare diseases (RD) and Alzheimer's disease (AD). This is followed by a systematic review of the underlying patent landscape to decipher trends and applications in patents for these diseases. To do so, we discuss prominent organisations involved in drug discovery research in AD and RD. This allows us to gain an understanding of the importance of AD and RD from specific organisational (pharmaceutical or university) perspectives. Next, we analyse the historical focus of patents in relation to individual therapeutic targets and correlate them with market scenarios allowing the identification of prominent targets for a disease. Lastly, we identified drug repurposing activities within the two diseases with the help of patents. This resulted in identifying existing repurposed drugs and novel potential therapeutic approaches applicable to the indication areas. The study demonstrates the expanded applicability of patent documents from legal to drug discovery, design, and research, thus, providing a valuable resource for future drug discovery efforts. Moreover, this study is an attempt towards understanding the importance of data underlying patent documents and raising the need for preparing the data for machine learning-based applications.

1. Introduction

Patent documents are considered crucial assets within the drug discovery domain as they allow the inventor rights over an invention for typically 20 years post filing. In the biomedical domain, these inventions could include information on drug formulation, dosage or efficacy, as well as information on the medicinal chemistry properties of leads or pre-clinical candidates. Moreover, the legal value of intellectual property in patent documents, and their distinctive function compared to scientific literature, makes patents crucial milestones in drug discovery and development [1]. Despite this, patent documents have been relatively untapped as a source to assist in scientific discovery and are used in the latter parts of drug discovery to assist investors in drafting new patents for their innovation or extension of existing patents in case of "me-too" drugs [2]. Also, the use of legal phraseology within patent documents requires experts in both patent drafting and patent analysis to evaluate and understand the underlying content. As a result, patent literature mining has developed into a specialised field that is closely linked with chemoinformatic analysis and commercial evaluation, meaning the appearance of patent sourced information is less common in publicly available scientific data resources. In the current context, we use the word "patent" to indicate both patent applications and granted patents.

Recently, there has been an increase in the attention given to patent documents as an aid to monitor advancements in drug discovery. This is seen in the field of oncology drug discovery, where patenting activity is intense, and patents offer a window into the latest techniques in translational cancer therapies [3–5]. To support these efforts, oncology-specific patent datasets have been catalogued by established patent offices such as the USPTO Cancer Moonshot Patent Data (https://www.uspto.gov/ip-policy/economic-research/research-datasets/cancer-moonshot-patent-data). Moreover, there have been efforts in mining and ingesting patent-related information into knowledge

[*] Corresponding authors at: Fraunhofer Institute for Translational Medicine and Pharmacology (ITMP), Schnackenburgallee 114, Hamburg 22525, Germany.
E-mail address: Yojana.Gadiya@itmp.fraunhofer.de (Y. Gadiya).

https://doi.org/10.1016/j.ailsci.2023.100069
Received 10 February 2023; Received in revised form 29 March 2023; Accepted 29 March 2023
Available online 31 March 2023

graphs. One such example is the Chinese patent medicine (CPM) knowledge graph, where a natural language model (BERT) was used to extract biological entities (specifically chemicals, diseases, and conditions) from patent documents [6]. The resultant graph assisted in the identification of disease gaps as well as summarising drug prescription trends in China.

In addition to the extraction of patent-related information, the importance of exploring patent literature has attracted new market interest from scientific information vendors. The value of an exhaustive patent exploration process is significant for the pharmaceutical industry, resulting in the establishment of commercial and open-source patent databases besides institutional offices such as the United States Patent and Trademark Office (USPTO), European Patent Office (EPO), and Japan Patent Office (JPO). The dominant commercial providers in this market are Clarivate (https://clarivate.com/products/ip-intelligence/ip-data-and-apis/derwent-world-patents-index/) and CAS-Scifinder (https://www.cas.org/solutions/cas-scifinder-discovery-platform/cas-scifinder). Simultaneously, public chemical databases have also incorporated patent information for their catalogued chemical collections. PubChem [7] has historically incorporated compound links to USPTO, while the European counterpart, ChEMBL, has developed a stand-alone patent database (SureChEMBL) [8] based upon annotated biomedical entities which allow for searching a larger chemical space to inform drug discovery. Within the drug development domain, SureChEMBL represents one of the largest open-access resources of patents with its ability to extract data from independent patent offices. Such databases have enabled researchers to systematically assess, identify, and explore patterns within patent documents [9,10].

Patent documents can serve as a vast data source for machine learning (ML) and artificial intelligence (AI) applications [11]. ML-based algorithms can be used to predict the success rate of patented compounds by incorporating clinical trial results. This could be also beneficial for patent applicants in making R&D strategic decisions related to competitors. However, the unstructured nature of patent documents and their actual digital media, which are often in PDF format when downloaded from public repositories, also poses a challenge and as a result, manual curation is required to extract valuable information, such as bioactivity data, from these documents [12]. Therefore, there is an increasing need for tools for patent automatic analysis utilising language models (LM) to identify and extract relevant metadata for exhaustive patent data search.

In 2022, we introduced PEMT, a patent enrichment tool that allows extraction of pharmaceutically relevant entities such as targets and chemicals of interest from patent documents [13]. This tool annotates modulators with patent information rather than providing an overview of the usability of the patents in drug discovery. This manuscript aims to guide readers through the tool's implementation and, in doing so, highlight the role patent literature can play in driving decisions in the drug discovery process. We utilised knowledge graphs (KGs), which are graphical databases composed of aggregated literature and experimental data, as the platform for demonstrating the utility of our analyses. For rare diseases, we utilised OrphaNet [14], while for Alzheimer's disease, we used the Human Brain Pharmacome [15]. By means of these KGs, we extracted data from the past two decades of chemical patent space and performed patent landscaping, which involved the identification of patterns within patent documents. We start by identifying the commercial and non-commercial pharmaceutical originators based on the patent activity of their affiliated owners. Next, we performed a retrospective overview of targeted proteins from patent documents, to identify their importance in the drug development process at specific periods. Additionally, we examined the impact clinical trials have had on the corresponding target prioritisation. Lastly, we examined the drug repurposing aspect of pre-clinical small molecules and clinical stage drugs, highlighting potential novel disease treatment options. Overall, this patent landscaping approach highlights the importance of analysing patents and their usefulness in making decisions during drug development efforts.

2. Methods

In the initial sections, we discuss the KGs employed as data sources, explain the functioning of the PEMT tool, and outline the curation process necessary for standardising information from patent documents. Finally, we provide information on the implementation details pertinent to the analysis.

2.1. Knowledge graph selection

We extracted data from two indication-specific semantically organised KGs: OrphaNet and Human Brain Pharamcome (HBP). OrphaNet (accessed on 2021–11–01) is an open-source network that focuses on rare diseases and includes information on biological entities such as orphan gene targets and drugs, as well as clinical entities such as clinical trials and biobanks [14]. In contrast, the Human Brain Pharmacome (HBP) is a publicly available biomedical knowledge graph representing neurodegenerative diseases with a particular focus on Alzheimer's disease [15]. The data in the graph is built on the Biological Expression Language (BEL) framework and includes information on proteins, biological processes, and other relevant features [16].

2.2. Experimental and patent data extraction

To expand the data resource with patent data, we used the PEMT tool [13], which systematically retrieves patent documents using a list of genes or proteins. To do so, it extracts manually annotated and verified chemical and biological modulators that have binding or functional effects on proteins from ChEMBL and subsequently captures all patent documents related to the modulators using SureChEMBL. The patent documents collected are then filtered based on their pharmaceutical relevance and status (i.e. whether the patents are active or expired). Thus, the tool removed patent documents older than 20 years and only included those documents that were tagged with specific IPC code classes referenced in the publication [13]. These codes are alphanumeric codes assigned by officials to patent documents and help in understanding the scope of the patent.

2.3. Patent data harmonisation pipeline

To aid systematic analysis, we developed data harmonisation and normalisation pipelines. In this pipeline, patent owner names were classified into three categories: "Organisation", "Acquired", and "Individuals". Organisations include patent owners that were part of companies and universities, while Acquired owners are those that were within legacy entities which have been acquired or merged into larger Organisations. An example would be Dupont Pharmaceuticals, which Bristol Myers Squibb (BMS) acquired. Individuals are individuals who own a patent. Additionally, similar organisations were grouped for consistency. For example, all variations of Bristol Myers Squibb were mapped to Bristol Myers Squibb. This normalisation step allows for identifying all relevant patents from the same owner and has not been performed previously.

2.4. Compound clinical stage annotation

To annotate compounds based on their clinical trial stage, we utilised the information captured within ChEMBL. For each compound registered in ChEMBL, an additional annotation is made based on whether the compound has been a clinical candidate and if so, at which stage. As a result, ChEMBL assigns phase numbers 0 - 4 where 0 indicates a compound being researched in the preclinical stage, 1 to 3 indicate the compound being tested in clinical trial phases I - III, and 4 indicates the compound has been approved by FDA and/or EMA. In our analysis, we make use of this ChEMBL data to identify compounds, with associated patents, that have been moved from the preclinical to the clinical setup within the drug discovery pipeline.

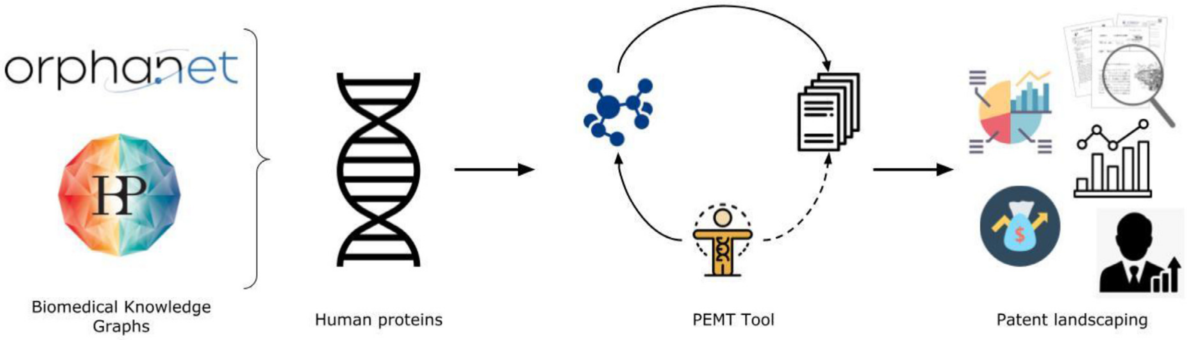


Fig. 1. Synopsis of the patenting landscape analysis. We collect all human proteins from the two biological networks relevant to diseases and symptoms. Using these proteins as a starting point, we run our PEMT tool and retrieve all associated modulators and patent documents based on predefined conditions. Finally, using the patent document dataset, also referred to as patent corpora, we conduct deductive and exploratory analysis to uncover patterns within the dataset.

2.5. Patent data clustering

To cluster patent owners, we made use of a hierarchical approach where two or more owners are clustered together based on a distance metric from singletons to cluster groups in a bottom-to-top manner. This type of clustering allows creation of dendrograms that help establish patent owner groups that belong to the same cluster. The distance metric used in the clustering approach was the distance correlation which enabled in measuring the dependency between the two owners (Eq. (1)). The clustering was achieved using Seaborn ‘clustermap‘ package (https://seaborn.pydata.org/generated/seaborn.clustermap.html).

$$x = 1 - \frac{(u - u') \cdot (v - v')}{||(u - u')||\cdot||(v - v')\,||} \tag{1}$$

Eq. (1). **Distance correlation.** The distance correlation is a metric used to clusters that have minimal distance together. In this equation, *u'* is the mean of vector *u* and *x* . *y* is the dot product of *x* and *y*.

2.6. Target based publication data collection

To extract publications relevant to targets of interest, we made use of open source platforms within NCBI Gene Browser (https://www.ncbi.nlm.nih.gov/gene/) to gather information on the human targets. Through the link between the gene platform and PubMed (https://pubmed.ncbi.nlm.nih.gov/), we collected all relevant publications related to a target. For our purposes, we clustered all the publication found in a year together and used this statistics as the basis for understanding the research trend on targets. Moreover, we rescaled the publication counts to be between 0 and 1 using the min-max normaliser (Eq. (2)).

$$x' = \frac{x - min(x)}{max(x) - min(x)} \tag{2}$$

Eq. (2). **Min-max normalizer.** The min-max normalise allows us to scale the values to a define range. In this equation, *x* is an original value and *x'* is the normalized value.

2.7. Data availability and implementation details

To run our PEMT tool, we extracted human proteins that modulate a disease or symptom from both graphs. The PEMT tool was run on a Windows operating system and was followed by the harmonisation pipeline mentioned in Section 2.3. Additionally, within the collected patent documents, we found 131 patents in rare diseases and 705 patents in Alzheimer's diseases with no associated owner. This is likely due to the quarterly updation of public databases like SureChEMBL, causing a lag in the time for updating. As a result, additional filtering was done to exclude patents with no owners from the study. Fig. 1 provides an overview of the process involved in collecting and analysing the patent dataset, referred to as the patent corpora, in the following sections. The data and the scripts used during this analysis can be found on GitHub at https://github.com/Fraunhofer-ITMP/Pharmaceutical-patent-landscaping.

3. Results

We analysed the patent extracted from 4314 to 10,237 proteins from OrphaNet and HBP respectively. In this analysis, we define “rare diseases” as diseases mentioned within the Orphanet database. In the following section, details four subsections through which we explain how patent literature can be utilised in drug discovery. First, we provide an overview of the information collected from the patent corpora for each disease. Next, we retrospectively analyse the patent corpora unveiling leading organisations in drug research and discovery (R&D) for individual diseases. Following this, we investigated case studies for selected targets based on biological networks and examined their patenting trend. Lastly, we illustrate the usefulness of patent documents in identifying drug repurposing opportunities.

3.1. Overview of the patent corpora

In this section, we investigate the patent corpora of each disease and provide a statistical overview of the information collected within the patent corpora. This summary will provide insights into the scope of patenting activity within the disease area, denoting whether the patents are filed for pharmacological relevance or for formulation purposes. We will also examine and compare the number of patent documents that have been approved versus those still in the application pipeline.

We retrieved 17,506 compound-patent pairs from the dataset, which included 585 compounds and 502 unique patent documents with the rare disease patent corpora. Out of the patents retrieved, 180 were granted patents, represented by documents with kind codes starting with B or E, while the rest were still in the application process, represented by documents with kind codes starting with A (Fig. 2A). Additionally, based on the IPC classification, it was revealed that more than half of the patents were related to C07D (https://www.wipo.int/classifications/ipc/en/ITsupport/Version20170101/transformations/ipc/20170101/en/htm/C07D.htm#C07D), which includes heterocyclic compounds. Following this, 12% of patent documents were in the IPC code C07J (steroids), 7% were in code C07K (peptides), and 6% were in code A61P (specific therapeutic activity of chemical compounds or medicinal preparations) (Fig. 2B).

Conversely, in the case of the Alzheimer's disease patent corpora, we retrieved 76,321 compound-patent pairs that included 22,930 compounds and 13,181 unique patent applications. From these patents, 3980 were granted (indicated by kind codes starting with either B or

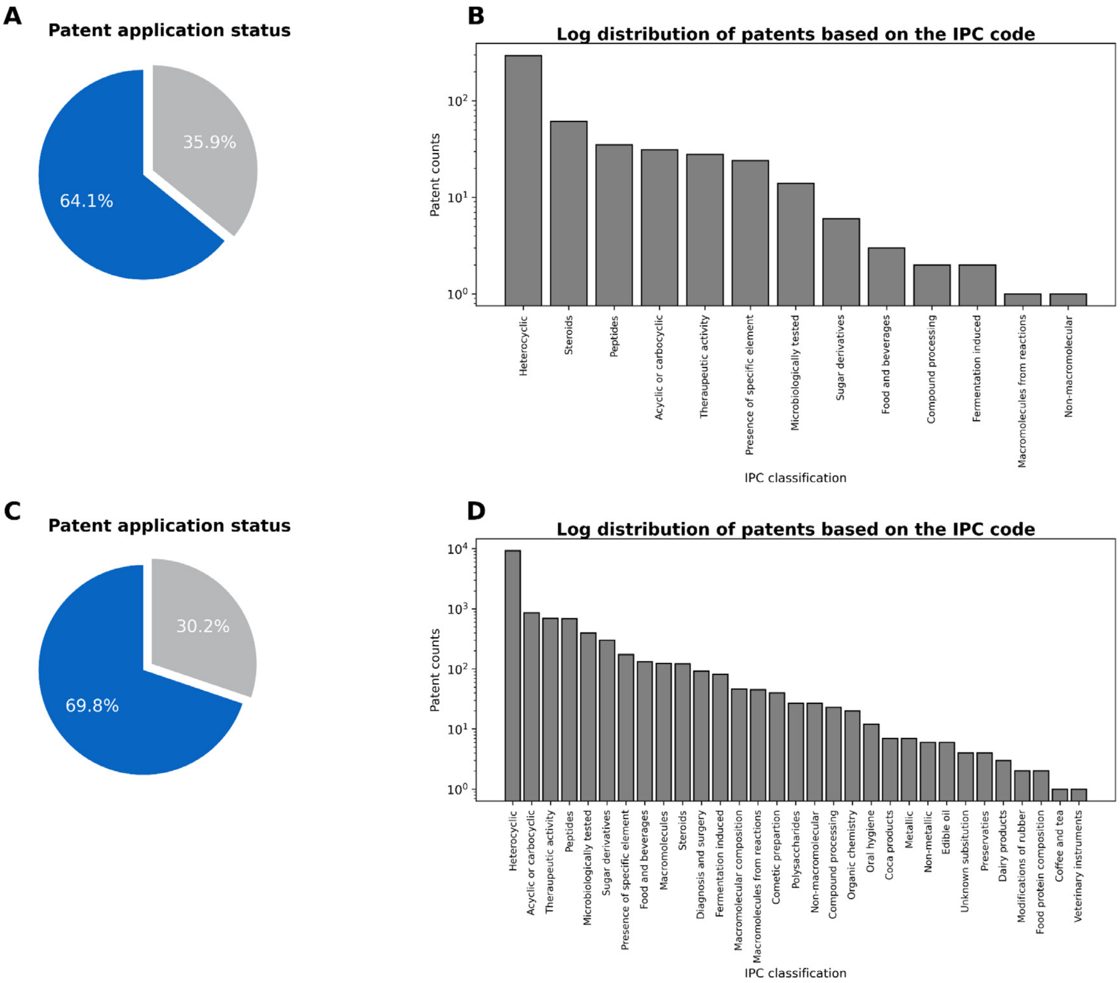


Fig. 2. Patenting Landscape overview for Rare and Alzheimer's Diseases. (**A**) Rare disease patent application ratio denoting whether the patents have been granted (in grey) or are still pending (in blue). (**B**) Logarithmic distribution of the rare disease patents based on the IPC classification category. (**C**) Alzheimer's disease patent application ratio denoting whether the patents have been granted (in grey) or are still pending (in blue). **D)** Logarithmic distribution of the Alzheimer's disease patents based on the IPC classification category.

E), while the remaining were still under the application process (indicated by kind codes starting with A) (Fig. 2C). Upon reviewing the IPC classification of the patents, it was found that 70% of the documents belonged to class C07D, representing patent applications mentioning heterocyclic compounds. Alongside this class, other notable IPCs include C07C (representing acyclic and carboxylic compounds) with 7% of patent documents, A61P (representing a specific therapeutic activity of chemical compounds or medicinal preparations) with 5% of patent applications, and C07K (representing peptides) with 5% of patent applications (Fig. 2D).

Thus, RD and AD patent cohorts revealed two recurring patterns: i) a lower number of granted patents and ii) a dominance of "C07D coded" patent applications. The observation that granted patents are approximately half of all applications filed is consistent with the statistics within the pharmaceutical domain generated by EPO (https://new.epo.org/en/statistics-centre#/technologyfields?code=16). This filed-to-granted patents ratio can be attributed to several reasons. One of the reasons is the difficulty and expenditure involved in the research and development around patents. With the limited number of individuals reviewing the documents, the search for potential competitiveness within pre-existing inventions is time-consuming (https://www.epo.org/learning/materials/inventors-handbook/protection/patents.html). Additionally, the prevalence of C07D patents is likely due to the diversity provided by this class of compounds, which makes them a desirable candidate for drug discovery patent innovation compared to other compound classes [17].

3.2. Trends in the patenting landscape from the patent owners' perspective

Next, we analysed the patenting activity for each disease to determine the major pharmaceutical companies and academic institutions that have made significant contributions to the R&D of potential drugs in the field. To do so, we leveraged the high-level patent owner classification (i.e., organisation, acquired, and individuals) and grouped patent owners based on the number of applications they have filed. Then, we reviewed the top patent owners and deduced their historic patenting activity in the field. We examined RD and AD separately to understand better the patent landscaping related to each disease.

3.2.1. Identifying key players in rare disease patenting over the past 20 years

For RD's, from 2000 to 2021, 369 organisations, 53 individuals, and 80 acquired organisations have played a role in the patent landscape. As illustrated in Fig. 3, patenting activity has grown over the years, with organisations being the most active, followed by individuals. The boost in the patenting of large organisations, mainly pharmaceutical companies, is ascribed to legislations like the Orphan Drug Act (ODA), introduced in 1983, which gave financial incentives for developing therapies for

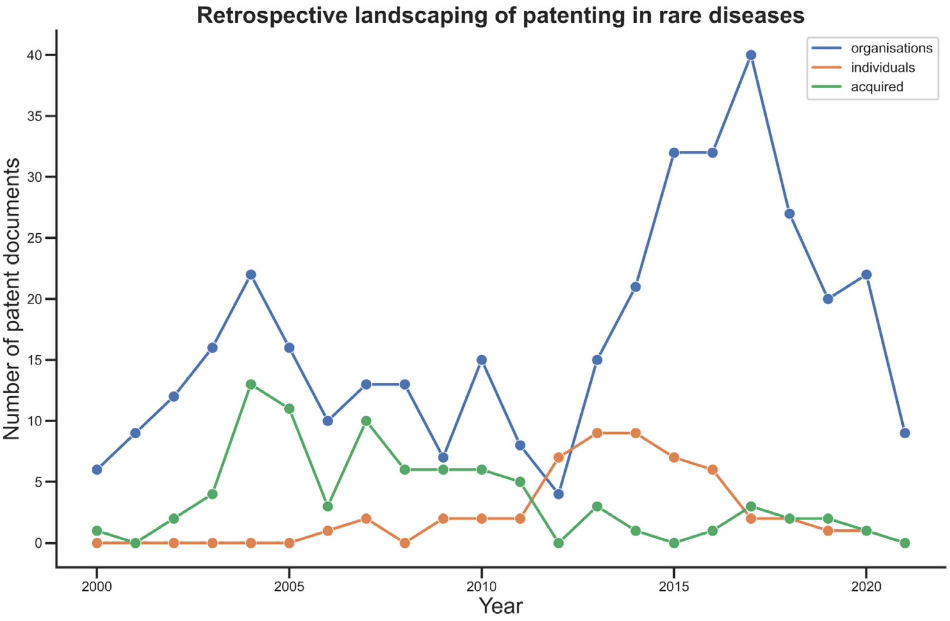


Fig. 3. The retrospective landscape of the rare diseases patent corpora in the past two decades. The line plots indicate the growth of the patent documents years based on assignee type: organisations (blue), individuals (orange), and acquired (green). As expected, a large contribution is evident for organisations compared to others. A large boost in the Individual owners is seen around 2012, after which there has been a decline due to the legislatures.

rare diseases [18]. On top of that, the surging patent activity within the field after 2012 was correlated to the establishment of the Patient Protection and Affordable Care Act (ACA) which catalysed engagement of investors in financing R&D for orphan drugs and diseases [19]. On the contrary, a decrease in independent patent applications by individuals was observed since its peak in 2013. This effect could be due to two reasons: firstly, changes made in US and EU legislatures during the early 2000s, moving the “professor’s privileges” at IP generation to the universities, and secondly, professors and professionals, due to their legislation, were either hired by commercial companies or became part of larger academic groups [20].

To get a historical perspective of these patents, we selected 10 top patent owners and organised their patent applications based upon the timestamp of each patent application date, also known as the priority date. This subset of owners consisted only of global pharmaceutical companies such as Pfizer, GlaxoSmithKline, etc. (Fig. 4). It is essential to note that many start-up and medium-sized companies have emerged and subsequently been absorbed by larger organisations providing them with the leverage of patent acquisition filed by the smaller companies. One example is Sterix Ltd., which had 33 patents and was acquired by Ipsen (in 2004), adding to the company’s patent portfolio. Furthermore, most top assignees had a decreasing patenting activity in the past years. For instance, global giant Takeda, a Japanese multinational pharmaceutical company, just had one compound, Luvadaxistat, out of 122 from the patent compound space between 2000 and 2021 that went into the clinical trial phase II for Friedreich’s ataxia [21]. Thus, indicating a reason for decreasing interest of Takeda in rare diseases patenting (Fig. 4). Regardless, the top 10 assignees accounted for 44% of the 502 patent applications analysed, and the top 20 assignees accounted for 61%. Additionally, we observed the cumulative positive patenting trend for the top patent inventors in rare disease. The trend between 2013 and 2021, however, can be observed with a peak around 2016 and a gradual decline in the later years (**Supplementary Figure 1**).

Additionally, we looked into the correlation between the patenting activity for the top owners based on the target portfolio to identify pharmaceutical companies that show similar target portfolios. We identified two clusters when correlating the target-based portfolio of the top 10 owners in rare diseases (Fig. 5). One of the clusters was between Sanofi and Ardelyx. We suspect this was due to their collaboration on sodium and potassium channel inhibitors (which include KCNK3, KCNK9 and SLC9A3), allowing them to apply for over 50 patents (https://ir.ardelyx.com/news-releases/news-release-details/ardelyx-licenses-nap2b-phosphate-inhibitor-program-kidney). A prominent target in the cluster space is DAO, with competitive patenting between large pharmaceutical companies like Pfizer, Takeda, and John and Johnson. A number of singletons are also seen, indicating the development of candidates targeting isolated targets (e.g. GSK, Cleave). Together this target portfolio approach allows for identifying top patented targets within the indication areas and comparing patent owners’ target portfolios.

As mentioned, no patent assignees in the top 10 were affiliated with an academic institute. To depict the difference, we divided organisations into two categories, industry and academia. Interestingly, a boost in patent applications from the academic sector was seen after 2012 has been tracked. (**Supplementary Figure 2A**). However, a general positive trend in patent applications by the industry is observed denoting increasing interest within pharmaceutical drug discovery for novel treatments for this disease. During the same time, patent law amendments took place in the United States of America, commonly referred to as the 2011 Patent Reform Act [22]. The reformed law pointed out the change in the prosecution of patent applications within the U.S. Patent and Trademark Office (USPTO) by moving from a “first-to-invent” to a “first-to-file” system. This law thereby redefined what was required in terms of prior art, modifying the application process in a significant way.

3.2.2. Identifying key players in Alzheimer’s disease patenting over the past 20 years

In Alzheimer’s disease, 10,017 organisations, 1778 individuals, and 1386 acquired organisations have contributed towards the patenting activity. Similar to the rare disease patent landscape, organisations are in the lead compared to individuals and acquired organisations (as shown in Fig. 6). This is attributed to the exceptionally high cost involved in the research and development of clinical trials, which cannot be afforded by publicly funded universities or small-scale organisations unless in partnerships with pharmaceuticals [23]. On the other hand, individuals showed an increase in patenting activity in 2011, with a peak in 2013 and then a decline afterwards, likely due to changes in patenting laws made by US and EU government bodies that allowed for commercial innovation rights to universities as well. Thus, individuals affiliated with universities lost their right to file patents to universities, while other in-

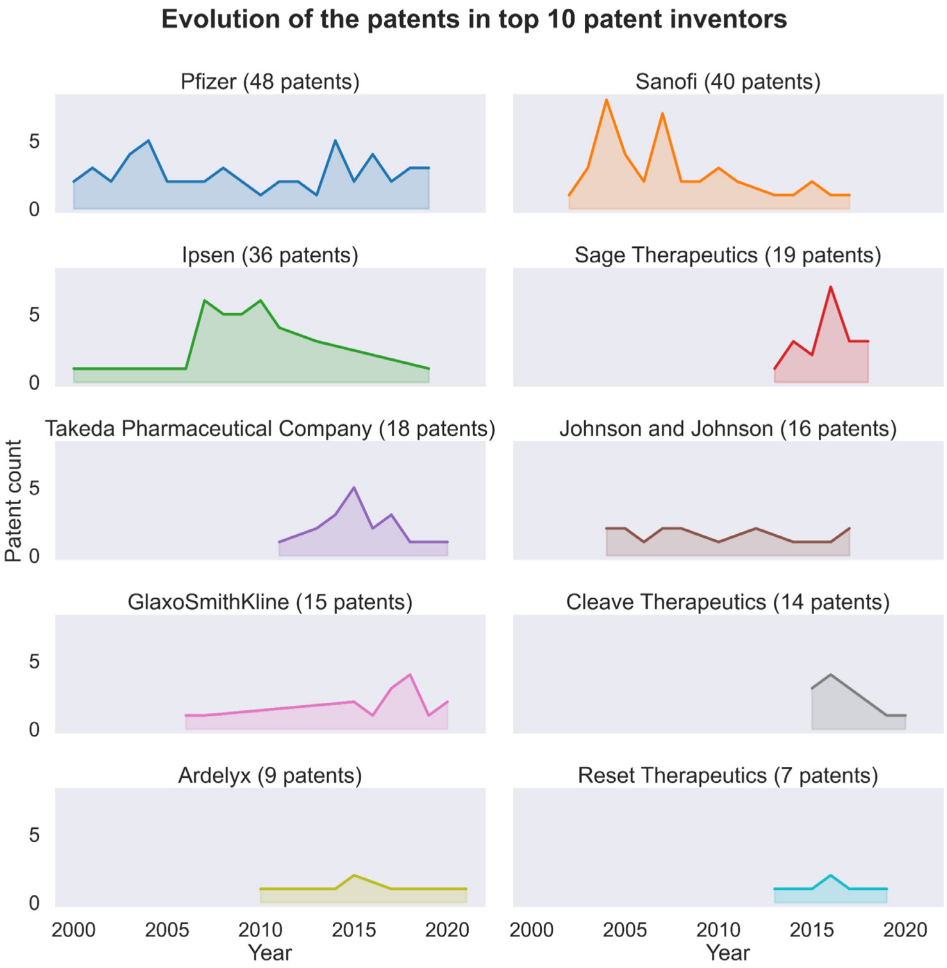


Fig. 4. Top owners-based patent analysis in the field of rare diseases. Patent landscaping of the top 10 owners in the rare disease domain in descending order. The top owners were deduced from the In the figure, it can be seen that large pharmaceutical companies historically active in rare diseases, like Sanofi or Takeda, declined their patenting activity in the past few years, allowing other biopharmaceuticals such as GlaxoSmithKline, Pfizer, and Sage therapeutics to enter the field.

