Chemical Research in Toxicology

Editors' Choice
Open Access

pubs.acs.org/crt

Review

# Machine Learning for Toxicity Prediction Using Chemical Structures: Pillars for Success in the Real World

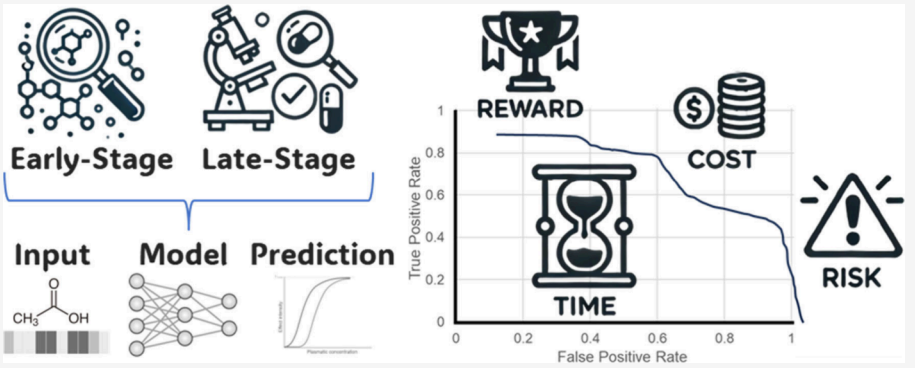
Srijit Seal,* Manas Mahale, Miguel García-Ortegón, Chaitanya K. Joshi, Layla Hosseini-Gerami, Alex Beatson, Matthew Greenig, Mrinal Shekhar, Arijit Patra, Caroline Weis, Arash Mehrjou, Adrien Badré, Brianna Paisley, Rhiannon Lowe, Shantanu Singh, Falgun Shah, Bjarki Johannesson, Dominic Williams, David Rouquie, Djork-Arné Clevert, Patrick Schwab, Nicola Richmond, Christos A. Nicolaou, Raymond J. Gonzalez, Russell Naven, Carolin Schramm, Lewis R Vidler, Kamel Mansouri, W. Patrick Walters, Deidre Dalmas Wilk, Ola Spjuth,* Anne E. Carpenter,* and Andreas Bender*

Cite This: https://doi.org/10.1021/acs.chemrestox.5c00033

Read Online

ACCESS | Metrics & More | Article Recommendations






**ABSTRACT:** Machine learning (ML) is increasingly valuable for predicting molecular properties and toxicity in drug discovery. However, toxicity-related end points have always been challenging to evaluate experimentally with respect to *in vivo* translation due to the required resources for human and animal studies; this has impacted data availability in the field. ML can augment or even potentially replace traditional experimental processes depending on the project phase and specific goals of the prediction. For instance, models can be used to select promising compounds for on-target effects or to deselect those with undesirable characteristics (e.g., off-target or ineffective due to unfavorable pharmacokinetics). However, reliance on ML is not without risks, due to biases stemming from nonrepresentative training data, incompatible choice of algorithm to represent the underlying data, or poor model building and validation approaches. This might lead to inaccurate predictions, misinterpretation of the confidence in ML predictions, and ultimately suboptimal decision-making. Hence, understanding the predictive validity of ML models is of utmost importance to enable faster drug development timelines while improving the quality of decisions. This perspective emphasizes the need to enhance the understanding and application of machine learning models in drug discovery, focusing on well-defined data sets for toxicity prediction based on small molecule structures. We focus on five crucial pillars for success with ML-driven molecular property and toxicity prediction: (1) data set selection, (2) structural representations, (3) model algorithm, (4) model validation, and (5) translation of predictions to decision-making. Understanding these key pillars will foster collaboration and coordination between ML researchers and toxicologists, which will help to advance drug discovery and development.

## INTRODUCTION

In recent years, machine learning (ML) approaches for toxicity prediction using chemical structures and sometimes additional data sources have attracted widespread interest, particularly in drug discovery.[1] There are constant innovations in ML for investigating biological systems and understanding their interactions with drugs, resulting in therapeutic activity and/or

Received: January 26, 2025
Revised: March 24, 2025
Accepted: March 25, 2025

adverse outcomes. Still, these improvements must be reflected in practice. Currently, the costs of bringing a drug to market are increasing,[2] while the overall success rates in clinical drug development remain poor.[3] ML models can be trained on data from empirical assays to predict the properties of compounds from molecular structure (see Box 1 for standard definitions for ML related to predictive models for molecular toxicity). These models are often termed Quantitative Structure Activity/Property Relationship (QSAR/QSPR) models. By enabling the assessment and triage of compound toxicity prior to synthesis and physical testing, these approaches can streamline the drug discovery and development process. Rapid *in silico* screening procedures allow a larger number of potential candidates to be assessed in a shorter time and at decreased cost.

ML models can be used to predict compound properties including efficacy, toxicity (both on-target and off-target biology), and Absorption, Distribution, Metabolism, and Excretion (ADME) or Pharmacokinetic (PK) properties. On-target toxicity occurs when a drug affects its intended target and this itself causes adverse effects, while off-target toxicity arises from interactions of the drug with unintended biological targets.[4] Certain PK properties and dosing schedules can help mitigate on-target toxicity by optimizing how the drug is absorbed, distributed, metabolized, and excreted, effectively controlling its exposure levels. Moreover, more potent compounds with favorable pharmacokinetic properties often requires smaller doses to achieve the desired efficacy. This reduction in the level of necessary exposure can decrease the likelihood of off-target toxicity, thereby improving the overall benefit-risk profile of the drug. Therefore, efficacy and PK/ADME properties drive target engagement and, thus, potential toxicity, all three need to be combined to arrive at a human-relevant toxicity estimation.

For ML-based QSAR/QSPR models, chemical structures of compounds are usually the basis for predictions of continuous (numerical) or categorical labels (properties) biologically relevant to toxicity.[5] In addition to predictive models, recent advancements in generative models have emerged as powerful tools for designing novel chemical structures while optimizing for properties like low toxicity.[6] This Perspective consolidates insights from both academia and industry on the application of ML models across various stages of drug discovery—ranging from early screening and alert systems in hit identification to compound optimization. We focus on discussing the role of ML models in predicting *in vitro* toxicity assays, *in vivo* animal study outcomes, and human-relevant toxicity throughout the drug discovery process.

## CHALLENGES IN TRANSLATING PRECLINICAL TOXICITY DATA TO HUMAN-RELEVANT PREDICTIONS

In the early stages of drug discovery, High-Throughput Screens[7] (HTSs) are commonly applied to screen extensive chemical libraries and find compounds with potential efficacy against a target or phenotype of interest (on-target effects), albeit with some arguments for[8] and against[9] its efficiency. Screens can be cellular or cell-free (isolated protein), with the former testing some aspects of permeability that the latter does not consider. After identifying hits, molecules are optimized to enhance their potency, selectivity, and ADME properties. Usually, *in vitro* assessment of toxicity, including off-target screens, are not conducted in a high-throughput manner; at early screening phases, unless there is an obvious off target activity, selective secondary-pharmacology screens are often run in multiple 'waves'.[10,11]

In later stages of the drug discovery pipeline, *in vitro* human and *in vivo* animal data are used to predict human-relevant toxicity, which is the ultimate goal. However, though widely used, these have shown variable accuracy. For instance, *in vitro* assays have been reported to detect only 50−60% of rare and idiosyncratic drug-induced liver injury (DILI) cases in humans,[12] while animal models often fail to fully capture human-specific toxicity due to species differences in drug metabolism and response.[13−16] While studies in preclinical species can give insights into the drug's effects on a whole organism, they have significant drawbacks. These *in vivo* models are also expensive and not scalable and raise ethical concerns. Therefore, there is a growing desire for alternative models that can more closely mimic human physiology. Complex models using human cells, such as *ex vivo* tissue slices and *in vitro* primary cells or spheroids/organoids, are increasingly used in an attempt to better predict human responses to drugs and decrease animal use in line with the FDA modernization act 2.0[17−19] and related regulations worldwide. There is ongoing work on New-Approach Methodologies (NAMs) that avoid animals for early screening and gaining regulatory acceptance;[20] over 250 regulatory-relevant NAMs have been proposed (https://nams.network/explore). These advancements align with the FDA Harmonization Act and the 3Rs (Replacement, Reduction, and Refinement) principles, which advocate for reducing the need for animal testing and instead opting for more ethical and scalable alternatives.[18] Tissue slices, spheroid or organoid models are usually more physiologically relevant than traditional 2D cultured cancer-derived cell lines. For example, HepaRG, a human hepatic *in vitro* cell line, is widely used because it closely mimics primary human hepatocytes.[21] These cells retain essential functions, such as major cytochrome P450 (CYP) enzymes, key phase 2 enzymes, nuclear receptors, and transporters, making them a valuable tool for studying hepatic drug metabolism and toxicity. However, they do not recapitulate the three-dimensional (3D) structure of the organ and some other aspects; hence a push toward the more expensive tissue slices or spheroids/organoids.[22] Overall, progress in replacing animal testing with human-based NAMs has advanced significantly.[23]

Many toxicity hurdles that slow down the development process are detected in animal experiments rather than in the less-expensive initial screening stages.[24] This underscores the urgent need for more reliable ML models for preclinical safety assessments. Since animal data remains a critical factor in determining a compound's viability, predicting human findings using animal data will remain critical in the near term. Still, advancing toward methods that offer better understanding, more predictive power, and cost-effectiveness represents clear scientific progress. By integrating *in vitro* systems with advanced omics technologies and ML, there is potential to reduce the reliance on animal testing. The key challenge lies in recapitulating biological complexity to improve predictivity, which is needed for better decision making.[25] Therefore, models that integrate PK/ADME with toxicity and predict *in vitro* to *in vivo* translation, or animal-to-human translation could accelerate drug discovery[26,27] by incorporating interspecies and intersex toxicodynamic and toxicokinetic properties.[28,29] Prioritizing such translational models may bridge the gap between *in vitro* or animal data and human data, enabling faster, cost-effective development of safer and more effective medicines.


https://doi.org/10.1021/acs.chemrestox.5c00033
Chem. Res. Toxicol. XXXX, XXX, XXX−XXX

## GUIDELINES FOR ML-BASED TOXICITY MODELS

For the use of ML-based models in drug discovery, model reliability is key. Deprioritization of compounds with undesired properties and progression of those without any early risk flags requires decision-making based on ML models. To ensure the credibility of QSAR/QSPR models, the Organisation for Economic Co-operation and Development (OECD) has defined five principles for model validation.[30,31] These include: (i) a defined end point, (ii) an unambiguous algorithm, (iii) a defined domain of applicability (in terms of chemical structures), (iv) appropriate measures of goodness-of-fit, robustness, and predictivity, and (v) a mechanistic interpretation, if possible. These five principles serve as a foundational framework for evaluating the quality and reliability of QSAR/QSPR models in regulatory contexts, ensuring that predictions are applicable to real-world scenarios. A complementary set of four guidelines was recently proposed for using ML models for ADME/PK for small molecule lead optimization.[32] They include frequently retraining models using data sets from multiple sources (global data) as well as new experimental data (local data), and ensuring models are interactive, interpretable, and integrated into chemists' tools. Bender et al. propose a general set of guidelines for the development and evaluation of ML tools, particularly focusing on supervised learning.[33] Their guidelines emphasize the importance of comprehensive data reporting, conducting retrospective evaluations, comparing models against baselines, performing prospective testing, and ensuring a thorough model interpretation. Additionally, they provide specific recommendations for reporting standards and evaluation metrics, aiming to enhance the reliability and applicability of ML models in addressing real-world chemical challenges.

The risks of reliance on ML include the potential for inaccurate or mis-calibrated models to select drug candidates with undesirable properties, including safety risks. Mis-calibration, here, refers to when the model predicted probabilities do not accurately reflect the true likelihood of observed outcomes. Mis-calibrated models can be a result of (a) low information content (Box 1) in the feature space being used to train models, (b) exposure to narrow chemical space in training data, which then leads to a narrow applicability domain (because of overfitting of models to a small chemical space), or (c) improper (and often nonexistent) prospective validation of models (for future projects), which can lead to overoptimizing and deployment of models that do not generalize.

Building on these principles in the context of ML-driven drug discovery, in this work, we propose five critical pillars for success using ML tools for toxicity prediction (Figure 1):

1. Selecting appropriate data sets that accurately represent the toxicity of interest ensures that model predictions are relevant.
2. Chemical structures must be encoded into relevant representations that capture essential molecular 'information' to generate ML-ready features.
3. Model algorithms must be suitable to learn the signal in the data and representation characteristics mentioned above.
4. Models must be validated to assess their predictive performance, both retrospectively and prospectively, within their applicability domain.
5. In practical model applications, it is crucial to consider project scenarios and desired outcomes to facilitate real-world drug discovery and development.

Adopting these five pillars would enhance the translatability and reliability of predictive models in toxicity prediction, potentially accelerating the design of safe and effective therapeutic candidates. In the subsequent sections, we provide a detailed examination of these five pillars.

When predicting toxicity, there is typically limited data available for the specific end point of interest. Although human toxicity prediction is the ultimate goal, animal studies are more common than in humans and can provide granular information, including dose response, time response, and target organ characterization. Recently, one of the largest data set with *in vivo* toxicity data for 80,000 compounds against a total of 59 acute systemic toxicity end points was analyzed by Sosnin et al. and later made publicly available by Jain et al.[43,44] However, *in vivo* data is only available at advanced stages of compound characterization and *in vitro* data is used for early screening. Thus, *in vitro* assays correlate with one or multiple aspects of the end point of interest, but *they are not the* in vivo *relevant end point itself*. Due to data availability, training data sets for ML models very often include data from such experimental *in vitro* assays that typically proxy for a human clinical end point (Figure 2a). Where mechanisms are better understood, well-studied proxy assays can be used, hence leading to both better predictivity and better interpretability (Figure 2b).[45,46] For example, these could involve functional or binding assays for proteins involved in known mechanisms of toxicity (for example, human ether-a-go-go-related, hERG channel inhibitors in cardiotoxicity[47] or imaging screens for organelle toxicity such as mitochondrial membrane depolarization[48,49]). Hence, the situation differs from case to case quite significantly, where better mechanistic understanding is present different (e.g., target-based) assays might be employed; but in either case the prediction of *in vivo relevant toxicity* is the goal (as opposed to just assay labels), in order to arrive at a prediction in the end that can be used for real-world decision making.

### Chart Box 1

**Box 1.** Some Common Definitions in ML applied to Predictive Toxicity

**Feature Space** is the space spanned by the set of variables or characteristics (features) an ML model uses to make predictions.[34]

**Information content** refers to the relevant information in the features, i.e., representations of chemical structure, that help improve the predictive power, leading to predictions with high accuracy.[35]

**Chemical Space** refers to the universe of all possible chemical compounds.[36] Limited exposure to chemical diversity in training data means the model has learned from a narrow set of examples, which may not represent the full range of compounds it will encounter.

**Overfitting** occurs when an ML model is too closely tailored to the training data, capturing random fluctuations or noise as if they were significant patterns.[37] Overfitting results in a model that performs well on training data with high statistical accuracies but does not generalize well and exhibits low predictivity on new, unseen data.

The **Applicability Domain** is the region of chemical space covered by the training set within which a model is expected to make reliable predictions.[38,39] A narrow applicability domain indicates that the model can only make accurate predictions for a small subset of compounds, where the model has enough representation by its training set chemicals, beyond which its predictions become unreliable[40]. Limited applicability domain can also be a result of overfitting, where the model learns the noise in the training data rather than the underlying relationships, limiting its generalizability.

**Prospective validation** involves testing the model on new data unavailable during development.[41,42] This type of validation assesses the model's ability to generalize and make accurate predictions on future projects or compounds. Prospective validation is effectively a real-world test of the model's predictive power.

## CHOOSING RELEVANT (PREDICTIVE) ASSAY END POINTS FOR *IN VIVO* TOXICITY

An ideal proxy end point strongly predicts the phenotype observed in humans. The design or selection of proxy end points



https://doi.org/10.1021/acs.chemrestox.5c00033
Chem. Res. Toxicol. XXXX, XXX, XXX−XXX

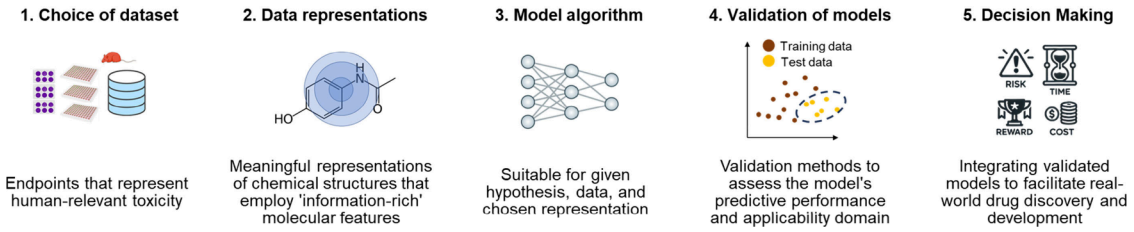



**Figure 1.** Five critical pillars deserving attention from researchers using ML tools for toxicity prediction are discussed in this review. Pillar 1: Choice of data end points, data sets, and data preparation.

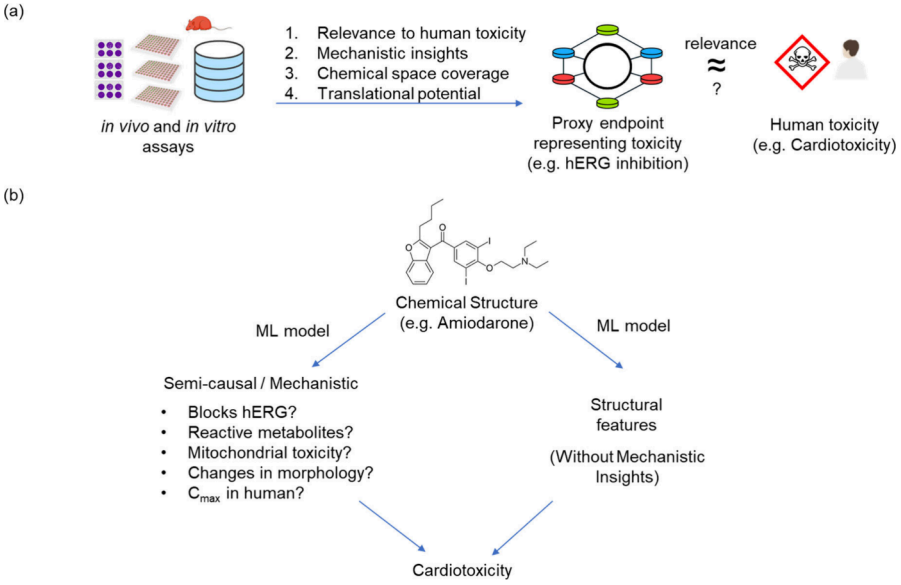



**Figure 2.** (a) The general criteria when choosing the *in vitro* or *in vivo* assay as a proxy end point that could represent human toxicity learned by an ML model. (b) The choice of proxy end point will influence what the model learns and whether mechanistic insights are possible.

