

Optimising DMTA through Academia-Industry Partnerships

Andrea Hunklinger *Universitat de Barcelona* handrean130@alumnes.ub.edu
Matthew Ball *Université de Strasbourg* mtball@unistra.fr
Bob van Schendel *Universiteit Leiden* bwvanschendel@gmail.com
Fabian Krüger *Helmholtz Munich* fabian.krueger@helmholtz-munich.de

Abstract

Industrial drug discovery pipelines, such as the common design–make–test–analyze (DMTA) cycle, are lengthy and resource intensive. While AI has begun to accelerate and optimize these processes, we argue that further progress could be achieved by incorporating academic innovations and fostering collaboration through joint research positions and challenge initiatives that evaluate models on industry-relevant datasets.

Keywords: Drug Discovery, Artificial intelligence, Machine learning

1 Introduction

The discovery and development of new medicines has always been shaped by advances in technology. Today, artificial intelligence (AI) is becoming one of the most transformative forces in this field, reshaping how researchers generate, evaluate, and optimize potential drug candidates [9]. The successful adoption of methods like AlphaFold, REINVENT, and AiZynthFinder shows how AI has begun to shape industrial practice, especially in the early discovery phases focused on identifying and refining new molecules [20, 25, 32]. This is the stage where small-molecule therapeutics, which continue to play a central role in pharmaceutical portfolios, are developed through the design–make–test–analyze (DMTA) cycle [29]. The DMTA cycle is a framework in which molecules are proposed, synthesized, tested for relevant properties, and analyzed to guide subsequent iterations [15].

However, despite these encouraging developments, the integration of academic AI research into the DMTA cycle remains limited. In practice, much of the innovation that reaches pharmaceutical companies originates not directly from universities but either through biotech startups that act as intermediaries or is developed in-house within pharmaceutical companies [33]. This gap can be explained by several barriers. The lack of standardized benchmarks makes it difficult to evaluate methods under conditions that reflect real-world use [44]. The Polaris consortium represents a promising step toward addressing these challenges by offering standardized benchmarks shaped by a board of academic and industrial leaders [44, 6]. However, such initiatives remain the exception rather than the norm, and their broader adoption will be crucial to closing the gap. Beyond benchmarking, academic implementations often lack the robustness,

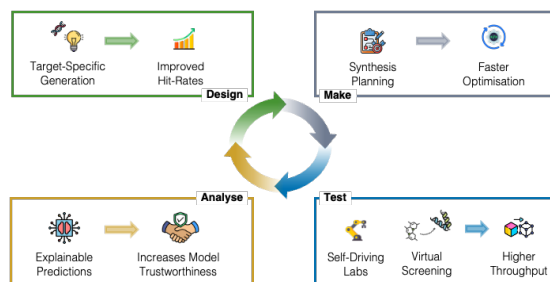


Figure 1: AI technologies have an impact in every aspect of the DMTA cycle an essential drug discovery pipeline.

documentation, and usability required in industrial settings, forcing companies to re-implement tools at considerable cost [10]. Insufficient interpretability adds another layer of difficulty, as domain-experts must be able to trust and understand AI outputs before integrating them into critical decision-making [17]. Finally, concerns about the confidentiality of costly, in-house datasets mean that valuable resources remain inaccessible to academics, leaving researchers to rely on smaller and less representative public collections when developing their tools [21]. In this paper, we first examine how AI is already influencing the individual stages of the DMTA cycle. We then outline solutions on how to better align academic and industrial priorities in order to foster the development of methods that are both scientifically novel and practically applicable.

2 AI advancements in the DMTA cycle

2.1 Design

In the drug discovery DMTA cycle, the design stage focuses on generating and optimizing novel molecular structures that are predicted to interact effectively with biological targets while maintaining favorable pharmacokinetic and safety profiles [15]. This stage integrates medicinal chemistry expertise with computational approaches to propose new compounds that balance potency, selectivity, and drug-like properties. In industry, AI tools are increasingly used to accelerate and guide this process. One example is the generative framework REINVENT, which employs recurrent neural networks and transformers to generate molecules [25]. Molecule generation can be steered toward desired properties using reinforcement learning, which requires scoring models to evaluate candidate molecules. These scoring models assign numerical values that reflect how well a molecule satisfies the target properties, and the generative model then learns to produce molecules that maximize these scores. Scoring models thus represent a core component of AI-assisted molecular design, with approaches spanning from traditional machine learning (ML) to modern deep learning techniques [28, 23]. A notable example in industrial use is QSARtuna, an automated platform that

builds robust predictive models for molecular property prediction and applies them to guide molecule scoring [27].

