frontiers | Frontiers in Chemistry

TYPE Review
PUBLISHED 29 November 2023
DOI 10.3389/fchem.2023.1292027

Cheminformatics and artificial intelligence for accelerating agrochemical discovery

Yannick Djoumbou-Feunang[1]*, Jeremy Wilmot[2], John Kinney[1], Pritam Chanda[1], Pulan Yu[2], Avery Sader[2], Max Sharifi[3], Scott Smith[1], Junjun Ou[2], Jie Hu[1], Elizabeth Shipp[4], Dirk Tomandl[5]† and Siva P. Kumpatla[6]†

[1]Corteva Agriscience, Farming Solutions and Digital, Indianapolis, IN, United States, [2]Corteva Agriscience, Crop Protection Discovery and Development, Indianapolis, IN, United States, [3]Corteva Agriscience, Regulatory and Stewardship, Indianapolis, IN, United States, [4]Corteva Agriscience UK Limited, Regulation Innovation Center, Abingdon, United Kingdom, [5]Atomwise, San Francisco, CA, United States, [6]Karyosoft Inc, Carmel, IN, United States

The global cost-benefit analysis of pesticide use during the last 30 years has been characterized by a significant increase during the period from 1990 to 2007 followed by a decline. This observation can be attributed to several factors including, but not limited to, pest resistance, lack of novelty with respect to modes of action or classes of chemistry, and regulatory action. Due to current and projected increases of the global population, it is evident that the demand for food, and consequently, the usage of pesticides to improve yields will increase. Addressing these challenges and needs while promoting new crop protection agents through an increasingly stringent regulatory landscape requires the development and integration of infrastructures for innovative, cost- and time-effective discovery and development of novel and sustainable molecules. Significant advances in artificial intelligence (AI) and cheminformatics over the last two decades have improved the decision-making power of research scientists in the discovery of bioactive molecules. AI- and cheminformatics-driven molecule discovery offers the opportunity of moving experiments from the greenhouse to a virtual environment where thousands to billions of molecules can be investigated at a rapid pace, providing unbiased hypothesis for lead generation, optimization, and effective suggestions for compound synthesis and testing. To date, this is illustrated to a far lesser extent in the publicly available agrochemical research literature compared to drug discovery. In this review, we provide an overview of the crop protection discovery pipeline and how traditional, cheminformatics, and AI technologies can help to address the needs and challenges of agrochemical discovery towards rapidly developing novel and more sustainable products.

KEYWORDS

agrochemicals, DMTA cycle, cheminformatics, artificial intelligence, machine learning, lead generation, lead optimization, sustainability

Check for updates

OPEN ACCESS

EDITED BY
Ziwen Wang,
Tianjin Normal University, China

REVIEWED BY
Xiuling Yu,
Linyi University, China
Yan Yang,
Taiyuan University of Technology, China

*CORRESPONDENCE
Yannick Djoumbou-Feunang,
✉ yannick.djoumboufeunang@corteva.com

†These authors have contributed equally to this work

RECEIVED 10 September 2023
ACCEPTED 09 November 2023
PUBLISHED 29 November 2023

CITATION
Djoumbou-Feunang Y, Wilmot J, Kinney J, Chanda P, Yu P, Sader A, Sharifi M, Smith S, Ou J, Hu J, Shipp E, Tomandl D and Kumpatla SP (2023), Cheminformatics and artificial intelligence for accelerating agrochemical discovery. *Front. Chem.* 11:1292027. doi: 10.3389/fchem.2023.1292027

1 Introduction

The development and application of computational tools has accelerated the pace of research and product development in diverse domains. Considering the impact computation has created, it was no exaggeration when it was stated that 'behind every great scientific finding in the modern age, from astronomy to zoology, there is a computer' (Perkel, 2021). Following decades of impressive growth, both pharmaceutical and agricultural industries have faced several challenges in bringing new products to market. Elevated costs (W. Zhang, 2018), increased regulatory requirements, and the need for differentiated products with novel modes of action (Sparks et al., 2018) are requiring unprecedented research and development investments to account for attrition in the pipeline and success in developing promising products (McDougall, 2016; Wouters et al., 2020).

Agrochemical product development, while having some parallels to pharmaceutical industry, has its own set of challenges that include addressing resistance development in pests (Siegwart et al., 2015; Hawkins et al., 2019), identifying sustainable chemistries (Whiteker, 2019), striking a balance with available genetically modified solutions, and competing with alternative and emerging technologies (Nishimoto, 2019). The data explosion and significant developments in data analytics that occurred throughout recent decades have provided means to address these challenges. In fact, this has further motivated the creation of newer, faster, and more scalable computational methods and tools for data generation, analysis, and hypothesis generation with the potential of decreasing the cost and time requirements for research and development of bioactive molecules.

Cheminformatics, also referred to as chemoinformatics, is the application of computer and informatics technologies to chemistry and has revolutionized the understanding of chemistry by improving the speed of development of novel products (Engel, 2006). It is a multidisciplinary field that employs tools and learnings from chemistry, biology, biochemistry, mathematics, statistics, and a host of other fields. Although the specific term cheminformatics has been in circulation for a little over two decades, its foundations can be traced back to the middle of last century when the conversion of chemical literature and mass spectra from print to electronic formats was initiated, database search systems were developed, and the widely used substructure matching algorithm came into existence (W. L. Chen, 2006; Ray and Kirsch, 1957). These seminal advances were followed by notable progress in subsequent decades that includes the development of chemical database retrieval systems and AI-based expert systems in the 1960s, creation of major chemical databases and development of binary fingerprints for substructure and similarity searches in the 1980s, introduction of new structural representation formats such as the Simplified Molecular-Input-Line-Entry System (SMILES) in the 1980s (Weininger, 1988) and the IUPAC International Chemical Identifier in the 2000s (Goodman et al., 2021), and the development of the first Machine Learning (ML) models to predict activity and physical properties in the 1990s (W. L. Chen, 2006). A key aspect of computational modeling that became a vital part of modern cheminformatics was the correlation of molecular structures with their biological function, which came to be known as Quantitative Structure-Activity Relationship (QSAR) (W. L. Chen, 2006). While linear models, partial least squares (PLS), and related traditional mathematical techniques enabled initial successes in QSAR modeling, the use of artificial neural networks (ANNs) for QSAR studies, first reported in 1990, was in prominence for several years until the onset of Random Forest and Support Vector Machine (SVM) approaches (Aoyama et al., 1990). The revolutionary successes of deep neural network (DNN) architectures in imaging (Baskin et al., 2016) brought about a renaissance of neural network architectures in a host of new and emerging tools for almost all steps in the discovery and development pipeline in both pharma and agricultural sectors since the mid-2010s. These include transformer-based ANNs for accurate conversion of chemical notations and the prediction of physicochemical properties, generative adversarial networks (GANs) for exploring chemical space as well as optimizing the functionality of known compounds, and deep learning (DL) and generative methods for intelligent navigation of small molecule space (Lo et al., 2019; Kell et al., 2020; Blanchard et al., 2021; Krasnov et al., 2021).

Several excellent reviews exist that describe the role of cheminformatics in drug discovery (Begam and Satheesh Kumar 2012; Lo et al., 2018; Lo et al., 2019; Chen and Kirchmair, 2020; Martinez-Mayorga et al., 2020). The motivation behind this article is to provide such a review for agrochemical discovery and development and to highlight how cheminformatics and AI tools are impacting the efficiency and speed of this process and in realizing the goals of developing sustainable and environmentally friendly products.

2 The crop protection discovery pipeline

The value of cheminformatics has been demonstrated in all stages of the pipeline used for the discovery of new crop protection active ingredients. In this review, we will refer to the crop protection pipeline (See Figure 1) as hit → active → lead generation → lead optimization as outlined by Loso *et al.* (Loso et al., 2017). Briefly, a 'hit' is defined as a compound that passes an activity threshold in the earliest tests (typically high-throughput screening) while an 'active' is a synthetically actionable compound with activity against target species that makes it a reasonable starting point for further exploration. A 'lead' molecule has an activity profile and novelty that warrant significant investment. Each stage of the pipeline has unique challenges that have the potential to be partially or entirely addressed with cheminformatics technology.

There are many approaches to begin the search for new hits to feed into the pipeline (Lamberth et al., 2013). Some examples include retrospective searches through databases for hints of activity from historical assays, known target site binders from the pharma literature, genome searches for cross-species target sites, pesticidal natural products (Lorsbach et al., 2019; Meyer et al., 2021; Sparks et al., 2021), novel fragments (Zhu et al., 2011), and competition inspired hit generation (Lahm et al., 2007). The success of any of these approaches hinges on the ability to quickly and accurately search across multiple chemical structure databases of millions of structures (company databases and literature) to billions or more in the case of virtual databases such as Enamine's REAL offerings (Grygorenko et al., 2020).

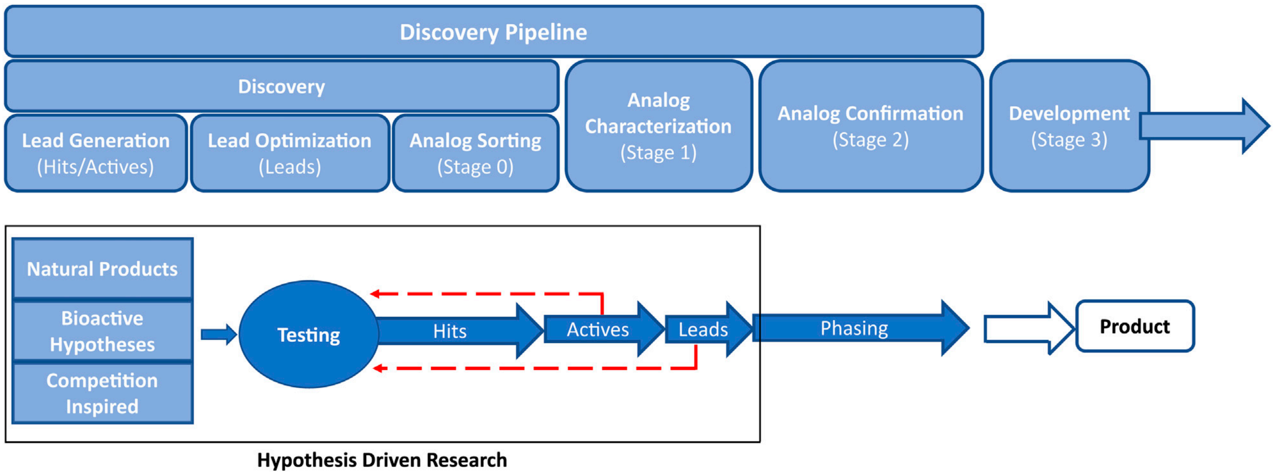


FIGURE 1
Overview of the crop protection discovery pipeline.

These search results should be easily narrowed down to those compounds that are predicted to not only have activity against target pests, but also have ag-like physicochemical properties (Zhang et al., 2018). It is also desirable to limit screening decks to diverse but relevant subsets, that ideally are small so they can be built upon in subsequent optimizations.

Once a hit has been identified, the advancement to an active typically involves broad exploration of nearby chemical space, Structure-Activity Relationship (SAR) exploration, and scaffold hops with testing to define the general areas of activity (e.g.: lepidopterans vs. coleopterans, broadleaf vs. grasses, ascomycetes vs. basidiomycetes, etc.). Compound sourcing at this stage is similar to hit generation such that compounds available within the company compound library, from commercial vendors, and from direct synthesis are all utilized. Broadly trained predictive models are generally still relevant, as the chemical space in which active generation takes place is still large and there are insufficient data points to create a meaningful active-specific predictive model.

At the active-to-lead stage of the pipeline, the SAR exploration becomes narrower, however, there is still significant probing of available chemical space. Typically, from this point forward all additional molecules are custom prepared, as commercial vendor chemical space is exhausted. It is at this point that there are usually enough molecules tested to generate area-specific predictive models. Physical property guidance becomes even more important. Initial work on target site identification begins, if it is not already known. The general pest spectrum is characterized, and a potential product concept is sketched out.

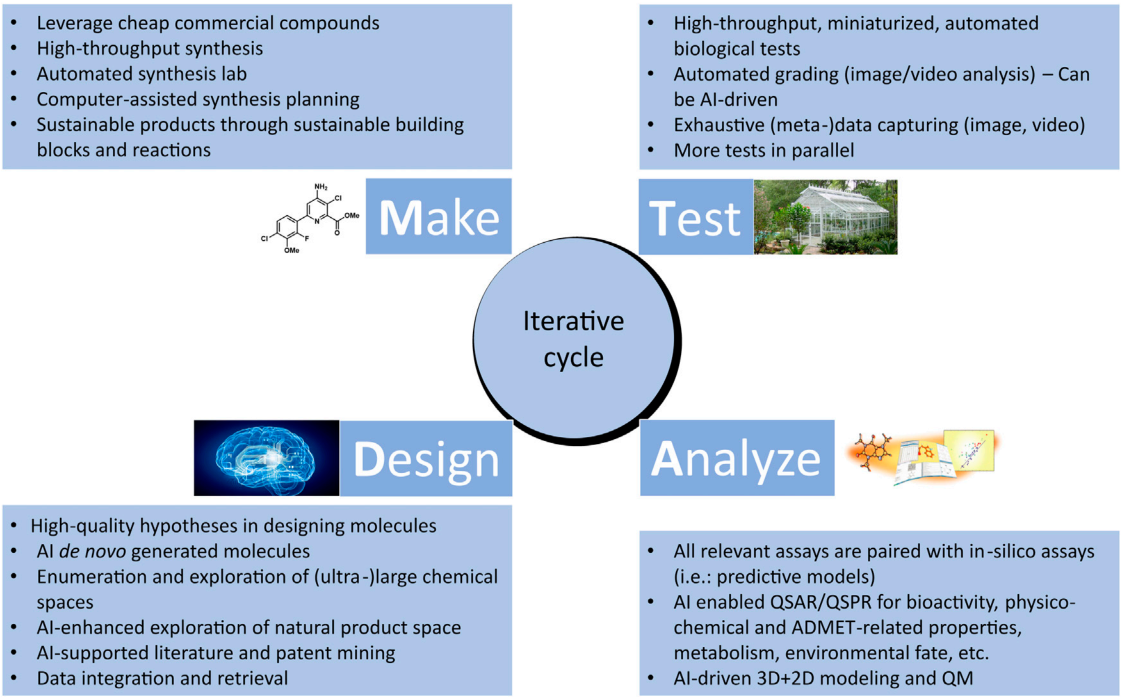
Once an active is advanced to lead, it is then considered a full-fledged project with significant resources made available. The SAR has been narrowed to the point that each portion of the molecule is deep-drilled due to the need for optimization of many parameters simultaneously (potency, selectivity, toxicity, environmental fate, cost of manufacture, etc.). Target site and mode of action confirmation become imperative, which can then further inform models.

3 Cheminformatics and AI for the design-make-test-analyze cycle

The Design-Make-Test-Analyze (DMTA) cycle is a central, iterative process consisting of interdependent steps that aim at efficiently designing, testing, and validating hypotheses, upon which data generated through experiments are analyzed, in order to discover new information that advances the discovery and optimization of leads (Plowright et al., 2012) (See Figure 2). Through several cycles, chemical hits are gradually optimized with respect to activity, selectivity, toxicity, and stability, into actives and eventually into more efficient lead molecules. Subsequently, selected lead molecules are further assessed using advanced models before development is initiated (Andersson et al., 2009). In this section, we will discuss how cheminformatics and AI are enhancing the pace and efficiency of DMTA cycle. In addition, we will highlight some of the challenges that need to be addressed to further accelerate the digitalization and automation as well as improve the success rate of DMTA processes in the discovery and optimization of sustainable agrochemicals.

3.1 Design

Molecular design closely ties to the design cycle of the widely accepted concept of iterative lead discovery. Its primary goal is to deliver new chemical entities with specified properties and potencies (Kuhn et al., 2016); however, those properties (e.g.: physicochemical, ADME-Tox properties) can vary greatly from the pharmaceutical industry (Tice, 2001). Molecular design includes two critical steps - generating a pool of candidates, and using molecular scoring strategies to select molecules from the collection for different disciplines in the agrochemical (agchem) industry, such as insect management, weed management (Gandy et al., 2015; Quareshy et al., 2018), and crop disease management, each of which has differing physiochemical property requirements (Avram et al., 2014;



- Leverage cheap commercial compounds
- High-throughput synthesis
- Automated synthesis lab
- Computer-assisted synthesis planning
- Sustainable products through sustainable building blocks and reactions

- High-throughput, miniaturized, automated biological tests
- Automated grading (image/video analysis) – Can be AI-driven
- Exhaustive (meta-)data capturing (image, video)
- More tests in parallel

- High-quality hypotheses in designing molecules
- AI *de novo* generated molecules
- Enumeration and exploration of (ultra-)large chemical spaces
- AI-enhanced exploration of natural product space
- AI-supported literature and patent mining
- Data integration and retrieval

- All relevant assays are paired with in-silico assays (i.e.: predictive models)
- AI enabled QSAR/QSPR for bioactivity, physico-chemical and ADMET-related properties, metabolism, environmental fate, etc.
- AI-driven 3D+2D modeling and QM

FIGURE 2
Overview of the iterative Design-Make-Test-Analyze (DMTA) cycle in agrochemical discovery. Cheminformatics and artificial intelligence play an increasingly significant role in each of the phases.

Zhang et al., 2018) (See Figure 3). At the hit generation stage, molecules with appropriate physical properties should be chosen since targets tend to gain mass during the active and lead generation process within the bounds of a given discipline. Since this stage also contains the largest possible chemical space, tools that accurately predict these properties quickly and display the results to a user alongside relevant activity-based metrics (e.g.: predicted assay activity, similarity to a query structure with known activity, relative location in known chemical space, etc.) in an intuitive and responsive manner are particularly important. Rapid searching and virtual screening of billions of compounds in modern commercial screening collections can be accomplished using tools such as fastROCS (OpenEye Scientific, 2023a), Ftrees (BioSolveIT, 2023a), and InfiniSee (BioSolveIT, 2023b) (See Table 1). The optimization of targets from hits to actives and leads should adhere as closely as possible to principles of green chemistry, namely low use rates, low ecological toxicity, minimal bioaccumulation, and thorough breakdown into benign fragments (Casida, 2012; Whiteker, 2019). Therefore, computational methods used at this stage such as QSAR and traditional ML and deep learning (DL) largely focus on molecule generation, docking, virtual screens, or molecular properties prediction, with molecule generation being a popular application of cheminformatics capability.

Historically, molecule generation included creating novel molecules from scratch and modifying structures based on scaffolds or fragments with demonstrated activity by bioisostere replacement, scaffold hopping/replacement, attaching functional groups, or linking multiple fragments. These functions have been implemented in popular tools such as DataWarrior (Sander et al., 2015), and KNIME (Berthold et al., 2008) (See Table 1). Early efforts in this area mostly prioritized the development of heuristic algorithms that focused on molecules predicted to be highly active and with desired properties (Sliwoski et al., 2013). The accumulation of data and advancement of ML methods are replacing these heuristics with evolving DL methods (Mater and Coote, 2019; Paul et al., 2021). Deep generative models (DGMs), leveraging the power of DNN architecture, are designed to learn latent representations of molecules even within a low-data setting and have a function to approximate the true distribution from which new compounds with desired molecular properties are sampled (Michael A. et al., 2021). Based on the architecture, these models can be categorized into (variational and adversarial) autoencoders (Blaschke et al., 2018; Richards and Groener, 2022), generative adversarial networks (GANs) (Méndez-Lucio et al., 2020; Abbasi et al., 2022), recurrent neural networks (RNN) with long short-term memory (LSTM) and gated recurrent unit (GRU) variants (Segler et al., 2018; He et al., 2021), and hybrid models combining deep generative models with reinforcement learning (RL) (Elton et al., 2019; Xue et al., 2019; Pereira et al., 2021) or autoencoders (Prykhodko et al., 2019). RL (Ståhl et al., 2019; Blaschke et al., 2020; Langevin et al., 2020) or conditional generative models (Kang and Cho, 2019; Sagar, 2020) speed up the process by generating only the molecules with desired properties or interesting scaffolds. Most DGMs take SMILES strings as inputs and then use a Variational Autoencoder (VAE) with Bayesian optimization in the latent space to generate molecules. Instead of generating molecules atom by atom, fragment-based language models can significantly reduce

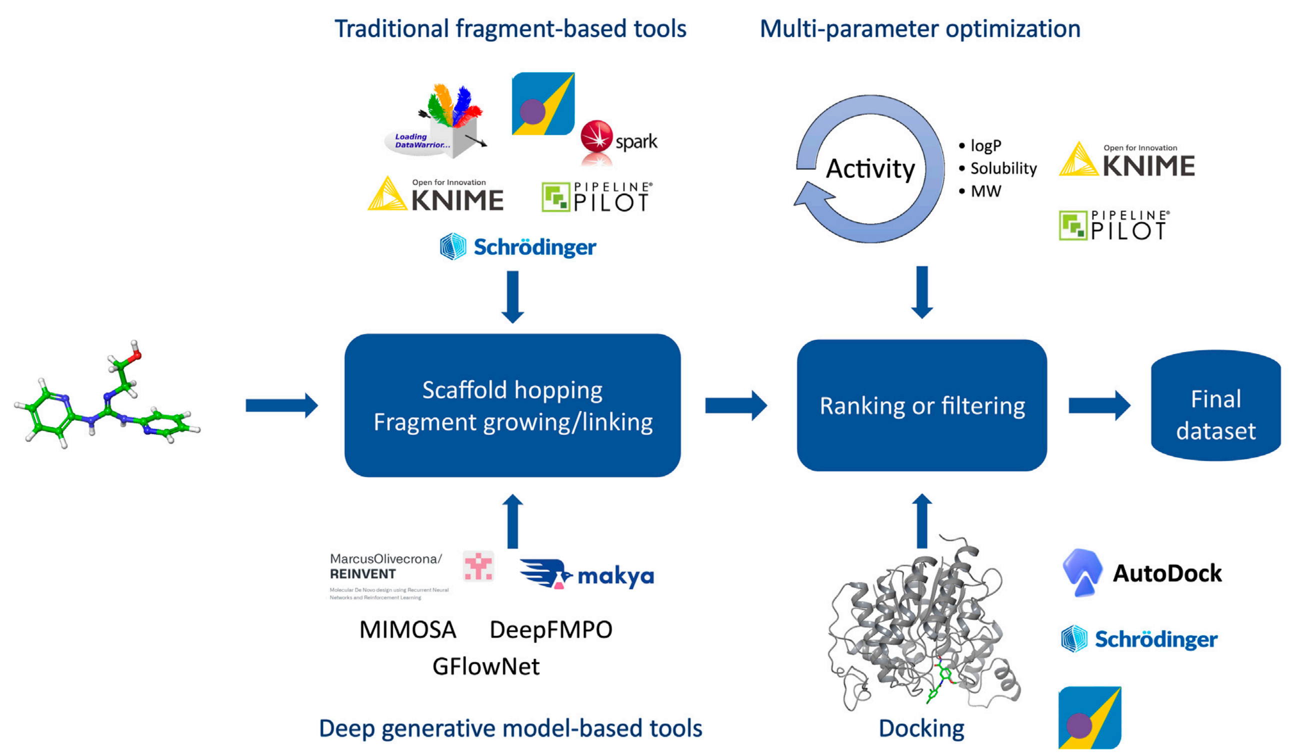


FIGURE 3
Overview of the molecular design workflow. Some examples of computational tools used in the design phase are provided.

TABLE 1 Examples of software tools used in molecular design.