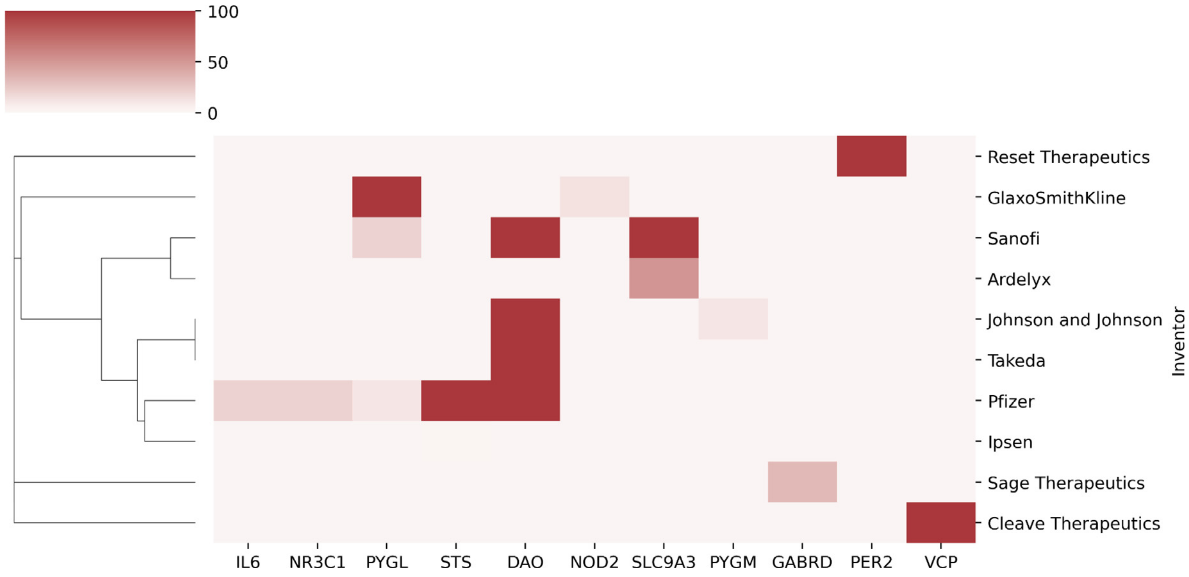


Fig. 5. Target portfolio of top 10 owners in rare diseases. The hierarchically-clustered heatmap denotes the correlation distance between the top 10 rare disease owners based on their respective target patent landscape. The colour scale in the heatmap represents the number of patent documents associated with these targets. Together this clustering approach allows for the identification of clusters, here between Sanofi and Ardelyx and Pfizer and Ipsen, thereby indicating overlap between the targeting portfolio of these pharmaceuticals. From the plot, we can also notice the focus on DAO inhibitors within the rare disease industry.

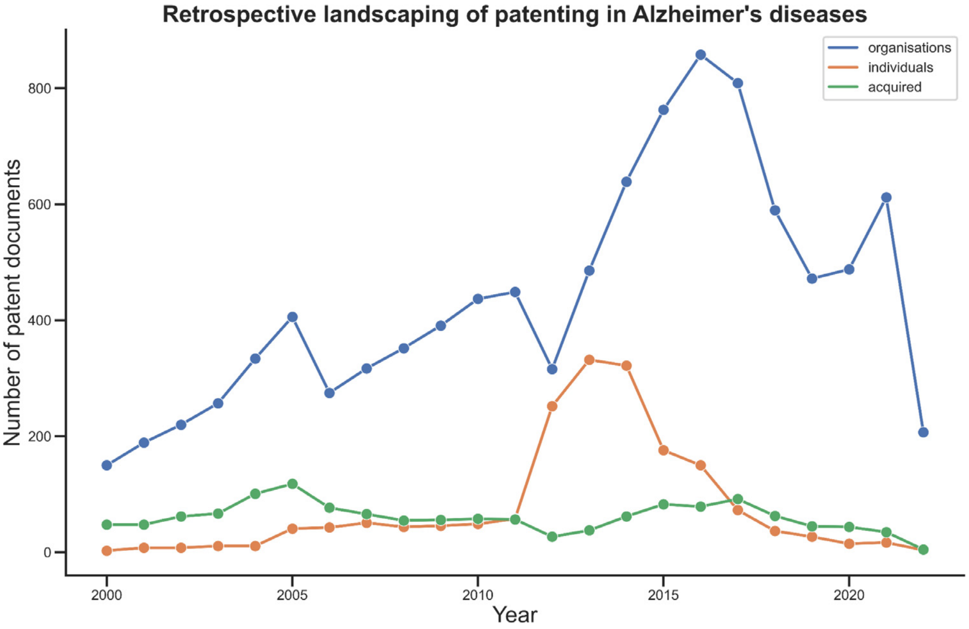


Fig. 6. Patent activity by owner type in the Alzheimer's field. The line plots indicate the growth of the patent documents years based on assignee type: organisations (blue), individuals (orange), and acquired (green). Pharmaceutical companies and universities are leading in patenting activity compared to others. Compared to the rare disease field, the contribution of the acquired owners is greater. A decline in Individual owner activity was observed post-2012 due to the legislature changes in US and EU.

dividuals attracted pharmaceutical companies towards them and soon were hired by these large companies.

We generated a retrospective perspective of the top 10 organisations found in Alzheimer's patent corpora (Fig. 7). Even though the total number of patent applications was higher for Alzheimer's disease compared to rare diseases, a similar trend was observed. The top patent owners include Roche, Bristol Myers Squibb, Pfizer, etc., which are global pharmaceutical companies. There were "patent cliffs", where a sudden loss of patent rights occurs due to concomitant expirations of several patents at the same time [24], within the underlying cohorts with decrements and increments in the patenting activity during the two decades. The increment denoted the increased interest of researchers in understanding the complex mechanism involved in Alzheimer's disease. Despite the high investment in basic and clinical research, the mechanisms involved in the modulation of Alzheimer's disease remained unclear, with a high attrition rate seen in clinical trials [25,26]. We would like to point out that the trend mentioned above for Alzheimer's disease would not be fully concordant with the current market trends as we did not consider any "biologicals" compounds, which have played an enormous role in drug discovery against Alzheimer's disease lately [27]. A similar trend as that of rare diseases was observed in the cumulative patenting of the top patent inventors in Alzheimer's disease. A gradual increase in activity was observed from 2013 with a peak around 2016 followed by a decline in the activity as the years came closer to the COVID-19 pandemic age (**Supplementary Figure 1**).

Next, we looked at the top patent owners with Alzheimer's disease and plotted their target portfolio to examine whether two or more owners follow a similar target portfolio. Prominent collaboration partners such as Novartis and Amgen have coordinated neuroscience efforts and hence can be seen within a cluster in Fig. 8 (https://www.genengnews.com/news/amgen-novartis-launch-neuroscience-drug-collaboration/). Moreover, individual companies such as Takeda are observed due to their dominance, in particular, targeting mechanisms such as LPAR5 and SLC6A9. It is noteworthy that despite having common target spaces with leading pharmaceuticals, each top owner has at least one other neurodegenerative-related target space where it is leading with respect to the patenting landscape. ROCK2 for BMS, TTK for Bayer, and GCG for Pfizer are examples of this case. Overall, this approach enables identification of patent owners with comparable target focus and uncovers patterns in neurodegenerative drug target prioritisation.

Similarly to the rare disease case, none of the top patent owners were from the academic sector, which is not surprising given the R&D cost involved in the case of Alzheimer's disease [23]. Thus, we looked into the patenting activity of the academic sector and compared their growth with the industry (**Supplementary Figure 2B**). Similar to rare diseases, a general positive trend in patent applications by the industry is observed. An incremental rise in academic patenting activity was observed due to the amendment of the 2011 Patent Reform Act that was highlighted previously. During the same period, there was establishment of an act in the United States, the National Alzheimer's Project Act (NAPA), that promoted research and development in the field of Alzheimer's (https://www.nia.nih.gov/about/nia-and-national-plan-address-alzheimers-disease). In Europe, collaborative projects involving public-private partnerships with key stakeholders from the pharmaceutical industry were developed. One such project was the European Prevention of Alzheimer's Dementia Consortium EPAD (https://ep-ad.org/) which involved data collection and analysis to speed up the drug discovery process for the disease, and support for improving clinical trial practice and design.

3.3. Elucidation of the research trends of targets using patent documents

Many factors contribute to the successful discovery of small molecule drugs for specific therapies. One of the essential factors is the mode of action (MoA) of the drug in the human body, which refers to the effect a drug has on a biological protein (also known as a target) that is relevant to mechanisms involved within diseased conditions. With the help of patent documents, we can investigate the target-based MoA claims within patent applications based on their effect in certain indication areas or on their importance over time from success achieved in clinical trials. Additionally, identifying potential disease targets, their prevalence, and their role in affected patient populations can be useful in developing new intellectual property. In this section, we will focus only on trends in the prioritisation of targets in patent literature over the years, while the therapeutic usage of targets has been previously described [28].

To investigate the research trend for targets, we first linked the patent information to targets. This was done by retrospectively connecting the patents to the targets through small molecule modulators. We would like to highlight that no prospective mapping of patents to targets based on their presence in the patent document was performed to con-

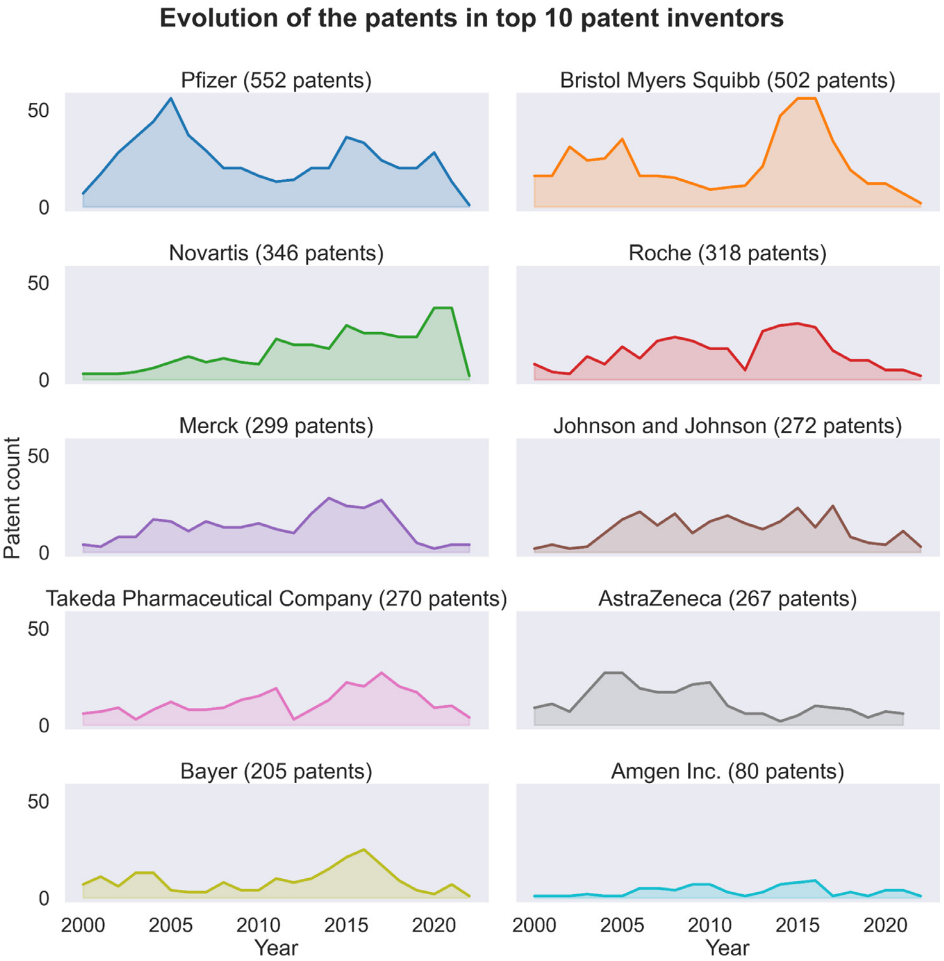


Fig. 7. Patenting activity from the top owners in Alzheimer's disease. In descending order, the plots showcase the patenting activity of the past two decades for the 10 top owners. The leaders in the field are global pharmaceutical companies such as Pfizer and BMS. All the top owners have been active within the field of Alzheimer's with consistent patenting activity.

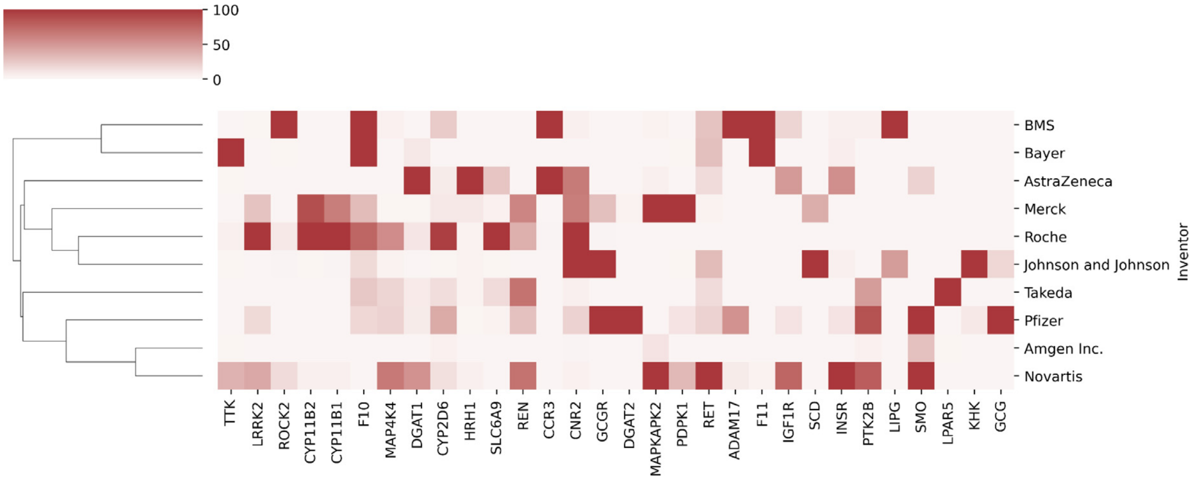


Fig. 8. Target portfolio of top 10 patent owners in Alzheimer's diseases. The hierarchically-clustered heatmap denotes the correlation distance between the top 10 Alzheimer's patent owners based on their respective target patent landscape. Each coloured box in the plot indicates the number of patents from the owner linked to the specific target. We can observe three main clusters in this plot: the first is that of BMS and Bayer, the second is that of Roche and Johnson & Johnson, and the third is composed of Amgen and Novartis.

firm the link. Similar to the patent assignee analysis, we looked into the prevalence of targets over time **(Supplementary Tables 1 and 2)**. Furthermore, we will clearly distinguish the target priorities between RD and AD to understand the involved targets' evolution. For both RD's and AD, only the top targets, represented by the number of corresponding patent documents, were used to understand the target portfolio focus within the pharmaceutical companies.

We first started to evaluate the variations of the targets that are prioritised each year, to form the basis for target-centric landscaping, enabling us to identify targets that have been continued, diverted (to other targets), or halted, with respect to research, during their development. During this analysis, it was observed that the target perspective of the patent landscape was characterised by waves, which meant that not all targets were prioritised simultaneously **(Supplementary Figure 3)**.

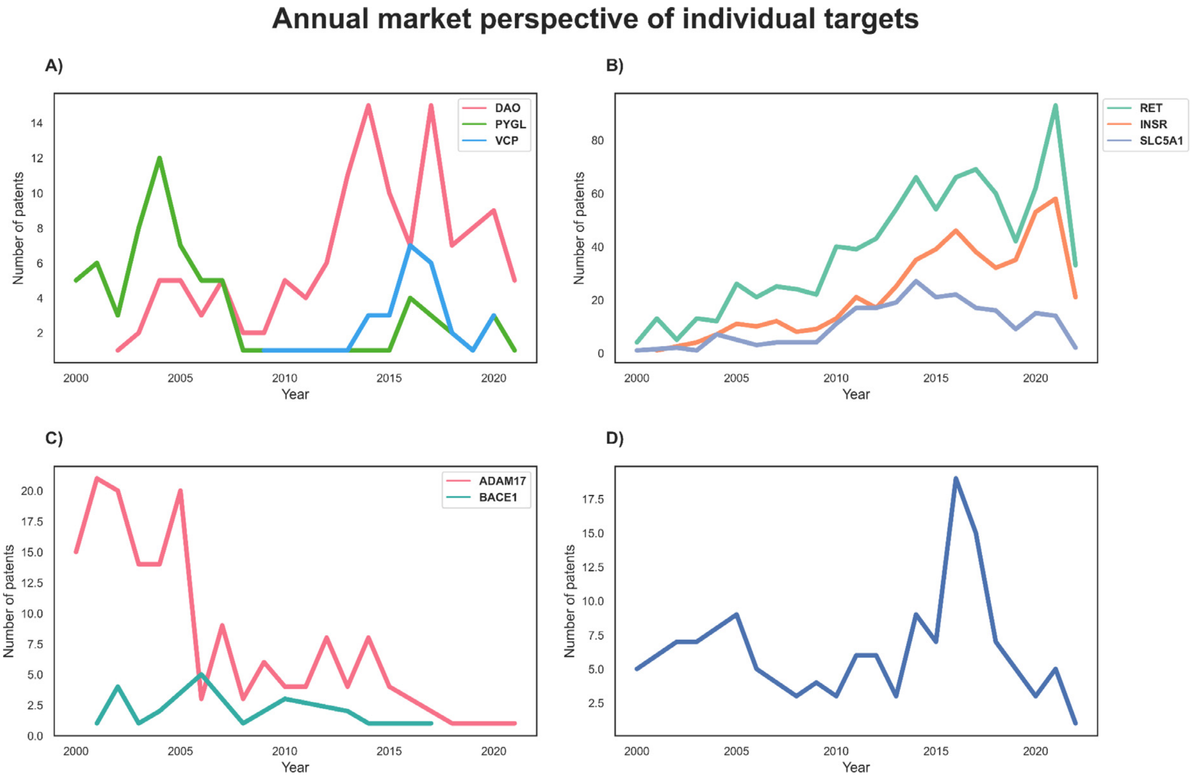


Fig. 9. The patenting frequency of selected targets within each disease domain. (A) Targets for rare diseases. In the case of rare diseases, we looked into three targets, namely DAO (red), PYGL (green), and VCP (blue), to understand the underlying market perception concerning patents. **(B) Targets for insulin metabolism-related neurodegenerative diseases.** In the case of Alzheimer's disease, we looked into a set of targets that play a role in insulin and sugar metabolism, namely RET (turquoise), INSR (orange), and SLC5A1 (violet). Each of the three targets follows a similar patenting profile. (**C) Secretase-related targets in Alzheimer's diseases.** The major secretase enzymes related to Alzheimer's disease include ADAM17 (pink) and BACE1 (teal), and both of them follow a different patenting profile despite being in the same class. (**D) Patenting profile for ELANE, a target within Alzheimer's disease.**

Thus, we looked at selected high profile targets for each disease and compared the patent activity with the corresponding therapeutical trial activities to understand the correlation between the agglomerated targets in a specific year and the preclinical research in drug discovery.

Glycogen phosphorylase L (PYGL) was one of the high incidence targets with 71 patent references. The mutation in the PYGL gene causes inhibition in the conversion of glycogen to glucose, thus, associating it with an autosomal recessive rare disease called glycogen storage disease type VI (GSDVI) [29]. This key link between the mutant PYGL and GSDVI was the reason for its early research and patenting activity. However, over the past years, interest in this target has gradually decreased, with almost no patent documents between 2014 and 2019 (Fig. 9A). In early 2019, small molecules called glucokinase activators (GKAs) were thought to be a potential solution for type II diabetes due to their ability to regulate glucose-6-phosphate and PYGL [30]. As research progressed on GKAs, it was found that they induced hypoglycemia, a condition characterised by decreased blood glucose levels, causing all clinical trials during that time to be terminated [31]. As a result, researchers refocused on the PYGL gene post-2014.

Another illustrative target in the ranked list of targets was ATPase valosin-containing protein (VCP/p97). This protein plays a role in intracellular homoeostasis by regulating protein metabolism and is associated with multiple diseases such as Paget disease and Amyotrophic Lateral Sclerosis and its types [32,33]. However, due to the complex nature of the gene's function in the body, research took an extended period, causing a dormant period in the patenting world. As soon as it gained interest, there was a steep increase in the number of patent documents, but this number soon fell (Fig. 9A). The key reason for this decline could be realistically attributed to VCP inhibitor CB-5083, a molecule developed by Cleave Bioscience, that failed at phase-1 due to off-target effects [34]. Subsequently, it was deduced that the covalent binding towards the target caused mutations that modified the morphology of protein, thus, making cells resistant to the drug [35]. Moreover, since the target is involved in cellular homoeostasis [36], a deeper understanding of the MoA of drugs is needed for such targets.

Another target that surfaced in the rare disease domain with the highest number of associated patent documents was D-amino acid oxidase (DAO). It is a neuromodulator that was found to be involved in psychotic disorders [37]. This target has been active over the years from both the patenting (Fig. 9A) and scientific literature, with several clinical trials. Additionally, the discovery of the target's role in schizophrenia and cognitive dysfunction, and the need for its inhibition in these conditions, is one of the reasons for its active patenting landscape [38]. However, in the past two years, there has been a decrease in patents. This decline in patent activity could be ascribed to the COVID-19 pandemic, as global attention and resources have been focused on addressing the outbreak [39].

Interestingly, in the context of neurodegenerative diseases, some of the most well-known targets for Alzheimer's, such as amyloid beta precursor protein (APP) and tau (MAPT) [40], were not found in the top target list **(Supplementary Table 2).** This was likely due to an absence of publicly available bioactivity data within resources such as ChEMBL on these targets, as antibodies and other biologicals are not considered. Contrary to these, coagulation factors, such as coagulation factor X (F10), which have a direct link to neuro-related diseases, were found in the top list [41] due to several novel small molecule modulators discovered. At the top of the target list was ret proto-oncogene (RET), with 886 patent application documents, demonstrating an active patenting profile (Fig. 9B). This activity profile can be attributed to the growing interest and research regarding insulin and sugar metabolism in the brain, where the receptor RET contributes towards metabolic homoeostasis [42]. Other targets that contribute to the same mechanisms include insulin receptors (INSR) and glucose transporters solute carrier family 5 member 1 (SLC5A1), which were also found in the list later with an active patenting profile (Fig. 9B).

Secretases such as alpha-secretase (ADAM17) have gained interest within the pharmaceutical industry since the early 2000s due to their role in inhibiting amyloid beta formation [43]. Despite high patenting activity in the early years, there was a sharp decrease in 2005, and interest in such targets has been declining since then (Fig. 9C). Possible reasons for this decline include the failure of clinical trials for secretase inhibitors Lundbeck's Flurizan and Lilly's semagacestat in later years [44,45]. The major drawback of these compounds was their inability to reach the desired concentration in the brain, as they could not pass the blood-brain barrier, thus, not reaching long residency time.

Other targets, such as neutrophil elastase (ELANE), also demonstrate patenting cliffs [25]. As described previously, this is correlated with either research activities on the target or clinical trial outcomes of drugs. The same pattern was observed in ELANE (Fig. 9D), which involved the development of elastase inhibitors and their advanced clinical trials [46]. The declining patenting activity was attributed to the identification of multiple mutations for the targets and the involvement of these mutants in multiple diseases. This target complexity thus required further research on both the biological and drug development sides.

Lastly, it is already known that there is a "lag" in the period between the filing of patent applications and the scientific research involved. We extrapolated this lag in the target space by looking into selected top targets and correlating the patent applications and research publications over time (**Supplementary Figure 4**). It can be observed that, on average, there is a lag of 5 years between the research involved on a target (with or without disease context) and the filing of patents relevant to the target. There are indeed certain outliers to this lag period such as INSR, which took at least 14 years to enter into the Alzheimer's patenting world. One reason for this could be due to the importance of the target on other diseases, causing it to be extensively researched in the early 2000s. However, only after 2009 did the study of INSR in neurodegeneration start following the recognition of type 2 diabetes as a risk factor for Alzheimer's [47].

3.5. Identification of drug repurposing scenarios of diseases using patent documents

As drug discovery is so resource intensive, a preliminary assessment of promising or under-exploited protein target and/or chemical scaffold combinations repurposing drugs could be a convenient alternative approach. We aimed to investigate whether drugs that have been repurposed for different therapeutic areas are also present in the patent corpora for the two disease areas. To achieve this, we clustered the patent documents based on the compounds they mentioned and ranked them in descending order of their counts. Since we were interested in repurposed compounds, we restricted the patent search to granted documents only. It is important to note that a patent application does not exclusively cover a single compound due to the presence of Markush structures [48]. These structures are a scaffold representation of a compound with the addition of variable side chains, thus making it more difficult for researchers to identify a specific compound via a direct search match, although commercial software tools (like SciFinder) have been developed to solve this problem. A single compound can also be covered by multiple patents, as the same compound can be used for varied innovative features (such as disease type or treatments, physical pharmaceutical forms, combinations with other therapeutic compounds, etc.), allowing for independent drafting for each area. Moreover, multiple patent drafting does not necessarily mean the compound is used for multiple purposes, as they may be patents filed for the compound's formulations or non-obvious structural modifications. Thus, a systematic case-by-case study needs to be done to identify such compounds and ranking of compounds based on patent documents can serve as a starting point for such analysis.

Aligning our definition of repurposed compounds to a patent perspective, we categorised a compound as repurposed if it was found in more than two granted patent documents (belonging to B or E kind code). This resulted in a reduced dataset of 145 repurposed compounds in rare diseases and 1928 in Alzheimer's diseases, out of 585 and 22,682 compounds, respectively. Moreover, from the entire compound dataset, only four compounds for rare disease compounds and 215 compounds for Alzheimer's diseases have been tested in clinical trials (as of 1st of Jan 2023 http://clinicaltrial.gov) (as shown in **Supplementary Figure 5)** with minimal overlap with the dataset of repurposed compounds. This indicates that most of the compounds within the granted patent space under study are still in the research or pre-clinical phase, likely due to the limited availability of bioassay data on clinically validated compounds in public data repositories such as ChEMBL.

Cleave Biosciences's CB-5083 is one of the RD repurposed drugs forming an interesting case study. The company had patented the compound in Europe (EP-2,875,018-B1) and the United States of America (US-10,010,554-B2) as an inhibitor of VCP for cancer treatment. Following a phase I clinical trial, which revealed off-target activities, clinical development was halted [49]. Despite the unsuccessful clinical trial, it has made its way to the RD community regarding its potential use in treating Paget disease, a VPC-dependant disease [32]. Another example is Fosdagrocorat, a compound patented by Pfizer as a glucocorticoid receptor modulator for osteoporosis (EP-2,114,888-B1). Research has linked the same receptor to rheumatoid arthritis [50], thus making it possible to repurpose the compound for this rare disease [51]. Upadacitinib, commonly known by the trade name Rinvoq, is an Abbvie-developed kinase inhibitor that has been approved by the FDA for the treatment of multiple diseases, such as rheumatoid arthritis [52] and atopic dermatitis. It inhibits the Spleen tyrosine kinase (Syk) (US-10,072,034-B2), which has recently been identified as a key modulator of tauopathy, a disorder involving the accumulation of Tau [53]. This discovery opens up the potential for the drug to be used to treat neurodegenerative diseases. Similar to rare diseases, we found a few compounds derived from the cancer domain. Sotorasib is a compound that inhibits a specific KRAS mutant, KRAS G12C and demonstrates anticancerous properties (US-10,519,146-B2) [54]. Several researchers have identified several KRAS mutants that play a role in brain and spinal-related haemorrhage [55]. This presents an opportunity to test the compound on multiple KRAS mutations and optimise it to create a mutant-specific or general-purpose compound. In summary, it is evident through the above examples that the mining of patent documents from a compound perspective can provide insights into how compounds that have been repurposed for multiple disorders beyond the scope of their original patents.

3.6. Limitations and future perspective of the approach

The current study illuminates the historical trends observed in neurodegenerative and rare disease drug discovery and research through the analysis of patents. However, it is important to note that the methodology employed has limitations due to the lack of inclusion of "biologicals" compounds such as vaccines or long peptides. One such limitation is the indirect relationship between proteins and patents, as the connections were made based on chemical modulators, and the initial proteins may not necessarily be mentioned in the patent. To address this, further analyses using natural language processing (NLP) methods to extract gene annotations in patent documents are the best alternative. Additionally, the modulators linked to the proteins were sourced from open-source data repositories, such as ChEMBL. As the deposition of clinical candidate structures is infrequent and often only occurs after early clinical trials, the patent analysis may miss important compounds currently in development. To mitigate this, we focused on highly active compounds with submicromolar activity to increase the chances of identifying "champion" compounds in a patent. Lastly, the time range chosen for the analysis (2000–2021) may have skewed the overall distribution of results, as certain compounds approved for treatment before 2000 were not included in the analysis. For instance, Jeffrey Cummings (2018) mentioned in his article that “Agents producing cognitive enhancement may have mechanisms independent of AD-specific pathology (e.g., 5-HT6 antagonists)” [56]. 5-hydroxytryptamine receptor 6 (5-HT6) antagonists like memantine and acetylcholinesterase inhibitors have been approved for AD treatment and only alleviate some AD symptoms by enhancing cholinergic signalling. These chemicals are not curative and do not surface in our analysis, as they were patented before 2000. Despite these limitations, the study aims to provide insight into the intellectual effort required within the pharmaceutical industry to identify competitors’ patents containing "clinical candidates" and establish a successful "me-too" strategy.

4. Conclusion

Finding safe and effective therapies for diseases with unmet patient needs is a crucial goal of pharmaceutical research. Traditionally, the discovery process was initiated by systematically reviewing the scientific literature for the most promising disease-linked therapeutic targets or pathways. Once a targeting mechanism has been selected, it would either be subsequently validated or eventually discarded and an alternative strategy adopted based on new experimental insights or literature analysis. Additionally, assessment of target opportunities and estimating their potential for successful translation in development candidates is a core strategic activity within the pharmaceutical industry. Such analysis is driven by machine learning models trained on large multimodular data, including toxicological effects and druggability characteristics. Moreover, the novelty associated with the chemical structure of candidate molecules is a critical aspect of patent assessment, as companies need to be aware of pre-existing successful scaffolds. This chemoinformatic integration of structural features is often a key factor in defining medicinal chemistry strategies when optimising leads and further progressing compounds to development stages.

This study uses the PEMT tool to extract, analyse, and explore the patent landscape for rare and Alzheimer’s diseases. The results revealed historical trends in the past 20 years and identified key patent holders and application dates. Such a retrospective trend paved the way for generating a longitudinal visualisation of the importance of the targets in both diseases. Additionally, the analysis showed the potential for using patent literature in drug repurposing and annotating the clinical significance of targets. Last, but not least, the study was able to identify and reflect the shift between the research and patenting period involved in drug discovery. Certainly, this effort cannot be assumed to reach the same proficiency as a professional patent analysis as it takes advantage of public databases, which themselves are limited, as discussed in the previous sections. This paper is focused on the integration of public bioactivity datasets with patent applications and their analysis. We will inspect through future analyses how the patent information can optimise drug search within disease mechanisms networks, which is out of the scope of the current paper. This is just the tip of the iceberg as huge effort and resources need to be dedicated to standardising the data and making them a potential source for future ML and LM models.