2.2 Make

Beyond the accelerated design of drug-like molecules, AI also holds significant potential in translating these designs into experimental reality. AI-driven retrosynthesis tools can predict synthetic routes, proposing combinations of reactants that yield desired products [14, 26, 40, 5]. These approaches not only streamline synthetic planning but also highlight alternative pathways that human bias might overlook. Yet, identifying reactants alone is insufficient. Reaction conditions, the physicochemical parameters that define the reaction environment, remain notoriously difficult to optimize [41, 7]. Here, ML approaches can provide valuable guidance, whether by suggesting promising starting points [13, 34, 43] or by prioritizing future experiments [18, 35, 36, 42] that accelerate the discovery of novel compounds. Bringing together these predictions of both reactants and conditions, ML tools can estimate the likelihood of reaction success, with the potential to increase yields while reducing wasted time, resources, and effort by prioritising reactions most likely to lead to positive outcomes.

2.3 Test

In the next step, synthesized molecules are tested in biological assays against target biomolecules that are linked to disease mechanisms, such as protein receptors, and AI can support target identification [30]. Increasingly, self-driving labs powered by AI are being developed to minimize human involvement during testing and to accelerate experimental cycles [16]. Complementing these advances, fully in-silico approaches such as high-throughput virtual screening [11] and digital twins [45] enable the exploration of molecular interactions before, or in some cases instead of, committing resources to physical experiments.

2.4 Analyze

In the analyze step of the DMTA cycle the experimental results are interpreted, structure-activity relationships (SAR) are extracted, and hypotheses are generated to guide the next design iteration. Traditionally, this involves medicinal chemists manually examining property trends to decide which scaffolds or analogs to prioritize and AI has already begun to support this stage [12]. Individual studies have shown that Explainable AI (XAI) [19, 17] can improve the quality of insights fed back into the design stage by uncovering interaction points of biomolecules and explaining why certain modifications could succeed or fail, for example by highlighting structural motifs [46] or using counterfactual analysis of molecule pairs [39]. However, the integration of XAI into industry pipelines is still far from common practice, even though it offers additional benefits such as increasing trust in predictions and supporting regulatory approval by providing transparent mechanisms of action and clear decision rationales [4].

3 Impact and collaboration

In AI research, industry tends to exceed academia in impact, while academia often leads in novelty [24]. In drug discovery, technological advances are more often generated in academic labs, with startups bridging the gap to bring them into big pharma [8]. This path is lengthy and typically involves financial debates, prolonged negotiations, and confidentiality agreements, which slow the translation of new technologies to industry standards. To shorten the path from idea to impact, integrated industry-academia doctoral programs, like the Horizon Europe-funded AIDD and AiChemist consortia [2, 1], are established, allowing students to conduct research within academic groups while also gaining experience in industry during their PhD. Projects are co-designed to meet both scientific novelty and translational relevance. The interdisciplinary combination of AI research with concrete drug discovery tasks helps the next generation of researchers see the bigger picture, enables direct discussion with domain experts in both fields, and can ultimately accelerate progress in drug discovery. Examples of how the AiChemist program has already contributed to bridging the gap between academia and industry include our recent work on addressing privacy concerns related to sharing industry-developed AI models [21], exploring the practical challenges of achieving impactful improvements in reaction condition prediction [7], demonstrating how explainability can enhance both model development and deployment [17], and developing a new foundation model for molecular scoring [23].

Beyond advancing these efforts, we call for more open yet industry-relevant datasets, particularly for reaction data [38, 31] and for benchmarking explainability methods [37, 17] as well as a more systematic engagement of the scientific community. A powerful precedent is the Critical Assessment of Structure Prediction (CASP) challenges [22], which provided a standardized and prestigious platform for benchmarking protein folding methods. By aligning public funding with industrial needs, CASP fostered rigorous comparison, incentivized innovation, and ultimately catalyzed the breakthrough model AlphaFold [20]. In a similar way, there is a need for a broader scope of challenges that address the different applications of AI in the DMTA cycle [3], including a retrosynthesis challenge, which to the best of our knowledge does not yet exist. We believe collaborative, challenge-driven research will accelerate innovation and transform both academic exploration and industry practice.