Name	Description	References/Examples
RDKit	Cheminformatics	RDKit, (2023)
MOE	Molecular design, cheminformatics, QM, MM, QSAR, MD	ULC, C.C.G. (2023)
Data Warrior	Molecular design, cheminformatics	Sander et al. (2015)
KNIME	Data analysis, visualization, machine learning, deep learning, workflow	Berthold et al. (2008)
Pipeline Pilot	Data analysis, visualization, machine learning, deep learning, workflow	Dassault Systèmes SE, (2023)
Maestro	Molecular design, cheminformatics, QM, MM, QSAR, MD	Schrödinger, (2023b)
Spark	Molecular design	Cresset, (2023)
DeepFMPO	Deep learning, generative *de novo* design, reinforcement learning	Ståhl et al. (2019)
REINVENT	Deep learning, generative *de novo* design, multi-parametric optimization, reinforcement learning	Blaschke, Arús-Pous et al. (2020)
Makya	Deep learning, generative *de novo* design, multi-parametric optimization	Iktos, (2023a)
LillyMol	Cheminformatics	Eli Lilly & Co, (2019)
FastROCS™	Virtual Screening, Lead Hopping & Shape Clustering	OpenEye Scientific (2023b)
Ftrees	Virtual Screening	BioSolveIT, (2023a)
InfiniSee	Virtual Screening	BioSolveIT, (2023b)
Others	Deep learning, generative *de novo* design, multi-parametric optimization	Ståhl et al. (2019); Chuang et al. (2020), Khemchandani et al. (2020); Grechishnikova, (2021); Krishnan et al. (2022)

chemically invalid or duplicate compounds (Podda et al., 2020) as well as achieve comparable performance with fewer parameters and less training data (Chen et al., 2020). To further reduce the rate of chemically invalid generated molecules, Krenn *et al.* (Krenn et al., 2020) have introduced SELF-referencIng Embedded Strings (SELFIES), a more robust string-based representation of molecules. They demonstrated that VAEs and GANs using SELFIES generated only chemically valid molecules. Moreover, the generated sets of molecules were orders of magnitude more diverse when using SELFIES compared to SMILES strings. SELFIES were implemented in PASITHEA (Shen et al., 2021), a deep generative tool that applies "inceptionism" to propose new molecules with desired properties. Other popular approaches, such as Graph Neural Network (GNNs) have also been used in the generation of molecules (Shi et al., 2020; Mercado et al., 2021). GNNs, such as graph convolutional networks (GCNs) or message passing neural networks (MPNN), take graph-structured data as input and output a latent representation for the input graph. To improve performance, DGMs can be combined with each other (Méndez-Lucio et al., 2020) or other traditional ML algorithms (Blanchard et al., 2021). Metrics such as speed, coverage of chemical space, novelty, diversity, Kullback–Leibler (KL) divergence, and Fréchet ChemNet distance (Brown et al., 2019; Polykovskiy et al., 2020; Jie et al., 2021), among others, are widely used to evaluate their performance. The resulting molecules are then screened by agchem-related physiochemical property filters or pesticide-likeness scores (Zhang et al., 2018), predictive models trained by machine learning methods (Ray et al., 2017), or docking with protein models or homology models (Durrant et al., 2009; Chevillard et al., 2018; Hefke et al., 2020). In contrast to drug-likeness scores such as Quantitative Estimate of Drug likeness (QED) or drug-likeness models, the pesticide-likeness scores or models should include not only parameters related to bioactivity but also environmental effects such as volatilization, wash-off, photolysis, ecological toxicity, bioaccumulation, and soil metabolism for sustainability as well as the biodiversity of pests and usage conditions (Avram et al., 2014; Ouyang et al., 2021). The use of DGMs for molecule generation (Fromer and Coley, 2022) is promising but the challenges remain in how to improve diversity, novelty, and synthesizability (Benhenda, 2017; Gao and Coley, 2020), among other factors, within a multiparameter optimization framework. The advent of DGM has provided the opportunity to significantly improve the automated generation of molecules with desired properties and/or scaffolds using tools such as REINVENT and COMA, and thus, accelerate the advancement of molecules throughout the pipeline (Arús-Pous et al., 2020; Blaschke et al., 2020; Choi et al., 2023). The use of models trained on domain-relevant data, including the generative model, and associated scoring functions (e.g.: pharmacophore scoring) can lead to higher discovery rates of actionable and synthesizable compounds. For instance, by integrating pharmacophore features (e.g.: aromaticity, hydrophobicity) into the training of a REINVENT agent network, Yoshimori *et al.* (Yoshimori et al., 2021) were able synthesize nine DDR1 inhibitors with nanomolar potency. Moreover, recent works have introduced deep learning-based, protein-target driven *de novo* design approaches where the generative model takes protein specific information (e.g.: primary structure) to generate candidate ligands optimized towards various parameters (e.g.: high binding affinity, low toxicity) (Born et al., 2021; Zhang et al., 2023). While the methods used different generative algorithms and representations, they were able to propose ligands to relevant protein targets. Overall, the success of *de novo* generative design projects requires that goals be clearly defined by discovery teams, and priority be put on sampling strategy and efficiency, as illustrated in a benchmarking study by Gao *et al.* (Gao et al., 2022).

In contrast to the approach of creating novel molecules or modifying existing scaffolds and fragments, it is often desirable to screen libraries of compounds for novel hits. Molecular docking is a structure-based method that uses a search algorithm to generate ligand binding poses and a scoring function to quantitatively rank them. A common pitfall lies in the generation of false positives during ranking, either by failure to predict the correct pose of true ligands or by failure to discern between true ligands and decoys (Warren et al., 2006). Machine learning methods have shown promise in addressing these issues. For example, a support vector machine (SVM) regression analysis was used to score targets of AKT serine/threonine kinase 1, which led to the discovery of nanomolar inhibitors not attained with classical scoring functions (Zhan et al., 2014). Convolutional Neural Networks (CNN) algorithms have shown success at improving binding pose prediction by extracting features from protein-ligand complexes by analyzing their three-dimensional images (Ragoza et al., 2017). Incorporating machine learning into docking protocols is not without its share of issues. Neither protein-ligand structures nor sufficient data to develop a training set are guaranteed in agrochemical discovery. Moreover, the use of DL algorithms has been shown to fail compared to standard docking protocols in some cases (Gentile et al., 2020). As an alternative to developing novel scoring functions, Jimenéz-Luna *et al.* employed DL to rationally choose between standard docking protocols for a given protein-ligand pair with modest success (Jiménez-Luna et al., 2020a). Machine learning methods applied to docking and structure-based virtual screening are in a constant state of improvement, however, their utility in agrochemical discovery remains to be proven.

3.2 Make

The synthesis of chemical compounds is executed during the lead optimization and regulatory assessment phases, as well as once the final product is ready for commercialization. Hundreds of ideas and hypotheses can be generated in a relatively cost- and time-efficient manner during the design phase; however, the capability to convert these ideas into real and testable compounds remains one of the bottlenecks in the discovery process (Andersson et al., 2009). Because of the substantial number of assays to be run on target species as well as non-target species such as crops, much greater quantities of compounds are generally required compared to pharmaceutical research. It is thus critical that the synthesis of compounds is efficient, especially once an active has been optimized into a lead molecule. Generally, the synthesis of molecules involves: 1) selection of efficient synthetic routes for target compounds; 2) acquisition of building blocks and reagents; and 3) execution of the synthesis and purification phases. Cheminformatics and AI tools can be used in each of these phases to accelerate the process and reduce failures in the making of the novel molecules (Venkatasubramanian and Mann, 2022).

The decision on how to synthesize a novel compound is not only essential within the DMTA cycle, but also one of the most intellectually challenging. It is even more critical when scaling from gram to metric ton scale. At that stage, it cannot be emphasized enough that optimal manufacturing routes must be time- and cost-effective, efficient, safe, and environmentally sustainable. Designing such routes requires scientific intuition, as well as depth and breadth of knowledge in synthetic chemistry. Since the 1960s, synthetic chemists have increasingly relied on computers to suggest the most promising synthetic routes and help plan their execution (Corey and Wipke, 1969; Cook et al., 2012). Computer-Assisted Synthesis Planning (CASP) primarily involves retrosynthesis, condition recommendation, and forward reaction prediction (Struble et al., 2019). Retrosynthesis aims at generating feasible pathways starting from the target compounds and ending with building blocks that can be easily acquired. Traditionally, it has been achieved using a knowledge-based approach, which iteratively applies *a priori* expert knowledge (including reaction templates and constraints) encoded as rules or heuristics (Marcou et al., 2015; Szymkuć et al., 2016). One example of retrosynthetic pathway prediction tools is Synthia™ (formerly Chematica) (Szymkuć et al., 2016; Grzybowski et al., 2018), which was used in a 2018 study to design multistep synthetic routes to eight structurally diverse targets with medicinal relevance that were successfully executed in the laboratory (Klucznik et al., 2018).

Despite its interpretability, the knowledge-based approach can be costly due to maintenance and expansion of knowledgebases and is not very applicable to novel chemistries (Kayala, 2011; Reng et al., 2018). Recent advances in deep/transfer learning have enabled the development of innovative approaches that can automatically learn from available data, suggest routes, and predict outcomes (Gao et al., 2018; Dai et al., 2020; Schwaller et al., 2021). Additionally, several hybrid approaches have been developed that implement rule-based algorithms to suggest possible reactions which are then ranked and selected using machine learning algorithms (Zhang and Aires-de-Sousa, 2005; Segler and Waller, 2017; Nicolau et al., 2020). The prediction of reaction conditions is helpful for the prioritization of safe and efficient reactions. The outcome of such predictions usually includes chemicals (e.g.: catalysts, reagents, and solvents) and physical properties (e.g.: pressure, temperature). Examples of such prediction models include expert systems (Marcou et al., 2015) and machine learning-based models (Gao et al., 2018; Walker et al., 2019; Maser et al., 2021). Forward reaction prediction helps validate each reaction step and identify by-products to facilitate purification. Additionally, the prediction of yield provides a measure of how efficient a reaction step or route is. Recently, several tools have been proposed that address the prediction of both reaction outcomes and yields (Coley et al., 2017; Haywood et al., 2021; Martinez et al., 2021). As with several other applications of predictive modeling, high-quality, comprehensively annotated data can be very scarce and sparse. Moreover, collected reaction datasets tend to omit less successful and failed reactions. However, these would provide more insights into the mechanisms and latent variables that can best describe the feasibility of chemical reactions and thus, improve prediction accuracy. The recently released Open Reaction Database (ORD) is an effort to promote the sharing of proprietary pre-competitive reaction data in a comprehensive yet structured format (Kearnes et al., 2021). The ORD allows users to upload, search, visualize, and eventually submit chemical reaction data through programmatic access and web interfaces. By adhering to those standards, researchers can contribute to the amount and diversity of high-quality data available to carry out diverse CASP projects.

Currently, several CASP tools are freely or commercially available. Most of them provide a graphical user interface that enhances user experience, with the capabilities of visualizing and interacting with the proposed reactions and pathways (See Table 2). So far, it is not immediately apparent whether rule-based or machine learning-based approaches consistently provide superior results. However, a significant advantage of machine learning-based tools is that they can be rapidly improved and scaled efficiently as more data become available. Additionally, they tend to be easier to generalize over a larger chemical space. This is especially the case for template-free reaction prediction models. The evaluation of predicted pathways is generally carried out manually by groups of chemists. In order to perform pathway evaluations in a systematic, reliable, and consistent manner, automated and scalable frameworks need to be designed, which could be improved with the availability of additional data. Mo et al. (2021) introduced a data-driven approach to evaluate the relative strategic levels of retrosynthesis routes. The resulting tree-LSTM model, built on 238K routes from patents, could not only recognize but also cluster similar pathways. More recently, PaRoutes was introduced as a framework for comparing the quality and diversity of predicted synthetic pathways (Samuel and Bjerrum, 2022). The authors suggested metrics that could serve as methods for comparing predictions. It is envisaged that the significant efforts in this space will help to define metrics and workflows for the comparative evaluation and prioritization of predicted pathways, which would eventually point to cases where one prediction algorithm or tool performs comparatively better than its peers. Moreover, it could enhance the identification of innovative synthetic routes.

The synthesis of target compounds requires that starting materials and reagents are available. Electronic catalogs and databases containing structural information and metadata about building blocks and reagents can be linked to CASP platforms (e.g.: ASKCOS (Coley et al., 2019)). Such libraries can be maintained manually, or by service providers such as eMolecules (eMolecules, 2023), and Chemspace (Chemspace, 2023), the latter of which is the world's largest available compound catalog containing over 1.6 billion in-stock and make-on-demand building blocks. Another strategy used by chemists is the enumeration of chemical virtual libraries, where complex molecules are virtually created, using cheminformatics tools, by applying selected, easily reproducible chemical reaction schemes on available building blocks. The resulting compounds can then be filtered based on several criteria (max price, delivery time, properties, predicted activity), and then synthesized. An example of such virtual libraries is the Proximal Lilly collection, which provides chemists with a diverse collection of compounds that can be synthesized in-house (Nicolaou et al., 2016). Once obtained, the compounds can be tested in various experimental assays. Currently, cheminformatics platforms are being developed to assist chemists from planning to compound ordering to automation of synthesis (Schwaller et al., 2021; IBM, 2023).

TABLE 2 Examples of software tools and resources used in the Make phase of the DMTA cycle.

Name	Description	References/Examples
ASKCOS	Machine-learning based; single- and multistep retrosynthesis; condition recommendation; forward reaction outcome and evaluation	Struble et al. (2019)
Synthia™ (Chematica)	Manual and computer-aided retrosynthesis; User-defined rules and filters; direct link and metadata to commercially available and known building blocks	Grzybowski et al. (2018); Klucznik et al. (2018)
IBM-RXN	Molecular Transformer-based models for retrosynthesis, and forward reaction prediction	Schwaller et al. (2021)
ICSYNTH	Retrosynthesis analysis; machine-learned chemical rules; not limited to organic reactions	Bøgevig et al. (2015)
Spaya	Machine learning-based tool for full retrosynthetic analysis	Iktos, (2023b)
Reaxys	Predictive retrosynthesis with deep neural networks train on Reaxys data	Elsevier, (2023)
ChemFinder™ Ultra	Database management and structure search; retrieval of chemical and biological data (documents, structures, reactions, properties, etc.); property calculation	Aldrich, (2023)
CAS SciFinder^n	Search engine; retrieval of chemical data (structures, reactions, properties, etc.); linked to the CAS Content Collection™	CAS, (2023)
ReactionSage™	AI-based reaction pathway prediction; retrosynthesis, and forward reaction prediction	KEBOTIX, (2023)

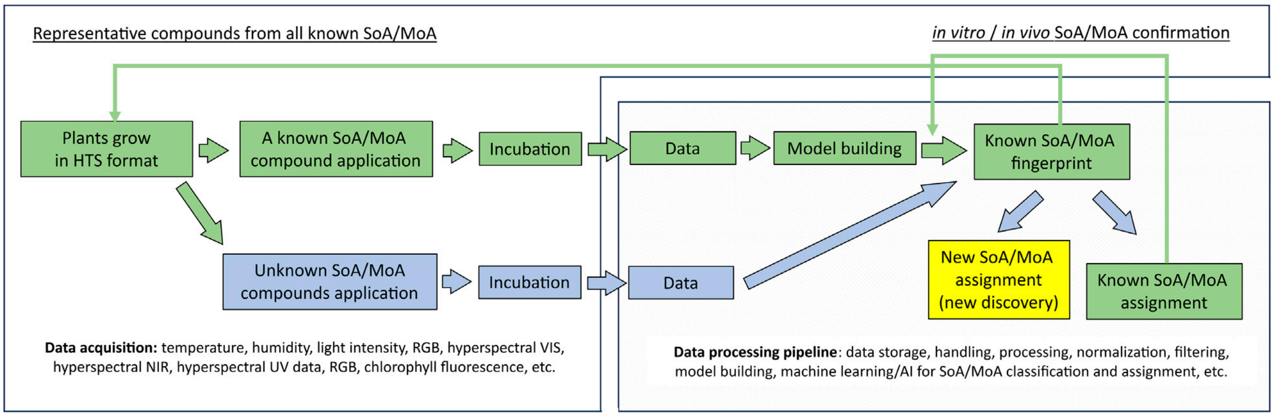

Representative compounds from all known SoA/MoA
in vitro / in vivo SoA/MoA confirmation
Plants grow in HTS format
A known SoA/MoA compound application
Incubation
Data
Model building
Known SoA/MoA fingerprint
Unknown SoA/MoA compounds application
Incubation
Data
New SoA/MoA assignment (new discovery)
Known SoA/MoA assignment
Data acquisition: temperature, humidity, light intensity, RGB, hyperspectral VIS, hyperspectral NIR, hyperspectral UV data, RGB, chlorophyll fluorescence, etc.
Data processing pipeline: data storage, handling, processing, normalization, filtering, model building, machine learning/AI for SoA/MoA classification and assignment, etc.

FIGURE 4
Workflow to identify herbicide SoA/MoA for screening compounds using high-throughput (HTS) assay, data rich imaging technologies, and Machine Learning/AI. Green colored workflow: compounds with known site of action (SoA)/mode of action (MoA) are used to build and validate the model to generate SoA/MoA fingerprints; Blue colored workflow is for new compounds with unknown SoA/MoA handling. Arrow indicates the workflow direction. After the model building and validation are completed with representative compounds from each known SoA/MoA, when new compounds with unknown SoA/MoA are tested and analyzed using this workflow, the compounds can either be assigned as new or unknown SoA/MoA, which warrants further efforts to use the DMTA cycle to discover novel herbicide with new mode of action; or be assigned as known SoA/MoA, after these assignment are confirmed by *in vitro* and/or *in vivo*, these data can be used for further model refinement.

3.3 Test

In agrochemical discovery, substantial amounts of data from a plethora of assays run on target and non-target species are obtained for further analysis. It is thus particularly important to enhance testing capabilities as well as data collection. The synergistic interaction between cheminformatics and biological tests in agrochemical discovery has not been commonly discussed, yet it has been essential. Data from medium and high throughput assays, such as *in vitro* enzyme assays, cell-based assays, metabolomics/genomics assays, and *in vivo* whole organism plate-based assays have been used as input for cheminformatics tools. The improvement of computational power, data storage capacity, data analysis capability, and the integration of these three in low-cost cloud computing services (e.g.: Amazon Web Services™ cloud computing platform (Amazon Web Services, 2023)) for cheminformatics tools have enabled new generation of data collection with existing assays, especially in whole organism level assays. For example, data rich hyperspectral/multispectral imaging (Thomas et al., 2018; Paulus and Mahlein, 2020; Klie et al., 2022) and video-based chemobehavioral phenotyping (Henry and Wlodkowic, 2020) provides enriched data to further enhance cheminformatics development such as building more sophisticated models and enabling the training of AI predictive models (Ozdemir and Polat, 2020). On the other hand, new and powerful predictive models can further support the automation in non-destructive data collection, mode-of-action prediction, and so forth, which can further increase the test throughput potential and derive

TABLE 3 Examples of relevant resources, cheminformatics software, and machine/deep learning tools utilized in the analyze phase of the DMTA cycle in agrochemical discovery. Abbreviations: Support vector regression (SVR), Liquid Chromatography – Mass Spectrometry (LC-MS), Graph Neural Network (GNN), Retention Time (RT), Deep Graph Learning (DGL), Natural Products (NP).

Name	Description	References/Examples
Structural classifications tools		
LeadScope	SAR analysis and visualization tool, with a focus on toxicological data	Roberts et al. (2000)
DataWarrior	General purpose SAR tool	Sander et al. (2015)
Pipeline Pilot	Data pipeline tool; capabilities for various *ad hoc* analyses	Dassault Systèmes SE, (2023)
KNIME	Data pipeline tool; capabilities for various *ad hoc* analyses	Berthold et al. (2008)
OpenEye Toolkit	Molecular toolkit; Low-level API tools custom structure analyses	OpenEye Scientific (2023a)
RDKit	Molecular toolkit; Low-level API tools custom structure analyses	RDKit, (2023)
Structure-Activity-Relationship Visualizations		
DataWarrior	General purpose SAR tool	Sander et al. (2015)
StarDrop™	Includes multi-parameter optimization and SAR tools	Optibrium, (2023)
TIBCO Spotfire®	Lead Discovery collection adds extensive cheminformatics capabilities, including predictive analytics	TIBCO, (2023)
Cheminformatics and AI-enabled Metabolomics		
Peakonly	DL-based model for LC-MS peak detection and integration	Melnikov et al. (2020)
ChromAlignNet	DL-based tool for peak-alignment of GC-MS data	Li and Wang (2019)
CFM-ID	Hybrid (AI-, rule-based) tool for LC-MS spectra prediction, peak annotation, and metabolite identification	Wang et al. (2021a), Djoumbou-Feunang et al. (2019b)
3D-MolMS	Tandem MS Spectra prediction	Hong et al. (2023)
MassFormer	Tandem MS Spectra prediction	Young et al. (2023)
SIRIUS	Computational platform for tandem MS data-based analysis of metabolites; provides molecule search, and class prediction capabilities	Dührkop et al. (2019)
MESSAR	Automated tool for metabolite substructure recommendation from tandem mass spectra	Liu et al. (2020)
ClassyFire	Structural classification of small and large molecules	Djoumbou-Feunang (2016)
NP-Classifier	DNN-based structural classification of natural products	Kim et al. (2020)
BioTransformer	Hybrid, comprehensive tool for metabolite prediction and identification in humans, gut microbiota, and environmental microbiota	Djoumbou-Feunang et al. (2019a)
ADMET Predictor	Machine learning-based prediction of human metabolites	SimulationsPlus, (2023)
QSAR Toolbox	AI-based prediction of chemical products from abiotic transformations and metabolism (microbial, rat liver S9, skin)	QSAR Toolbox, (2023)
OASIS Times	AI-based prediction of chemical products from abiotic transformations as well as *in vitro* (gut, lung, rat liver S9) and *in vivo* (rat) metabolites	OASIS, (2021)
GLORYx	Machine learning-based prediction of human metabolites	de Bruyn Kops et al. (2021)
MetaTrans	Deep-learning-based, rule-free tool for prediction of small molecule metabolites in humans	Litsa et al. (2020)
Retip	ML-based retention time prediction	Bonini et al. (2020)
GNN-RT	GNN-based liquid chromatography retention time prediction	Yang Q. et al. (2021)
DeepCCS	Deep Learning tool for the prediction of collision cross-section values	Plante (2019)
Spectral Databases	Spectral databases commonly used for metabolite identification	(NIST, (2023); Guijas et al. (2018); Wang et al. (2021b); Mehta, (2020); Wishart et al. (2018); Wang et al. (2016)
Programming libraries and cheminformatics tools for predictive modeling		
Scikit-learn	General Python-based programming library	Pedregosa et al. (2011)
PyTorch	General Python-based programming library for deep learning, including explainable DL	PyTorch, (2023)
Tensorflow	General Python-based programming library for deep learning, including explainable DL	Abadi et al. (2015)
DeepChem	Python-based programming library for deep chemistry	Ramsundar et al. (2019)
Chemprop	Python programming package implementing Message Parsing Neural Networks (MPNN) for the prediction of molecular properties as well as chemical reactions; provides uncertainty quantification capabilities	Yang K. et al. (2019)
DGL-Lifesci	Python programming library for graph neural network-based learning for chemistry and biology	Li Y. et al. (2021)
MolPMoFit	Transfer learning approach (and model) for molecular property (QSAR/QSPR) prediction	Li and Fourches (2020)
Chemformer		Irwin, Dimitriadis et al. (2022)

(Continued on following page)

TABLE 3 (*Continued*) Examples of relevant resources, cheminformatics software, and machine/deep learning tools utilized in the analyze phase of the DMTA cycle in agrochemical discovery. Abbreviations: Support vector regression (SVR), Liquid Chromatography – Mass Spectrometry (LC-MS), Graph Neural Network (GNN), Retention Time (RT), Deep Graph Learning (DGL), Natural Products (NP).

Name	Description	References/Examples
	A Python library for molecular optimization, property prediction, reaction and retrosynthetic prediction	
DESlib	A Python library for dynamic classifier and ensemble selection	Cruz et al. (2020)
SHAP	A Python programing library for Shapley Additive exPlanations	Lundberg and Lee, (2017); Rodríguez-Pérez and Bajorath, (2021)
Alibi Explain	Implements several algorithms for inspecting and explaining machine learning models	Klaise et al. (2021)
GNN-Explainer	A Python library for the explanation of GNN-based predictions	Ying et al. (2019)
CIME	A library for web-based exploratory analysis of chemical model explanations	Humer et al. (2022)

extra value from each test, especially for *in vivo* tests (Mishra et al., 2017; Klie et al., 2022). For example, Klie et al. recently disclosed a workflow that can be used to predict herbicidal site of action and/or mode of action of novel chemistries using classical machine learning and/or AI (See Figure 4).

3.4 Analyze

The Analyze phase of the DMTA cycle is a continual process during the entirety of a project's timeline. The analyses are used in the Design phase to help determine what compounds to synthesize, in the Test phase to help evaluate and plan additional tests, and in an oversight role to determine whether to continue a project or not. Cheminformatics plays a significant role in helping the researcher answer key questions in all these phases. The following subsections describe four critical aspects of the analyze phase and illustrate their overall impact in the DMTA cycle, while providing a brief description of several tools that enhance the analyses (See Table 3). These include structure classifications, predictive modeling, SAR visualizations, and metabolomics.

3.4.1 Structural classifications

Structure classification tools allow the partitioning of compounds into groups that can be used in a variety of visual and statistical tools to highlight areas of particular interest. This capability is at the heart of commercial software tools such as LeadScope (Roberts et al., 2000) and open-source software tools such as DataWarrior (Sander et al., 2015).

Some of the most common classification approaches include the identification of ring systems and frameworks, and clustering based on structural fingerprints. The compounds in each cluster can be further classified by determining the "Maximum Common Substructure", i.e., the largest substructure that is found in each compound in the cluster. These classifications tend to be "unsupervised", driven solely by the nature of the structures on hand, and thus are easy to accomplish with the use of modern cheminformatics toolkits (See Table 3).

A semi-manual approach, R-Group Decomposition (RGD), involves the identification of specific core structures in a molecule set, then determining the substituents that are attached in specific locations on the core (Agrafiotis et al., 2011; Naveja and Vogt, 2021). This technique usually involves an iterative analysis in order to describe as many compounds in the project as possible. In the end, the researcher is left with a set of molecular partitions and descriptors that generally align with the synthetic sources of the molecules. One important use of the resulting RGD table is to track the specific compounds that have been made and which of these have been tested in which assays. It also helps to quickly spot and track gaps in the already designed libraries, which is particularly important given that most researchers work on many projects simultaneously and over many years.

