Artificial Intelligence in the Life Sciences 3 (2023) 100056

Contents lists available at ScienceDirect

Artificial Intelligence in the Life Sciences

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

Perspective

Machine learning for small molecule drug discovery in academia and industry

Andrea Volkamer[a], Sereina Riniker[b], Eva Nittinger[c], Jessica Lanini[d], Francesca Grisoni[e,f], Emma Evertsson[c], Raquel Rodríguez-Pérez[d,*], Nadine Schneider[d,*]

[a] Data Driven Drug Design, Center for Bioinformatics, Saarland University, 66123 Saarbruecken, Germany
[b] Laboratory of Physical Chemistry, ETH Zurich, Vladimir-Prelog-Weg 2, 8093 Zurich, Switzerland
[c] Medicinal Chemistry, Research and Early Development, Respiratory and Immunology (R&I), BioPharmaceuticals R&D, AstraZeneca, Gothenburg, Sweden
[d] Novartis Institutes for Biomedical Research, Novartis Campus, CH-4002 Basel, Switzerland
[e] Eindhoven University of Technology, Dept. Biomedical Engineering, Institute for Complex Molecular Systems, 5600 Eindhoven, the Netherlands
[f] Centre for Living Technologies, Alliance TU/e, WUR, UU, UMC Utrecht, 3584 Utrecht, the Netherlands

ARTICLE INFO

Keywords:
Machine learning
Drug discovery
QSAR
Compound properties
Compound optimization
Drug design
Academia
Pharmaceutical industry
Model life cycle
DMTA
Model deployment

ABSTRACT

Academic and pharmaceutical industry research are both key for progresses in the field of molecular machine learning. Despite common open research questions and long-term goals, the nature and scope of investigations typically differ between academia and industry. Herein, we highlight the opportunities that machine learning models offer to accelerate and improve compound selection. All parts of the model life cycle are discussed, including data preparation, model building, validation, and deployment. Main challenges in molecular machine learning as well as differences between academia and industry are highlighted. Furthermore, application aspects in the design-make-test-analyze cycle are discussed. We close with strategies that could improve collaboration between academic and industrial institutions and will advance the field even further.

1. Introduction

The predictions of bioactivity and physical properties are some of the most important applications of machine learning (ML) and artificial intelligence (AI) in drug discovery. This field is broadly known as quantitative structure-activity and property relationships (QSAR, QSPR) and is a necessary component of numerous drug discovery projects (for an overview of QSAR and its history see Tyrchan et al. [1]). Academia and industry are both playing central roles in shaping the field of molecular machine learning for drug discovery [2–4]. ML is used to make better decisions faster and to accelerate the design-make-test-analyze (DMTA) cycle of novel molecular entities [5]. In pharmaceutical industry, models are usually implemented in a result-oriented fashion to save resources and time, in the spirit of finding the most promising drug candidates or '*fail early and cheap*' [6]. To drive decision-making, model reproducibility, confidence, and robustness are of utmost importance. Another key aspect is the democratization of models and data science practices to enable scientists of different domains to work on a shared goal. Although pharmaceutical industry is becoming progressively more active in fundamental ML and AI research, the focus on the final model application remains the central pillar.

On the contrary, model development and proofs-of-concept are often at the core of academic endeavors. In this case, the main goal is advancing current algorithms, generating knowledge on how to improve the state-of-the-art, and expanding methods' applicability to novel problems [7]. Academic studies typically focus on pushing the boundaries of ML in drug discovery, e.g., by borrowing inspiration from other fields such as natural language processing (NLP) [8] or geometric deep learning (DL) [9]. This is usually achieved using static data sets or existing benchmarks (e.g., GuacaMol [10], FS-Mol [11], MoleculeNet [12]), but only seldom via prospective experimental validation [7]. Projects are also influenced by the duration of a PhD thesis or postdoctoral work, causing rapid turnovers in methods, high competition, and, potentially, *'state-of-the-art chasing'* [13] at the expense of true progress [14]. As it is well known, academia is more publication-oriented, as an effect of the *'publish-or-perish'* culture [15].

Another important aspect is the training and education of young scientists. Dedicated ML/AI programs and departments [16] are expected

[*] Corresponding authors.
E-mail addresses: raquel.rodriguez_perez@novartis.com (R. Rodríguez-Pérez), nadine-1.schneider@novartis.com (N. Schneider).

https://doi.org/10.1016/j.ailsci.2022.100056
Received 15 December 2022; Received in revised form 31 December 2022; Accepted 31 December 2022
Available online 5 January 2023

to complement the more traditional natural science education, and increase the cheminformatics literacy of future generations [17]. This aspect holds promise to turn data science into a key skill of wet-lab scientists, to accelerate interdisciplinary science, and to further strengthen its role within and outside of academia in the future.

In this perspective, we discuss diverse aspects important to ML models in the area of small-molecule drug discovery, including model building and applications. We highlight the different challenges and opportunities that researchers in industry and academia face, and potential synergies. Since the field of ML/AI is very broad, we focus on models that are trained on experimental data as opposed to calculated data, although many aspects are common and transferable. The industry perspective mainly covers large pharmaceutical companies and not necessarily common practices in smaller biotech companies or start-ups.

2. Key steps of the model life cycle

Realizing actionable predictive models in drug discovery can be summarized as an iterative cycle of four different steps, as shown in Fig. 1. Step 1 covers data preparation (Section 2.1), which requires understanding the experimental data for reliable curation. Step 2 constitutes the actual model design and building process (Section 2.2), including architecture definition and hyperparameter tuning. Step 3 refers to performance validation of the model and is crucial for prospective model usage (Section 2.3). Finally, step 4 is the model deployment phase (Section 2.4), where the model is made available to users.

2.1. Training data

ML model performance heavily relies on the quality of the experimental data used for training, including size (*how many datapoints are in the data set?*), chemical and property space coverage (*how broad is the chemical space covered and what is the dynamic range in the data set?*), diversity (*how biased or clustered is the data set?*), and errors (*how noisy is the data set?*). Public and proprietary data from pharmaceutical industry (called ‘in-house’ in the following) typically differ when it comes to these characteristics.

Data availability in the public domain has dramatically increased thanks to major databases such as ChEMBL [18,19] and PubChem [20]. However, those data sets are generally smaller than in-house data sets from pharmaceutical industry. Fig. 2 summarizes the size of different proprietary and public data sets utilized for modelling in research studies [18,21–28]. Exemplified on these absorption, distribution, metabolism, and excretion (ADME) data sets, in-house data (Bayer, Novartis, Merck, and Boehringer Ingelheim) always contained more compound measurements than public data sets (Fig. 2).

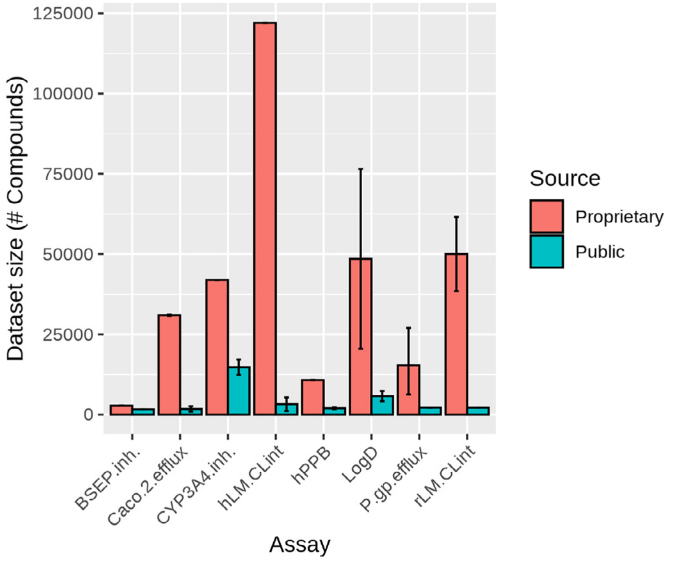

Fig. 2. Data set sizes (i.e., number of compounds) in ADME data sets used for modelling from in-house and public sources. Data are reported for bile salt export pump inhibition (BSEP.inh.) [23,27], Caco-2 permeability efflux (Caco.2.efflux) [18,21,22,27], CYP3A4 inhibition (CYP3A4.inh.) [21,27,28], metabolic stability in human liver microsomes (hLM.CLint)[18,21,27], human plasma protein binding (hPPB) [18,21,27], octanol/water distribution coefficient (LogD) [18,22,25,27], P-glycoprotein substrates (P.gp.efflux) [21,22,24,27], and metabolic stability in rat liver microsomes (rLM.CLint) assays [21,26,27]. For endpoints present in different sources, average data set size is shown, whereas error bars reported the minimum and maximum size values.