**Table 1. Commonly Used *In Vitro* Assays That Can Be Used as Proxies for Human-Relevant Toxicity End Points**[a]

| Toxicity | Assays for Proxy End points | Comment | Commonly Used Data sets in Published Predictive Models |
|---|---|---|---|
| Hepatotoxicity | HepG2 cell viability assays, CYP inhibition assays,[62] mitochondrial membrane depolarization, transporter and reactive metabolite assays | Assess liver function and potential for liver damage through cell viability and enzyme inhibition | Gates Library screen for HepG2 cell viability,[63] CYP P450 Inhibition[64] |
| Cardiotoxicity | Human induced pluripotent stem cell-derived cardiomyocytes (hiPSC-CMs) assays,[65] hERG, Nav1.5, and Cav1.2 assays | Evaluate effects on cardiac cells and potential for arrhythmia through cardiac cell function and ion channel activity | qHTS for Inhibitors of hERG,[66] hERGCentral[67] |
| Nephrotoxicity | Kidney proximal tubule cell assays,[68,69] HK-2 cell cytotoxicity assays[68,69] | Assess kidney cell damage and function, focusing on tubule cells and cytotoxicity | *in vitro* models based on HK-2 cells[70] |
| Neurotoxicity | Neuronal cell viability assays,[71,72] Neurite outgrowth assays,[71,72] and RNAi screen for *Drosophila* neurons | Evaluate potential damage to neurons and effects on neurite growth, indicating neurotoxicity | *in silico*, *in vitro*, and *in vivo* end points of neurotox including sedation, ataxia, and seizure detection[73] |
| Genotoxicity and Carcinogenicity | Ames test, Micronucleus assay,[74] Comet assay, Cell transformation assays[75−79] | Assess DNA damage and mutagenic potential, indicating genetic toxic risks. Evaluate the potential for causing cancer through cell transformation and chromosomal damage | Ames data set collated by Xu et al. [80] CCRIS (Chemical Carcinogenesis Research Information System) Data[81] |
| Endocrine Disruption | Reporter gene assays (e.g., for estrogen receptor activity), H295R steroidogenesis assay[82,83] | Assess interference with hormone activity, focusing on receptor binding and hormone production | Endocrine-related *in vitro* assays from ToxCast[84] |
| Skin Sensitization | Local lymph node assay (LLNA),[85] guinea pig maximization test (GPMT) | Evaluate the potential for causing skin irritation or damage using human skin models | AOP-related assays in SkinSensDB[86]− KeratinoSens/LuSens, human Cell Line Activation Test (h-CLAT), Local lymph node assay (LLNA) |
| Ocular Toxicity | Bovine corneal opacity and permeability (BCOP) assay, HET-CAM assay[87] | Assess the potential for causing eye irritation or damage through corneal and membrane assays | GHS classifications based on Draize rabbit eye test[88] |

[a]Data from these assays are commonly used in the ML models. For more details on the secondary pharmacology panel of targets, see Bowes et al.[11] and Brennan et al.[55]

requires careful consideration to ensure reliability, translatability, and relevance to the application area the model's predictions aim to serve.[50] For example, an assay commonly used to gauge a common cardiotoxicity mechanism detects


https://doi.org/10.1021/acs.chemrestox.5c00033
Chem. Res. Toxicol. XXXX, XXX, XXX−XXX

**Table 2. Data Sets Examined by Authors of This Study That Are Recommended for Machine Learning Models as They Represent Large Data Sets with Well-Defined End Points**

| Data Set | Type of Data (*in vivo*/*in vitro*) | Data Type in Original Publication | Data Points and End Point | Comments | Link to Data Set | Source |
|---|---|---|---|---|---|---|
| Novartis SPD | *in vitro* - targets | Quantitative/Numerical | 1958 drugs assayed for 200 secondary pharmacology targets linked to ADEs; 150k $AC_{50}$ values with 121k unique entries; Cmax(free) for 940 drugs; Dose–response measurements | Internal data (*in vitro* safety pharmacology assays) | https://zenodo.org/records/7378746 | 61 |
| Biogen ADME | *in vitro* - ADME | Quantitative/Numerical | 3521 pharmaceuticals; 6 *in vitro* ADME end points—Human and rat liver microsomal stability, MDR1-MDCK efflux ratio, solubility, and human and rat plasma protein binding assays | Internal prospective data sets (six ADME *in vitro* end points over 20 time points) | https://github.com/molecularinformatics/Computational-ADME | 41 |
| Activity Cliff Eindhoven | *in vitro* - Activity Cliff/Bioactivity | Qualitative/Categorical | 48,707 molecules for 30 targets, including kinases, nuclear receptors, G-protein-coupled receptors, transferases, and proteases | Curated data on 30 macromolecular targets from ChEMBL; Nonproprietary; Small molecules | https://github.com/molML/MoleculeACE | 89 |
| Activity Cliffs Zhang | *in vitro* - Activity Cliff/Bioactivity | Qualitative/Categorical | 400 K Matched Molecular Pairs (MMPs) against 190 targets (142,307 activities), including over 20K MMP-cliffs and 380 K non-AC MMPs | Collected from publicly available database ChEMBL; Nonproprietary; Available to Buy; Small molecules | https://drugai.github.io/ACNet/ | 90 |
| Activity Cliffs Morris | *in vitro* - Activity Cliff/Bioactivity | Quantitative/Numerical | Three binding affinity data sets testing for small molecule inhibitors: 8883 compounds for dopamine receptor D2; 3605 compounds for factor Xa; 1924 compounds for SARS-CoV-2 main protease | Collected from sources: ChEMBL compounds with $K_i$ values for DRD2 and FXA, Covid Moonshot IC50 compounds from COVID moonshot | https://github.com/MarkusFerdinandDablander/QSAR-activity-cliff-experiments | 91 |
| AZ ADMET - Clearance (Hepatocytes) | *in vitro* - *Pharmacokinetic* | Quantitative/Numerical | 408 compounds with intrinsic clearance measured in human hepatocytes; 837 compounds with intrinsic clearance measured in rat hepatocytes | Internal Experimental data determined at AstraZeneca | https://www.ebi.ac.uk/chembl/assay_report_card/CHEMBL3301372/, https://www.ebi.ac.uk/chembl/assay_report_card/CHEMBL3301371/ | 92 |
| AZ ADMET - Clearance (Microsomal) | *in vitro* - *Pharmacokinetic* | Quantitative/Numerical | 1102 compounds with intrinsic clearance measured in human liver microsomes | Internal Experimental data determined at AstraZeneca | https://www.ebi.ac.uk/chembl/assay_report_card/CHEMBL3301370/ | |
| AZ ADMET - PPB | *in vitro* - *Pharmacokinetic* | Quantitative/Numerical | 1614 compounds with percentage bound to plasma by equilibrium dialysis | Internal Experimental data determined at AstraZeneca | https://www.ebi.ac.uk/chembl/assay_report_card/CHEMBL3301365/ | |
| Genetox (Ames/IVMN) - US EPA | *in vitro* - Genetox | Qualitative/Categorical | 8442 compounds; measures presence of *in vitro* micronucleus (IVMN) | Compiled from several public sources (e.g., TOXNET, COSMOS, eChemPortal) | https://github.com/g-patlewicz/genetox | 93 |
| Genetox (Ames/IVMN) - US EPA | *in vitro* - Genetox | Qualitative/Categorical | 211 Ames negative, 682 Ames positive and other equivocal and *in vitro* micronucleus (IVMN) measurements | Compiled from several public sources (e.g., TOXNET, Carcinogenicity and Genotoxicity eXperience (CGX) database) | EURL-ECVAMNegative Ames - https://www.sciencedirect.com/science/article/pii/S1383571820300693?via%3DihubPositive Ames - https://data.europa.eu/data/datasets/jrc-eurl-ecvam-genotoxicity-carcinogenicity-ames?locale=en | 94,95 |
| Ames - ISSTOX | *in vitro* - Genetox | Qualitative/Categorical | 5536 compounds evaluated for Ames test | Compiled from public data sets | https://data.mendeley.com/datasets/ktc6gbfsbh/2 | 96 |
| Ames - Benchmark Data Set for *in silico* Prediction of Ames Mutagenicity | *in vitro* - Mutagenicity | Qualitative/Categorical | 6512 compounds evaluated for Ames test | Compiled from several public sources (e.g., CCRIS, VITIK, GeneTox) | https://pubs.acs.org/doi/10.1021/ci900161g | 97 |
| Ames -*in silico* Prediction of Chemical Ames Mutagenicity | *in vitro* - Mutagenicity | Qualitative/Categorical | 8348 compounds evaluated for Ames test | Compiled from several public literature sources | https://pubs.acs.org/doi/10.1021/ci300400a | 80 |
| Prenatal developmental toxicity | *in vivo* - prenatal developmental toxicity | Qualitative/Categorical | Chemicals with developmental toxicity classifications. | Compiled from public repositories and literature sources | https://pubs.acs.org/doi/10.1021/acs.est.2c01040?goto=supporting-info | 98 |



https://doi.org/10.1021/acs.chemrestox.5c00033
Chem. Res. Toxicol. XXXX, XXX, XXX−XXX

Human Ether-a-Go-Go-Related Gene (hERG) channel inhibitors. However, the hERG assay is merely a proxy end point for some cardiotoxicity signals such as prolonging the QT interval of ECG in a beating heart.[51] It does not convey the potential severity of the induced cardiotoxicity by itself; rather, the extent of QT prolongation (extended interval between the heart contracting and relaxing) and the risk of progressing to TdP (Torsades de Pointes, a type of atypical heart rhythm) vary significantly among compounds known to inhibit hERG (even at very similar activity against hERG itself), illustrating the importance of taking into account the ADME properties of drugs in toxicity prediction.[52] Some proxy end points have been very well established, such as skin sensitization for cosmetics, where the European Union (EU) has already phased out animal-based experiments.[53] Key Events (KEs) are well-defined for skin sensitization, which often involves chemical reactivity, offering detailed mechanistic insight into how chemicals trigger skin sensitization. The OECD has specified a series of *in vitro* assays to accurately measure many of the early key events in skin sensitization. For instance, KE1, which involves the covalent binding of haptens to skin proteins, is effectively assessed by the Direct Peptide Reactivity Assay (DPRA), Amino Acid Derivative Reactivity Assay (ADRA), and the kinetic DPRA assay, and these are well-established methods from the OECD.[54] Proxy end points such as activity against protein targets are mechanistically understood and help gather information on the bioactivity profile of the compounds (Table 1). Secondary pharmacology (off-target effect) is usually screened for using a panel of targets with a relatively well-established relationship to human toxicity (for details, see Bowes et al.[11] and Brennan et al.[55]).

Selecting an end point to be predicted goes hand in hand with selecting data sets containing values for that end point. Companies have their own in-house toxicity data and there are also publicly available data sets, such as reviewed in Vo et al.,[56] Cavasotto et al.,[57] Huang et al.,[58] and Schapin et al.[59] Regardless of the source, data sets often suffer from technical issues that require careful preprocessing and data curation. For instance, there might be invalid chemical structures, inconsistent chemical representations, or undefined stereochemistry. Moreover, data sets might include irrelevant toxicity end points or high assay artifact rates, which do not allow for proper ML benchmarking.[60] Here, we hence present a list of data sets structured by the type of organ-based toxicity they aim to predict (Table 2) that were carefully examined and represent clearly defined end points, with outcomes that are likely to be reproducible. Data sets such as the Novartis Secondary Pharmacology Database (SPD)[61] and Biogen ADME[41] provide a large number of data points from consistently measured state-of-the-art *in vitro* assays and commercially available compounds, which might be particularly useful for newcomers seeking to understand the available data landscape when building and benchmarking ML models. Nevertheless, the precise use of a data set for a particular purpose requires case-by-case consideration and can hence not be generically answered here.

## CHEMICAL DIVERSITY AND APPLICABILITY DOMAIN

The diversity of the represented chemical space is a key aspect to consider when choosing a data set for modeling. The extrapolation of ML models to novel chemistry (not observed during model training) is difficult to achieve in many cases, due to the size of chemical space available and the "local" behavior of properties for highly similar compound. Therefore, the data set for the selected proxy end point should ideally cover a wide range of chemical structures on which the model might be used. For this reason, models are often regularly updated by using assay results from the latest round of project compounds.

A local model built with a small data set lacking structural diversity will have a narrow applicability domain, i.e., strong performance only within a particular area of chemical space. Nevertheless, this might be very useful for a particular project. Local models typically have poor performance on novel chemical scaffolds; that is, scaffolds that represent different features from the chemicals used for training a model.[99,100] Generally, most ML models predicting toxicity (but also bioactivity and other domains) are in practice local models with limited applicability in the larger chemical space, and extrapolation can only be expected to a limited extent due to the high dimensionality of input space and limited data.[101,102] In theory, with large and diverse data sets, ML models ought to have stronger performance and generalizability. Such models are referred to as global models, and are typically trained with data from different series or scaffolds.[100] However, despite their conceptually greater generalization potential, global models for toxicity prediction are often infeasible. One reason is data availability or large experimental variability as data sets are often published with inadequate metadata to be able to effectively combine measurements from different laboratories and/or with different equipment. One example is hERG inhibition, where we often see a massive variability in measured IC50 values due to how compounds were measured.[103] Despite this variability, many charged hERG blockers share a basic nitrogen pharmacophore, which classification models—commonly used for this end point—can recognize when predicting across a broad chemical space.[104,105] Global models work only in cases where the data available is sufficiently large and where the relationship of the end point with chemical structure is sufficiently simple for a global model to capture predictive trends (in the given descriptor space employed, see next section). For example global models may work better in the prediction of logD where large data sets are available[106,107] (compared to models predicting toxicity outcomes). LogD is in principle largely additive in nature; that is, the same group on different scaffolds makes similar contributions (which then enables the model to extrapolate to new cases). Mannhold et al. observed RMSE = 1 for prediction of logP/logD7.4 for over 95,000 Pfizer compounds and mentioned this result as failure.[108] The time split validation model for logD in the article reported an RMSE of 0.92 and is still worse than the 0.3 expected experimental error of this property. End points that are correlated with logD also show some additivity, but for examples where lipophilicity is not the driver of behavior (e.g., high logD, low unbound intrinsic clearance), usually ML models tends to predict incorrectly. A drug's biological activity (potency), in contrast, tends to be more specific to the overall chemical structure (due to the dependency of bioactivity on a particular spatial arrangement of features in a protein binding pocket), and while it has a lipophilicity component, it is more often influenced by nonadditive effects from group combinations and shows less additivity across targets. Validating models within their applicability domain (and establishing this applicability domain in the first place) is an important topic that is discussed later in this work.


https://doi.org/10.1021/acs.chemrestox.5c00033
Chem. Res. Toxicol. XXXX, XXX, XXX−XXX

## STANDARDIZING CHEMICAL STRUCTURES

Preprocessing chemical data should include a clearly outlined and reproducible method to standardize molecules and generate ML-ready representations,[109] considering validity, stereochemistry, solvents and counterions, tautomerism, and protonation states. Each of these factors plays a critical role in the molecule's interaction with biological targets and also for the descriptors that are generated from the chemical structures as input for ML models.

Enantiomers can have vastly different pharmacokinetic and pharmacodynamic properties, impacting efficacy and safety. For example, (R)- and (S)-fluoxetine show differences in metabolism, receptor binding, and side effects.[110,111] At the biological receptor level, enantiomers can also show stereoselectivity; D-asparagine tastes sweet whereas L-asparagine does not.[112] At an experimental level, the presence of multiple enantiomers within a single screening compound sample reduces the potential concentration of active species and complicates the identification of hits from the assays. Separating the activity of various isomers from a racemic mixture is difficult because individual enantiomers must be tested to find which causes the true biological activity. At the modeling level, some molecular fingerprint representations do not encode stereochemistry, which needs to be taken into account (see next section). Moreover, in public databases, stereochemistry annotation is usually poor and absolute configuration is often unknown,[113] particularly for natural products. Unfortunately, in our experience, where stereochemistry information is present, some of it is often arbitrarily chosen, as enantiomers may get separated early in preprocessing. One common approach in modeling is to remove stereochemistry information altogether before generating representations, and as most drug molecules are usually not chiral, this approach generally retains information content of the majority of the training data set (losing information only where stereochemistry has an impact on the respective end point). Unfortunately, some processing workflows involve selecting a single, arbitrary enantiomer, which is clearly not optimal. Another approach in modeling is determining all stereocenters, correcting invalid stereochemistry, and generating canonical isomeric SMILES that retain stereochemistry.[114,115] In cases where stereochemistry has an impact on the respective end point, and where sufficient data is available for the model to learn this relationship, this approach may sometimes be advantageous.

Tautomers are structural isomers that are readily interconvertible due to the transfer of one or more protons.[116] Tautomerism can explain certain bioactivity, for example, the presence of a double bond at the $\alpha,\beta$-position adjacent to the keto carbonyl in $\alpha$-ketoamide derivatives allows the formation of inactive tautomers that exist in equilibrium, and this leads to reduced efficacy against Dengue virus proteases.[117] Thus, considering tautomers is important from a modeling perspective, as they vary in physicochemical properties such as logP, hydrophobicity, and solubility. Many attempts to predict low-energy tautomers have been unsatisfactory.[118] The standardized molecule from tautomer enumeration will typically not represent the "correct" form in the solution, and exhaustive searches can yield nonphysical forms that cannot exist in the solution. The best approach in a standardization pipeline, in practice, can often be to be consistent throughout because often being *consistent* is more important than being *correct* for subsequent model generation steps. Significant efforts have been made in cheminformatics to study tautomerism in large data sets, in particular, on selecting the most consistent tautomer and capturing their molecular descriptors.[119−121] Tools like RDKit and others provide methods to enumerate tautomers; however, they often handle tautomerism differently, and not all types of tautomerism are implemented in every software package, leading generally to different results. For example, PubChem standardizer generates a canonical tautomer during compound standardization, but ChEMBL does not.[114,115] Overall, tautomers are often not trivial to handle when preprocessing structures for predictive modeling.