3.4.2 Predictive modeling

One of the most important activities in the DMTA cycle, and in the analyze phase, involves the study of quantitative relationships between molecular structures and various endpoints, including but not limited to biological activity (QSAR), physicochemical properties (Quantitative Structure Property Relationships; QSPR), and biodegradation (Quantitative Structure Biodegradation Relationship; QSBR). Leveraging diverse datasets generated throughout the test phase, among other sources, machine learning, and especially predictive modeling, have gradually matured over the last few decades into an essential component of discovery and regulatory processes for pharmaceuticals and agrochemicals (Naik et al., 2009). They both deliver mathematically sound, reliable, cost- and time-efficient, and more accessible "*in silico* assays" that can predict relevant endpoints, and be automatically improved with increasing data, in an adaptive environment (J. C. Dearden, 2016; Yang K. et al., 2019; Shen and Nicolaou, 2019). Most innovation in this space has occurred in pharmaceutical research, and agrochemical research has followed suit. Unfortunately, as illustrated by a relatively low number of related publications (See Figure 5), the adoption of an AI-driven discovery paradigm is not nearly as rapid in the relatively smaller space of agrochemical discovery, leaving untapped an increasing reservoir of innovative opportunities to accelerate research and development. Yet, urgent needs for novel and safer crop protection agents, along with the resolutions of regulatory agencies (e.g.: U.S. EPA (United States Environmental Protection Agency, 2023), EFSA (European Food and Safety Agency, 2023)) to aggressively reduce animal testing (Barlow et al., 2009), highlight the need for predictive tools that provide different lines of evidence and support the use of New Approach Methodologies (NAMs) in various scientific tasks, such as chemical risk assessment (U.S. EPA, 2021; Kavlock et al., 2018).

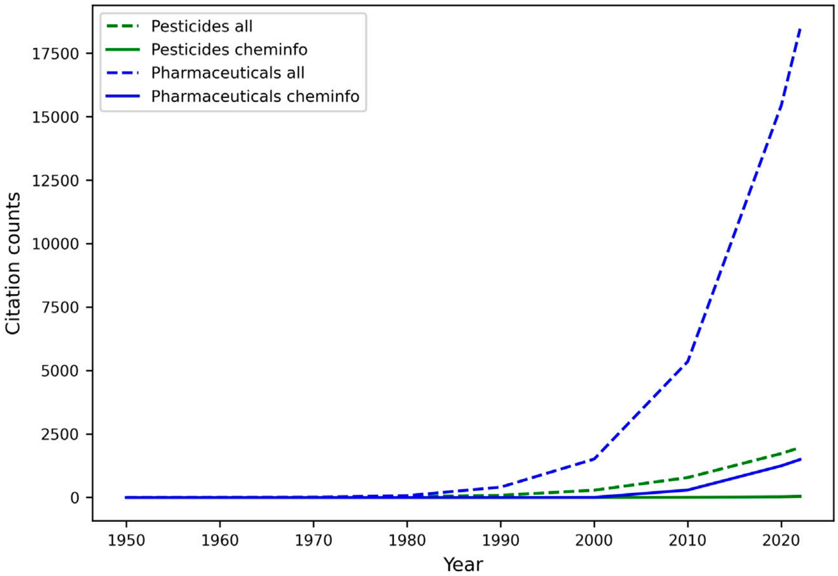


FIGURE 5
Comparative analysis of machine learning and cheminformatics-related publication counts for pesticide and pharmaceutical discovery. Publications (articles, reviews, reports, and dissertations only) were retrieved in May 2022 from the web of science literature database, upon mining the title and abstracts for specific keywords. The search for cheminformatics papers was limited to publications containing either of the keywords in the title and/or abstract: "cheminformatics" (or a derivative), "QSAR", "QSPR", "QSBR". The list was expanded to include papers from selected scientific journals whose title contains either "cheminformatics" (e.g.: journal of cheminformatics) or "QSAR". ML papers included various (groups of) keywords related to ML tasks, and metrics. The keyword "all" refers to publications including machine learning- and cheminformatics-related terms. Keywords describing the molecule class included drug, pharmaceutical, agrochemical, pesticide, insecticide, herbicide, fungicide, nematicide, and their derivatives.

Ideally, a crop protection agent must display optimal properties with respect to efficacy, metabolic stability, activity spectrum, uniqueness of its mode of action, and sustainability, among other parameters. High-Throughput Screening (HTS) is a key component of the discovery pipeline that provides scientists with diverse types of data exploitable for decision-making. In contrast to drug discovery, however, most screening assays in "agchem" discovery are phenotypic and run against whole organisms, especially in preliminary stages, when the target site is unknown. An advantage of such assays is that they incorporate the cellular complexity of biology as they highlight molecules that are both intrinsically active and bioavailable (Bender and Cortés-Ciriano, 2021a). However, they only provide little insight on the mode of action, which may be species-dependent (FRAC, 2023; HRAC, 2023; IRAC, 2023), and very little on the fundamental mechanisms that make a compound more or less active or completely inactive (e.g.: ADME properties), thus posing challenges for subsequent optimization. Bottlenecks resulting from these limitations include, among others, poor translation of activity against selected targets from the greenhouse to the field, limited systemic activity, and discrepancies between *in vivo* and *in vitro* activities (Zhang et al., 2018) Since activity, ADME-Tox, environmental fate, and other relevant mechanisms are influenced by the molecule's physicochemical properties (e.g.: lipophilicity (LogD), water solubility (WS), UV stability, pKa), QSPR-based tools that rapidly and accurately predict such properties are indispensable for rapid exploration of the immense chemical space, efficient selection of promising candidates, and decision making. For instance, physicochemical property prediction tools support the estimation of ag-likeness, typically defined with various degrees of specificity. Herbicides and sap-feeding insecticides, which need to be transported through the plant's xylem typically display high WS and low LogD, while chewing insecticides display high LogD and low WS that limit their uptake by, and mobility within plant leaves (Zhang et al., 2018). Commercial fungicides, however, occupy a relatively broader range with respect to those properties. Several guidelines have been proposed by Tice (2001), Zhang et al. (2018), and others (Hao et al., 2011; Avram et al., 2014) to assess ag-likeness based on various molecular (e.g.: constitutional, physicochemical) descriptors. Examples of most commonly used open-source cheminformatics packages for the computation of such molecular descriptors include, among others, RDKit (RDKit, 2023), Mordred (Moriwaki et al., 2018), the PaDEL-Descriptor software (Yap, 2011), and the Chemistry Development Kit (CDK) (Willighagen et al., 2017). While these tools typically provide a diverse set of descriptors, they often either lack certain physicochemical properties used in ag-likeness rules (e.g.: UV-stability, pKa) or provide different implementations compared to those used in the rules (e.g.: XLogP vs. ALogP vs. MLogP). Freely available and interactive web platforms such as InsectiPAD (Chen-Yang et al., 2019), FungiPAD (M.-y. Wang et al., 2019), and HerbiPAD (Huang, 2020) provide capabilities to explore pre-computed physicochemical properties and evaluate pesticide-likeness of chemicals. However, these are limited to only a few hundred chemicals and cannot be easily integrated into *in silico* workflows. OPERA (Mansouri et al., 2018) is an open-source/open data, standalone, limited collection of QSPR/QSAR models that predict several toxicity (e.g.: androgen receptor activity) and

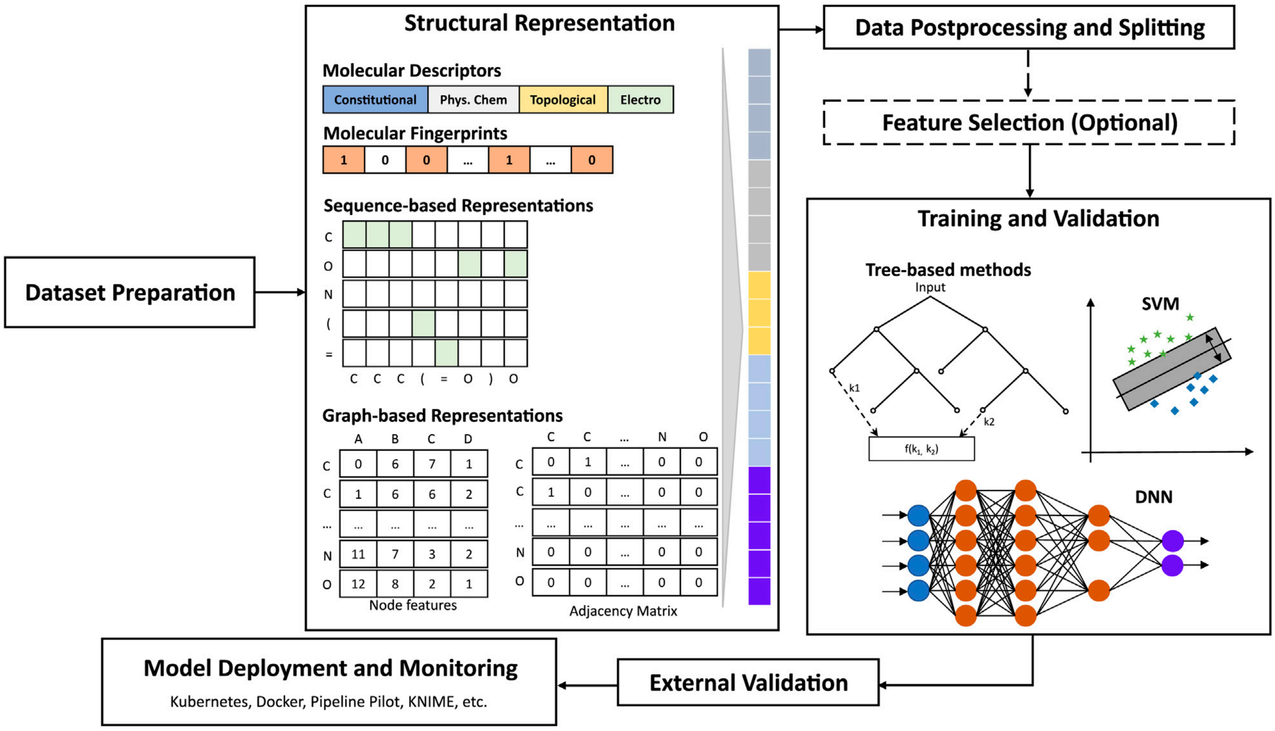


FIGURE 6
Schematic overview of the QSAR/QSPR modeling and deployment process.

environmental fate (biodegradation half-life) endpoints, along with other fate-related properties (e.g.: water solubility). In general, the combination of QSPR-, ag-likeness-, and other endpoint prediction models (e.g.: QSAR), can guide stepwise virtual screening programs, as demonstrated in several studies (Oršolić et al., 2021; Lewer et al., 2022). These tools provide much needed capabilities for ligand-based discovery, especially in early stages, where targets and/or modes of action are unknown.

At later stages of the discovery pipeline, leads must still be optimized with respect to activity against target and non-target species (Martin et al., 2017), favorable/unfavorable modes of action (Kienzler et al., 2017), efficacy, metabolism in target and non-target species (Clark, 2018; Diéguez-Santana et al., 2022), abiotic degradation, and (eco-)toxicity (Devillers et al., 2015; Venko et al., 2018), among other parameters. Some examples include the implementation of a 3D-QSAR approach for the prediction of acetylcholinesterase inhibition of pesticides (Lee and Barron, 2016), the integration of mode of action information into classification and regression QSAR models for the prediction of acute toxicity in honeybees (Carnesecchi et al., 2020a), and the development of OECD-compliant models that accurately predict biodegradation rates of organic compounds (Tang et al., 2020). Additionally, QSAR/QSPR models in later stages could enhance for instance, the improvement of activity and ADME-Tox profiles, and the promotion of more sustainable crop protection agents with minimal risk of resistance (Oršolić et al., 2021). Given the structural differences between pesticides and drugs, it is worth noting that the QSAR/QSPR tools used at each stage of the pipeline, should be either generalizable enough, or at the very least, applicable to the local or global agchem space of interest. Unfortunately, most predictive models available either commercially or open-source are trained on datasets significantly biased towards drugs and drug-like molecules. Moreover, many of the relevant published studies focus on small samples (<500 compounds), thus describing local models. Consequently, crop protection discovery scientists are often forced to a tradeoff between using such tools with less certainty, adapting them towards agrochemicals, or building entirely new predictive models (See Figure 6).

Developing valid, OECD-compliant (Benfenati et al., 2011; Belfield et al., 2023; OECD, 2023) predictive models depends on several key factors: 1) high-quality datasets; 2) proper mathematical representations of molecules that capture key elements essential for the learning task, and powerful computational methods to capture the complex patterns of association between the molecular representations and target endpoints; 3) rigorous performance evaluation criteria, and 4) adequate methods for explainability and uncertainty estimation. In the following, "QSAR" is used as a general term for the quantitative relationship between chemical structures and relevant endpoints (activity, properties, biodegradability, toxicity, etc.).

3.4.2.1 High-quality datasets

Predictive models typically require training on sufficiently large and diverse datasets. Modern high-throughput techniques for the measurement of proxy points (e.g.: LogD), along with increasingly powerful automated text mining and data extraction technologies (Han et al., 2010; Tarasova et al., 2019; NextMove Software, 2022; Shavalieva et al., 2022) have enhanced the acquisition of physicochemical and biological data through internal laboratories (Zhang et al., 2018), CROs, and large-scale data mining projects, sometimes resulting in the publication of FAIR-compliant data (Wilkinson et al., 2016). However, many data-related issues still impede the development of accurate "Ag-adapted" models. These issues include: 1) relatively smaller number of data collected through whole organism assays; 2) relatively smaller coverage of ag-like compounds and Ag-relevant assay data (e.g.: non-target toxicity, bioremediation, plant metabolism, etc.) in public and private databases (Lewis et al., 2016; Gaulton et al., 2017; Williams et al., 2017; Wishart et al., 2017; Kim H. et al., 2021); and 3) the inconsistencies in experimental settings, which are often not taken into consideration during data curation. These limitations can impede the modeling of complex biochemical characteristics or activities and limit the exploration of algorithms such as deep neural networks that require vast amounts of high-quality data. When applicable, scientists often implement different techniques to circumvent these obstacles, that include but are not limited to oversampling/undersampling (Idakwo et al., 2020), cross-validation and cross-testing (Korjus et al., 2016), ensemble learning (Hung and Chang, 2021), data augmentation (Cortes-Ciriano and Bender, 2015; Bjerrum, 2017), transfer learning (Shen and Nicolaou, 2019), multi-task learning (Xu et al., 2017; Martin and Zhu, 2021), representation learning (Kim S. et al., 2021), and self-supervised learning (Dillard, 2021).

3.4.2.2 Mathematical representations and machine learning methods

The hypothesis underlying QSAR studies is that structurally similar molecules tend to behave similarly and to exhibit similar physicochemical properties. Therefore, the selection of molecular representations that are predictive of the molecular property endpoint (e.g.: activity, physicochemical property) is critical for any machine learning task (Bender and Cortés-Ciriano, 2021b). Ideally, such representations shall efficiently express the structural composition of, and subtle nuances between molecules, in a faithful and consistent manner (Chuang et al., 2020). Moreover, interpretable representations would facilitate the human understanding of relevant patterns learned. One can distinguish between fixed, more interpretable representations (e.g.: whole molecule descriptors, atomic descriptors, quantum properties, dictionary- and hash-based fingerprints), and learned, more parsimonious, less interpretable representations (e.g.: convolution- or sequence-based embeddings). Readers are referred to reviews that provide detailed descriptions, comparisons, and applications of structural representations for QSAR (Shen and Nicolaou, 2019; David et al., 2020).

Traditional machine learning approaches typically involve a challenging combinatorial optimization process which consists of selecting a set of most relevant features or feature combinations from a variety of pre-calculated, fixed representations (Goodarzi et al., 2012; Mao et al., 2021) that serve as input to build predictive models that implement one or many algorithms (e.g.: Random Forest, SVMs) (Wu et al., 2020; Yang L. et al., 2021). Molecular fingerprints are often used in addition or as alternatives to the common 2D/3D (e.g.: constitutional and topological) descriptors. In fact, several studies have demonstrated that models based exclusively on fingerprints can outperform 2D/3D-descriptor-based models on various tasks (Venkatraman, 2021). For instance, Li et al. (2017) developed binary and tertiary classification models to predict pesticide aquatic toxicity against rainbow trout and *Lepomis* species, using only fingerprints. The best models implemented SVMs or ANNs on MACCS (Durant et al., 2002) or Graph-only fingerprints and achieved accuracies of 0.89 or higher. Examples of open-source packages that compute molecular fingerprints (FPs) include RDKit, CDK, and the PaDEL descriptor software. Limitations of molecular fingerprints include, among others, limited applicability domain of dictionary-based FPs, and sparsity, possible data loss, and bit collision for hash-based FPs. Moreover, the best fingerprint type can vary depending on the problem, and even between different train-test splits of the same dataset (Sandfort et al., 2021). To address these, several methods have been proposed, such as variants of circular fingerprints, and the combination of various fingerprint features (Capecchi et al., 2020; Sandfort et al., 2021). Fingerprints and molecular descriptors are by no means mutually exclusive. In fact, in many cases, the combination of both types of descriptors can lead to better results (Shi et al., 2018; Tian et al., 2021).

The success of ANNs in computer vision and natural language processing (NLP) in the 2000s has renewed interest in these algorithms, which had fallen out of favor due to many practical issues (e.g.: speed, overfitting, memory requirements). As early as 2008, Sparks *et al.* proposed a new ANN-based QSAR approach capable of suggesting structural modifications that dramatically improved the biological efficacy of Spinosyn analogs (Sparks et al., 2008), where other machine learning methods had failed. This innovation contributed to the design and registration of Spinetoram, a semi-synthetic insecticide. In 2015, Ma et al. (2015) demonstrated that deep neural networks trained using a set of atom pair-, and donor-acceptor pair-descriptors for molecular representation could routinely outperform the most-commonly used random forest models, with a 10% mean R^2 improvement (Ma et al., 2015) on various datasets. These success stories contributed significantly to the renewed interest in deep learning (DL) for chemistry (Chen et al., 2018). Chemical structures can be represented as graphs, or word sequences (e.g.: SMILES (Weininger, 1988)). Therefore, several algorithms have been developed to adapt DL algorithms, once prominent mostly in computer vision, NLP, and network modeling to the world of chemistry.

Prominent DL architectures for molecular property prediction include Graph Convolutional and Sequence-based models (See Figure 6). Graph convolutional networks (GCNs) take as input molecules encoded as graphs where nodes represent heavy atoms and edges represent covalent bonds between them (Gilmer et al., 2017; Lee and Min, 2022). Sequence-based models borrow ideas from NLP to utilize molecular representations such as SMILES strings for learning relationships between different parts of a molecule (akin to learning relationships between different words in a sentence) through recurrent neural network-based architectures such as LSTM and GRU (Goh et al., 2017). Several graph- and sequence-based DL algorithms have been implemented in DL packages such as DeepChem (Ramsundar et al., 2019), Chemprop (Heid et al., 2023), and DGL-LifeSci (Li M. et al., 2021). Over the last 5 years, significantly increased performances in molecular property prediction using DL relative to traditional machine learning models have been reported, with applications ranging from ADME-Tox modeling to bioactivity prediction (Montanari et al., 2019; Zhou et al., 2019; Feinberg et al., 2020; Stokes et al., 2020). More recently, several variations of graph-based and sequence-based (SMILES) algorithms have been demonstrated to achieve 14%–133% better performance than traditional machine learning algorithms in the prediction of relevant properties, in single- or multi-task settings (Honda et al., 2019; Yang X. et al., 2019; Sun et al., 2020).

A key advantage of DL algorithms is their capability of learning molecular representations in a supervised or unsupervised mode, with varying degrees of generalizability, depending on the intended use (Chuang et al., 2020). These representations, also known as molecular embeddings, can be trained using a variety of algorithms (e.g.: neural-based autoencoders, graph-neural networks, self-attention) to extract diverse information about physicochemical properties, structural properties, bioactivity, and other endpoints (Koutroumpa, 2023). The resulting DL models not only learn their own expert feature representations directly from the data, but they also learn how to weigh these features to deliver accurate predictions. Several frameworks have been implemented and published, which consist of pretraining models for sequence- or graph-based molecular representations in a self-supervised or unsupervised framework, using large unlabeled datasets (e.g.: ChEMBL (Gaulton et al., 2017), ZINC (Irwin and Stoichet, 2005)). The models can then be fine-tuned for more specific tasks. This methodology is particularly amenable to transfer learning, which has been very well exploited in both graph computing and NLP spaces. For instance, Ashtawy *et al. (*Ashtawy et al., 2021*)* pre-trained a GNN molecular representation model that performs comparably or better than supervised models when fine-tuned over several ADMET related tasks. Li and Fourches proposed MolPMoFiT, a transfer learning approach based on self-supervised pre-training and task-specific fine-tuning for QSPR/QSAR modeling (Li and Fourches, 2020). MolPMoFiT was used to build predictive models for small datasets that showed comparable or better performances on several datasets compared to state-of-the-art D-MPNN, Random Forest, and other Feed Forward Network models. Lately, inspired by their success in NLP, attention-based transformer models (Honda et al., 2019; Irwin et al., 2022) have emerged as more powerful architectures for encoding molecular representations to predict reactions or properties. For example, to learn molecular embeddings, Irwin et al. (Irwin et al., 2022) pre-trained several Bidirectional Auto-Regressive Transformer (BART) models on >100 million datasets from the ZINC-15 dataset (Irwin and Stoichet, 2005). In a multi-task learning framework, the models were rapidly trained on several sequence-to-sequence (e.g.: direct synthesis) and discriminative (e.g.: activity) prediction tasks, yielding task-specific models with comparable or better performance compared to the baseline. Several other sequence-based (e.g.: Bidirectional Encoder Representations from Transformers (BERT), Siamese RNNs, and graph-based (e.g.: D-MPNNs) frameworks for representation and transfer learning have been developed and implemented to build predictive models with improved performances (Yang K. et al., 2019; Payne et al., 2020; Fernández-Llaneza et al., 2021) (See Table 4). Moreover, to leverage the advantages and alleviate the limitations of various molecular representations, it is common to build hybrid architectures by combining them, as illustrated by several recent publications (Hasebe, 2021; Li M. et al., 2021).

Training DNNs typically requires larger amounts of training data compared to traditional ML models. NLP-based algorithms can benefit from numerous augmentation methods, including SMILES randomization (Bjerrum, 2017; Arús-Pous et al., 2019) and other SMILES-derived encodings (Lambard and Gracheva, 2020) that can lead to improvements even in low-data regimes. Representation and transfer learning provide opportunities to lower data size requirements for the development of accurate predictive models. Increasingly popular techniques include one-shot-, few-shot-, and meta-learning, which learn rich molecular representations from relatively small datasets (Altae-Tran et al., 2017; Nguyen et al., 2020; Wang F. et al., 2021; Fernández-Llaneza et al., 2021; Guo et al., 2021) and self-supervised learning methods that leverage large unlabeled datasets (Dillard, 2021; Li P. et al., 2021). Finally, neural prediction models implement active learning approaches that can effectively sample the set of possible training candidates given a fixed training budget, thereby offering a systemic approach for exploring the data that is at the core of drug discovery research (Konze et al., 2019; Reker, 2019).

Overall, the methods mentioned above help modeling several endpoints of utmost importance in agrochemical discovery that have traditionally been difficult to tackle. For instance, the prediction of activity translation, which is typically limited to small datasets given the low number of molecules tested, especially in higher tiers, could be addressed using approaches that perform well in low-data regimes. The key methods discussed above are summarized in Table 4. For a comprehensive review of molecular representations and machine/deep learning methods used for molecular property prediction, readers are referred to other publications (Lo et al., 2018; Lo et al., 2019; Sun et al., 2020; Wieder et al., 2020; Mao et al., 2021; Dhamercherla et al., 2022).

3.4.2.3 Rigorous performance evaluation criteria

The success stories referenced throughout this review highlight not only the importance of AI in crop protection discovery, but also the fact that so far, no single (ML or DL) algorithm or molecular representation (Sabando et al., 2021; Orosz et al., 2022) is found to be best suited for most modeling tasks. It is thus important to define means for adequate comparative evaluations of a model as it would provide a fair model assessment and facilitate the selection of the most suitable algorithms and approaches for future modeling tasks (Liu et al., 2018b). Examples of high-quality datasets that are used for training and comparative evaluations include, among others, the Tox21 (Huang et al., 2016; Mayr et al., 2016), PubChem BioAssay (Wang et al., 2012), and MoleculeNet (Wu et al., 2018) datasets, which are available either in raw formats or as encoded objects in various DL packages such as DeepChem (Ramsundar et al., 2019), Chemprop (Yang X. et al., 2019), and DGL-LifeSci (Li P. et al., 2021). It is highly desirable that such packages also include datasets for Ag-relevant molecular endpoints. In recent years, several comparative evaluations (with respect to accuracy, computational efficiency, etc.) of traditional and deep learning algorithms have been published (Jiang et al., 2021; Rao et al., 2021). In several experiments, traditional machine learning using traditional molecular representations approaches significantly outperformed deep

TABLE 4 Examples of key AI-driven algorithms and methods for prediction of molecular properties.