Our work demonstrates that the PEMT tool provides a comprehensive and informative patent cohort to understand the research in the field of small molecule drugs. This landscape not only aligns with scientific literature but also allows for competitive intelligence analysis in pharmaceutical R&D. Overall, this patent landscaping approach from a drug discovery perspective highlights the importance of analysing patents and their usefulness in making decisions for future drug development efforts. The possibility of tracking patenting activities from publicly available resources might help define strategies for target ranking within a therapeutic focus. Our methodology will contribute towards simplifying and streamlining the process of target prioritisation in drug discovery campaigns at a pharmaceutical company.

Funding

This work was supported by the German Federal Ministry of Education and Research (BMBF) [01GM2003A] (project ‘TreatKCNQ’).

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.

CRediT authorship contribution statement

Yojana Gadiya: Project administration, Formal analysis, Writing – original draft. **Philip Gribbon:** Writing – review & editing. **Martin Hofmann-Apitius:** Writing – review & editing. **Andrea Zaliani:** Project administration, Formal analysis, Writing – review & editing, Writing – original draft.

Data availability

The data and code is available on GitHub at https://github.com/Fraunhofer-ITMP/Pharmaceutical-patent-landscaping.

Acknowledgements

The authors thank Daniel Domingo-Fernández, Bruce Schultz, and the anonymous reviewers for their comments and suggestions on the manuscript. We would also like to thank Bruce Schultz for providing the protein data for the HBP network, hosted by Fraunhofer SCAI.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ailsci.2023.100069.

References

[1] Walker RD. Patents as a source of scientific and technical information in developing nations. World Pat Inf 1988;10(1):5–10. doi:10.1016/0172-2190(88)90210-4.

[2] Morgan MR, Roberts OG, Edwards AM. Ideation and implementation of an open science drug discovery business model–M4K Pharma. Wellcome Open Res 2018;3. doi:10.12688/wellcomeopenres.14947.1.

[3] Scianna M, Munaron L. Computational approaches for translational oncology: concepts and patents. Recent Pat Anticancer Drug Discov 2016;11(4):384–92. doi:10.2174/1574892811666161003111543.

[4] Pan CL, Chen FC. Patent trend and competitive analysis of cancer immunotherapy in the United States. Hum Vaccin Immunother 2017;13(11):2583–93. doi:10.1080/21645515.2017.1361074.

[5] Zhang T, Chen J, Jia X. Identification of the key fields and their key technical points of oncology by patent analysis. PLoS ONE 2015;10(11):e0143573. doi:10.1371/journal.pone.0143573.

[6] Xiong W, Cao J, Zhou X, Du J, Nie B, Zeng Z, Li T. Design and evaluation of a prescription drug monitoring program for Chinese patent medicine based on knowledge graph. EvidBased Complement Alter Med 2021;2021. doi:10.1155/2021/9970063.

[7] Li Q, Cheng T, Wang Y, Bryant SH. PubChem as a public resource for drug discovery. Drug discovery today 2010;15(23–24):1052–7. doi:10.1016/j.drudis.2010.10.003.

[8] Papadatos G, Davies M, Dedman N, Chambers J, Gaulton A, Siddle J, Overington JP. SureChEMBL: a large-scale, chemically annotated patent document database. Nucleic Acids Res. 2016;44(D1):D1220–8. doi:10.1093/nar/gkv1253.

[9] Mucke HA. What patents tell us about drug repurposing for cancer: a landscape analysis. In: Seminars in cancer biology, 68. Academic Press; 2021. p. 3–7. doi:10.1016/j.semcancer.2019.09.010.

[10] Falaguera MJ, Mestres J. Congenericity of claimed compounds in patent applications. Molecules 2021;26(17):5253. doi:10.3390/molecules26175253.

[11] Choi Y, Park S, Lee S. Identifying emerging technologies to envision a future innovation ecosystem: a machine learning approach to patent data. Scientometrics 2021;126:5431–76. doi:10.1007/s11192-021-04001-1.

[12] Leach AR, Magarinos MP, Gaulton A, Felix E, Kiziloren T, Arcila R, Oprea TI. Illuminating the Druggable Genome through Patent Bioactivity Data. *bioRxiv*, 2022-07 2022. doi:10.1101/2022.07.15.500187.

[13] Gadiya, Y., Zaliani, A., Gribbon, P., & Hofmann-Apitius, M. PEMT: a patent enrichment tool for drug discovery. *Bioinformatics*. doi:10.1093/bioinformatics/btac716.

[14] Weinreich SS, Mangon R, Sikkens JJ, Teeuw ME, Cornel MC. Orphanet: a European database for rare diseases. Ned Tijdschr Geneeskd 2008;152(9):518–19.

[15] Lage-Rupprecht V, Schultz B, Dick J, Namysl M, Zaliani A, Gebel S, Hofmann-Apitius M. A hybrid approach unveils drug repurposing candidates targeting an Alzheimer pathophysiology mechanism. Patterns 2022;3(3):100433. doi:10.1016/j.patter.2021.100433.

[16] Hoyt CT, Konotopez A, Ebeling C. PyBEL: a computational framework for Biological Expression Language. Bioinformatics 2018;34(4):703–4. doi:10.1093/bioinformatics/btx660.

[17] Kabir E, Uzzaman M. A review on biological and medicinal impact of heterocyclic compounds. Res Chem 2022:100606. doi:10.1016/j.rechem.2022.100606.

[18] Bagley N, Berger B, Chandra A, Garthwaite C, Stern AD. The Orphan Drug Act at 35: observations and an outlook for the twenty-first century. Innov. Policy Econ. 2019;19(1):97–137. doi:10.1086/699934.

[19] Stella P, Gold-von Simson G. Pharmaceutical pricing, cost containment and new treatments for rare diseases in children. Orphanet J Rare Dis 2014;9(1):1–4. doi:10.1186/s13023-014-0152-2.

[20] Hvide HK, Jones BF. University innovation and the professor's privilege. Am Eco Rev 2018;108(7):1860–98. doi:10.1257/aer.20160284.

[21] Wang H, Norton J, Xu L, DeMartinis N, Sen R, Shah A, Lynch D. Results of a randomized double-blind study evaluating luvadaxistat in adults with Friedreich ataxia. Ann Clin Transl Neurol 2021;8(6):1343–52. doi:10.1002/acn3.51373.

[22] Ashworth AL. Race You to the patent office: how the new patent reform act will affect technology transfer at universities. Alb LJ Sci Tech 2012;23:383.

[23] Cummings JL, Goldman DP, Simmons-Stern NR, Ponton E. The costs of developing treatments for Alzheimer's disease: a retrospective exploration. Alzheimer's Dement 2022;18(3):469–77. doi:10.1002/alz.12450.

[24] Song CH, Han JW. Patent cliff and strategic switch: exploring strategic design possibilities in the pharmaceutical industry. Springerplus 2016;5(1):1–14. doi:10.1186/s40064-016-2323-1.

[25] Anderson, R.M., Hadjichrysanthou, C., Evans, S., & Wong, M.M. (2017). Why do so many clinical trials of therapies for Alzheimer's disease fail?. *The Lancet*, 390(10110), 2327–9. 10.1016/S0140-6736(17)32399-1.

[26] Koynova R, Tenchov B. Natural product formulations for the prevention and treatment of Alzheimer's disease: a patent review. Recent Pat Drug Deliv Formul 2018;12(1):23–39. doi:10.2174/1872211312666171207152326.

[27] Senior M. Fresh from the biotech pipeline: fewer approvals, but biologics gain share. Nat. Biotechnol. 2023;1. doi:10.1038/s41587-022-01630-6.

[28] Zdrazil, B., Richter, L., Brown, N., & Guha, R. (2020). Moving targets in drug discovery. Sci Rep, 10(1), 1–15. 10.1038/s41598-020-77033-x.

[29] Burwinkel B, Bakker HD, Herschkovitz E, Moses SW, Shin YS, Kilimann MW. Mutations in the liver glycogen phosphorylase gene (PYGL) underlying glycogenosis type VI (Hers disease). Am J Hum Gene 1998;62(4):785–91. doi:10.1086/301790.

[30] Matschinsky FM. Assessing the potential of glucokinase activators in diabetes therapy. Nat Rev Drug disc 2009;8(5):399–416. doi:10.1038/nrd2850.

[31] Nakamura A, Terauchi Y. Present status of clinical deployment of glucokinase activators. J Diabetes Investig 2015;6(2):124–32. doi:10.1111/jdi.12294.

[32] Costantini S, Capone F, Polo A, Bagnara P, Budillon A. Valosin-Containing Protein (VCP)/p97: a Prognostic Biomarker and Therapeutic Target in Cancer. Int J Mol Sci 2021;22(18):10177. doi:10.3390/ijms221810177.

[33] Scarian E, Fiamingo G, Diamanti L, Palmieri I, Gagliardi S, Pansarasa O. The role of VCP mutations in the spectrum of amyotrophic lateral sclerosis-frontotemporal dementia. Front Neurol 2022;271. doi:10.3389/fneur.2022.841394.

[34] Le Moigne R, Aftab BT, Djakovic S, Dhimolea E, Valle E, Murnane M, Rolfe M. The p97 inhibitor CB-5083 is a unique disrupter of protein homeostasis in models of multiple myeloma. Mol. Cancer Ther. 2017;16(11):2375–86. doi:10.1158/1535-7163.MCT-17-0233.

[35] Bastola P, Wang F, Schaich MA, Gan T, Freudenthal BD, Chou TF, Chien J. Specific mutations in the D1–D2 linker region of VCP/p97 enhance ATPase activity and confer resistance to VCP inhibitors. Cell Death Discov 2017;3(1):1–9. doi:10.1038/cddiscovery.2017.65.

[36] Ahlstedt BA, Ganji R, Raman M. The functional importance of VCP to maintaining cellular protein homeostasis. Biochem. Soc. Trans. 2022;50(5):1457–69. doi:10.1042/BST20220648.

[37] Mitchell J, Paul P, Chen HJ, Morris A, Payling M, Falchi M, de Belleroche J. Familial amyotrophic lateral sclerosis is associated with a mutation in D-amino acid oxidase. Proc Natl Acad Sci 2010;107(16):7556–61. doi:10.1073/pnas.0914128107.

[38] Sacchi S, Rosini E, Pollegioni L, Molla G. D-amino acid oxidase inhibitors as a novel class of drugs for schizophrenia therapy. Curr pharma des 2013;19(14):2499–511. doi:10.2174/1381612811319140002.

[39] Riccaboni M, Verginer L. The impact of the COVID-19 pandemic on scientific research in the life sciences. PLoS ONE 2022;17(2):e0263001. doi:10.1371/journal.pone.0263001.

[40] Kodamullil AT, Zekri F, Sood M, Hengerer B, Canard L, McHale D, Hofmann-Apitius M. Trial watch: tracing investment in drug development for Alzheimer disease. Nature reviews. Drug discovery 2017;16(12):819. doi:10.1038/nrd.2017.169.

[41] De Luca C, Virtuoso A, Maggio N, Papa M. Neuro-coagulopathy: blood coagulation factors in central nervous system diseases. Int J Mol Sci 2017;18(10):2128. doi:10.3390/ijms18102128.

[42] Zhao M, Jung Y, Jiang Z, Svensson KJ. Regulation of energy metabolism by receptor tyrosine kinase ligands. Front Physiol 2020;11:354. doi:10.3389/fphys.2020.00354.

[43] Qian M, Shen X, Wang H. The distinct role of ADAM17 in APP proteolysis and microglial activation related to Alzheimer's disease. Cell. Mol. Neurobiol. 2016;36(4):471–82. doi:10.1007/s10571-015-0232-4.

[44] Wan HI, Jacobsen JS, Rutkowski JL, Feuerstein GZ. Translational medicine lessons from flurizan's failure in Alzheimer's disease (AD) trial: implication for future drug discovery and development for AD. Clin Transl Sci 2009;2(3):242. doi:10.1111/j.1752-8062.2009.00121.x.

[45] Doody RS, Raman R, Farlow M, Iwatsubo T, Vellas B, Joffe S, Mohs R. A phase 3 trial of semagacestat for treatment of Alzheimer's disease. N Engl J Med 2013;369(4):341–50. doi:10.1056/NEJMoa1210951.

[46] Ahmad S, Saleem M, Riaz N, Lee YS, Diri R, Noor A, Elsebai MF. The natural polypeptides as significant elastase inhibitors. Front Pharmacol 2020;11:688. doi:10.3389/fphar.2020.00688.

[47] Wang H, Wang R, Zhao Z, Ji Z, Xu S, Holscher C, Sheng S. Coexistences of insulin signaling-related proteins and choline acetyltransferase in neurons. Brain Res. 2009;1249:237–43. doi:10.1016/j.brainres.2008.10.046.

[48] Geyer P. Markush structure searching by information professionals in the chemical industry–our views and expectations. World Pat Inf 2013;35(3):178–82. doi:10.1016/j.wpi.2013.05.002.

[49] Leinonen H, Cheng C, Pitkänen M, Sander CL, Zhang J, Saeid S, Palczewski K. A p97/valosin-containing protein inhibitor drug CB-5083 has a potent but reversible off-target effect on phosphodiesterase-6. J Pharmacol Exp Ther 2021;378(1):31–41. doi:10.1124/jpet.120.000486.

[50] Schlaghecke R, Kornely E, Wollenhaupt J, Specker C. Glucocorticoid receptors in rheumatoid arthritis. Arthritis Rheum Off J Am Coll Rheum 1992;35(7):740–4. doi:10.1002/art.1780350704.

[51] Stock T, Fleishaker D, Wang X, Mukherjee A, Mebus C. Improved disease activity with fosdagrocorat (PF-04171327), a partial agonist of the glucocorticoid receptor, in patients with rheumatoid arthritis: a Phase 2 randomized study. Int J Rheum Dis 2017;20(8):960–70. doi:10.1111/1756-185X.13053.

[52] Duggan S, Keam SJ. Upadacitinib: first approval. Drugs 2019;79(16):1819–28. doi:10.1007/s40265-019-01211-z.

[53] Schweig JE, Yao H, Coppola K, Jin C, Crawford F, Mullan M, Paris D. Spleen tyrosine kinase (SYK) blocks autophagic Tau degradation in vitro and in vivo. J Biol Chem 2019;294(36):13378–95. doi:10.1074/jbc.RA119.008033.

[54] Nakajima EC, Drezner N, Li X, Mishra-Kalyani PS, Liu Y, Zhao H, Singh H. FDA approval summary: sotorasib for KRAS G12C-mutated metastatic NSCLC. Clin Cancer Res 2022;28(8):1482–6. doi:10.1158/1078-0432.CCR-21-3074.

[55] Hong T, Yan Y, Li J, Radovanovic I, Ma X, Shao YW, Wang Y. High prevalence of KRAS/BRAF somatic mutations in brain and spinal cord arteriovenous malformations. Brain 2019;142(1):23–34. doi:10.1093/brain/awy307.

[56] Cummings J. Lessons learned from Alzheimer disease: clinical trials with negative outcomes. Clin Transl Sci 2018;11(2):147. doi:10.1111/cts.12491.

A.6 Causal reasoning over knowledge graphs leveraging drug-perturbed and disease-specific transcriptomic signatures for drug discovery

Reprinted with permission from "Domingo-Fernández, D., Gadiya, Y., Patel, A., Mubeen, S., Rivas-Barragan, D., Diana, C. W., Biswapriya B. M., Healey, D., Rokicki, J., and Colluru, V. (2022). Causal reasoning over knowledge graphs leveraging drug-perturbed and disease-specific transcriptomic signatures for drug discovery. *PLoS computational biology*, 18(2), e1009909.".

PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Causal reasoning over knowledge graphs leveraging drug-perturbed and disease-specific transcriptomic signatures for drug discovery

Daniel Domingo-Fernández[1]*, Yojana Gadiya[1], Abhishek Patel[1], Sarah Mubeen[2], Daniel Rivas-Barragan[3], Chris W. Diana[1], Biswapriya B. Misra[1], David Healey[1], Joe Rokicki[1], Viswa Colluru[1]*

[1] Enveda Biosciences, Boulder, Colorado, United States of America, [2] Bonn-Aachen International Center for IT, Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn, Germany, [3] Barcelona Supercomputing Center, Barcelona, Spain

* daniel.domingo-fernandez@envedabio.com (DDF); viswa.colluru@envedabio.com (VC)

OPEN ACCESS

Citation: Domingo-Fernández D, Gadiya Y, Patel A, Mubeen S, Rivas-Barragan D, Diana CW, et al. (2022) Causal reasoning over knowledge graphs leveraging drug-perturbed and disease-specific transcriptomic signatures for drug discovery. PLoS Comput Biol 18(2): e1009909. https://doi.org/10.1371/journal.pcbi.1009909

Editor: Carl Herrmann, Heidelberg University, GERMANY

Received: October 21, 2021

Accepted: February 9, 2022

Published: February 25, 2022

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pcbi.1009909

Data Availability Statement: All links to relevant data are within the manuscript under the

Abstract

Network-based approaches are becoming increasingly popular for drug discovery as they provide a systems-level overview of the mechanisms underlying disease pathophysiology. They have demonstrated significant early promise over other methods of biological data representation, such as in target discovery, side effect prediction and drug repurposing. In parallel, an explosion of -omics data for the deep characterization of biological systems routinely uncovers molecular signatures of disease for similar applications. Here, we present RPath, a novel algorithm that prioritizes drugs for a given disease by reasoning over causal paths in a knowledge graph (KG), guided by both drug-perturbed as well as disease-specific transcriptomic signatures. First, our approach identifies the causal paths that connect a drug to a particular disease. Next, it reasons over these paths to identify those that correlate with the transcriptional signatures observed in a drug-perturbation experiment, and anti-correlate to signatures observed in the disease of interest. The paths which match this signature profile are then proposed to represent the mechanism of action of the drug. We demonstrate how RPath consistently prioritizes clinically investigated drug-disease pairs on multiple datasets and KGs, achieving better performance over other similar methodologies. Furthermore, we present two case studies showing how one can deconvolute the predictions made by RPath as well as predict novel targets.

Author summary

Different types of interactions between various biological elements (e.g., proteins, drugs and diseases) can be modeled using networks for various applications, including drug discovery and finding novel use cases of known drugs. Nevertheless, we are far from having a complete picture of all possible biological interactions that can occur in humans, and so,

Implementation Details section and can be found at https://github.com/enveda/RPath.

Funding: DDF, YG, AP, CWD, BBM, DH, JR, and VC have been funded by Enveda Biosciences. This work has been funded by Enveda Biosciences (https://www.envedabio.com/). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. SM and DRB received no specific funding for this work.

Competing interests: I have read the journal's policy and the authors of this manuscript have the following competing interests: DDF, YG, AP, CWD, BBM, DH, JR, and VC are employees of Enveda Biosciences Inc. during the course of this work and have real or potential ownership interest in Enveda Biosciences Inc.

current networks modeling human biology remain incomplete. To try and compensate for this shortcoming, researchers are beginning to use both knowledge of biological interactions, alongside experimental data. In this work, we show how we can deduce which drugs may be good candidates for treatments by using networks to estimate how a drug can affect a disease, and overlaying elements in our network with those in experimental datasets. These experimental datasets can help guide us through the network, showing us which interactions are likely occurring and which are not. Finally, we show that the approach we take can also help us to come up with new research questions and determine which proteins a drug must actually target to produce a therapeutic effect in a patient.

This is a *PLOS Computational Biology* Methods paper.

Introduction

The representation of biomolecular interactions occurring within cells is often intuitively organized in the form of biological networks. These networks can be used to inherently model biological processes through the use of nodes denoting biological entities and edges representing their relationships. While homogeneous networks, such as protein-protein interaction networks, can represent relationships between a single entity type, knowledge graphs (KGs) can incorporate a broad range of biological scales, from the genetic and molecular level (e.g., proteins, drugs, and biochemicals), to biological concepts (e.g., phenotypes and diseases). These KGs can then be utilized for several applications in drug discovery, such as providing insights into molecular mechanisms and therapeutic targets [1–2], side effect prediction in the early stages of drug development [3], target prioritization [4], and drug repositioning [5].

Given the flexibility of KGs, multiple heterogeneous relation types can be modeled to represent biological processes that are governed by interactions occurring between component entities [6]. Even though a variety of relation types (e.g., literature co-occurrence, associations, etc.) can be leveraged by network-topology algorithms for various applications, causal relations are particularly useful as they can be used to infer the effect of any given node on another by reasoning over the KG [7]. Nonetheless, not all interactions included in a given KG are necessarily biologically relevant as they may be context-specific, such as to a particular cell type, tissue or disease. Furthermore, as the complete human interactome remains unknown, KGs modeling PPIs are also incomplete and the interactions which are modeled tend to be biased towards well-studied proteins and their relationships [8]. One approach to address these challenges is to jointly leverage prior knowledge in KGs with data-driven *-omics* experiments [9–13].

Experimental datasets have been widely employed by recent drug repurposing approaches to identify drug candidates for a given disease using the anti-correlation in biological processes or pathways at the transcriptomic- or proteomic- level between drugs and diseases as a proxy [13–16]) (see [17] for a recent review and **S8 Table** for a survey of such methods). While these approaches use prior knowledge in the form of pathways (gene sets), this concept has yet to be applied on KGs for drug discovery. However, by mapping the signatures of an *-omics* experiment to a KG, we can not only verify which causal interactions are observed within a specific context, but also prioritize and identify the mechanism of action of a drug for a given disease with high precision.

Currently, there exist numerous algorithms that leverage causal relations for the interpretation of *-omics* data. In general, these algorithms operate by assessing the concordance between transcriptomic or proteomic signatures and the predicted causal effects encoded in these relations [18–19]. For instance, the Reverse Causal Reasoning (RCR) [20] and Network Perturbation Amplitude (NPA) algorithms [21–22] assess and score this concordance employing causal graphs consisting of up-stream and down-stream proteins (nodes) representing regulations occurring in biological pathways. Subsequently, the scores obtained from these algorithms can be used for the interpretation of *-omics* data commonly derived from contrast experiments. Although the interpretations obtained from these algorithms may be relevant for several downstream applications, such as drug target prediction, disease characterization, and side effect prediction, the algorithms themselves cannot be directly used for these applications. Additionally, these algorithms have been specifically designed for bipartite graphs, thus, simplifying biological pathways to a single relation between two proteins.

While traditionally, these algorithms were applied on small causal networks, they have recently begun to be applied on large-scale KGs, given the increasing availability of causal information, including proteins, drugs and phenotypes. For instance, a recent algorithm we published, drug2ways, reasons over all paths between a drug and a disease in a KG to predict the effect of the drug as the cumulative effect of all directed interactions between these two nodes [23]. Reasoning over all paths overcomes the limitation of earlier algorithms that exclusively account for shortest paths on protein-protein interaction networks, oversimplifying the effect exerted by one node on another, as all other paths between the two nodes are ignored [24–25]. Nonetheless, paths in large-scale KGs can grow exponentially, many of which may not be relevant in a true biological context. Thus, incorporating signatures from context-specific experimental datasets along with prior knowledge in a KG can enable us to reason over the entire network-structure and ensure only paths which can be observed in a biologically meaningful context are retained. In doing so, we can address several of the limitations of the above-mentioned methods for drug discovery.

Here, we present RPath, a novel algorithm that prioritizes drugs for a particular disease by reasoning over causal paths in a KG, guided by both drug-perturbed and disease-specific transcriptomic signatures (**Fig 1**). We demonstrate how RPath is able to recover a large proportion of clinically investigated drug-disease pairs on multiple transcriptomic datasets and KGs, performing better than other network-based methods. Furthermore, we show two additional applications where we illustrate how our approach can also assist in hypothesis generation and target prioritization.

Results

This section is divided into three subsections that outline the different applications of RPath presented in this manuscript. First, we demonstrate how RPath can be used to identify potential drug candidates for various diseases using a variety of KGs and datasets, outperforming numerous link prediction methods. Next, we leverage the inherent interpretability of KGs to generate hypotheses for the predictions made by RPath. Finally, we outline how RPath can be reversed-engineered and alternatively used to predict targets for a given disease.

Identification of drug candidates

To demonstrate the ability of our algorithm to accurately identify drug candidate for a given disease, we evaluated its performance to recover clinically investigated drug-disease pairs using two distinct KGs and four transcriptomic datasets (i.e., two each containing numerous

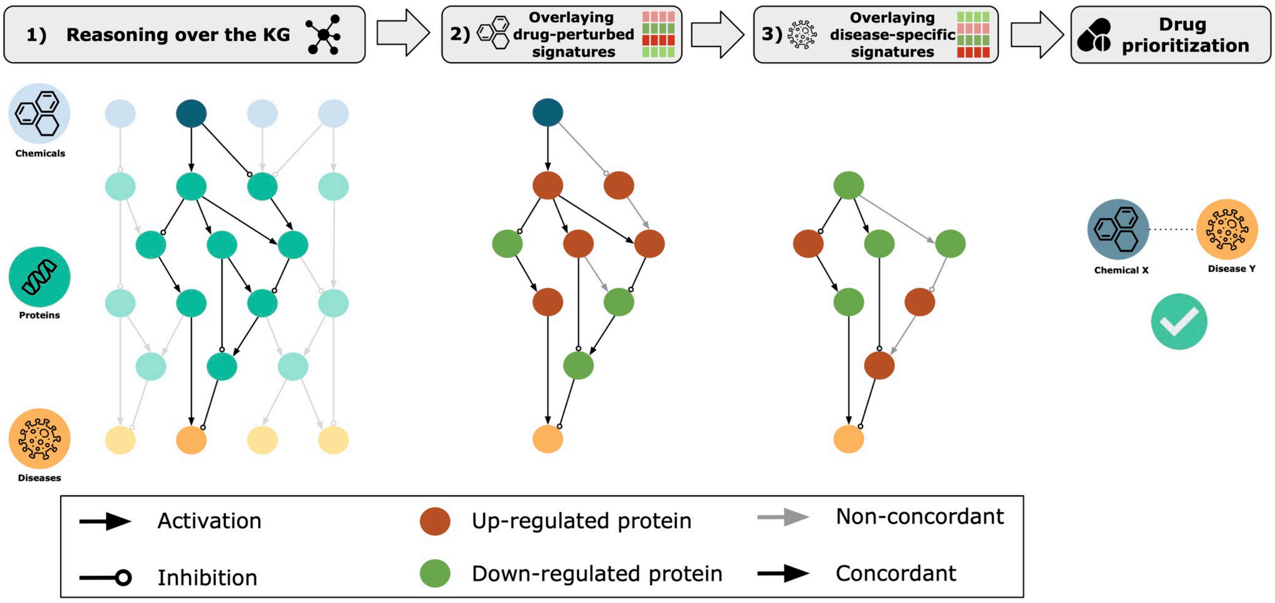


Fig 1. Schematic representation of the RPath algorithm. Step 1) All acyclic paths of a given length between a drug and a disease in the KG are calculated. If there exist causal acyclic paths connecting the drug and the disease, a subgraph involving all these paths is inferred. This subgraph represents the proposed mechanism of action by which the drug may be a therapeutic target of the given disease. **Step 2)** Transcriptomic signatures observed from a drug-perturbed experiment are overlaid onto each corresponding node present in these paths. Then, RPath traverses through each path and evaluates whether the inferred direction of regulation (i.e., activation or inhibition) at every step is concordant with the up- and down- regulations (i.e., red and green nodes, respectively) observed in the transcriptomic signatures. **Step 3)** In a similar manner, transcriptomic signatures observed within a specific disease context are overlaid onto each corresponding node in the concordant paths from the previous step (if any). Next, RPath evaluates whether the disease transcriptomic signatures contradict the paths that were concordant with the drug signatures. If this is the case, the specific drug-disease pair is prioritized.

https://doi.org/10.1371/journal.pcbi.1009909.g001

drug-perturbed and disease transcriptomic experiments). In this task, RPath consistently prioritized a significantly larger number of clinically investigated drug-disease pairs across all datasets and in both KGs compared with the precision expected by chance (i.e., probability of randomly picking a positive label among drug-disease combinations that are connected through a path) (Table 1).

The highest precision values were found for the L1000-GEO datasets with 80% and 66.67% for the OpenBioLink and custom KGs, respectively. In the remaining datasets, the precision was approximately 50%, except for the CREEDS-Open Targets datasets in the custom KG that exclusively yielded a single drug-disease pair which was not in clinical trials. While the precision expected by chance approximately varied between 10% and 42%, RPath consistently achieved higher precision values across nearly all datasets and KGs, ranging between 50% and 80% (e.g., more than five times higher for the L1000-GEO dataset in OpenBioLink running

Table 1. Evaluation of RPath in multiple datasets across the two KGs using precision. Each row corresponds to the results of running RPath on a specific drug-disease dataset combination. The second and fourth columns show the performance that is expected to be achieved by chance.

-	OpenBioLink KG		Custom KG	
Dataset combination	**Precision (TP/TP+FP)**	**Expected precision by chance**	**Precision (TP/TP+FP)**	**Expected precision by chance**
L1000 [26]–GEO [27]	80% (4/5)	17.42%%	66.67% (2/3)	13.74%
L1000 –Open Targets [28]	54.55% (6/11)	15.01%	50% (2/4)	9.62%
CREEDS [29]–Open Targets	50% (1/2)	32.66%	0% (0/1)	24.40%
CREEDS–GEO	50% (1/2)	41.15%	50% (1/2)	34.08%

https://doi.org/10.1371/journal.pcbi.1009909.t001

RPath (80%) vs. chance (17.42%)). Notably, the number of prioritized drug-disease pairs were constrained for two reasons: i) RPath requires transcriptomic information for a given drug and disease and, ii) the pair must also be present in the KG **(see S1 Table for details)**. Furthermore, apart from the low number of drug-disease pairs that fulfilled these criteria, RPath filters the majority of pairs with paths between them after overlaying the transcriptomic signatures in Steps 1 and 2 **(see Fig 1)** of the algorithm (**S2 Table**). For example, in the case of the CREEDS-GEO datasets and the OpenBioLink KG, the total number of diseases was 10, resulting in only a couple of drug-disease pairs being prioritized. Nonetheless, we were still able to validate our methodology across multiple datasets and KGs, observing that RPath performed significantly better than chance at identifying clinically investigated drug-disease pairs.