The protonation state of a molecule affects, among other biological properties, binding to transporters and enzymes, permeability across biological membranes, and pharmacokinetic and pharmacodynamic properties.[122] Due to effects on distribution of a molecule, protonation state may be even more important for toxicity prediction than other end points. For example, drugs such as proton pump inhibitors, with a p$K_a$ between 3.8 and 4.9, selectively accumulate in the acidic secretory canaliculi of parietal cells and are converted into active forms only under specific protonation conditions.[123] When predicting pharmacokinetic parameters, a relevant pH must be considered for protonating compounds during compound standardization (such as the pH of blood, or the pH in the GI tract in case of absorption, etc.).[124,125] The impact of protonation on featurization varies according to the descriptor used. For example, MACCS keys (Molecular ACCess System) have a key (key no. 49) that indicates the presence or absence of charge, while Morgan fingerprints are invariant to protonation states. Descriptors such as the Hammett electronic parameter ($\sigma$), logD, and the Taft constant are also impacted by the protonation state of the molecule. Hence, protonation states of molecules are important, but challenging aspects of preparing small molecules for modeling.[126]

## CHECKING END POINT DATA FOR INCONSISTENCIES OR AMBIGUITIES

The next stage of preparing data for modeling involves assessing the biological data (end point space). Often more so than traditional ML data sets (for example, from the image or speech domain), biological data sets are conditional on experimental factors such as compound concentration, cell type, time points, and *in vivo* pharmacokinetics.[127] Therefore, data from such end points require careful curation to remove inconsistencies or ambiguities in the measured end points and minimize the risk of introducing biases into ML models. For example, the Kendall Tau test can measure the ordinal association between two measured quantities, helping to identify inconsistencies in ranked data.[128] For example, Landrum et al. demonstrated that in nearly 65% of instances, multiple IC50 assays for the same compound-target pair showed discrepancies of over 3-fold, with 27% of cases differing by more than 10-fold, and a Kendall's Tau correlation of 0.51 between multiple assays measuring the same compound against the same target.[129] Outlier detection methods, such as $Z$-scores or the Tukey IQR method, can highlight anomalies that may indicate errors.[130] For example, Kalliokoski et. al recommends if discrepancies are observed in ΔpIC50 are greater than 2.5, there is a high probability they may contain errors and those inconsistent experimental records can be removed before modeling.[131] For example, models designed to predict specific toxicities, such as mitochondrial membrane depolarization or DNA damage from *in vitro* assays, often inadvertently capture and predict broader indicators of overall


https://doi.org/10.1021/acs.chemrestox.5c00033
Chem. Res. Toxicol. XXXX, XXX, XXX−XXX

**Table 3. Most Commonly Used Representations in ML Models, Such as Those Used to Predict Compound Toxicity**

| Representation | Description | Examples | Advantages | Limitations | Implications for ML Model Performance |
|---|---|---|---|---|---|
| Structural keys | Bit format indicating the presence/absence of specific substructures; was intended for cataloging compounds and substructure filtering | MACCS keys[155] | High interpretability; often direct correspondence to molecular substructures | Limited chemical space coverage; no 3D information present; missing information on some substructures | Limited coverage of chemical space, because it was not originally intended to fully represent structure (but empirically often performing rather well) |
| Chemical 2D Fingerprints | Encodes chemical structures into a fixed-size bit string, capturing molecular connectivity | Extended Connectivity Fingerprints (ECFPs)[161] | Compresses feature space; captures broader and more specific chemical space than fixed substructure keys | No 3D information present; algorithms are predefined to generate fingerprints from connectivity | Balances coverage and compactness. It is often possible to obtain structural contributions to model predictions. Often favored in benchmarks, partially due to data set biases (especially analogue bias) |
| 2D Descriptors | Properties calculated from simple two-dimensional (2D) chemical structures | Hydrophobicity, E-state indices, polar surface area, molecular weight and other physicochemical properties[143] | Provides insight into molecular properties; interpretable because related to measurables | Does not capture spatially dependent information of molecular structure | Enhances model interpretability and predictive power (for example, logD is often predictive of complex toxicity) |
| 3D Descriptors | Properties calculated from the three-dimensional structure of molecules, considering spatial arrangements and capturing stereochemistry | Solvent accessible surface area with positive partial charge in the structure,[166] Weighted Holistic Invariant Molecular Descriptors (WHIMs) and Grid-Independent Descriptors (GRINDs) | Captures detailed structural information; potential to encode other 3D properties such as surface area | Computationally intensive due to conformational sampling and complex encodings; usually, bioactive conformations are unknown | Improves model precision by leveraging detailed 3D structural information, essential for tasks requiring stereochemical details (but has conformational sampling trade-off) |
| 2D graph-based representations[153] | Encode the two-dimensional connectivity of molecules | Message-Passing Neural Networks (MPNNs), graph transformers | Alternative to 2-dimensional fingerprints and leverages deep learning architectures | Computationally efficient compared to 3D models, but lack spatial information | Useful for predicting molecular properties related to connectivity |
| 3D graph-based representations[153] | Encode the three-dimensional structure of molecules, considering spatial arrangements | Message-Passing Neural Networks (MPNNs), graph transformers | Captures stereochemistry and has potential to encode other properties such as surface area | Computationally intensive and increased memory requirements | Potential to improve model precision by leveraging connectivity from structural information |
| Descriptors from experiments | Experimental data that capture changes in a biological system upon perturbation with chemicals | Cell Painting, Transcriptomics, Proteomics, etc. | Captures biological *response* space that may not be captured by structure-derived representations | Requires experiments with actual synthesized compounds, processing, quality control, etc. to prepare feature space | Potential to improve the applicability domain of models and identify activity cliffs, as well as interpretability in cases where features are found to be predictive for end point of interest |
| Learned/pretrained representations | Automatically extract features from input data (whether chemical structures or experimental data) using pretrained encoders (supervised/semisupervised) and convert them to a latent space | ChemProp, CLOOME | Captures complex patterns and relationships; adaptable to various data types; improves with more data | May require large data sets for training and generation of features that are suitable for future use cases; can be less interpretable than traditional descriptors | Enhances model accuracy and generalizability; potential to integrate diverse data sources for a comprehensive understanding of molecular behavior |



https://doi.org/10.1021/acs.chemrestox.5c00033
Chem. Res. Toxicol. XXXX, XXX, XXX−XXX

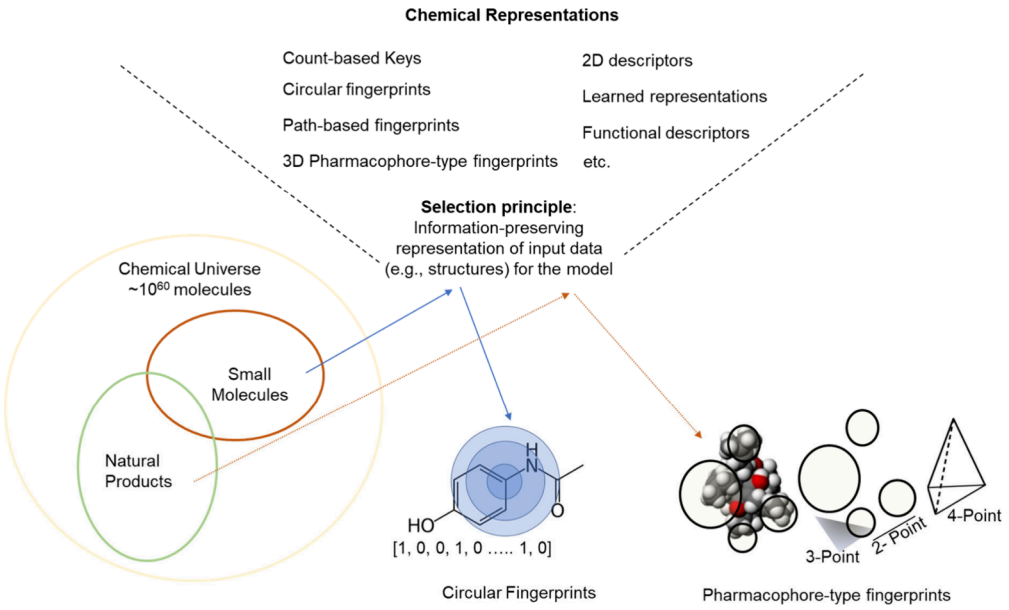



**Figure 3.** An example selection process for chemical representations. Chemical representations vary in their ability to capture the complexity of the molecular structures. When building models for specific categories, like small molecules or natural products, it is important to choose a suitable representation, based on factors such as predictive performance. For instance, circular fingerprints often work well for small molecules, while pharmacophore-based fingerprints may be better suited for natural products. However, studies show that no single representation consistently outperforms others across all QSAR data sets.[151,152]

cell viability instead.[132] Such models may have been trained without confirmatory assay data—which combines results from a screening assay with a cytotoxicity counter screen to exclude highly cytotoxic compounds.[133,134] Some modeling methods, such as Huber regression,[135,136] are less sensitive to outliers, and given the noise in biological data, these can be preferable (an analogy being that median-averaging is less sensitive to outliers than mean-averaging). Other models, such as Probabilistic Neural Networks,[137] can take large variations into account while training, and can generate predictions along with confidence intervals or prediction ranges that reflect the uncertainty or variability in the data. Overall, curating the biological end points to avoid biases or inaccuracies that could compromise model performance is essential for ensuring that models produce meaningful and actionable predictions.

## THE CONDITIONAL NATURE OF DATA AND CATEGORIZATION

A given compound is not simply "toxic" or "non-toxic", as famously quoted by Paracelsus in 1538, "What is there that is not poison? All things are poison, and nothing is without poison. Solely the dose determines that a thing is not a poison".[138] The effective dose, Cmax (maximum concentration in the blood), and concentrations in relevant organs can vary by orders of magnitude depending on a molecule's efficacy and its ADME properties.[139] Therefore, any measurement of toxicity—whether through a simplified binary classification or a detailed dose−response curve—is useful only when it reflects clinically relevant doses. Given that toxicity (as other effects) can vary significantly depending on the experimental conditions, such as the dosage, duration of exposure, and the biological model used (organism, sex, target tissue, etc.), this 'conditionality' of biological data often represents a problem when training predictive models; it is rarely practical to capture experimental conditions in the model itself, given the sparsity of data across experiment variables.

Converting toxicity data sets into simplistic binary prediction problems can hence simplify ML strategies but serve as poor representations of biological realities. For example, selective estrogen receptor modulators stimulate estrogen receptors at very low (picomolar) concentrations, but at much higher (micromolar) concentrations, they impact microtubules, leading to a distinct activity profile.[140] Binarization of continuous values leads to loss of information, especially for compounds that lie close to the threshold for binarization.[127] Mervin et al. presented a compelling case for employing particular modeling methods to better use binary labels based on readouts that were originally continuous in nature, given reasonable constraints upon modeling objective and data sets.[141] The presence of activity cliffs (where small changes in chemical structure lead to major changes in bioactivity) in data sets makes the distribution of labels highly heteroscedastic (uneven over the range of values measured). Given the unequal distribution of measurement error across the range of activity values, the authors argue that regression is also not favorable for *in silico* target prediction and suggest using Probabilistic Random Forest models. These models were shown to improve bioactivity predictions close to the classification threshold by taking into account the experimental uncertainty. To avoid choosing a single particular threshold for binary classes, one could incorporate multiclass categories such as "highly toxic", "toxic", "nontoxic", and "ambiguous". Multiclass categorization can use dose, exposure, or margins to set thresholds for classification calls. For example, models using *in vivo* toxicity data might set a threshold of total $C_{max}$ value less than 10 $\mu$M but greater than 1 $\mu$M, which helps to set the context around the category. Representing intermediate activity or ambiguity can be particularly valuable in capturing the complexities of biological data and preventing the misclassification of borderline cases.

To summarize Pillar 1, the chosen data set must be relevant to the toxicity one wants to predict, represent chemical space sufficiently diverse and relevant for the intended use case, be


https://doi.org/10.1021/acs.chemrestox.5c00033
Chem. Res. Toxicol. XXXX, XXX, XXX−XXX

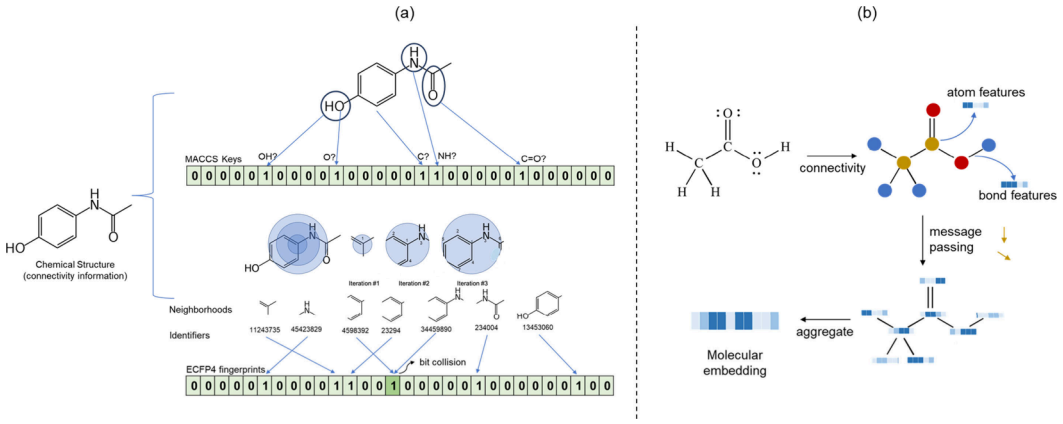



**Figure 4.** (a) Encoding chemical structures using Molecular ACCess System (MACCS) keys, top, and Extended Connectivity Fingerprints (ECFP), bottom. MACCS keys, originally used for cataloging, directly map specific predefined features with bits representing distinct rules, such as the presence of atoms, bonds, etc. ECFP encodes a broader range of substructures. (b) Learned representations encapsulate the learned embeddings of atoms within a molecule aggregated to form a comprehensive molecular embedding. Typically, they focus on molecular topology and connectivity with nodes as atoms and edges as bonds with an iterative message-passing process.

high in information content, provide mechanistic insights (ideally), and be appropriately preprocessed to remove inconsistencies. Well-curated data (with a standardized framework for organizing and categorizing information where possible) reliably train models only when the encoded features capture the most relevant chemical and biological properties. Together, these elements enable the development of accurate, reliable, and generalizable predictive models that are more likely to be applicable in a real-world drug discovery context.

## PILLAR 2: ENCODING CHEMICAL STRUCTURES INTO INFORMATION-RICH AND/OR MEANINGFUL FEATURES

ML models for molecular toxicity prediction that use chemical structures as their input are based on the fact that a molecule's properties depend on its chemical structure[142,143] and that, generally, molecules with 'similar' chemical structures (which can be defined in a wide variety of ways) tend to exhibit similar properties. In classical ML applied to these tasks, quantitative structure−activity−property relationship (QSAR/QSPR) models are trained on predefined feature spaces representing chemical structures (Table 3) and used to predict toxicity labels. As such, these models can only learn relationships captured by the feature representation chosen by the ML practitioner; the choice of representation strongly influences the ability to predict particular molecular toxicities (Figure 3 and described in more detail in subsequent sections).[1,144] Descriptors derived from a molecule's structure are used to represent its chemical and physical properties in this particular feature space, and hence model performance (and interpretability) is intrinsically linked to feature choice. For example, physicochemical properties of a structure are related to ADME end points such as permeability or solubility; omitting features such as charge or molecular weight will generally decrease model performance for those end points. Alternatively, features can be learned directly from chemical structures, such as structural keys, chemical fingerprints, etc., but each featurization method is based on inherent assumptions that might not apply in the universal chemical space, encompassing approximately $10^{60}$ compounds.[36,145] Further, structure−property dependencies are often not smoothly related: exceptions to the general structure−property rule are called activity cliffs—structurally similar compounds that exhibit significantly different properties.[89] Activity cliffs test the limits of structure-only prediction and require experiment-based inputs for successful prediction, whether biological or physicochemical assays differentiate structurally similar compounds. Overall, the choice of features to represent a molecule is nontrivial.

Besides chemical structure-based descriptors, phenotypic data are sometimes available to aid predictions. These omics descriptors can be collected from cells treated with each chemical, including readouts like the Cell Painting assay, transcriptomics, proteomics and metabolomics.[5,146−148] These are relatively unbiased or hypothesis-free descriptors, and unlike computed structural features, they report on the actual interaction of the chemical with a biological system. The experimental costs for phenotypic assays can be high, but descriptors from them have been shown to improve the chemical applicability domain of models to predict toxicity.[146,149,150] Nevertheless, representations from phenotypic data sets might not always be relevant to predict a specific biological end point; that is, they have limitations in their biological applicability domain (i.e., mode of action coverage). The signal present in phenotypic data needs to be established and validated for each end point of interest, in many cases empirically. Whereas Pillar 1 covered compound standardization, the focus of this Pillar is on representations directly derived from chemical structures, given that this representation does not require experimentally determining compound effects, and it is hence universally applicable.

## MOLECULAR KEYS AND FINGERPRINTS

One common representation of molecules using only their chemical structure as input is molecular keys (Figure 4a, top).[153] Structural keys such as Molecular ACCess System (MACCS) keys are a straightforward representation, originally developed for cataloging compounds, but also performing surprisingly well in predictive models.[154] These keys encode molecular structures into a 166-bit format (publicly available) or a 960-bit format (in the commercial implementation) by recording the presence or absence of specific predefined substructures or chemical patterns within a molecule, such as the presence of ring systems, particular functional groups, etc.[155] Because each bit in structural keys corresponds to a predefined feature, often a


https://doi.org/10.1021/acs.chemrestox.5c00033
Chem. Res. Toxicol. XXXX, XXX, XXX−XXX

**Table 4. Commonly Used 3D Descriptors**

| 3D Descriptor | Description | Source |
|---|---|---|
| MoRSE Descriptors | 3D molecular representations of structure based on electron diffraction descriptors | 173 |
| Gravitational Index Descriptors | Based on analogies to gravitational physics, it measures the distribution of atomic mass within a molecule | 174 |
| 3D Autocorrelation Descriptors (3DAc) | The relative position of atoms (or atom properties) based on the separation between atom pairs in Euclidean distance | 175 |
| RDF (Radial Distribution Function) | Provides information about the probability distribution of interatomic distances within a molecule | 176 |
| WHIM Descriptors | Weighted Holistic Invariant Molecular descriptors are derived from the atomic coordinates of a molecule and capture features like size, shape, symmetry, and atom distribution | 177 |
| Flexophore/Pharmacophores | Uses a reduced graph with atom types and encodes molecular flexibility through node distance histograms across diverse conformers | 178 |

molecular substructure, it is easy to interpret what features are vital for the model's predictions. Thus, encoding molecules using structural keys facilitates interpretation due to the direct correspondence between specific values in the fingerprint and particular predefined features.[156,157] However, these predefined features limit the chemical space represented, because structural keys were not intended to capture all relevant aspects of the chemical structure.