Class	Method description	References/Examples
Dynamic Selection	Techniques for dynamic selection of classifiers based on individual sample	Cruz et al. (2018)
Ensemble learning	Combination of multiple learners for performance improvement	Svetnik et al. (2003); Sheridan et al. (2016); Kwon et al. (2019); Davronov and Adilova, (2021)
Fully Connected DL	Fully connected deep learning network for Single-task QSAR analysis	Ma et al. (2015)
GCN	Multitask graph convolutional networks	Montanari et al. (2019)
Fully Connected DL	Fully connected deep learning network for Multi-task QSAR analysis	Kearnes et al. (2016)
GCN	PotentialNet family of graph convolutions for protein-ligand binding affinity	Feinberg et al. (2018); Feinberg et al. (2020)
GNN	Molecular Contrastive Learning	Wang et al. (2021c)
MPNN	Message passing neural networks for molecular property prediction	(Yang X. et al. (2019); Stokes et al. (2020); Chen et al. (2021); Heid et al. (2023)
Graph Transformer	Molecular encoding using hybrid MPNN-Transformer architectures	Rong et al. (2021)
NLP inspired	Autoencoder-based Molecular encoding and QSAR	Winter et al. (2019)
NLP inspired	Transformer based encoder model	Honda et al. (2019); Payne et al. (2020); Irwin, Dimitriadis et al. (2022)
NLP inspired	Transfer learning for NLP based classification tasks	Li and Fourches (2020)
NLP inspired	Siamese RNNs for QSAR Prediction	Fernández-Llaneza et al. (2021)
Active Learning	Retrosynthetic and combinatorial synthesis coupled with Active Learning	Konze et al. (2019)
Explainable Artificial Intelligence	Methods to provide interpretability to ML/DL models. These include approaches for explaining their predictions, quantifying their uncertainty, and estimating their applicability domains	Interpretability Ribeiro et al. (2016); Lundberg and Lee, (2017); Nori et al. (2019); Rodríguez-Pérez and Bajorath, (2021)
		Uncertainty estimation Liu et al. (2018a); Cortés-Ciriano and Bender, (2019); Gawlikowski et al. (2021); Zhong et al. (2022)
		Applicability domain Liu and Wallqvist, (2019), R. P. Sheridan, (2015); Schroeter et al. (2007); Supratik et al. (2018)

learning models using unsupervised molecular representations, showing a different trend than studies referenced in the previous section. Interestingly, in a recent study combining less expensive traditional algorithms, such as Gaussian processes and random forests, Green et al. (Green et al., 2023) demonstrated that fixed [e.g.: ECFP (Rogers and Hahn, 2010]) or learned representations [e.g.: Mol2vec (Jaeger et al., 2018)] could often yield better overall results compared to fully deep-learning-based approaches, both for property- and ADMET-related predictive modeling tasks. The overall takeaway is that the potential of DL has not yet been fully exploited in chemistry. In contrast to other areas like computer vision, there is still a lot to uncover and prove. Moreover, traditional ML algorithms and molecular representation techniques will not be obsolete soon. It can be expected, as pointed by Bender and Cortés-Ciriano (Bender and Cortes-Ciriano, 2021a), that learned representations could become more useful in high-data regimes, whereas expert-chosen representations will probably remain more useful when data is scarce. Benchmarking would help establish guidelines in the setup and hyperparameter tuning, and in identifying trends that guide the selection of appropriate algorithms, molecular embeddings, and predictive models. Additionally, meta learning (Olier et al., 2018) can help understanding the relationships between the performance of ML algorithms and measurable properties, as well as selecting the best predictive models (Cruz et al., 2018; Olier et al., 2018; Cruz et al., 2020). Furthermore, given that ligand-based models are prone to false positives, more research is needed to develop algorithms that systematically identify gaps where the model learned a trivial relationship that is not generalizable.

3.4.2.4 Explainability and uncertainty estimation of predictive models

Besides high performance (as measured by various metrics) and scalability, it is highly desirable that predictive models be explainable. The ability to assess the contribution of a molecule's various structural features and physicochemical properties, among other features, towards quantitative or qualitative output variables is critical for designing, assessing, and optimizing molecules. Unfortunately, the black box nature of most ML (especially DL) approaches, makes it difficult to interpret the prediction from QSAR models, and thus, impedes their widespread adoption. In recent years, explainable AI (XAI) has been the focus of numerous drug discovery research projects (Jiménez-Luna et al., 2020b). In the area of QSAR, one can distinguish among feature-, atom/fragment-, compound-, and graph-based approaches for model explanation (Rodríguez-Pérez and Bajorath, 2021) (See Table 4). While atom-/fragment-based and graph-based approaches could, for instance, highlight substructures that contribute to soil degradation of a specific molecule, feature-based approaches could explain how specific

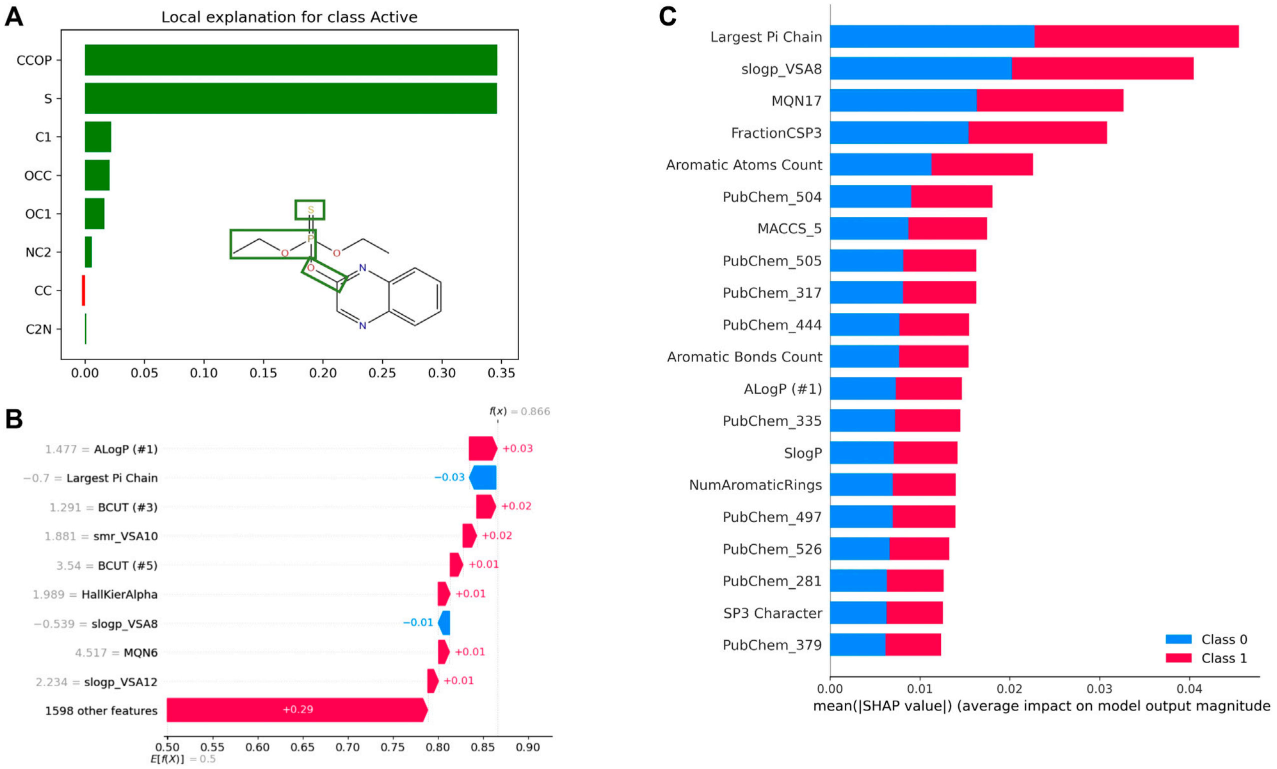


FIGURE 7
Examples of QSAR model explanations using feature attribution methods LIME and SHAP. **(A)** Explanation of Quinalphos' predicted aryl hydrocarbon receptor (AhR) agonistic activity using LIME. The explanation displays the contribution of various tokens to the predicted output from an LSTM model trained on the Tox21 AhR dataset using SMILES representations. **(B)** SHAP Waterfall Plot to explain the contributions of various features towards AhR activity of a single molecule as predicted by a random forest model trained on the same dataset. This time, the model uses physico-chemical properties computed by RDKit as input features. For this example, the ALogP contributes the most towards predicted activity (class 1), while the Largest Pi Chain features contributes the most towards predicted inactivity (class 0). **(C)** Global interpretation of the AhR random forest model with SHAP. On average the "Largest Pi Chain" features contributes the most towards predicted activity and predicted inactivity.

molecular properties influence the toxicity against honeybees, for example. Commonly used methods include feature attribution and graph-convolution-based methods. Feature attribution methods, such as SHAP, LIME, and DeepLIFT, determine the importance of every input feature towards a prediction (See Figure 7). Various subgraph identification, and attention-based approaches have been developed to provide explainability to DNN models (Karpov et al., 2020; Weber et al., 2021). For instance, GNN-explainer, which provides explanations for every graph-based machine learning, was able to correctly identify several functional groups known to be mutagenic to *Salmonella typhimurium* (Ying et al. 2019). In recent years, several benchmarks have been published for comparing the interpretability of various XAI methods using traditional (e.g.: Random Forest, SVMs) and deep learning models (Sanchez-Lengeling et al., 2020; Klaise et al., 2021; Matveieva and Polishchuk, 2021).

The interpretability of a model can not only provide insights into the relationship between features and the modeled outcome, but also helps to select the best features to model similar tasks, resulting in better performance. However, as recommended by Muratov et al. (2020), model explanations must be used with caution. Scientists should only be confident in a predictive model if it is generalizable enough to perform well on unseen data, and the molecules of interest are within the model's domain of applicability. It is, therefore, important that the predictive model be deployed along with tools or capabilities to define its domain of applicability for the assessment of compounds of interest, and to estimate the uncertainty of its predictions. Several approaches (e.g.: ensemble, probabilistic, and distance) that are applicable to different types of machine learning algorithms have been developed to quantify prediction errors and estimate applicability domains (Schroeter et al., 2007; Cortés-Ciriano and Bender, 2019; Gawlikowski et al., 2021). As demonstrated by Zhong et al. (2022), uncertainty estimation can also be used to increase the applicability domain of QSAR models, which is critical, especially in low-data regimes. Overall, it is believed that implementing methods for uncertainty estimation and model explainability could help tackle some of the most challenging, unaddressed problems, such as the prediction of activity translation and the prioritization of molecules between different experimental tiers, as the number of datapoints becomes increasingly smaller and more realistic experimental settings are employed for testing, thus increasing the complexity of modeling tasks.

3.4.3 Structure-activity-relationship (SAR) visualizations

Adding information from our structure classifications (e.g.: R-group decomposition described above) and predictive models

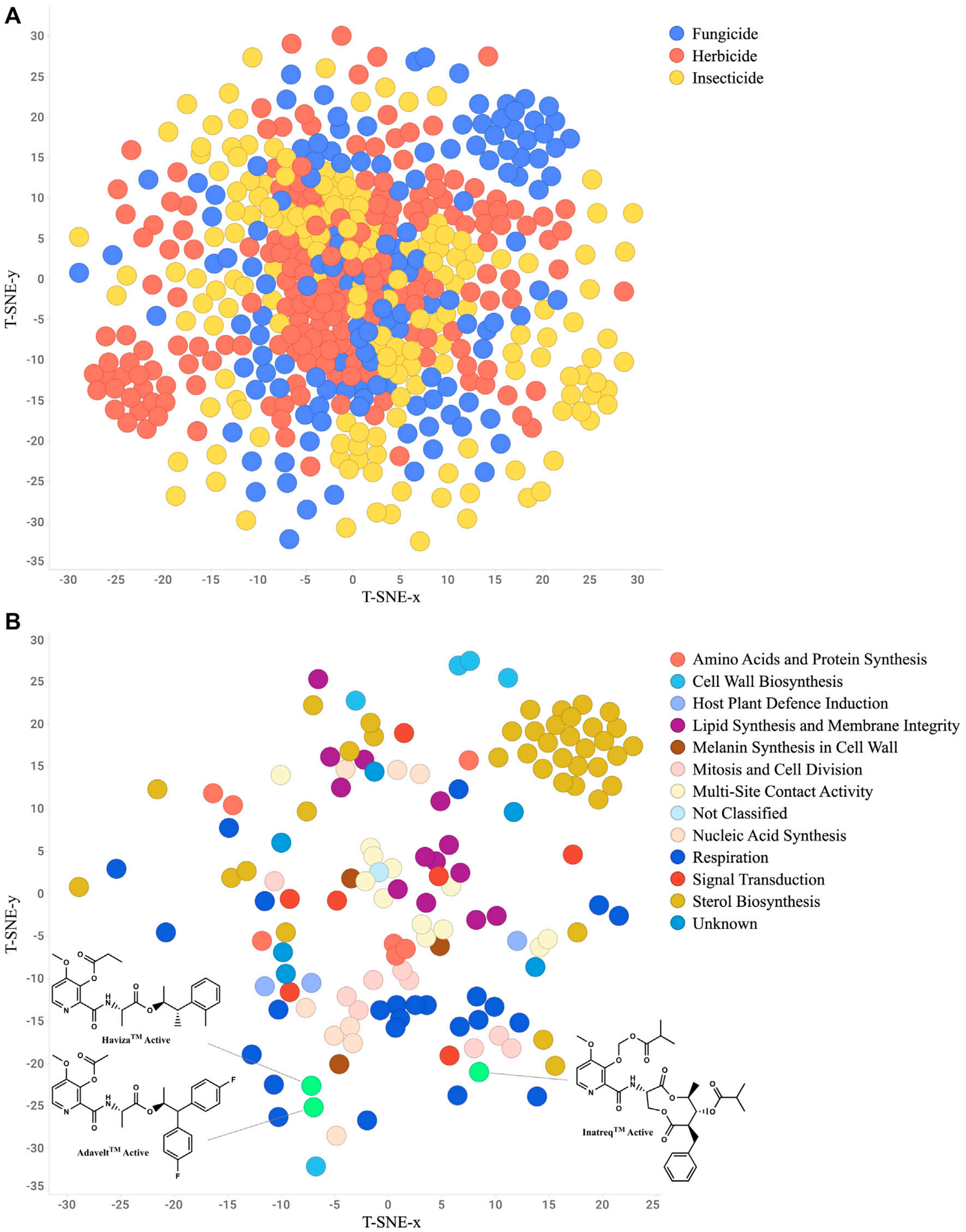


FIGURE 8
(Continued).

to any plot of biological activity automatically gives rise to SAR plots. It is now possible to look for themes and features that drive towards the desired activity profiles. Modern data visualization tools such as Tibco Spotfire with Lead Discovery (Perkin Elmer) (Elmer, 2023) and Tableau (Tableau Software, 2023) make it easy to construct interactive displays that allow the researcher to explore the connections between the structural features and classes, and the biological data. There are several good examples of additional SAR visualizations in the literature including SARNEA (Lounkine et al., 2010), SAR Matrices, which can overcome the inflexibility of R-group decomposition (Yoshimori et al., 2019), and SAR Maps (Agrifiotis et al., 2007). SAR Matrices can support bioactivity prediction, and large-scale database building for analog searching, among other applications (Yoshimori and Bajorath, 2020).

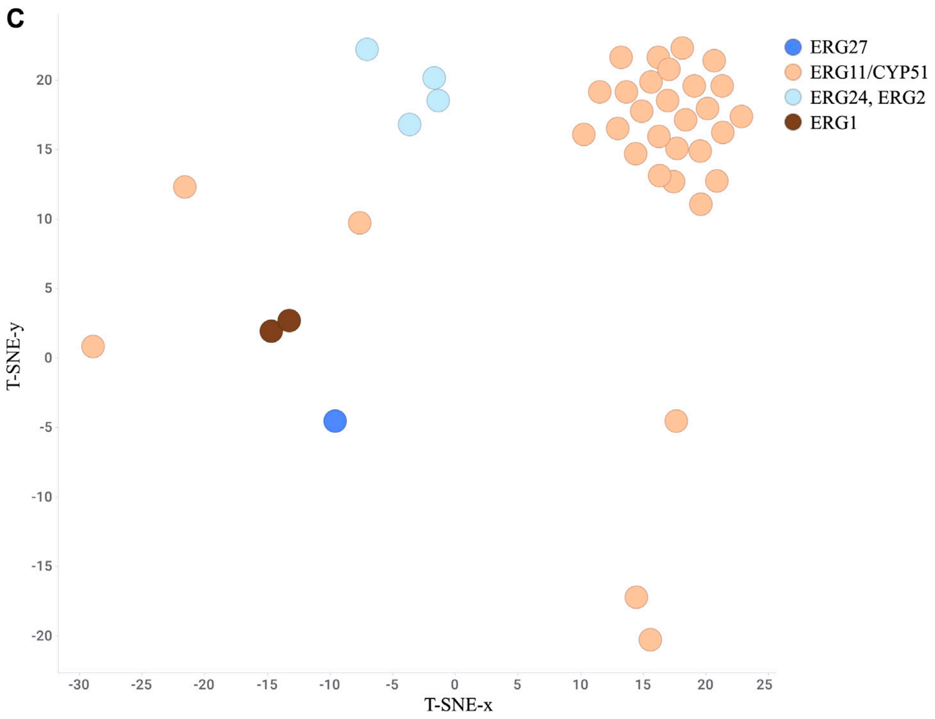


FIGURE 8
(Continued). T-SNE analysis of all FRAC, IRAC, and WRAC compounds in the ChemBL database using the cuML library using circular fingerprints using the OpenEye toolkit (atom type, 2 bond radius) folded into a 512 bit array **(A)** then filtered to fungicides after the analysis **(B)** and further filtered to sterol biosynthesis inhibitors **(C)**.

One valuable visualization for the researcher is a "Chemistry-Space Map". This is often called a star-field map due to its similarity to nighttime sky. Each compound is mapped in a 2D or 3D space in such a way as to group the most similar compounds together while still showing the relationships to more dissimilar compounds. The layout is created using a set of structure descriptors and then analyzed using a numerical approach such as t-distributed stochastic neighbor embedding (t-SNE) (Karlov et al., 2019; Andronov et al., 2021), Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018), or Tree MAP (Probst and Reymond, 2020). These maps provide a useful structure-based organization of the project chemistry which can then be analyzed further by layering on the biological results (Janssen et al., 2019). For example, Gonçalves *et al.* (Gonçalves et al., 2021) utilized a combination of t-SNE and k-means methods to compare several hundred novel isoxazolines to commercialized isoxazoline insecticides, clearly identifying areas of novelty. Additionally, Wang et al. (Wang et al. 2022) mapped approved drugs with similar commercial herbicides to suggest isoteric replacements for novel herbicide chemotypes. Figure 8A depicts a t-SNE analysis of all FRAC, HRAC, and IRAC compounds available in ChEMBL. As this dataset does not contain the newest picolinamide fungicides Inatreq™, Adavelt™, and Haviza™, we added these structures and filtered to only fungicides to highlight this technique's ability to qualitatively identify novel areas of chemistry (Figure 8B). The newer, non-macrocyclic picolinamides are clearly distinct from the natural product-derived macrocycle. Further filtering to only sterol biosynthesis inhibitors (Figure 8C) provides clusters at each site of action which generally overlap with chemical class.

3.4.4 Metabolism

Metabolomics is a field of "omics" research that focuses on the high-throughput characterization and identification of small products of cell metabolism, called metabolites, within biological matrices (Wishart, 2008). In agriculture, the metabolite content and its alterations are related to developmental and differentiation processes, plant and fruit maturation processes, as well as resistance to external stimuli such as pathogen attacks and other environmental factors (Ibarra-Estrada et al., 2016). From an active/lead discovery standpoint, how a molecule is metabolized within a biological species influences its mode of action, bioactivity, and toxicity, among other parameters (Aliferis and Chrysayi-Tokousbalides, 2011). Compared to pharmaceutical discovery, metabolomics studies in agrochemical discovery encompass a larger and more complex set of biological systems, as pesticides are applied on and around crops. These systems include pests (insects, weeds, fungi, fish, etc.), the crops to be protected, and the surrounding environment (non-target organisms, soil, water, etc.). While in the early design phases scientists may often promote molecules with no major toxic metabolites to rapidly discard toxic molecules, or control the metabolism/activation process of the future active ingredient (Jeschke, 2016), it becomes important, as molecules progress through the lead optimization stage, to identify all metabolites for toxicity and environmental fate (U.S. EPA, 2023a). Moreover, less toxic yet non-degradable compounds are of concern too, due to bioaccumulation potential. For these reasons, a deep understanding of small molecule metabolism and environmental fate at early discovery stages would allow rapid screening of millions of compounds to prioritize the more promising ones, thus leaving more time and resources for synthesis and other phases of the DMTA cycle (Clark, 2018) or simply to accelerate the whole discovery process. Thus, over the last two decades, cheminformatics and AI-based tools have become ubiquitous in the development of cost-effective tools for automated metabolite elucidation and metabolic data interpretation.

A major bottleneck in metabolomics is the structural elucidation of small molecules detected in metabolism and environmental fate studies (Wishart, 2008). Typically, this involves the acquisition of spectra (mass or nuclear magnetic resonance (NMR)) from biological or environmental samples, followed by their processing and matching against reference spectral databases to identify the corresponding chemical structure(s) (See Table 3). Unfortunately, current libraries are neither comprehensive nor structurally diverse enough to support the retrieval of all known metabolites, which often leads to extremely low identification rates (<2%) in untargeted metabolomics experiments (da Silva et al., 2015). To address this data scarcity and many other challenges of structure elucidation, *in silico* approaches usually follow the assumption that structurally similar compounds possess similar fragmentation patterns and properties under similar conditions (Schollée et al., 2017). To assess and leverage this "quantitative structure-fragmentation relationship" (QSFR), they generally combine cheminformatics concepts such as molecular fingerprints (Rogers and Hahn, 2010), structure-based classification (Djoumbou et al., 2016; Kim et al., 2020), chemical-informed clustering (R. Ash, 2019), and structure/reaction representation languages (Daylight Chemical Information Systems, Inc, 2019a; Daylight Chemical Information Systems, Inc, 2019b) with machine learning approaches, ranging from random forest to DNNs (Liebal et al., 2020). Notable contributions to improve metabolite identification workflows include, among others, tools for spectra pre-processing (Li and Wang, 2019; Melnikov et al., 2020), the prediction of MS and NMR spectra from molecular structures (Castillo et al., 2016; Wang et al., 2021b; Hong et al., 2023), as well as the prediction of molecular structures and features from MS spectra (Wang et al., 2021c; Dührkop et al., 2021) (See Table 3). Altogether, such ML tools can be used to propose chemical structures without database search or expanding reference databases (Guijas et al., 2018; Wishart et al., 2018; Djoumbou-Feunang et al., 2019b) to boost the identification rates.

The identification of (major) metabolites and metabolic pathway, and the expansion of metabolome databases in particular, could be further facilitated by using metabolism and environmental fate prediction tools, which can suggest biologically feasible and relevant structures. Such tools are thus relevant not only in early (e.g.: molecule design, lead optimization, ADME-Tox profiling) but also in late (e.g.: metabolism and environmental fate studies, ecotoxicological risk assessment) stages of the agrochemical discovery process (Clark, 2018). Currently, most of the available software tools focus solely on human/mammalian metabolism (Litsa et al., 2020; SimulationsPlus, 2023), and environmental microbial metabolism (Wicker et al., 2016). Only a few tools (Djoumbou-Feunang et al., 2019a; QSAR Toolbox, 2023) allow comprehensive prediction for several biological systems including human hepatic and gut microbial, environmental microbial, etc. Unfortunately, most of these tools have been developed using training data mostly comprising drugs and drug-like molecules and perform less well on ag-relevant chemistries. Moreover, xenobiotics metabolism is difficult to predict, as several factors (polymorphisms, expression levels, reference data scarcity, etc.) affect the training of predictive models. Another limitation to the comprehensive prediction of metabolism and biodegradation products obtained from agrochemicals is that software tools that predict plant metabolism (Karp et al., 2019; Pathway Tools, 2023), and abiotic transformations (U.S. EPA, 2023b) of small xenobiotics are scarce at best. Furthermore, the increasing emphasis on pollinator-friendly farming implies that there is an urgency to develop and share computational tools and resources that can enable the prediction/elucidation of metabolites in pollinator species (e.g.: bees), and how metabolic alterations affect them (du Rand, 2015). These limitations must be addressed by harnessing data from publicly available regulatory reports and private data. Furthermore, scientists should leverage recent works in the area of synthetic reaction prediction, as the methodologies and algorithms could apply to metabolism prediction as well. For instance, Litsa *et al.* developed transformer-based, template-free model for mammalian drug metabolite prediction, with comparable performance to template-based models such as BioTransformer and SyGMa (Litsa et al., 2020).