Finally, we benchmarked RPath against 11 alternative methods [30–31] that have been used to predict drug-disease links in a KG with the same characteristics as the ones used in this work. The precision of these methods varied between 5% and 43% **(S3 Table)**. Furthermore, since the majority of these methods prioritize a drug and a disease based on their network proximity (e.g., shortest paths and number of shared nodes), these methods recurrently prioritized the same set of drug-disease pairs. Thus, these methods could not be used to prioritize drugs outside the vicinity of disease-associated proteins since only a minority of drug-disease pairs are connected by a single protein, but the majority of them contain longer paths that are not considered by these methods (**S2 Table).** Lastly, we also conducted permutation experiments where we permuted both the binarized gene expression values (i.e., +1, -1, 0) observed in the transcriptomic datasets and the edges of the KGs, while maintaining their underlying structure. The results of our experiments showed how the number of prioritized drug-disease pairs significantly decreases when permuted datasets and KGs are employed and that none of these few prioritized pairs were clinically investigated (**S9 Table**).

Interpretation of the mechanisms of action of the proposed drug candidates

In a case study, we sought to explore the results obtained by running RPath on the custom KG. Of the prioritized drug-disease pairs (see **S3 Text**), we studied the paths between two of the pairs to demonstrate how our approach can potentially be used to deconvolute the mechanism of action of some drugs **(Fig 2).** We selected bicalutamide and ponatinib as these two anti-cancer drugs were the top-ranked prioritized drugs and already approved for prostate cancer and acute myeloid leukemia, respectively. Furthermore, since the mechanisms of action of these drugs have been widely studied, we can compare the mechanistic paths identified by RPath against known interactions and pathways reported in scientific literature.

First, we investigated ponatinib, a multi-targeted tyrosine-kinase inhibitor, which is used to treat acute myeloid leukaemia (AML) (**Fig 2B**). Among the targets of this drug present in the concordant paths for this pair, we were able to identify fms-like tyrosine kinase 3 (FLT3), which is mutated in approximately 20% of AML patients [32] and several members of the FGFR family proteins. Furthermore, we observed other proteins including KDR, LYN, and SRC, all of which are kinase-associated targets in AML. As a downstream target of these proteins, we found JAK2, a well-studied player in myeloproliferative diseases, with known mutations and hypermethylation events. We further identified the transcription factor, CEBPA, that is critical for normal development of granulocytes and is also implicated in AML [33] and the SPI1 gene, from which circSPI1, a circular RNA derived from the gene, has recently been shown to be highly expressed in AML patients [34]. Other proteins that are inhibited as a result of the signaling cascade triggered by ponatinib include KIT, which is implicated in cell death in AML [35]. RAS family members NRAS and KRAS, both of which are associated with the

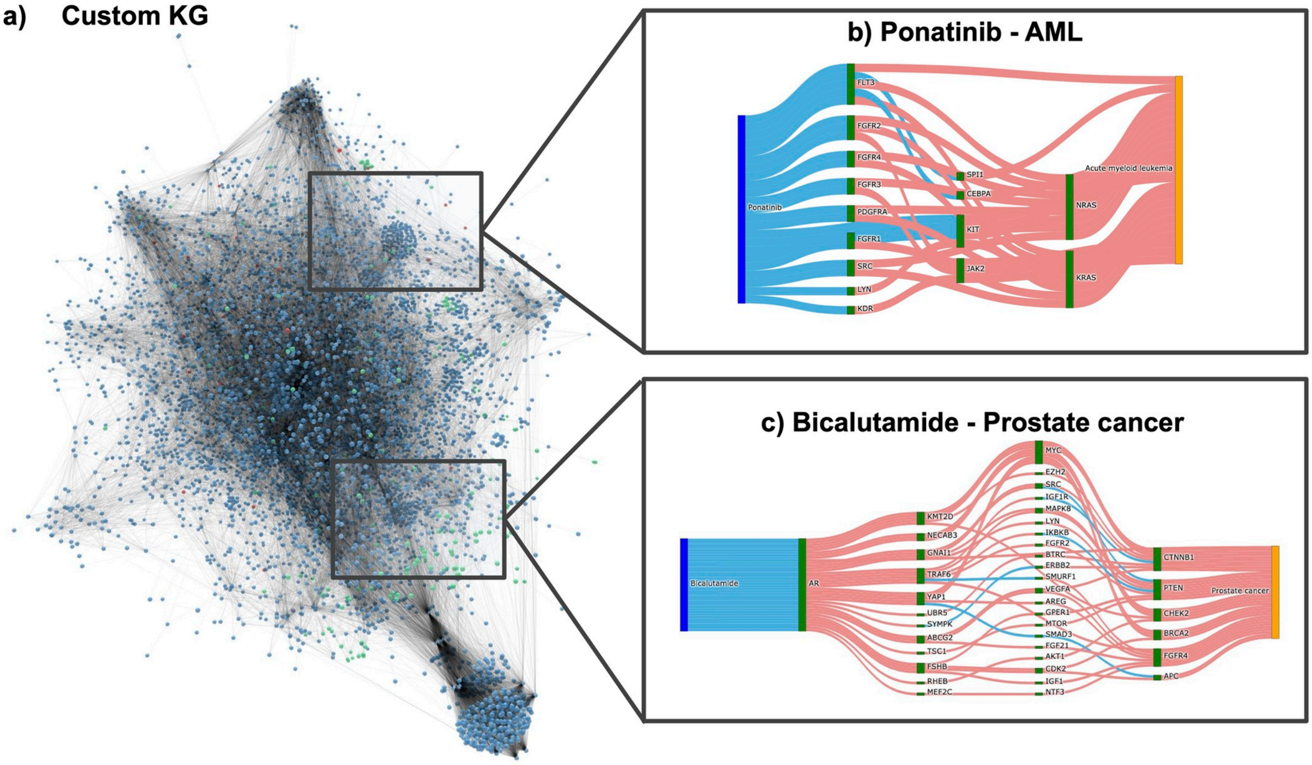


Fig 2. Devoncoluting the mechanism of action of a drug through RPath. By investigating all the paths of a given length between a drug and a disease in a KG, we can analyze the different mechanisms that are proposed by RPath. **a)** Visualization of the custom KG. Proteins are colored in blue, diseases in red and drugs in green. Sankey diagram illustrating a sample of the paths between ponatinib and AML **(b)** and bicalutamide and prostate cancer **(c)** for the custom KG. Activatory relations in the Sankey diagrams are colored in red and inhibitory relations in blue.

https://doi.org/10.1371/journal.pcbi.1009909.g002

prognosis of solid tumors and hematological malignancies, including AML [36] were also implicated.

The second studied drug-disease pair is bicalutamide, used for the treatment of prostate cancer. Bicalutamide is an anti-androgen medication that binds to the androgen receptor (AR), as illustrated in Fig 2C. The paths between bicalutamide and prostate cancer point to several downstream targets of this drug, including the epigenetic regulator KMT2D, which is known to sustain prostate carcinogenesis by epigenetic mechanisms [37], and NECAB3, known to enhance the activity of HIF1A, thus promoting glycolysis under normoxic conditions and enhancing tumorigenicity in cancer cells [38]. Furthermore, we were able to identify CTNNB1, which plays a role in the development of numerous prostate cancers [39]. Interestingly, we also observed novel players that have not yet been reported in the literature, such as GNAI1, SYMPK, UBR5, and MEF2C that may provide new insights on the mechanism of action of this drug.

Target prioritization

Prior to the identification of a therapeutic drug candidate for any given disease, a crucial first step is often to identify biologically relevant protein targets. Ideally, the perturbation of a particular protein target in a disease state should result in the reversal of the observed phenotype. In a similar manner to the above-mentioned applications, by reasoning over the KG guided by disease signatures, RPath can be used for target prioritization. Since, as per our knowledge,

Table 2. Top 5 prioritized protein target-disease pairs. These results were obtained by running RPath over both KGs with the GEO and Open Targets datasets using the same path length as the drug discovery task **(see Methods)**. Pairs were prioritized based on the number of concordant paths. The vast majority of pairs were prioritized using the disease transcriptomic signatures from the GEO dataset given its larger coverage of measured genes compared to Open Targets (**S4 Table**).

Protein target	Disease	Concordant paths	Nodes in the concordant paths	KG	Transcriptomic dataset
NOG	AML	18,456	1,008	Custom KG	GEO
PRKCA	AML	12,861	669	Custom KG	GEO
CXCL8 / IL-8	AML	7,234	465	Custom KG	GEO
NOG	Plasma cell myeloma	5743	616	Custom KG	GEO
CDC42	Medulloblastoma	5,651	91	OpenBioLink	GEO

https://doi.org/10.1371/journal.pcbi.1009909.t002

there are no large datasets that contain information about known targets for a wide variety of indications, we could not conduct a validation strategy similar to the analyses presented in the subsection *Identification of drug candidates*. Instead, we focused on evaluating the top prioritized protein targets across all diseases using literature evidence (**Table 2**).

Among the top protein target-disease pairs proposed by RPath, two have already been associated with AML, including PRKCA, for which several drugs already exist [40–41] and CXCL8/IL-8 [42–44]. Furthermore, CDC42, which has been proposed as a candidate target for medulloblastoma, plays a role in several cancers. Specifically, CDC42 has been shown to act as a regulator of medulloblastoma-associated genes [45] and compounds for its inhibition have also been proposed [46].

Discussion

In this work, we present a novel methodology that leverages prior knowledge from causal relations across multiple biological modalities in KGs and assesses their concordance with transcriptomic signatures for drug discovery. In the past, several algorithms have been primarily introduced for the interpretation of transcriptomic signatures by reasoning over shortest paths [24–25] or bipartite graphs [20–22]. Though these algorithms could also be indirectly applied for drug discovery, they present some shortcomings: i) they operate on homogeneous causal graphs with a single entity type (e.g., protein nodes), ii) they are solely conducted on single contrast experiments (e.g., drug-treated vs. control), and iii) they do not fully exploit all possible paths in these causal graphs. RPath addresses these shortcomings by reasoning over all possible causal paths in a multimodal KG and leveraging both drug and disease transcriptomic signatures. First, our algorithm reasons over the ensemble of paths between a given drug and a disease in a KG. Second, it evaluates the concordance of these paths against the transcriptomic changes experimentally observed for that drug. Third, it assesses whether the effect of these paths is opposite to the transcriptomic signatures observed within the disease context. In a final step, the algorithm identifies potential drug candidates as those whose paths correlate with drug-perturbed transcriptomic signatures and are anti-correlated to the disease transcriptomic signatures. We have validated our methodology in eight independent analyses, finding that RPath consistently identifies a large proportion of clinically investigated drug-disease pairs over multiple datasets and KGs. Additionally, we conducted several robustness experiments and benchmarked the algorithm against 11 network-based methodologies. Finally, we also showed how our approach can be used to deconvolute the mechanism of action of a drug as well as to prioritize protein targets for a given disease.

We acknowledge a few shortcomings in our work that are worth discussion. Firstly, we were limited by the availability of high-quality annotated transcriptomic datasets for drugs and diseases, as only four of the approximately 30 datasets that we identified met our requirements. Furthermore, the coverage of measured genes varied largely across experiments. For instance, while the average number of genes measured in the Open Targets dataset was approximately 900, that number dropped to 500 in the CREEDS dataset (**S4 Table**). In contrast, the total number of proteins in the KGs were in the range of several thousands. As RPath requires that signatures from both these drug and disease datasets be mapped to the KG, most of the proteins in the KGs could not be quantified. Thus, we allowed for up to one error when calculating the concordance in the path between a drug and a disease. Furthermore, two other reasons justified an error within the path. Firstly, introducing an error limits the impact of an arbitrary fold change cut-off, which ultimately determines the up-/down-regulation of each protein. Secondly, some paths might contain causal relations that do not reflect a change at the transcription level of the affected protein (e.g., phosphorylation of a protein kinase) [18–19]. We expect that this challenge we faced of quantifying proteins in our KGs will be overcome by high-quality, consistent datasets such as those generated in large pharmaceutical enterprises and emerging data-driven biotech companies looking to leverage large-scale computational technologies. Another characteristic of our approach is that the identification of a potential drug for a given disease requires knowledge of the protein target and the effect of the drug on it. However, this information is not always available or must be inferred using computational approaches. Finally, the interpretation of the mechanism of action of a proposed drug with the help of scientific literature comes with the caveat that the individual interactions were themselves derived from the literature. Nonetheless, it is still possible to interpret the mechanism of action of a drug irrespective of the aforementioned limitation as the paths of the proposed drug-disease pairs include only those which are concordant with observed data-driven transcriptomic signatures.

While we have demonstrated our novel algorithm across multiple datasets and KGs, we envision multiple other applications. Firstly, by incorporating time series data into the analysis, we can determine how the paths between the drug and the disease are altered over time following the concept outlined by [47]. Secondly, although we have demonstrated our methodology using transcriptomic data, other modalities can be used if the KG contains causal relations for these entities (e.g., metabolomics). Additionally, although we have employed transcriptomic signatures in this work, we acknowledge that RNA levels may not directly reflect the functional activity of proteins. However, given the growth in the availability of proteomic data, we envisage the application of our approach on proteomic experiments from databases such as PRIDE [48], ProteomicsDB [49], and L1000 [26] in the future. Furthermore, although a multimodal KG may lack the context within which each relation occurs, RPath inherently takes this into account by removing the paths which do not match the observed transcriptomic signatures. However, the algorithm could also be applied on a disease-specific KG in order to model the pathophysiological mechanisms characteristic of a given phenotype [50–51].

Methods

Theoretical background

We denote a KG as a set of nodes and edges, where nodes correspond to three distinct biological entities (i.e., chemicals, proteins, and diseases) connected through causal relations, representing activatory or inhibitory effects. Causal relations within the KG connect drug-protein, protein–protein, and protein–disease nodes. A (directed) path in a KG is defined as a sequence of two or more biological entities connected through causal relations. Paths in the KG can be either cyclic or simple. A cyclic path refers to paths in which one or more nodes repeat, whereas a simple path corresponds to a path in which no nodes appear more than once. The length of a path is defined by the number of edges that connect the nodes within the path.

RPath algorithm

The algorithm used in our framework, RPath, reasons over the paths in a KG to identify all possible effects a given drug can have on a disease (**Fig 1**). Each of these paths can be divided into three main sequential parts that attempt to represent the mechanism of action of a drug: i) the drug activates/inhibits a protein target (drug-protein edge), ii) the protein target triggers a signaling cascade (a set of protein-protein edges), and iii) the signaling cascade reverts the disease condition (protein-disease edge). Furthermore, since every causal edge contains information on the effect each node exerts on another (i.e., activation or inhibition), we can infer the direction of regulation (i.e., up-/down-regulated) for each node at each step of a path [24–25].

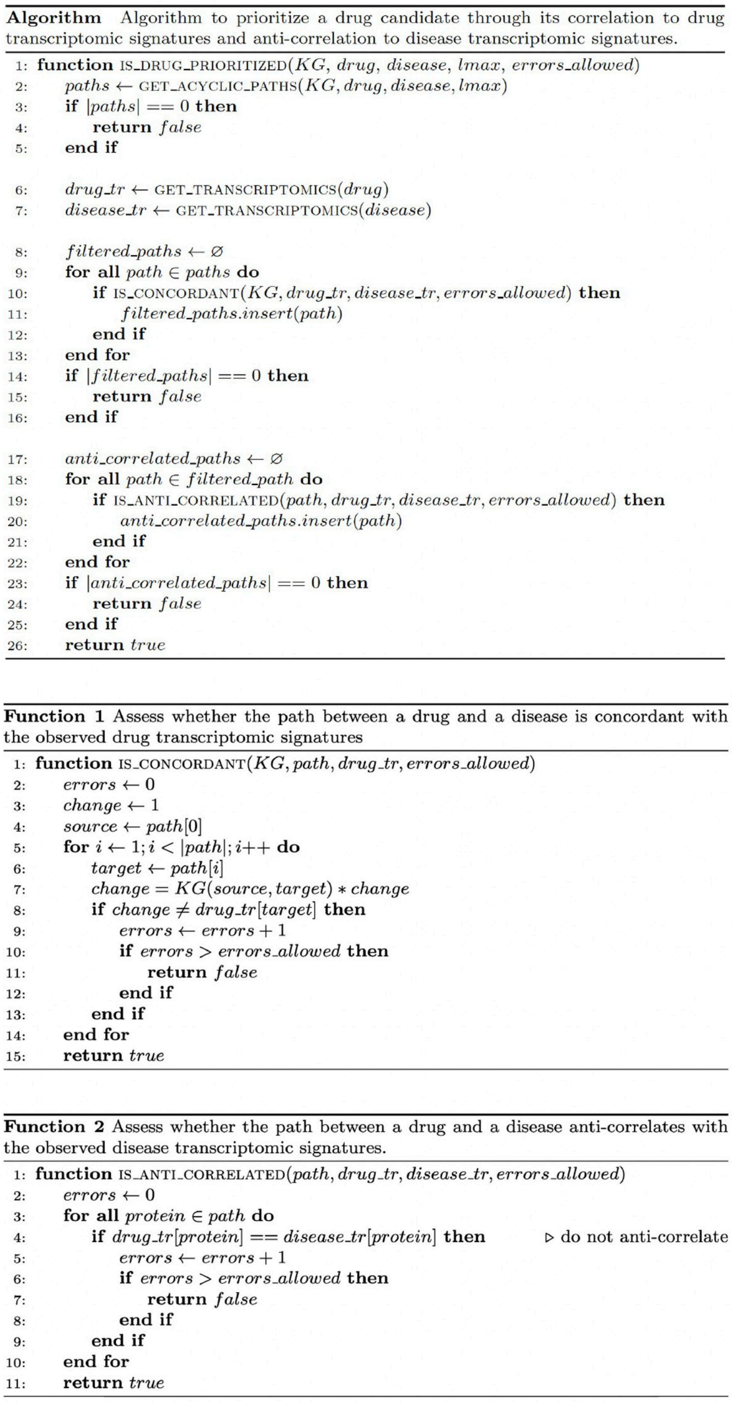
Once the causal acyclic paths between a particular drug and disease in the KG have been calculated (**Fig 1; step 1**), the next step of RPath is to overlay transcriptomic signatures from a drug-perturbed experiment (**Fig 1; step 2**). We hypothesize that because a number of paths might represent the biologically relevant mechanism of action of this drug, the observed transcriptomic signatures for proteins in the KG should be concordant with the inferred up- or down-regulations at every step of the path. For example, if in a given path, a drug inhibits a protein target and that target activates a signaling cascade, we expect the inhibition of the protein target as well as the inhibition of the proteins downstream of the target. We would like to note that a gene is considered to be differentially expressed if its expression is significantly altered with respect to a reference sample (i.e., control). Keeping this in mind, a cut-off is applied to each measured gene in the experimental dataset based on the fold change; this measurement is used to define whether the gene is up-/down-regulated or unchanged.

Similarly, the final step of RPath involves overlaying disease-specific transcriptomic signatures to the nodes in the paths of the KG (**Fig 1; step 3**). We hypothesize that, in contrast to the overlaying of drug-perturbed signatures, transcriptomic signatures in a disease context should be anti-correlated to both the drug-perturbed signatures as well as the inferred up- or down-regulations for every node in the path. This final step is inspired by previous work that exploited the anti-correlation between drug and disease signatures at the pathway level for drug repurposing [15–16]. In summary, RPath aims at prioritizing a specific drug for a given disease if i) there exist causal paths between the drug and disease in the KG, ii) the causal effects on these paths are aligned with the transcriptomic changes observed in the drug-perturbed experiment, and iii) both the drug signatures and the paths are anti-correlated with the transcriptomic dysregulations observed in the disease. **Fig 3** outlines the pseudocode of the described logic of the algorithm.

As an additional application, the algorithm can be modified following the same logic for target prioritization **(see S1 Fig for the pseudocode).** This variant of the algorithm begins from a disease of interest and calculates all paths from the disease to all proteins for a given path length (e.g., a path length of 6). Next, it calculates the concordance between the paths for each potential protein target and the transcriptomic signatures of the given disease to assess whether there are proteins that could be key up-stream regulators of the observed phenotype. We would like to note that this application exponentially increases the running time of the algorithm as it requires querying paths from a disease to several thousands of proteins in the KG, as opposed to only a handful of chemicals.

Datasets and validation

In this subsection, we present drug-perturbed and disease-specific transcriptomic datasets as well as the KGs used to demonstrate our methodology. We then introduce the strategy we follow to validate our methodology.



Algorithm Algorithm to prioritize a drug candidate through its correlation to drug transcriptomic signatures and anti-correlation to disease transcriptomic signatures.

```
function IS_DRUG_PRIORITIZED($KG$, $drug$, $disease$, $lmax$, $errors\_allowed$)
    $paths \leftarrow$ GET_ACYCLIC_PATHS($KG, drug, disease, lmax$)
    if $|paths| == 0$ then
        return $false$
    end if

    $drug\_tr \leftarrow$ GET_TRANSCRIPTOMICS($drug$)
    $disease\_tr \leftarrow$ GET_TRANSCRIPTOMICS($disease$)

    $filtered\_paths \leftarrow \varnothing$
    for all $path \in paths$ do
        if IS_CONCORDANT($KG, drug\_tr, disease\_tr, errors\_allowed$) then
            $filtered\_paths.insert(path)$
        end if
    end for
    if $|filtered\_paths| == 0$ then
        return $false$
    end if

    $anti\_correlated\_paths \leftarrow \varnothing$
    for all $path \in filtered\_path$ do
        if IS_ANTI_CORRELATED($path, drug\_tr, disease\_tr, errors\_allowed$) then
            $anti\_correlated\_paths.insert(path)$
        end if
    end for
    if $|anti\_correlated\_paths| == 0$ then
        return $false$
    end if
    return $true$
```

Function 1 Assess whether the path between a drug and a disease is concordant with the observed drug transcriptomic signatures

```
function IS_CONCORDANT($KG, path, drug\_tr, errors\_allowed$)
    $errors \leftarrow 0$
    $change \leftarrow 1$
    $source \leftarrow path[0]$
    for $i \leftarrow 1; i < |path|; i{+}{+}$ do
        $target \leftarrow path[i]$
        $change = KG(source, target) * change$
        if $change \neq drug\_tr[target]$ then
            $errors \leftarrow errors + 1$
            if $errors > errors\_allowed$ then
                return $false$
            end if
        end if
    end for
    return $true$
```

Function 2 Assess whether the path between a drug and a disease anti-correlates with the observed disease transcriptomic signatures.

```
function IS_ANTI_CORRELATED($path, drug\_tr, disease\_tr, errors\_allowed$)
    $errors \leftarrow 0$
    for all $protein \in path$ do
        if $drug\_tr[protein] == disease\_tr[protein]$ then        ▷ do not anti-correlate
            $errors \leftarrow errors + 1$
            if $errors > errors\_allowed$ then
                return $false$
            end if
        end if
    end for
    return $true$
```

Fig 3. Pseudocode of the RPath algorithm. Given a KG, drug, disease and a defined path length (i.e., *lmax)*, the core function of the algorithm, *is_drug_prioritized*, returns whether a drug should be prioritized or not. For this, the function calculates all acyclic paths between a drug-disease pair in the KG. For each path found, drug-perturbed (i.e., drug_tr) and disease-specific (disease_tr) transcriptomic signatures are overlaid onto their corresponding protein nodes. The function then prioritizes the drug if at least one path is concordant with the observed drug-perturbed transcriptomic signatures (evaluated via **Function 1**, *is_concordant*) and the same path is anti-correlated with the observed disease-specific transcriptomic signatures (evaluated via **Function 2**, *is_anti_correlated*). Paths which match both the drug-perturbed signatures and contradict disease-specific signatures are then returned by RPath as promising drug candidates.

https://doi.org/10.1371/journal.pcbi.1009909.g003

Drug-perturbed and disease transcriptomic datasets. We identified four databases that were suitable for our approach (**S5 Table**); drug-perturbed transcriptomic data were obtained from CREEDS [29] and L1000 [26] while disease transcriptomic data were collected from Open Targets [28] and GEO [27]. All experimental datasets from these resources (downloaded on 15.02.2021) contained gene expression changes measured in humans. Drugs and diseases from datasets obtained from these databases were then mapped to PubChem compound identifiers and the Mondo Disease Ontology (MONDO), respectively, for consistency with the entities of the knowledge graphs presented in the next subsection. Similarly, gene identifiers in all datasets were harmonized to ENTREZ. Of the four databases, datasets from L1000 contained a binarized value for the direction of dysregulation for every gene (i.e., up-regulation and down-regulation), while for the remaining databases, fold changes were binarized for significantly dysregulated genes using $|\log_2 \text{fold change}| = 1$ as a cutoff (**S1 Text**). As fold change thresholds tend to be arbitrary selected [52], we opted to select a threshold of 1 as opposed to a more stringent one (e.g., $|\log_2 \text{fold change}| > 2$) to ensure a larger number of dysregulated genes would be retained. Finally, we conducted a systematic search for databases that contained either a large number of drug-perturbed or disease-specific transcriptomic datasets. While this search initially resulted in 27 candidate databases **(see S5 Table for details about each dataset)**, the majority of them were not suitable for our study as they either contained too few transcriptomic datasets or the drugs/diseases in these datasets were not in the KGs used to demonstrate our methodology.

Knowledge graphs. We demonstrate our methodology using two established publicly available KGs that contain causal relations across drugs, proteins, and diseases: OpenBioLink KG [53] and a custom KG [23]. Both KGs are originally generated from a compedia of independent databases; thus, containing unique causal interactions depending on the source databases they include. As outlined in the algorithm, the KGs are required to encompass three types of causal edges: drug-protein (i.e., drug activates/inhibits protein), protein–protein (i.e., protein activates/inhibits protein), and protein–disease (i.e., protein activates/inhibits disease). Furthermore, the original node identifiers for drugs and diseases in both KGs were respectively mapped to PubChem compound identifiers and MONDO to be consistent with the transcriptomic datasets. Next, we removed drugs and diseases that were not present in any of the four transcriptomic datasets presented in the previous subsection as the paths between these drug-disease pairs cannot be validated. **Fig 4** shows the final statistics of both KGs after the previously outlined filtering steps. **S4 and S6 Tables** summarize the overlap between the genes

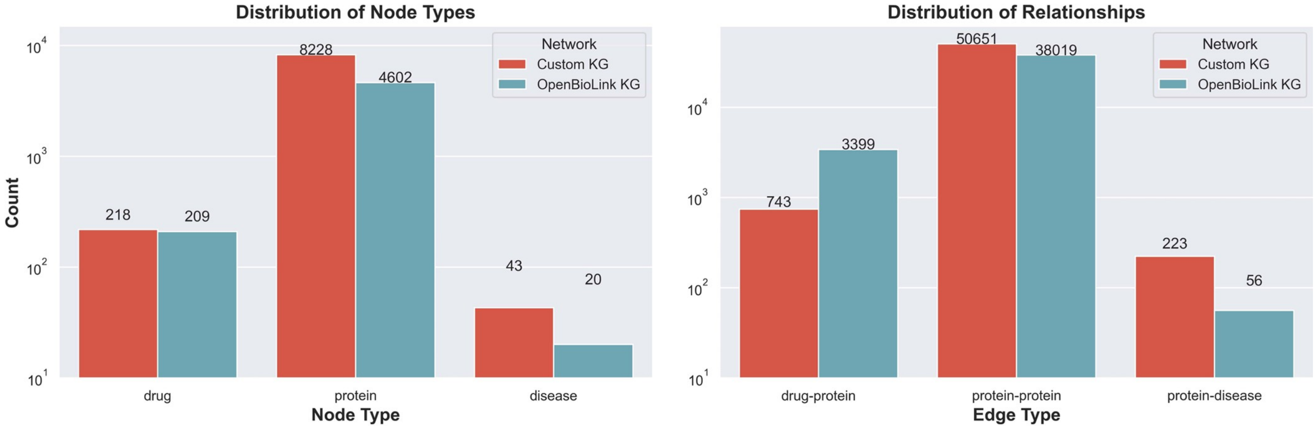


Fig 4. Distribution of node and edge types in the custom and OpenBioLink KGs. The properties of each of the two networks are detailed in S7 Table.

https://doi.org/10.1371/journal.pcbi.1009909.g004

measured in each of the four transcriptomic datasets and their corresponding protein nodes in the KGs.

Validation. In line with other network-based approaches designed for drug discovery [54–55], we used drug-disease pairs that have been clinically investigated as positive labels, extracting this information from ClinicalTrials.gov (accessed on 28.09.2020). Clinical trials are commonly used as a proxy for highly validated, medically relevant biological interactions, independently of whether the clinical trial was successful or not, as multiple *in vitro* and *in vivo* studies must first validate the interaction in order for the drug to proceed to a clinical trial. Notably, this assumption may result in a worse performance being reported than the actual performance of the algorithm as some of the drug-disease pairs that are considered as negative labels may in fact be positive ones.

Since drugs and diseases in ClinicalTrials.gov are formalized using MeSH identifiers, we harmonized these identifiers to the ontologies used in the KG (i.e., PubChem compounds for drugs and MONDO for diseases) (mappings are available at the GitHub repository). After the harmonization, any drug or disease in the KGs that was not present in any clinical trial or did not have any path to any disease in the KG was subsequently removed, as its corresponding node could not be used in the presented validation. Details on the harmonization procedure are provided in **S1 Table.**

To test the robustness of RPath in identifying these clinically investigated drug-disease pairs, we conducted eight independent analyses, one for each of the combinations of the two drug datasets, the two disease datasets, and the two KGs (e.g., CREEDS-GEO-OpenBioLink, L1000-GEO-OpenBioLink, etc.). For each of these eight analyses, we ran RPath over a given KG to prioritize drug-disease pairs among all possible drug-disease combinations. We would like to note that these pairs prioritized by the algorithm are those whose paths are both correlated with drug-perturbed transcriptomic signatures and anti-correlated with disease transcriptomic signatures. Furthermore, we make two assumptions in the design of the algorithm. Firstly, paths with cycles or a length greater than 7 edges between a given drug and disease are not considered, assuming that the effects exerted by paths beyond this length are less biologically relevant [23]. Secondly, we allow for at most one error between the transcriptomic data and a given path (see pseudocode of the algorithm in **Fig 3**). We refer to an error in the path as the disagreement between the type of causal interaction (i.e., activation or inhibition) and the direction of dysregulation of genes in the transcriptomic datasets (i.e., up or down -regulation), or the absence of the drug-perturbed and/or disease-specific transcriptomic signature. We restricted the number of allowed errors to at most one as, without any errors, running the algorithm over the two KGs with most dataset combinations will not yield any prioritized pairs, and permitting more than one error will result in an exponential increase in the number of prioritized pairs. For example, in the latter case, for a path of length 5 (i.e., a sequence of 3 proteins), an allowance of two errors (e.g., missing expression values for 2 of the 3 proteins) would still result in the prioritization of the drug-disease pair if the remaining protein both correlated with the drug and anti-correlated with the disease, obfuscating results.