An alternative to structural keys are chemical fingerprints, such as extended connectivity fingerprints (ECFPs),[158] which represent the chemical structure of a molecule as a count-based vector or fixed-length binary vector (Figure 4a, bottom). Count-based fingerprints in principle benefit by counting the presence of multiple substructures, which can be especially useful in the case of, for example, repeat units (peptide backbones, terpenes, sugar attachments, etc.). Although count-based fingerprints capture more information, not all ML methods can leverage them as they are always very large ($\sim 2^{32}$-1).[159] Thus, counts are usually converted into binary bits, and therefore some details about the molecule's specific structural or chemical features are lost. This loss of information does not necessarily mean lower predictive power as the signal and predictive ability will be retained if the relevant information is retained. As for interpretability, it is possible to store a correspondence table between bits and chemical substructures to allow the unambiguous interpretation of the bits in a chemical fingerprint.[160] Due to the vast number of possible substructures, most commonly employed chemical fingerprints use a folding and hashing algorithm to encode the chemical structure into a fixed-size bit string.[161] Here, the same bit could correspond to multiple substructures, known as a hash collision (Figure 4a). The transformation is usually one way that introduces ambiguity, as the relationship between the original substructures and the resulting bits is not easily discernible. Overall, molecular fingerprints have been successful in toxicity prediction tasks and capture certain aspects of the chemical structures being analyzed; they are able to encode wider aspects of the chemical structure than predefined keys.

## 2D DESCRIPTORS

2D continuous descriptors are numerical properties calculated from a molecule's 2D graph representation, typically derived from the connection table, molecular topology or chemical graphs, such as hydrophobicity, topological polar surface area[162] and E-state indices.[163] This level of information includes details about how atoms are connected within the molecule but does not account for the actual spatial arrangement of atoms in three-dimensional space. Physicochemical descriptors are often highly relevant to the effects of drugs within biological systems and they are readily interpretable, since they correspond to measurable properties of a molecule.[164] Further, 2D descriptors are more computationally efficient to calculate, given that they are derived solely from molecular connectivity matrices. In summary, using 2D descriptors for molecular representation can add information to chemical structural information, such as chemical fingerprints, and enhance predictive accuracy.[165] The quality of data, but then also problem-relevant and consistent preprocessing, is a key factor controlling model quality in cheminformatics models.

## 3D DESCRIPTORS

3D descriptors are calculated using the three-dimensional structure of a molecule, taking into account the spatial and geometric properties, including bond angles, distances, and overall geometry (Table 4). 3D descriptors can distinguish different three-dimensional conformations, but require accurate 3D structures, which are not always available or reliable, and the generation thereof leads to 'combinatorial explosion' (a large number of possible conformations). 3D information is needed in some cases such as for docking, where a spatial fit of the ligand into the target is performed.[167] Studies have reported 3D descriptors to be more effective than 2D descriptors for predicting biological targets of ligands with low structural similarity, given the ability of 3D descriptors to capture pharmacophoric alignments.[168] Overall, 3D descriptors can capture critical stereochemical and conformational information that can enhance predictions involving complex interactions and stereochemistry; however, they intrinsically need more computing resources to calculate, in particular, due to conformational sampling. When it comes to the information content that they contain, the signal-to-noise ratio needs to be considered (which does not always increase with descriptor complexity). Moreover, benchmark data sets often contain analogues that can already be well predicted by 2D descriptors, thereby intrinsically favoring 2D descriptors (and disfavoring 3D ones that are computationally expensive).

Among the challenges when incorporating 3D descriptors into ML models, the most common is the conformation-dependent nature of most 3D descriptors; this dependence results in substantial variations of the descriptors based on the conformer selected for descriptor calculation. In principle, an ML model could leverage a set of descriptors derived from an ensemble of conformations to capture the inherent flexibility of molecules. However, typical data sets do not map activity to different conformations and 3D representations are mostly incompatible with most contemporary ML algorithms, which are designed to map a single instance to a single label. Multi-Instance Learning (MIL) offers a solution by representing each


https://doi.org/10.1021/acs.chemrestox.5c00033
Chem. Res. Toxicol. XXXX, XXX, XXX−XXX

ligand as a bag containing multiple instances (conformers), and the task is to predict a property or activity associated with the entire bag rather than individual instances. MIL models are still in early stages of development and their application in toxicity prediction is yet to be validated.[169,170]

Studies combining 2D and 3D descriptors sometimes yield better performance due to their complementary information.[171] In theory, 3D information such as stereochemistry is responsible for the selectivity in binding affinity of molecules to many protein targets.[111] In practice, for certain bioactivity prediction tasks, 2D descriptors might outperform 3D due to a more favorable signal-to-noise ratio in descriptor space (e.g., fewer irrelevant conformations), and/or possibly the analogue bias in benchmark data sets.[152,171,172] Thus, the choice between 2D and 3D descriptors is specific to the problem statement and data set's chemical space and hence needs situation-specific exploration.

## GRAPH-BASED REPRESENTATIONS

Graph Neural Network (GNN) architectures for molecular toxicity and activity prediction have emerged as an often more expressive alternative than using binary fingerprints with classical ML models (Figure 4b). GNNs can directly learn molecular representations from chemical structures on the fly as the model is trained, which eliminates the need for manual feature engineering and enables customizing representations to specific data sets or tasks. These representations are learned through iterative rounds of so-called 'message passing' (Figure 4b). Briefly, molecules are initially featurized as graphs, with atoms as nodes and bonds as edges, and atoms and edges are initially described by simple physicochemical properties. While 2D GNNs use only simple atom and bond properties to provide node and edge features, 3D GNNs (Figure 5) additionally

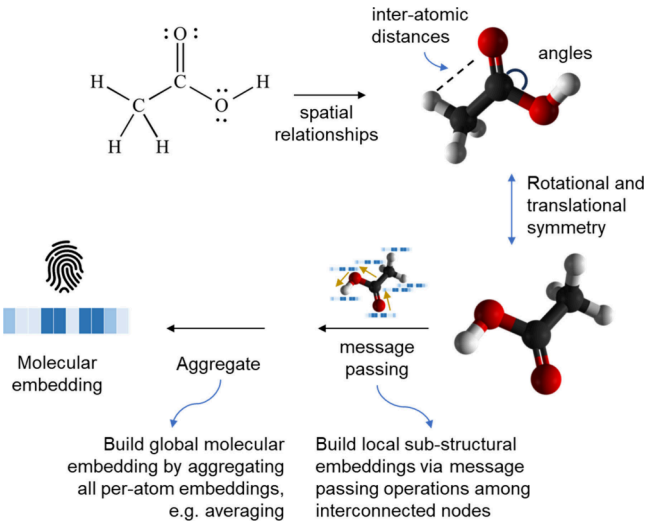



**Figure 5.** 3D Graph Neural Network (GNN) architectures applied to molecular representations. 3D geometric GNN representations incorporate spatial relationships and geometric transformations to capture three-dimensional information, such as bond angles.

incorporate 3D coordinates, providing spatial information for atom and edge features, and may leverage relative coordinates in their message passing rules to capture the molecule's geometric configuration. Then, atom and bond representations are made progressively more abstract by transforming and combining the descriptors of those atoms and bonds that are neighbors in the graph. In this way, a message passing on the graph makes atoms and bonds incorporate information about their chemical environment. Finally, the representations of atoms and bonds can be combined to obtain representations for the whole molecule.[179,180] Additional chemical- and surface-related information can also be provided as initial input features to GNNs. Alternatively, transformer architectures have also emerged as a powerful approach for graph-based modeling, with a flexible mechanism to capture long-range dependencies in molecular structures.[181,182]

The use of message passing in GNNs ensures that all operations performed by the model are independent of any particular ordering of the constituent nodes, leading to model reasoning that is purely based on the aggregated information from a node's local neighborhood. GNNs can serve as an encoder that processes a molecular graph, with nodes representing atoms and edges representing bonds (Figure 5), to generate a compact vector capturing the chemical structural information in the context of the pretraining task.[183] The GNN, pretrained on large chemical spaces, refines these representations to better capture features influencing target outcomes like compound toxicity. Pretraining can be supervised, using labeled data sets to predict molecular properties, or semisupervised, leveraging labeled and unlabeled data to refine latent space representations. Message-passing neural networks (e.g., Chemprop) have been proposed as encoders to generate latent space representations of molecules[180] and used to predict various properties including toxicity (discussed later as representation learning).[184,185] An emerging class of geometric GNN architectures for molecular toxicity prediction represents molecular graphs in three-dimensional space and additionally incorporates inductive biases that make them invariant to global rigid transformations of molecules—rotations and translations of the 3D coordinates.[186−188]

Gao et al. showed that models with access to three-dimensional structural information on the protein−ligand complex outperform approaches that use 2D fingerprint representations for the prediction of protein−ligand binding affinity.[152] Broadly, for those situations where the binding pose is relevant for an effect, using three-dimensional binding pose-dependent properties may provide a more comprehensive representation of the compound (albeit with the need to handle conformational information). However, protein flexibility is a significant factor in scenarios where binding pockets are less defined or allosteric binding occurs; the conformational changes in the protein induced by ligand binding can significantly affect the binding mode and affinity. For example, enzymes like cytochrome P450s (CYPs) as well as transporters exhibit a high degree of flexibility, further complicating the accurate prediction of ligand−protein interactions.[189] 3D ligand information is also being used in protein−ligand cofolding, which then predicts binding, and this is rapidly evolving with recent advances in AlphaFold[190] and RosettaFold.[191]

In reality, three-dimensional molecules have several possible conformers, with proteins and ligands existing in multiple conformations and ensembles, and their interactions can shift these conformational equilibria. As such, a single three-dimensional graph embedding may fail to consider molecules' dynamic behavior and flexibility, which can adopt various conformations in different environments, such as rotational freedom around single bonds. For example, the conformers of acetylcholine, such as the anti and gauche forms, demonstrate variable affinity to nicotinic and muscarinic receptor subtypes.[192] In contrast, only the anti form is preferred for catalysis


https://doi.org/10.1021/acs.chemrestox.5c00033
Chem. Res. Toxicol. XXXX, XXX, XXX−XXX

by acetylcholinesterase, an enzyme critical in the breakdown of acetylcholine and, thus, in regulating neurotransmission.[193,194] However, flexibility is not always critical to prediction for ML models—in majority of data, if an inhibitor fits the pharmacophore and effectively binds to inhibit the target, it may be sufficient to train an ML model without considering the molecule's flexibility. Where importance of molecules' dynamic behavior and flexibility is needed based on the specific context and the nature of the interaction being studied, one could use higher dimensional descriptors (multidimensional QSAR[195,196]) or use multiple instance learning (MIL), learning meaningful molecular representations relevant to prediction tasks from multiple ligand conformers of three-dimensional chemical structures.[197,198] By leveraging the diverse conformations that a molecule can adopt, MIL enhances the accuracy of toxicity predictions.[199]

Whether to use 2D or 3D geometric GNNs for molecular toxicity prediction is a question of data availability and the expressivity of representations. For tasks where the conformational state of a molecule is known to be important, such as protein−ligand interactions, 3D GNNs that incorporate protein−ligand poses from docking (providing information about the ligand's approximate binding mode) have been shown to outperform models based on 2D fingerprints that also encode protein−ligand extended connectivity, underscoring the critical role of 3D information in accurately predicting protein−ligand binding affinity.[200−202]

The advantage of 3D information compared to compressed representations has not been completely established in practice for the prediction of toxicity end points; Cremer et al. showed that 2D graph-based models produce comparable results to a 3D equivariant graph transformer model (which leverages geometry of conformers) and 3D geometry-based graph neural networks on ToxCast and Tox21 assays.[203] Still, the authors largely attribute the comparable results to the limited data set size in toxicity data sets; also toxicity can be driven by individual chemical substructures (e.g., reactive chemical groups, electrophilicity, etc.), where 2D descriptors capture relevant information rather well. The relative paucity of 3D protein structure data means that overfitting is common in more expressive geometric GNNs, leading to most published work in molecular toxicity prediction relying on standard 2D GNNs.[204]

## REPRESENTATION LEARNING

Representations can be learned on a given data set and then used as a feature on another prediction task, for example, in a strategy called transfer learning. Representations that have been pretrained for this purpose are relatively novel to cheminformatics compared to chemical fingerprints. Unlike hand-engineered or experimental descriptors, representation learning automatically extracts relevant features from raw data, capturing complex patterns and relationships, and is particularly powerful when handling large data sets.[106] By contrast, they are often not beneficial on smaller data sets.[205] Encoders are typically pretrained on related tasks to extract relevant features from raw data into a lower-dimensional and abstract representation known as the latent space. The compressed vector space retains essential features, making it easier for the models to handle complex data. Learned representations can be used downstream as feature vector inputs to another model or can be fine-tuned for a specific task.

There are various strategies to learn representations, such as message passing neural networks, discussed in the prior section. Another approach to automatically extract relevant features from raw data involves comparing and contrasting samples (including the possibility of using 3D information).[170,206] This method, known as contrastive learning, trains a model to distinguish between similar and dissimilar pairs of data points, typically by minimizing the distance between representations of similar pairs (positive pairs) and maximizing the distance between representations of dissimilar pairs (negative pairs). Thus, contrastive learning involves creating an embedding space where similar data points are located close to each other and dissimilar points are distant. For example, in the context of compound bioactivity data, the model might be trained to recognize that two different chemical scaffolds with similar bioactivity should have similar representations, while scaffolds with different bioactivity should have distinct representations. In addition to the strategies above, another potent method for learning representations involves using neural machine translation between two semantically equivalent but syntactically distinct molecular structure notations, such as InChI and SMILES. This technique, based on continuous and data-driven descriptors,[207] compresses the shared substantive information from both notations into a condensed, information-rich vector representation.

One of the advantages of representation learning is its ability to integrate multiple data types, such as chemical structures, biological activities, gene expression, and phenotypic profiles.[208,209] Liu et al. introduced the Information Alignment (InfoAlign) approach, which uses the information bottleneck method to learn enhanced molecular representations by integrating chemical structure data with cell morphology and gene expression data.[209] On the Biogen ADME data set, InfoAlign was tested across five end points: MDR1-MDCK efflux ratio (ER), solubility at pH 6.8, rat liver microsomal intrinsic clearance, human plasma protein binding (hPPB) percent unbound, and rat plasma protein binding (rPPB) percent unbound. InfoAlign outperformed traditional chemical fingerprints and other contrastive learning methods like CLOOME[210] and InfoCORE[211] reducing mean average errors by 6.33%, and thus learning better representations of molecules by integrating biological data.[5] Overall, representation learning allows the integration of various task-relevant biological data in encoding chemical space.

## FEATURE SELECTION OR REDUCTION

Feature selection or reduction is a key step in data analysis and modeling, serving as a way to reduce model variance (via the elimination of noisy features[212]) or bias (via the inclusion of relevant features).[213] Thus, feature selection methods can enhance the signal-to-noise ratio and enable the model to concentrate on the most significant predictors. An effective and efficient feature selection method considers both feature relevance (high information content) and redundancy (correlations between features).[214] Including irrelevant or redundant features can add noise (albeit models like partial least-squares regression, as one of few exceptions, can still work with highly redundant features), causing the model to overfit by learning the noise rather than the true underlying patterns resulting in poor generalization of new data and decreased model performance.[214] However, eliminating useful features during the selection process degrades performance by reducing the model's ability to capture important patterns or relationships in the data. The difficulty in practice is that data sets are limited in size, and hence feature relevance can only be assessed on the given data;


https://doi.org/10.1021/acs.chemrestox.5c00033
Chem. Res. Toxicol. XXXX, XXX, XXX−XXX

**Table 5. Table of Commonly Used Methods for Feature Selection (from the set of original features) and Reduction (new features)**

| Method | Description | Strategy for Selection/Reduction | Advantages | Disadvantages | Category | Reference |
|---|---|---|---|---|---|---|
| Minimum Redundancy Maximum Relevance (MRMR) | Selects features highly correlated with the target but minimally redundant with each other | Relation to output space | Balances relevance and redundancy, suitable for high-dimensional data | May struggle with nonlinear relationships, computationally expensive for large data sets | Supervised, Filter | 218 |
| Recursive Feature Elimination (RFE) | Iteratively removes features by training models and evaluating their performance | Relation to output space | Provides an optimal subset of features, integrates well with different model types | Computationally expensive for large data sets | Supervised, Wrapper | 219 |
| Mutual Information | Measures the information gained about one variable through the other, useful for feature ranking | Relation to output space | Nonparametric, can capture nonlinear relationships | Computationally expensive, sensitive to noise in the data, can be inaccurate as well, because the true MI often involves a very high-dimensional integral[225] | Supervised, Filter | 220 |
| Boruta | Iteratively compares feature importance with that of randomly shuffled features to determine significance | Variance in input space and relation to output space | Robust to overfitting, easy to interpret | Computationally expensive, may select more features than necessary | Supervised, Wrapper | 221 |
| Least Absolute Shrinkage and Selection Operator (LASSO) | Applies L1 regularization to linear regression to penalize coefficients, forcing some to zero | Relation to output space | Simple to implement, handles high-dimensional data well, and reduces overfitting | May select too few features if high correlation exists among predictors; difficult to interpret because it is an emergent property of the model fitting and not based on a metric calculated per-feature | Supervised, Embedded | 222 |
| Elastic Net | Combines L1 (LASSO) and L2 (ridge) regularization to select relevant features and handle correlated predictors | Relation to output space | Handles correlated features well, useful for data sets with more predictors than observations | Can be difficult to tune the mixing parameter between L1 and L2 regularization | Supervised, Embedded | 226 |
| ReliefF | Estimates feature importance by repeatedly sampling instances and comparing nearest neighbors | Variance in input space and relation to output space | Captures interactions between features, works well with noisy data sets | Sensitive to the choice of nearest neighbors, may not perform well with imbalanced data | Supervised, Filter | 227 |
| Genetic Algorithms | Evolves a population of potential solutions (feature subsets) using genetic operators like mutation and crossover | Natural selection and relation to output space | Flexible and adaptable, can escape local optima in feature selection | Computationally expensive, requires careful tuning of parameters like mutation rate | Unsupervised, Embedded | 228 |
| Principal Component Analysis (feature reduction method) | Dimensionality reduction technique that projects data onto orthogonal axes to maximize variance. Lower-ranking features can be discarded. | Variance in input space | Reduces dimensionality, improves computational efficiency, and removes multicollinearity | Information loss occurs, and the principal components often lack direct interpretability | Unsupervised, Filter | 223 |



https://doi.org/10.1021/acs.chemrestox.5c00033
Chem. Res. Toxicol. XXXX, XXX, XXX−XXX

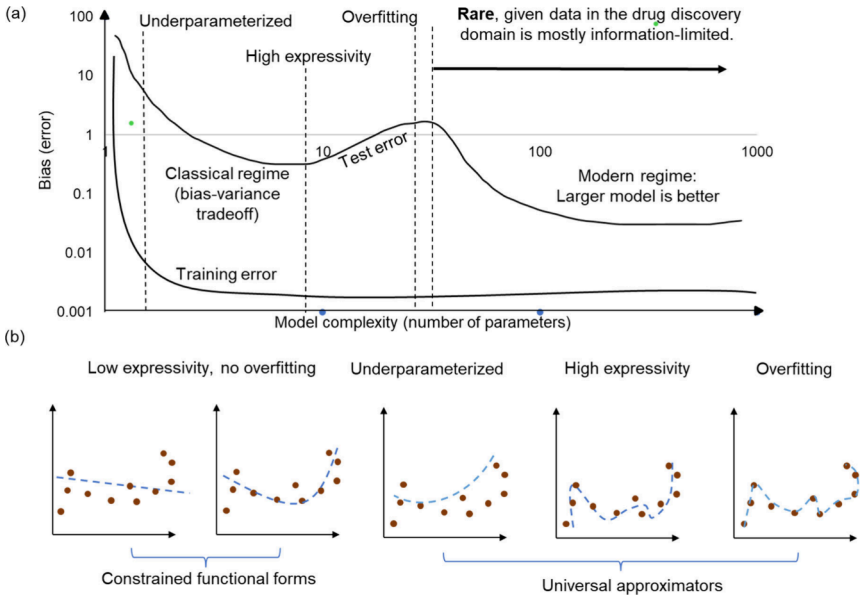



**Figure 6.** (a) Theoretical depiction of how increased model complexity reduces bias but increases variance, highlighting the tipping point where overfitting begins. Yet, recent research reveals a "double descent" regime (toward the right), as described in the main text. (b) Data points (in brown) alongside fitted models (blue dashed lines) to visualize the difference in the fit quality across various model complexities, from simple models with constrained functional forms to neural networks that are universal approximators with high expressivity but prone to overfitting.

relevance of features might change when new data becomes available.