To increase data set size, public data are generally pooled from multiple sources [29]), which in turn increases heterogeneity. Merging data sources implies major efforts and bears the risk of biases, redundancies, and error accumulation [30], and poses challenges both in academic and industrial settings. In industry, assay protocols are standardized and typically include multiple measurements per compound, giving

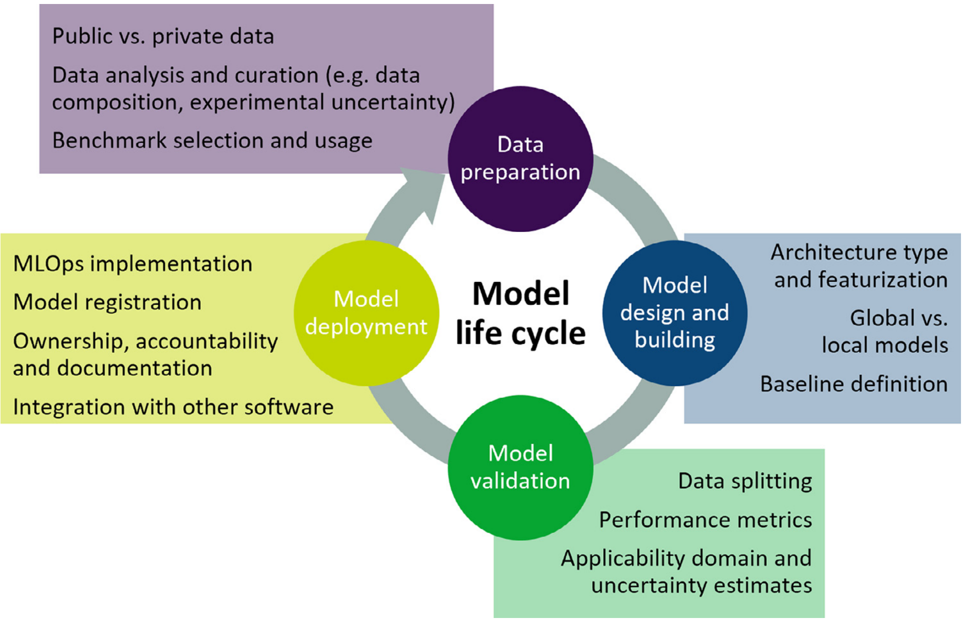
Public vs. private data
Data analysis and curation (e.g. data composition, experimental uncertainty)
Benchmark selection and usage
Data preparation
MLOps implementation
Model registration
Ownership, accountability and documentation
Integration with other software
Model deployment
Model life cycle
Model design and building
Architecture type and featurization
Global vs. local models
Baseline definition
Model validation
Data splitting
Performance metrics
Applicability domain and uncertainty estimates

Fig. 1. Overview of the model life cycle in molecular machine learning. Summarized are the key steps of predictive model building in drug discovery and considerations at different stages.

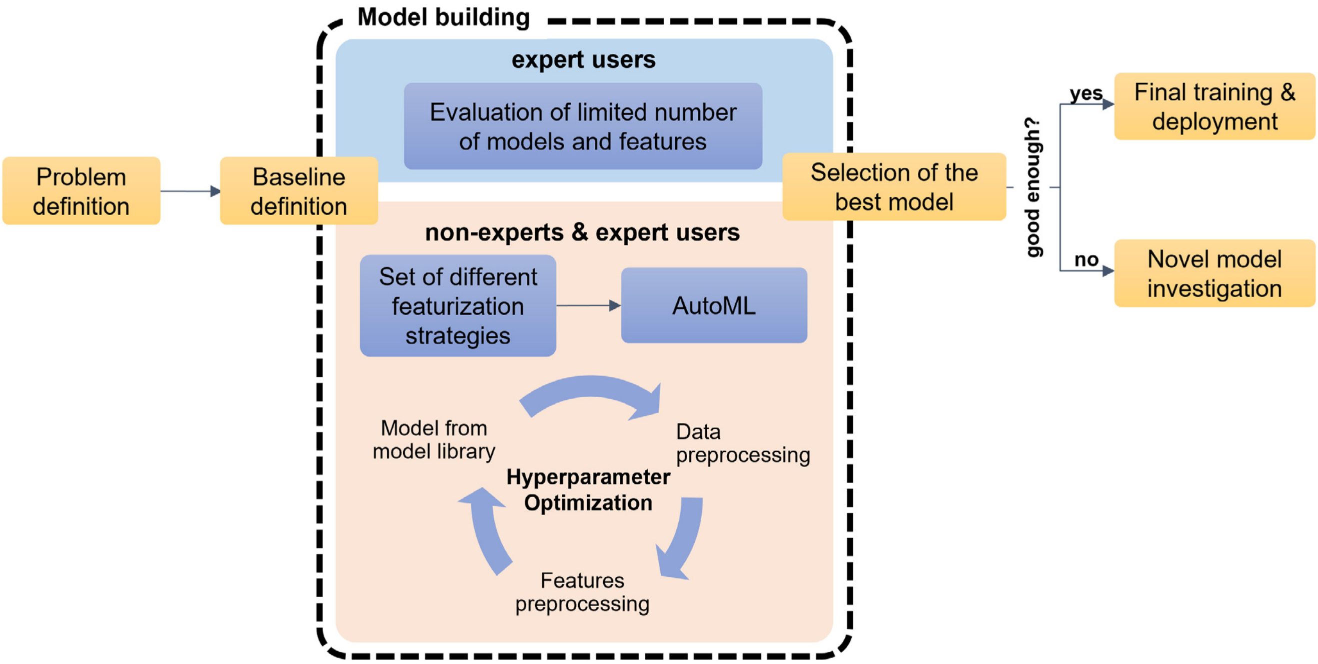
Model building
expert users
Evaluation of limited number of models and features
Problem definition
Baseline definition
Selection of the best model
good enough?
yes
Final training & deployment
no
Novel model investigation
non-experts & expert users
Set of different featurization strategies
AutoML
Model from model library
Data preprocessing
Hyperparameter Optimization
Features preprocessing

Fig. 3. Exemplary overview of model building in industry. First. the problem is defined and some prediction baselines are implemented. Here, different modelling strategies including featurization and ML algorithms should be tested. Previously, this was done by an expert using a limited number of combinations. Nowadays, AutoML implementations allow non-experts and expert users to execute a broader method screening and extensive hyperparameter optimization. Finally, the best model is selected and deployed.

rise to more homogeneous and consistent data sets. However, bringing diverse data sources together might also constitute a major effort due to legacy systems, change of protocols over time or different conventions in annotations (e.g. units or molecule identifiers). Given these challenges with experimental data, curation and homogenization are crucial steps for successful ML applications [11]. Discussions with experimentalists may help detecting outliers, properly combining data from diverse sources, or defining other modelling aspects. Moreover, having replicates to analyze experimental variability and error across the measurement range also provides information about the maximum accuracy that a ML model can achieve and facilitates interpretation of model outputs [31,32].

Even though in-house data sets are typically larger than publicly available ones, these are still heavily biased with respect to the vastness of chemical space. Compound data do not represent a systematic screen of the relevant chemical space but rather collections of clusters centered around specific chemical series targeting given proteins. Thus, structural diversity might be small even in large data sets since many compounds could share the same chemical scaffold. Such data clusters and intrinsic biases highly impact model performance and may cause overly optimistic ML results even with robust validation techniques [33]. Among others, class imbalance or non-uniform property data distributions pose considerable challenges for classification and regression models, respectively [34]. For instance, for bioactivity predictions, the desired outcome typically constitutes the minority class in real-world applications (the situation may be opposite or less imbalanced for ADME data sets). However, public bioactivity data sets often suffer from a lack of 'negative' or 'inactive' data. For example, the ChEMBL database [18,19] contains an unrealistic ratio of active to inactive compounds compared to high-throughput screens (HTS). In particular, 11% and 27% of actives were reported in Cáceres et al. [35], Valsecchi et al. [36], respectively. Nevertheless, HTS hit rates typically range from 0.1 to 2% [37,38]. In contrast to HTS, project-specific data may have higher percentages of actives but typically still below public data sets. This lack of negative public data can be addressed by adding putative negative examples or *decoys* selected according to different strategies [35,39–42]. Due to these reasons, public benchmarking data sets try to mimic but often do not represent real-world data. To facilitate academic research and improve method development, sharing in-house data (e.g., ADME properties of no longer sensitive data) might be beneficial. A recent example is a data set with 1162 PDE10A inhibitors, including binding affinities with time stamps, 77 crystal structures, and docking poses, by Roche [43].

Another – not often discussed – aspect of public benchmarking sets is that they are sometimes not used for the applications they were intended for. A prominent example are the DUD data sets [39], which were originally developed to benchmark docking methods and are now frequently used for ligand-based ML models' evaluation [44,45]. However, such benchmarks may be trivial for some applications such as ligand-based ML-based activity predictions [40,46,47]. Hence, appropriate data set selection for model training and testing is crucial for model performance assessment, as discussed in Section 2.3.

2.2. Model design and building

Model design starts with the definition of the problem and prediction task (see Fig. 3) [48]. In industry, the intend is to use a model prospectively for experiment selection, compound prioritization, or, more generally, assisting in drug design. Thus, users and practical applications need to be considered already during model development and evaluation stages. This is a key difference compared to academia, where the focus is shifted more towards theory and method development (e.g., implementing a model with improved performance with respect to published algorithms or generating new strategies that can be used by others in future applications). Typically, academia is pushing the scientific frontier by developing new tools and improving understanding, while industry is employing them to generate new molecules with desirable properties. Nevertheless, users should also be considered in academia, e.g., when sharing code, a software package, or ML model. In such cases, the final model users are often not clearly defined and discussions with them prior to publication are uncommon or not possible. Therefore, user expectations, stable deployment, and future updates are less frequently considered during model building.

