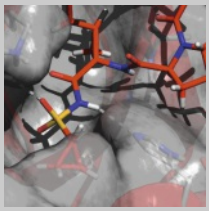
doi:10.2533/chimia.2023.786 Chimia 77 (2023) 786–788



Medicinal Chemistry and Chemical Biology Highlights

Division of Medicinal Chemistry and Chemical Biology

A Division of the Swiss Chemical Society

Trends in Medicinal Chemistry: KNIME Workflows, QSAR Models, LLMs and Chemical Search Strategies

Seyedeh Maryam Salehi*, Melchor Sanchez-Martinez, Kristina Goncharenko, and Natalja Kurbatova

*Correspondence: Dr. S. M. Salehi, E-mail: Maryam.Salehi@zifornd.com
Zifo Technologies Ltd, St Alban-Anlage 66, 4052 Basel, Switzerland,

Abstract: Through a lens encompassing KNIME workflows, QSAR models, LLMs, and chemical substructure search strategies, the article navigates the essential considerations driving innovation and progress in industrial cheminformatics for medicinal chemistry and drug discovery.

Keywords: Chemical search · KNIME workflows · LLM · QSAR

Automation – KNIME Workflows

Automation plays a pivotal role in modern cheminformatics by streamlining complex laboratory processes and data analysis. A recent notable tendency is the rising popularity of KNIME (Konstanz Information Minor) workflows for data transformation and analysis in computer-aided drug design (CADD), where processing extensive datasets of chemicals and proteins is essential.[1–5]

KNIME is an open-source platform offering robust data pipelining, analysis and reporting capabilities. It enables the construction of data workflows, execution of specific analyses, and interactive data visualisation based on pre-implemented code units (nodes) that eliminate extensive coding, making it accessible to a broad audience. The active KNIME community continually contributes to CADD-related nodes, expanding its capabilities.[6] Moreover, some of the most popular tools in CADD, offered as standalone software or libraries for coding languages, are accessible as KNIME nodes.

The key benefit of KNIME in CADD is its capacity to facilitate semi-automated drug discovery pipelines, integrating programmatically accessible open-source repositories like CHEMBL, PubChem, UniProt, and DrugBank.[7] KNIME workflows streamline data utilisation and accelerate the journey from data collection to discovering hidden data relationships. Besides, they are highly reproducible and adaptable, catering to individual project needs. Because of that KNIME workflows ranging from hit identification to ADMET prediction, are gaining increasing attention over time.[8–10]

Predictive Modelling – Toxicity Prediction

A long-standing and ongoing trend in the modern drug discovery is the application of the machine-learning technique called QSAR (Quantitative Structure–Activity Relationship) modelling. QSAR, a computational technique with roots dating back to the 1960s,[11] remains highly relevant, evident from the numerous recent papers indexed in PubMed.[12–14] There are 825 results in Search Results – PubMed (nih.gov) at 09/29/2023 using 'QSAR' as keyword.

The rekindling of interest in QSAR models can be attributed significantly to the support from regulatory authorities, notably the FDA (Food and Drug Administration) and the EMA (European Medicines Agency).[15,16] These agencies are actively championing the reduction of animal testing and advocating for the adoption of alternative methods.

QSAR models learn patterns from a dataset of molecules with known activities and properties. Subsequently, they apply these learned patterns to predict the behavior of new molecules. These models find versatile applications, notably in chemical safety assessment, where they predict chemical toxicity.[17–19] Accessing this data before production aids sustainable product development and promotes green chemistry. The advantages of QSAR modelling extend to reducing both time and costs associated with drug and molecule discovery and production. This reduction conserves resources and addresses ethical concerns, particularly regarding animal testing. While animal testing is necessary to assess the safety and efficacy of new drugs and chemicals, there are growing concerns about the ethical and practical implications of this practice.[20] Avoiding unnecessary animal testing has become a societal debate and a goal of regulatory agencies.