Feature selection or reduction (as well as any optimization of model hyperparameters) should be done on the training data only, while the test set, on which future performance of the model is estimated, should not be used for feature selection; this would represent information leakage.[215,216] Feature selection methods can evaluate features based on statistical criteria without involving a predictive model ('filters'), use a machine learning model to iteratively select the best subset of features by testing performance ('wrappers'), or be embedded within the learning process of a predictive model ('embedded methods').[217]

Several methods address the feature selection step (Table 5).[160] For instance, Minimum Redundancy Maximum Relevance (MRMR) aims to select features that are highly correlated with the target variable while being minimally redundant with each other.[218] Recursive Feature Elimination (RFE) iteratively removes less significant features based on model performance, providing an optimal subset of features and integrating well with different model types.[219] Other techniques like Mutual Information measure the 'information gain' provided by each feature (that is, information provided about the target variable), capturing nonlinear relationships and aiding in feature ranking.[220] Boruta is an all-relevant feature selection method: that is, selecting all features with a meaningful relationship to the target, whether redundant or unique. Boruta iteratively compares the importance of actual features with that of randomly shuffled features to determine significance, making it robust to overfitting and easy to interpret.[221] Context-dependent sparse feature selection methods, such as LASSO and other L1 methods, help remove low-relevance features, which minimizes overfitting and improves model interpretability and predictive power.[222] Principal Component Analysis (PCA) is often employed for feature (dimensionality) reduction by selecting a finite number of components that explain the majority of the *variance*, but it does *not* explicitly select features based on their relevance to a specific task, such as predicting a target variable such as toxicity. While some have proposed using PCA for feature selection,[223] it is usually not ideal for feature selection where task-specific relevance is key.[224] Overall, in ML models, feature selection or reduction can improve the signal in the data by removing irrelevant or redundant features, thereby enhancing model performance as well as simplifying model interpretation due to the reduced number of features (provided that suitable features exist in the first place).

## PILLAR 3: THE CHOICE OF MODEL ALGORITHM

After choosing a suitable representation of chemicals, the choice of modeling algorithm is the next critical factor determining the performance and usefulness of an ML system for molecular toxicity prediction. In the context of ML, a model is a mathematical system designed to make predictions based on input data. Models use a set of features (as outlined in the previous section) to capture the essential characteristics of the molecules under study.[143] Most models make predictions for a data point (chemical) by using that data point's features and a set of model parameters that need to be fitted (trained) to the data. In a simple linear model, these parameters determine the weight of each descriptor in making a prediction; however, in more complex models, those relationships can be more difficult to interpret. Fitting a model involves adjusting these parameters to minimize the error between the model's predictions and observed outcomes from the training data. For a review on classical ML methods, the reader is referred to Mitchell[229] and Lavecchia et al.[230]

## BIAS-VARIANCE TRADE-OFFS

As with selecting a feature representation, the choice of ML models also presents a trade-off between having lower prediction power (or poor ability to generalize to novel chemical space) and overfitting (the ability to fit the training set well, but at the cost of being less able to generalize to unseen structures). Usually, ML models with constrained functional forms can only


https://doi.org/10.1021/acs.chemrestox.5c00033
Chem. Res. Toxicol. XXXX, XXX, XXX−XXX

**Table 6. Evaluation of Commonly Implemented ML Algorithms Concerning Their Simplicity, Interpretability, Robustness to Noise, Functional Form, Expressivity, and Bias-Variance Trade-Off When Used in Toxicity Prediction**

| ML Model Type | Complexity and Interpretability | Robustness to Noise | Functional Forms Approximated | Expressivity (ability capture complexity, variability of data) | Bias-Variance Trade-Off |
|---|---|---|---|---|---|
| Linear Models | Low complexity; high interpretability | Low | Linear relationships | Low | High bias, low variance—good for simple relationships but may underfit complex data |
| *k*-Nearest Neighbors (kNNs) | Low complexity; high interpretability[238] | Low | Local properties; distance-based | Depends on the input data; usually high | Low bias, high variance (with the small *k*-values that are typically necessary in moderate data regimes) |
| Decision Trees (DTs) | Moderate complexity; high interpretability | Moderate | Hierarchical decision rules | Moderate | Can vary; e.g., deep trees have low bias but high variance |
| Random Forests (ensembles of DTs) | High complexity; moderate interpretability | High | Ensemble of decision rules | High | Reduced variance through averaging, better generalization, and flexible control of bias and variance, but can overfit if not properly tuned |
| Gradient Boosting Machines (GBMs, typically ensembles of DTs) | High complexity; moderate interpretability | High | Ensemble of decision rules | Very high | Flexible control of bias and variance but can overfit if not properly tuned |
| Support Vector Machines (SVMs) | Moderate complexity; low interpretability | High | Linear and nonlinear (with kernels) | High | Good balance, but the choice of kernel affects the complexity |
| Neural Networks (NNs, multilayer perceptron/feedforward) | Complexity and interpretability depend on architecture (Shallow NNs have single hidden layers); while Deep NNs have multiple hidden layers, and hence higher complexity and lower interpretability than classical models | Depends on architecture (dropout, regularizers) | Virtually any functional form | Very high | Flexible control of bias, high variance—powerful but prone to overfitting (requires regularization), requires large data sets |
| Graph Neural Networks (GNNs) | High complexity; low-to-moderate interpretability | High | Complex graph structures | Very high | Balances bias and variance through structure-aware processing; can be prone to overfitting with highly expressive models, requires large data sets |



https://doi.org/10.1021/acs.chemrestox.5c00033
Chem. Res. Toxicol. XXXX, XXX, XXX−XXX

**Table 7. Reviews of the Predictive Model Algorithms Used in Predicting Toxicity**

| Review | Year | Algorithms Described/Discussed | Reference |
|---|---|---|---|
| Mitchell et al. | 2014 | Discusses common supervised learning algorithms: Artificial Neural Networks (ANNs), Random Forest (RF), Support Vector Machines (SVMs), *k*-Nearest Neighbors (kNNs), and Naïve Bayes (NB) classifiers. | 229 |
| Ekins | 2014 | Discusses Bayesian models, SVM, kNN, and RF for predicting various toxicities, such as hepatotoxicity and cardiotoxicity. | 241 |
| Lavecchia et al. | 2015 | Covers algorithms like SVM, Decision Trees (DT), RF, NB classifiers, kNN, and ANN. | 230 |
| Raies and Bajic | 2016 | Discusses rule-based systems, structural alerts, read-across methods, dose−response models, pharmacokinetic/pharmacodynamic models, and QSAR models. | 242 |
| Baskin | 2018 | Explains methods like multiple linear regression, kNN, SVM, DT, RF, and deep learning, along with unsupervised methods like Kohonen's self-organizing maps. | 243 |
| Yang et al. | 2018 | Describes ML approaches for predicting chemical toxicity, including SVM, RF, deep learning, and structural alerts for toxic substructure identification. | 244 |
| Vamathevan et al. | 2019 | Discusses ML applications in drug discovery, highlighting target identification, clinical trials, and challenges like data quality and model interpretability to reduce clinical failure rates. | 245 |
| Ciallella et al. | 2019 | Reviews AI applications in computational toxicology, covering data-driven and mechanism-driven models such as Adverse Outcome Pathways using public data sets and high-throughput screening. | 246 |
| Jiménez-Luna et al. | 2020 | Focuses on deep learning, explainable AI (XAI) techniques like feature attribution, gradient-based methods, surrogate models, and instance-based approaches for drug discovery model interpretability. | 247 |
| Wang et al. | 2021 | Covers regression models, kNN, DT, NB, SVM, RF, ensemble learning, ANN, Deep Neural Networks (DNNs), and CNN for modeling. | 248 |
| Dara et al. | 2022 | Describes methods such as SVM, RF, Multi-Layer Perceptron (MLP), deep learning, Autoencoders, and Reinforcement Learning in drug discovery. | 249 |
| Cavasotto et al. | 2022 | Reviews recent ML advances in toxicity prediction, noting challenges, methods, and relevant databases. | 57 |
| Tran et al. | 2023 | Focuses on key toxicity properties (e.g., hERG inhibition, drug-induced liver injury) and ML models like RF, SVM, and DNN in toxicity prediction. | 250 |
| Guo et al. | 2023 | Covers ML and deep learning models like SVM, RF, kNN, ensemble learning, MLP, Convolutional Neural Networks (CNNs), and Graph Convolutional Networks (GCNs) for toxicity prediction. | 251 |
| Tonoyan et al. | 2024 | Highlights opportunities in supervised, unsupervised, and reinforcement learning for toxicology applications. | 252 |

approximate a limited set of target functions, limiting their expressivity but preventing overfitting, which hence represents a trade-off (Figure 6). For example, linear regression models assume a linear relationship between the input features and labels, which limits the expressivity of the model to linear functions. Thus, linear regression models can be insufficient for capturing the complexity and variability present in domains where nonlinear patterns are common but can be a useful choice when data is scarce and have the advantage of being readily interpretable. On the other hand, deep learning models, such as multilayer feed-forward neural networks are more flexible (universal) approximators and can capture complex nonlinear functions but require sufficient size, data, training, and hyperparameter tuning (network architecture, etc.), and come at a cost of interpretability.[231]

The trade-off between poor predictive power and overfitting is a fundamental principle of ML known as the bias-variance trade-off (Figure 6).[232] Both bias and variance relate to the model's ability to learn feature-target relationships from the training data set. Bias refers to the model's tendency to not capture (and hence overgeneralize) real feature-target relationships that are present in the training set. Bias can be high when a simple model is used, such as a linear model or a small neural network, or when the model or training algorithm includes assumptions about the shape and nature of the data that may not be true for the data seen during deployment. Variance, on the other hand, refers to the model's tendency to capture spurious relationships in the training data set (such as noise). For example, the presence of a rare functional group in a single toxic compound (where this functional group, however, is not related to toxicity) might lead the model to incorrectly predict that other compounds with this functional group are toxic. Overly simplistic models exhibit high bias and fail to learn from the data adequately but are less likely to overfit to noise or spurious correlations in the training data and can have the ability to extrapolate better to new data. Due to their simplicity, interpretability, and robustness to noise, classical models, such as linear models (Table 6), have been widely used with binary fingerprints to predict toxicity.[1,233] In contrast, complex models can model intricate relationships but are more likely to learn spurious correlations and overfit.[234]

Recent deep learning research has challenged and modified the traditional understanding of the bias-variance trade-off by introducing the concept of double descent (Figure 6a).[232,235] When a model becomes too complex, it initially leads to overfitting and high variance (Figure 6b). As the number of parameters approaches the number of observations, small changes in the data cause large changes in the model. However, as complexity is increased further, the error surprisingly decreases again, an effect known as 'double descent'.[235] At some point, highly overparametrized models can hence often generalize well, despite having at the same time a large number of parameters. Double descent challenges the traditional U-shaped curve of bias-variance trade-off, revealing a more nuanced relationship between model complexity, training time and data set size with the generalization error.[236] Nevertheless, its relevance to toxicology data, especially in low-data scenarios, may at the current point in time be limited. There are only a few instances where this effect has been observed in the drug discovery domain. For example, when training NeuralDock to predict the binding energy and affinity of a protein-small molecule pair based on protein pocket 3D structure and small molecule topology, the authors observed the second descent because the 46 million parameters was much higher than the roughly 2000 data points used in training; the authors noted that the second descent contributed to its high accuracy.[237] However, in smaller data sets typically used in ML models for compound toxicity, encountering double descent is unlikely.

Algorithms described in the prior section (Table 6) are widely employed in molecular and toxicity prediction (Table 7), and are based on features like molecular structure, chemical


https://doi.org/10.1021/acs.chemrestox.5c00033
Chem. Res. Toxicol. XXXX, XXX, XXX−XXX

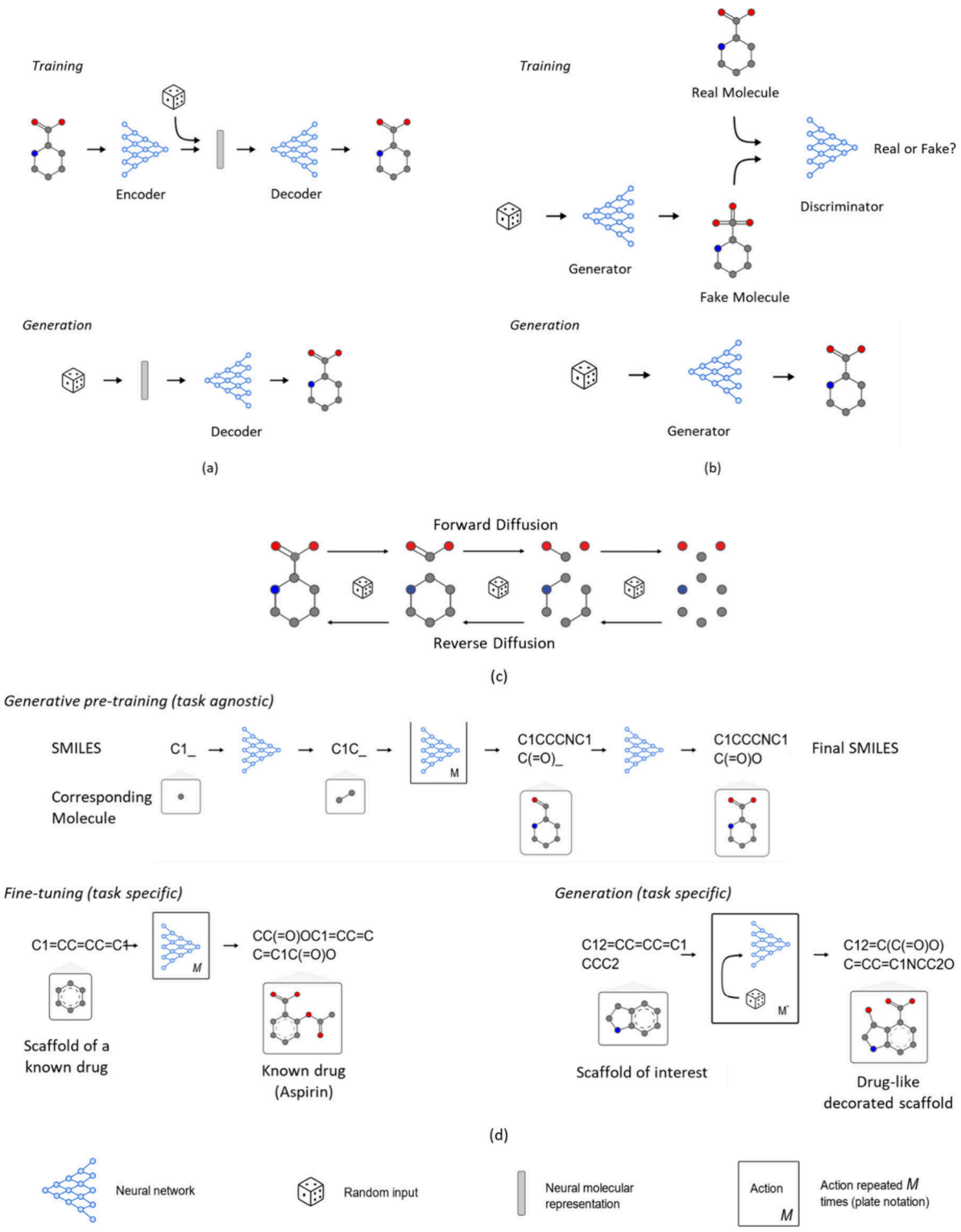



**Figure 7.** (a) Variational Autoencoders (VAEs) encode molecules into a latent space representation and then decode to generate new molecules, aiding in the exploration of chemical space. (b) A conceptual framework behind Generative Adversarial Networks (GANs), illustrating how a generator creates new molecule designs and a discriminator evaluates their realism, facilitating the generation of novel compounds. (c) Diffusion Models, a class of generative models that learn to generate data by transforming noise. (d) Large Language Models (LLMs) are also used to design molecules, emphasizing their ability to predict molecular properties and generate new compounds based on learned patterns.

properties, and biological data. For further reading, we recommend the compendium of articles in "Introduction to the Special Issue: AI Meets Toxicology" and "Analysis of Tox24 challenge results", which offers a comprehensive exploration of AI/ML applications in toxicology, presenting advanced methodologies and diverse case studies that showcase the full potential of these technologies.[239,240]


https://doi.org/10.1021/acs.chemrestox.5c00033
Chem. Res. Toxicol. XXXX, XXX, XXX−XXX

## READ-ACROSS MODELS

While in some cases the quantitative or qualitative prediction of a compound property across all chemical space and all output property space is desired, especially in the safety area, sometimes other approaches are employed. One such approach is known as 'read-across', which relies on structural similarity *and contextual information* to predict the toxicity of a chemical substance.[253,254] Read-across is particularly popular due to its application context and the legal environment, which impose distinct requirements, such as interpretability and confidence in assigning compound properties. The process involves first identifying molecules that belong to the same chemical series or category as a query compound, based on the assumption that structurally similar compounds will exhibit similar biological activities due to shared mechanisms of action. Once appropriate analogs are identified, toxicological data from these substances are "read across" (data gap filling) to the target chemical to fill data gaps and inform safety assessments. A $k$-nearest neighbor chemical structure-based strategy (often with $k = 1$) is one example of a read-across strategy. Given the regulatory environment and cultural practice in toxicology, 'machine learning is not everything': approaches can and should be tailored to the particular use case.