Supervised, semi-supervised or unsupervised techniques can be selected based on the question at hand. Data set characteristics also influence the choice of a regression or classification model, for numerical value and categorical predictions, respectively [49–51]. Thus, the selection of the modelling approach, ML algorithm (e.g. random forest, deep neural networks), and molecular featurization (e.g. fingerprints, calculated descriptors, graphs) is a critical step after problem definition and data curation [52–54]. Due to fast emerging AI technologies, the use of novel algorithms and features is influenced by code availability and data science expertise. In industry, the integration of new ML methodologies might fall behind evergreens like random forests, which provide comparable performance and easier interpretation for many applications. In contrast, given academia's focus on innovation, sometimes simpler and well-performing models might be overlooked due to the state-of-the-art chasing. For that reason, defining strong and valid baseline models is key to detect whether simple models or decision rules are sufficient to achieve similar performance or if more complex modelling and featurization methods are required. Baselines are equally crucial in industry given the focus on robustness and reproducibility more than novelty and complexity [32,55–57].

During model building, investigating a variety of models and features in a systematic way is complex and time-consuming but key for the identification of the best model (in academia) and a useful model (in industry). While this is often part of academic investigations and there is increasing demand for it from journals, there might be less time allocated for hyperparameter tuning and model refinement in industry due to project pressure. New ML libraries, often provided by academic groups - such as AutoML [58] or Optuna [59] to enhance hyperparameter tuning, or DeepChem [60]/MoleculeNet [12], Therapeutic Data Common [61] and AutoSklearn [62] for model building and benchmarking - can help to accelerate and democratize the benchmarking process (see Fig. 3). Some of these tools provide also a collection of open-source benchmark data sets including tasks such as lipophilicity, toxicity (Tox21, ToxCast), or binding affinity (PDBBind) prediction [12].

The selection of the modelling approach is also influenced by the available data set size and composition. For instance, large data sets are often available for some physicochemical and ADME properties that are measured across discovery projects in industry. Models based on such data sets with broader chemical space coverage (often referred to as 'global models') usually generalize better [63]. For project-specific assays, e.g., compound activity towards a protein target, less data is available and is biased towards few chemical series. In this case, 'local models' with a potentially narrower applicability domain are generated. Distinction between global and local models becomes more difficult when trained on public data due to diverse data sources and heterogeneity, as discussed in the previous section.

2.3. Performance evaluation

Even though model validation is an essential step in academia and industry, they focus on different aspects. Model generalizability can only be estimated with a proper choice of data splitting procedures (i.e., training and test data selection) and metrics, and it is key to avoid overly optimistic or pessimistic results [64–69]. Random splits typically give a too optimistic view of the prospective performance of a model [70–72], while time splits have emerged as a preferred approach in industry to evaluate the prospective model performance [71]. Such evaluation mimics how a ML model will be used in practice, i.e. to predict compounds that have not been synthesized or measured (new chemical matter under exploration). Unfortunately, such temporal information is not available in public data, preventing time splits in most academic settings. Some of the data sets popular for academic benchmarks were discussed in Section 2.1. Moreover, an increasing body of literature points to the importance of assessing the model performance in the presence of 'structure-activity/property discontinuities', such as non-additivity [73,74] and activity cliffs [75–77].

Before model deployment or after re-training, a minimum quality criterion should be ensured. The selection of a superior model might thereby differ depending on the considered performance metric. Thus, using multiple metrics [67] as well as ensuring that the metric fits to the application [78] may be beneficial. For instance, a new ML methodology for bioactivity predictions could be evaluated and selected based on balanced accuracy or area under the receiver operating characteristic curve (ROC). However, if such ML model is subsequently applied for virtual screening, only the top-ranked predictions are relevant. In this case, recall in the 1% of top-ranked compounds or enrichment would be more informative. In industry, focus is put on the fact that models are evaluated using the same metrics that the actual model users (i.e., project teams) will check once the model is deployed [23,31].

Furthermore, often ML (especially DL) models are treated as black boxes. To gain trust in the models, methods to better understand what the model learned are required. Explainable AI [79–82] approaches – often researched in academics groups – are developed to shed light on model decisions [83]. Such investigations are valuable to better understand what the model learned and to judge the robustness of ML models [84,85], as well as to learn data-driven features that might be related to a specific effect, e.g., toxicophores [86].

2.4. Model deployment

Figure 4 illustrates the last step of the model-life cycle, which is the deployment of the trained ML model and includes several impor-

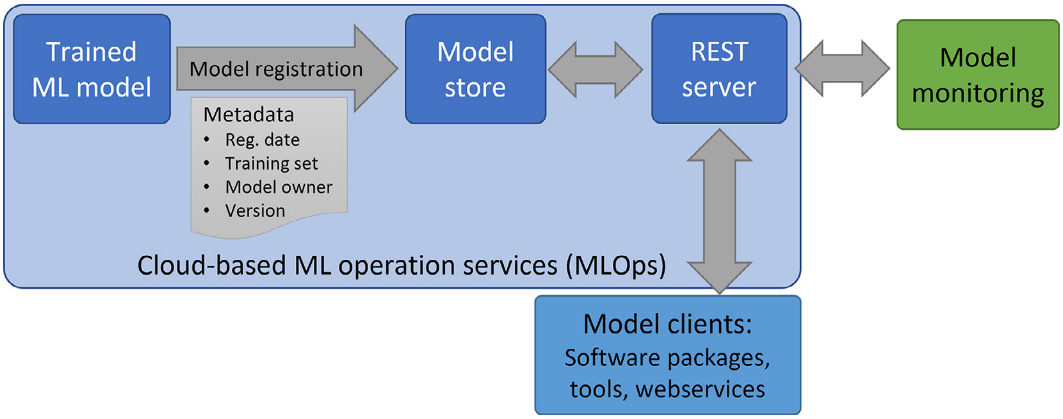

Fig. 4. Exemplary model deployment procedure. Modern MLOps scheme where a trained ML model is registered in a model store, including metadata such as registration (reg.) date, the used training set, model owner, model version or featurization employed, among others. A REST endpoint, facilitating integration into software packages, tools or web services, can be created so that model clients can request predictions. Model monitoring is also considered, including performance estimation on new data points.

tant tasks: model registration, documentation and guidelines, integration into existing tools and workflows, accessibility for non-data scientists, model ownership and accountability, monitoring, and model maintenance. In industry, deployment is necessary to make a model actionable in the drug discovery process.

In the past, technical deployment often included setting up a server, creating a web service, or hard-wiring the models to other tools. This certainly needs a special expertise, which may not be part of a data scientist profile, and implies high maintenance costs. Nowadays, with an increasing importance of predictive models, cloud-based solutions with sophisticated ML operation services (MLOps) implementations help to bridge this gap. A current model deployment procedure is illustrated in Fig. 4. Usually, this includes model registration (into a model store or database to include important metadata like registration date, training set, molecule standardization protocol, or model owner), technical monitoring, and REST (Representational State Transfer) endpoints for each model, which facilitates maintenance and integration into other software packages, tools, or web services. Model usage increases when integrated in tools used daily by project teams, including non-data scientists. Hence, easy accessibility helps democratizing the use of ML models for decision-making.

Computational approaches can become increasingly obscure in terms of accountability and ownership [87–89] due to many data scientists operating on the same pipeline, the inevitability of software bugs and errors, and the increasing use of ML algorithms that might elude full human understanding. Model builders are typically the owners and are responsible to create documentation (including information about training data, modelling approach, applicability domain estimates, and how predictions are reported). The established monitoring protocol should also track the performance of global models in a local setting, to allow project members to judge the reliability of the model data for their purposes. Thus, model ownership and accountability inform the user community about responsibilities and facilitate expert support in case of issues or questions.

A non-negligible aspect for influencing decision-making in drug discovery is the integration of these models into other design platforms accessible to medicinal chemist teams. Continuous model advertisement and education also becomes essential to keep potential users informed and to increase adoption. The modelers or computational chemists connected to a project, who are often involved in the decision-making, also constitute important training resource.

Compared to industry, models in academia are less frequently deployed and come with a shorter life span. Model maintenance and accountability is challenging due to shorter-term contracts (namely graduate students and postdocs). This might lead to premature obsolescence unless detailed documentation and knowledge transfer is achieved before the model builder leaves the research group. Given the push of FAIR (findable, accessible, interoperable, and reusable) science (both data [90] and software [91]) and the growing spirit of open-source developments, these potential issues can be mitigated. However, by default there is limited training in this area, and financial support to hire software engineers in academic groups is largely absent.

Free training workshops (e.g., MolSSi or Software Carpentries) and material (e.g., TeachOpenCADD [92], novel formats as the 'tutorial' category from the LiveCoMS journal) may help to narrow this gap, until the funding/hiring situation changes in academic settings. In addition, libraries and tools are available to support research groups in maintenance tasks (e.g., continuous integration), documentation (e.g., docstrings, read the docs) and readability (e.g. black or PEP8 style guide for Python code). This increases code transparency and can contain elements of accountability. Platforms such as GitHub or GitLab allow users – besides version control – signalling potential issues and proposing solutions to bugs and bottlenecks. Input data or model sharing, which requires more storage, can be achieved via open-repositories like Zenodo [93].