Information Search – LLMs

Almost in any scientific discipline, including cheminformatics, the use of Large Language Models (LLMs) has revolutionized document and information search within the scientific domain.[21–25] LLMs (like Generative Pre-trained Transformer (GPT)) have the capability to comprehensively analyse huge repositories of scientific literature and databases, making them indispensable tools for researchers. Through natural language processing, LLMs can quickly sift through mountains of text to pinpoint specific chemical compounds, reaction mechanisms, and experimental results, drastically expediting the research process. Moreover, LLMs can generate highly relevant and context-aware summaries, aiding scientists in distilling key insights from complex scientific documents. As a result, the integration of LLMs in cheminformatics has not only enhanced the efficiency of information retrieval but has also paved the way for deeper data-driven discoveries and innovations in the field.

Chemical Search – Cloud Databases and GPUs

Another intriguing facet of chemistry, both ancient and contemporary, is the chemical search within (cloud) databases, a domain shaped by the emergence of cloud services (Fig. 1).

Chemical (sub)structure search is vital in drug discovery, but it can be time-consuming, especially for large databases.[26–28] A common approach involves a two-stage process, beginning with fingerprint screening ('classical' chemical search) followed by subgraph isomorphism detection.[29–31] However, fingerprint screening, particularly with larger sizes (>4,096 bits), can be resource intensive.

Graphics Processing Units (GPUs), known for their parallel processing capabilities, have become pivotal in cheminformatics

Can you show us your Medicinal Chemistry and Chemical Biology Highlight?
Please contact: Prof. Dr. Kathrin Lang, Dept. of Chemistry and Applied Bioscience, ETH Zurich, Vladimir Prelog Weg 1-5/10, CH-8093 Zurich, E-mail: kathrin.lang@org.chem.ethz.ch

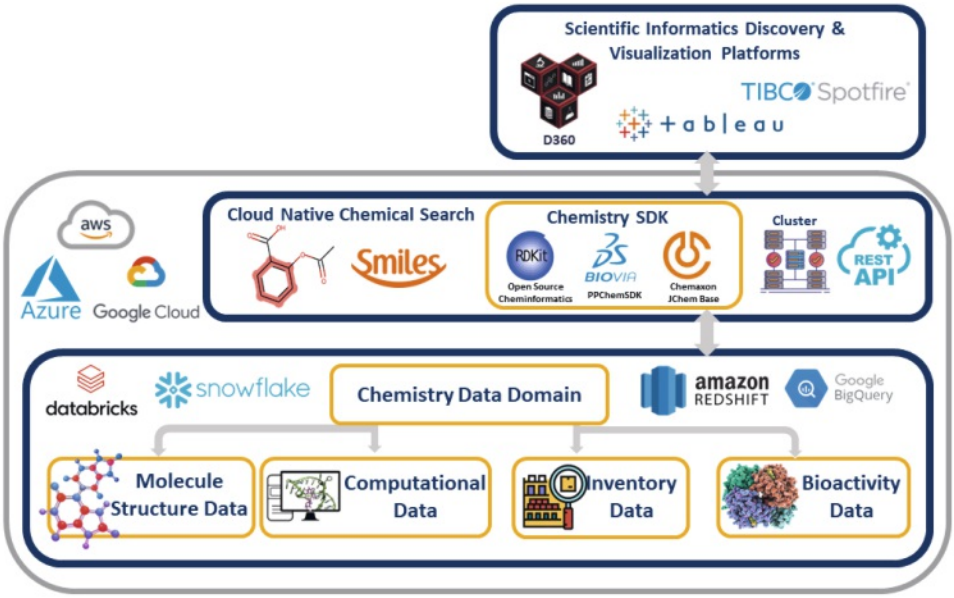


Fig. 1. Efficient cloud-based chemical search leveraging external algorithms and cloud resources.

and bioinformatics.[32–37] Recent GPU advancements, including expanded memory capacity and tools like CuPy, facilitate efficient handling of vast datasets, minimizing data transfer bottlenecks.[38] To address the challenges of cloud databases and chemical substructure search, a promising solution involves conducting searches outside the database. This approach leverages usage of chemistry SDKs like RDKit,[39] GPU clusters for scalability, modern task orchestration techniques, tools like CuPy, and efficient information retrieval from the database.

Medicinal Chemistry Use Cases

Cheminformatics is a very valuable tool for medicinal chemists, enabling efficient data management, compound screening, rational design, and optimisation of potential drug candidates. It accelerates drug discovery and development while minimising costs and experimental efforts.