## GENERATIVE MODELS

Related to *predictive* models are *generative* models, which in addition to learning patterns from a training set are able to generate novel chemical structures and which can include toxicity as one factor for optimization. Unlike predictive models, where one predicts the label for a given molecule, generative models hence 'design' the molecule itself given the desired label.[255] These models are typically optimized over iterations using scoring functions that predict desired properties including novelty, activity, and synthesizability in addition to toxicity.[256,257] Recently, generative models have been used for this, including Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), and Large Language Models (LLMs).[256,258]

## VARIATIONAL AUTOENCODERS

Variational autoencoders (VAEs) are a class of generative models that learn representations through compression and reconstruction (Figure 7a).[259] In a classical autoencoder, an encoder network maps data points from a high-dimensional input space (e.g., 2D or 3D molecular structure space) to a lower-dimensional latent space. Then, a decoder network is reconstructed back to the original input space, attempting to recover the data points that were fed into the encoder network. In contrast, in a VAE, the encoder maps a point from the input space not to a specific point in the latent space but to the parameters of the Gaussian distribution in the latent space, and the decoder tries to restore the point in the input space based on the result of sampling the Gaussian distribution. The presence of the sampling stage makes the VAE a generative model. Since reconstruction is required, but at the same time the latent space is smaller than the input space, the VAE encoder network must learn a compressed representation of the data that captures its most important features and ignores irrelevant ones. Because the ideal distribution in the trained latent space is known, novel molecules can be generated with the trained model by sampling points in the latent space and passing them through the decoder.

VAEs are popular general models for molecular representation learning and generation.[260,261] Molecular generation with VAEs can be biased to optimize various objective functions such as docking scores[262,263] and drug-likeness.[264,265] In the context of toxicity prediction, conditional VAEs[265] have been applied to explore low-toxicity regions of chemical space, using a scalarized objective that combines labels related to different forms of toxicity (cardiotoxicity, mutagenicity, pulmonary toxicity and skin sensitivity).[255] Other studies have used arithmetic in the latent space to predict polypharmacology, adding or subtracting latent representations corresponding to treatment with different substances in order to simulate cellular states (as described by cell morphology and gene expression) when subjected to multiple compounds.[266]

## GENERATIVE ADVERSARIAL NETWORKS

Generative Adversarial Networks (GANs) are a class of generative models consisting of a generator and a discriminator, two neural networks, that are trained simultaneously through adversarial training.[267] A conceptual framework behind a GAN is depicted in Figure 7b. The generator network in Figure 7b is trained to create novel molecular structures, while the discriminator network evaluates whether these structures are realistic. At each training iteration, the discriminator is shown a batch of molecules that either were sampled randomly from a set of real molecules ("real") or were artificially generated by the generator ("fake"). Using these examples, the discriminator must learn to determine whether a molecule is real or fake. In turn, the generator must learn to fool the discriminator by proposing molecules that mimic those in the real set. We can bias the generator to suggest compounds with certain properties, for example low toxicity, by designing the set of real molecules so that it features those properties.

Recent examples of works that explore the application of GANs to toxicity in medicinal chemistry are MedGAN, ToxGan and TransOrGAN. MedGAN was developed to generate new quinoline scaffold molecules from complex molecular graphs.[268] The best model included in this study generated 25% valid compounds, of which 92% were quinolines of up to 50 atoms. Around 22−31% of the generated molecules were predicted to be nontoxic across the 12 Tox21 end points.[269] That being said, the generation of quinolines is a task also easily achievable by a medicinal chemist, and the impact of such models in practice remains to be seen. In another study, ToxGan was used to predict new animal study results from historical data.[270] Instead of training on molecular structures, the authors used repeated dose transcriptomic profiles from *in vivo* rat studies (Open TG-GATES) to train a GAN that generated transcriptomic profiles for compounds of interest. They could generate highly similar transcriptomic profiles that showed over 87% agreement in Gene Ontology with the actual gene expression data. Finally, TransOrGAN allowed molecular mapping of gene expression profiles based on rat RNA-seq data from 288 samples in nine different organs of both sexes and four developmental stages.[271] By inferring transcriptomic profiles between any two of the nine organs studied, TransOrGAN achieved an average cosine similarity of 0.984 between synthetic transcriptomic profiles and the corresponding real profiles.

The success of frameworks like MedGAN, ToxGan, and TransOrGAN demonstrates the principle of GANs in molecular design, offering a promising avenue for generating novel, drug-like compounds taking all available data into account in an unbiased fashion. As with other ML models, however, there are some ethical concerns with generative AI, as designing


https://doi.org/10.1021/acs.chemrestox.5c00033
Chem. Res. Toxicol. XXXX, XXX, XXX−XXX

molecules with the reverse loss function has been shown to rediscover potent chemical toxicants.[272]

## DIFFUSION MODELS

Diffusion models have gained significant interest in drug discovery recently.[273] They are a class of generative models that learn to generate data by transforming noise.[274,275] Rather than training a generator to fool a discriminator, diffusion models learn to transform noise to data by learning to reverse a diffusion process that corrupts the data. The diffusion process, which is chosen by the ML practitioner and is not trained, starts with a point from the data set and progressively corrupts it to create a sequence of increasingly noisy iterates (Figure 7c). The process is chosen such that the final iterate is completely noise, i.e., has no information from the original data remaining such that it does not depend on the initial data point and has a distribution that can be easily sampled from. A neural network is then trained to reverse the diffusion process by predicting the previous (less noisy) iteration from any of the noisy iterates. This neural network is used to generate data by starting with a sample from the noise distribution and progressively predicting each previous, less-noisy iteration until it is left with an approximate sample from the original data distribution.

Diffusion models are used for molecular machine learning in two main ways. First, a 2D molecular structure consisting of atom and bond types can be generated.[276] Second, given a pre-existing 2D molecular graph, the diffusion model generates 3D atom coordinates. For example, 3D structures can be generated for either low-energy conformers of the molecule,[277] or for other molecular arrangements such as a ligand's binding pose in a protein binding pocket.[278] These two approaches can also be combined to generate both 2D molecular graph and 3D structure at once.[279] In each case, the diffusion process (to corrupt the data) and the neural network (to reconstruct the data) must be designed to suit the structure of the data, whether they are discrete atoms and bonds or continuously varying atom coordinates.

Another important use of diffusion models is to avoid off-target toxicity by designing ligands that selectively bind to targets over specific off-targets,[280] although it remains to be seen that such model improvement also translates to future applications (e.g., novel chemical space). Overall, diffusion models hold promise for predicting and understanding molecule binding and biological pathways *in silico*; however, care needs to be taken to properly evaluate models in a prospective setting.

## LARGE LANGUAGE MODELS

A key trend in machine learning, especially with the rise of deep learning architectures, has been the scaling of expressive power and representation learning by harnessing available large-scale data sets. While traditional supervised learning relies on labeled data, which can be limited and costly to obtain, self-supervised learning enables models to learn from unlabeled data through carefully designed pretext tasks. These pretext tasks allow models to extract meaningful representations from vast amounts of real-world data regardless of annotation availability, which allows labeled data to be preserved for a later training stage where desired. This approach has been particularly transformative in language modeling, where popular models such as ChatGPT generate natural language sequences by effectively using unannotated text. Autoregressive models, such as Recurrent Neural Networks (RNNs), have also been applied to generate molecular representations like SMILES strings or molecular fingerprints (Figure 7c). These models work by predicting the next token in a sequence based on the previous tokens, making them well-suited for tasks where the structure of the molecule is represented as a sequence. When trained on large data sets of molecular sequences, these models can learn the underlying patterns and relationships that define valid chemical structures, allowing the generation of novel molecules with similar properties.

Building on the foundation of autoregressive models, the rise of Large Language Models (LLMs) for Natural Language Processing (NLP)[281−283] has prompted applying similar architectures to molecular tasks given established conventions of treating molecular structures as sequences, especially for the purposes of generation and representation learning. LLMs are first pretrained on a large corpus of molecular data to learn general representations, which can then be fine-tuned for specific downstream tasks, such as molecular property prediction or reaction prediction. This two-step process—pretraining and fine-tuning—enables LLMs to generalize across different molecular tasks, making them a powerful tool for drug discovery and cheminformatics. LLMs for chemistry are sometimes called Chemical Language Models (CLMs). Many chemical databases describe and store molecules in sequence representations such as SMILES. This has facilitated the development of generative models that treat molecules as sequences,[260,261,284] some of which have achieved considerable success in academia and industry.[285]

The idea of pretraining and fine-tuning in earlier QSAR models (before LLMs were used in chemistry) was inspired by neurophysiology and is the basis of the ASNN method..[286] Fine-tuning approaches in QSAR (the so-called 'Library Model') increased accuracy of the logP model to predict logD7.4 values.[287,288] One LLM architecture that has been particularly successful in NLP is the GPT (Generative Pretrained Transformer) architecture. The GPT model consists of a stack of neural blocks, each of which is inspired by the decoder of the transformer.[289] This decoder block consists of a self-attention layer and a feed-forward layer, both wrapped in residual connections[290] to avoid gradient degradation during back-propagation due to the very deep architecture. The training of GPT is performed in two stages: pretraining and fine-tuning (Figure 7c).[281] During pretraining, the model learns to generate sequences from a training corpus of sequences one element at a time (since sequences are typically tokenized, elements are usually tokens in a token vocabulary; in NLP, tokens are common word fragments or subwords, whereas in molecular tasks, tokens may be common strings of characters in SMILES, e.g., strings of characters that correspond to common functional groups in SMILES representation). At each pretraining iteration, the model's parameters are modified to maximize the conditional likelihood of a certain token in a sequence given all previous tokens in that sequence. In this way, pretraining is carried out in an unsupervised manner and it does not require manual annotations, which are often costly to obtain. In the second stage, fine-tuning, the models' parameters are adapted to improve performance in downstream supervised tasks with labels. Fine-tuning to specific tasks can be done through several strategies, including full-parameter tuning, adapter layers,[291] or Low-Rank Adaptation (LoRA).[292] An example downstream task in NLP is sentiment analysis, where sentences are labeled with positive or negative sentiment labels. An example downstream molecular task is scaffold decoration (Figure 7d). In order to


https://doi.org/10.1021/acs.chemrestox.5c00033
Chem. Res. Toxicol. XXXX, XXX, XXX−XXX

teach a chemical language model how to decorate scaffolds to obtain drug-like candidates with predicted low toxicity, the scaffolds of approved drugs may be paired with drug-like candidates. PromptSmiles is a recent CLM that specializes in scaffold decoration and fragment linking.[293]

LLMs and CLMs hold great potential for molecular generative AI and property prediction, including exploration of predicted low-toxicity chemical spaces, taking information on all available data into account. Generative pretraining strategy enables this class of models to learn from large unlabeled databases (e.g., Enamine REAL SPACE[294]), which is suitable for the domain of medicinal chemistry where collecting experimental labels from toxicity studies involving animal or human clinical trials is costly. LLMs and CLMs have shown high flexibility and performance in chemical tasks. Natural-language models are able to capture information about chemistry and an ability to assist with chemical tasks in natural language, including synthesis, both with[295] and without[296] fine-tuning for chemistry. Surprisingly, they have also been shown to produce molecular representations competitive with featurizations such as fingerprints, even when trained on natural language rather than chemical sequences such as SMILES.[297] Specifically, representations from the NLP model LLaMA2[283] achieved a mean AUC-ROC of 0.77 on the toxicity classification benchmark Tox21 that was comparable to Morgan fingerprints that achieved a mean AUC-ROC of 0.76 and representations from the CLM MoLFormer-XL, trained on 1.1 billion SMILES and comparatively small (just 12 layers),[284] achieved a mean AUC-ROC of 0.78.[297] This suggests that models for natural language hold potential in toxicity prediction (at least in the way benchmarking has been performed in those studies). It should be kept in mind that validation on limited data sets such as the above does not necessarily translate to future use cases (e.g., in new chemical space) due to analogue bias in data sets, etc. In addition, it is worth noting that data limitations currently (and likely in the foreseeable future) prevent the development of extremely large models for medicinal chemistry trained on molecular sequences such as SMILES, in contrast to LLMs such as GPT and LLaMA that can be trained on natural language (and hence a much larger information corpus). This is because even unlabeled databases represent a comparatively small sample of the chemical space (which is in the order of $10^{60}$ as opposed to around $10^{10}$ to $10^{15}$ in current databases).[298,299]

## PILLAR 4: VALIDATION OF PREDICTIVE MODELS

While evaluation on validation sets is used to guide training in many ML models, once a model is trained on a data set, it must also be evaluated on an appropriate, clearly defined test set with the proper evaluation metrics.

The choice of the 'test set' is crucial here, since one aims to extrapolate to the performance in *future* use cases from the performance obtained on the test. On the other hand, given the size of chemical space, and that in most cases properties behave differently in different areas of chemical space, this is no trivial feat, as illustrated in Figure 8 below. While evaluation of the model is performed on an 'external test set' (which may not be entirely external, since it is derived from an existing, split data set), future use cases are by definition in *different* areas of chemical space, and hence performance extrapolation is in practice often *not* trivial to perform. The closer to the future use case, the better the choice of the test data set in general.

We discuss two major ways to assess the models in the following: (1) retrospective validation, which holds out test compounds from the same data source, for example, by splitting the set into training and test data, and (2) prospective validation, running separate experiments on a new set of compounds.[300] Beyond validation type, realistically assessing the performance of ML models in predictive modeling depends on selecting appropriate numerical evaluation *metrics* to understand the model's accuracy and generalizability.[33] The chemical space in which the model can make reliable predictions (known as the *applicability domain*[301]) should also be clearly defined following OECD principles (as discussed earlier in this work). We discuss various ways to define the applicability domain of models and common approaches to understanding models via feature importance and mechanistic analysis.

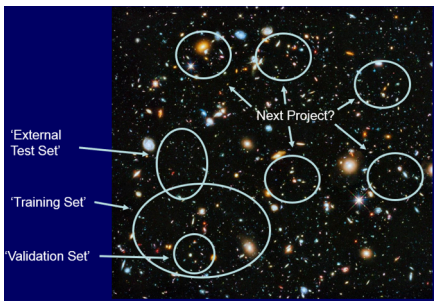



**Figure 8.** Chemical space visualized as a universe of molecules: stars represent compounds with clusters showing similar properties. The training set forms a constellation guiding predictions, while scattered stars of the test set highlight challenges in extrapolating to unseen regions, emphasizing the vastness and diversity of chemical space.

## CHOOSING EVALUATION METRICS FOR CLASSIFICATION

ML models must use relevant performance metrics that are calculated on definitive predictions, along with confidence measures, applicability domain measures, etc.[302,303] in order to be helpful for decision making. There are many evaluation metrics for predictive models; each emphasizes a different aspect (or a combination of aspects) of model behavior. Bender et al.[33] provide a list of recommended metrics in the evaluation of ML models in the chemical sciences. Here, we focus on some common metrics that should be considered when predicting assay outcomes from high-thoughoutput screens, toxicity-based assays, and regression outputs.