From this set of prioritized drug-disease pairs, we expect to retrieve a larger proportion of clinically investigated drug-disease pairs (i.e., positive labels) than expected by chance (i.e., proportion of positive labels in the dataset that also have a path between the drug and disease). Here, it is important to note that, as in any drug discovery task, there is a class label imbalance where the vast majority of the drug-disease pairs are negative labels while the proportion of positive labels ranges from anywhere between 9% and 41% for each of the eight analyses (**S4 Table**). Furthermore, this type of validation falls into the so-called early retrieval problem. In other words, from the thousands of drug-disease pairs that are tested, we are exclusively prioritizing the top-ranked pairs that have been equally prioritized by the algorithm. This small subset represents the interesting drug-disease pairs that would be further investigated in the drug discovery process. In such cases, it is inadequate to apply metrics such as receiver operating characteristic (ROC) curves as they operate on a full ranked list. Therefore, it does not necessarily evaluate the ability of a model to prioritize the most promising drug-disease pairs candidates [56]. Additionally, considering that not all drug-disease pairs have been clinically studied, a number of the negative labels might be falsely classified as positive. To address these issues, we evaluated the performance of RPath based on the ratio of true positives that appear in the prioritized drug-disease pairs (i.e., precision was used as the performance metric). As a baseline, we assessed whether the prioritized drug-disease pairs found through the algorithm contain a larger proportion of positive labels (i.e., drug-disease pairs investigated in clinical trials) than expected on average by chance.

As a benchmark, we compared RPath against 11 equivalent approaches that can be used to prioritize drug-disease pairs based solely on network structure, as outlined by [26] and [27] (**S2 Text**). The choice of these approaches is motivated by the fact that, as per our knowledge, there are no network-based methods that operate on multimodal KGs using transcriptomic signatures for the prioritization of drug-disease pairs. Additionally, we conducted a validation experiment where we simultaneously randomly permuted the directionality of the genes measured in the transcriptomic datasets and the KGs using the XSwap algorithm [57] while both preserving network structure and the original gene expression distributions. Using these, we then rerun the eight analyses to compare the significance of our results [57].

Implementation details

The RPath algorithm and the benchmarked methods are implemented in Python leveraging NetworkX (v2.5) (https://networkx.github.io). Network visualizations were done using WebGL, D3.js, Three.js, Matplotlib and igraph. Source code, documentation, and data are available at https://github.com/enveda/RPath. The validation presented in the paper can be reproduced by running the Jupyter notebooks available at https://github.com/enveda/RPath/tree/master/notebooks.

Supporting information

S1 Text. Processing of transcriptomic datasets.
(DOCX)

S2 Text. Benchmarked methods.
(DOCX)

S3 Text. Prioritized pairs.
(DOCX)

S1 Table. Clinical trial information mapped to the OpenBioLink and custom KGs. For each possible pairing of a drug-disease database (i.e., column 2) with entities that could be mapped to either the OpenBioLink or custom KG, we report the proportion of drug-disease pairs contained in ClinicalTrials.gov (i.e., column 3). We consider these clinically investigated drug-disease pairs as positive labels for the validation of our approach. The number of unique drugs (PubChem compound identifiers) and diseases (MONDO identifiers) are reported in columns 4 and 5, respectively, while the total number of possible combinations of these unique drugs and diseases are presented in column 6.
(XLSX)

S2 Table. Percentage of drug-disease pairs at different steps for every dataset combination and KG.
(XLSX)

S3 Table. Evaluation of the 11 benchmark methods using precision as a metric. None of the benchmarked methods achieve a precision greater than 50%. Furthermore, we would like to note that most of the drug-disease pairs prioritized by each of these methods are the same since they are based on network proximity. Thus, if a drug and a disease share a large number of nodes, they will consistently be prioritized by most of these methods.
(XLSX)

S4 Table. Statistics on the genes measured in the four transcriptomic datasets used.
(XLSX)

S5 Table. Investigated datasets.
(XLSX)

S6 Table. Overlap of the transcriptomic dataset and the KGs. The total number of drugs and diseases in the datasets which can be mapped to the KG as well as the proportions of them that are present in the KG are given in columns 3 and 4, respectively. Similarly, column 5 displays the total number of mapped proteins as well as the proportions of proteins that are present in the KG. Details about each individual drug/disease are available at https://github.com/enveda/RPath/blob/master/data/drug_disease_overview.tsv.
(XLSX)

S7 Table. Properties of the OpenBioLink and custom KGs.
(XLSX)

S8 Table. Drug repurposing approaches exploiting anticorrelation of transcriptomic signatures.
(XLSX)

S9 Table. Permutation experiments across the four dataset combinations using permuted KGs and gene expression datasets.
(XLSX)

S1 Fig. Pseudocode of the RPath algorithm designed for target prioritization.
(TIF)

Acknowledgments

We are very grateful for the publicly available content provided by several databases which we have used in this work (i.e., the KGs and transcriptomic datasets).

Author Contributions

Conceptualization: Daniel Domingo-Fernández.

Data curation: Daniel Domingo-Fernández, Sarah Mubeen, Daniel Rivas-Barragan.

Formal analysis: Daniel Domingo-Fernández, Yojana Gadiya, Abhishek Patel, Sarah Mubeen.

Funding acquisition: David Healey, Joe Rokicki, Viswa Colluru.

Investigation: Daniel Domingo-Fernández, Yojana Gadiya, Abhishek Patel, Biswapriya B. Misra.

Methodology: Daniel Domingo-Fernández, Yojana Gadiya, Joe Rokicki.

Project administration: Daniel Domingo-Fernández, Chris W. Diana, Biswapriya B. Misra, David Healey, Joe Rokicki, Viswa Colluru.

Resources: Daniel Domingo-Fernández, Chris W. Diana, Viswa Colluru.

Software: Daniel Domingo-Fernández, Yojana Gadiya, Daniel Rivas-Barragan.

Supervision: Daniel Domingo-Fernández, Chris W. Diana, David Healey, Joe Rokicki, Viswa Colluru.

Validation: Daniel Domingo-Fernández, Yojana Gadiya, Biswapriya B. Misra.

Visualization: Daniel Domingo-Fernández, Yojana Gadiya, Chris W. Diana.

Writing – original draft: Daniel Domingo-Fernández, Sarah Mubeen, Biswapriya B. Misra, David Healey.

Writing – review & editing: Daniel Domingo-Fernández, Sarah Mubeen, Daniel Rivas-Barragan, Biswapriya B. Misra, David Healey, Viswa Colluru.

References

1. Fotis C, Antoranz A, Hatziavramidis D, Sakellaropoulos T, Alexopoulos LG. Network-based technologies for early drug discovery. *Drug discovery today*. 2018 Mar 1; 23(3):626–35. https://doi.org/10.1016/j.drudis.2017.12.001 PMID: 29294361
2. Bharadhwaj VS, Ali M, Birkenbihl C, Mubeen S, Lehmann J, Hofmann-Apitius M, et al. Domingo-Fernández D. CLEP: a hybrid data-and knowledge-driven framework for generating patient representations. *Bioinformatics*. 2021 Oct 1; 37(19):3311–8. https://doi.org/10.1093/bioinformatics/btab340
3. Zitnik M, Agrawal M, Leskovec J. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics*. 2018 Jul 1; 34(13):i457–66. https://doi.org/10.1093/bioinformatics/bty294 PMID: 29949996
4. Sang S, Yang Z, Liu X, Wang L, Lin H, Wang J, et al. GrEDeL: A knowledge graph embedding based method for drug discovery from biomedical literatures. *IEEE Access*. 2018 Dec 12; 7:8404–15. https://doi.org/10.1109/ACCESS.2018.2886311
5. Nelson W, Zitnik M, Wang B, Leskovec J, Goldenberg A, Sharan R. To embed or not: network embedding as a paradigm in computational biology. *Frontiers in genetics*. 2019 May 1; 10:381. https://doi.org/10.3389/fgene.2019.00381 PMID: 31118945
6. Bonner S, Barrett IP, Ye C, Swiers R, Engkvist O, Bender A, et al. A review of biomedical datasets relating to drug discovery: A knowledge graph perspective. *arXiv preprint* arXiv:2102.10062. 2021 Feb 19.
7. MacLean F. Knowledge graphs and their applications in drug discovery. *Expert opinion on drug discovery*. 2021 Sep 2; 16(9):1057–69. https://doi.org/10.1080/17460441.2021.1910673 PMID: 33843398
8. Schaefer MH, Serrano L, Andrade-Navarro MA. Correcting for the study bias associated with protein–protein interaction measurements reveals differences between protein degree distributions from different cancer types. *Frontiers in genetics*. 2015 Aug 4; 6:260. https://doi.org/10.3389/fgene.2015.00260 PMID: 26300911
9. Vella D, Marini S, Vitali F, Di Silvestre D, Mauri G, Bellazzi R. MTGO: PPI network analysis via topological and functional module identification. *Scientific reports*. 2018 Apr 3; 8(1):1–3. https://doi.org/10.1038/s41598-017-17765-5 PMID: 29311619
10. Liu A, Trairatphisan P, Gjerga E, Didangelos A, Barratt J, Saez-Rodriguez J. From expression footprints to causal pathways: contextualizing large signaling networks with CARNIVAL. *NPJ systems biology and applications*. 2019 Nov 11; 5(1):1–0. https://doi.org/10.1038/s41540-019-0118-z PMID: 31728204
11. Belyaeva A, Cammarata L, Radhakrishnan A, Squires C, Yang KD, Shivashankar GV, Uhler C. Causal network models of SARS-CoV-2 expression and aging to identify candidates for drug repurposing. *Nature communications*. 2021 Feb 15; 12(1):1–3. https://doi.org/10.1038/s41467-020-20314-w PMID: 33397941
12. Winkler S, Winkler I, Figaschewski M, Tiede T, Nordheim A, Kohlbacher O. De novo identification of maximally deregulated subnetworks based on multi-omics data with DeRegNet. *bioRxiv*. 2021 Jan 1. https://doi.org/10.1101/2021.05.11.443638

13. Iorio F, Bosotti R, Scacheri E, Belcastro V, Mithbaokar P, Ferriero R, et al. Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proceedings of the National Academy of Sciences*. 2010 Aug 17; 107(33):14621–6. https://doi.org/10.1073/pnas.1000138107 PMID: 20679242

14. Sirota M, Dudley JT, Kim J, Chiang AP, Morgan AA, Sweet-Cordero A, et al. Discovery and preclinical validation of drug indications using compendia of public gene expression data. *Science translational medicine*. 2011 Aug 17; 3(96):96ra77–. https://doi.org/10.1126/scitranslmed.3001318 PMID: 21849665

15. Peyvandipour A, Saberian N, Shafi A, Donato M, Draghici S. A novel computational approach for drug repurposing using systems biology. *Bioinformatics*. 2018 Aug 15; 34(16):2817–25. https://doi.org/10.1093/bioinformatics/bty133 PMID: 29534151

16. Emon MA, Domingo-Fernández D, Hoyt CT, Hofmann-Apitius M. PS4DR: a multimodal workflow for identification and prioritization of drugs based on pathway signatures. *BMC bioinformatics*. 2020 Dec; 21(1):1–21. https://doi.org/10.1186/s12859-019-3325-0 PMID: 31898485

17. Samart K, Tuyishime P, Krishnan A, Ravi J. Reconciling multiple connectivity scores for drug repurposing. *Briefings in Bioinformatics*. 2021 Nov; 22(6):bbab161. https://doi.org/10.1093/bib/bbab161 PMID: 34013329

18. Hill SM, Nesser NK, Johnson-Camacho K, Jeffress M, Johnson A, Boniface C, et al. Context specificity in causal signaling networks revealed by phosphoprotein profiling. *Cell systems*. 2017 Jan 25; 4(1):73–83. https://doi.org/10.1016/j.cels.2016.11.013 PMID: 28017544

19. Babur Ö, Luna A, Korkut A, Durupinar F, Siper MC, Dogrusoz U, et al. Causal interactions from proteomic profiles: Molecular data meet pathway knowledge. *Patterns*. 2021 Jun 11; 2(6):100257. https://doi.org/10.1016/j.patter.2021.100257 PMID: 34179843

20. Catlett NL, Bargnesi AJ, Ungerer S, Seagaran T, Ladd W, Elliston KO, et al. Reverse causal reasoning: applying qualitative causal knowledge to the interpretation of high-throughput data. *BMC bioinformatics*. 2013 Dec; 14(1):1–4. https://doi.org/10.1186/1471-2105-14-340

21. Martin F, Thomson TM, Sewer A, Drubin DA, Mathis C, Weisensee D, et al. Assessment of network perturbation amplitudes by applying high-throughput data to causal biological networks. *BMC systems biology*. 2012 Dec; 6(1):1–8. https://doi.org/10.1186/1752-0509-6-54 PMID: 22651900

22. Martin F, Sewer A, Talikka M, Xiang Y, Hoeng J, Peitsch MC. Quantification of biological network perturbations for mechanistic insight and diagnostics using two-layer causal models. *BMC bioinformatics*. 2014 Dec; 15(1):1–24. https://doi.org/10.1186/1471-2105-15-238 PMID: 25015298

23. Rivas-Barragan D, Mubeen S, Guim Bernat F, Hofmann-Apitius M, Domingo-Fernández D. Drug2ways: Reasoning over causal paths in biological networks for drug discovery. *PLoS computational biology*. 2020 Dec 2; 16(12):e1008464. https://doi.org/10.1371/journal.pcbi.1008464 PMID: 33264280

24. Chindelevitch L, Ziemek D, Enayetallah A, Randhawa R, Sidders B, Brockel C, et al. Causal reasoning on biological networks: interpreting transcriptional changes. *Bioinformatics*. 2012 Apr 15; 28(8):1114–21. https://doi.org/10.1093/bioinformatics/bts090 PMID: 22355083

25. Krämer A, Green J, Pollard J Jr, Tugendreich S. Causal analysis approaches in ingenuity pathway analysis. *Bioinformatics*. 2014 Feb 15; 30(4):523–30. https://doi.org/10.1093/bioinformatics/btt703 PMID: 24336805

26. Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*. 2017 Nov 30; 171(6):1437–52. https://doi.org/10.1016/j.cell.2017.10.049 PMID: 29195078

27. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets—update. *Nucleic acids research*. 2012 Nov 26; 41(D1):D991–5. https://doi.org/10.1093/nar/gks1193

28. Ochoa D, Hercules A, Carmona M, Suveges D, Gonzalez-Uriarte A, Malangone C, et al. Open Targets Platform: supporting systematic drug–target identification and prioritisation. *Nucleic Acids Research*. 2021 Jan 8; 49(D1):D1302–10. https://doi.org/10.1093/nar/gkaa1027 PMID: 33196847

29. Wang Z, Monteiro CD, Jagodnik KM, Fernandez NF, Gundersen GW, Rouillard AD, et al. Extraction and analysis of signatures from the Gene Expression Omnibus by the crowd. *Nature communications*. 2016 Sep 26; 7(1):1–1. https://doi.org/10.1038/ncomms12846

30. Abbas K, Abbasi A, Dong S, Niu L, Yu L, Chen B, et al. Application of network link prediction in drug discovery. *BMC bioinformatics*. 2021 Dec; 22(1):1–21. https://doi.org/10.1186/s12859-020-03881-z PMID: 33388027

31. Coşkun M, Koyutürk M. Node similarity-based graph convolution for link prediction in biological networks. *Bioinformatics*. 2021 Dec 1; 37(23):4501–8. https://doi.org/10.1093/bioinformatics/btab464 PMID: 34152393

32. Smith CC, Wang Q, Chin CS, Salerno S, Damon LE, Levis MJ, et al. Validation of ITD mutations in FLT3 as a therapeutic target in human acute myeloid leukaemia. *Nature*. 2012 May; 485(7397):260–3. https://doi.org/10.1038/nature11016 PMID: 22504184

33. Pabst T, Mueller BU. Complexity of CEBPA dysregulation in human acute myeloid leukemia. *Clinical Cancer Research*. 2009 Sep 1; 15(17):5303–7. https://doi.org/10.1158/1078-0432.CCR-08-2941 PMID: 19706798

34. Wang X, Jin P, Zhang Y, Wang K. CircSPI1 acts as an oncogene in acute myeloid leukemia through antagonizing SPI1 and interacting with microRNAs. *Cell death & disease*. 2021 Mar 19; 12(4):1–3. https://doi.org/10.1038/s41419-021-03566-2 PMID: 33741901

35. Heo SK, Noh EK, Kim JY, Jeong YK, Jo JC, Choi Y, et al. Targeting c-KIT (CD117) by dasatinib and radotinib promotes acute myeloid leukemia cell death. *Scientific reports*. 2017 Nov 10; 7(1):1–2. https://doi.org/10.1038/s41598-016-0028-x PMID: 28127051

36. Mascaux C, Iannino N, Martin B, Paesmans M, Berghmans T, Dusart M, et al. The role of RAS oncogene in survival of patients with lung cancer: a systematic review of the literature with meta-analysis. *British journal of cancer*. 2005 Jan; 92(1):131–9. https://doi.org/10.1038/sj.bjc.6602258 PMID: 15597105

37. Lv S, Ji L, Chen B, Liu S, Lei C, Liu X, et al. Histone methyltransferase KMT2D sustains prostate carcinogenesis and metastasis via epigenetically activating LIFR and KLF4. *Oncogene*. 2018 Mar; 37 (10):1354–68. https://doi.org/10.1038/s41388-017-0026-x PMID: 29269867

38. Nakaoka HJ, Hara T, Yoshino S, Kanamori A, Matsui Y, Shimamura T, et al. NECAB3 promotes activation of hypoxia-inducible factor-1 during normoxia and enhances tumourigenicity of cancer cells. *Scientific reports*. 2016 Mar 7; 6(1):1–3. https://doi.org/10.1038/s41598-016-0001-8 PMID: 28442746

39. Gerstein AV, Almeida TA, Zhao G, Chess E, Shih IM, Buhler K, et al. APC/CTNNB1 (β-catenin) pathway alterations in human prostate cancers. *Genes, Chromosomes and Cancer*. 2002 May; 34(1):9–16. https://doi.org/10.1002/gcc.10037 PMID: 11921277

40. Konopatskaya O, Poole AW. Protein kinase Cα: disease regulator and therapeutic target. *Trends in pharmacological sciences*. 2010 Jan 1; 31(1):8–14. https://doi.org/10.1016/j.tips.2009.10.006 PMID: 19969380

41. Takami M, Katayama K, Noguchi K, Sugimoto Y. Protein kinase C alpha-mediated phosphorylation of PIM-1L promotes the survival and proliferation of acute myeloid leukemia cells. *Biochemical and biophysical research communications*. 2018 Sep 10; 503(3):1364–71. https://doi.org/10.1016/j.bbrc.2018.07.049 PMID: 30017192

42. Campbell LM, Maxwell PJ, Waugh DJ. Rationale and means to target pro-inflammatory interleukin-8 (CXCL8) signaling in cancer. *Pharmaceuticals*. 2013 Aug; 6(8):929–59. https://doi.org/10.3390/ph6080929 PMID: 24276377

43. Schinke C, Giricz O, Li W, Shastri A, Gordon S, Barreyro L, et al. IL8-CXCR2 pathway inhibition as a therapeutic strategy against MDS and AML stem cells. *Blood, The Journal of the American Society of Hematology*. 2015 May 14; 125(20):3144–52. https://doi.org/10.1182/blood-2015-01-621631 PMID: 25810490

44. Kuett A, Rieger C, Perathoner D, Herold T, Wagner M, Sironi S, et al. IL-8 as mediator in the microenvironment-leukaemia network in acute myeloid leukaemia. *Scientific reports*. 2015 Dec 17; 5(1):1–1. https://doi.org/10.1038/srep18411 PMID: 26674118

45. Nalla AK, Asuthkar S, Bhoopathi P, Gujrati M, Dinh DH, Rao JS. Suppression of uPAR retards radiation-induced invasion and migration mediated by integrin β1/FAK signaling in medulloblastoma. *PloS one*. 2010 Sep 24; 5(9):e13006. https://doi.org/10.1371/journal.pone.0013006 PMID: 20886051

46. Hong L, Kenney SR, Phillips GK, Simpson D, Schroeder CE, Nöth J, et al. Characterization of a Cdc42 protein inhibitor and its use as a molecular probe. *Journal of Biological Chemistry*. 2013 Mar 22; 288 (12):8531–43. https://doi.org/10.1074/jbc.M112.435941 PMID: 23382385

47. Coker EA, Mitsopoulos C, Workman P, Al-Lazikani B. SiGNet: A signaling network data simulator to enable signaling network inference. *Plos one*. 2017 May 17; 12(5):e0177701. https://doi.org/10.1371/journal.pone.0177701 PMID: 28545060

48. Perez-Riverol Y, Csordas A, Bai J, Bernal-Llinares M, Hewapathirana S, Kundu DJ, et al. The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic acids research*. 2019 Jan 8; 47(D1):D442–50. https://doi.org/10.1093/nar/gky1106 PMID: 30395289

49. Samaras P, Schmidt T, Frejno M, Gessulat S, Reinecke M, Jarzab A, et al. ProteomicsDB: a multi-omics and multi-organism resource for life science research. *Nucleic acids research*. 2020 Jan 8; 48 (D1):D1153–63. https://doi.org/10.1093/nar/gkz974 PMID: 31665479

50. Boué S, Talikka M, Westra JW, Hayes W, Di Fabio A, Park J, et al. Causal biological network database: a comprehensive platform of causal biological network models focused on the pulmonary and vascular systems. *Database*. 2015 Jan 1; 2015. https://doi.org/10.1093/database/bav030 PMID: 25887162

51. Domingo-Fernández D, Kodamullil AT, Iyappan A, Naz M, Emon MA, Raschka T, et al. Multimodal mechanistic signatures for neurodegenerative diseases (NeuroMMSig): a web server for mechanism enrichment. *Bioinformatics*. 2017 Nov 15; 33(22):3679–81. https://doi.org/10.1093/bioinformatics/btx399 PMID: 28651363

52. Bui TT, Lee D, Selvarajoo K. ScatLay: utilizing transcriptome-wide noise for identifying and visualizing differentially expressed genes. *Scientific reports*. 2020 Oct 15; 10(1):1–1. https://doi.org/10.1038/s41598-019-56847-4 PMID: 31913322

53. Breit A, Ott S, Agibetov A, Samwald M. OpenBioLink: a benchmarking framework for large-scale biomedical link prediction. *Bioinformatics*. 2020 Jul 1; 36(13):4097–8. https://doi.org/10.1093/bioinformatics/btaa274 PMID: 32339214

54. Malas TB, Vlietstra WJ, Kudrin R, Starikov S, Charrout M, Roos M, et al. Drug prioritization using the semantic properties of a knowledge graph. *Scientific reports*. 2019 Apr 18; 9(1):1–0. https://doi.org/10.1038/s41598-018-37186-2 PMID: 30626917

55. Gysi DM, Do Valle Í, Zitnik M, Ameli A, Gan X, Varol O, et al. Network medicine framework for identifying drug-repurposing opportunities for COVID-19. *Proceedings of the National Academy of Sciences*. 2021 May 11; 118(19). https://doi.org/10.1073/pnas.2025581118

56. Berrar D, Flach P. Caveats and pitfalls of ROC analysis in clinical microarray research (and how to avoid them). *Briefings in bioinformatics*. 2012 Jan 1; 13(1):83–97. https://doi.org/10.1093/bib/bbr008 PMID: 21422066

57. Hanhijärvi S, Garriga GC, Puolamäki K. Randomization techniques for graphs. InProceedings of the 2009 SIAM International Conference on Data Mining 2009 Apr 30 (pp. 780–791). *Society for Industrial and Applied Mathematics*. https://doi.org/10.1137/1.9781611972795.67

A.7 Predicting antimicrobial class specificity of small molecules using machine learning

Reprinted with permission from "Gadiya, Y., Genilloud, O., Bilitewski, U., Brönstrup, M., von Berlin, L., Attwood, M., Gribbon, P., and Zaliani, A. (2024). Predicting antimicrobial class specificity of small molecules using machine learning. *Journal of Chemical Information and Modeling*". 65 (5), pp.2416-2431.

JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

Open Access

pubs.acs.org/jcim

Article

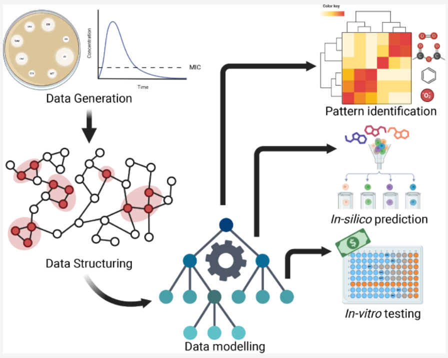
Predicting Antimicrobial Class Specificity of Small Molecules Using Machine Learning

Yojana Gadiya,* Olga Genilloud, Ursula Bilitewski, Mark Brönstrup, Leonie von Berlin, Marie Attwood, Philip Gribbon, and Andrea Zaliani

Cite This: *J. Chem. Inf. Model.* 2025, 65, 2416−2431

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: While the useful armory of antibiotic drugs is continually depleted due to the emergence of drug-resistant pathogens, the development of novel therapeutics has also slowed down. In the era of advanced computational methods, approaches like machine learning (ML) could be one potential solution to help reduce the high costs and complexity of antibiotic drug discovery and attract collaboration across organizations. In our work, we developed a large antimicrobial knowledge graph (AntiMicrobial-KG) as a repository for collecting and visualizing public *in vitro* antibacterial assay. Utilizing this data, we build ML models to efficiently scan compound libraries to identify compounds with the potential to exhibit antimicrobial activity. Our strategy involved training seven classic ML models across six compound fingerprint representations, of which the Random Forest trained on the MHFP6 fingerprint outperformed, demonstrating an accuracy of 75.9% and Cohen's Kappa score of 0.68. Finally, we illustrated the model's applicability for predicting the antimicrobial properties of two small molecule screening libraries. First, the EU-OpenScreen library was tested against a panel of Gram-positive, Gram-negative, and Fungal pathogens. Here, we unveiled that the model was able to correctly predict more than 30% of active compounds for Gram-positive, Gram-negative, and Fungal pathogens. Second, with the Enamine library, a commercially available HTS compound collection with claimed antibacterial properties, we predicted its antimicrobial activity and pathogen class specificity. These results may provide a means for accelerating research in AMR drug discovery efforts by carefully filtering out compounds from commercial libraries with lower chances of being active.

1. INTRODUCTION

Since their discovery, antibiotics have been primarily used to treat bacterial infections due to their ease of administration and potent antibacterial activity.[1] However, decades of liberal antibiotic use have led to a significant loss of effective treatment options. There is growing evidence that antimicrobial resistance (AMR) is an emerging threat to human health worldwide.[2] This has been highlighted in two comprehensive studies: the Antimicrobial Resistance report from 2016[3] and the Global Burden of AMR study in 2019,[4,5] among others.[6] To counter this, annual surveillance studies have generated and collated data, providing regional health institutions with opportunities to adapt and modify local prescribing trends and implement antimicrobial stewardship initiatives. Despite significant local, regional, and global efforts, the AMR burden remains at an all-time high. One major reason for the actual low rate of antibiotic authorization is the prolonged time required to develop drugs through the traditional drug discovery pipeline, coupled with the attrition of compounds that fail to reach the market.[7] Machine learning (ML), which can enable time-effective and efficient decision-making when presented with vast amounts of data, has the potential to improve drug discovery.

There is a pressing need to understand bacterial resistance phenotypes to develop effective AMR drugs. This necessity has driven efforts to manage AMR infections in both clinical and community settings. In clinical environments, omics experiments like whole-genome sequencing for antimicrobial susceptibility testing (WGS-AST) have shown the potential to provide rapid, consistent, and accurate predictions of known resistance phenotypes while offering rich surveillance data.[8−11] Meanwhile, judicious and controlled usage of existing medications in the community has also increased to combat this issue.[12−14] Understanding bacterial resistance mechanisms and characterizing the intrinsic pharmacokinetics and pharmacodynamics (PK/PD) features of drugs is essential in antibiotic drug discovery. Optimizing aspects such as drug metabolic stability, systemic half-life, and bioavailability often provides dosage advantages, enhancing the safety profile and

Received: December 17, 2024
Revised: January 23, 2025
Accepted: January 24, 2025
Published: February 23, 2025

https://doi.org/10.1021/acs.jcim.4c02347

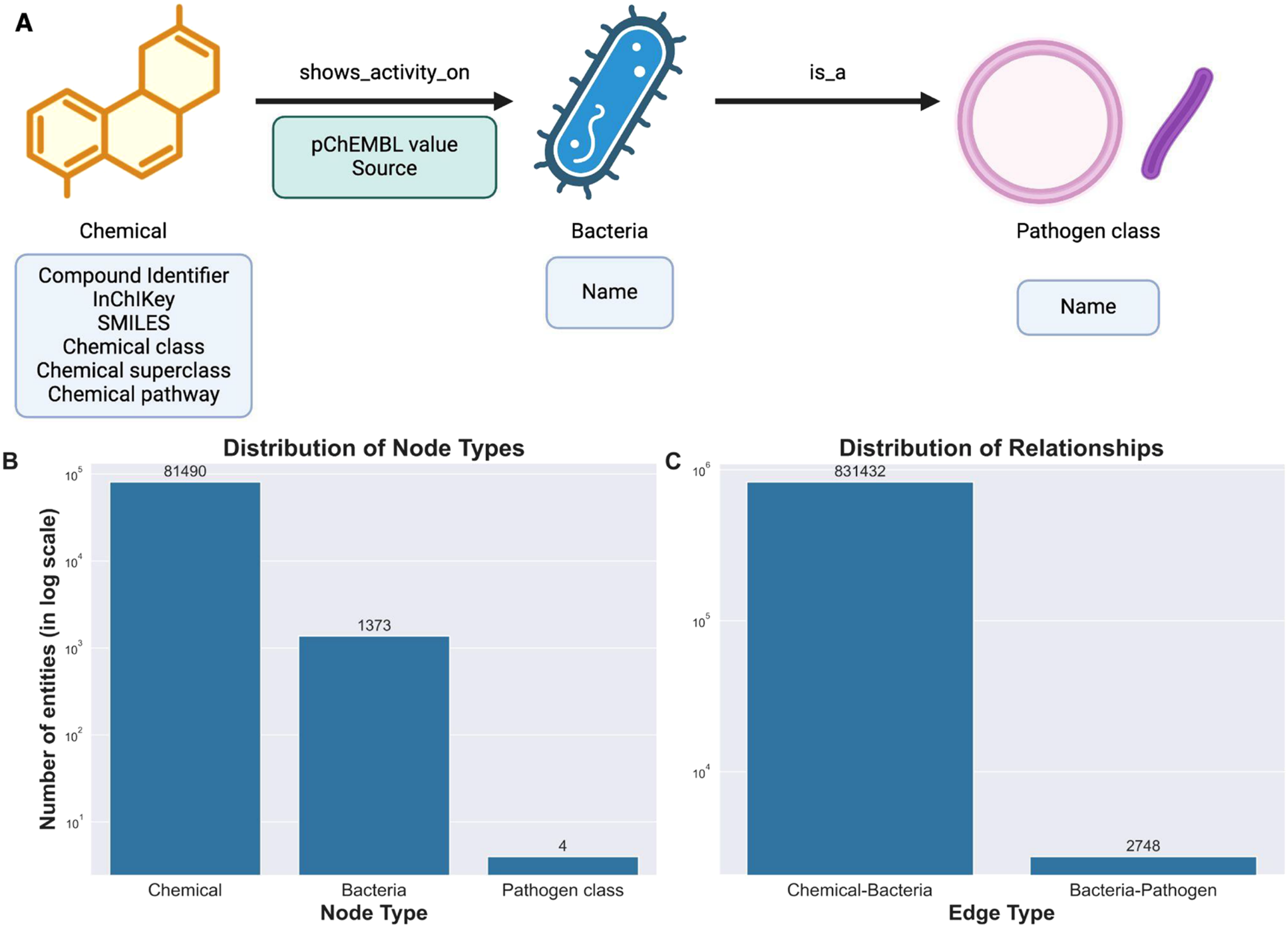


Figure 1. AntiMicrobial-KG schema and overview statistics. A) The schema of the AntiMicrobial-KG connecting the chemicals to bacteria and bacteria to pathogen class. Created with BioRender.com. B) The distribution of the number of nodes found in AntiMicrobial-KG. C) The distribution of edges in the AntiMicrobial-KG.

minimizing the risks of rapid resistance emergence.[15−17] Therefore, such optimization processes require a swift turnover of novel synthetic candidates, increasingly identified using assistance from ML algorithms and models combined with de novo design methods.[18−21]

The application of ML in AMR drug discovery has shown promising results, advancing from preclinical to clinical research stages.[22] Researchers have utilized ML algorithms to discover novel synergistic drug interactions from millions of potential combinations, thereby accelerating the development of combination therapies.[23−26] ML-informed computational approaches such as docking have been used to identify the activity of potential antibacterials against known microbial targets, as demonstrated by Chio et al.[27] and Alves et al.[28] Similarly, predicting the activity of antimicrobial peptides against AlphaFold-predicted structures of microbial targets has been explored by Karnati et al.,[29] among others.[30,31] Susceptibility-related data is another source of data utilized by ML models to aid in selecting appropriate antibiotic therapy regimens for patients in clinical settings.[32,33] These models help optimize treatment decisions by analyzing patterns in microbial resistance and patient-specific factors. Together, these efforts represent a shift in drug discovery, offering new avenues for the development of effective antimicrobial agents to combat the growing threat of AMR. Furthermore, by training mathematical models on empirical data sets, ML algorithms can predict the antibacterial activity of new compounds when presented with previously unseen data.[19,34] The availability of public data sets, advances in computer engineering, and the proliferation of free and open-source ML libraries have deeply impacted this approach. However, many ML-based approaches serve the limitation of being predominantly focused on the phenotypic effects of drugs on target organisms[22,35] rather than detailed molecular descriptions. This has led to underutilization of the chemo-physical characteristics of compounds[36] and insufficient attention to structural features[37] and general pharmacophoric features.[38] Leveraging these aspects can provide a more comprehensive understanding of drug behavior and efficacy, enhancing the development of novel therapeutic agents.