4 Conclusions

We showed how AI is already integrated in the design-make-test-analyze cycle, a common drug discovery framework and advocated for more collaborative and industry-relevant datasets, challenges and research positions.

References

- [1] Explainable ai for molecules - aichemist. <https://cordis.europa.eu/project/id/101120466>. Accessed: 2025-10-01.
- [2] Advanced machine learning for innovative drug discovery. <https://cordis.europa.eu/project/id/956832>. Accessed: 2025-10-01.
- [3] S. Ackloo, R. Al-awar, R. E. Amaro, C. H. Arrowsmith, H. Azevedo, R. A. Batey, Y. Bengio, U. A. K. Betz, C. G. Bologna, J. D. Chodera, W. D. Cornell, I. Dunham, G. F. Ecker, K. Edfeldt, A. M. Edwards, M. K. Gilson, C. R. Gordijo, G. Hessler, A. Hillisch, A. Hogner, J. J. Irwin, J. M. Jansen, D. Kuhn, A. R. Leach, A. A. Lee, U. Lessel, M. R. Morgan, J. Moul, I. Muegge, T. I. Oprea, B. G. Perry, P. Riley, S. A. L. Rousseaux, K. S. Saikatendu, V. Santhakumar, M. Schapira, C. Scholten, M. H. Todd, M. Vedadi, A. Volkamer, and T. M. Willson. CACHE (Critical Assessment of Computational Hit-finding Experiments): A public-private partnership benchmarking initiative to enable the development of computational methods for hit-finding. *Nat Rev Chem*, 6(4):287–295, Apr. 2022. ISSN 2397-3358. doi: 10.1038/s41570-022-00363-z.
- [4] C. S. Ajmal, S. Yerram, V. Abishek, V. P. M. Nizam, G. Aglave, J. D. Patnam, R. S. Raghuvanshi, and S. Srivastava. Innovative Approaches in Regulatory Affairs: Leveraging Artificial Intelligence and Machine Learning for Efficient Compliance and Decision-Making. *AAPS J*, 27(1):22, Jan. 2025. ISSN 1550-7416. doi: 10.1208/s12248-024-01006-5. URL <https://doi.org/10.1208/s12248-024-01006-5>.
- [5] T. Akhmetshin, D. Zankov, P. Gantzer, D. Babadeev, A. Pinigina, T. Madzhidov, and A. Varnek. Synplanner: An end-to-end tool for synthesis planning. *Journal of Chemical Information and Modeling*, 65(1): 15–21, Jan. 2025. ISSN 1549-9596. doi: 10.1021/acs.jcim.4c02004. URL <https://doi.org/10.1021/acs.jcim.4c02004>.
- [6] J. R. Ash, C. Wognum, R. Rodríguez-Pérez, M. Aldeghi, A. C. Cheng, D.-A. Clevert, O. Engkvist, C. Fang, D. J. Price, J. M. Hughes-Oliver, and et al. Practically significant method comparison protocols for machine learning in small molecule drug discovery. *ChemRxiv*, 2024. doi: 10.26434/chemrxiv-2024-6dbwv-v2.
- [7] M. Ball, D. Horvath, T. Kogej, M. Kabeshov, and A. Varnek. Predicting reaction conditions: a data-driven perspective. *Chemical Science*, page 10.1039.D5SC03045E, 2025. ISSN 2041-6520, 2041-6539. doi: 10.1039/D5SC03045E. URL <https://xlink.rsc.org/?DOI=D5SC03045E>.