For binary classification models, predictions are classified into two classes (e.g., positive and negative negative). A confusion matrix helps to visualize the performance of a classification model by organizing predictions into four categories:

- True Positives (TP): Correct predictions of the positive class
- True Negatives (TN): Correct predictions of the negative class
- False Positives (FP): Incorrect positive predictions of negative class instances
- False Negatives (FN): Incorrect negative predictions of positive class instances

Some classification models produce a probability score for each instance. A threshold is applied to convert these probabilities into definitive class predictions (e.g., predicting a compound as toxic or nontoxic). By adjustment of a threshold, different trade-offs can be made between capturing more true positives or avoiding false positives.


https://doi.org/10.1021/acs.chemrestox.5c00033
Chem. Res. Toxicol. XXXX, XXX, XXX−XXX

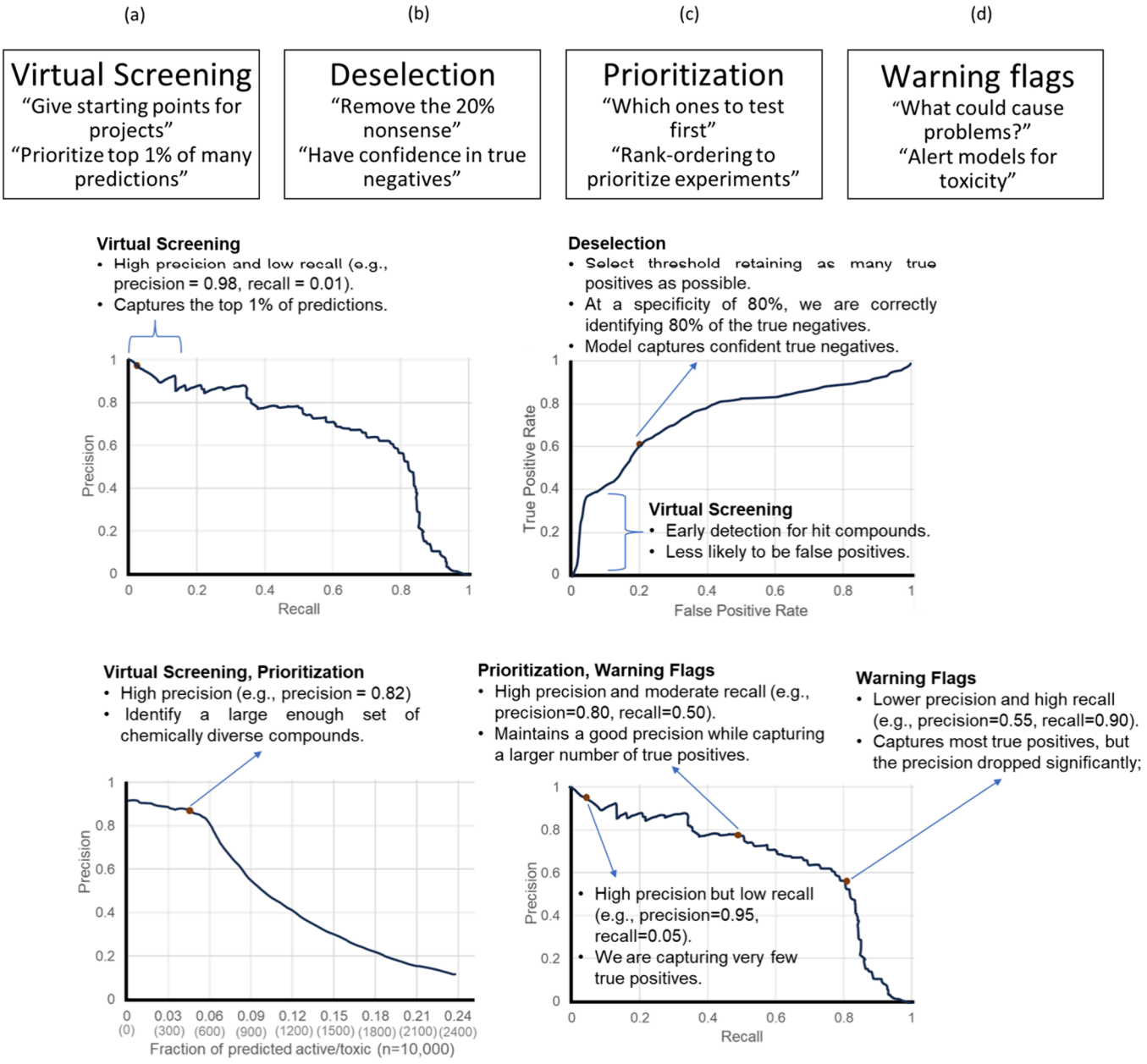



**Figure 9.** Evaluating ML model predictions in various settings: the primary goal is to prioritize specific predictions for that setting rather than rely on predicted probabilities across a range of possible thresholds. Strategies include (a) selection, (b) deselection, (c) rank-ordering, and (d) flagging toxic compounds. The choice of evaluation metrics in regression

From the confusion matrix, a range of metrics can be derived to evaluate different aspects of the model's performance, some of which we will describe in the following. Accuracy is a fundamental metric that measures the proportion of correct predictions (both true positives and true negatives) among the total predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

However, in contexts like drug discovery and toxicity prediction, data sets are often highly imbalanced, with significantly more inactive compounds than active ones. This imbalance poses challenges because a model could achieve high accuracy by predicting the majority class all of the time, offering little practical value. For example, if the inactive class comprises 99% of the total data set, a model solely predicting this class would be 99% correct overall, but the model would still be of no practical use (since it never predicts a compound as active for a given end point).

To address the limitations of accuracy in imbalanced data sets, metrics such as balanced accuracy becomes relevant, which consider the accuracy on the *class* level equally, i.e. this metric pays the same attention to both classes, independent of the number of data points they contain, thereby removing the above bias (while introducing of course another).

$$Balanced\ Accuracy = \frac{Sensitivity + Specificity}{2}$$

Precision measures the proportion of true positives among all positive predictions, emphasizing the model's ability to avoid false positives, and describing the 'trust' (likelihood of a positive data label) for any positive prediction the model makes.


https://doi.org/10.1021/acs.chemrestox.5c00033
Chem. Res. Toxicol. XXXX, XXX, XXX−XXX

$$Precision = \frac{TP}{TP + FP}$$

Recall (or sensitivity) assesses the proportion of true positives identified out of all actual positives, focusing on the model's capacity to capture, to retrieve, all relevant instances.

$$Recall = \frac{TP}{TP + FN}$$

The F1 Score is the harmonic mean of precision and recall. Given its ability to account for both false positives and false negatives, the F1 score is widely used and particularly useful in imbalanced data sets. On the other hand, it does not consider the true negatives in this measure, and the same F1 score can be obtained by very different precision and recall values, and hence model behaviors. It can be seen that every metric has particular trade-offs of which aspects of model performance it pays attention to.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Because annotated toxicity data sets are typically smaller than in other domains, chance plays a significant role in producing hits. Cohen's Kappa[304] measures the agreement between actual and predicted values, adjusted for chance agreement, making it valuable for imbalanced data sets. However, it weighs predictions of the positive and negative classes equally, whereas in toxicity prediction, false positives (overpredictions) are often tolerated more than false negatives (not detecting potential toxic compounds).

It is also common practice to visualize model performance using Receiver Operating Characteristic (ROC) and Precision Recall (PR) curves (Figure 9). The ROC curve displays the relationship between the True Positive Rate (TPR, also known as recall or sensitivity) and the false positive rate (FPR, 1 − specificity) across various thresholds. The Area Under the ROC curve (AUC-ROC) provides a summary of the model's overall performance. However, in imbalanced data sets, AUC-ROC can be misleading, as the model might achieve a high score simply by prioritizing the majority class. PR curves plot precision against recall at different thresholds. The Area Under the PR curve (AUC-PR) is especially useful for imbalanced data sets, as it focuses on positive class performance.

To translate predictions from models to decision making in drug discovery, it is often relevant to visually inspect the PRC and ROC curves obtained (Figure 9). Although often reported in benchmarking studies where large numbers of predictions and tasks are summarized, AUC-ROC and AUC-PR consider the *overall quality* of models based on a full range of thresholds for classification, so they are less helpful for a particular real-world use case where a particular decision needs to be made. In the real world, we may need, e.g., ML models in virtual screening (Figure 9a; early selection), removing the bottom 20% of least suitable predictions (Figure 9b; say compounds certain to be unsuitable due to high clearance), prioritizing specific predictions, such as identifying the top 1% of predictions (Figure 9c), or rank-ordering which compounds to test first (Figure 9c). For these decisions, it is often important that models should be evaluated based on the actual predictions of classes rather than model predicted probabilities.

For virtual screening[305] (Figure 9a), what matters most for ML models is prioritizing compounds from hit triage toward validation. From an evaluation perspective, in this setting a model needs to be better at early detection (high sensitivity at low false positive rates), and the modeler needs to determine the threshold for prediction according to the practical use case, such as the number of compounds that can be selected for experimental testing.[306] There are metrics, such as the Enrichment Factor, designed to emphasize the early recognition compared to the larger set of compounds considered in the ROC curve.[307] Enrichment factor refers to the ratio between the proportion of true positives among the top-ranked candidates and the proportion in the full screening set for some number or fraction of top-ranked candidates. A higher enrichment factor indicates that the model is effective at early detection, identifying active compounds among the top-ranked candidates. From an AUC-ROC perspective, the enrichment factor is primarily dependent on the initial segment of the curve. It is often more practical to work with absolute numbers, such as selecting 100,000 compounds for testing (or screening 100 × 96-well plates), to align with the specific capacity or resources available. Another way to achieve these objectives is by analyzing the regions of the PRC and ROC curves. For example, ranking the top 1% of compounds predicted correctly as active (true positives), one could identify the region of the PRC curve where the highest precision is achieved, while recall is relatively able to capture 1% of positives. Typically, the top 1% of compounds would be at the beginning of the PRC (Figure 9a), typically at the top left of the curve, where the model makes predictions with the highest precision because the model is most confident about these predictions. Since it is often hard to determine what parts of the PR or ROC curves map to percentage of compounds predicted active, another way to evaluate models for prioritizing compounds would be to plot precision as a function of the percentage of compounds predicted active (with exact numbers tested on a secondary axis).[308] Overall, we would want to identify a set of chemically diverse compounds, while a model that has high precision. To do so, modelers could choose a threshold that has acceptable precision and that is large enough to perform diversity selection.

Deselection involves setting a threshold on the ROC curve to correctly identify a large proportion of undesired compounds as true negatives while retaining as many desired compounds as true positives as possible (Figure 9b). To ensure a high recall of the undesired class (e.g., 90%), we aim to maximize true negative rate specifically subject to this constraint.

Prioritizing (or ranking) compounds based on their likelihood of being true positives involves analyzing the Precision-Recall Curve (PRC) to determine where precision and recall maximize the true positive rate (Figure 9c). The goal is to identify a set of compounds that is sufficiently large to support a selection of the desired size while also ensuring a high confirmation rate. This selection typically occurs in a region of the PRC where the model maintains good precision while recalling a significant number of true positives. An alternative way to evaluate the model's ability to rank compounds is to plot precision as a function of the number of compounds predicted to be positive (Figure 9), which is often easier to interpret.[308] For finer granularity, a rank-based metric such as Kendall's tau can also be useful.

For toxicity prediction tasks, when the goal is often to flag compounds with likely toxicity, prioritizing a higher recall of toxic compounds is often more appropriate, even if it results in lower precision, which is associated with more false positives (Figure 9d), because advancing toxic compounds is highly costly. This approach is related to a deselection setting, where


https://doi.org/10.1021/acs.chemrestox.5c00033
Chem. Res. Toxicol. XXXX, XXX, XXX−XXX

we aim for a high recall of desired compunds. On the other hand, if the model is used to deselect compounds, a high true negative rate (removing *mostly* toxic compounds) is desirable to preserve as many safe compounds as possible. It can be seen that models present a trade-off, and not all that is desirable can be obtained at the same time; being able to, say, have a higher recall of desired compounds comes at the cost of more false-positive predictions, etc. This underlines the importance of choosing a suitable performance metric and classification threshold for a particular use case, as opposed to trying to find a generic answer to the problem.

Models should also be compared with baseline models (such as majority class or average predictor or randomly shuffled labels using Y-Scrambling[309]), in order to establish how much value the proposed model provides. Baseline models can provide surprisingly good performance in many cases, such as on imbalanced data sets (where one can predict the majority class), or otherwise biased data sets, where the baseline classifier just picks up the underlying bias of the data set.[310]

For classification models, it is necessary to compare models to chance prediction and baseline models, in particular, given that even baseline models can obtain surprisingly good performance. Cohen's Kappa measures the agreement between actual and predicted values, adjusted for agreement occurring by chance.[304,311] In terms of applications, Cohen's Kappa was used as the metric to identify the optimal decision threshold that maximizes the balanced accuracy of the classifier predicting structure−activity data from 138 public data sets corresponding to pharmaceutical targets, addressing the issue of class imbalance, where the model may overpredict the majority class and underpredict the minority class.[312] The authors found that optimizing thresholds significantly enhances prediction outcomes for inhibitors of Tau fibril formation, particularly when the initial model is already well-predictive (as measured by AUC-ROC). This optimization improved the True Positive Rate (TPR, also known as recall and sensitivity) from 4% to 43% and Cohen's Kappa from 0.06 to 0.4. However, for predicting inhibitors of Marburg virus binding or entry into cells, where the initial model had lower predictive power, the improvements were more modest, with TPR increasing from 0.1% to 5% and Cohen's Kappa from 0.002 to 0.08. Where there are several classes to predict (such as low, medium, and high solubility), Kappa can be used to rank classification models; this was used in recent Kaggle solubility challenge.[313] Overall, Kappa-based optimization is recommended for machine learning classification models in toxicity prediction, particularly when the initial model performance is strong.

Overall, the threshold chosen to generate definitive model predictions from model-predicted probabilities is context-dependent for each task. The final model should be chosen based on the desired balance between the performance metric most relevant for a particular use case.

To evaluate a regression model, it is essential to check at least three aspects of the prediction: (a) correlation—how well the model captures the relationship between the real target values and predicted data, (b) goodness of fit—how well the model fits the data overall and how much of the variance in the output variable is explained by the model, and (c) dispersion of errors—how close the predicted values are to the actual target values. Distribution- and point-based metrics are generally used to evaluate regression models.

The simplest way to evaluate a regression model is by measuring correlation, which tells us about the strength and direction of the relationship between the model's predictions and the actual values. A common metric for this purpose is the Pearson correlation coefficient ($r$), which ranges from −1 to 1. If $r$ is close to 1, it indicates a strong positive relationship, meaning that the model's predictions closely match the actual target values. A value close to 0 (or negative) suggests little to no linear relationship or even an inverse relationship, indicating that the model may not effectively capture the trend in the data. The squared Pearson correlation ($r^2$) measures the proportion of the variance in the actual values that is predictable from the model's predictions. Another useful correlation metric is Spearman correlation, which is used when we are more interested in the ranking of values rather than their exact predicted values. Unlike Pearson correlation, Spearman correlation focuses on the rank order of the predictions compared to the actual values and is thus less sensitive to outlier values. This makes Spearman correlation useful for nonlinear relationships between predictions and actual target values, and in practice it can be a very useful metric to establish the ability to prioritize, e.g., compounds to be tested experimentally in an assay, where the number of compounds to select is fixed (and only depends on the ordered list of model output).

The second aspect of model evaluation is assessing the goodness of fit, which tells us how well the regression model fits the data points. There are multiple metrics commonly used to evaluate goodness of fit such as coefficient of determination ($R^2$), coefficient of determination on regression through the origin ($R_o^2$), and modified $R^2$ ($R_m^2$).[314,315] An $R^2$ value closer to 1 indicates that the model explains a large percentage of the variability in the data, suggesting a good fit. It is well established that regression through the origin ($R_o^2$) is of less significance to predictive ML models, where we wish to compare observed versus predicted values. Global metrics like $R^2$ are often less informative because they are not measures of the decisional impact; often, errors near the decisional boundary are most important instead of equally weighting everywhere. Alexander et al. demonstrate that $R^2$ (as defined in equation below), defined as the proportion of total variability explained by the model (the ratio of explained variability to total variability, subtracted from unity), is practically useful, particularly when the goal is to minimize residuals between model predictions and actual values.[316]

$$R^2 = 1 - \frac{\Sigma(y - \hat{y})^2}{\Sigma(y - \bar{y})^2}$$

where $y$ is the observed response variable, $\bar{y}$ is its mean, and $\hat{y}$ is the corresponding predicted value.

While $r^2$ and $R^2$ are sometimes equivalent, they are not always the same. The key difference arises when the mean of the model's predictions does not match the mean of the real data. In such cases, $R^2$ will decrease because the model's predictions are biased, either systematically overestimating or underestimating the targets. In contrast, $r^2$ remains unaffected by this bias as it only captures the strength of the linear relationship between the predictions and targets. Therefore, $r^2$ may still be high even if the model's predictions are consistently offset. Whether to prioritize $R^2$ or $r^2$ depends on the goal of the model. If the primary interest is in capturing the underlying trends without regard to systematic bias, $r^2$ may be more informative. However, if the exact accuracy of predictions is critical, $R^2$ is more appropriate as it penalizes bias.


https://doi.org/10.1021/acs.chemrestox.5c00033
Chem. Res. Toxicol. XXXX, XXX, XXX−XXX

While metrics like the coefficient of determination ($R^2$) are standard for evaluating how well a regression model captures the overall trend and variance of the data distribution (measure of distribution), Alexander et al. suggest that it is also important to assess the accuracy of predictions at individual data points (measure of dispersion).[316] $R^2$ measures how well the predicted values replicate the variability of the true data but does not detect point-wise deviations where predictions are systematically higher or lower than actual values.

To capture these point-wise deviations, point-based error metrics are used that are, when it comes to anticipating the error of a prediction in future use cases, at least as important as the above distribution-based metrics. Studies often use Mean Absolute Error (MAE), which measures the average magnitude of the errors without considering their direction, offering a straightforward measure of how much, on average, the predictions deviate from the true values at each point.

$$MAE = = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)$$

where for $n$ predictions from a sample of $n$ data points, $Y_i$ is the observed value of the variable being predicted, and $\hat{Y}_i$ is the predicted value.

The most common metrics used is the Mean Squared Error (MSE), which calculates the average of the squared differences between predicted and actual values at each data point. A lower MSE indicates that the model's predictions are generally close to the actual values on a point-by-point basis.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

where for $n$ predictions from a sample of $n$ data points, $Y_i$ is the observed value of the variable being predicted, and $\hat{Y}_i$ is the predicted value.

The Root-Mean-Square Error (RMSE) is the square root of the MSE and provides an error measure in the same units as the target variable, enhancing interpretability.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}$$

where for $n$ predictions from a sample of $n$ data points, $Y_i$ is the observed value of the variable being predicted, and $\hat{Y}_i$ is the predicted value.

In some regression tasks, it can be useful to evaluate model performance using classification metrics by converting continuous targets into binary labels based on a threshold. For instance, if the target values are bimodal or contain a small fraction of extreme values, we can convert them into binary or multiclass labels and compare the actual data to the predicted labels. Also in case of data with large experimental error, such as microsomal stability assays, categorical predictions based on a regression model can be useful.[317] Metrics such as accuracy, precision, recall, and F1 score can then be used to assess the model's ability to correctly classify target values. This approach is particularly helpful when the continuous target values are noisy within each mode, and the primary focus is on accurately making a binary decision: identifying high-risk or high-value cases.

From the perspective of ML models in toxicity prediction, a model with a low $R^2$ may still be helpful in practical drug discovery if the measure of dispersion, such as the Root-Mean-Square Error (RMSE), is low and acceptable for the intended use. For example, *in vivo* pharmacokinetic parameters are organism-level data that are considered difficult to model using chemical descriptors: models usually have a low $R^2$.[124,125] Yet, such models where the RMSE is deemed within acceptable thresholds are used internally in pharmaceutical companies. In predictive *in vivo* PK property models, the number of compounds predicted within 2- or 3-fold error (prediction fold error) is a common evaluation metric.[318] This level of 'error' is acceptable due to the significant inherent variability in biological systems, as well as differences in the output variable that span multiple orders of magnitude, meaning that even the experimental data can exhibit substantial variability between individuals, and predictions falling within a 2- to 3-fold range of observed values are considered reasonable in PK prediction, as the model's performance aligns with the actual variability in the data.[319,320]

Overall, we recommend measures of dispersion like RMSE and Mean/Median Absolute Error (MAE) along with fit-to-data-distribution metrics like ($R^2$) as metrics in evaluating toxicity models for real-world applications. (For a more detailed of those aspects, see ref 33.) In addition, it is important to keep in mind that the impact of experimental assay error should not be overlooked in the pursuit of achieving the 'best' performance and that model error needs to be seen in context of experimental error. Once a model reaches the known theoretical accuracy (and error) of the experimental method (the data from which the model was built on), the model could be considered as good as the underlying experiment.[321] Striving for performance beyond that point, on the same test set, may only reflect the error and noise inherent in the data and hence lead to model overfitting (which indeed seems to be the case for several published models[322]).