Our work addresses the previously mentioned limitations of existing ML models for predicting the antimicrobial activity of compounds. As part of the IMI AMR Accelerator project, COMBINE (https://amr-accelerator.eu/project/combine/), we gathered data on small molecules and their minimal inhibitory concentration (MIC) values from publicly available resources, creating a comprehensive database, the AntiMicrobial-KG. In this study, our primary goal was to develop ML models for predicting the activity of small molecules in the antimicrobial field, focusing on advanced preclinical drug

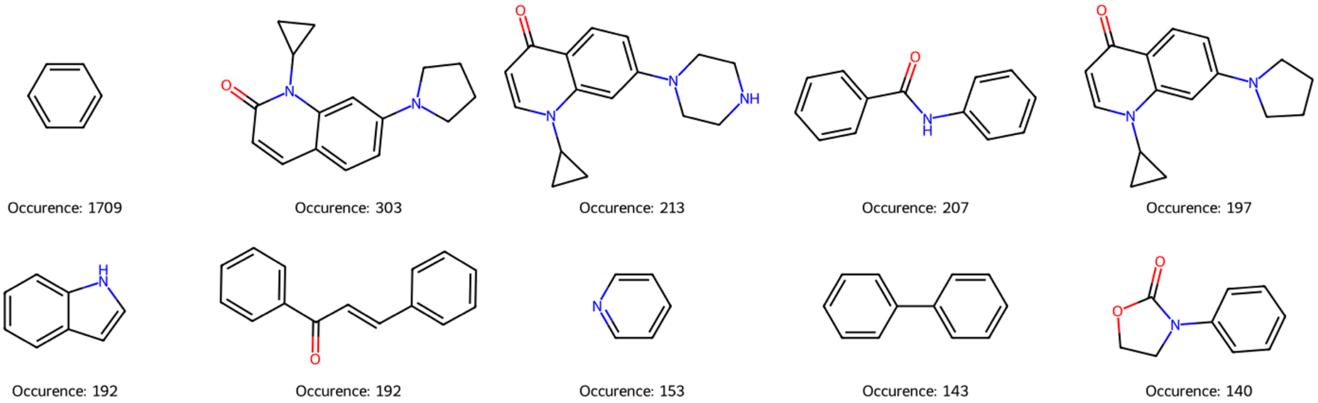


Figure 2. Top 10 ubiquitous generic Murcko scaffolds identified in AntiMicrobial-KG. For each scaffold, the occurrence number in the AntiMicrobial-KG is depicted at the bottom.

discovery. This manuscript highlights the value of the utilization of structural-based molecular features for training ML models. Inherently, the insights from the model for these structural features will assist in developing promising compounds. We also built models to predict the broad-spectrum activity (antibacterial or antifungal) of novel compounds, facilitating efficient screening of compound libraries for experimental validation. Thus, by systematically integrating high-quality data to build transparent ML models, we can effectively demonstrate their applicability in assisting with AMR drug discovery.

2. RESULTS

This section begins by presenting the schema of the AntiMicrobial-KG generated. In Section 2.2, we investigated the performance of the models across different training data sets (SMOTE vs non-SMOTE) generated using six chemical fingerprinting approaches (ECFP8, RDKit, MACCS, MHFP6, ErG, and ChemPhys). These data sets were analyzed with six different models from PyCaret: Naive Bayes (NB), Logistic Regression (LR), LightGBM, Decision Tree (DT), Random Forest (RF), and XGBoost. From these models, we select the best-performing one and, in Section 2.3, evaluate it on external data sets, including commercial libraries, to identify potential antibacterial chemicals. Finally, subsection 2.4 discusses the limitations of our employed strategy.

2.1. AntiMicrobial-KG as a Data Warehouse for MIC Bioassay End Points. We built the AntiMicrobial-KG on the property-graph-based schema that is compliant with Neo4J. This data structure enabled systematic organization and accessibility of data, enhancing its reusability and compliance with FAIR principles. The graph consists of three types of nodes: Chemicals, Bacteria, and Pathogen class. It also includes two types of relationships (or edges), namely "shows_activity_on" (connecting chemicals to bacteria) and "is_a" (connecting bacteria to pathogen class), as illustrated in Figure 1A. At present, the AntiMicrobial-KG incorporates 82,867 nodes, with 81,490 chemicals tested across 1,373 bacteria and 831,432 relationships, constructing a harmonious interaction network. The distribution of the nodes and relationships in the AntiMicrobial-KG are summarized in Figure 1B,C, respectively. To enable efficient chemical substructure searches, metadata for chemical representation in the form of SMILES and InChIKey are stored in the KG. Additionally, chemical classification from NPClassifier into classes, pathways, and superclasses for each chemical can be searched within the database. This comprehensive structure supports sophisticated querying and analysis, facilitating deeper insights into antimicrobial resistance patterns.

Next, we accessed the structural heterogeneity of the chemical space within the AntiMicrobial-KG by reducing the chemicals to their Murcko scaffolds. This investigation revealed a diverse chemical space within the AntiMicrobial-KG involving 24,506 unique Murcko scaffolds across 81,490 chemicals. We further classified the scaffolds based on their occurrence as ubiquitous and scarce within our data set. As exemplified in Figure 2, Michael acceptor substructures were more commonly found in ubiquitous chemical scaffolds than in scarce ones (Figure S1). Michael acceptors are compounds that contain an α,β-unsaturated carbonyl group, which can participate in Michael addition reactions. Molecular scaffolds containing chemically reactive groups such as pyridones, benzochromanones, and azetidinones target nucleophilic centers, often allowing for irreversible binding, a commonly observed property of natural products.[39] Moreover, Michael acceptor compounds are known to exhibit a variety of biological activities, including antimicrobial and antifungal properties, due to their ability to modify critical proteins and enzymes in microorganisms.[40−42]

Following the analysis of chemical diversity, we inspected the chemical space of the AntiMicrobial-KG by examining several aspects: Rule-of-Five (Ro5) compliance, the presence or absence of structural alerts, and the distribution of chemicals into classes, superclasses, and pathways across bacterial strains. We found that 75.2% (i.e., 55,814) of chemicals in the AntiMicrobial-KG violated at least one of the Ro5 guidelines, suggesting that a significant portion of the data set consists of either natural products (NPs) originated chemicals or chemicals with poor oral bioavailability (Table S1). To validate this, we assessed the NP-likeness of the chemicals and observed a substantial number of synthetic chemicals (25,313) compared to natural products (8,872), indicating that both NP origin and the poor bioavailability of synthetic compounds contribute to the Ro5 violations (Figure S2). We also investigated the presence of structural alerts that include specific substructures known to be associated with toxicity or other undesirable properties. The overall ratio of chemicals with structural alerts to those without was approximately 7:3 across all pathogen classes. For chemicals with no structural alerts, the breakdown by pathogen class is as follows: 30.78%

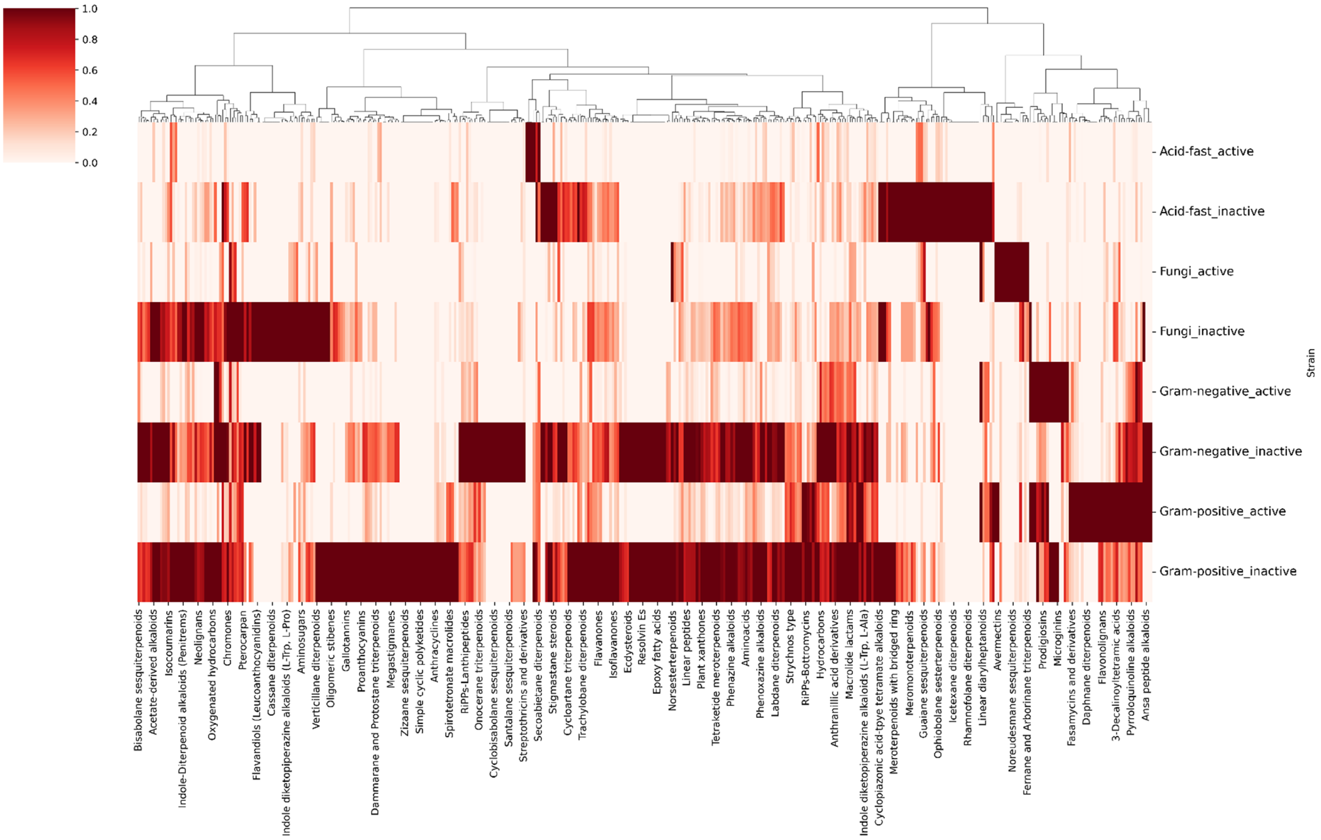


Figure 3. Cluster map of chemical classes across the four pathogen classes, each divided into two subclasses. From the cluster map, certain classes of chemicals are shown to be dominant (shown with dark red patches) across each of the pathogen categories, while others are underrepresented (shown with light red patches). Additionally, chemical class patterns are different between active and inactive chemicals for the same pathogen class.

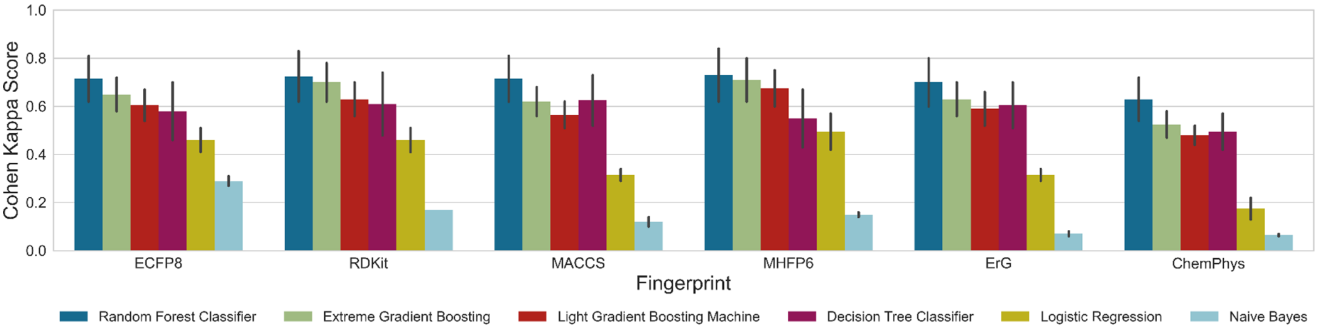


Figure 4. Comparison of the classic ML models across a spectrum of model–fingerprint combinations. The evaluation metric shown in this figure is based on the hold-out or validations set (20% of data), and models are trained on 60% of the data.

chemicals for Gram-positives, 27.93% for Gram-negatives, 31.39% for fungi, and 32.75% for acid-fast bacteria (Table S1). We further explored the characterization of chemicals within the AntiMicrobial-KG, grouping them into 411 chemical classes, 67 superclasses, and 7 biosynthetic pathways. This analysis revealed distinct patterns differentiating active from inactive chemicals across various pathogens, specifically at the chemical class level (Figure 3). For instance, 64 chemical classes, including tetracyclic and daphnane diterpenoids, steroidal alkaloids, and rotenoids, were exclusively found in Gram-positive active chemicals. Similarly, 34 classes, such as aeruginosins, pentacyclic guanidine alkaloids, and fasamycins and derivatives, were specific to Gram-negative active chemicals. For fungal pathogens, 42 chemical classes, including rhizoxins, ergostane steroids, and lipopeptides, were identified as active. Additionally, for acid-fast pathogens, 31 chemical classes, such as artemisinin, furans, polysaccharides, and tropane alkaloids, were uniquely associated with the active compounds. From a biosynthetic perspective, the polyketide pathway was the only pathway distinctly associated with Gram-positive active compounds compared to the inactive ones.

2.2. MHFP6 with Random Forest Outperforms Other Models.

We began by developing a workflow to train and test a cohort of six (i.e., Naive Bayes, Logistic regression, Light Gradient Boosting Machine, Decision Tree, Random Forest and eXtreme Gradient Boosting) classical and explainable machine learning (ML) models. The implementation architecture encompasses data harmonization, model comparison, model selection, and optimization (see Section 4.4 for more details). Our model comparison strategy allowed us to perform two independent assessments: first, to identify which model-fingerprint combination provided the best prediction results based on the Cohen's Kappa score, and second, to evaluate the impact of implementing a SMOTE-based data set on enhancing model prediction capabilities. Upon analyzing the model-fingerprint pair combinations, we found that, regardless of the fingerprinting approach used, the Random Forest and eXtreme Gradient Boosting (XGBoost) models consistently outperformed other models (Figure 4). In contrast, Naive Bayes and Logistic Regression models were the least effective for multiclass classification. Notably, there was a similar pattern in model outcomes (Cohen's Kappa scores) between models using the RDKit fingerprint and those using MHFP6. Finally, models trained with ChemPhys and ErG fingerprints were the lowest performers, indicating that representing chemicals with fewer than 1,000 features (as these fingerprints do) is insufficient for capturing essential chemical characteristics compared to other fingerprints that utilize over 1,000 features. This insight underscores the importance of comprehensive feature representation in improving the predictive performance of ML models. The poor performance of the Naive Bayes models (Cohen's Kappa score <0.3) raises concerns about the implementation of the model in PyCaret and the predefined parameters used for training. The simplicity of the Naive Bayes model and its underlying assumptions about data distribution (e.g., Gaussian distribution for continuous features) and feature independence, which is rarely true in real-world data, further exacerbates its ineffectiveness in complex multiclass classification tasks. When comparing the SMOTE-trained models with classically trained models, we observed that the SMOTE-trained models exhibited an increase in performance, with an average improvement of 10% in Cohen's Kappa score (Table S2). To further evaluate the robustness of model predictions with SMOTE-based training, we trained the top two models (Random Forest and XGBoost) across all fingerprints using both the classic and SMOTE data sets (Figure S3). This comparison consistently demonstrated that SMOTE training improved performance across various fingerprints, mirroring the initial observation.

We identified Random Forest and XGboost as the top-performing models. These models were then subjected to hyperparameter optimization using SMOTE-trained data to enhance their predictive power. The specific hyperparameters optimized for training these two models are detailed in Table S3. The optimized models were then tested on the remaining 20% of test data. Among the two, Random Forest outperformed XGBoost, showing approximately 5% improvement across all metrics (Table S4). Table 1 highlights the performance of various fingerprints, with MHFP6 emerging as the best performer. This superior performance can be attributed to its unique chemical representation, which encodes the chemical using more than 2,000-bit vectors. This extensive encoding likely captures more detailed structural information, contributing to its enhanced predictive accuracy. In both models, the ChemPhys fingerprint showed the lowest performance, with 68.6% accuracy in random forests and 63.4% in XGBoost. This could be attributed to the lower number of features (i.e., 29 in ChemPhys vs 2048 in MHFP6) that correspond to the chemical. This small number might not be entirely sufficient to distinguish the activity of chemicals in final label classes (i.e., Gram-positive, Gram-negative, Acti-fast, Fungi, and Inactive).

Table 1. Random Forest Model Performance on the Test Data[a]

Descriptors	Accuracy	Cohen's Kappa	Macro precision	Macro recall	Macro F1
MHFP6	**0.759**	**0.678**	**0.775**	**0.756**	**0.764**
ECFP8	0.752	0.670	0.765	0.753	0.758
RDKIT	0.743	0.659	0.752	0.746	0.749
MACCS	0.732	0.645	0.728	0.739	0.733
ErG	0.725	0.636	0.727	0.725	0.726
ChemPhys	0.686	0.585	0.671	0.680	0.675

[a]The table shows the average metric reported across all the five classes (Gram-positive, Gram-negative, Acid-fast, Fungi, and Inactive).

All random forest trained fingerprint pairs demonstrated an average AUC-ROC score of 0.82. The MHFP6 fingerprint achieved the highest AUC-ROC score of 0.845, closely followed by the ECFP8-trained model with an AUC-ROC score of 0.843 (Figure 5A). In contrast, the ChemPhys-trained model showed a lower AUC-ROC score of 0.79. For the best model pair, we further examined its ability to correctly classify chemicals into their respective pathogen classes using a confusion matrix (Figure 5B). The true positive rate of Acid-fast pathogens was the highest at 0.81, while that of Gram-negative pathogens was 0.65. We also observed that overall, the model demonstrated a high precision or positive predictive value (PPV) and high negative predictive value (NPV) for all labels, indicating the model is robust and can correctly identify true positives and true negatives in the data (Table S5). Interestingly, 16% of Gram-negative active chemicals (i.e., 370 of 2,271 chemicals) were incorrectly classified as actives for the Gram-positive pathogen, and 5% of the Gram-positive active chemicals (i.e., 246 of 6,952 chemicals) were incorrectly classified as actives for Gram-negative pathogen. This misclassification can be attributed to two limitations of our current approach. According to our model, a chemical is favorable to be active only against a single pathogen class, which is disparate to real-world scenarios wherein a "broad spectrum" activity (i.e., activity across multiple pathogens) of chemicals is noticed. Second, the misclassification is due to the inability of the model to distinguish the MHFP6 fingerprint landscape of the two or more groups. One potential way to mitigate this issue is to increase the data set used for training data, enabling the model to understand the MHFP6 chemical manifold space for better distinction.

Lastly, we inspected the feature importance of the model. Unfortunately, for most chemical fingerprints, the bit vectors are not easily translatable into chemical features that could directly inform future drug and lead optimizations (Figure S4). To address this, we leveraged the ChemPhys-trained model to correlate the target class with specific chemical properties, providing clues for improving antibacterial and antifungal drugs. Feature importance analysis from this model can be fundamental for interpreting the "trend rules" the model relies on to classify chemicals. These known "rules" can now be validated on a bigger chemical space than before and be used to generate novel hypotheses on important chemical properties. It is key to note that despite the ChemPhys-trained models demonstrating the lowest performance metrics overall, they still achieved an accuracy of 68.6%. Local label-specific

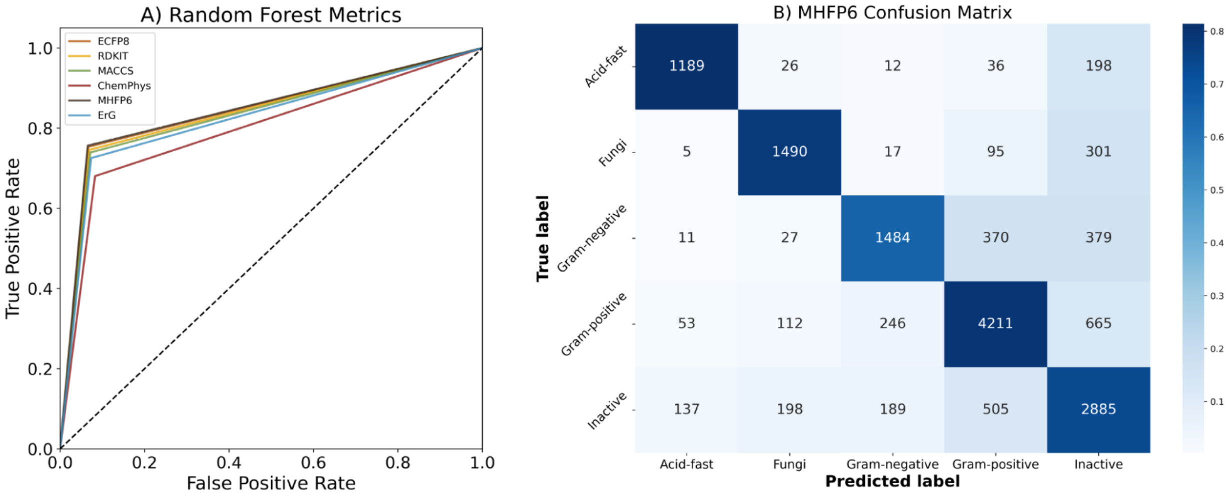


Figure 5. A) AUC-ROC sources of individual fingerprints trained with random forest. B) The confusion matrix of the best model is the MHFP6-trained random forest. The heatmap is color-coded based on the true positive rate (TPR) for each class, with annotations indicating the number of correctly classified chemicals.

feature importance analysis of this model revealed that chemical properties patterns allowed for clustering chemicals into three major buckets: Antibacterials (Gram-positive and Gram-negative pathogens), Antituberculosis with acid-fast pathogens, and Antifungals with close patterns to those of inactive chemicals. Individual chemical properties such as hydrogen bond donors (HBD), the number of rings, LogP, and fraction SP3 played crucial roles in class predictability (Figure 6). The heatmap allowed the identification of known and novel trend rules for antimicrobial activity. First, SLogP, a critical parameter for solubility and cell permeability and a known factor influencing AMR activity,[43] was identified by the model as a positive distinguishing feature between antibacterial and antifungal chemicals. Second, the presence and number of amide bonds, another well-established criterion for antibacterial activity particularly for Gram-negative strains,[44,45] was also identified and utilized by our model as a key distinguishing factor. Third, the hydrogen bond donor (HBD) count (exact or Lipinski filtered) is more sensitive for Gram-type pathogens than for fungi or acid-fast. To the best of our knowledge, this specific aspect has not been highlighted in any previous antimicrobial activity prediction models. Another example of novel trends we identified was the role of aliphatic rings (carbocyclic or heterocyclic). The model used the number of these rings to differentiate Gram-positive active chemicals from Gram-negative active chemicals. Additionally, the number of saturated rings helped identify antifungal compounds, while the presence of saturated heterocycles was linked to antituberculosis activity. Finally, a number of different descriptors dealing with structural complexity, polar surface, and atomic nature have been chosen to dissect inactive compounds.

2.3. Testing the Model with External Compound Libraries.

Finally, we tested the best-performing model, Random Forest trained with MHFP6 fingerprint, on commercial compound libraries to evaluate their *in-silico* antimicrobial activity. This testing phase was crucial for assessing the model's practical relevance and potential effectiveness in identifying novel or repurposed compounds useful for combating AMR. Additionally, by applying the model to a diverse range of commercially available compounds, we aimed to determine ML model robustness and reliability in real-world scenarios.

The complete EU-OS library, comprising the ECBL and Bioactive sets, was screened against seven microbial pathogens: *Candida auris* DSM21092, *Staphylococcus aureus* ATCC 29213, *Pseudomonas aeruginosa* group, *Candida albicans* ATCC 64124, *Enterococcus faecalis* ATCC 29212, *Aspergillus fumigatus* ATCC 46645, and *Escherichia coli* ATCC 25922. To classify the compounds based on their activity, an arbitrary threshold of ≥50% inhibition (at 50 μM) was applied to distinguish active from inactive compounds. Using this criterion, over 95,000 compounds from the ECBL set were labeled as inactive, while approximately 1,000 compounds were classified as active (Figure 7). Specifically, 983 compounds were found to be active against *Staphylococcus aureus*, 1 compound against *Pseudomonas aeruginosa*, 946 compounds against *Candida auris*, 34 compounds against *Enterococcus faecalis*, 198 compounds against *Aspergillus fumigatus*, 131 compounds against *Candida albicans*, and 26 compounds against *Escherichia coli*. From the pathogen class perspective, 99.7% of the compounds were inactive, and 0.3% of the compounds were active for Fungi. For Gram-negative pathogens, 99.9% of the compounds were inactive, and 0.01% of the compounds were active, while for Gram-positive pathogens, 98.9% of the compounds were inactive, and 1.1% of the compounds were active.

Analogously, for the Bioactive set, approximately 4,700 compounds were classified as inactive, while the remaining 300 compounds were active (Figure 8). Among these, 313 compounds were found to be active against *Staphylococcus aureus*, 13 compounds against *Pseudomonas aeruginosa*, 275 compounds against *Candida auris*, 123 compounds against *Enterococcus faecalis*, 145 compounds against *Aspergillus fumigatus*, 120 compounds against *Candida albicans*, and 52 compounds against *Escherichia coli*. From the pathogen class perspective, 95% of the compounds were inactive, and 5% of the compounds were active for Fungi. For Gram-negative pathogens, 98.9% of the compounds were inactive, and 1.1% of the compounds were active, while for Gram-positive pathogens, 94% of the compounds were inactive, and 6% of the compounds were active. Compared to the ECBL set, the

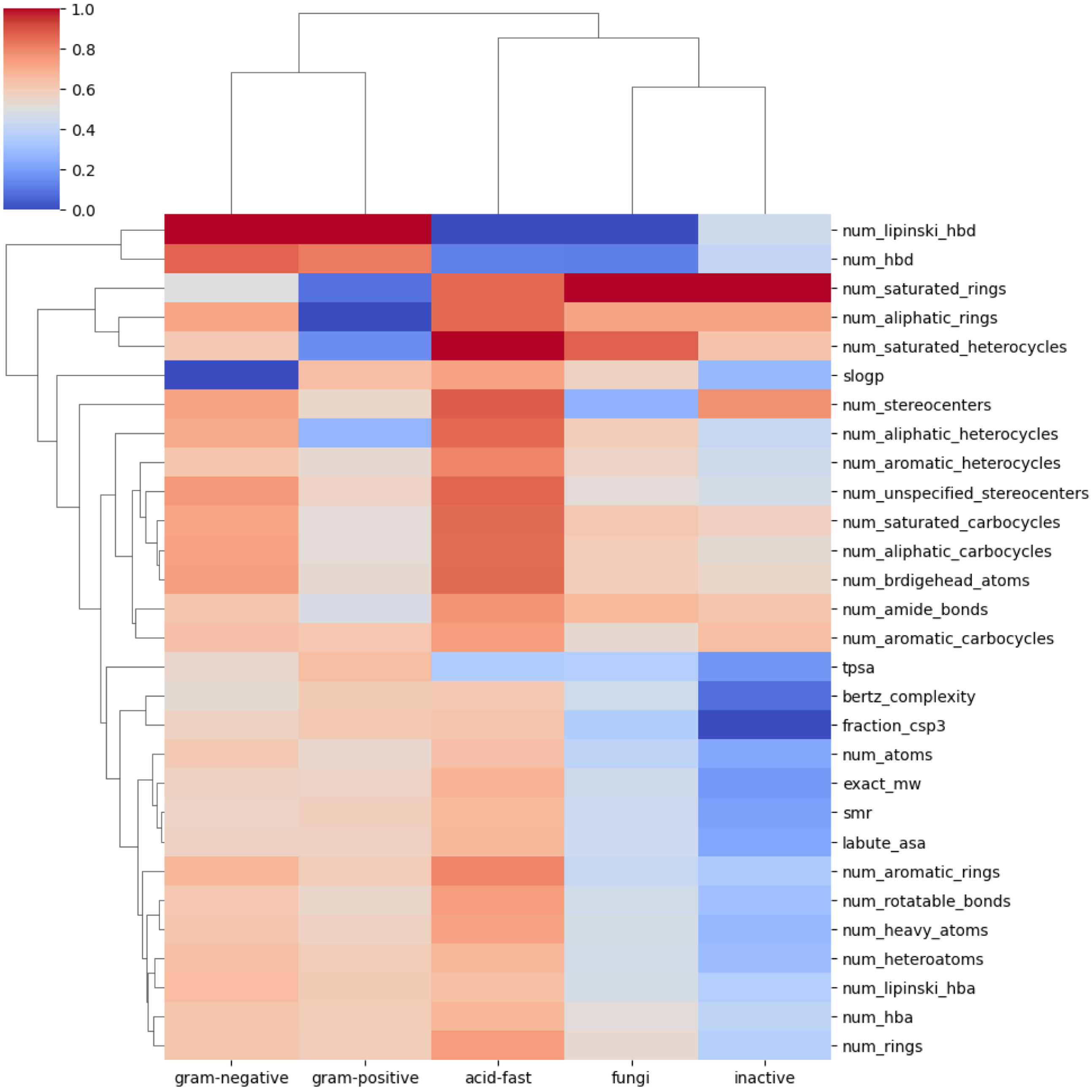


Figure 6. Heatmap of the ChemPhys importance across the five label classes. The physicochemical descriptors are ranked in a top-down manner based on their importance in describing the target or activity class of the data set. The rows indicate the ChemPhys properties of the chemical, the columns indicate the activity class or label, and the color intensity indicates the importance of the normalized feature for each ChemPhys—activity pair.