The analysis and interpretation of metabolic data is another very cumbersome task in metabolomics. Such analyses are conducted to study the response of plants and target species to external stimuli, and identify metabolic/biosynthetic pathways, among other tasks. Some of the most promising approaches rely on statistical analysis, as well as substructure- and network-based methods, which are often combined with machine/deep learning. Readers are referred to publications by De Souza *et al.* (De Souza et al., 2020), Ramos *et al.* (Ramos et al., 2019), and Beniddir *et al.* (Beniddir et al., 2021) for a comprehensive review of such methods. Furthermore, several recent articles provide a comprehensive review on the applications of AI in metabolomics (Uppal et al., 2016; Heinemann, 2019; Liebal et al., 2020; Pomyen et al., 2020). For a list of software tools and resources that enhance metabolomics, readers are referred to Table 3.

4 Impact of cheminformatics on sustainability and environment-friendly programs

Insect population decline (Belsky and Joshi, 2019; Sánchez-Bayo and Wyckhuys, 2019) has, partly due to use of agrochemicals, led to the development of novel strategies to promote ecological-resilience and sustainable crop protection. Examples of such strategies include designing new broad-utility nitrogen stabilizers with improved safety, and herbicides with novel, un-exploited, or proven efficient mode of action for effective and durable control of driver weeds in crops. Within just a few decades, the term sustainability has gained in popularity and significance. Recent agricultural methods are far more efficient than those farmers used a few decades ago, primarily due to advancements in technology such as using big data in agricultural practices to characterize chemical toxicity and impacts on human wellbeing and ecosystem health, and advancements in plant genotyping methods and sequencing as promising tools for plant breeding and genetics research. The insights and recommendations derived from these advanced data analytics and bioinformatics tools together with the adoption of precision agriculture technologies guide us toward having safer, efficient, and more environmentally friendly alternatives.

Mathematical and *in silico* models are being used in food (from farm to fork) and agriculture sectors for sustainable and resilient systems with the general goal of providing safer food and transitioning to more sustainable farming. For instance, DynamiCrop is a plant-uptake multi-compartment mathematical model (DynamiCROP, 2023) used for the assessment of human exposure from pesticide residues in food crops. The model uses databases of reference plant dissipation half-lives of 333 pesticides in crops to estimate the amount and traces of pesticides in multiple crops and also to estimate human health impacts due to the uptake of pesticides (Peter et al., 2014). Apart from dynamic mathematical models, *in silico* models and tools in the agriculture industry are being considered inexpensive and fast alternative approaches to toxicological and ecological assessments (Supratik et al., 2017). Short-term toxicity assays such as Ames-mutagenicity and carcinogenicity (Benfenati et al., 2019), skin sensitization (Borba et al., 2021), and bee toxicity (Carnesecchi et al., 2020b) are examples of toxicity assessments that can be assisted by *in silico* models for prediction of such toxicological testing where prediction on a new pesticide candidate can be made merely by using the chemical structure.

Natural products have long been used as pesticides and have broadly served as a source of inspiration for synthetic organic fungicides, herbicides, and insecticides. Natural products are produced by biosynthetic enzymes and pathways. Cheminformatics tools can enhance structural characterization and activity specification of pesticidal natural products, and thus, make substantial contributions to the renewed field of natural product discovery (Chen and Kirchmair, 2020). During the hit-to-lead and lead optimization phases, ML approaches have been applied to natural products for predicting bioactivity and their protein targets (Olğaç et al., 2017; Cockroft et al., 2019). For example, STarFish (Cockroft et al., 2019) uses publicly available natural product databases and implements a stacked ensemble approach that combines multiple ML classification models to predict the protein target for the bioactive natural product. A recent publication uses machine learning classifiers to predict antibacterial or antifungal activity directly from natural product biosynthetic gene clusters (Walker and Clardy, 2021).

An example of commercially successful natural pesticides are Spinosyns, a large family of substances produced from fermentation of a soil inhabiting bacterium (*Saccharopolyspora spinosa*) (Kirst, 2010). Two insecticidal products have been commercialized from spinosyns: 1) Spinosad, a naturally occurring mixture of spinosyn A and spinosyn D, and 2) Spinetoram, a semi-synthetic derivative of spinosyns (Sparks et al., 2021). Spinosad received an expedited review and has been registered for integrated pest management and insect control by EPA since early 1997 (National Pesticide Infomation Center, 2023). It is valuable in control of insect pests, while minimizing the impact on beneficial insects (Sarfraz et al., 2005), and has a favorable environmental profile as it does not persist in the environment. Moreover, since Spinosad adsorbs to the soil with a higher affinity (especially in soil-clay), leaching through unsaturated soil to groundwater resources is minimized (Sparks et al., 2021). Molecular modeling using cheminformatics and AI tools has contributed, among others, to the discovery of spinosyns, and in particular, to the development of the semi-synthetic product, Spinetoram, with improved insecticidal efficacy and expanded spectra (Sparks et al., 2008; Sparks et al., 2021), which is considered a new milestone in the age of natural product-based insecticide discovery and crop protection research.

5 Current challenges, and future perspectives

As discussed throughout this review, cheminformatics and artificial intelligence can significantly enhance the design and development of novel and more sustainable crop protection agents. Yet, several factors still limit the wider adoption of *in silico* tools throughout the process. The scarcity of standardized, high-quality agrochemical datasets remains a challenge that hampers several processes, including but not limited to knowledge extraction (e.g.: via semantic-based querying), and predictive modeling.

A rising concern in agrochemical research is the environmental impact of pesticides. This has led to a renewed interest in natural products as pesticides. Besides the methods presented here, metagenomics-based approaches have been developed that provide a means to mine and link the metabolome and genome of species of interest. These techniques can be especially useful in the identification of natural agrochemicals and species-specific target genes. Moreover, the study of gene/protein mutations in resistant pests can be investigated to identify the underlying mechanism and suggest actions for the design of more potent agrochemicals or propose new modes of action. Validation of the proposed hypothesis can only be achieved if protein structures are available for large and diverse sets of ag-relevant species (e.g., pests, pollinators). This is, however, a bottleneck in agchem research that impedes the discovery of novel protein targets and modes of action, as well as the generation and optimization of molecules. In 2021, significant milestones were achieved in the prediction of protein structures based on sequence information with the publication of AlphaFold (Jumper et al., 2021) and RoseTTAFold (Baek et al., 2021). These predictive models could be used, along with other *in silico* solutions, to develop theoretical models and annotate protein-ligand complexes (Simonovsky and Meyers, 2020; Agaarwal et al., 2021; Hekkelman et al., 2023), which could be deposited in public databases (Varadi et al., 2022). For specific discovery projects, relevant models can then be probed using solutions described throughout the paper for hit/active identification, target selection, mode-of-action detection, and the design of *in vitro* protein assays. Furthermore, molecules selected through virtual screening or (generative) *de novo* design, can be synthesized efficiently, and tested *in vitro* to provide data for further (QSAR) analysis.

The rapid development and efficient use of cheminformatics and ML tools requires capabilities to generate, process, store, and transfer data between various environments. Moreover, developing the best tools usually requires probing larger spaces with respect to data, algorithms, and parameters. Fortunately, frameworks for distributed storage and processing (Dask, 2023; Spark, 2023), containerization (Docker, 2023), and orchestration (Kubernetes, 2023), among others, can be used for the development of scalable cheminformatics and ML solutions, and seamless integration with diverse pipelines. Ideally, deployed tools should be coupled with user interfaces providing capabilities for visualization, data manipulation, and better user experience for chemists and biologists. Moreover, efficient means for feedback loops and timely updates of software tools are needed to further engage users (Volkamer et al., 2023). Platform-as-a-service (PaaS), and Model-as-a-service (MaaS) cloud computing solutions can enhance the monitoring, use, and automated release of *in silico* solutions. However, such platforms cannot be designed and used efficiently without significant expertise and close collaborations between all parties involved (scientists and engineers). To alleviate these challenges, chemistry aware *in silico* tools such Torx® (Torx Software Ltd, 2023) and LiveDesign® (Schrödinger, 2023a) have been developed to help manage the complex workflows of compound synthesis, hypotheses tracking, assay cascades, computational analyses, and compound genealogy in a collaborative way.

While major developments have occurred in both experimental and predictive discovery procedures, it is clear that a paradigm shift towards a more AI-involved approach requires a gradual implementation of inter-connected automated software and hardware solutions capable of generating, prioritizing, and validating explainable hypotheses (with or without human biases) upon integrative data analysis for better decision making. This is a fairly complex task, which cannot be achieved independently, and requires cross-collaborations, not only within but between institutions. Several public private consortia have been created over the last 5 years with the aim of developing comprehensive computational discovery and synthesis platforms. Examples include the MLPDS (MLPDS, 2023), the MELLODDY (MELLODDY, 2023), and the ATOM consortia (ATOM, 2023). While these alliances have predominant membership from pharmaceutical companies, we strongly believe that the agrochemical industry would benefit from joining.

Author contributions

YD-F: Conceptualization, Data curation, Supervision, Writing–original draft, Writing–review and editing. JW: Conceptualization, Supervision, Writing–original draft, Writing–review and editing. JK: Conceptualization, Writing–original draft, Writing–review and editing. PC: Conceptualization, Writing–original draft, Writing–review and editing. PY: Conceptualization, Writing–original draft, Writing–review and editing. AS: Conceptualization, Writing–original draft, Writing–review and editing. MS: Conceptualization, Writing–original draft, Writing–review and editing. SS: Conceptualization, Writing–original draft, Writing–review and editing. JO: Conceptualization, Writing–original draft, Writing–review and editing. JH: Conceptualization, Writing–original draft, Writing–review and editing. ES: Conceptualization, Writing–original draft, Writing–review and editing. DT: Conceptualization, Writing–original draft, Writing–review and editing. SK: Conceptualization, Supervision, Writing–original draft, Writing–review and editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Acknowledgments

We would like to thank Corteva Agriscience for supporting the preparation and submission of this article. We are especially thankful to Mark Pobanz, Guomin Shan, C.W. Arnett, Vidyadhar Hegde, Nicole Janovick, and Brian Lutz for their review and approval of the manuscript.

Conflict of interest

YD-F, JW, JK, PC, PY, AS, MS, SS, JO, and JH were employed by Corteva Agriscience. ES was employed by Corteva Agriscience UK Limited. DT was employed by Atomwise. SK was employed by Karyosoft Inc.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., et al. (2015). TensorFlow: large-scale machine learning on heterogeneous systems. [Online] Available at: https://www.tensorflow.org/. doi:10.5281/zenodo.4724125

Abbasi, M., Santos, B. P., Pereira, T. C., Sofia, R., Monteiro, N. R. C., Simões, C. J. V., et al. (2022). Designing optimized drug candidates with generative adversarial network. *J. Cheminformatics* 14 (40), 40. doi:10.1186/s13321-022-00623-6

Agaarwal, R., Gupta, A., Chelur, V., Jawahar, C. V., and Deva Priyakumar, U. (2021). DeepPocket: ligand binding site detection and segmentation using 3D convolutional neural networks. *J. Chem. Inf. Model.*

Agrafiotis, D. K., Wiener, J. J. M., Skalkin, A., and Kolpak, J. (2011). Single R-group polymorphisms (SRPs) and R-cliffs: an intuitive framework for analyzing and visualizing activity cliffs in a single analog series. *J. Chem. Inf. Model.* 51 (5), 1122–1131. doi:10.1021/ci200054u

Agrifiotis, D. K., Shemanarev, M., Connolly, P. J., Farnum, M., and Lobanov, V. S. (2007). SAR maps: a new SAR visualization technique for medicinal chemists. *J. Med. Chem.*, 5936–5937. doi:10.1021/jm070845m

Aldrich, S. (2023). ChemFinder™ ultra. [Online] Available at: https://www.sigmaaldrich.com/US/en/product/aldrich/z511595.

Aliferis, K. A., and Chrysayi-Tokousbalides, M. (2011). Metabolomics in pesticide research and development: review and future perspectives. *Metabolomics* 7, 35–53. doi:10.1007/s11306-010-0231-x

Altae-Tran, H., Ramsundar, B., Pappu, A. S., and Pande, V. (2017). Low data drug discovery with one-shot learning. *ACS Central Sci.* 3, 283–293. doi:10.1021/acscentsci.6b00367

Amazon Web Services (2023). Amazon web services. [Online] Available at: https://aws.amazon.com/.

Andersson, S., Armstrong, A., Björe, A., Bowker, S., Chapman, S., Davies, R., et al. (2009). Making medicinal chemistry more effective—application of Lean Sigma to improve processes, speed and quality. *Drug Discov. Today* 11-12, 598–604. doi:10.1016/j.drudis.2009.03.005

Andronov, M., Fedorov, M. V., and Sosnin, S. (2021). Exploring chemical reaction space with reaction difference fingerprints and parametric t-SNE. *ACS Omega* 6, 30743–30751. doi:10.1021/acsomega.1c04778

Aoyama, T., Suzuki, Y., and Ichikawa, H. (1990). Neural networks applied to structure-activity relationships. *J. Med. Chem.* 33, 905–908. doi:10.1021/jm00165a004

Arús-Pous, J., Johansson, S. V., Prykhodko, O., Bjerrum, E. J., Tyrchan, C., Reymond, J. L., et al. (2019). Randomized SMILES strings improve the quality of molecular generative models. *J. Cheminformatics* 11, 71. doi:10.1186/s13321-019-0393-0

Arús-Pous, J., Patronov, A., Bjerrum, E. J., Tyrchan, C., Reymond, J. L., Chen, H., et al. (2020). SMILES-based deep generative scaffold decorator for de-novo drug design. *J. Cheminformatics* 12 (38), 38. doi:10.1186/s13321-020-00441-8

Ash, R. J., Kuenemann, M. A., Rotroff, D., Motsinger-Reif, A., and Fourches, D. (2019). Cheminformatics approach to exploring and modeling trait-associated metabolite profiles. *J. Cheminformatics* 11 (43), 43. doi:10.1186/s13321-019-0366-3

Ashtawy, H., Anderson, B., Sorenson, J., and Wallach, I. (2021). Atomwise. [Online] Available at: https://blog.atomwise.com/pretraining-graph-neural-networks-on-ultra-large-chemical-libraries-to-learn-generalizable-admet-predictors.

ATOM (2023). Accelerating therapeutics for opportunities in medicine (ATOM). [Online] Available at: https://atomscience.org/.

Avram, S., Funar-Timofei, S., Borota, A., Chennamaneni, S. R., Manchala, A. K., and Muresan, S. (2014). Quantitative estimation of pesticide-likeness for agrochemical discovery. *J. Cheminformatics* 6, 42. doi:10.1186/s13321-014-0042-6

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., et al. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science* 373 (6557), 871–876. doi:10.1126/science.abj8754

Barlow, S., Chesson, A., Collins, J. D., Flynn, A., Hardy, A., Jany, K.-D., et al. (2009). Opinion of the Scientific Committee on a request from EFSA on existing approaches incorporating replacement, reduction and refinement of animal testing: applicability in food and feed risk assessment. *EFSA J.* 1052, 1–77. doi:10.2903/j.efsa.2009.1052

Baskin, I. I., Winkler, D., and Terko, I. V. (2016). A renaissance of neural networks in drug discovery. *Expert Opin. Drug Discov.* 11 (8), 785–795. doi:10.1080/17460441.2016.1201262

Begam, B., and Satheesh Kumar, J. (2012). A study on cheminformatics and its applications on modern drug discovery. *Procedia Eng.* 38, 1264–1275. doi:10.1016/j.proeng.2012.06.156

Belfield, S. J., Cronin, M. T., Enoch, S. J. F. J. W., and Firman, J. W. (2023). Guidance for good practice in the application of machine learning in development of toxicological quantitative structure-activity relationships (QSARs). *PLOS ONE* 18, e0282924. doi:10.1371/journal.pone.0282924

Belsky, J., and Joshi, N. K. (2019). Impact of biotic and abiotic stressors on managed and feral bees. *Insects* 10 (8), 233. doi:10.3390/insects10080233

Bender, A., and Cortés-Ciriano, I. (2021a). Artificial intelligence in drug discovery: what is realistic, what are illusions? Part 1: ways to make an impact, and why we are not there yet. *Drug Discov. Today* 26 (2), 511–524. doi:10.1016/j.drudis.2020.12.009

Bender, A., and Cortés-Ciriano, I. (2021b). Artificial intelligence in drug discovery: what is realistic, what are illusions? Part 2: a discussion of chemical and biological data. *Drug Discov. Today* 26 (4), 1040–1052. doi:10.1016/j.drudis.2020.11.037

Benfenati, E., Diaza, R. G., Cassano, A., Pardoe, S., Gini, G., Mays, C., et al. (2011). The acceptance of in silicomodels for REACH: requirements, barriers, and perspectives. *Chem. Central J.* 5, 58. doi:10.1186/1752-153x-5-58

Benfenati, E., Roncaglioni, A., Raitano, G., and Vian, M. (2019). *In silico* model for mutagenicity (Ames test), taking into account metabolism. *Mutagenesis* 34 (1), 41–48. doi:10.1093/mutage/gey045

Benhenda, M. (2017). *ChemGAN challenge for drug discovery: can AI reproduce natural chemical diversity? arXiv preprint arXiv:1708.08227.*

Beniddir, M. A., Kang, B., Genta-Jouve, G., Huber, F., Rogers, S., van der Hooft, J., et al. (2021). Advances in decomposing complex metabolite mixtures using substructure- and network-based computational metabolomics approaches. *Nat. Prod. Rep.* 38, 1967–1993. doi:10.1039/d1np00023c

Benjamin, S.-L., Carlos, O., Gabriel, L. G., and Alan, A.-G. (2017). *Optimizing distributions over molecular space. An objective-reinforced generative adversarial network for inverse-design chemistry (ORGANIC).* s.l.:s.n.

Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., et al. (2008). KNIME: the konstanz information miner. *s.l.* Springer.

BioSolveIT (2023a). FTrees. [Online] Available at: https://www.biosolveit.de/products/#FTrees (Accessed, 2023).

BioSolveIT (2023b). InfiniSee. [Online] Available at: https://www.biosolveit.de/infiniSee (Accessed, 2023).

Bjerrum, E. J. (2017). *SMILES enumeration as data augmentation for neural network modeling of molecules. arXiv preprint arXiv:1703.07076.*

Blanchard, A. E., Standley, C., and Bhowmik, D. (2021). Using GANs with adaptive training data to search for new molecules. *J. Cheminform* 13, 14. Article No. 14. doi:10.1186/s13321-021-00494-3

Blaschke, T., Arús-Pous, J., Chen, H., Margreitter, C., Tyrchan, C., Engkvist, O., et al. (2020). REINVENT 2.0: an AI tool for *de novo* drug design. *J. Chem. Inf. Model.* 60, 5918–5922. doi:10.1021/acs.jcim.0c00915

Blaschke, T., Olivecrona, M., Engkvist, O., Bajorath, J., and Chen, H. (2018). Application of generative autoencoder in *de novo* molecular design. *Mol. Inf.* 37, 1700123. doi:10.1002/minf.201700123

Bøgevig, A., Federsel, H. J., Huerta, F., Hutchings, M. G., Kraut, H., Langer, T., et al. (2015). Route design in the 21st century: the ICSYNTH software tool as an idea generator for synthesis prediction. *Org. Process Res. Dev.* 19 (2), 357–368. doi:10.1021/op500373e

Bonini, P., Kind, T., Tsugawa, H., Barupal, D. K., and Fiehn, O. (2020). Retip: retention time prediction for compound annotation in untargeted metabolomics. *Anal. Chem.* 92 (11), 7515–7522. doi:10.1021/acs.analchem.9b05765

Borba, J. V. B., Braga, R. C., Alves, V. M., Muratov, E. N., Kleinstreuer, N., Tropsha, A., et al. (2021). Pred-skin: a web portal for accurate prediction of human skin sensitizers. *Chem. Res. Toxicol.* 34 (2), 258–267. doi:10.1021/acs.chemrestox.0c00186

Born, J., Manica, M., Cadow, J., Markert, G., Mill, N. A., Filipavicius, M., et al. (2021). Data-driven molecular design for discovery and synthesis of novel ligands: a case study on SARS-CoV-2. *Mach. Learn. Sci. Technol.* 2, 025024. doi:10.1088/2632-2153/abe808

Brown, N., Fiscato, M., Segler, M., and Vaucher, A. (2019). GuacaMol: benchmarking models for *de novo* molecular design. *J. Chem. Inf. Model.* 59 (3), 1096–1108. doi:10.1021/acs.jcim.8b00839

Capecchi, A., Probst, D., and Reymond, J.-L. (2020). One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. *J. Cheminformatics* 12, 43. doi:10.1186/s13321-020-00445-4

Carnesecchi, E., Toma, C., Roncaglioni, A., Kramer, N., Benfenati, E., and Dorne, J. L. C. (2020a). Integrating QSAR models predicting acute contact toxicity and mode of action profiling in honey bees (*A. mellifera*): data curation using open source databases, performance testing and validation. *Sci. Total Environ.* 735, 139243. doi:10.1016/j.scitotenv.2020.139243

Carnesecchi, E., Toropov, A. A., Toropova, A. P., Kramer, N., Svendsen, C., Dorne, J. L., et al. (2020b). Predicting acute contact toxicity of organic binary mixtures in honey bees (*A. mellifera*) through innovative QSAR models. *Sci. Total Environ.* 704, 135302. doi:10.1016/j.scitotenv.2019.135302

CAS (2023). *Cas SCIFINDER^n*. [Online] Available at: https://www.cas.org/solutions/cas-scifinder-discovery-platform/cas-scifinder.

Casida, J. E. (2012). The greening of pesticide–environment interactions: some personal observations. *Environ. Health Perspect.* 120 (4), 487–493. doi:10.1289/ehp.1104405

Castillo, A. M., Bernal, A., Dieden, R., Patiny, L., and Wist, J. (2016). "Ask Ernö": a self-learning tool for assignment and prediction of nuclear magnetic resonance spectra. *J. Cheminformatics* 8 (26), 26. doi:10.1186/s13321-016-0134-6

Chemspace, (2023). Chemspace. [Online] Available at: https://chem-space.com/ (Accessed, 2023).

Chen, H., Engkvist, O., Wang, Y., Olivecrona, M., and Blaschke, T. (2018). The rise of deep learning in drug discovery. *Drug Discov. Today* 23, 1241–1250. doi:10.1016/j.drudis.2018.01.039

Chen, J., Zheng, S., Song, Y., Rao, J., and Yang, Y. (2021). *Learning attributed graph representations with communicative message passing transformer. arXiv.*

Chen, W. L. (2006). Chemoinformatics: past, present, and future. *J. Chem. Inf. Model.* 46, 2230–2255. doi:10.1021/ci060016u

Chen, Y., and Kirchmair, J. (2020). Cheminformatics in natural product-based drug discovery. *Mol. Inf.* 39, e2000171. doi:10.1002/minf.202000171

Chen, Z., Min, M. R., Parthasarathy, S., and Ning, X. (2020). *Molecule optimization via fragment-based generative models. arXiv preprint arXiv:2012.04231.*

Chen-Yang, J., Wang, F., Hao, G.-F., and Yang, G.-F. (2019). InsectiPAD: a web tool dedicated to exploring physicochemical properties and evaluating insecticide-likeness of small molecules. *J. Chem. Inf. Model.* 59 (2), 630–635. doi:10.1021/acs.jcim.8b00843

Chevillard, F., Rimmer, H., Betti, C., Pardon, E., Ballet, S., van Hilten, N., et al. (2018). Binding-site compatible fragment growing applied to the design of β2-adrenergic receptor ligands. *J. Med. Chem.* 61, 1118–1129. doi:10.1021/acs.jmedchem.7b01558

Choi, J., Seo, S., and Park, S. (2023). COMA: efficient structure-constrained molecular generation using contractive and margin losses. *J. Cheminformatics* 15 (8), 8. doi:10.1186/s13321-023-00679-y

Chuang, K. V., Gunsalus, L. M., and Keiser, M. J. (2020). Learning molecular representations for medicinal chemistry: miniperspective. *J. Med. Chem.* 63, 8705–8722. doi:10.1021/acs.jmedchem.0c00385

Clark, R. D. (2018). Predicting mammalian metabolism and toxicity of pesticides *in silico*. *Pest Manag. Sci.* 74 (9), 1992–2003. doi:10.1002/ps.4935

Cockroft, N. T., Cheng, X., and Fuchs, J. R. (2019). STarFish: a stacked ensemble target fishing approach and its application to natural products. *Chem. Inf. Model.* 59, 4906–4920. doi:10.1021/acs.jcim.9b00489

Coley, C. W., Barzilay, R., Jaakkola, T. S., Green, W. H., and Jensen, K. F. (2017). Prediction of organic reaction outcomes using machine learning. *ACS Central Sci.* 3 (5), 434–443. doi:10.1021/acscentsci.7b00064

Coley, C. W., Thomas, D. A., Lummiss, J. A. M., Jaworski, J. N., Breen, C. P., Schultz, V., et al. (2019). A robotic platform for flow synthesis of organic compounds informed by AI planning. *Science* 365 (6453), eaax1566. doi:10.1126/science.aax1566

Cook, A., Johnson, A. P., Law, J., Mirzazadeh, M., Ravitz, O., and Simon, A. (2012). Computer-aided synthesis design: 40 years on. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 2 (1), 79–107. doi:10.1002/wcms.61

Corey, E., and Wipke, W. (1969). Computer-assisted DEsign of complex organic syntheses. *Science* 166 (3902), 178–192. doi:10.1126/science.166.3902.178

Cortes-Ciriano, I., and Bender, A. (2015). Improved chemical structure–activity modeling through data augmentation. *J. Chem. Inf. Model.* 55 (12), 2682–2692. doi:10.1021/acs.jcim.5b00570

Cortés-Ciriano, I., and Bender, A. (2019). Deep confidence: a computationally efficient framework for calculating reliable prediction errors for deep neural networks. *J. Chem. Inf. Model.* 59 (3), 1269–1281. doi:10.1021/acs.jcim.8b00542

Cresset (2023). [Online] Available at: http://www.cresset-group.com/spark.