Finally, applications of the models with subsequent measurements of experimental data are typically missing in academia, or can only be realized through collaborations (see Section 4). Nevertheless, there are several good examples of ML models developed in academia that have been made available long-term including updates, e.g., retrosynthesis prediction (ASKCOS [94]), toxicity prediction (emoltox [95], SwissADME [96], OCHEM [97], FAME2 [98]), structure-based analysis (PlayMolecule [99], OpenFold [100]), or learned descriptors (CDDD [101]).

3. Model application aspects

In this section, important considerations during the application phase will be discussed once the model life cycle (see Fig. 1 and Sections 2.1–2.4) has been completed at least once.

3.1. Design-make-test-analyze (DMTA) cycle

Looking at the DMTA cycle – often invoked in industry – predictive models are used in all phases of drug design (Fig. 5). Ideally, proposed compounds with better predicted properties are more likely to be synthesized and prioritized than compounds outside the desired property range. On small scale, the model influences individual decisions on

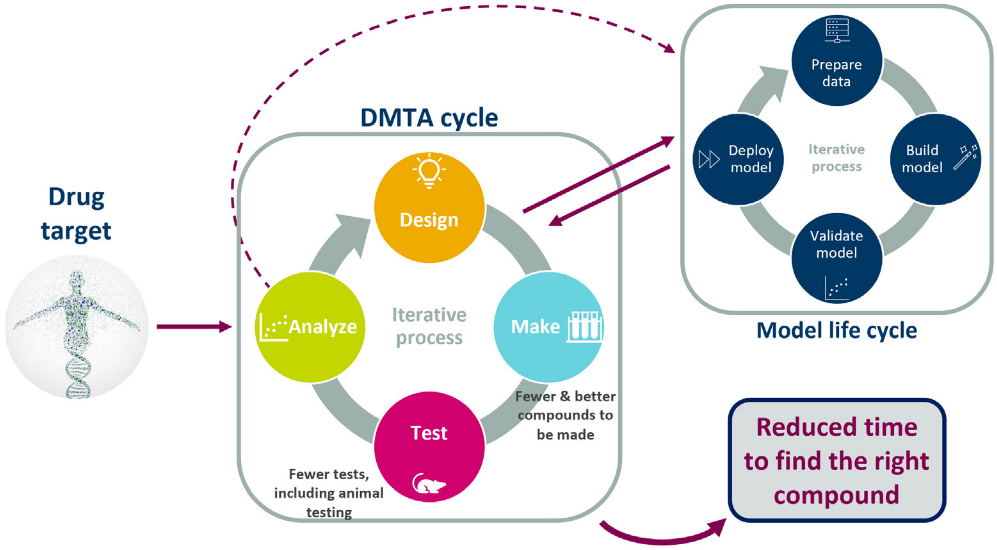

Fig. 5. Integration of design-make-test-analyze (DMTA) cycle with the key steps of the model life cycle, which include data preparation, model building, model validation, and deployment.

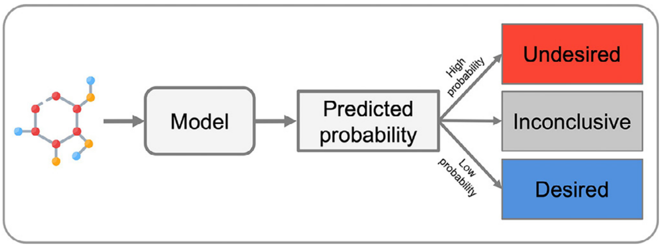

Fig. 6. Scheme of an ML model with intuitive classification outputs and confidence estimation. A classification model can be developed to predict the probability of a certain assay result, e.g. probability that a compound is insoluble. However, instead of providing the model's probability to the user, a more intuitive output can be generated. Exemplary ML model reporting 'undesired' (e.g. insoluble), 'desired' (e.g. soluble), and 'inconclusive' (high uncertainty) categories. Adapted from Rodríguez-Pérez and Gerebtzoff [23] with permission from *Artificial Intelligence in the Life Sciences*.

which compound to make next, whereas on large scale they can also be used for predicting complete virtual libraries to focus making and testing towards the most promising compounds in wet-lab experiments. Models are also used in the analysis phase for experimental selection. If enough evidence for trusting the model - e.g. low error or high confidence values - is present, experimental testing could be avoided and focus shifted to testing other molecules predicted with low confidence [102]. Ultimately, new experimental data and obtained knowledge is fed back to the model in the re-building or re-training stages (see Section 2.2).

In academia – unless executed in a collaborative effort (see Section 4) – the complete DMTA cycle is rarely fully executed. Few examples exist where at least one or two cycles were performed, mainly in the context of active learning [103].

3.2. Model outputs

Reporting meaningful, intuitive, and stable model outputs is essential. The model output should have a clear interpretation for non-data scientists. For instance, the output could be the probability of a given outcome (e.g., desired solubility range) or direct prediction of the experimental value (e.g., solubility in µM units). In case of classification models, meaningful property thresholds should be considered such that predictions enable a discrimination between 'desirable' and 'undesirable' property ranges [23,31]. To define such thresholds and ranges, discussions with project teams and assay experts are of utmost importance. Interestingly, prediction reporting might also be different across applications. Numerical property predictions might be preferred for generative chemistry or library design applications, whereas compound prioritization might be simplified by considering an 'undesired' category [31]. Fig. 6 shows an exemplary classification model with 'undesired', 'desired', and 'inconclusive' (uncertain) categories. A critical assessment of models' output is highly relevant and ideally should include the prediction and an uncertainty estimate [104–106]. However, model applicability and predictions' uncertainty is challenging to assess and is an active and key research topic, especially in academia [104–107]. It has become increasingly crucial in industry to gain the trust of model users and improve decision-making based on ML [23].

3.3. Monitoring and model re-training

In industry, the performance of global models is generally evaluated on data from ongoing projects. Performance can vary depending on how many project compounds have been used for model training or whether the chemical space of the project is covered by the global model. Since drug discovery projects in industry constantly generate new data points, there is a need of model re-training to ensure stable performance and continuous increase of the applicability domain (see Sections 2.2 and 2.3). Exemplified for the prediction of bile salt export pump inhibitors [23], Fig. 7 illustrates the effect of model re-training on prospective performance. Policies on model re-training should be agreed with the user community, e.g., frequency of re-training and strategies for tracing predictions. Each company has its own policies. In some cases, model updates are automatically triggered when a given amount of new data points are available. Predictions for the same compounds might change after each model update, which might cause irritation within the user community who typically prefers consistent predictions over time. Furthermore, prospective performance evaluation becomes difficult if all new data is used for training. A more conservative approach is to only retrain models when the performance decreases over a period of time and across diverse projects. The monitoring protocol can help to identify when the predictive performance is decreasing and model re-training is justified. Such a strategy provides stable predictions and enables a better understanding of current model quality by checking prospective prediction performance.

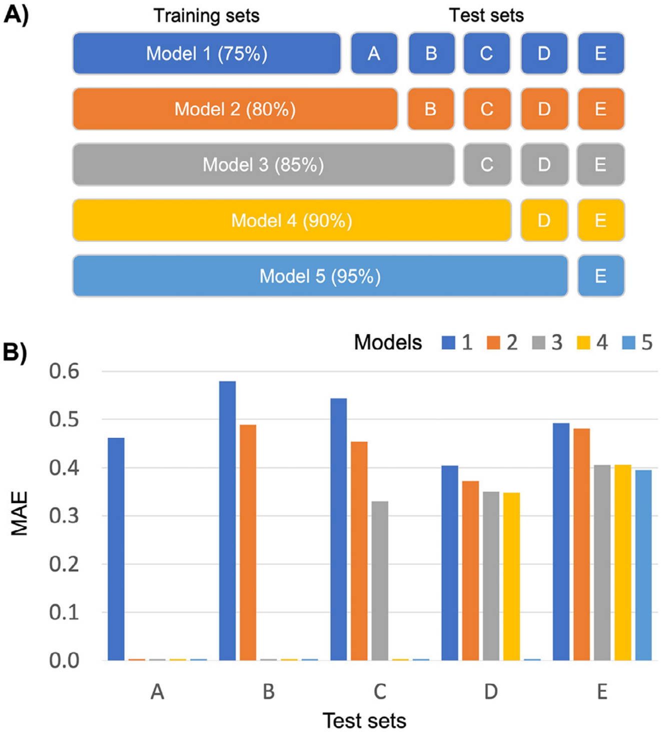

Fig. 7. Effect of model re-training for the prediction of bile salt export pump inhibition. (**A**): Scheme of a ML set-up for a prospective validation (time split). Five models were trained with different time splits: Model 1 (75/25%, dark blue), model 2 (80/20%, orange), model 3 (85/15%, gray), model 4 (90/10%, yellow), and model 5 (95/5%, cyan). Test sets were divided into subsets corresponding to 5% of the data and are labeled with a letter (from A to E) according to the measurement date. Models were evaluated on their prospective test sets. (**B**): Mean absolute error (MAE) values for the five models A–E on the test sets. Adapted from Rodríguez-Pérez and Gerebtzoff [23] with permission from *Artificial Intelligence in the Life Sciences*. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Importantly, an increasing usage of the models for experiment selection influences data generation and thus model re-training. For example, using a reliable solubility prediction model will lead to less insoluble compounds being made in the future; by this the label distribution in the new training set will change and class imbalance will be reinforced. This process is also known as feedback loops and can cause real challenges for future versions of the model [108]). Further development of methods for interpretation could help to improve models in the next training cycle [109].