In the quest to discover new drugs, not all compounds serve as suitable starting points due to unfavorable pharmacokinetic characteristics that can, for instance, hinder a drug's absorption, distribution, metabolism, and excretion (ADME). Consequently, these compounds are frequently omitted from datasets intended for virtual screening. Fig. 2 is an illustrative KNIME workflow designed for hit identification eliminating molecules that exhibit lower drug-like properties from a dataset. The workflow involves steps such as data acquisition from the CHEMBL database, data transformation, machine learning-driven virtual screening, and visualisation of results.

A significant proportion of drug candidates fail during clinical trials due to ADMET-related problems. ADMET predictive modelling can help reduce the failure rate by identifying unsuitable candidates (poor ADMET profiles) as early as possible, thus saving time and resources.[40,41] Moreover, medicinal chemists can use ADMET predictions not only to discard compounds but also to guide chemical modifications and optimise the properties of lead compounds. This optimisation can, for instance, enhance bioavailability, reduce toxicity, or improve potency. In this regard, QSAR models can be applied to predict various ADMET properties of chemical compounds.[42,43] Examples of applications of these models can include skin sensitisation and prediction of mutagenicity or hepatotoxicity, among others.[44–47]

Cheminformatics tools help medicinal chemists store, organise and retrieve large amounts of chemical and biological data. These tools streamline the management of information critical to drug discovery. Within them, LLMs have emerged recently as a hot topic for almost every scientific discipline, including medicinal chemistry. LLMs are instrumental in early-stage drug discovery, leveraging vast textual data like scientific literature and patents to identify potential therapeutic compounds.[23,24,48] LLMs have yielded results useful in solving interesting problems like molecular property prediction, molecule optimisation, compound discovery or target prediction.[48] Notably, in the field of novel target prediction, LLMs excel in named entity recognition tasks, offering more consistent and less noisy results aligned with desired outcomes. An example can be OncoRTT, a model aimed to predict oncology-related therapeutic targets using BERT embeddings and omics features.[49]

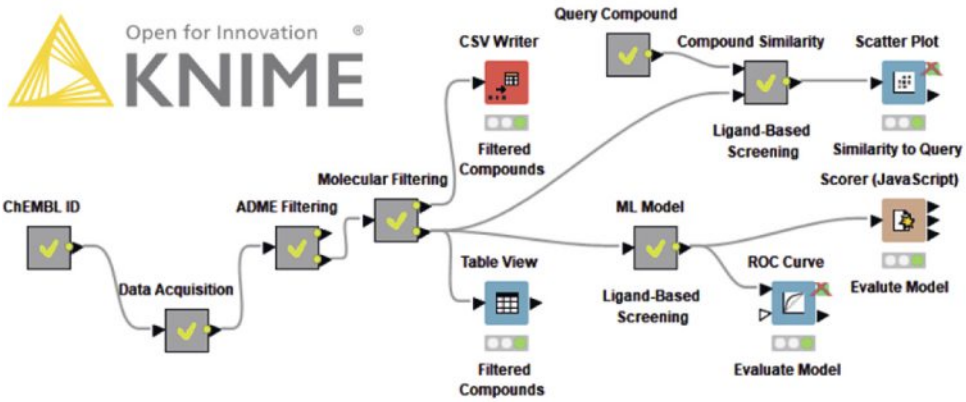


Fig. 2. An illustrative KNIME workflow for drug discovery. Image adapted from https://hub.knime.com/volkamerlab/spaces/Public/latest/TeachOpen CADD /TeachOpen CADD~xYhrR1mfFcGNxz7I

Chemical (substructure) search is crucial in medicinal chemistry and small molecule drug discovery, enabling precise molecule retrieval from databases based on specific chemical features. Although systems designed for this purpose have existed for many years, quick searches still present a substantial challenge.[50] This fact and the rapid growth of virtual libraries make it necessary to improve current search techniques.[51] A useful approach is the use of GPUs. Concretely balance GPU-based fingerprint screening with CPU parallelisation of the subgraph isomorphism screen, with horizontal scaling to distribute work across multiple machines .[52]

Conclusions

In this overview, we looked into some aspects of cheminformatics that play pivotal roles in modern drug discovery. The illustrative KNIME workflow showcased the significance of data-driven hit identification. ADMET prediction has emerged as a critical tool for early-stage candidate evaluation, potentially reducing costly clinical trial failures. QSAR models enhance ADMET predictions, aiding medicinal chemists in compound optimisation. Cheminformatics and AI/ML tools, including LLMs, facilitate data management and information retrieval, with LLMs proving instrumental in early-stage drug discovery and novel target prediction. Lastly, chemical substructure search techniques, augmented by GPUs, address the challenge of rapid library growth, ensuring precise molecule retrieval and advancing search capabilities in medicinal chemistry and drug discovery.