Cruz, R. M., Sabourin, R., and Cavalcanti, G. D. (2018). Dynamic classifier selection: recent advances and perspectives. *Inf. Fusion* 41, 195–216. doi:10.1016/j.inffus.2017.09.010

Cruz, R. M. O., Hafemann, L. G., Sabouri, R., and Cavalcanti, G. D. C. (2020). DESlib: a Dynamic ensemble selection library in Python. *J. Mach. Learn. Res.* 21 (8), 1–5.

Dai, H., Li, C., Coley, C. W., Dai, B., and Song, L. (2020). Retrosynthesis prediction with conditional graph logic network. *Adv. Neural Inf. Process. Syst.* Vol. 32.

da Silva, R. R., Dorrestein, P. C., and Quinn, R. A. (2015). Illuminating the dark matter in metabolomics. *Proc. Natl. Acad. Sci. U. S. A.* 112 (41), 12549–12550. doi:10.1073/pnas.1516878112

Dask (2023). Dask. [Online] Available at: https://www.dask.org/.

Dassault Systèmes, S. E. (2023). [Online] Available at: https://www.3ds.com/products-services/biovia/.

David, L., Thakka, A., Mercado, R., and Engkvist, O. (2020). Molecular representations in AI-driven drug discovery: a review and practical guide. *J. Cheminformatics* 12, 56. doi:10.1186/s13321-020-00460-5

Davronov, R., and Adilova, F. (2021). A comparative analysis of the ensemble methods for drug design. *AIP Conf. Proc.* 2365. doi:10.1063/5.0057487

Daylight Chemical Information Systems, Inc. (2019a). SMARTS: a language for describing molecular patterns. [Online] Available at: https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html (Accessed June, 2021).

Daylight Chemical Information Systems, Inc. (2019b). *Smirks - a reaction transform language*. [Online] Available at: https://www.daylight.com/dayhtml/doc/theory/theory.smirks.html (Accessed June, 2021).

Dearden, J. C. (2016). The history and development of quantitative structu re-activity relationships (QSARs). *Int. J. Quantitative Structure-Property Relat.* 1 (1), 1–44. doi:10.4018/ijqspr.2016010101

de Bruyn Kops, C., Šícho, M., Mazzolari, A., and Kirchmair, J. (2021). GLORYx: prediction of the metabolites resulting from phase 1 and phase 2 biotransformations of xenobiotics. *Chem. Res. Toxicol.* 34 (2), 286–299. doi:10.1021/acs.chemrestox.0c00224

De Souza, L. P., Alseekh, S., Brotman, Y., and Fernie, A. R. (2020). Network-based strategies in metabolomics data analysis and interpretation: from molecular networking to biological interpretation. *Expert Rev. Proteomics* 17 (4), 243–255. doi:10.1080/14789450.2020.1766975

Devillers, J., Bro, E., and Millot, F. (2015). Prediction of the endocrine disruption profile of pesticides. *SAR QSAR Environ. Res.* 26 (10), 831–852. doi:10.1080/1062936x.2015.1104809

Dhamercherla, S., Jadav, S. S., Babu, M. C., and Ahsan, M. J. (2022). Machine learning in drug discovery: a review. *Artif. Intell. Rev.* 55, 1947–1999. doi:10.1007/s10462-021-10058-4

Diéguez-Santana, K., Nachimba-Mayanchi, M. M., Puris, A., Gutiérrez, R. T., and González-Díaz, H. (2022). Prediction of acute toxicity of pesticides for Americamysis bahia using linear and nonlinear QSTR modelling approaches. *Environ. Res.* 214, 113984. doi:10.1016/j.envres.2022.113984

Dillard, L. (2021). *Self-supervised learning for molecular property prediction*. ChemRxiv.

Djoumbou-Feunang, Y., Eisner, R., Knox, C., Chepelev, L., Hastings, J., Owen, G., et al. (2016). ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. *J. Cheminformatics* 8 (61), 61. doi:10.1186/s13321-016-0174-y

Djoumbou-Feunang, Y., Fiamoncini, J., Gil-de-la-Fuente, A., Greiner, R., Manach, C., and Wishart, D. S. (2019a). BioTransformer: a comprehensive computational tool for small molecule metabolism prediction and metabolite identification. *J. Cheminformatics* 11, 2. doi:10.1186/s13321-018-0324-5

Djoumbou-Feunang, Y., Pon, A., Karu, N., Zheng, J., Li, C., Arndt, D., et al. (2019b). CFM-ID 3.0: significantly improved ESI-MS/MS prediction and compound identification. *Metabolites* 9 (4), 72. doi:10.3390/metabo9040072

Docker, (2023). Docker. [Online] Available at: https://www.docker.com/.

Dührkop, K., Fleischauer, M., Ludwig, M., Aksenov, A. A., Melnik, A. V., Meusel, M., et al. (2019). SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nat. Methods* 16, 299–302. doi:10.1038/s41592-019-0344-8

Dührkop, K., Nothias, L. F., Fleischauer, M., Reher, R., Ludwig, M., Hoffmann, M. A., et al. (2021). Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. *Nat. Biotechnol.* 39, 462–471. doi:10.1038/s41587-020-0740-8

du Rand, E. E., Smit, S., Beukes, M., Apostolides, Z., Pirk, C. W., and Nicolson, S. W. (2015). Detoxification mechanisms of honey bees (*Apis mellifera*) resulting in tolerance of dietary nicotine. *Sci. Rep.* 5, 11779. doi:10.1038/srep11779

Durant, J. L., Leland, B. A., Henry, D. R., and Nourse, J. G. (2002). Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.* 42 (6), 1273–1280. doi:10.1021/ci010132r

Durrant, J. D., Amaro, R. E., and McCammon, J. A. (2009). AutoGrow: a novel algorithm for protein inhibitor design. *Chem. Biol. drug Des.* 73, 168–178. doi:10.1111/j.1747-0285.2008.00761.x

DynamiCROP (2023). DynamiCROP. [Online] Available at: https://www.dynamicrop.org/ (Accessed, 2023).

Accessed Eli Lilly & Co (2019). LillyMol public code. [Online]Available at: https://github.com/EliLillyCo/LillyMol, 2021).

Accessed Elmer, P. (2023). Lead discovery premium. [Online]Available at: https://perkinelmerinformatics.com/products/research/lead-discovery-premium/, 2023).

Elsevier, (2023). Reaxys. [Online] Available at: https://www.elsevier.com/solutions/reaxys.

Elton, D. C., Boukouvalas, Z., Fuge, M. D., and Chung, P. W. (2019). Deep learning for molecular design—a review of the state of the art. *Mol. Syst. Des. Eng.* 4, 828–849. doi:10.1039/c9me00039a

eMolecules (2023). eMolecules. [Online] Available at: https://www.emolecules.com/ (Accessed, 2023)

Engel, T. (2006). Basic overview of chemoinformatics. *J. Chem. Inf. Model.* 46 (6), 2267–2277. doi:10.1021/ci600234z

European Food; Safety Agency (2023). European food and safety agency. [Online] Available at: https://www.efsa.europa.eu/en (Accessed, 2023).

Feinberg, E. N., Joshi, E., Pande, V. S., and Cheng, A. C. (2020). Improvement in ADMET prediction with multitask deep featurization. *J. Med. Chem.* 63 (16), 8835–8848. doi:10.1021/acs.jmedchem.9b02187

Feinberg, E. N., Sur, D., Wu, Z., Husic, B. E., Mai, H., Li, Y., et al. (2018). Potential net for molecular property prediction. *ACS Cent. Sci.* 4, 1520–1530. doi:10.1021/acscentsci.8b00507

Fernández-Llaneza, D., Ulander, S., Gogishvili, D., Nittinger, E., Zhao, H., and Tyrchan, C. (2021). Siamese recurrent neural network with a self-attention mechanism for bioactivity prediction. *ACS Omega* 6 (16), 11086–11094. doi:10.1021/acsomega.1c01266

FRAC (2023). FRAC. [Online] Available at: https://www.frac.info/(Accessed, 2023).

Fromer, J. C., and Coley, C. W. (2022). *Computer-aided multi-objective optimization in small molecule discovery. arXiv.*

Gandy, M. N., Corral, M. G., Mylne, J. S., and Stubbs, K. A. (2015). An interactive database to explore herbicide physicochemical properties. *Org. Biomol. Chem.* 13 (20), 5586–5590. doi:10.1039/c5ob00469a

Gao, H., Struble, T. J., Coley, C. W., Wang, Y., Green, W. H., and Jensen, K. F. (2018). Using machine learning to predict suitable conditions for organic reactions. *ACS Central Sci.* 4 (11), 1465–1476. doi:10.1021/acscentsci.8b00357

Gao, W., and Coley, C. W. (2020). The synthesizability of molecules proposed by generative models. *J. Chem. Inf. Model.* 60, 5714–5723. doi:10.1021/acs.jcim.0c00174

Gao, W., Fu, T., Sun, J., and Coley, C. W. (2022). *Sample efficiency matters: a benchmark for practical molecular optimization. arxiv.*

Gaulton, A., Hersey, A., Nowotka, M., Bento, A. P., Chambers, J., Mendez, D., et al. (2017). The ChEMBL database in 2017. *Nucleic Acids Res.* 45 (D1), D945–D954. doi:10.1093/nar/gkw1074

Gawlikowski, J., Njieutcheu, C. R. T., Ali, M., Lee, J., Humt, M., Feng, J., et al. (2021). *A survey of uncertainty in deep neural networks. arXiv.*

Gentile, F., Agrawal, V., Hsing, M., Ton, A. T., Ban, F., Norinder, U., et al. (2020). Deep docking: a deep learning platform for augmentation of structure based drug discovery. *ACS Central Sci.* 6 (6), 939–949. doi:10.1021/acscentsci.0c00229

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. (2017). *Neural message passing for quantum chemistry.* s.l., s.n., pp. 1263-1272.

Goh, G. B., Hodas, N. O., Siegel, C., and Vishnu, A. (2017). *SMILES2Vec: an interpretable general-purpose deep neural network for predicting chemical properties. arXiv.*

Gonçalves, I. L., Machado das Neves, G., Porto Kagami, L., Eifler-Lima, V. L., and Merlo, A. A. (2021). Discovery, development, chemical diversity and design of isoxazoline-based insecticides. *Bioorg. Med. Chem.* 30, 115934. doi:10.1016/j.bmc.2020.115934

Goodarzi, M., Dejaegher, B., and Heyden, Y. V. (2012). Feature selection methods in QSAR studies. *J. AOAC Int.* 95 (3), 636–651. doi:10.5740/jaoacint.sge_goodarzi

Goodman, J. M., Pletnev, I., Thiessen, P., Bolton, E., and Heller, S. R. (2021). InChI version 1.06: now more than 99.99% reliable. *J. Cheminformatics* 13 (1), 40. doi:10.1186/s13321-021-00517-z

Grechishnikova, D. (2021). Transformer neural network for protein-specific *de novo* drug generation as a machine translation problem. *Sci. Rep.* 11 (1), 321. doi:10.1038/s41598-020-79682-4

Green, J., Cabrera Diaz, C., Jakobs, M. A. H., Dimitracopoulos, A., van der Wilk, M., and Greenhalgh, R. D. (2023). *Current methods for drug property prediction in the real world. arxiv.*

Grygorenko, O. O., Radchenko, D. S., Dziuba, I., Chuprina, A., Gubina, K. E., and Moroz, Y. S. (2020). Generating multibillion chemical space of readily accessible screening compounds. *iScience* 23 (11), 101681. doi:10.1016/j.isci.2020.101681

Grzybowski, B. A., Szymkuć, S., Gajewska, E. P., Molga, K., Dittwald, P., Wołos, A., et al. (2018). Chematica: a story of computer code that started to think like a chemist. *Chem* 4, 390–398. doi:10.1016/j.chempr.2018.02.024

Guijas, C., Montenegro-Burke, J. R., Domingo-Almenara, X., Palermo, A., Warth, B., Hermann, G., et al. (2018). METLIN: a technology platform for identifying knowns and unknowns. *Anal. Chem.* 90 (5), 3156–3164. doi:10.1021/acs.analchem.7b04424

Guo, Z., Zhang, C., Yu, W., Herr, J., Wiest, O., Jiang, M., et al. (2021). *Few-shot graph learning for molecular property prediction. arXiv.*

Han, L., Suzek, T. O., Wang, Y., and Bryant, S. H. (2010). The Text-mining based PubChem Bioassay neighboring analysis. *BMC Bioinforma.* 11 (549), 549. doi:10.1186/1471-2105-11-549

Hao, G., Dong, Q., and Yang, G. (2011). A comparative study on the constitutive properties of marketed pesticides. *Mol. Inf.* 30 (6-7), 614–622. doi:10.1002/minf.201100020

Hasebe, T. (2021). Knowledge-Embedded message-passing neural networks: improving molecular property prediction with human knowledge. *ACS Omega* 6, 27955–27967. doi:10.1021/acsomega.1c03839

Hawkins, N. J., Bass, C., Dixon, A., and Neve, P. (2019). The evolutionary origins of pesticide resistance. *Biol. Rev. Camb. Philosophical Soc.* 94 (1), 135–155. doi:10.1111/brv.12440

Haywood, A. L., Redshaw, J., Hanson-Heine, M. W. D., Taylor, A., Brown, A., Mason, A. M., et al. (2021). Kernel methods for predicting yields of chemical reactions. *J. Chem. Inf. Model.* 62, 2077–2092. doi:10.1021/acs.jcim.1c00699

He, J., You, H., Sandström, E., Nittinger, E., Bjerrum, E. J., Tyrchan, C., et al. (2021). Molecular optimization by capturing chemist's intuition using deep neural networks. *J. Cheminformatics* 13, 26. doi:10.1186/s13321-021-00497-0

Hefke, L., Hiesinger, K., Zhu, W. F., Kramer, J. S., and Proschak, E. (2020). Computer-Aided fragment growing strategies to design dual inhibitors of soluble epoxide hydrolase and LTA4 hydrolase. *ACS Med. Chem. Lett.* 11, 1244–1249. doi:10.1021/acsmedchemlett.0c00102

Heid, E., Greenman, K. P., Chung, Y., Li, S-C., Graff, D. E., Vermeire, F. H., et al. (2023). *Chemprop: machine learning package for chemical property prediction.* ChemRxiv.

Heinemann, J. (2019). Machine learning in untargeted metabolomics experiments. *Methods Microb. Biol.* 1859, 287–299. doi:10.1007/978-1-4939-8757-3_17

Hekkelman, M. L., de Vries, I., Joosten, R. P., and Perrakis, A. (2023). AlphaFill: enriching the AlphaFold models with ligands and co-factors. *Nat. Methods* 20, 205–213. doi:10.1038/s41592-022-01685-y

Henry, J., and Wlodkowic, D. (2020). High-throughput animal tracking in chemobehavioral phenotyping: current limitations and future perspectives. *Behav. Process.* 180, 104226. doi:10.1016/j.beproc.2020.104226

Honda, S., Shi, S., and Ueda, H. R. (2019). *SMILES transformer: pre-trained molecular fingerprint for low data drug discovery. arXiv preprint arXiv:1911.04738.*

Hong, Y., Li, S., Welch, C. J., Tichy, S., Ye, Y., and Tang, H. (2023). 3DMolMS: prediction of tandem mass spectra from 3D molecular conformations. *Bioinformatics* 39 (6), btad354. doi:10.1093/bioinformatics/btad354

HRAC (2023). HRAC. [Online] Available at: https://www.hracglobal.com/ (Accessed, 2023).

Huang, J.-J., Wang, F., Ouyang, Y., Huang, Y., Jia, C., Zhong, H., et al. (2020). HerbiPAD: a free web platform to comprehensively analyze constitutive property and herbicide-likeness to estimate chemical bioavailability. *Pest Manag. Sci.* 77 (3), 1273–1281. doi:10.1002/ps.6140

Huang, R. X. M. N. D. T., Xia, M., Nguyen, D. T., Zhao, T., Sakamuru, S., Zhao, J., et al. (2016). Tox21Challenge to build predictive models of nuclear receptor and stress response pathways as mediated by exposure to environmental chemicals and drugs. *Front. Environ. Sci.* 3. doi:10.3389/fenvs.2015.00085

Humer, C., Heberle, H., Montanari, F., Wolf, T., Huber, F., Henderson, R., et al. (2022). ChemInformatics Model Explorer (CIME): exploratory analysis of chemical model explanations. *J. Cheminformatics* 14 (21), 21. doi:10.1186/s13321-022-00600-z

Hung, N., and Chang, J. M. (2021). *Complementary ensemble learning. arXiv.*

Ibarra-Estrada, E., Soto-Hernández, R. M., and Palma-Tenango, M. (2016). "Metabolomics as a tool in agriculture," in Metabolomics: Fundamentals and Applications. *s.l.* (London, United Kingdom: IntechOpen Limited).

IBM (2023). *IBM RoboRXN.* [Online] Available at: https://rxn.res.ibm.com/rxn/robo-rxn/welcome (Accessed, 2023).

Idakwo, G., Thangapandian, S., Luttrell, J., Li, Y., Wang, N., Zhou, Z., et al. (2020). Structure–activity relationship-based chemical classification of highly imbalanced Tox21 datasets. *J. Cheminformatics* 12, 66. doi:10.1186/s13321-020-00468-x

Iktos (2023a). Makya. [Online]Available at: https://iktos.ai/makya/.

Iktos (2023b). Spaya. [Online]Available at: https://iktos.ai/spaya/.

IRAC (2023). IRAC. [Online]Available at: https://irac-online.org/ (Accessed, 2023).

Irwin, J. J., and Stoichet, B. K. (2005). ZINC - a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* 45 (1), 177–182. doi:10.1021/ci049714+

Irwin, R., Dimitriadis, S., He, J., and Bjerrum, E. J. (2022). Chemformer: a pre-trained transformer for computational chemistry. *Mach. Learn. Sci. Technol.* 3 (1), 015022. doi:10.1088/2632-2153/ac3ffb

Jaeger, S., Fulle, S., and Turk, S. (2018). Mol2vec: unsupervised machine learning approach with chemical intuition. *J. Chem. Inf. Model.* 58 (1), 27–35. doi:10.1021/acs.jcim.7b00616

Janssen, A. P. A., Grimm, S. H., Wijdeven, R. H. M., Lenselink, E. B., Neefjes, J., van Boeckel, C. A. A., et al. (2019). Drug discovery maps, a machine learning model that visualizes and predicts kinome-inhibitor interaction landscapes. *J. Chem. Inf. Model* 59, 1221–1229. doi:10.1021/acs.jcim.8b00640

Jeschke, P. (2016). Propesticides and their use as agrochemicals. *Pest Manag. Sci.* 72, 210–225. doi:10.1002/ps.4170

Jiang, D., Wu, Z., Hsieh, C. Y., Chen, G., Liao, B., Wang, Z., et al. (2021). Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models. *J. Cheminformatics* 13, 12. doi:10.1186/s13321-020-00479-8

Jie, Z., Rocío, M., Ola, E., and Hongming, C. (2021). Comparative study of deep generative models on chemical space coverage, v18. s.n: s.l.

Jiménez-Luna, J., Cuzzolin, A., Bolcato, G., Sturlese, M., and Moro, S. (2020a). A deep-learning approach toward rational molecular docking protocol selection. *Molecules* 25 (11), 2487. doi:10.3390/molecules25112487

Jiménez-Luna, J., Grisoni, F., and Schneider, G. (2020b). Drug discovery with explainable artificial intelligence. *Nat. Mach. Intell.* 2, 573–584. doi:10.1038/s42256-020-00236-4

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596 (7873), 583–589. doi:10.1038/s41586-021-03819-2

Kang, S., and Cho, K. (2019). Conditional molecular design with deep generative models. *J. Chem. Inf. Model.* 59, 43–52. doi:10.1021/acs.jcim.8b00263

Karlov, D. S., Sosnin, S., Tetko, I., and Fedorov, M. V. (2019). Chemical space exploration guided by deep neural networks. *RSC Adv.* 9, 5151–5157. doi:10.1039/c8ra10182e

Karp, P. D., Midford, P. E., Billington, R., Kothari, A., Krummenacker, M., Latendresse, M., et al. (2019). Pathway Tools version 23.0 update: software for pathway/genome informatics and systems biology. *Briefings Bioinforma.* 22 (1), 109–126. doi:10.1093/bib/bbz104

Karpov, P., Godin, G., and Tetko, I. V. (2020). Transformer-CNN: Swiss knife for QSAR modeling and interpretation. *J. Cheminformatics* 12, 17. doi:10.1186/s13321-020-00423-w

Kavlock, R. J., Bahadori, T., Barton-Maclaren, T. S., Gwinn, M. R., Rasenberg, M., and Thomas, R. S. (2018). Accelerating the pace of chemical risk assessment. *Chem. Res. Toxicol.* 31 (5), 287–290. doi:10.1021/acs.chemrestox.7b00339

Kayala, M. A. A. C.-A. C. J. H., Azencott, C. A., Chen, J. H., and Baldi, P. (2011). Learning to predict chemical reactions. *J. Chem. Inf. Model.* 51, 2209–2222. doi:10.1021/ci200207y

Kearnes, S., Goldman, B., and Pande, V. (2016). *Modeling industrial ADMET data with multitask networks. arXiv preprint arXiv:1606.08793.*

Kearnes, S. M., Maser, M. R., Wleklinski, M., Kast, A., Doyle, A. G., Dreher, S. D., et al. (2021). The open reaction database. *J. Am. Chem. Soc.* 143, 18820–18826. doi:10.1021/jacs.1c09820

KEBOTIX (2023). KEBOTIX. [Online] Available at: https://www.kebotix.com/solutions.

Kell, D. B., Samanta, S., and Swainston, N. (2020). Deep learning and generative methods in cheminformatics and chemical biology: navigating small molecule space intelligently. *Biochem. J.* 477, 4559–4580. doi:10.1042/bcj20200781

Khemchandani, Y., O'Hagan, S., Samanta, S., Swainston, N., Roberts, T. J., Bollegala, D., et al. (2020). DeepGraphMolGen, a multi-objective, computational strategy for generating molecules with desirable properties: a graph convolution and reinforcement learning approach. *J. Cheminformatics* 12, 53. doi:10.1186/s13321-020-00454-3

Kienzler, A., Barron, M. G., Belanger, S. E., Beasley, A., and Embry, M. R. (2017). Mode of action (MOA) assignment classifications for ecotoxicology: an evaluation of approaches. *Environ. Sci. Technol.* 51, 10203–10211. doi:10.1021/acs.est.7b02337

Kim, H., Wang, M., Leber, C. A., Nothias, L. F., Reher, R., Kang, K. B., et al. (2020). *NPClassifier: a deep neural network-based structural classification tool for natural products*. ChemRvix.

Kim, H., Lee, J., Ahn, S., and Jongsuk, R. L. (2021). A merged molecular representation learning for molecular properties prediction with a web-based service. *Sci. Rep.* 11 (11028), 11028. doi:10.1038/s41598-021-90259-7

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., et al. (2021). PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Res.* 49 (D1), D1388–D1395. doi:10.1093/nar/gkaa971

Kirst, H. A. (2010). The spinosyn family of insecticides: realizing the potential of natural products research. *J. antibiotics* 63, 101–111. doi:10.1038/ja.2010.5

Klaise, J., Van Looveren, A., Vacanti, G., and Coca, A. (2021). Alibi explain: algorithms for explaining machine learning models. *J. Mach. Learn. Res.* 22 (181), 1–7.

Klie, S., Schröder, F., and Busch, M. (2022). *Method for screening of a chemical substance*. Germany. Patent No. WO2022128366A1.