Bioactive set exhibited a higher experimental hit rate. This could be attributed to several factors, such as the novelty of the ECBL compounds, which may not yet be optimized for cellular permeability. Additionally, the Bioactive set is inherently biased toward "activity," even if on different protein targets, which could explain its higher hit rate. Another potential factor is the greater structural diversity present in the Bioactive set compared to the ECBL (Figure S5). The ECBL, by design, aimed to sample a mini-family of 6–8 compounds around each collection component to ensure a "mini SAR" (structure—activity relationship) in case a hit was identified.

Model predictions for the ECBL set indicated that 90% of the compounds were inactive, with 3.89% active against Gram-positive bacteria, 2.7% against acid-fast bacteria, 1.53% against fungi, and 0.99% against Gram-negative bacteria. In contrast, for the Bioactive set, 81.8% of the compounds were inactive, while 11% were active against Gram-positive bacteria, 3.13% against acid-fast bacteria, 1.99% against fungi, and 2.13% against Gram-negative bacteria. Next, we evaluated the hit rate of the model predictions in comparison to the experimental results, as summarized in Table 2. This analysis focused on examining the number of active compounds identified by both the model and the experiments. In most cases, the model's hit rate exceeded that of the experimental results. Furthermore, a significant difference was observed when comparing the two libraries. For the ECBL library, the difference between the

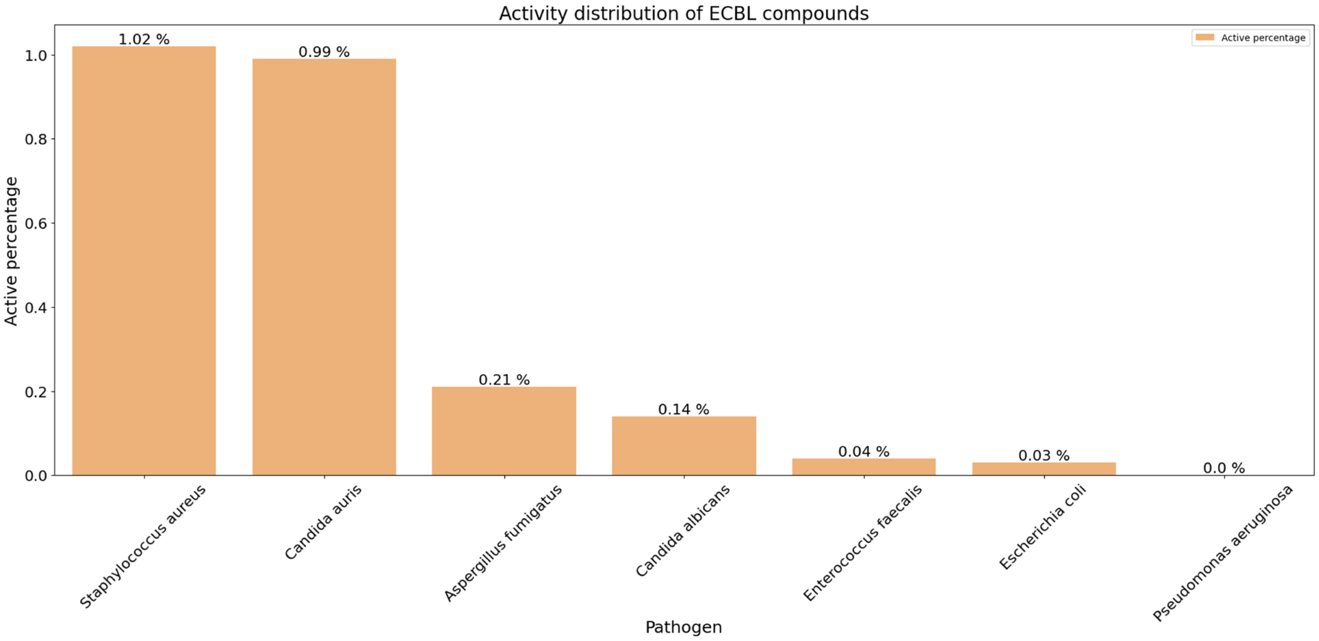


Figure 7. The percentage of experimental active compounds in the ECBL Library for each bacterial strain. The threshold was set to 50% inhibition with compound activity value ≥50% classified as actives, and compounds with activity value <50% were classified as inactive.

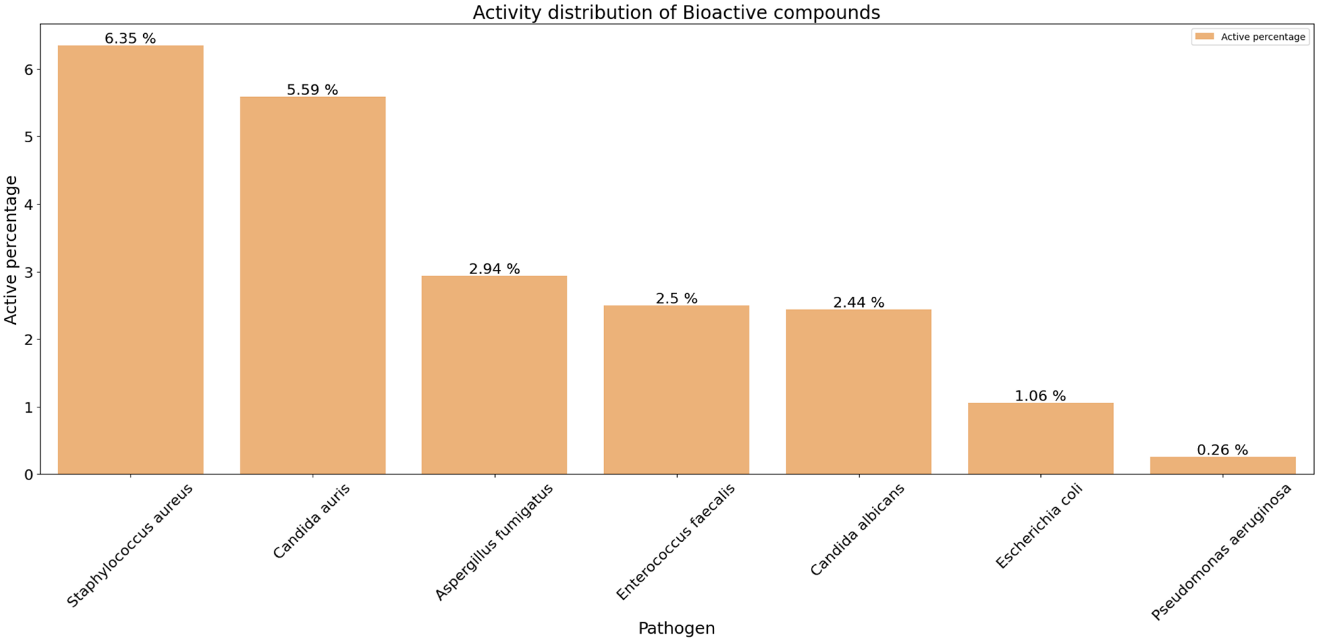


Figure 8. The percentage of experimental active compounds in the Bioactive Library across each bacterial strain. The threshold was set to 50% inhibition with compound activity value ≥50% classified as actives, and compounds with activity value <50% were classified as inactive.

Table 2. Comparison of the HitRate Associated with the Two Library Sets: ECBL and Bioactive[a]

Pathogen class	Experimental strain	ECBL model HitRate (experimental HitRate)	Bioactive model HitRate (experimental HitRate)
Gram-positive	*Staphylococcus aureus*	1.09% (1.02%)	14.81% (6.35%)
Gram-positive	*Enterococcus faecalis*	0.16% (0.04%)	8.52% (2.49%)
Gram-negative	*Pseudomonas aeruginosa*	0% (0.001%)	5.71% (0.26%)
Gram-negative	*Escherichia coli*	0.32% (0.03%)	10.48% (1.06%)
Fungi	*Candida auris*	1.29% (0.98%)	42.86% (5.58%)
Fungi	*Aspergillus fumigatus*	0.41% (0.21%)	36.73% (2.94%)
Fungi	*Candida albicans*	0.34% (0.14%)	30.61% (2.43%)

[a] For each strain tested in EU-OS, the HitRate for the model predictions and experimental results are reported.

Table 3. Impact Quantification of the Model Predictions on the Bioactive Library[a]

Pathogen class	Experimental strain	True active compound percentage	Screening library percentage
Gram-positive	*Staphylococcus aureus*	25.6%	10.96%
Gram-positive	*Enterococcus faecalis*	37.4%	10.96%
Gram-negative	*Pseudomonas aeruginosa*	46.1%	2.13%
Gram-negative	*Escherichia coli*	21.1%	2.13%
Fungi	*Candida auris*	15.3%	1.98%
Fungi	*Aspergillus fumigatus*	24.8%	1.98%
Fungi	*Candida albicans*	25%	1.98%

[a] For each strain tested in EU-OS, the number of actives found with a small percentage of screening library is shown.

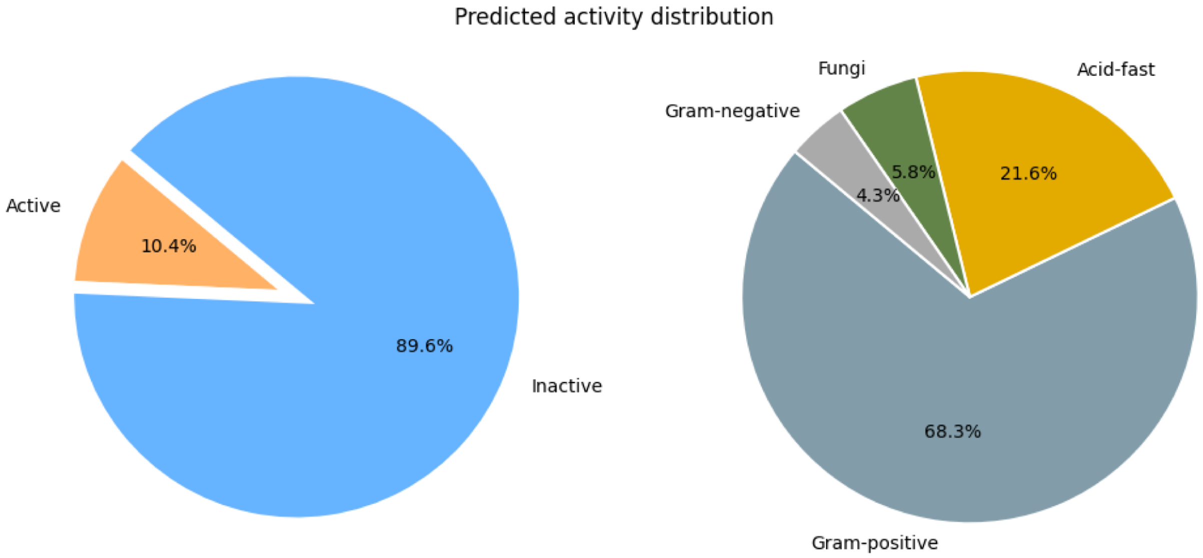


Figure 9. Overview of hits in Enamine Antibacterial Library based on model predictions. The left pie chart summarizes the overall predictions, while the right chart shows the distribution of predicted active compounds (10%) in the different pathogen classes.

model-predicted hits and experimental hits was 0.17, whereas for the Bioactive library, this difference was markedly higher, at 18.37. The highest hit rate was observed for the Fungal strains, followed by the Gram-positive and Gram-negative, respectively. We also observed a similar trend in the hit rate difference for each pathogen class (Table S6). As with any ML model, the higher hit rate observed for the Bioactive library can be attributed to the similarity of its chemical space to the model's training data set.

In addition to assessing the hit rate success of the model, we quantified its potential impact on future predictions, particularly in terms of cost savings. A key consideration in this analysis was addressing the question: How much money could be saved by leveraging our ML models? To explore this, we compared the compounds predicted to be active by the model against the entire compound library. In a typical scenario, such as with the EU-OS library, all compounds would need to be tested against all pathogens to identify a hit—defined as a compound exhibiting activity against a specific pathogen class or strain. However, by utilizing the model prior to experimental screening, a significantly smaller subset of compounds needs to be tested to achieve a comparable hit rate. For example, in the EU-OS library, the highest hit rate for *Staphylococcus aureus* was achieved with a library size of 100,000 compounds (Figures 7 and 8). The model was able to identify nearly 25% of these active compounds by validating only 11% of the library (Table 3). On a broader scale, for Gram-positive pathogens, testing just 11% of the compound library was sufficient to capture over 30% of the hits, on average. Similarly, for Gram-negative pathogens and Fungi, as little as 2% of the compound library needed to be tested to identify active hits. This highlights the efficiency and cost-effectiveness of incorporating ML models into the screening process, significantly reducing the experimental workload while maintaining high hit rates.

On the other hand, for the 32,000 compounds from Enamine, we found ~90% of the compounds to demonstrate inactivity for antibacterial, antifungal, and antituberculosis activities. For the remaining 10%, many compounds are predicted to show activity against Gram-positive (~68%, 2,276 compounds) and acid-fast (~21%, 719 compounds) pathogens (Figure 9). In addition to looking at prediction probabilities like that of EU-OS prediction analysis, we looked at the cosine similarity between the MHFP6 fingerprints of library compounds and the training data set to assess the applicability domain of each predicted data set. A median cosine similarity of 0.71 was reported for the compound library, with six compounds demonstrating a perfect cosine similarity of 1 to the training data sets. If we focus on high-confidence predictions (prediction probabilities >0.5), we retained 27 compounds from the library with predicted activity against Gram-positive (8 of 33 compounds), Fungi (11 of 33 compounds), Gram-negative (1 of 33 compounds), and Acid-fast (7 of 33 compounds pathogens (Table S7). These findings underscore the model's robustness and ability to identify subset antibacterial compounds from large commercial libraries. By focusing on compounds with high-confidence predictions (in this case >0.5), we can potentially prioritize candidates for further experimental validation and development or subset more specific novel antimicrobial libraries, potentially accelerating the discovery of new treatments for AMR. The predictions for all the compounds in the Enamine library can be found on Zenodo (see **Data Availability**).

3. DISCUSSION AND CONCLUSION

Antimicrobial resistance (AMR) has been exponentially emerging as a major threat and is projected to be the leading cause of death by 2050.[46−48] A recent study highlighted that about 5 million deaths per year are associated with AMR, with an increasing burden across low and middle-income countries.[49−51] With this growing crisis, there is an urgent need for advancements in delivering potential drugs to tackle antibiotic resistance. To assist this effort, we have developed machine learning models designed to streamline and accelerate the AMR drug discovery workflow, thereby supporting researchers in identifying effective treatments more efficiently.

Previous attempts in this direction were limited by their scope, such as focusing solely on antibacterial peptide prediction[4,52] or bacterial species specificity.[53,54] Additionally, such studies often demonstrate a limited use of training data sets from a single source or in-house data. Our approach aims to overcome these limitations by generating a comprehensive data set covering a broader scope, enhancing the potential to discover effective antimicrobial treatments. We illustrated that integrating multiple public data set collections is possible using FAIR principles. This approach ensured the consistent use of ontologies and controlled vocabularies, resulting in a medicinal chemistry-driven small-molecule bioactivity graph, the AntiMicrobial-KG. The KG consisted of experimental bioassay data for 81,490 compounds stored in a property-based graph format. The bioactivity end points were tested across 1,373 species, categorized into four broad pathogen classes: Gram-positive, Gram-negative, Acid-fast, and Fungi. Through the KG, we deciphered certain trends between the compounds tested for microbes and fungi. As expected, a larger number of compounds were tested against Gram-positive (∼21,000 compounds) and Gram-negative (∼9,000 compounds) pathogens compared to the Acid-fast pathogens (∼6,000 compounds). Moreover, the KG revealed a substantial presence of compounds with Michael acceptors, which have been previously known to exhibit antimicrobial and antifungal properties.[40−42] Additionally, a dominance of specific chemical classes was observed. For instance, avermectins and sesquiterpenoids were predominantly found in compounds tested against Fungi, while azo and azoxy alkaloids were prevalent in compounds tested for Gram-positive pathogens.

Following this, we showcased one potential use of the AntiMicrobial-KG in the context of the current trend of ML-based predictions. To do so, first, we converted the compounds in AntiMicrobial-KG into the descriptive data set using chemoinformatics techniques. Chemoinformatics offers more than 4,000 descriptors based on 2D and 3D representations of a compound.[55] However, many of these descriptors are highly correlated, necessitating strategic selection of the most representative ones or, ideally, the most interpretable ones. Given the debate about the superiority of 2D descriptors over 3D ones, we chose to avoid 3D descriptor-based representation.[56] Complementary to compound classical descriptors, compound structural features can also be represented through fingerprints.[57] Hence, we decided to investigate multiple fingerprinting methods based on their compactness and ability to represent complementary information such as physicochemical properties, structural connectivity, and pharmacophoric contents. Our exploration of the ML-based prediction approaches yielded various conclusions. First, addressing the imbalances in compounds tested across the five activity groups (Gram-positive, Gram-negative, Acid-Fast, Fungi, and Inactive) with the SMOTE technique significantly improved model predictions. Second, training models through systematic, progressive selection criteria assist in identifying the best model for a specific use case. Historically, ML models were trained based on community recognition (such as random forests) or black-box models (such as deep neural networks), with a few researchers comparing across a range of models. For instance, we found an average of 6-fold difference between the least-performant model (Naive Bayes) and the top-performing one (Random Forest), thus assisting us in selecting the best model from the cohort. Analogously, we used several molecular fingerprints to determine the best model-fingerprint combination. For our use case, the MHFP6-Random forest combination performed the best of all possible combinations. Lastly, while classical structural and pharmacophoric-based fingerprints outperform physicochemical properties, training models on the physicochemical representation of compounds can offer valuable insights into drug development and optimization.[58] For instance, our ChemPhys-Random Forest model highlighted hydrogen bond donor and LogP as key characteristics influencing antifungal and antibacterial activity. This finding aligns with existing research, reinforcing its relevance.[59] Additionally, we found that the number of aliphatic rings in a compound could contribute positively toward Gram-positive activity and negatively toward Gram-negative activity. Lastly, we demonstrated the applicability of our model predictions on two compound libraries commonly used in drug discovery: the EU-OPENSCREEN library and the Enamine Antibacterial collection. Our goal was to showcase the practical use of ML models for the preliminary screening of compound activities across libraries. The strategy involved improving filtering options, allowing users to select molecules with the highest probabilities of activity, thus reducing screening costs. Additionally, this approach can help commercial library vendors enhance their collections by including more active molecules. More ambitiously, we aimed to confirm and highlight key chemical features in the screened compounds that might positively distinguish active from inactive compounds for each pathogen class. Any novel insights in this area would be valuable, given that the training set used here represents one of the most extensive chemical spaces ever assembled from public sources to our knowledge. Moreover, these general predictions can be further refined to target specific bacterial species using existing ML models.[60]

Despite the promising results, our current approach has, however, limitations at multiple stages, from data processing to modeling and prediction steps. First, the pathogen classes assigned to the compounds were based on the highest recorded MIC activity, meaning each compound was linked to only one pathogen class. This association implies high risks and could be misleading, as many "broad-spectrum" compounds show activity across multiple pathogen classes. Additionally, the majority of activities analyzed were driven by cellular responses, which were biased by differences in

Table 4. Number of the Chemicals Collected from the Different Data Resources in the Four Categories (Gram-Positive, Gram-Negative, Acid-Fast, and Fungi)[a]

Resource name	# Gram-positive	# Gram-negative	# Acid-fast	# Fungi	# Unique total
ChEMBL (v.34)	43,976	33,310	17,974	11,488	66,575
CO-ADD	147	113	-	112	286
SPARK	3,085	13,826	162	2	14,629

[a] It is key to note that the same chemical can be analyzed in multiple strains, and hence, the final total (i.e. # Unique total) of the chemicals is calculated based on distinct InChiKey representations. Chemicals in each of these categories may be active or inactive against the pathogen class.

cellular permeability across pathogens, potentially skewing the results in an untracked way. Another challenge lies in data collection assembly. While some sources, such as ChEMBL, ensure data quality through manual curation, others, like SPARK and CO-ADD, may employ quality control methods of inhomogeneous levels. These distinct approaches might not necessarily be synergistic, thus yielding variations in overall data quality. On the modeling side, the lack of benchmark and validation data sets of the same size also posed limitations. The absence of structural diversity in the model's validation sets may have hampered the model's ability to produce more generalizable results. Nevertheless, the dynamic nature of antibiotic research, with its ever-growing number of published data sets combined with the predictive models we present here, offers a valuable tool for investigators to test and predict outcomes on their data sets. Lastly, while the inference of physicochemical "trend rules" through feature importance analysis is tied to the training data set, one could doubt that the model's findings reflect real inherent patterns in the data but only patterns present in the training set. However, this is similar to other widely accepted rules, such as Lipinski's[61] or Veber's,[62] which were also derived from such data-driven inferences. In reality, being the chemical space size from which our "trend rules" are derived, the largest published data set to date, the approach offers a more robust foundation for future research than any former attempts in this direction. Last but not least, our approach in training ML models using chemical fingerprints allows interested users, academic or commercial, to contribute to the future development of the models using the shared Web site for predictions of their molecules while preserving compound structural information. Moreover, with the current model, researchers could quickly test (*in-silico*) chemical libraries for antimicrobial activity and subsequently either validate the results from the model or screen new chemical space and provide the results for training the model for better accuracy in the future.

4. METHODS

4.1. Aggregation of Antibacterial Experimental Data.

We created an Antimicrobial Resistant Knowledge Graph (AntiMicrobial-KG), an exhaustive data warehouse of experimentally validated antibacterial chemicals covering Gram-positive, Gram-negative, acid-fast bacteria and fungi. The construction of the AntiMicrobial-KG involved collecting minimum inhibitory concentration (MIC) data from three public data resources: CO-ADD,[63] ChEMBL,[64] and SPARK[65] (Table 4). Since each resource used unique identifiers for chemicals and the bacterial species they were tested on, we implemented a two-step process to harmonize the data set. Initially, we manually classified pathogens into four classes: Gram-positive, Gram-negative, Acid-fast, and Fungi. Subsequently, we standardized the chemicals to extract database identifiers and their corresponding representations in the form of SMILES, InChIKeys, and InChI. We selected only those chemical-bacterial pairs that fulfilled two conditions: a) exact MIC50 value (i.e., those results with greater than or less than signs were omitted, and only those with exact values with sign (=) or less than equal to sign (≤) retained) and b) experiments reported with standard result units (i.e., μg/mL) were selected. The original source of the tested chemicals was retained, allowing the traceability of chemical-bacteria pairs to their origin. Moreover, to enable efficient comparison across these different experiments and resources, we standardized experimental values into a logarithmic scale metric similar to that implemented within ChEMBL (known as the pChEMBL value).[64] A pChEMBL cutoff of greater than 0 was used to avoid certain erroneous activity results like negative values. This approach facilitated consistent and comparable analysis of antimicrobial resistance data across diverse data sets and experimental conditions.

4.1.1. CO-ADD. Led by the University of Queensland, Australia, CO-ADD (http://db.co-add.org/) is a collaborative crowdsourcing approach aimed at advancing antibiotic drug discovery.[63] The repository is strategically curated with screening experimental data collected for ESKAPE pathogens (*Enterococcus faecium*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Acinetobacter baumannii*, *Pseudomonas aeruginosa*, and *Enterobacter species*) along with various fungal strains.[66] This bacterial profiling data set covers over 9,000 small molecules and peptide-based chemicals. CO-ADD is distinguished by its experimental approach, where the same chemical is tested across its entire strain library. This allows for in-depth chemical-strain specificity studies, a feature not commonly found in other resources where chemicals are typically tested for a single strain or strain class. As a result, the AntiMicrobial-KG integrates 286 chemicals from this resource that fit the previously described conditions (i.e., experimental end points with exact values and pChEMBL > 0).

4.1.2. ChEMBL. ChEMBL version 34 (https://www.ebi.ac.uk/chembl/) is a large-scale database repository for bioactivity data, focusing on small molecules and their effects on biological entities, including proteins, cell lines, and entire organisms.[64] The database provides detailed information on the resulting bioactivities of these interactions. ChEMBL is one of the most comprehensive resources for bioassays, capturing a wide range of data, including binding (B), functional (F), adsorption (A), distribution (D), metabolism (M), excretion (E), and toxicity (T) related chemical bioactivity recorded in literature and patent documents. With over 1.6 million bioassays from multiple species documented, ChEMBL serves as a critical resource for integrating extensive bioactivity data into the AntiMicrobial-KG. From this repository, we selected approximately 63,000 data points involving around 66,000 chemicals in 1,317 strains.

4.1.3. SPARK. The Shared Platform for Antibiotic Research (SPARK), now integrated and maintained by the CO-ADD community, was initially created by the Pew Charitable Trusts to expand research around antibiotics targeting Gram-negative bacteria.[65] Through this initiative, stakeholders from industry, academia, and government collaborated to develop an open platform for data sharing with the support of Collaborative Drug Discovery (CDD). Industrial partners, including Novartis and Merck, contributed published and unpublished experimental data to CDD, where domain experts, including microbiologists, curated the data to ensure consistency and establish evidence links for the recorded activities. The curation process involved harmonizing the data to maintain uniformity and reliability. The curated data was then reintegrated into CDD, creating a valuable resource for researchers and drug developers. Over time, SPARK expanded beyond Gram-negative bacterial strains to include a broader range of bacterial and fungal strains. For the SPARK data, an additional preprocessing step was performed to ensure compliance with tidy principles, wherein the standard relation and standard value are present in two columns.[67] Within the AntiMicrobial-KG, we incorporated data for approximately 14,629 chemicals from SPARK.

Pooling together the public bioassay resources, we constructed our AntiMicrobial-KG. As illustrated in Figure 10, some chemicals appeared in multiple databases, necessitating an additional standardization step to deduplicate chemicals tested on the same strain class. To address this, we retained only the chemical with the assay value closest to the median of all reported pMIC values. This process reduced the collected data set from 81,486 to 74,202 chemicals. Notably, all assays in the final collection possessed nonzero values, a characteristic prominently observed in SPARK data sets. In addition to storing chemical activity data, we also incorporated chemical classifications using the NPClassifier (https://npclassifier.ucsd.edu/).[68] This tool employs a chemical classification ontology to annotate properties, such as the pathways (for e.g., alkaloids, terpenoids, etc.) and superclasses (for e.g., glycerophospholipids, fatty amines, beta-lactams, etc.) in which the chemicals are involved. This ensured that the AntiMicrobial-KG not only captures bioactivity data but also insights into the functional roles of the chemicals, thereby enhancing overall utility.

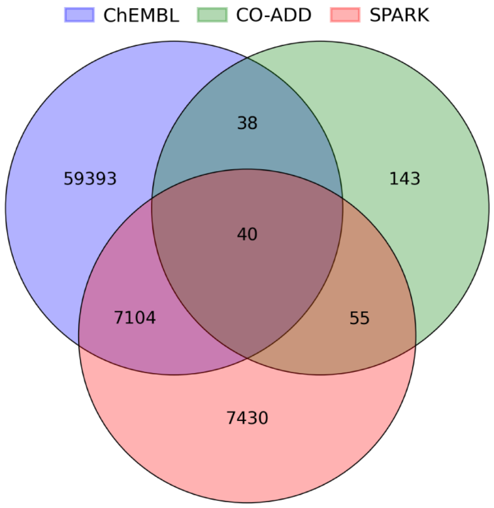


Figure 10. Set diagram showing the overlap between the chemicals found across the three data resources. A total of 40 chemicals are reported in all databases, while other chemicals remain unique to their original data repositories.

4.2. Fingerprint Generation for Model Training. To train the models, chemical structures from the AntiMicrobial-KG are encoded into fingerprint vectors. A fingerprint is a classic representation of a chemical, serving as a molecular descriptor that relies on factors such as molecular connectivity, physicochemical properties, and functional group annotations. For this purpose, we employed the open-source tool RDKit (https://www.rdkit.org/). Utilizing RDKit, we converted 2D chemical structures into various classic fingerprints, including Morgan or Extended-connectivity fingerprints with a bond order of 4 (ECFP4),[69] topological or RDKit fingerprint with 1−7 atoms (RDKit), Molecular ACCess System (MACCS),[70] and 2D Pharmacophore fingerprint ErG.[71] Each of these fingerprints represents a one-hot encoding vector of bit sizes 1024, 2048, 167, and 315, respectively. Additionally, we incorporated the MinHash fingerprint, up to six bonds (MHF6, 2048 bits), for its proven efficacy in chemical retrieval.[72] Alongside these binary vector fingerprints that provide limited information about the chemical features, we added a physicochemical fingerprint (ChemPhys) with 29 molecular properties. The main benefit of this fingerprint was its ease of interpretability, which allowed for insights into molecular optimization strategy. These properties include generic descriptors such as SLogP, surface molecular resonance (SMR), Labute accessible surface area (ASA), polar surface area (TPSA), molecular weight (MW), number of Lipinski hydrogen bond donors (HBD) and acceptors (HBA), number of rotatable bonds, number of HBD, number of HBA, number of amide bonds, number of heteroatoms, number of heavy atoms, the total number of atoms, and various counts of rings (total, aromatic, saturated and aliphatic). Specific descriptors included fragment complexity, rotatable bond count (terminal and nonterminal), number of bridgehead carbon atoms shared between rings, number of stereocenters (specified and unspecified), and counts of heterocycles (aromatic and saturated) and carbocycles (aromatic, saturated, and aliphatic).