Received: October 5, 2023

[1] S. M. Kohlbacher, G. Ibis, C. Permann, S. Bryant, T. Langer, T. Seidel, *Mol. Inform*. **2023**, 42, e2200245, https://doi.org/10.1002/minf.202200245.

[2] T. Dekker, M. A. C. H. Janssen, C. Sutherland, R. W. M. Aben, H. W. Scheeren, D. Blanco-Ania, F. P. J. T. Rutjes, M. Wijtmans, I. J. P. de Esch, *ACS Med. Chem. Lett*. **2023**, *14*, 583, https://doi.org/10.1021/acsmedchemlett.2c00503.

[3] A. Afantitis, G. Melagraki, *Curr. Med. Chem*. **2020**, *27*, 6442, https://doi.org/10.2174/0929867327382010141028l4.

[4] S. Kralj, M. Jukič, U. Bren, *Int. J. Mol. Sci*. **2022**, *23*, 5727, https://doi.org/10.3390/ijms23105727.

[5] S. Chines, C. Ehrt, M. Potowski, F. Biesenkamp, L. Grützbach, S. Brunner, F. van den Broek, S. Bali, K. Ickstadt, A. Brunschweiger, *Chem. Sci.* **2022**, *13*, 11221, https://doi.org/10.1039/d2sc02474h.

[6] D. Sydow, M. Wichmann, J. Rodríguez-Guerra, D. Goldmann, G. Landrum, A. Volkamer, *J. Chem. Inf. Model*. **2019**, *59*, 4083, https://doi.org/10.1021/acs.jcim.9b00662.

[7] D. Palazzotti, M. Fiorelli, S. Sabatini, S. Massari, M. L. Barreca, A. Astolfi, *J. Chem. Inf. Model*. **2022**, *62*, 6309, https://doi.org/10.1021/acs.jcim.2c01199.

[8] G. Falcón-Cano, C. Molina, M. Á Cabrera-Pérez, *J. Chem. Inf. Model.* **2020**, *60*, 2660, https://doi.org/10.1021/acs.jcim.0c00019.

[9] O. Casanova-Alvarez, A. Morales-Helguera, M. Á. Cabrera-Pérez, R. Molina-Ruiz, C. Molina, *J. Chem. Inf. Model*. **2021**, *61*, 3213, https://doi.org/10.1021/acs.jcim.0c01439.

[10] D. Sydow, M. Wichmann, J. Rodríguez-Guerra, D. Goldmann, G. Landrum, A. Volkamer, *J. Chem. Inf. Model*. **2019**, *59*, 4083, https://doi.org/10.1021/acs.jcim.9b00662.

[11] J. C. Dearden, *Int. J. Quant. Struct.-Prop. Relatsh. IJQSPR* **2016**, *1*, 1, https://doi.org/10.4018/IJQSPR.2016010101.

[12] E. N. Muratov, J. Bajorath, R. P. Sheridan, I. V. Tetko, D. Filimonov, V. Poroikov, T. I. Oprea, I. I. Baskin, A. Varnek, A. Roitberg, O. Isayev, S. Curtalolo, D. Fourches, Y. Cohen, A. Aspuru-Guzik, D. A. Winkler, D. Agrafiotis, A. Cherkasov, A. Tropsha, *Chem. Soc. Rev.* **2020**, *49*, 3525, https://doi.org/10.1039/D0CS00098A.

[13] S. K. Niazi, Z. Mariam, *Int. J. Mol. Sci*. **2023**, *24*, 11488, https://doi.org/10.3390/ijms241411488.