Klucznik, T., Mikulak-Klucznik, B., McCormack, M. P., Lima, H., Szymkuć, S., Bhowmick, M., et al. (2018). Efficient syntheses of diverse, medicinally relevant targets planned by computer and executed in the laboratory. *Chemistry* 4 (3), 522–532. doi:10.1016/j.chempr.2018.02.002

Konze, K. D., Bos, P. H., Dahlgren, M. K., Leswing, K., Tubert-Brohman, I., Bortolato, A., et al. (2019). Reaction-based enumeration, active learning, and free energy calculations to rapidly explore synthetically tractable chemical space and optimize potency of cyclin-dependent kinase 2 inhibitors. *J. Chem. Inf. Model.* 59 (9), 3782–3793. doi:10.1021/acs.jcim.9b00367

Korjus, K., Hebart, M. N., and Vicente, R. (2016). An efficient data partitioning to improve classification performance while keeping parameters interpretable. *PLoS One* 11 (8), e0161788. doi:10.1371/journal.pone.0161788

Koutroumpa, N.-M., Papavasileiou, K. D., Papadiamantis, A. G., Melagraki, G., and Afantitis, A. (2023). A systematic review of deep learning methodologies used in the drug discovery process with emphasis on *in vivo* validation. *Int. J. Mol. Sci.* 24 (7), 6573. doi:10.3390/ijms24076573

Krasnov, L., Khokhlov, I., Fedorov, M. V., and Sosnin, S. (2021). Transformer-based artificial neural networks for the conversion between chemical notations. *Sci. Rep.* 11, 14798. Article No. 14798. doi:10.1038/s41598-021-94082-y

Krenn, M., Häse, F., Nigam, A., Friederich, P., and Aspuru-Guzik, A. (2020). Self-referencing embedded strings (SELFIES): a 100% robust molecular string representation. *Mach. Learn. Sci. Technol.* 1 (4), 045024. doi:10.1088/2632-2153/aba947

Krishnan, S. R., Bung, N., Bulusu, G., Roy, A., and Roy, A. (2022). *De novo* structure-based drug design using deep learning. *J. Chem. Inf. Model.* 62 (21), 5100–5109. doi:10.1021/acs.jcim.1c01319

Kubernetes (2023). *Kubernetes*. [Online] Available at: https://kubernetes.io/.

Kuhn, B., Guba, W., Hert, J., Banner, D., Bissantz, C., Ceccarelli, S., et al. (2016). A real-world perspective on molecular design. *J. Med. Chem.* 59, 4087–4102. doi:10.1021/acs.jmedchem.5b01875

Kwon, S., Bae, H., Jo, J., and Yoon, S. (2019). Comprehensive ensemble in QSAR prediction for drug discovery. *BMC Bioinformatic* 21. doi:10.1186/s12859-019-3135-4

Lahm, G. P., Stevenson, T. M., Selby, T. P., Freudenberger, J. H., Cordova, D., Flexner, L., et al. (2007). Rynaxypyr™: a new insecticidal anthranilic diamide that acts as a potent and selective ryanodine receptor activator. *Bioorg. Med. Chem. Lett.* 17 (22), 6274–6279. doi:10.1016/j.bmcl.2007.09.012

Lambard, G., and Gracheva, E. (2020). SMILES-X: autonomous molecular compounds characterization for small datasets without descriptors. *Mach. Learn. Sci. Technol.* 1 (2), 025004. doi:10.1088/2632-2153/ab57f3

Lamberth, C., Jeanmart, S., Luksch, T., and Plant, A. (2013). Current challenges and trends in the discovery of agrochemicals. *Science* 341 (6147), 742–746. doi:10.1126/science.1237227

Langevin, M., Minoux, H., Levesque, M., and Bianciotto, M. (2020). Scaffold-Constrained molecular generation. *J. Chem. Inf. Model.* 60, 5637–5646. doi:10.1021/acs.jcim.0c01015

Lee, M., and Min, K. (2022). A comparative study of the performance for predicting biodegradability classification: the quantitative structure–activity relationship model vs the graph convolutional network. *ACS Omega* 7 (4), 3649–3655. doi:10.1021/acsomega.1c06274

Lee, S., and Barron, M. G. (2016). A mechanism-based 3D-QSAR approach for classification and prediction of acetylcholinesterase inhibitory potency of organophosphate and carbamate analogs. *J. Computer-Aided Mol. Des.* 30, 347–363. doi:10.1007/s10822-016-9910-7

Lewer, J. M., Stickelman, Z. R., Huang, J. H. P. J. F., and Kostal, J. (2022). Structure-to-process design framework for developing safer pesticides. *Sci. Adv.* 8 (13), eabn2058. doi:10.1126/sciadv.abn2058

Lewis, K. A., Tzilivakis, J., Warner, D. J., and Green, A. (2016). An international database for pesticide risk assessments and management. *Hum. Ecol. Risk Assess. Int. J.* 22 (4), 1050–1064. doi:10.1080/10807039.2015.1133242

Li, F., Fan, D., Wang, H., Yang, H., Li, W., Tang, Y., et al. (2017). *In silico* prediction of pesticide aquatic toxicity with chemical category approaches. *Toxicol. Res.* 6, 831–842. doi:10.1039/c7tx00144d

Li, M., and Wang, X. (2019). Peak alignment of gas chromatography–mass spectrometry data with deep learning. *J. Chromatogr. A* 1604, 460476. doi:10.1016/j.chroma.2019.460476

Li, M., Zhou, J., Hu, J., Fan, W., Zhang, Y., Gu, Y., et al. (2021). DGL-LifeSci: an open-source toolkit for deep learning on graphs in life science. *ACS Omega* 6, 27233–27238. doi:10.1021/acsomega.1c04017

Li, P., Wang, J., Qiao, Y., Chen, H., Yu, Y., Yao, X., et al. (2021). An effective self-supervised framework for learning expressive molecular global representations to drug discovery. *Briefings Bioformatics* 22 (6), bbab109. doi:10.1093/bib/bbab109

Li, X., and Fourches, D. (2020). Inductive transfer learning for molecular activity prediction: next-Gen QSAR Models with MolPMoFiT. *J. Cheminformatics* 12 (1), 27. doi:10.1186/s13321-020-00430-x

Li, Y., Xu, Y., and Yu, Y. (2021). CRNNTL: convolutional recurrent neural network and transfer learning for QSAR modeling in organic drug and material discovery. *Molecules* 26 (23), 7257. doi:10.3390/molecules26237257

Liebal, U. W., Phan, A. N. T., Sudhakar, M., Raman, K., and Blank, L. M. (2020). Machine learning applications for mass spectrometry-based metabolomics. *Metabolites* 10 (6), 243. doi:10.3390/metabo10060243

Litsa, E. E., Das, P., and Kavraki, L. E. (2020). Prediction of drug metabolites using neural machine translation. *Chem. Sci.* 11 (47), 12777–12788. doi:10.1039/d0sc02639e

Liu, R., Glover, K. P., Feasel, M. G., and Wallqvist, A. (2018a). General approach to estimate error bars for quantitative structure-activity relationship predictions of molecular activity. *J. Chem. Inf. Model.* 58 (8), 1561–1575. doi:10.1021/acs.jcim.8b00114

Liu, R., Madore, M., Glover, K. P., Feasel, M. G., and Wallqvist, A. (2018b). Assessing deep and shallow learning methods for quantitative prediction of acute chemical toxicity. *Assess. Deep Shallow Learn. Methods Quantitative Predict. Acute Chem. Toxic.* 164 (2), 512–526. doi:10.1093/toxsci/kfy111

Liu, R., and Wallqvist, A. (2019). Molecular similarity-based domain applicability metric efficiently identifies out-of-domain compounds. *J. Chem. Inf. Model.* 59 (1), 181–189. doi:10.1021/acs.jcim.8b00597

Liu, Y., Mrzic, A., Meysman, P., De Vijlder, T., Romijn, E. P., Valkenborg, D., et al. (2020). MESSAR: automated recommendation of metabolite substructures from tandem mass spectra. *PLoS ONE* 15 (1), e0226770. doi:10.1371/journal.pone.0226770

Lo, Y., Ren, G., Honda, H., and Davis, K. (2019). "Artificial intelligence-based drug design and discovery," in ChemInformatics and its Applications. *s.l.* (IntechOpen.

Lo, Y.-C., Rensi, S. E., Torng, W., and Altman, R. B. (2018). Machine learning in chemoinformatics and drug discovery. *Drug Discov. Today* 23, 1538–1546. doi:10.1016/j.drudis.2018.05.010

Lorsbach, B. A., Sparks, T. C., Cicchillo, R. M., Garizi, N. V., Hahn, D. R., and Meyer, K. G. (2019). Natural products: a strategic lead generation approach in crop protection discovery. *Pest Manag. Sci.* 75 (9), 2301–2309. doi:10.1002/ps.5350

Loso, M. R., Garizi, N., Hegde, V. B., Hunter, J. E., and Sparks, T. C. (2017). Lead generation in crop protection research: a portfolio approach to agrochemical discovery. *Pest Manag. Sci.* 73 (4), 678–685. doi:10.1002/ps.4336

Lounkine, E., Wawer, M., Wassermann, A. M., and Bajorath, J. (2010). SARANEA: a freely available program to mine Structure–Activity and Structure–Selectivity relationship information in compound data sets. *J. Chem. Inf. Model* 50, 68–78. doi:10.1021/ci900416a

Lundberg, S., and Lee, S.-I. (2017). *A unified approach to interpreting model predictions. arXiv*.

Ma, J., Sheridan, R. P., Liaw, A., Dahl, G. E., and Svetnik, V. (2015). Deep neural nets as a method for quantitative structure-activity relationships. *J. Chem. Inf. Model.* 55 (2), 263–274. doi:10.1021/ci500747n

Mansouri, K., Grulke, C. M., Judson, R. S., and Williams, A. J. (2018). OPERA models for predicting physicochemical properties and environmental fate endpoints. *J. Cheminformatics* 10 (1), 10. doi:10.1186/s13321-018-0263-1

Mao, J., Akhtar, J., Zhang, X., Sun, L., Guan, S., Li, X., et al. (2021). Comprehensive strategies of machine-learning-based quantitative structure-activity relationship models. *iScience* 24 (9), 103052. doi:10.1016/j.isci.2021.103052

Marcou, G., Aires de Sousa, J., Latino, D. A. R. S., de Luca, A., Horvath, D., Rietsch, V., et al. (2015). Expert system for predicting reaction conditions: the michael reaction case. *Michael React. Case* 55 (2), 239–250. doi:10.1021/ci500698a

Martin, E. J., and Zhu, X.-W. (2021). Collaborative profile-QSAR: a natural platform for building collaborative models among competing companies. *J. Cheminformatics Model.* 61 (4), 1603–1616. doi:10.1021/acs.jcim.0c01342

Martin, T. M., Lilavois, C. R., and Barron, M. G. (2017). Prediction of pesticide acute toxicity using two-dimensional chemical descriptors and target species classification. *SAR QSAR Environ. Res.* 28 (6), 525–539. doi:10.1080/1062936x.2017.1343204

Martinez, J. I., Martinez Alvarado, J. I., Shields, B. J., and Doyle, A. G. (2021). Predicting reaction yields via supervised learning. *Accounts Chem. Res.* 54 (8), 1856–1865. doi:10.1021/acs.accounts.0c00770

Martinez-Mayorga, K., Madariaga-Mazon, A., Medina-Franco, J., and Maggiora, G. (2020). The impact of chemoinformatics on drug discovery in the pharmaceutical industry. *Expert Opin. Drug Discov.* 15 (3), 293–306. doi:10.1080/17460441.2020.1696307

Maser, M. R., Cui, A. Y., Ryou, S., DeLano, T. J., Yue, Y., and Reisman, S. E. (2021). Multilabel classification models for the prediction of cross-coupling reaction conditions. *J. of Chem. Inf. Model.* 61 (1), 156–166. doi:10.1021/acs.jcim.0c01234

Mater, A. C., and Coote, M. L. (2019). Deep learning in chemistry. *J. Chem. Inf. Model.* 59, 2545–2559. doi:10.1021/acs.jcim.9b00266

Matveieva, M., and Polishchuk, P. (2021). Benchmarks for interpretation of QSAR models. *J. Cheminformatics* 13, 41. doi:10.1186/s13321-021-00519-x

Mayr, A., Klambauer, G., Unterthiner, T., and Hochreiter, S. (2016). DeepTox: toxicity prediction using deep learning. *Front. Environ. Sci.* 3. doi:10.3389/fenvs.2015.00080

McDougall, P. (2016). *The cost of agrochemical product discovery, development and registration in 1995.* 2000, 2005-8 and 2010-14, s.l.: s.n.

McInnes, L., Healy, J., Saul, N., and Grosberger, L. (2018). UMAP: Uniform Manifold approximation and projection. *J. Open Source Softw.* 3 (29), 861. doi:10.21105/joss.00861

Mehta, S. (2020). *Massbank of north America (mona): an open-access, autocurating mass spectral database for compound identification in metabolomics*. s.l., s.n.

MELLODDY (2023). *Melloddy*. [Online] Available at: https://www.melloddy.eu/objectives (Accessed, 2023).

Melnikov, A., Tsentalovich, Y., and Yanshole, V. (2020). Deep learning for the precise peak detection in high resolution LC-MS data. *Anal. Chem.* 92 (1), 588–592. doi:10.1021/acs.analchem.9b04811

Méndez-Lucio, O., Baillif, B., Clevert, D. A., Rouquié, D., and Wichard, J. (2020). *De novo* generation of hit-like molecules from gene expression signatures using artificial intelligence. *Nat. Commun.* 11, 10. doi:10.1038/s41467-019-13807-w

Mercado, R., Rastemo, T., Lindelöf, E., Klambauer, G., Engkvist, O., Chen, H., et al. (2021). Graph networks for molecular design. *Mach. Learn. Sci. Technol.* 2, 025023. doi:10.1088/2632-2153/abcf91

Meyer, K. G., Bravo-Altamirano, K., Herrick, J., Loy, B. A., Yao, C., Nugent, B., et al. (2021). Discovery of florylpicoxamid, a mimic of a macrocyclic natural product. *Bioorg. Med. Chem.* 50, 116455. doi:10.1016/j.bmc.2021.116455

Michael, A., Greg, R., David, S. W., and Leonard, J. F. (2021). *Deep generative models enable navigation in sparsely populated chemical space*. s.n: s.l.

Mishra, P., Asaari, M. S. M., Herrero-Langreo, A., Lohumi, S., Diezma, B., and Scheunders, P. (2017). Close range hyperspectral imaging of plants: a review. *Biosyst. Eng.* 164, 49–67. doi:10.1016/j.biosystemseng.2017.09.009

MLPDS (2023). MLPDS. [Online] Available at: https://mlpds.mit.edu/.

Mo, Y., Guan, Y., Verma, P., Guo, J., Fortunato, M. E., Lu, Z., et al. (2021). Evaluating and clustering retrosynthesis pathways with learned strategy. *Chem. Sci.* 12, 1469–1478. doi:10.1039/d0sc05078d

Montanari, F., Kuhnke, L., Laak, A. T., and Clevert, D.-A. (2019). Modeling physico-chemical ADMET endpoints with multitask graph convolutional networks. *Molecules* 25 (1), 44. doi:10.3390/molecules25010044

Moriwaki, H., Tian, Y.-S., Kawashita, N., and Takagi, T. (2018). Mordred: a molecular descriptor calculator. *J. Cheminformatics* 10, 4. doi:10.1186/s13321-018-0258-y

Muratov, E. N., Bajorath, J., Sheridan, R. P., Tetko, I. V., Filimonov, D., Poroikov, V., et al. (2020). QSAR without borders. *Chem. Soc. Rev.* 49, 3525–3564. doi:10.1039/d0cs00098a

Naik, P., Singh, S. T., and Singh, H. (2009). Quantitative structure–activity relationship (QSAR) for insecticides: development of predictive *in vivo* insecticide activity models. *SAR QSAR Environ. Res.* 20 (5-6), 551–566. doi:10.1080/10629360903278735

National Pesticide Infomation Center (2023). *National pesticide infomation center*. [Online] Available at: http://npic.orst.edu/factsheets/spinosadgen.html.

Naveja, J. J., and Vogt, M. (2021). Automatic identification of analogue series from large compound data sets: methods and applications. *Molecules* Vol. 26, 5291. doi:10.3390/molecules26175291

Accessed NextMove Software (2022). LeadMine. [Online] Available at: https://www.nextmovesoftware.com/leadmine.html, 2022).

Nguyen, C. Q., Kreatsoulas, C., and Branson, K. M. (2020). *Meta-learning initializations for low-resource drug discovery*. ChemrXiv [Preprint]. Available at: https://chemrxiv.org/engage/chemrxiv/article-details/60c748d0ee301c1d5cc7997e.

Nicolaou, C., Watson, I. A., and Wang, J. (2016). The proximal lilly collection: mapping, exploring and exploiting feasible chemical space. *J. Chem. Inf. Model.* 56 (7), 1253–1266. doi:10.1021/acs.jcim.6b00173

Nicolau, C. A., Watson, I. A., LeMasters, M., Masquelin, T., and Wang, J. (2020). Context aware data-driven retrosynthetic analysis. *J. of Chem. Inf. Model.* doi:10.1021/acs.jcim.9b01141

Nishimoto, R. (2019). Global trends in the crop protection industry. *J. Pestic. Sci.* 44 (3), 141–147. doi:10.1584/jpestics.d19-101

NIST (2023). NIST20: updates to the NIST tandem and electron ionization spectral libraries. [Online] Available at: https://www.nist.gov/programs-projects/nist20-updates-nist-tandem-and-electron-ionization-spectral-libraries (Accessed June, 2023).

Nori, H., Jenkins, S., Koch, P., and Caruana, R. (2019). *InterpretML: a unified framework for machine learning interpretability. arXiv*.

OASIS (2021). OASIS TIMES. [Online] Available at: http://oasis-lmc.org/products/software/times.aspx [Accessed September 2021].

OECD (2023). OECD. [Online] Available at: https://www.oecd.org/chemicalsafety/risk-assessment/validationofqsarmodels.htm (Accessed, 2023).

Olğaç, A., Orhan, I. E., and Banoglu, E. (2017). *Future medicinal chemistry*.

Olier, I., Sadawi, N., Bickerton, G. R., Vanschoren, J., Grosan, C., Soldatova, L., et al. (2018). Meta-QSAR: a large-scale application of meta-learning to drug design and discovery. *Mach. Learn.* 107, 285–311. doi:10.1007/s10994-017-5685-x

OpenEye Scientific (2023a). FastROCS toolkit. [Online] Available at: https://www.eyesopen.com/molecular-modeling-fastrocs (Accessed, 2023).

OpenEye Scientific (2023b). OpenEye toolkits. [Online] Available at: https://docs.eyesopen.com/toolkits/python/index.html.

Optibrium (2023). StarDrop. [Online] Available at: https://optibrium.com/stardrop/.

Orosz, Á., Héberger, K., and Rácz, A. (2022). Comparison of descriptor- and fingerprint sets in machine learning models for ADME-tox targets. *Front. Chem.* Vol. 10, 852893. doi:10.3389/fchem.2022.852893

Oršolić, D., Pehar, V., Šmuc, T., and Stepanić, V. (2021). Comprehensive machine learning based study of the chemical space of herbicides. *Sci. Rep.* 11, 11479. doi:10.1038/s41598-021-90690-w

Ouyang, Y., Huang, J. j., Wang, Y. L., Zhong, H., Song, B. A., and Hao, G. f. (2021). Silico resources of drug-likeness as a mirror: what are we lacking in pesticide-likeness? J. Agric. Food Chem. *Sept.* 69, 10761–10773. doi:10.1021/acs.jafc.1c01460

Ozdemir, A., and Polat, K. (2020). Deep learning applications for hyperspectral imaging: a systematic review. *J. Inst. Electron. Comput.* 2 (1), 39–56. doi:10.33969/jiec.2020.21004

Pathway Tools (2023). Pathway tools software. [Online] Available at: http://bioinformatics.ai.sri.com/ptools/.

Paul, D., Sanap, G., Shenoy, S., Kalyane, D., Kalia, K., and Tekade, R. K. (2021). Artificial intelligence in drug discovery and development. *Drug Discov. Today* 26, 80–93. doi:10.1016/j.drudis.2020.10.010

Paulus, S., and Mahlein, A.-K. (2020). Technical workflows for hyperspectral plant image assessment and processing on the greenhouse and laboratory scale. *GigaScience* 9 (8), giaa090. doi:10.1093/gigascience/giaa090

Payne, J., Srouji, M., Yap, D. A., and Kosaraju, V. (2020). *BERT learns (and teaches) chemistry. arXiv*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, D., et al. (2011). Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830.

Pereira, T., Abbasi, M., Ribeiro, B., and Arrais, J. P. (2021). Diversity oriented Deep Reinforcement Learning for targeted molecule generation. *J. Cheminformatics* 13, 21. doi:10.1186/s13321-021-00498-z

Perkel, J. M. (2021). Ten computer codes that transformed science. *Nature* 589, 344–348. doi:10.1038/d41586-021-00075-2

Peter, F., Gillespie, B. W., Juraske, R., and Jolliet, O. (2014). Estimating half-lives for pesticide dissipation from plants. *Environ. Sci. Technol.* 48, 8588–8602. doi:10.1021/es500434p

Plante, P.-L., Francovic-Fontaine, É., May, J. C., McLean, J. A., Baker, E. S., Laviolette, F., et al. (2019). Predicting ion mobility collision cross-sections using a deep neural network: DeepCCS. *Anal. Chem.* 91 (8), 5191–5199. doi:10.1021/acs.analchem.8b05821

Plowright, A. T., Johnstone, C., Kihlberg, J., Pettersson, J., Robb, G., and Thompson, R. A. (2012). Hypothesis driven drug design: improving quality and effectiveness of the design make-test-analyse cycle. *Drug Discov. Today* 1-2, 56–62. doi:10.1016/j.drudis.2011.09.012

Podda, M., Bacciu, D., and Micheli, A. (2020). *A deep generative model for fragment-based molecule generation*, 2240–2250. s.l., s.n.

Polykovskiy, D., Zhebrak, A., Sanchez-Lengeling, B., Golovanov, S., Tatanov, O., Belyaev, S., et al. (2020). Molecular sets (moses): a benchmarking platform for molecular generation models. *Front. Pharmacol.* 11, 565644. doi:10.3389/fphar.2020.565644

Pomyen, Y., Wanichthanarak, K., Poungsombat, P., Fahrmann, J., Grapov, D., and Khoomrung, S. (2020). Deep metabolome: applications of deep learning in metabolomics. *Comput. Struct. Biotechnol. J.* 18, 2818–2825. doi:10.1016/j.csbj.2020.09.033

Probst, D., and Reymond, J.-L. (2020). Visualization of very large high-dimensional data sets as minimum spanning trees. *J. Cheminformatics* 12, 12. doi:10.1186/s13321-020-0416-x

Prykhodko, O., Johansson, S. V., Kotsias, P. C., Arús-Pous, J., Bjerrum, E. J., Engkvist, O., et al. (2019). A *de novo* molecular generation method using latent vector based generative adversarial network. *J. Cheminformatics* 11, 74. doi:10.1186/s13321-019-0397-9

PyTorch (2023). PyTorch. [Online] Available at: https://pytorch.org/.

QSAR Toolbox (2023). QSAR Toolbox. [Online] Available at: https://qsartoolbox.org/ (Accessed, 2023).

Quareshy, M., Prusinska, J., Li, J., and Napier, R. (2018). A cheminformatics review of auxins as herbicides. J. Exp. Bot. *5 January* 69 (2), 265–275. doi:10.1093/jxb/erx258

Ragoza, M., Hochuli, J., Idrobo, E., Sunseri, J., and Koes, D. R. (2017). Protein-ligand scoring with convolutional neural networks. *J. Chem. Inf. Model.* 57 (4), 942–957. doi:10.1021/acs.jcim.6b00740

Ramos, A. F., Evanno, L., Poupon, E., Champy, P., and Beniddir, M. A. (2019). Natural products targeting strategies involving molecular networking: different manners, one goal. *Nat. Product. Rep.* 36 (7), 960–980. doi:10.1039/c9np00006b

Ramsundar, B., Eastman, P., Walters, P., and Pande, V., 2019. *Deep learning for the life sciences: applying deep learning to genomics, microscopy, drug discovery, and more*; 1st Edition. s.l.:O'Reilly Media.

Rao, J., Zheng, S., Song, Y., Chen, J., Li, C., Xie, J., et al. (2021). *MolRep: a deep representation learning library for molecular property prediction*. bioRxiv [Preprint]. doi:10.1101/2021.01.13.426489

Ray, L., and Kirsch, R. (1957). Finding chemical records by digital computers. *Science* 126, 814–819. doi:10.1126/science.126.3278.814

Ray, P. C., Kiczun, M., Huggett, M., Lim, A., Prati, F., Gilbert, I. H., et al. (2017). Fragment library design, synthesis and expansion: nurturing a synthesis and training platform. *Drug Discov. Today* 22, 43–56. doi:10.1016/j.drudis.2016.10.005

RDKit (2023). *RDKit: open-source cheminformatics software. [Online]* (Accessed, 2023).

Reker, D. (2019). *Practical considerations for active machine learning in drug discovery*. Technologies: Drug Discovery Today, 73–79.

Reng, F., Lai, L., and Pei, J. (2018). Computational chemical synthesis analysis and pathway design. *Front. Chem.* 6, 199. doi:10.3389/fchem.2018.00199

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). *Explaining the predictions of any classifier. arXiv*.Why should I trust you?