In academia, studies to identify the right model updating strategy are of high interest. One example is the conformal prediction framework [110], which adds a calibration step on top of the ML model to enable confidence estimation. Mitigating the effects of data drifts in chemical toxicity data – between assays, over time, and from public to proprietary data – on ML performance was studied by assessing the conformal predictor calibration in collaboration with industry partners [111–113]. Morger et al. [112,113] found that updating the calibration set only with the newer data resulted in a higher performance gain than re-training the whole model. This finding may depend on the data set, including size and composition, and needs to be explored further.

4. Collaborations between academia and industry: challenges and opportunities

Industry and academia have been considered in our analysis up to this point as two separated worlds. Naturally, there is a lot of research overlap and both parties benefit from one another. In the following discussion about academia-industry collaborations, we focus on two central aspects: (i) how do we handle reproducibility of publications and the FAIR guidelines when working with proprietary data, and (ii) how can collaborations, including researchers' education, be funded. Interesting thoughts on how academia and industry can work together in drug discovery in general are also provided in the review by Tralau-Stewart and co-workers [114].

4.1. Strategies to facilitate publishing

Some goals and constraints of the academic and industry partners in a collaboration can be partly opposed. In academia, it is important that research outcomes are published in a timely manner. Nowadays, publication typically involves that data sets become freely available and software developments are made open source to ensure reproducibility. Many journals have adapted their guidelines along these lines in the last few years, e.g., Bajorath et al. [115], and Nature Machine Intelligence [116]. However, while it is often relatively straightforward for academic groups to follow the FAIR guidelines [90,91], publishing data sets is often impossible for researchers in industry. A recent study [7] pointed out a remarkable difference in the number of published papers in the domain of ML/AI for drug discovery between industry and academia, with only 7% of the total number of articles with industry affiliations (between 1986 to 2021). This might be attributable to the diverging goals between industry and academia, i.e., intellectual property concerns as well as less pressure to publish in industry. Nevertheless, many companies expect their researchers to publish general scientific findings and code developments. The issues around data sets can pose, however, a non-negligible problem for collaborations between academia and industry, for instance when the idea of the collaboration is to validate/test method developments on real-world proprietary data sets. The latter is an important requisite to make ML models developed in academia more applicable in industry.

A possible solution for this dilemma is to use both public and in-house data when testing/validating a ML model. The public data set can be made available and can be used for in-depth analysis and reproduction of results, whereas applicability of the ML model on real-world data can be demonstrated without sharing the proprietary data [22]. Examples for this way of publishing results from academia-industry collaborations are [86,117–119].

Federated learning is an alternative approach developed for industry-industry collaborations [120]. A recent example is the MELLODDY consortium [121], aiming at using data from different pharmaceutical companies collectively to train ML models while protecting proprietary data of the contributors. So far, models were only usable by the consortium partners. It is yet to be seen whether this approach can be adapted to facilitate publishing in academia-industry collaborations. A different federated learning approach was developed by LHASA [122], where models are trained on proprietary data (each company separately) followed by predictions on non-sensitive public data. The consolidated labels from the different ML models for the public data set serve in turn as input for the training of a final model. As the final model is in principle publishable, this may be a more applicable approach for academia-industry collaborations.

4.2. Strategies for training and funding

In addition to the classical scheme where companies fund PhD or postdoctoral positions of individual research groups, different set-ups for larger consortia with multiple academic and/or industry partners have been explored. Recent education initiatives involving academia-industry partnerships (e.g., BIGCHEM [123] and the AIDD [124] Marie Skłodowska-Curie Doctoral Networks initiatives) are leading to a progressively more relevant role of industry in the education of the next generation of scientists in cutting-edge domains of research.

While the use of commercial software in other areas of computational drug discovery means that training in university cannot directly prepare the students for future work, this has changed in the area of ML/AI due to the increasing usage of the same open-source tools to build, test, and deploy ML models in academia and industry. Using the same tools leads to vast opportunities for cross-fertilization and simplifies collaborations. MIT and several industry partners set up a collaboration to advance research in machine learning for pharmaceutical discovery and synthesis (MLPDS), where the industry partners were funding researchers at MIT while also being strongly involved in driving the research into a relevant direction for real-world applications.

Another recent example is the CACHE (critical assessment of computational hit-finding experiments) [125] initiative between academic and industry partners to improve prospective testing of new methods for hit finding. Blind challenges in general – where for example companies or academic groups make new data sets available (e.g., D3R [126,127], SAMPL [128,129], Tox21 [130], DREAM [131], or the ongoing Kaggle EUOS/SLAS joint challenge for compound solubility prediction [132]) – are a very important way to test new methods in drug discovery (including ML models) with real-world examples.

5. Conclusions

Academia and industry are both driving the field of molecular ML research. Overall there has been great advances in the field of molecular ML, and models have permeated almost every step in the DMTA cycle. Due to the fast emergence of new ML algorithms, the field has needed and needs to adapt quickly, including changes in collaboration – sharing data, protocols, code, and models – and multidisciplinary scientists' education. Besides the common goal of generating valid and trustworthy models to help designing safer drugs faster, there are differences in the way academia and industry approach this aim, often attached to the given circumstances (e.g. resources availability) and overarching goals (e.g. publications or intellectual property aspects).

First, available training data sets are typically larger and more homogeneous (consistent measurements and experimental protocols) in industry, even though project-specific data can also be limited in size as well as chemical space coverage. Although the data situation in academia has improved substantially over the past decade, freely available data sets are usually still smaller in size and often collected from different sources bearing a higher risk of introducing noise and heterogeneity. In addition, due to the bias towards publishing 'positive' results, ratios of 'positive'/'negative' data points in data sets differ between the public and private sectors. Nonetheless, the usage of publicly available data in academic settings allows to promote open-science and reproducibility of scientific findings. This is not always possible in industry settings. Future deposition of negative experimental results as well as funded campaigns to generate additional data for model building might help in this regard.

Table 1
Summary of the important aspects regarding molecular machine learning in industry and academia.

Industry	Academia
Larger, more homogeneous in-house data sets Often multiple measurements per compound Easy access to experimental protocols Project-specific data sets may be small, biased, and imbalanced	Data situation has improved but still more heterogeneous, imbalanced, and smaller data sets with mostly single measurements per compound and less “negative examples” More real-world data sets needed Data are usually available for free sharing and re-utilization, thus contributing to open science.
Model design and building	
Focus on final application Problem definition can be done with final model users Tendency to simpler and robust models Strong benchmark baselines are crucial	Focus on theory to advance the field rather than a concrete application More creative and fundamental work on innovative modelling approaches Simpler but robust models potentially overlooked due to “state-of-the-art chasing” Benchmark baselines are good practice
Performance evaluation	
Time split possible and should be gold standard Application of the model dictates the required level of quality and robustness	Random or cluster-based splits are used in absence of temporal information in public data Standard ML metrics are typically used to evaluate novel methods Availability of standardized benchmarking platforms to build upon and contribute to
Deployment	
Model accessibility by users is crucial Ownership and accountability is considered	Models are less often deployed and often have a shorter life span Difficulties to keep models accessible and updated Model accountability is harder to maintain More open source libraries and services available following programming best practices increase accountability and maintenance of models
Model application aspects	
Performance monitoring to assess the need of re-training or architecture changes Model users and practical applications need to be considered Model output should have a clear interpretation for non-expert users and should be adapted according to the application Model advertisement (with guidelines) and documentation is required to inform users	Model users are (often) not clearly definable Communication to model users (typically non-data scientists) is less common/required Push of FAIR coding and data guidelines helps to improve documentation and user aspects

When designing a new model to achieve a given task, e.g., prioritizing compounds active against a given target, the incentives in academia and industry often differ in their purpose. While academic research aims to push the boundaries of ML techniques to advance method development, models in pharmaceutical industry are built to solve a specific problem or question, where simplicity and robustness are more in the focus. When it comes to model validation – besides typical cross-validation performance reporting – evaluation on external compounds sets is highly recommended. Prospective testing is more common in pharmaceutical companies, whereas academic groups usually realize this through collaborations. In academia, random or cluster-based splits are used as an alternative in the absence of temporal information. At the same time, the availability of open-access benchmarks and related data makes it easier for academic research to leverage and build upon previously designed evaluation systems, thereby ensuring comparability and transparency.