[14] H. C. S. Chan, H. Shan, T. Dahoun, H. Vogel, S. Yuan, *Trends Pharmacol. Sci*. **2019**, *40*, 592, https://doi.org/10.1016/j.tips.2019.06.004.

[15] F. T. Musuamba, R. Bursi, E. Manolis, K. Karlsson, A. Kulesza, E. Courcelles, J. Boissel, R. Lesage, C. Crozatier, E. M. Voisin, C. F. Rousseau, T. Marchal, R. Alessandrello, L. Geris, *CPT Pharmacomet. Syst. Pharmacol*. **2020**, *9*, 195, https://doi.org/10.1002/psp4.12504.

[16] H. Hong, M. Chen, H. W. Ng, W. Tong, 'QSAR Models at the US FDA/NCTR', in 'In Silico Methods for Predicting Drug Toxicity. Methods in Molecular Biology', vol 1425, Ed. E. Benfenati, Humana Press, New York, NY, **2016**, pp 431–459. https://doi.org/10.1007/978-1-4939-3609-0_18.

[17] K. Mikkelsen, J. B. Sørli, M. Frederiksen, N. Hadrup, *Toxicology* **2023**, *495*, 153612, https://doi.org/10.1016/j.tox.2023.153612.

[18] P. Rodríguez-Belenguer, V. Mangas-Sanjuan, E. Soria-Olivas, M. Pastor, *J. Chem. Inf. Model.* **2023**, https://doi.org/10.1021/acs.jcim.3c00945.

[19] R. Lui, D. Guan, S. Matthews, *Chem. Res. Toxicol.* **2023**, *36*, 1248, https://doi.org/10.1021/acs.chemrestox.2c00385.

[20] K. M. Sullivan, J. R. Manuppello, C. E. Willett, *SAR QSAR Environ. Res*. **2014**, *25*, 357, https://doi.org/10.1080/1062936X.2014.907203.

[21] S. S. Sohail, *Ann. Biomed. Eng*. **2023**, https://doi.org/10.1007/s10439-023-03335-6.

[22] S. Liu, J. Wang, Y. Yang, C. Wang, L. Liu, H. Guo, C. Xiao, *arXiv* May 29, **2023**, https://doi.org/10.48550/arXiv.2305.18090.

[23] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, D. S. W. Ting, *Nat. Med*. **2023**, *29*, 1930, https://doi.org/10.1038/s41591-023-02448-8.

[24] A. Clyde, A. Ramanathan, R. Stevens, 'Large Language Models for Science', in 'Artificial Intelligence for Science'; WORLD SCIENTIFIC, **2022**; pp 643–669, https://doi.org/10.1142/9789811265679_0034.

[25] A. Egli, *Clin. Infect. Dis*. **2023**, ciad407, https://doi.org/10.1093/cid/ciad407.

[26] C. Merlot, D. Domine, C. Cleva, D. J. Church, *Drug Discov. Today* **2003**, *8*, 594, https://doi.org/10.1016/s1359-6446(03)02740-5.

[27] S. Kim, *Curr. Protoc*. **2021**, *1*, e217, https://doi.org/10.1002/cpz1.217.

[28] J. B. Baell, G. A. Holloway, *J. Med. Chem*. **2010**, *53*, 2719, https://doi.org/10.1021/jm901137j.

[29] D. C. I. Systems, 'Fingerprints - Screening and Similarity,' https://Www.Daylight.Com/Dayhtml/Doc/Theory/Theory.Finger.Html, accessed 22 May 2023.

[30] A. Dalke, *J. Cheminformatics* **2019**, *11*, 76, https://doi.org/10.1186/s13321-019-0398-8.

[31] M. Kratochvíl, J. Vondrášek, J. Galgonek, *J. Cheminformatics* **2018**, *10*, 27, https://doi.org/10.1186/s13321-018-0282-y.

[32] P. Liu, D. K. Agrafiotis, D. N. Rassokhin, E. Yang, *J. Chem. Inf. Model*. **2011**, *51*, 1807, https://doi.org/10.1021/ci200164g.

[33] Q. Liao, J. Wang, I. A. Watson, *J. Chem. Inf. Model*. **2011**, *51*, 1017, https://doi.org/10.1021/ci200061p.