Richards, R. J., and Groener, A. M. (2022). *Conditional β-VAE for de novo molecular generation. arXiv*.

Roberts, G., Myatt, G. J., Johnson, W. P., Cross, K. P., and Blower, P. E. (2000). LeadScope: software for exploring large sets of screening data. *J. Chem. Inf. Comput. Sci.* 40, 1302–1314. doi:10.1021/ci0000631

Rodríguez-Pérez, R., and Bajorath, J. (2021). Explainable machine learning for property predictions in compound optimization. *J. Med. Chem.* 64, 17744–17752. doi:10.1021/acs.jmedchem.1c01789

Rogers, D., and Hahn, M. (2010). Extended-connectivity fingerprints. *J. Chem. Inf. Model.* 50 (5), 742–754. doi:10.1021/ci100050t

Rong, Y., Bian, Y., Xu, T., Xie, W., Wei, Y., Huang, W., et al. (2021). *Self-supervised graph transformer on large-scale molecular data. arXiv*.

Sabando, M. V., Ponzoni, I., Milios, E. E., and Soto, A. J. (2021). *Using molecular embeddings in QSAR modeling: does it make a difference? arXiv*.

Sagar, A. (2020). *Generate novel molecules with target properties using conditional generative models. arXiv preprint arXiv:2009.12368*.

Samuel, G., and Bjerrum, E. (2022). PaRoutes: a framework for benchmarking retrosynthesis route predictions. *Chemrxiv*. doi:10.26434/chemrxiv-2022-wk8c3

Sánchez-Bayo, F., and Wyckhuys, K. A. (2019). Worldwide decline of the entomofauna: a review of its drivers. *Biol. Conserv.* 232, 8–27. doi:10.1016/j.biocon.2019.01.020

Sanchez-Lengeling, B., Wei, J., Lee, B., Reif, E., Wang, P., Qian, W., et al. (2020). Evaluating attribution for graph neural networks. *NeurIPS Proc.*

Sander, T., Freyss, J., von Korff, M., and Rufener, C. (2015). DataWarrior: an open-source program for chemistry aware data visualization and analysis. *J. Chem. Inf. Model.* 55, 460–473. doi:10.1021/ci500588j

Sandfort, F., Strieth-Kalthoff, F., Kühnemund, M., Beecks, C., and Glorius, F. (2021). A structure-based platform for predicting chemical reactivity. *Chem* 6 (6), 1379–1390. doi:10.1016/j.chempr.2020.02.017

Sarfraz, M., Dosdall, L., and Keddie, B. (2005). *Spinosad: a promising tool for integrated pest management*. Outlooks on Pest Management, 78–84. Available at: https://www.ingentaconnect.com/content/resinf/opm/2005/00000016/00000002/art00009;jsessionid=6j6p5poc8j7nh.x-ic-live-03

Schollée, J. E., Schymanski, E. L., Stravs, M. A., Gulde, R., Thomaidis, N. S., and Hollender, J. (2017). Similarity of high-resolution tandem mass spectrometry spectra of structurally related micropollutants and transformation products. *J. Am. Soc. Mass Spectrom.* 28, 2692–2704. doi:10.1007/s13361-017-1797-6

Schrödinger (2023a). Live design. [Online] Available at: https://www.schrodinger.com/products/livedesign/drug-discovery.

Schrödinger (2023b). Maestro. [Online] Available at: https://www.schrodinger.com/products/maestro.

Schroeter, T. S., Schwaighofer, A., Mika, S., Ter Laak, A., Suelzle, D., Ganzer, U., et al. (2007). Estimating the domain of applicability for machine learning QSAR models: a study on aqueous solubility of drug discovery molecules. *J. Computer-Aided Mol. Des.* 21, 485–498. doi:10.1007/s10822-007-9125-z

Schwaller, P., Laino, T., and Vaucher, A. (2021). IBM RXN: new AI model boosts mapping of chemical reactions. [Online] Available at: https://www.ibm.com/blogs/research/2021/01/roborxn-designs-molecules/ (Accessed May, 2021).

Segler, M. H. S., Kogej, T., Tyrchan, C., and Waller, M. P. (2018). Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Central Sci.* 4, 120–131. doi:10.1021/acscentsci.7b00512

Segler, M. H. S., and Waller, M. P. (2017). Neural-symbolic machine learning for retrosynthesis and reaction prediction. *Chem. – A Eur. J.* 23, 5966–5971. doi:10.1002/chem.201605499

Shavalieva, G., Papadokonstantakis, S., and Peters, G. (2022). Prior knowledge for predictive modeling: the case of acute aquatic toxicity. *J. if Chem. Inf. Model.* 62 (17), 4018–4031. doi:10.1021/acs.jcim.1c01079

Shen, C., Krenn, M., Eppel, S., and Aspuru-Guzik, A. (2021). Deep molecular dreaming: inverse machine learning for de-novo molecular design and interpretability with surjective representations. *Mach. Learn. Sci. Technol.* 2, 03LT02. doi:10.1088/2632-2153/ac09d6

Shen, J., and Nicolaou, C. A. (2019). Molecular property prediction: recent trends in the era of artificial intelligence. *Drug Discov. Today Technol.* 32, 29–36. doi:10.1016/j.ddtec.2020.05.001

Sheridan, R., Wang, W. M., Liaw, A., Ma, J., and Gifford, E. M. (2016). Extreme gradient boosting as a method for quantitative structure–activity relationships. *J. Chem. Inf. Model.* 56, 2353–2360. doi:10.1021/acs.jcim.6b00591

Sheridan, R. P. (2015). The relative importance of domain applicability metrics for estimating prediction errors in QSAR varies with training set diversity. *J. Chem. Inf. Model.* 55 (6), 1098–1107. doi:10.1021/acs.jcim.5b00110

Shi, C., Xu, M., Zhu, Z., Zhang, W., Zhang, M., and Tang, J. (2020). *GraphAF: a flow-based autoregressive model for molecular graph generation*. s.n: s.l.

Shi, J., Zhao, G., and Wei, Y. (2018). Computational QSAR model combined molecular descriptors and fingerprints to predict HDAC1 inhibitors. *Med. Sci.* 34 (F1), 52–58. doi:10.1051/medsci/201834f110

Siegwart, M., Graillot, B., Blachere Lopez, C., Besse, S., Bardin, M., Nicot, P. C., et al. (2015). Resistance to bio-insecticides or how to enhance their sustainability: a review. *Front. Plant Sci.* 6, 381. doi:10.3389/fpls.2015.00381

Simonovsky, M., and Meyers, J. (2020). DeeplyTough: learning structural comparison of protein binding sites. *J. Chem. Inf. Model.* 60 (4), 2356–2366. doi:10.1021/acs.jcim.9b00554

SimulationsPlus (2023). ADMET Predictor® metabolism module. [Online] Available at: https://www.simulations-plus.com/software/admetpredictor/metabolism/.

Sliwoski, G., Kothiwale, S., Meiler, J., and Lowe, J. (2013). Computational methods in drug discovery. *Pharmacol. Rev.* 66, 334–395. doi:10.1124/pr.112.007336

Spark, A. (2023). Apache Spark. Available at: spark.apache.org

Sparks, T. C., Crouse, G. D., Benko, Z., Demeter, D., Giampietro, N. C., Lambert, W., et al. (2021). The spinosyns, spinosad, spinetoram, and synthetic spinosyn mimics - discovery, exploration, and evolution of a natural product chemistry and the impact of computational tools. *Pest Manag. Sci.* 77, 3637–3649. doi:10.1002/ps.6073

Sparks, T. C., Crouse, G. D., Dripps, J. E., Anzeveno, P., Martynow, J., DeAmicis, C. V., et al. (2008). Neural network-based QSAR and insecticide discovery: spinetoram. *J. Computer-Aided Mol. Des.* 22, 393–401. doi:10.1007/s10822-008-9205-8

Sparks, T. C., Hunter, J. E., Lorsbach, B. A., Hanger, G., Gast, R. E., Kemmitt, G., et al. (2018). Crop protection discovery: is being the first best? *J. Agric. Food Chem.* 66 (40), 10337–10346. doi:10.1021/acs.jafc.8b03484

Ståhl, N., Falkman, G., Karlsson, A., Mathiason, G., and Boström, J. (2019). Deep reinforcement learning for multiparameter optimization in *de novo* drug design. *J. Chem. Inf. Model.* 59, 3166–3176. doi:10.1021/acs.jcim.9b00325

Stokes, J. M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N. M., et al. (2020). A deep learning approach to antibiotic discovery. *Cell* 180 (4), 475–483. doi:10.1016/j.cell.2020.04.001

Struble, T. J., Alvarez, J. C., Brown, S. P., Chytil, M., Cisar, J., DesJarlais, R. L., et al. (2019). Current and future roles of artificial intelligence in medicinal chemistry synthesis. *J. Med. Chem.* 63 (16), 8667–8682. doi:10.1021/acs.jmedchem.9b02120

Sun, M., Zhao, S., Gilvary, C., Elemento, O., Zhou, J., and Wang, F. (2020). Graph convolutional networks for computational drug development and discovery. *Briefings Bioinforma.* 21 (3), 919–935. doi:10.1093/bib/bbz042

Supratik, K., Kunal, R., and Leszczynsk, J. (2017). "On applications of QSARs in food and agricultural sciences: history and critical review of recent developments," in Advances in QSAR modeling. Challenges and Advances in computational Chemistry and physics. *s.l.* (Cham: Springer), 203–302.

Supratik, K., Roy, K., and Leszczynski, J. (2018). Applicability domain: a step toward confident predictions and decidability for QSAR modeling. *Methods Mol. Biol.* 1800, 141–169. doi:10.1007/978-1-4939-7899-1_6

Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., and Feuston, B. P. (2003). Random forest: a classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Comput. Sci.* 43 (6), 1947–1958. doi:10.1021/ci034160g

Szymkuć, S., Gajewska, E. P., Klucznik, T., Molga, K., Dittwald, P., Startek, M., et al. (2016). Computer-assisted synthetic planning: the end of the beginning. *Angew. Chem. Int. Ed.* 55, 5904–5937. doi:10.1002/anie.201506101

Tableau Software, L. L. C. (2023). Tableau. [Online] Available at: https://www.tableau.com/(Accessed, 2023).

Tang, W., Li, Y., Yu, Y., Wang, Z., Xu, T., Chen, J., et al. (2020). Development of models predicting biodegradation rate rating with multiple linear regression and support vector machine algorithms. *Chemosphere* 253, 126666. doi:10.1016/j.chemosphere.2020.126666

Tarasova, O. A., Biziukova, N. Y., Filimonov, D. A., Poroikov, V. V., and Nicklaus, M. C. (2019). Data mining approach for extraction of useful information about biologically active compounds from publications. *J. Chem. Inf. Model.* 59 (9), 3635–3644. doi:10.1021/acs.jcim.9b00164

Thomas, S., Kuska, M. T., Bohnenkamp, D., Brugger, A., Alisaac, E., Wahabzada, M., et al. (2018). Benefits of hyperspectral imaging for plant disease detection and plant protection: a technical perspective. *J. Plant Dis. Prot.* 125 (1), 5–20. doi:10.1007/s41348-017-0124-6

Tian, S., Cao, X., Greiner, R., Li, C., Guo, A., and Wishart, D. S. (2021). CyProduct: a software tool for accurately predicting the byproducts of human cytochrome P450 metabolism. *J. Chem. Inf. Model.* 26, 3128–3140. doi:10.1021/acs.jcim.1c00144

TIBCO (2023). TIBCO Spotfire®. [Online] Available at: https://www.tibco.com/products/tibco-spotfire.

Tice, C. M. (2001). Selecting the right compounds for screening:does Lipinski's Rule of 5 for pharmaceuticalsapply to agrochemicals? *Pest Manag. Sci.* 57, 3–16. doi:10.1002/1526-4998(200101)57:1<3::aid-ps269>3.0.co;2-6

Torx Software Ltd (2023). TORX. [Online] Available at: https://www.torx-software.com/.

ULC, C.C.G. (2023). Molecular operating environment (MOE). [Online] Available at: https://www.chemcomp.com/.

United States Environmental Protection Agency (2023). *United States environmental protection agency*. [Online] Available at: https://www.epa.gov/ (Accessed, 2023).

Uppal, K., Walker, D. I., Liu, K., Li, S., Go, Y. M., and Jones, D. P. (2016). Computational metabolomics: a framework for the million metabolome. *Chem. Res. Toxicol.* 29 (12), 1956–1975. doi:10.1021/acs.chemrestox.6b00179

U.S. EPA (2021). Guidance for reviewing pesticide environmental fate studies. [Online] Available at: https://www.epa.gov/pesticide-science-and-assessing-pesticide-risks/guidance-reviewing-pesticide-environmental-fate (Accessed, 2021).

U.S. EPA (2023a). CTS: chemical transformation simulator. [Online] Available at: https://qed.epa.gov/cts/ (Accessed, 2023).

U.S. EPA (2023b). US. EPA. [Online] Available at: https://www.epa.gov/chemical-research/epa-new-approach-methods-work-plan-reducing-use-vertebrate-animals-chemical (Accessed, 2023).

Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., et al. (2022). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acid Res.* 50 (D1), D439–D444. doi:10.1093/nar/gkab1061

Venkatasubramanian, V., and Mann, V. (2022). Artificial intelligence in reaction prediction and chemical synthesis. *Curr. Opin. Chem. Eng.* 36, 100749. doi:10.1016/j.coche.2021.100749

Venkatraman, V. (2021). FP-ADMET: a compendium of fingerprint-based ADMET prediction models. *J. Cheminformatics* 13, 75. doi:10.1186/s13321-021-00557-5

Venko, K., Drgan, V., and Novič, M. (2018). Classification models for identifying substances exhibiting acute contact toxicity in honeybees (*Apis mellifera*)$. *SAR QSAR Environ. Res.* 29 (9), 743–754. doi:10.1080/1062936x.2018.1513953

Volkamer, A., Riniker, S., Nittinger, E., Lanini, J., Grisoni, F., Evertsson, E., et al. (2023). *Artificial intelligence in the life sciences.*Machine learning for small molecule drug discovery in academia and industry.

Walker, A. S., and Clardy, J. (2021). A machine learning bioinformatics method to predict biological activity from biosynthetic gene clusters. *J. Chem. Inf. Model.* 61 (6), 2560–2571. doi:10.1021/acs.jcim.0c01304

Walker, E., Kammeraad, J., Goetz, J., Robo, M. T., Tewari, A., and Zimmerman, P. M. (2019). Learning to predict reaction conditions: relationships between solvent, molecular structure, and catalyst. *J. of Chem. Inf. Model.* 59 (9), 3645–3654. doi:10.1021/acs.jcim.9b00313

Wang, F., Liigand, J., Tian, S., Arndt, D., Greiner, R., and Wishart, D. S. (2021a). CFM-ID 4.0: more accurate ESI-MS/MS spectral prediction and compound identification. *Anal. Chem.* 17 (94), 11692–11700. doi:10.1021/acs.analchem.1c01465

Wang, M., Carver, J. J., Phelan, V. V., Sanchez, L. M., Garg, N., Peng, Y., et al. (2016). Sharing and community curation of mass spectrometry data with global natural products social molecular networking. *Nat. Biotechnol.* 34 (8), 828–837. doi:10.1038/nbt.3597

Wang, M.-y., Wang, F., Hao, G.-F., and Yang, G.-F. (2019). FungiPAD: a free web tool for compound property evaluation and fungicide-likeness analysis. *J. Agric. Food Chem.* 67 (7), 1823–1830. doi:10.1021/acs.jafc.8b06596

Wang, Y., Abuduweili, A., Yao, Q., and Dou, D. (2021b). *Property-aware relation networks for few-shot molecular property prediction. arXiv.*

Wang, Y., Wang, J., Cao, Z., and Farimani, A. B. (2021c). *Molecular contrastive learning of representations via graph neural networks. arXiv.*

Wang, Y., Xiao, J., Suzek, T. O., Zhang, J., Wang, J., Zhou, Z., et al. (2012). PubChem's BioAssay database. *PubChem's BioAssay Database* 40, D400–D412. doi:10.1093/nar/gkr1132

Wang, Y., Xiong, Y., Garcia, E. A. L., and Butch, C. J. (2022). Drug chemical space as a guide for new herbicide development: a cheminformatic analysis. *J. Agric. Food Chem.* 70, 9625–9636. doi:10.1021/acs.jafc.2c01425

Warren, G. T., Andrews, C. W., Capelli, A. M., Clarke, B., LaLonde, J., Lambert, M. H., et al. (2006). A critical assessment of docking programs and scoring functions. *J. Med. Chem.* 49 (20), 5912–5931. doi:10.1021/jm050362n

Weber, J. K., Morrone, J. A., Bagchi, S., Pabon, J. D. E., Kang, S. g., Zhang, L., et al. (2021). Simplified, interpretable graph convolutional neural networks for small molecule activity prediction. *J. Computer-Aided Mol. Des.* 36, 391–404. doi:10.1007/s10822-021-00421-6

Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Introd. Methodol. encoding rules* 28 (1), 31–36. doi:10.1021/ci00057a005

Whiteker, G. T. (2019). Applications of the 12 principles of green chemistry in the crop protection industry. *Org. Process Res. Dev.* 23 (10), 2109–2121. doi:10.1021/acs.oprd.9b00305

Wicker, J., Lorsbach, T., Gütlein, M., Schmid, E., Latino, D., Kramer, S., et al. (2016). enviPath--The environmental contaminant biotransformation pathway resource. *Nucleic Acid Res.* 4 (44), D502–D508. (D)). doi:10.1093/nar/gkv1229

Wieder, O., Kohlbacher, S., Kuenemann, M., Garon, A., Ducrot, P., Seidel, T., et al. (2020). A compact review of molecular property prediction with graph neural networks. *Drug Discov. Today Technol.* 37, 1–12. doi:10.1016/j.ddtec.2020.11.009

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3, 160018. doi:10.1038/sdata.2016.18

Williams, A. J., Grulke, C. M., Edwards, J., McEachran, A. D., Mansouri, K., Baker, N. C., et al. (2017). The CompTox Chemistry Dashboard: a community data resource for environmental chemistry. *J. Cheminformatics* 9, 61. doi:10.1186/s13321-017-0247-6

Willighagen, E., Mayfield, J. W., Alvarsson, J., Berg, A., Carlsson, L., Jeliazkova, N., et al. (2017). The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. *J. Cheminformatics* 9, 33. doi:10.1186/s13321-017-0220-4

Winter, R., Montanari, F., Noé, F., and Clevert, D.-A. (2019). Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. *Chem. Sci.* 10 (6), 1692–1701. doi:10.1039/c8sc04175j

Wishart, D. S. (2008). Metabolomics: applications to food science and nutrition research. *Trends Food Sci. Technol.* 19, 482–493. doi:10.1016/j.tifs.2008.03.003

Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., et al. (2017). DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* 46 (D1), D1074–D1082. doi:10.1093/nar/gkx1037

Wishart, D. S., Feunang, Y. D., Marcu, A., Guo, A. C., Liang, K., Vázquez-Fresno, R., et al. (2018). HMDB 4.0: the human metabolome database for 2018. *Nucleic Acid Res.* 4 (46), D608–D617. doi:10.1093/nar/gkx1089

Wouters, O. J., McKee, M., and Luyten, J. (2020). Estimated research and development investment needed to bring a new medicine to market, 2009-2018. *JAMA* 323 (9), 844–853. doi:10.1001/jama.2020.1166

Wu, Z., Ramsundar, B., Feinberg, E., Gomes, J., Geniesse, C., Pappu, A. S., et al. (2018). MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.* 9, 513–530. doi:10.1039/c7sc02664a

Wu, Z., Zhu, M., Kang, Y., Leung, E. L. H., Lei, T., Shen, C., et al. (2020). Do we need different machine learning algorithms for QSAR modeling? A comprehensive assessment of 16 machine learning algorithms on 14 QSAR data sets. *Briefings Bioinforma.* 22 (3). doi:10.1093/bib/bbaa321

Xu, Y., Ma, J., Liaw, A., Sheridan, R. P., and Svetnik, V. (2017). Demystifying multitask deep neural networks for quantitative structure–activity relationships. *J. Chem. Inf. Model.* 57 (10), 2490–2504. doi:10.1021/acs.jcim.7b00087

Xue, D., Gong, Y., Yang, Z., Chuai, G., Qu, S., Shen, A., et al. (2019). Advances and challenges in deep generative models for *de novo* molecule generation. *WIREs Comput. Mol. Sci.* 9, e1395. doi:10.1002/wcms.1395

Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., et al. (2019). Analyzing learned molecular representations for property prediction. *J. Chem. Inf. Model.* 59 (8), 3370–3388. doi:10.1021/acs.jcim.9b00237

Yang, L., Sang, C., Wang, Y., Liu, W., Hao, W., Chang, J., et al. (2021). Development of QSAR models for evaluating pesticide toxicity against Skeletonema costatum. *Chemosphere* 285, 131456. doi:10.1016/j.chemosphere.2021.131456

Yang, Q., Ji, H., Lu, H., and Zhang, Z. (2021). Prediction of Liquid chromatographic retention time with graph neural networks to assist in small molecule identification. *Anal. Chem.* 93 (4), 2200–2206. doi:10.1021/acs.analchem.0c04071

Yang, X., Wang, Y., Byrne, R., Schneider, G., and Yang, S. (2019). Concepts of artificial intelligence for computer-assisted drug discovery. *Chem. Rev.* 119 (18), 10520–10594. doi:10.1021/acs.chemrev.8b00728

Yap, C. W. (2011). PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *J. Comput. Chem.* 32 (7), 1466–1474. doi:10.1002/jcc.21707

Ying, Z., Bourgeois, D., You, J., Zitnik, M., and Leskovec, J. (2019). GNNExplainer: generating explanations for graph neural networks. *Adv. Neural Inf. Process. Syst.* 32, 9240–9251.

Yoshimori, A., Asawa, Y., Kawasaki, E., Tasaka, T., Matsuda, S., Sekikawa, T., et al. (2021). Design and synthesis of DDR1 inhibitors with a desired pharmacophore using deep generative models. *ChemMedChem* 16, 955–958. doi:10.1002/cmdc.202000786

Yoshimori, A., and Bajorath, J. (2020). The SAR matrix method and an artificially intelligent variant for the identification and structural organization of analog series, SAR analysis, and compound design. *Mol. Inf.* 39 (12), e2000045. doi:10.1002/minf.202000045

Yoshimori, A., Tanoue, T., and Bajorath, J. (2019). Integrating the structure–activity relationship matrix method with molecular grid maps and activity landscape models for medicinal chemistry applications. *ACD Omega* 4, 7061–7069. doi:10.1021/acsomega.9b00595

Young, A., Wang, B., and Röst, H. (2023). *MassFormer: tandem mass spectrum prediction for small molecules using graph transformers*. [Online] Available at:. doi:10.48550/arXiv.2111.0482

Zhan, W., Li, D., Che, J., Zhang, L., Yang, B., Hu, Y., et al. (2014). Integrating docking scores, interaction profiles and molecular descriptors to improve the accuracy of molecular docking: toward the discovery of novel Akt1 inhibitors. Eur. J. Med. Chem. *March* 75 (21), 11–20. doi:10.1016/j.ejmech.2014.01.019

Zhang, Q. Y., and Aires-de-Sousa, J. (2005). Structure-based classification of chemical reactions without assignment of reaction centers. *J. Chem. Inf. Model.* 45 (6), 1775–1783. doi:10.1021/ci0502707

Zhang, W. (2018). Global pesticide use: profile, trend, cost/benefit and more. *Proc. Int. Acad. Ecol. Environ. Sci.* 8 (1), 1–27.

Zhang, Y., Li, S., Xing, M., Yuan, Q., He, H., and Sun, S. (2023). Universal approach to *de novo* drug design for target proteins using deep reinforcement learning. *ACS Omega* 8 (6), 5464–5474. doi:10.1021/acsomega.2c06653

Zhang, Y., Lorsbach, B. A., Castetter, S., Lambert, W. T., Kister, J., Wang, N. X., et al. (2018). Physicochemical property guidelines for modern agrochemicals. *Pesticide Manag. Sci.* 74, 1979–1991. doi:10.1002/ps.5037

Zhong, S., Lambeth, D. R., Igou, T. K., and Chen, Y. (2022). Enlarging applicability domain of quantitative Structure–Activity relationship models through uncertainty-based active learning. *ACS ES&T Eng.* 2, 1211–1220. doi:10.1021/acsestengg.1c00434

Zhou, Y., Cahya, S., Combs, S. A., Nicolaou, C. A., Wang, J., Desai, P. V., et al. (2019). Exploring tunable hyperparameters for deep neural networks with industrial ADME data sets. *J. Chem. Inf. Model.* 59 (3), 1005–1016. doi:10.1021/acs.jcim.8b00671

Zhu, Y., Loso, M. R., Watson, G. B., Sparks, T. C., Rogers, R. B., Huang, J. X., et al. (2011). Discovery and characterization of sulfoxaflor, a novel insecticide targeting sap-feeding pests. *J. Agric. Food Chem.* 59, 2950–2957. doi:10.1021/jf102765x