Since academia is more focused on basic research and methodological enhancements, the model cycle often ends with a proof-of-concept study. In contrast, model deployment, accessibility, and stability are crucial in industry since the models are used in active drug discovery projects. More collaboration efforts between academia and industry to share data and code might lessen the gap between exploratory and applied research work. Some examples of private-public collaborations were mentioned showcasing constellations in which science can be advanced in real-world project set-ups while keeping sensitive data private.

For an overview of all different aspects discussed in this perspective see also Table 1.

With the increasing ease at which new ML developments can be translated into industrial applications, we expect that the mentioned boundaries/differences, progressively blur. Active research areas include uncertainty estimation and ML explainability. ML for molecule discovery has come a long way, but its impact will become even more apparent when correct model decisions can be identified (low uncertainty) and rationalized (explained).

Author contributions

All authors have contributed equally to conceptualization and writing of this perspective.

Declaration of Competing Interest

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

Acknowledgements

The authors thank Vigneshwari Subramanian, Nikolaus Stiefl, Grégori Gerebtzoff and Gregory Landrum for proofreading the manuscript and valuable suggestions.

References

[1] Tyrchan C, Nittinger E, Gogishvili D, Patronov A, Kogej T. Chapter 4—Approaches using ai in medicinal chemistry. In: Akitsu T, editor. Computational and data-driven chemistry using artificial intelligence. Elsevier; 2022. p. 111–59. doi:10.1016/B978-0-12-822249-2.00002-5. ISBN 978-0-12-822249-2.

[2] Green DVS. Using machine learning to inform decisions in drug discovery: an industry perspective. In: *Machine learning in chemistry: data-driven algorithms, learning systems, and predictions*, 1326. American Chemical Society; 2019. p. 81–101.

[3] Stephenson N, Shane E, Chase J, Rowland J, Ries D, Justice N, Zhang J, Chan L, Cao R. Survey of machine learning techniques in drug discovery. Curr Drug Metab 2019;20:185–93.

[4] Brown N, Ertl P, Lewis R, Luksch T, Reker D, Schneider N. Artificial intelligence in chemistry and drug design. J Comput-Aided Mol Des 2020;34:709–15.

[5] Schneider G. Automating drug discovery. Nature Rev Drug Discov 2018;17:97–113.

[6] Hughes JP, Rees S, Kalindjian SB, Philpott KL. Principles of early drug discovery. Br J Pharm 2011;162:1239–49.

[7] Mak K-K, Balijepalli MK, Pichika MR. Success stories of AI in drug discovery – where do things stand? Expert Opin Drug Discov 2022;17:79–92.

[8] Öztürk H, Özgür A, Schwaller P, Laino T, Ozkirimli E. Exploring chemical space using natural language processing methodologies for drug discovery. Drug Discov Today 2020;25:689–705.

[9] Atz K, Grisoni F, Schneider G. Geometric deep learning on molecular representations. Nat Mach Intel 2021;3:1023–32.

[10] Brown N, Fiscato M, Segler MH, Vaucher AC. GuacaMol: benchmarking models for de novo molecular design. J Chem Inf Model 2019;59:1096–108.

[11] Stanley M, Bronskill JF, Maziarz K, Misztela H, Lanini J, Segler M, et al. FS-mol: a few-shot learning dataset of molecules. In: 35th conference on NeurIPS datasets and benchmarks track (round 2); 2021.

[12] Wu Z, Ramsundar B, Feinberg EN, Gomes J, Geniesse C, Pappu AS, et al. MoleculeNet: a benchmark for molecular machine learning. Chem Sci 2018;9:513–30.

[13] Church KW, Kordoni V. Emerging trends: sota-chasing. Nat Lang Eng 2022;28:249–69.

[14] Raji I.D., Bender E.M., Paullada A., Denton E., Hanna A.. AI and the everything in the whole wide world benchmark. arXiv preprint:arXiv:2111.153662021

[15] Moosa IA. Publish or perish: perceived benefits versus unintended consequences. Edward Elgar Publishing; 2018.

[16] Zhang D., Mishra S., Brynjolfsson E., Etchemendy J., Ganguli D., Grosz B., Lyons T., Manyika J., Niebles J.C., Sellitto M., et al. The AI index 2021 annual report. 2021arXiv preprint:arXiv:2103.06312

[17] Sydow D., Rodr-guez-Guerra J., Volkamer A.. Teaching Computer-Aided Drug Design Using TeachOpenCADD; chap. 10. 2021, p. 135–158.

[18] Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, et al. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res 2012;40:D1100–7.

[19] Gaulton A, Hersey A, Nowotka M, Bento AP, Chambers J, Mendez D, Mutowo P, Atkinson F, Bellie LJ, Cibrián-Uhalte E, Davies M, Dedman N, Karlsson A, Magarinos MP, Overington JP, Papadatos G, Smit I, Leach AR. The ChEMBL database in 2017. Nucleic Acids Res 2017;45:D945–54.

[20] PubChem: National Center for Biotechnology Information (NCBI). Accessed: 2022-10-19; http://www.pubchem.ncbi.nlm.nih.gov.

[21] Aleksic S, Seeliger D, Brown JB. ADMET predictability at boehringer ingelheim: state-of-the-art, and do bigger datasets or algorithms make a difference? Mol Inf 2021;41:2100113.

[22] Hamzic S, Lewis R, Desrayaud S, Soylu C, Fortunato M, Grégori G, Rodríguez-Pérez R. Predicting in vivo compound brain penetration using multi-task graph neural networks. J Chem Inf Model 2022;62:3180–90.

[23] Rodríguez-Pérez R, Gerebtzoff G. Identification of bile salt export pump inhibitors using machine learning: predictive safety from an industry perspective. AI Life Sci 2021;1:100027.

[24] Sheridan RP. Stability of prediction in production ADMET models as a function of version: why and when predictions change. J Chem Inf Model 2022;62:3477–85.

[25] Montanari F, Kuhnke L, Laak AT, Clevert D-A. Modeling physico-chemical ADMET endpoints with multitask graph convolutional networks. Molecules 2020;25:44.

[26] Lim MA, Yang S, Mai H, Cheng AC. Exploring deep learning of quantum chemical properties for absorption, distribution, metabolism, and excretion predictions. J Chem Inf Model 2022. doi:10.1021/acs.jcim.2c00245.

[27] Venkatraman V. FP-ADMET: a compendium of fingerprint-based ADMET prediction models. J Cheminform 2021;13:75.

[28] Veith H, Southall N, Huang R, James T, Fayne D, Artemeko N, et al. Comprehensive characterization of cytochrome P450 isozyme selectivity across chemical libraries. Nat Biotechnol 2009;27:1050–5.

[29] Kramer C, Kalliokoski T, Gedeck P, Vulpetti A. The experimental uncertainty of heterogeneous public K_i data. J Med Chem 2012;55:5165–73.

[30] Yonchev D, Dimova D, Stumpfe D, Vogt M, Bajorath J. Redundancy in two major compound databases. Drug Discov Today 2018;27:1337–45.

[31] Rodríguez-Peréz R, Trunzer M, Schneider N, Faller B, Gerebtzoff G. Multispecies machine learning predictions of in vitro intrinsic clearance with uncertainty quantification analyses. Mol Pharm 2023;20:383-394. doi:10.1021/acs.molpharmaceut.2c00680. in press

[32] Sheridan RP, Karnachi P, Tudor M, Xu Y, Liaw A, Shah F, Cheng AC, Joshi E, Glick M, Alvarez J. Experimental error, kurtosis, activity cliffs, and methodology: what limits the predictivity of quantitative structure–activity relationship models? J Chem Inf Model 2020;60:1969–82.

[33] Volkov M, Turk J-A, Drizard N, Martin N, Hoffmann B, Gaston-Mathé Y, Rognan D. On the frustration to predict binding affinities from protein–ligand structures with deep neural networks. J Med Chem 2022;65:7946–58.

[34] Esposito C, Landrum GA, Schneider N, Stiefl N, Riniker S. GHOST: adjusting the decision threshold to handle imbalanced data in machine learning. J Chem Inf Model 2021;61:2623–40.

[35] Cáceres EL, Mew NC, Keiser MJ. Adding stochastic negative examples into machine learning improves molecular bioactivity prediction. J Chem Inf Model 2020;60:5957–70.

[36] Valsecchi C, Grisoni F, Motta S, Bonati L, Ballabio D. NURA: a curated dataset of nuclear receptor modulators. Tox Appl Pharm 2020;407:115244.

[37] Bradley D. Dealing with a data dilemma. Nat Rev Drug Discov 2008;7:632–3.

[38] Rodríguez-Pérez R, Miyao T, Jasial S, Vogt M, Bajorath J. Prediction of compound profiling matrices using machine learning. ACS Omega 2018;3:4713–23.

[39] Irwin JJ. Community benchmarks for virtual screening. J Comput-Aided Mol Des 2008;22:193–9.

[40] Riniker S, Landrum GA. Open-source platform to benchmark fingerprints for ligand-based virtual screening. J Cheminf 2013;5:26.

[41] Kurczab R, Smusz S, Bojarski AJ. The influence of negative training set size on machine learning-based virtual screening. J Cheminf 2014;6:32.

[42] Réau M, Langenfeld F, Zagury J-F, Lagarde N, Montes M. Decoys selection in benchmarking datasets: overview and perspectives. Front Pharm 2018;9:11.