- [8] L. Berghauser Pont, E. Fleming, R. Moss, L. Robke, K. Smietana, and S. Wurzer. Innovation sourcing in biopharma: Four practices to maximize success. <https://www.mckinsey.com/industries/life-sciences/our-insights/innovation-sourcing-in-biopharma-four-practices-to-maximize-success>, 2022.
- [9] H. Chen, O. Engkvist, Y. Wang, M. Olivecrona, and T. Blaschke. The rise of deep learning in drug discovery. *Drug discovery today*, 23(6):1241–1250, 2018.
- [10] H. Cho, J. Gray, and Y. Sun. Quality-aware academic research tool development. In *2012 19th Asia-Pacific Software Engineering Conference*, volume 2, pages 66–72. IEEE, 2012.
- [11] F. J. N. Ferreira and A. S. Carneiro. AI-Driven Drug Discovery: A Comprehensive Review. *ACS Omega*, 10(23):23889–23903, June 2025. ISSN 2470-1343, 2470-1343. doi: 10.1021/acsomega.5c00549.
- [12] A. Gangwal and A. Lavecchia. Artificial Intelligence in Natural Product Drug Discovery: Current Applications and Future Perspectives. *J. Med. Chem.*, 68(4):3948–3969, Feb. 2025. ISSN 0022-2623. doi: 10.1021/acs.jmedchem.4c01257. URL <https://doi.org/10.1021/acs.jmedchem.4c01257>. Publisher: American Chemical Society.
- [13] H. Gao, T. J. Struble, C. W. Coley, Y. Wang, W. H. Green, and K. F. Jensen. Using machine learning to predict suitable conditions for organic reactions. *ACS Central Science*, 4(11):1465–1476, Nov. 2018. ISSN 2374-7943, 2374-7951. doi: 10.1021/acscentsci.8b00357. URL <https://pubs.acs.org/doi/10.1021/acscentsci.8b00357>.
- [14] S. Genheden, A. Thakkar, V. Chadimová, J.-L. Reymond, O. Engkvist, and E. Bjerrum. Aizynthfinder: a fast, robust and flexible open-source software for retrosynthetic planning. *Journal of Cheminformatics*, 12(1): 70, Dec. 2020. ISSN 1758-2946. doi: 10.1186/s13321-020-00472-1. URL <https://jcheminf.biomedcentral.com/articles/10.1186/s13321-020-00472-1>.
- [15] G. M. Ghiandoni, E. Evertsson, D. J. Riley, C. Tyrchan, and P. C. Rathi. Augmenting dmta using predictive ai modelling at astrazeneca. *Drug discovery today*, 29(4):103945, 2024.
- [16] F. Häse, L. M. Roch, and A. Aspuru-Guzik. Next-Generation Experimentation with Self-Driving Laboratories. *Trends in Chemistry*, 1(3):282–291, June 2019. ISSN 2589-5974. doi: 10.1016/j.trechm.2019.02.007.
- [17] A. Hunklinger and N. Ferruz. Toward the explainability of protein language models for sequence design. *arXiv preprint arXiv:2506.19532*, 2025.

- [18] F. Häse, L. M. Roch, C. Kreisbeck, and A. Aspuru-Guzik. Phoenix: A bayesian optimizer for chemistry. *ACS Central Science*, 4(9): 1134–1145, 2018. ISSN 2374-7943. doi: 10.1021/acscentsci.8b00307. URL <https://doi.org/10.1021/acscentsci.8b00307>.
- [19] J. Jiménez-Luna, F. Grisoni, and G. Schneider. Drug discovery with explainable artificial intelligence. *Nat Mach Intell*, 2(10):573–584, Oct. 2020. ISSN 2522-5839. doi: 10.1038/s42256-020-00236-4. URL <https://www.nature.com/articles/s42256-020-00236-4>. Publisher: Nature Publishing Group.
- [20] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Židek, A. Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873): 583–589, 2021.
- [21] F. P. Krüger, J. Östman, L. Mervin, I. V. Tetko, and O. Engkvist. Publishing neural networks in drug discovery might compromise training data privacy. *Journal of Cheminformatics*, 17(1):38, 2025.
- [22] A. Kryshchak, T. Schwede, M. Topf, K. Fidelis, and J. Moult. Critical Assessment of Methods of Protein Structure Prediction (CASP) – Round XIV. *Proteins*, 89(12):1607–1617, Dec. 2021. ISSN 0887-3585. doi: 10.1002/prot.26237.
- [23] F. P. Krüger, N. Österbacka, M. Kabeshov, O. Engkvist, and I. Tetko. Molencoder: Towards optimal masked language modeling for molecules. *ChemRxiv*, 2025. doi: 10.26434/chemrxiv-2025-h4w9d.
- [24] L. Liang, H. Zhuang, J. Zou, and D. E. Acuna. The complementary contributions of academia and industry to AI research, Sept. 2024.
- [25] H. H. Loeffler, J. He, A. Tibo, J. P. Janet, A. Voronov, L. H. Mervin, and O. Engkvist. Reinvent 4: modern ai-driven generative molecule design. *Journal of Cheminformatics*, 16(1):20, 2024.
- [26] K. Maziarz, A. Tripp, G. Liu, M. Stanley, S. Xie, P. Gaiński, P. Seidl, and M. Segler. Re-evaluating retrosynthesis algorithms with syntheseus. *Faraday Discussions*, 256:568–586, 2025. ISSN 1359-6640, 1364-5498. doi: 10.1039/D4FD00093E. URL <http://arxiv.org/abs/2310.19796>. arXiv:2310.19796 [cs].
- [27] L. Mervin, A. Voronov, M. Kabeshov, and O. Engkvist. Qsartuna: an automated qsar modeling platform for molecular property prediction in drug design. *Journal of Chemical Information and Modeling*, 64(14):5365–5374, 2024.
- [28] 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, et al. Qsar without borders. *Chemical Society Reviews*, 49(11):3525–3564, 2020.