To ensure the robustness of all feature columns (i.e., the bit vector representing the chemical fingerprint), we conducted a check for null variance in each feature across the chemicals. The reason for performing a null variance check is to reduce noise for model training and to confirm that every feature provided meaningful information, thus contributing to the chemical characterization for the model's understanding. By integrating these diverse fingerprinting techniques, we created a comprehensive and detailed representation of the chemical structures, facilitating effective model training and subsequent predictions.

4.3. Train and Test Data Sets. For training and evaluating our machine learning (ML) model, we split the chemical data in AntiMicrobial-KG into train and test sets. Since our goal was to predict the activity of chemicals, the data set was organized into chemical-pathogen class pairs. Compounds were first classified as active (pChEMBL > 5) or inactive (pChEMBL ≤ 5) based on their pChEMBL values. For those in the active category, pathogen class selectivity was determined using a "best-of-four" approach, where the pathogen class (Gram-positive, Gram-negative, fungi, or acid-fast) with the highest potency, as indicated by its pChEMBL value, was selected for each chemical. If the compound was found to be inactive against all pathogens, it was labeled in the "inactive" category.

The initial distribution of chemicals across the five categories revealed class imbalances: 21,148 chemicals in Gram-positive, 9,083 in Gram-negative, 7,631 in fungi, 5,845 in acid-fast and 15,654 in inactive classes. To address this imbalance and enhance the robustness of our ML model, we employed an oversampling strategy during data set splitting. Specifically, we used the Synthetic Minority Oversampling Technique (SMOTE), which involves either undersampling of the majority class (in this case, the Gram-positive class) or oversampling of the minority class (in this case, the fungi class) to create synthetic sampled to balance the class.[73] This resulted in a balanced data set with 21,148 chemical-pathogen class pairs in each class.

For our training-testing splits, we allocated approximately 80% of chemical-pathogen pairs to the training set and the remaining 20% to the test set. Importantly, SMOTE was applied only to the training data, ensuring that the test set remained an accurate representation of the real-world data distribution. This approach allowed us to train more robust and reliable ML models capable of predicting chemical activities across different pathogen classes.

4.4. Development and Evaluation of Machine Learning Models. For this multiclass classification task, we developed a reproducible pipeline to test multiple models and optimize the best-performing one. This pipeline was designed with versatility in mind, allowing it to be applied to any use case, provided the data is available in the requisite format. We began by leveraging PyCaret's (https://pycaret.org/) model comparison pipeline to streamline the selection process for the best-performing ML models from a cohort. In this pipeline, we trained six classic ML models: Naive Bayes (NB), Logistic regression (LR), Light Gradient Boosting Machine (LightGBM), Decision Tree (DT), Random Forest (RF), and eXtreme Gradient Boosting (XGBoost) with a 5-fold cross-validation strategy. The choice of these models was driven by their transparent nature, as opposed to the black-box nature of certain advanced ML models like neural networks. For training these models, the initial training data (80%) was further split into train (60%) and validation or hold-out set (20%). The trained models were ranked based on their performance in predicting the hold-out set, using the Cohen-Kappa score as the performance metric. It is important to note that PyCaret's model training pipeline internally applies various data preprocessing steps, including imputation, scaling, and normalization. To further refine the best model, we performed an additional optimization strategy using hyper-parameter optimization (HPO) with Optuna (https://optuna.org/). This HPO pipeline involved 15 trial runs aiming to maximize the Cohen-Kappa score. Similar to PyCaret, the score for optimization was determined from the 5-fold cross-validation strategy. This two-layered approach (model selection and model refinement) ensures the development of a robust and globally optimized model for evaluation in our multiclass classification task.

In addition to the Cohen Kappa score, we calculated other prediction metrics such as accuracy and Area Under the Curve of Receiver Operating Characteristic (AUC-ROC). These metrics were used not only to select the best model from our collection but also to compare our model with existing ones in the field of AMR. Once the optimal model was identified, it was used to predict the outcomes of the test set (the remaining 20% of the data set), demonstrating its effectiveness and applicability. Furthermore, we used Shapley values to identify the most influential features affecting the strain class within the data set. Identifying the top ten most influential features enhances the transparency and interpretability of the model, regardless of the algorithm employed. This robust and transparent approach ensures that our model is not only highly accurate but also interpretable, providing valuable insights into the chemical features driving antimicrobial resistance.

4.5. External Compound Libraries Used for Predictions and Validation. Additionally, external data sets were collected to evaluate and predict the antibacterial, antifungal, and antituberculosis activity of compounds. This involved the compounds from two sources: a) the subset of Enamine compound library (with 32,000 compounds) specially optimized for AMR, the Antibacterial Library (https://enamine.net/compound-libraries/targeted-libraries/antibacterial-library) and b) EU Openscreen (EU-OS) European Chemical Biology Library (https://ecbd.eu/) with 101,024 compounds. The Enamine library was designed with privileged scaffolds from known antibacterial drugs and their underlying physicochemical properties in mind. The EU-OS library was designed based on four different chemoinformatic-rich content approaches and is not directed toward antibacterial but rather is a collection that includes a range of novel and diverse scaffolds. To distinguish these two, we call them the ECBL set (96,092 compounds) and the Bioactive set (4,927) and the results of the reports for these sets independently.

To compare the model predictions with the experimental predictions, we make use of the hit rate. The HitRate for a specific pathogen class is described as

$$\text{HitRate} = \frac{T_{\text{active}}}{T_{\text{active}} + T_{\text{inactive}}}$$

where T_{active} represents the bioactive compounds, and $T_{inactive}$ denotes the inactive compounds for the specific pathogen class. For experimental results, a compound is considered active if it shows 50% or greater inhibition. This metric allows for a quantitative comparison of the hits identified by the model with the experimental "gold standard" results.

4.6. Experimental Testing of EU-OS EBCL Library. In addition to the model predictions, the library was tested in a high-content screening viability assay against seven microbial pathogens: three fungi (*C. auris* DSM21092, *C. albicans* ATCC 64124, *A. fumigatus* ATCC 46645), two Gram-positive (*S. aureus* ATCC 29213 and *E. faecalis* ATCC 29212), and two Gram-negative (*P. aeruginosa* and *E. coli*). A compound concentration of 50 μM was used, and the results were reported as percentage inhibition. All results, along with the experimental protocols, can be found in the ECBD database (https://ecbd.eu/) with the following assay identifiers:

- *C. auris* DSM21092 - EOS300072
- *C. albicans* ATCC 64124 - EOS300076
- *A. fumigatus* ATCC 46645 - EOS300074
- *S. aureus* ATCC 29213 - EOS300078
- *E. faecalis* ATCC 29212 - EOS300080
- *P. aeruginosa* group—EOS300155
- *E. coli* ATCC 25922 - EOS300158

4.7. Implementation. The data set was generated using Python version 3.9. For the generation of the chemical fingerprints and molecular scaffolds, we took advantage of RDKit as implemented in Python. The stratification of the data into train-test splits was performed using the Scikit-learn library (https://scikit-learn.org), and SMOTE was implemented using the Imbalance-learn library (https://imbalanced-learn.org). To ensure the reproducibility of the sampling procedures, the data splitting was seed-coded. The models were built in PyCaret (v.3.2.0) and Scikit-learn (v.1.3.2). Feature importance was determined using the TreeInterpretor (https://pypi.org/project/treeinterpreter/) python package. For the AntiMicrobial-KG, we provide the opportunity to generate the graph with Neo4J (https://neo4j.com/), a commercially available graph database analysis and visualization tool. The Web site for displaying the database information was built using Streamlit (https://streamlit.io/) and is hosted on SciLifeLab's SERVE (https://serve.scilifelab.se/) instance.

■ ASSOCIATED CONTENT

Data Availability Statement

All the Python scripts used in this manuscript for model training, exploratory analysis, and KG generation are available on GitHub at https://github.com/IMI-COMBINE/broad_spectrum_prediction. The data collected with the AntiMicrobial-KG and the models can also be found on Zenodo at https://zenodo.org/records/13868088. Also, we have the AntiMicrobial-KG Web site at https://antimicrobial-kg.serve.scilifelab.se/to allow users to search the database and use our pretrained models for their compound activity prediction.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.4c02347.

> Figure S1: Nonpromiscuous scaffolds found in AntiMicrobial-KG. Figure S2: NP-likeness distribution of chemicals found in AntiMicrobial-KG. Figure S3: Comparison of SMOTE vs non-SMOTE for the top two performing models. Figure S4: Feature importance of MHFP6 fingerprints. Figure S5: Feature importance of MHFP6 fingerprints. Table S1: Summary of Ro5 violations and structural alerts in AMR-KG. Table S2: Evaluation metrics on hold-out set for all chemical fingerprints across six ML models. Table S3: Hyperparameters used for optimizing the top two models: Random forest and XGBoost. Table S4: XGBoost model performance on the test data. Table S5: Random Forest model performance on the test data. Table S6: EU-OS library prediction for each pathogen class. Table S7: High confidence prediction of Enamine library (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Yojana Gadiya − *Fraunhofer Institute for Translational Medicine and Pharmacology (ITMP), Hamburg 22525, Germany; Bonn-Aachen International Center for Information Technology (B-IT), University of Bonn, Bonn 53113, Germany;* orcid.org/0000-0002-7683-0452; Email: Yojana.Gadiya@itmp.fraunhofer.de

Authors

Olga Genilloud − *Fundación MEDINA, Centro de Excelencia En Investigación de Medicamentos Innovadores En Andalucía, Armilla 18016, Spain;* orcid.org/0000-0002-4202-1219

Ursula Bilitewski − *Helmholtz Centre for Infection Research, Braunschweig 38124, Germany*

Mark Brönstrup − *Helmholtz Centre for Infection Research, Braunschweig 38124, Germany; German Center for Infection Research, Hannover 38124, Germany;* orcid.org/0000-0002-8971-7045

Leonie von Berlin − *Fraunhofer Institute for Translational Medicine and Pharmacology (ITMP), Hamburg 22525, Germany*

Marie Attwood − *PK/PD Laboratory, North Bristol, NHS Trust, Southmead Hospital, Bristol BS10 5NB, U.K.*

Philip Gribbon − *Fraunhofer Institute for Translational Medicine and Pharmacology (ITMP), Hamburg 22525, Germany*

Andrea Zaliani − *Fraunhofer Institute for Translational Medicine and Pharmacology (ITMP), Hamburg 22525, Germany*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.4c02347

Author Contributions

Y.G.: Conceptualization, Methodology, Writing—Original draft preparation and Visualization; A.Z.: Conceptualization, Supervision, Methodology, Writing—Original draft preparation and Visualization; L.B.: Visualization; O.G.: Investigation; U.B.: Investigation; M.B.: Investigation; M.K.: Writing—Reviewing and Editing; P.G.: Writing—Reviewing and Editing.

Funding

This work is part of the COMBINE project that has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 853967. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation program and EFPIA. The work reflects the author's view and that neither IMI nor the European Union, EFPIA, or any Associated Partners are responsible for any use that may be made of the information contained therein.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We would like to thank the authors of the resources used in our work for making their datasets available to the scientific community. We would like to acknowledge Dr. Frederik Deroose, Dr. Anders Karlén, Dr. Axel Pahl, Dr. Maria Kuzikov and all other members of the AMR Accelerator Scientific Interest Group on Machine Learning group for their feedback for improving the manuscript. We would also like to thank the anonymous reviewers for their comments and suggestions for improving the readability of the manuscript.

■ REFERENCES

(1) Palmer, J. D.; Foster, K. R. The evolution of spectrum in antibiotics and bacteriocins. *Proc. Natl. Acad. Sci. U. S. A.* **2022**, *119* (38), No. e2205407119.

(2) Peterson, E.; Kaur, P. Antibiotic resistance mechanisms in bacteria: Relationships between resistance determinants of antibiotic producers, environmental bacteria, and clinical pathogens. *Front. Microbiol.* **2018**, *9*, 2928.

(3) O'Neill, J. *Tackling drug-resistant infections globally: Final report and recommendations*; Government of the United Kingdom. https://apo.org.au/node/63983.

(4) Murray, C. J.; Ikuta, K. S.; Sharara, F.; Swetschinski, L.; Aguilar, G. R.; Gray, A.; Han, C.; Bisignano, C.; Rao, P.; Wool, E.; et al. Global

burden of bacterial antimicrobial resistance in 2019: A systematic analysis. *Lancet* **2022**, *399* (10325), 629−655.

(5) Mestrovic, T.; Aguilar, G. R.; Swetschinski, L. R.; Ikuta, K. S.; Gray, A. P.; Weaver, N. D.; Han, C.; Wool, E. E.; Hayoon, A. G.; Hay, S. I.; et al. The burden of bacterial antimicrobial resistance in the WHO European region in 2019: A cross-country systematic analysis. *Lancet Public Health* **2022**, *7* (11), No. e897−e913.

(6) Essack, S. Y.; Lenglet, A. Bacterial antimicrobial resistance burden in Africa: Accuracy, action, and alternatives. *Lancet Glob. Health* **2024**, *12* (2), No. e171−e172.

(7) Ventola, C. L. The antibiotic resistance crisis: Part 1: Causes and threats. *Pharm. Ther.* **2015**, *40* (4), 277−283.

(8) Zankari, E.; Hasman, H.; Kaas, R. S.; Seyfarth, A. M.; Agersø, Y.; Lund, O.; Larsen, M. V.; Aarestrup, F. M. Genotyping using whole-genome sequencing is a realistic alternative to surveillance based on phenotypic antimicrobial susceptibility testing. *J. Antimicrob. Chemother.* **2013**, *68* (4), 771−777.

(9) Genestet, C.; Hodille, E.; Berland, J. L.; Ginevra, C.; Bryant, J. E.; Ader, F.; Lina, G.; Dumitrescu, O. Study Group Whole-genome sequencing in drug susceptibility testing of Mycobacterium tuberculosis in routine practice in Lyon, France. *Int. J. Antimicrob. Agents* **2020**, *55* (4), 105912.

(10) Ellington, M. J.; Ekelund, O.; Aarestrup, F. M.; Canton, R.; Doumith, M.; Giske, C.; Grundman, H.; Hasman, H.; Holden, M. T.; Hopkins, K. L.; et al. The role of whole genome sequencing in antimicrobial susceptibility testing of bacteria: Report from the EUCAST Subcommittee. *Clin. Microbiol. Infect.* **2017**, *23* (1), 2−22.

(11) Su, M.; Satola, S. W.; Read, T. D. Genome-based prediction of bacterial antibiotic resistance. *J. Clin. Microbiol.* **2019**, *57* (3), 10−128.

(12) Lundstrom, T. S.; Sobel, J. D. Antibiotics for gram-positive bacterial infections: vancomycin, teicoplanin, quinupristin/dalfopristin, and linezolid. *Infect. Dis. Clin.* **2000**, *14* (2), 463−474.

(13) Abu-Farha, R.; Gharaibeh, L.; Alzoubi, K. H.; Nazal, R.; Zawiah, M.; Binsaleh, A. Y.; Shilbayeh, S. A. Awareness, perspectives and practices of antibiotics deprescribing among physicians in Jordan: a cross-sectional study. *J. Pharm. Policy Pract.* **2024**, *17* (1), 2378484.

(14) Bdair, I. A.; Bdair, O. A.; Maribbay, G. M. L.; Elzehiri, D.; Hassan, E. S.; Tolentino, A. D.; Emara, M. M.; Ali, E. K.; Abdelmeged, R. M.; Fadul, M. O. Public awareness towards antibiotics use, misuse and resistance in Saudi community: A cross-sectional population survey. *J. Appl. Pharm. Sci.* **2024**, *14*, 217−226.

(15) Rodríguez-Gascón, A.; Solinís, M. Á.; Isla, A. The role of PK/PD analysis in the development and evaluation of antimicrobials. *Pharmaceutics* **2021**, *13* (6), 833.

(16) Sy, S. K.; Derendorf, H. Pharmacokinetics I: PK-PD approach, the case of antibiotic drug development. *Clinical pharmacology. Clin. Pharmacol.* **2016**, 185−217.

(17) Sou, T.; Hansen, J.; Liepinsh, E.; Backlund, M.; Ercan, O.; Grinberga, S.; Cao, S.; Giachou, P.; Petersson, A.; Tomczak, M.; et al. Model-informed drug development for antimicrobials: translational PK and PK/PD modeling to predict an efficacious human dose for apramycin. *Clin. Pharmacol. Ther.* **2021**, *109* (4), 1063−1073.

(18) Lin, T. T.; Yang, L. Y.; Lin, C. Y.; Wang, C. T.; Lai, C. W.; Ko, C. F.; Shih, Y. H.; Chen, S. H. Intelligent de novo design of novel antimicrobial peptides against antibiotic-resistant bacteria strains. *Int. J. Mol. Sci.* **2023**, *24* (7), 6788.

(19) Jukič, M.; Bren, U. Machine learning in antibacterial drug design. *Front. Pharmacol.* **2022**, *13*, 864412.

(20) Melo, M. C.; Maasch, J. R.; de la Fuente-Nunez, C. Accelerating antibiotic discovery through artificial intelligence. *Commun. Biol.* **2021**, *4* (1), 1050.

(21) Anahtar, M. N.; Yang, J. H.; Kanjilal, S. Applications of machine learning to the problem of antimicrobial resistance: an emerging model for translational research. *J. Clin. Microbiol.* **2021**, *59* (7), 10−128.

(22) Sakagianni, A.; Koufopoulou, C.; Feretzakis, G.; Kalles, D.; Verykios, V. S.; Myrianthefs, P.; Fildisis, G. Using machine learning to predict antimicrobial resistance――a literature review. *Antibiotics* **2023**, *12* (3), 452.

(23) Tyers, M.; Wright, G. D. Drug combinations: a strategy to extend the life of antibiotics in the 21st century. *Nat. Rev. Microbiol.* **2019**, *17* (3), 141−155.

(24) Lv, J.; Liu, G.; Ju, Y.; Sun, Y.; Guo, W. Prediction of synergistic antibiotic combinations by graph learning. *Front. Pharmacol.* **2022**, *13*, 849006.

(25) Qin, J.; Yang, Y.; Ai, C.; Ji, Z.; Chen, W.; Song, Y.; Zeng, J.; Duan, M.; Qi, W.; Zhang, S.; et al. Antibiotic combinations prediction based on machine learning to multicentre clinical data and drug interaction correlation. *Int. J. Antimicrob. Agents* **2024**, *63* (5), 107122.

(26) Cantrell, J. M.; Chung, C. H.; Chandrasekaran, S. Machine learning to design antimicrobial combination therapies: Promises and pitfalls. *Drug Discovery Today* **2022**, *27* (6), 1639−1651.

(27) Chio, H.; Guest, E. E.; Hobman, J. L.; Dottorini, T.; Hirst, J. D.; Stekel, D. J. Predicting bioactivity of antibiotic metabolites by molecular docking and dynamics. *J. Mol. Graphics Modell.* **2023**, *123*, 108508.

(28) Alves, M. J.; Froufe, H. J.; Costa, A. F.; Santos, A. F.; Oliveira, L. G.; Osório, S. R.; Abreu, R. M.; Pintado, M.; Ferreira, I. C. Docking studies in target proteins involved in antibacterial action mechanisms: Extending the knowledge on standard antibiotics to antimicrobial mushroom compounds. *Molecules* **2014**, *19* (2), 1672−1684.

(29) Karnati, P.; Gonuguntala, R.; Barbadikar, K. M.; Mishra, D.; Jha, G.; Prakasham, V.; Chilumula, P.; Shaik, H.; Pesari, M.; Sundaram, R. M.; Chinnaswami, K. Performance of Novel Antimicrobial Protein Bg_9562 and In Silico Predictions on Its Properties with Reference to Its Antimicrobial Efficiency against Rhizoctonia solani. *Antibiotics* **2022**, *11* (3), 363.

(30) Wong, F.; Krishnan, A.; Zheng, E. J.; Stärk, H.; Manson, A. L.; Earl, A. M.; Jaakkola, T.; Collins, J. J. Benchmarking AlphaFold-enabled molecular docking predictions for antibiotic discovery. *Mol. Syst. Biol.* **2022**, *18* (9), No. e11081.

(31) Zhao, F.; Qiu, J.; Xiang, D.; Jiao, P.; Cao, Y.; Xu, Q.; Qiao, D.; Xu, H.; Cao, Y. deepAMPNet: A novel antimicrobial peptide predictor employing AlphaFold2 predicted structures and a bi-directional long short-term memory protein language model. *PeerJ* **2024**, *12* (12), No. e17729.

(32) Feretzakis, G.; Sakagianni, A.; Loupelis, E.; Kalles, D.; Skarmoutsou, N.; Martsoukou, M.; Christopoulos, C.; Lada, M.; Petropoulou, S.; Velentza, A.; et al. Machine learning for antibiotic resistance prediction: A prototype using off-the-shelf techniques and entry-level data to guide empiric antimicrobial therapy. *Healthcare Inform. Res.* **2021**, *27* (3), 214−221.

(33) Weis, C. V.; Jutzeler, C. R.; Borgwardt, K. Machine learning for microbial identification and antimicrobial susceptibility testing on MALDI-TOF mass spectra: A systematic review. *Clin. Microbiol. Infect.* **2020**, *26* (10), 1310−1317.

(34) Carracedo-Reboredo, P.; Liñares-Blanco, J.; Rodríguez-Fernández, N.; Cedrón, F.; Novoa, F. J.; Carballal, A.; Maojo, V.; Pazos, A.; Fernandez-Lozano, C. A review on machine learning approaches and trends in drug discovery. *Comput. Struct. Biotechnol. J.* **2021**, *19*, 4538−4558.

(35) Ivanenkov, Y. A.; Zhavoronkov, A.; Yamidanov, R. S.; Osterman, I. A.; Sergiev, P. V.; Aladinskiy, V. A.; Aladinskaya, A. V.; Terentiev, V. A.; Veselov, M. S.; Ayginin, A. A.; et al. Identification of novel antibacterials using machine learning techniques. *Front. Pharmacol.* **2019**, *10*, 913.

(36) Fjell, C. D.; Jenssen, H.; Hilpert, K.; Cheung, W. A.; Panté, N.; Hancock, R. E.; Cherkasov, A. Identification of novel antibacterial peptides by chemoinformatics and machine learning. *J. Med. Chem.* **2009**, *52* (7), 2006–2015.

(37) Nguyen, M.; Brettin, T.; Long, S. W.; Musser, J. M.; Olsen, R. J.; Olson, R.; Shukla, M.; Stevens, R. L.; Xia, F.; Yoo, H.; Davis, J. J. Developing an in silico minimum inhibitory concentration panel test for Klebsiella pneumoniae. *Sci. Rep.* **2018**, *8* (1), 421.

(38) Gurvic, D.; Leach, A. G.; Zachariae, U. Data-driven derivation of molecular substructures that enhance drug activity in gram-negative bacteria. *J. Med. Chem.* **2022**, *65* (8), 6088−6099.

(39) Liang, S. T.; Chen, C.; Chen, R. X.; Li, R.; Chen, W. L.; Jiang, G. H.; Du, L. L. Michael acceptor molecules in natural products and their mechanism of action. *Front. Pharmacol.* **2022**, *13*, 1033003.

(40) Sherzad Othman, S. Synthesis of Novel Michael Adducts and Study of their Antioxidant and Antimicrobial Activities. *Chem. Rev. Lett.* **2022**, *5* (4), 226−233.

(41) Strharsky, T.; Pindjakova, D.; Kos, J.; Vrablova, L.; Smak, P.; Michnova, H.; Gonec, T.; Hosek, J.; Oravec, M.; Jendrzejewska, I.; et al. Trifluoromethylcinnamanilide michael acceptors for treatment of resistant bacterial infections. *Int. J. Mol. Sci.* **2022**, *23* (23), 15090.

(42) Lee, K. M.; Le, P.; Sieber, S. A.; Hacker, S. M. Degrasyn exhibits antibiotic activity against multi-resistant Staphylococcus aureus by modifying several essential cysteines. *Chem. Commun.* **2020**, *56* (19), 2929−2932.

(43) Ciura, K.; Fedorowicz, J.; Andrić, F.; Žuvela, P.; Greber, K. E.; Baranowski, P.; Kawczak, P.; Nowakowska, J.; Bączek, T.; Sączewski, J. Lipophilicity determination of antifungal isoxazolo [3, 4-b] pyridin-3 (1 H)-ones and their N1-substituted derivatives with chromatographic and computational methods. *Molecules* **2019**, *24* (23), 4311.

(44) Kokot, M.; Weiss, M.; Zdovc, I.; Senerovic, L.; Radakovic, N.; Anderluh, M.; Minovski, N.; Hrast, M. Amide containing NBTI antibacterials with reduced hERG inhibition, retained antimicrobial activity against gram-positive bacteria and in vivo efficacy. *Eur. J. Med. Chem.* **2023**, *250*, 115160.

(45) Limwongyut, J.; Moreland, A. S.; Nie, C.; Read de Alaniz, J.; Bazan, G. C. Amide Moieties Modulate the Antimicrobial Activities of Conjugated Oligoelectrolytes against Gram-negative Bacteria. *ChemistryOpen* **2022**, *11* (2), No. e202100260.

(46) Tang, K. W.; Millar, B. C.; Moore, J. E. Antimicrobial resistance (AMR). *Br. J. Biomed. Sci.* **2023**, *80*, 11387.

(47) The Lancet. Antimicrobial resistance: An agenda for all. *Lancet* **2024**, *403*, 2349.

(48) Fongang, H.; Mbaveng, A. T.; Kuete, V. Global burden of bacterial infections and drug resistance. In *Advances in Botanical Research*; Academic Press, 2023, Vol. *106*, pp. 1−20. DOI: .

(49) Brüssow, H. The antibiotic resistance crisis and the development of new antibiotics. *Microb. Biotechnol.* **2024**, *17* (7), No. e14510.

(50) Bournez, C.; Riool, M.; de Boer, L.; Cordfunke, R. A.; de Best, L.; van Leeuwen, R.; Drijfhout, J. W.; Zaat, S. A.; van Westen, G. J. CalcAMP: A new machine learning model for the accurate prediction of antimicrobial activity of peptides. *Antibiotics* **2023**, *12* (4), 725.

(51) Bajiya, N.; Choudhury, S.; Dhall, A.; Raghava, G. P. AntiBP3: A Method for Predicting Antibacterial Peptides against Gram-Positive/Negative/Variable Bacteria. *Antibiotics* **2024**, *13* (2), 168.

(52) Li, J. T.; Wei, Y. W.; Wang, M. Y.; Yan, C. X.; Ren, X.; Fu, X. J. Antibacterial activity prediction model of traditional Chinese medicine based on combined data-driven approach and machine learning algorithm: Constructed and validated. *Front. Microbiol.* **2021**, *12*, 763498.

(53) Nsubuga, M.; Galiwango, R.; Jjingo, D.; Mboowa, G. Generalizability of machine learning in predicting antimicrobial resistance in E. coli: A multi-country case study in Africa. *BMC Genomics* **2024**, *25* (1), 287.

(54) Todeschini, R.; Consonni, V. *Molecular descriptors for chemoinformatics*; John Wiley & Sons, 2009.

(55) Bahia, M. S.; Kaspi, O.; Touitou, M.; Binayev, I.; Dhail, S.; Spiegel, J.; Khazanov, N.; Yosipof, A.; Senderowitz, H. A comparison between 2D and 3D descriptors in QSAR modeling based on bioactive conformations. *Mol. Inf.* **2023**, *42* (4), 2200186.

(56) Duan, J.; Dixon, S. L.; Lowrie, J. F.; Sherman, W. Analysis and comparison of 2D fingerprints: Insights into database screening performance using eight fingerprint methods. *J. Mol. Graphics Modell.* **2010**, *29* (2), 157−170.

(57) Richter, M. F.; Hergenrother, P. J. The challenge of converting Gram-positive-only compounds into broad-spectrum antibiotics. *Ann. N. Y. Acad. Sci.* **2019**, *1435* (1), 18−38.

(58) Yuan, G.; Guan, Y.; Yi, H.; Lai, S.; Sun, Y.; Cao, S. Antibacterial activity and mechanism of plant flavonoids to gram-positive bacteria predicted from their lipophilicities. *Sci. Rep.* **2021**, *11* (1), 10471.

(59) Boulaamane, Y.; Molina Panadero, I.; Hmadcha, A.; Atalaya Rey, C.; Baammi, S.; El Allali, A.; Maurady, A.; Smani, Y.; Garg, N. Antibiotic discovery with artificial intelligence for the treatment of Acinetobacter baumannii infections. *Msystems* **2024**, *9* (6), No. e00325−24.

(60) Badura, A.; Krysiński, J.; Nowaczyk, A.; Buciński, A. Application of artificial neural networks to prediction of new substances with antimicrobial activity against Escherichia coli. *J. Appl. Microbiol.* **2021**, *130* (1), 40−49.

(61) Lipinski, C. A. Lead-and drug-like compounds: The rule-of-five revolution. *Drug Discovery Today: Technol.* **2004**, *1* (4), 337−341.

(62) Veber, D. F.; Johnson, S. R.; Cheng, H. Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D. Molecular properties that influence the oral bioavailability of drug candidates. *J. Med. Chem.* **2002**, *45* (12), 2615−2623.

(63) Cooper, M. A. A community-based approach to new antibiotic discovery. *Nat. Rev. Drug Discovery* **2015**, *14* (9), 587−588.

(64) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40* (D1), D1100−7.

(65) Thomas, J.; Navre, M.; Rubio, A.; Coukell, A. Shared platform for antibiotic research and knowledge: A collaborative tool to SPARK antibiotic discovery. *ACS Infect. Dis.* **2018**, *4* (11), 1536−1539.

(66) Semenov, V. V.; Raihstat, M. M.; Konyushkin, L. D.; Semenov, R. V.; Blaskovich, M. A. T.; Zuegg, J.; Elliott, A. G.; Hansford, K. A.; Cooper, M. A. Antimicrobial screening of a historical collection of over 140 000 small molecules. *Mendeleev Commun.* **2021**, *31* (4), 484−487.

(67) Wickham, H. Tidy data. *J. Stat. Softw.* **2014**, *59*, 1−23.

(68) Kim, H. W.; Wang, M.; Leber, C. A.; Nothias, L. F.; Reher, R.; Kang, K. B.; Van Der Hooft, J. J.; Dorrestein, P. C.; Gerwick, W. H.; Cottrell, G. W. NPClassifier: A deep neural network-based structural classification tool for natural products. *J. Nat. Prod.* **2021**, *84* (11), 2795−2807.

(69) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50* (5), 742−754.

(70) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42* (6), 1273−1280.

(71) Stiefl, N.; Watson, I. A.; Baumann, K.; Zaliani, A. ErG: 2D pharmacophore descriptions for scaffold hopping. *J. Chem. Inf. Model.* **2006**, *46* (1), 208−220.

(72) Probst, D.; Reymond, J. L. A probabilistic molecular fingerprint for big data settings. *J. Cheminf.* **2018**, *10*, 1−2.

(73) Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; Kegelmeyer, W. P. SMOTE: Snthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321−357.