[43] Tosstorff A, Rudolph MG, Cole JC, Reutlinger M, Kramer C, Schaffhauser H, Nilly A, Flohr A, Kuhn B. A high quality, industrial data set for binding affinity prediction: performance comparison in different early drug discovery scenarios. J Comput-Aided Mol Des 2022;36:753–65.

[44] Wallach I, Heifets A. Most ligand-based classification benchmarks reward memorization rather than generalization. J Chem Inf Model 2018;58:916–32.

[45] Chen L, Cruz A, Ramsey S, Dickson CJ, Duca JS, Hornak V, et al. Hidden bias in the DUD-E dataset leads to misleading performance of deep learning in structure-based virtual screening. PLoS One 2019;14:e0220113.

[46] Riniker S, Landrum GA. Similarity maps – a visualization strategy for molecular fingerprints and machine-learning methods. J Cheminf 2013;5:43.

[47] Sieg J, Flachsenberg F, Rarey M. In need of bias control: evaluating chemical data for machine learning in structure-based virtual screening. J Chem Inf Model 2019;59:947–61.

[48] Gopal M. Applied machine learning. McGraw-Hill Education; 2019.

[49] Biship CM. Pattern recognition and machine learning (information science and statistics). Springer New York; 2007.

[50] Goodfellow, I., Bengio, Y., Courville, A. (2016). *Deep learning*. MIT press.

[51] Sutton RS, Barto AG. Reinforcement learning: an introduction. MIT Press; 2018.

[52] Raghunathan S, Priyakumar UD. Molecular representations for machine learning applications in chemistry. Int J Quantum Chem 2022;122:e26870.

[53] Wigh DS, Goodman JM, Lapkin AA. A review of molecular representation in the age of machine learning. WIREs Comput Mol Sci 2022:e1603.

[54] Kimber TB, Chen Y, Volkamer A. Deep learning in virtual screening: recent applications and developments. Int J Mol Sci 2021;22:4435.

[55] Lin J. The neural hype and comparisons against weak baselines. In: ACM SIGIR forum, vol. 52; 2019. p. 40–51.

[56] Mucherino A, Papajorgji PJ, Pardalos PM. *K*-nearest neighbor classification. In: Data mining in agriculture. Springer; 2009. p. 83–106.

[57] Matveieva M, Polishchuk P. Benchmarks for interpretation of QSAR models. J Cheminf 2021;13:41.

[58] Karmaker SK, Hassan MM, Smith MJ, Xu L, Zhai C, Veeramachaneni K. Automl to date and beyond: challenges and opportunities. ACM Comput Surv (CSUR) 2021;54:175.

[59] Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: a next-generation hyperparameter optimization framework. In: Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining; 2019. p. 2623–31.

[60] Ramsundar B, Eastman P, Walters P, Pande V, Leswing K, Wu Z. Deep learning for the life sciences. O'Reilly Media; 2019.

[61] Huang K., Fu T., Gao W., Zhao Y., Roohani Y., Leskovec J., Coley C.W., Xiao C., Sun J., Zitnik M.. Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. 2021.

[62] Feurer M, Eggensperger K, Falkner S, Lindauer M, Hutter F. Auto-sklearn 2.0: hands-free automML via meta-learning. J Mach Learn Res 2022;23:1–61.

[63] Sahigara F, Mansouri K, Ballabio D, Mauri A, Consonni V, Todeschini R. Comparison of different approaches to define the applicability domain of QSAR models. Molecules 2012;17(5):4791–810.

[64] Rodríguez-Pérez R, Bajorath J. Evaluation of multi-target deep neural network models for compound potency prediction under increasingly challenging test conditions. J Comput-Aided Mol Des 2021;35:285–95.

[65] Tropsha A. Best practices for QSAR model development, validation, and exploitation. Mol Inf 2010;29:476–88.

[66] Puzyn T, Mostrag-Szlichtyng A, Gajewicz A, Skrzyński M, Worth AP. Investigating the influence of data splitting on the predictive ability of QSAR/QSPR models. Struct Chem 2011;22:795–804.

[67] Bender A, Schneider N, Segler M, Walters WP, Engkvist O, Rodrigues T. Evaluation guidelines for machine learning tools in the chemical sciences. Nat Rev Chem 2022;6:428–42.

[68] Alexander DL, Tropsha A, Winkler DA. Beware of r^2: simple, unambiguous assessment of the prediction accuracy of QSAR and QSPR models. J Chem Inf Model 2015;55:1316–22.

[69] Todeschini R, Ballabio D, Grisoni F. Beware of unreliable Q^2! a comparative study of regression metrics for predictivity assessment of QSAR models. J Chem Inf Model 2016;56:1905–13.

[70] Golbraikh A, Shen M, Xiao Z, Xiao Y-D, Lee K-H, Tropsha A. Rational selection of training and test sets for the development of validated QSAR models. J Comput-Aided Mol Des 2003;17:241–53.

[71] Sheridan RP. Time-split cross-validation as a method for estimating the goodness of prospective prediction. J Chem Inf Model 2013;53:783–90.

[72] Andrada MF, Vega-Hissi EG, Estrada MR, Garro Martinez JC. Impact assessment of the rational selection of training and test sets on the predictive ability of QSAR models. SAR QSAR Environ Res 2017;28:1011–23.

[73] Gogishvili D, Nittinger E, Margreitter C, Tyrchan C. Nonadditivity in public and inhouse data: implications for drug design. J Cheminf 2021;13:47.

[74] Kwapien K, Nittinger E, He J, Margreitter C, Voronov A, Tyrchan C. Implications of additivity and nonadditivity for machine learning and deep learning models in drug design. ACS Omega 2022;7:26573–81.

[75] Schneider N, Lewis RA, Fechner N, Ertl P. Chiral cliffs: investigating the influence of chirality on binding affinity. ChemMedChem 2018;13:1315–24.

[76] Winkler DA, Le TC. Performance of deep and shallow neural networks, the universal approximation theorem, activity cliffs, and QSAR. Mol Inf 2017;36:1600118.

[77] van Tilborg D, Alenicheva A, Grisoni F. Exposing the limitations of molecular machine learning with activity cliffs. J Chem Inf Model 2022;62:5938–51.

[78] Li Z, Yoon J, Zhang R, Rajabipour F, Srubar III WV, Dabo I, Radlińska A. Machine learning in concrete science: applications, challenges, and best practices. npj Comput Mater 2022;8:127.

[79] Rodríguez-Pérez R, Bajorath J. Explainable machine learning for property predictions in compound optimization. J Med Chem 2021;64:17744–52.

[80] Jiménez-Luna J, Grisoni F, Schneider G. Drug discovery with explainable artificial intelligence. Nat Mach Intel 2020;2:573–84.