- [29] H. X. Ngo and S. Garneau-Tsodikova. What are the drugs of the future? *MedChemComm*, 9(5):757–758, 2018.
- [30] F. W. Pun, I. V. Ozerov, and A. Zhavoronkov. AI-powered therapeutic target discovery. *Trends in Pharmacological Sciences*, 44(9):561–572, Sept. 2023. ISSN 0165-6147. doi: 10.1016/j.tips.2023.06.010.
- [31] P. Raghavan, B. C. Haas, M. E. Ruos, J. Schleinitz, A. G. Doyle, S. E. Reisman, M. S. Sigman, and C. W. Coley. Dataset design for building models of chemical reactivity. *ACS Central Science*, 9(12):2196–2204, Dec. 2023. ISSN 2374-7943, 2374-7951. doi: 10.1021/acscentsci.3c01163. URL <https://pubs.acs.org/doi/10.1021/acscentsci.3c01163>.
- [32] L. Saigiridharan, A. K. Hassen, H. Lai, P. Torren-Peraire, O. Engkvist, and S. Genheden. Aizynthfinder 4.0: developments based on learnings from 3 years of industrial application. *Journal of cheminformatics*, 16(1):57, 2024.
- [33] A. Schuhmacher, M. Hinder, A. Dodel, O. Gassmann, and D. Hartl. Investigating the origins of recent pharmaceutical innovation. *Nat Rev Drug Discov*, 22(10):781–782, 2023.
- [34] P. Schwaller, A. C. Vaucher, T. Laino, and J.-L. Reymond. Prediction of chemical reaction yields using deep learning. *Machine Learning: Science and Technology*, 2(1):015016, Mar. 2021. ISSN 2632-2153. doi: 10.1088/2632-2153/abc81d. URL <https://dx.doi.org/10.1088/2632-2153/abc81d>.
- [35] B. J. Shields, J. Stevens, J. Li, M. Parasram, F. Damani, J. I. M. Alvarado, J. M. Janey, R. P. Adams, and A. G. Doyle. Bayesian reaction optimization as a tool for chemical synthesis. *Nature*, 590(7844):89–96, Feb. 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03213-y. URL <https://www.nature.com/articles/s41586-021-03213-y>.
- [36] J. W. Sin, S. L. Chau, R. P. Burwood, K. Püntener, R. Bigler, and P. Schwaller. Highly parallel optimisation of chemical reactions through automation and machine intelligence. *Nature Communications*, 16(1):6464, 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-61803-0. URL <https://www.nature.com/articles/s41467-025-61803-0>.
- [37] P. Song. Explainable AI in Drug Development - ML Journey, 2025. URL https://mljourney.com/explainable-ai-in-drug-development/?utm_source=chatgpt.com.
- [38] F. Strieth-Kalthoff, F. Sandfort, M. Kühnemund, F. R. Schäfer, H. Kuchen, and F. Glorius. Machine learning for chemical reactivity: The importance of failed experiments. *Angewandte Chemie International Edition*, 61(29):e202204647, 2022. ISSN 1521-3773. doi: 10.1002/anie.202204647. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/anie.202204647>.