[34] L. A. Baumgardner, A. K. Shanmugam, H. Lam, J. K. Eng, D. B. Martin, *J. Proteome Res*. **2011**, *10*, 2882, https://doi.org/10.1021/pr200074h.

[35] C. Ma, L. Wang, X.-Q. Xie, *J. Chem. Inf. Model*. **2011**, *51*, 1521, https://doi.org/10.1021/ci1004948.

[36] M. Maggioni, M. D. Santambrogio, J. Liang, *Procedia Comput. Sci.* **2011**, *4*, 2007, https://doi.org/10.1016/j.procs.2011.04.219.

[37] X. Yan, Q. Gu, F. Lu, J. Li, J. Xu, *Mol. Divers*. **2012**, *16*, 759, https://doi.org/10.1007/s11030-012-9403-0.

[38] R. Okuta, Y. Unno, D. Nishino, S. Hido, 'Crissman. CuPy : A NumPy-Compatible Library for NVIDIA GPU Calculations', **2017**.

[39] G. Landrum, 'RDKit', Q2, Https://Www.Rdkit.Org/, **2010**.

[40] L. L. G. Ferreira, A. D. Andricopulo, *Drug Discov. Today* **2019**, *24*, 1157, https://doi.org/10.1016/j.drudis.2019.03.015.

[41] S. Zhang, Z. Yan, Y. Huang, L. Liu, D. He, W. Wang, X. Fang, X. Zhang, F. Wang, H. Wu, H. Wang, *Bioinformatics* **2022**, *38*, 3444, https://doi.org/10.1093/bioinformatics/btac342.

[42] N. Pillai, A. Dasgupta, S. Sudsakorn, J. Fretland, P. D. Mavroudis, *Drug Discov. Today* **2022**, *27*, 2209, https://doi.org/10.1016/j.drudis.2022.03.017.

[43] V. Gallego, R. Naveiro, C. Roca, D. Ríos Insua, N. E. Campillo, *Mol. Divers*. **2021**, *25*, 1461, https://doi.org/10.1007/s11030-021-10266-8.

[44] Y. Wang, Q. Xiao, P. Chen, B. Wang, *Int. J. Mol. Sci*. **2019**, *20*, 4106, https://doi.org/10.3390/ijms20174106.

[45] G. S. Chayawan, D. Baderna, C. Toma, A. Y. Caballero Alfonso, A. Gamba, E. Benfenati, *Toxicology* **2022**, *468*, 153111, https://doi.org/10.1016/j.tox.2022.153111.

[46] X. Yang, Z. Zhang, Q. Li, Y. Cai, *Sci. Rep*. **2021**, *11*, 8030, https://doi.org/10.1038/s41598-021-87035-y.

[47] J. V. B. Borba, R. C. Braga, V. M. Alves, E. N. Muratov, N. Kleinstreuer, A. Tropsha, C. H. Andrade, *Chem. Res. Toxicol*. **2021**, *34*, 258, https://doi.org/10.1021/acs.chemrestox.0c00186.

[48] S. B. Brahmavar, A. Srinivasan, T. Dash, S. R. Krishnan, L. Vig, A. Roy, R. Aduri, *bioRxiv* September 17, **2023**, p 2023.09.14.557698. https://doi.org/10.1101/2023.09.14.557698.

[49] M. A. Thafar, S. Albaradei, M. Uludag, M. Alshahrani, T. Gojobori, M. Essack, X. Gao, *Front. Genet.* **2023**, *14*, https://doi.org/10.3389/fgene.2023.1139626.

[50] D. K. Agrafiotis, V. S. Lobanov, M. Shemanarev, D. N. Rassokhin, S. Izrailev, E. P. Jaeger, S. Alex, M. Farnum, *J. Chem. Inf. Model*. **2011**, *51*, 3113, https://doi.org/10.1021/ci200413e.

[51] A. Cherkasov, *Nat. Chem. Biol*. **2023**, *19*, 667, https://doi.org/10.1038/s41589-022-01233-x.

[52] A. J. Whitehouse, M. Sanchez-Martinez, S. M. Salehi, N. Kurbatova, E. Dean, 'An Open-Source Approach to GPU-Accelerated Substructure Search', manuscript submitted for publication, **2023**.