[81] Yang CC. Explainable artificial intelligence for predictive modeling in healthcare. J Health Inf Res 2022;6:228–39.
[82] Rodríguez-Pérez R, Bajorath J. Chemistry-centric explanation of machine learning models. Artif Intel Life Scie 2021;1:100009.
[83] Ahmed I, Jeon G, Piccialli F. From artificial intelligence to explainable artificial intelligence in industry 4.0: a survey on what, how, and where. IEEE Trans Ind Inf 2022;18:5031–42.
[84] Sheridan RP. Interpretation of QSAR models by coloring atoms according to changes in predicted activity: how robust is it? J Chem Inf Model 2019;59:1324–37.
[85] Jiménez-Luna J, Skalic M, Weskamp N. Benchmarking molecular feature attribution methods with activity cliffs. J Chem Inf Model 2022;62:274–83.
[86] Webel HE, Kimber TB, Radetzki S, Neuenschwander M, Nazaré M, Volkamer A. Revealing cytotoxic substructures in molecules using deep learning. J Comput-Aided Mol Des 2020;34:731–46.
[87] De Laat PB. Algorithmic decision-making based on machine learning from big data: can transparency restore accountability? Philos Technol 2018;31:525–41.
[88] Nissenbaum H.. Accountability in a computerized society. Sci Eng Ethics1996; 2:25–42.
[89] Maini P., Yaghini M., Papernot N.. Dataset inference: ownership resolution in machine learning. arXiv preprint:arXiv:2104.107062021;
[90] Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR guiding principles for scientific data management and stewardship. Sci Data 2016;3(1):1–9.
[91] Chue Hong N.P., Katz D.S., Barker M., Lamprecht A.-L., Martinez C., Psomopoulos F.E., Harrow J., Castro L.J., Gruenpeter M., Martinez P.A., Honeyman T.. FAIR principles for research software (FAIR4RS principles)2021;.
[92] Sydow D, Rodr-guez-Guerra J, Kimber TB, Schaller D, Taylor CJ, Chen Y, Leja M, Misra S, Wichmann M, Ariamajd A, Volkamer A. TeachOpenCADD 2022: open source and FAIR Python pipelines to assist in structural bioinformatics and cheminformatics research. Nucleic Acids Res 2022. doi:10.1093/nar/gkac267.
[93] European Organization For Nuclear Research, OpenAIRE. Zenodo. 2013. https://www.zenodo.org/. 10.25495/7GXK-RD71
[94] Coley CW, Thomas DA, Lummiss JAM, Jaworski JN, Breen CP, Schultz V, Hart T, Fishman JS, Rogers L, Gao H, Hicklin RW, Plehiers PP, Byington J, Piotti JS, Green WH, Hart AJ, Jamison TF, Jensen KF. A robotic platform for flow synthesis of organic compounds informed by AI planning. Science 2019;365:eaax1566.
[95] Ji C, Svensson F, Zoufir A, Bender A. eMolTox: prediction of molecular toxicity with confidence. Bioinf 2018;34:2508–9.
[96] Daina A, Michielin O, Zoete V. SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. Sci Rep 2017;7:42717.
[97] Sushko I, Novotarskyi S, Körner R, Pandey AK, Rupp M, Teetz W, Brandmaier S, Abdelaziz A, Prokopenko VV, Tanchuk VY, Todeschini R, Varnek A, Marcou G, Ertl P, Potemkin V, Grishina M, Gasteiger J, Schwab C, Baskin II, Palyulin VA, Radchenko EV, Welsh WJ, Kholodovych V, Chekmarev D, Cherkasov A, de Sousa JA, Zhang Q-Y, Bender A, Nigsch F, Patiny L, Williams A, Tkachenko V, Tetko IV. Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information. J Comput-Aided Mol Des 2011;25:533–54.
[98] Šícho M, de Bruyn Kops C, Stork C, Svozil D, Kirchmair J. FAME 2: simple and effective machine learning model of cytochrome P450 regioselectivity. J Chem Inf Model 2017;57:1832–46.
[99] PlayMolecule. https://www.playmolecule.com/Accessed: 2022-10-11.
[100] Openfold – democratizing ai for biology. https://www.openfold.io/Accessed: 2022-11-25.
[101] Winter R, Montanari F, Noé F, Clevert D-A. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. Chem Sci 2019;10:1692–701.
[102] Ahlberg E, Bendtsen C, Carlsson L, Fredlund L, Jones B, Noeske T, Oprisiu I, Strömstedt P-E, Wernevik J, Winiwarter S. Use of in silico models for compound property prediction to reduce the in vitro screening burden. Tox Lett 2017;280:S285.
[103] Reker D, Schneider G. Active-learning strategies in computer-assisted drug discovery. Drug Discov Today 2015;20:458–65.
[104] Mervin LH, Trapotsi M-A, Afzal AM, Barret IP, Bender A, Engkvist O. Probabilistic random forest improves bioactivity predictions close to the classification threshold by taking into account experimental uncertainty. J Cheminf 2021;13:62.
[105] Mervin LH, Johansson S, Semenova E, Giblin KA, Engkvist O. Uncertainty quantification in drug design. Drug Discov Today 2021;26:474–89.
[106] Hirschfeld L, Swanson K, Yang K, Barzilay R, Coley CW. Uncertainty quantification using neural networks for molecular property prediction. J Chem Inf Model 2020;60:3770–80.
[107] Bajorath J. Understanding uncertainty in deep learning builds confidence. AI Life Sci 2022;2:100033.
[108] Sculley D, Holt G, Golovin D, Davydov E, Phillips T, Ebner D, Chaudhary V, Young M, Crespo J-F, Dennison D. Hidden technical debt in machine learning systems. Adv NeurIPS 2015;28.
[109] Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: a review of machine learning interpretability methods. Entropy 2021;23:1–45.
[110] Shafer G, Vovk V. A tutorial on conformal prediction. J Mach Learn Res 2008;9:371–421.
[111] McShane SA, Ahlberg E, Noeske T, Spjuth O. Machine learning strategies when transitioning between biological assays. J Chem Inf Model 2021;61:3722–33.
[112] Morger A, Svensson F, McShane SA, Gauraha N, Norinder U, Spjuth O, Volkamer A. Assessing the calibration in toxicological in vitro models with conformal prediction. J Cheminf 2021;13.
[113] Morger A, de Lomana MG, Norinder U, Svensson F, Kirchmair J, Mathea M, Volkamer A. Studying and mitigating the effects of data drifts on ML model performance at the example of chemical toxicity data. Sci Rep 2022;12:7244.
[114] Tralau-Stewart CJ, Wyatt CA, Kleyn DE, Ayad A. Drug discovery: new models for industry – academic partnerships. Drug Discov Today 2009;14:95–101.
[115] Bajorath J, Coley CW, Landon MR, Walters WP, Zheng M. Reproducibility, reusability, and community efforts in artificial intelligence research. Artif Intel Life Sci 2021;1:100002.
[116] Research reuse. repeat. Nat Mach Intell 2020;2:729. doi:10.1038/s42256-020-00277-9.
[117] Riniker S, Wang Y, Jenkins JL, Landrum GA. Using information from historical high-throughput screens to predict active compounds. J Chem Inf Model 2014;54:1880–91.
[118] Morger A, Mathea M, Achenbach JH, Wolf A, Buesen R, Schleifer K-J, Landsiedel R, Volkamer A. KnowTox: pipeline and case study for confident prediction of potential toxic effects of compounds in early phases of development. J Cheminf 2020;12:24.
[119] Esposito C, Wang S, Lange UEW, Oellien F, Riniker S. Combining machine learning and molecular dynamics to predict *P*-glycoprotein substrates. J Chem Inf Model 2020;60:4730–49.
[120] Rieke N, Hancox J, Li W, et al. The future of digital health with federated learning. npj Digit Med 2020;3:119.
[121] Oldenhof M., Ács G., Pejo B., Schuffenhauer A., Holway N., Sturm N., Dieckmann A., Fortmeier O., Boniface E., Mayer C., Gohier A., Schmidtke P., Niwayama R., Kopecky D., Mervin L., Rathi P.C., Friedrich L., Formanek A., Antal P., Rahaman J., Zalewski A., Heyndrickx W., Oluoch E., Stössel M., Vanco M., Endico D., Gelus F., de Boisfossé T., Darbier A., Nicollet A., Blottière M., Telenczuk M., Nguyen V.T., Martinez T., Boillet C., Moutet K., Picosson A., Gasser A., Djafar I., Simon A., Arany A., Simm J., Moreau Y., Engkvist O., Ceulemans H., Marini C., Galtier M.. Industry-scale orchestrated federated learning for drug discovery. arXiv preprint:arXiv:2210.088712022
[122] Fowkes A., Sartini A., Plante J., Davies R., Werner S., Hanser T.. Aligning data from public and proprietary sources to develop federated QSAR models. https://www.lhasalimited.org/Public/Library/2021/Effiris%20QSAR%202021.pdf.
[123] Bigchem project, Marie Skłodowska-Curie grant agreement No 676434. https://www.bigchem.eu/; 2022. Accessed: 2022-09-15.
[124] Advanced machine learning for innovative drug discovery (AIDD) project, Marie Skłodowska-Curie grant agreement no 956832. https://www.bigchem.eu/; 2022. Accessed: 2022-09-15.
[125] Ackloo S, Al-awar R, Amaro RE, Arrowsmith CH, Azevedo H, Batey RA, et al. CACHE (critical assessment of computational hit-finding experiments): a publicprivate partnership benchmarking initiative to enable the development of computational methods for hit-finding. Nat Rev Chem 2022;6:287–95.
[126] Gaieb Z, Liu S, Gathiaka S, Chiu M, ang C Shao HW, Feher VA, et al. D3R grand challenge 2: blind prediction of protein–ligand poses, affinity rankings, and relative binding free energies. J Comput-Aided Mol Des 2018;32:1–20.
[127] Parks CD, Gaieb Z, Chiu M, Yang H, Shao C, Walters WP, et al. D3R grand challenge 4: blind prediction of protein–ligand poses, affinity rankings, and relative binding free energies. J Comput-Aided Mol Des 2020;34:99–119.
[128] Bannan CC, Burley KH, Chiu M, Shirts MR, Gilson MK, Mobley DL. Blind prediction of cyclohexane/water distribution coefficients from the SAMPL5 challenge. J Comput-Aided Mol Des 2016;30:927–44.
[129] Amezcua M, Khoury LE, Mobley DL. SAMPL7 host guest challenge overview: assessing the reliability of polarizable and non-polarizable methods for binding free energy calculations. J Comput-Aided Mol Des 2021;35:1–35.
[130] Attene-Ramos M, Miller N, Huang R, Michael S, Itkin M, Kavlock RJ, Austin CP, Shinn P, Simeonov A, Tice RR, Xia M. The Tox21 robotic platform for the assessment of environmental chemicals – from vision to reality. Drug Discov Today 2013;18:716–23.
[131] Keller A, Gerkin RC, Guan Y, Dhurandhar A, Turu G, Szalai B, Mainland JD, Ihara Y, Yu CW, Wolfinger R, Vens C, Schietgat L, Grave KD, Norel R, Stolovitzky G, Cecchi GA, Vosshall LB, Meyer PDREAM Olfaction Prediction Consortium. Predicting human olfactory perception from chemical features of odor molecules. Science 2017;355:820–6.
[132] 1st EUOS/SLAS joint challenge: Compound solubility. https://www.kaggle.com/competitions/euos-slas/overview Accessed: 2022-11-27