- [39] H. Sturm, J. Teufel, K. A. Isfeld, P. Friederich, and R. L. Davis. Mitigating Molecular Aggregation in Drug Discovery With Predictive Insights From Explainable AI. *Angewandte Chemie*, 137(29):e202503259, July 2025. ISSN 1521-3757. doi: 10.1002/ange.202503259. URL <https://onlinelibrary.wiley.com/doi/10.1002/ange.202503259>. Publisher: John Wiley & Sons, Ltd.
- [40] Z. Tu, S. J. Choure, M. H. Fong, J. Roh, I. Levin, K. Yu, J. F. Joung, N. Morgan, S.-C. Li, X. Sun, H. Lin, M. Murnin, J. P. Liles, T. J. Struble, M. E. Fortunato, M. Liu, W. H. Green, K. F. Jensen, and C. W. Coley. Askcos: Open-source, data-driven synthesis planning. *Accounts of Chemical Research*, 58(11):1764–1775, 2025. ISSN 0001-4842. doi: 10.1021/acs.accounts.5c00155. URL <https://doi.org/10.1021/acs.accounts.5c00155>.
- [41] V. Voinarovska, M. Kabeshov, D. Dudenko, S. Genheden, and I. V. Tetko. When yield prediction does not yield prediction: An overview of the current challenges. *Journal of Chemical Information and Modeling*, 64(1): 42–56, Jan. 2024. ISSN 1549-9596. doi: 10.1021/acs.jcim.3c01524. URL <https://doi.org/10.1021/acs.jcim.3c01524>.
- [42] J. Y. Wang, J. M. Stevens, S. K. Kariofillis, M.-J. Tom, D. L. Golden, J. Li, J. E. Tabora, M. Parasram, B. J. Shields, D. N. Primer, B. Hao, D. Del Valle, S. DiSomma, A. Furman, G. G. Zipp, S. Melnikov, J. Paulson, and A. G. Doyle. Identifying general reaction conditions by bandit optimization. *Nature*, 626(8001):1025–1033, Feb. 2024. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-024-07021-y. URL <https://www.nature.com/articles/s41586-024-07021-y>.
- [43] X. Wang, C.-Y. Hsieh, X. Yin, J. Wang, Y. Li, Y. Deng, D. Jiang, Z. Wu, H. Du, H. Chen, Y. Li, H. Liu, Y. Wang, P. Luo, T. Hou, and X. Yao. Generic interpretable reaction condition predictions with open reaction condition datasets and unsupervised learning of reaction center. *Research*, 6: 0231, Jan. 2023. ISSN 2639-5274. doi: 10.34133/research.0231. URL <https://spj.science.org/doi/10.34133/research.0231>.
- [44] C. Wognum, J. R. Ash, M. Aldeghi, R. Rodríguez-Pérez, C. Fang, A. C. Cheng, D. J. Price, D.-A. Clevert, O. Engkvist, and W. P. Walters. A call for an industry-led initiative to critically assess machine learning for real-world drug discovery. *Nature Machine Intelligence*, 6(10):1120–1121, 2024.
- [45] J. Wu and V. H. Koelzer. Towards generative digital twins in biomedical research. *Computational and Structural Biotechnology Journal*, 23:3481–3488, Dec. 2024. ISSN 2001-0370. doi: 10.1016/j.csbj.2024.09.030.
- [46] L. Zhang, G. Domeniconi, C.-C. Yang, S.-g. Kang, R. Zhou, and G. Cong. CASTELO: clustered atom subtypes aided lead optimization—a combined

machine learning and molecular modeling method. *BMC Bioinformatics*, 22(1):338, June 2021. ISSN 1471-2105. doi: 10.1186/s12859-021-04214-4. URL <https://doi.org/10.1186/s12859-021-04214-4>.

5 Acknowledgements

The authors thank the Horizon Europe funding programme under the Marie Skłodowska-Curie Actions Doctoral Networks grant agreement *Explainable AI for Molecules - AiChemist*, no. 101120466 and the contributors to *Flaticon.com*.