## RETROSPECTIVE MODEL VALIDATION

In order to optimize hyperparameters of models, as well as to obtain an estimate of future performance, models must be validated on compounds not seen during model training before real-world use (see Box 2 for details on validation strategies). Models are typically validated by splitting the data to create train, validation and test sets (Figure 10), where the primary purpose of the training set is to fit model parameters; the validation set is used to optmise hyperparameters (however, preoptimized hyperparameters can also suffer from overfitting[323]) and the test set is used to estimate future model performance.

More specifically, cross-validation is often used to provide a better performance estimate than that given by a single train-test split and to estimate the uncertainty in this performance estimate. Cross-validation (Figure 10a) involves dividing the data set into multiple subsets and systematically training and validating the model across different combinations of these subsets.[234] In each iteration, one subset of the data is used for training, and the remaining subset of the data is used for testing. Performance metrics are each averaged across splits to provide an estimate of the model's performance, and the variance of each metric across splits is used to provide error bars for this performance estimate. If there is no need to select and optimize hyperparameters for an algorithm, then a cross-validation is sufficient. However, when optimizing parameters, a simple cross-validation often leads to an overestimation of model performance.[324] In any case, the caveat remains that performance of the model is still evaluated only on the chemistry


https://doi.org/10.1021/acs.chemrestox.5c00033
Chem. Res. Toxicol. XXXX, XXX, XXX−XXX

**Chart Box 2**

**Box 2.** Types of model validation

*Validation Strategies*

**Cross-validation** divides data into subsets, trains the model on some compounds, and test on others, repeating the process to assess the model.

**Nested cross-validation** combines hyperparameter tuning within an inner loop (using the 'validation set') and model performance assessment in an outer loop (using the 'test set'), with the aim to prevent overfitting.

**Held-out (external) test set** uses an independent dataset held-out from the original dataset, and not involved in training, to evaluate the model's generalization to unseen data. Whether performance measures obtained are predictive for future use cases depends on the similarity of the test set to future project compounds.

*Splitting Strategies*

**Random split** divides the data into training and test sets randomly; this method does not account for chemical diversity or data distribution shifts and can allow significant information leakage if similar compounds (e.g. from chemical series) are present n both sets.

**Scaffold split** divides the data based on chemical scaffolds where compounds with similar core structures are kept in the same sets, ensuring that the model is tested on scaffolds unseen during training. This split assesses the model's ability to generalize to new chemical scaffolds (which, however, may still be closer or further away from future project compounds) .

**Time-split** divides data based on a temporal order, often used in longitudinal or historical datasets. Compounds discovered or synthesized earlier are used for training, while newer compounds are used for testing. This method assesses the model's real-world applicability in predicting properties of future project compounds.

**Leave-one-cluster-out** excludes all samples in a cluster from training, reserving them for testing to evaluate the model's ability to generalize across different groups of compounds. The clustering can be user-defined such as via chemical scaffolds or physicochemical properties such as logD.

**Out-of-distribution sets** compare the model against compounds from different distributions to check generalization beyond the training scope, e.g., small molecules vs. natural compounds, etc.

**Post-hoc test set** refers to data from experiments that are conducted after the model was finalized. These test sets are an unbiased assessment of model performance on completely new data that did not exist until the model was finalized. On the other hand, future project compounds may still differ significantly also from the post-hoc test set employed.

*Baseline Models*

**Baseline models** serve as a benchmark to compare more complex models, ensuring that the new model significantly improves over trivial or basic predictions. Examples include predicting the majority class or an average outcome as a baseline.

**Y-scrambling**, also known as **randomization testing**, validates if a model learned beyond random predictions. This is done by randomly shuffling the target variable while keeping the feature variables intact (before feature selection etc.) . Y-scrambling is repeated a large number of times, and it can determine whether the model captures the relationship between the features and the target variable or fits random noise with the same general properties as the data: If models obtained with a 'scrambled' output variable are as good as the 'real' model then this real model does not detect any more meaningful input-output relationships that those which can be obtained in a randomized (scrambled) dataset (which should logically not give rise to a meaningful model). However, outperforming Y-scrambling does not necessarily indicate a good model. Y-scrambling serves only as a baseline to assess the lower-bound performance, not as a criterion for model evaluation.

contained in the test set, which may not represent future use cases. Hence, nested cross-validation (Figure 10d) extends this concept by incorporating an inner loop for hyperparameter tuning ('validation set') and an outer loop for model performance assessment ('test set'). This approach aims to prevent overfitting by ensuring that hyperparameter tuning does not bias the model's performance evaluation. Once a model has been optimized with respect to hyperparameters, it is typically retrained on all training data and validated on held-out data retrospectively (Figure 10c). A held-out ('external') test set is where a part of the original data set is reserved only for testing a single time, ensuring that the model was never exposed to it during prior steps (such as feature selection, model selection/training, nested-cross validation, etc.). Holding out test compounds from the same source data, however, means the test data is being drawn from the same underlying distribution as the training data, which can represent 'data leakage' (given that similar compounds are present in training and test set) and lead to an overestimation of model performance. If a model performs similar across different folds of the (nested) cross-validation it can be seen as 'stable', and if it performs well on a held-out test set, the model can be considered reliable *in the respective chemical space and conditions covered by this test set* (which may or may not be predictive for future use cases, particularly if the training set and test set contained highly similar compounds, allowing information leakage, but future use cases do not). Given future use cases are generally unknown, it is difficult to obtain a reliable estimate of model performance for this situation.

There are different approaches for data splitting, which impact how the training and test sets are formed.[325] The simplest is randomized data splitting, which assigns each data point randomly to training or test sets. Although this often still being used, it poses dangers: training data sets may differ in distribution from the data seen during model deployment—for example, the training set may contain clusters of compounds from the same chemical series, whereas the data the model is deployed on may include novel series—and thus other strategies would more realistically reflect expected model performance in real-world use.[326−328] For instance, "scaffold splitting" groups compounds based on their molecular scaffolds—the core structures of molecules without side chains.[326] By ensuring that compounds with the same scaffold are all in either the training or test set, metrics evaluated with this train-test split measure the model's ability to generalize to new chemical scaffolds not seen during training. While this is conceptually valid, 'new chemical scaffolds' may still be more similar, or less similar, in the test set compared to the training set relative to future use cases. Related approaches split data based on other methods of clustering molecules based on various measures of similarity to avoid information leakage from training to test. A more stringent variant of data split is leave-one-cluster-out (Figure 10b) validation, where each *cluster* forms a separate test split. In effect, this splitting strategy carves out whole *areas* of chemical space for the test set. A temporal (time) split, on the other hand, involves dividing the data based on the time of data acquisition or compound synthesis. The model is trained on older data and tested on newer data, simulating real-world scenarios where models are applied to future compounds, designed based on measured properties of past compounds.[328] This method evaluates the model's predictive performance over time and can reveal if model performance suffers due to the underlying data distribution changing over time. This approach is meant to measure model performance on 'new project compounds'; however, given that 'new project compounds' may still be located in very different areas of chemical space, this may or may not resemble future applications of the model. Step-forward splits are another approach of model validation, where data are split based on an ordering based on a molecular property, which could be a physicochemical property like logD or molecular weight. Molecules falling within some range of the property (often the highest or lowest values) are held out from training and are used for testing.[327,329] This strategy mimics the application of a model to data that may have different properties to the training set (for example, molecules in different ranges of physicochemical space) and can thus help in understanding how the model will perform in practical, forward-looking applications. It can be summarized that various data splitting strategies for model training exist; but due to the size of chemical space and unknown future use cases, obtaining a reliable model performance estimate is difficult in practice.

## PROSPECTIVE MODEL VALIDATION ON RELEVANT CHEMICAL SPACE

While the above retrospective data splitting strategies can give some estimate of performance in future use cases, the real proof of applicability of a model rests in *prospective* applications; after all, it is always easier to predict the past, than to predict the future. Prospective validation involves testing the model's



https://doi.org/10.1021/acs.chemrestox.5c00033
Chem. Res. Toxicol. XXXX, XXX, XXX−XXX

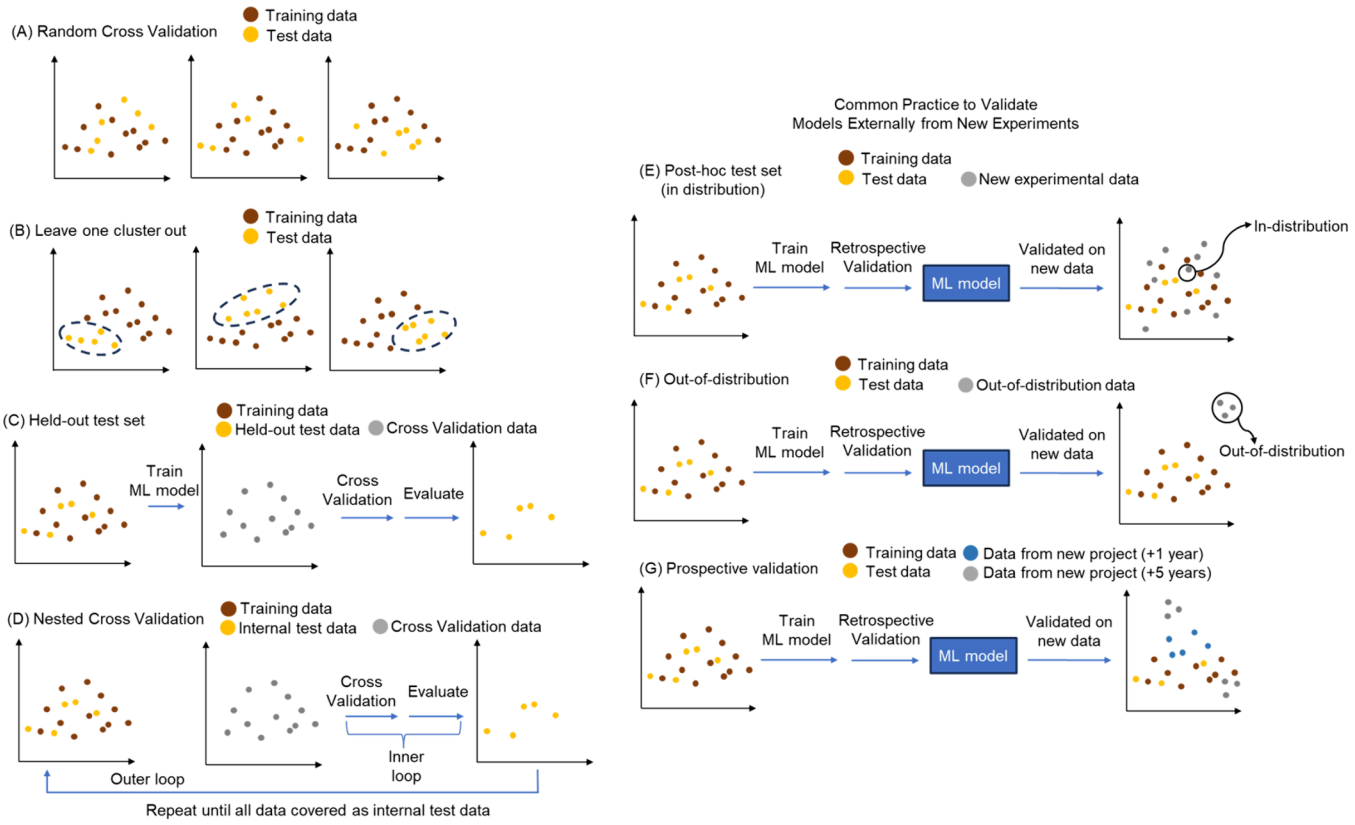



**Figure 10.** Comparative overview of model validation techniques for ML models in chemical space for toxicity prediction: Internal validation methods include (a) cross-validation, ensuring model reliability through repetitive training and testing on different data segments, (b) leave-one-cluster-out, assessing generalization across similar compound groups, (c) held-out test set, evaluating unseen data held-out from the data set; and (d) nested cross-validation, optimizing model parameters while preventing overfitting. External validation techniques comprise (e) a posthoc test set, providing an unbiased assessment of completely new experimental data; (f) an out-of-distribution set, testing model robustness against novel compound distributions such as new class compounds; and (g) prospective validation, confirming real-world applicability on future projects. Note: This figure offers an oversimplified representation of chemical space. While the 2D space can be visualized on two axes, chemical space is multidimensional, underpopulated, and behaves differently in every area. Thus, validation in chemical space is more complex than visualized here, and this figure is intended for illustrative purposes only.

predictions against new, unseen data on future projects or compounds, to confirm their reliability in the chemical space required for the project. While some of such cases are represented to model applications in true future projects, also some 'mixed' approaches are possible. A posthoc test set (Figure 10e) consists of validation of predictions for molecules (say, newly synthesized compounds) that the model has never encountered during any stage of its development. These molecules should be kept hidden from the model developers until the final evaluation phase. An out-of-distribution validation (Figure 10f) includes compounds that are significantly different from the training data distribution—such as new classes of compounds, novel scaffolds, or data obtained under different experimental conditions. The difference to retrospective validation is here fluent in a way, and artificially 'difficult' retrospective cases can be more difficult for the model to handle than, e.g., close analogues in a true prospective set. Finally, prospective validation on future projects (Figure 10g) would involve applying the model to predict outcomes in future projects or compounds synthesized after the model's development, to test the model's predictive power in live scenarios, reflecting its utility in actual practice. And this, in combination with what is described in the next pillar, is actually what matters in reality: Not whether a validation is 'retrospective' or 'prospective', but rather whether we can improve decision making (which of course will then usually be in the prospective domain).

The overall process of model training and validation is explained by using a hypothetical case here: Take for example, a model aiming to predict mitochondrial membrane depolarization, which will be applied in a discovery project for a novel class of chemical compounds, synthetic cannabinoids. An ML model could have been trained and internally validated using nested cross-validation on a data set for small molecules and associated mitochondrial membrane depolarization readouts (a toxicity measure); this set of compounds could include some known cannabinoids. The nesting can leave one-cluster-out to ensure that the optimization results in a model that can generalize across different subclasses of compounds, including cannabinoids. Once optimized, the fitted model should be validated retrospectively using a held-out test set, including some novel synthetic cannabinoids that were not part of the training data (out-of-distribution) to assess its predictive accuracy in the particular chemical space. If the performance is acceptable for practical application, the model could be applied to a drug discovery project targeting the development of cannabinoid-based therapeutics. A prospective validation would be attempted against experimental data from the new cannabinoid compounds developed during the project. Such prospective


https://doi.org/10.1021/acs.chemrestox.5c00033
Chem. Res. Toxicol. XXXX, XXX, XXX−XXX

validation assesses whether the models perform as expected in practical project applications.

Once validated, a model can be retrained on all available data, including the test data, before it is deployed again. This should expand the model's coverage of the chemical space since more data are included for model training. Given that the test set typically represents a small addition to a much larger training data set and the optimized model parameters remain the same as the before, the inclusion of the test data in the retraining now is unlikely to significantly alter predictions in other areas of the chemical space. Instead, it primarily enhances the model's accuracy within the newly added chemical space. On the other hand, using the test data in training the final deployed model means that further performance evaluation of the model is difficult unless new test data becomes available. Thus, as new data become available from experiments, the model is retrained, re-evaluated, updated if suitable, and deployed for use. This process is particularly important in the case of project work, where information about the chemical space of current interest can greatly increase model performance. For an example of implementation of suitable validation protocols, see OCHEM (https://ochem.eu), a platform developed more than 15 years ago.

## DEFINING THE APPLICABILITY DOMAIN

Optimizing a model for a given evaluation metric needs to be complemented with a measure how well the model is able to predict for future uses cases - similar to experiments, which result, e.g., in a mean and an associated standard deviation, predictions need to come with a prediction value, and an associated confidence measure into this prediction.[330] Models must, therefore, balance optimizing the performance metric with the so-called 'Applicability Domain' (AD). The AD defines the range of input/output values or conditions under which the model's predictions are considered reliable and valid. It essentially sets the boundaries within which the model can be expected to perform accurately (to various quantitative degrees), based on the data it was trained on and the model training and validation process used. Models trained on a smaller data set of more project-relevant compounds may outperform those trained on larger data sets; this was observed in the context of virtual screening, where removing half of the molecules with the lowest applicability scores indeed *improved* performance.[331] One approach to evaluating performance with respect to the applicability domain involves using test sets split into distance bins (where distance refers to structural similarity to the training compounds) to monitor performance. Furthermore, bins representing chemical space with poor performance suggest new experiments, whose data might improve the model performance in that space. Various alternatives have been explored to define AD in chemical space (Table 8), each catering to particular scenarios.[40] It should be noted that while a large number of such approaches exist, given the size of chemical space and that it behaves very differently locally, it is very difficult to come with reliable error (and applicability domain) estimates in practice.

## PILLAR 5: TRANSLATING TO DECISION-MAKING

As we have seen, improving ML models for molecular toxicity prediction relies on data, chemical representation, model architectures, and validation. However, improving drug discovery (as a process) involves translating the outcomes of

**Table 8. Methods to Estimate the Applicability Domain of a Predictive Model for Small Molecule Compound Properties[331]**

| Method | Description | When to Use | When Not to Use | Reference |
|---|---|---|---|---|
| Range-based methods | Define boundaries based on the range of descriptor values in the training set | When the descriptor values are well-defined and bounded within a specific range | When the data contains many outliers | 40 |
| Convex hull methods | Establish a geometric boundary around training data points | When the data points form a well-defined area in the descriptor space | When the data is multidimensional, sparse, or irregularly distributed | 332 |
| Distance-based methods | Measure the distance between a query molecule and the closest training data points | When the distances between data points in the descriptor space are meaningful and consistent | When the descriptor space is a binary vector or too high-dimensional | 333 |
| Density-related methods | Evaluate the density of data points around a query molecule | When data consists of a few chemical series, the density of data points around a query molecule indicates prediction reliability | When the data is uniformly distributed or lacks significant density variations | 40 |
| One-class SVM | Train a model to identify the region in the descriptor space occupied by the training set | Identifying whether a query point belongs to the same distribution as the training set | When the training data is highly heterogeneous or contains multiple clusters | 334 |
| Uncertainty quantification and confidence estimation | Assess prediction confidence by analyzing model outputs | Report the confidence of the model's predictions | When the computational cost of uncertainty estimation is prohibitively high | 335,336 |
| Reliability-density neighborhood approaches | Combine density and distance metrics to define AD | Combining density and distance metrics can provide a more nuanced definition of AD | When the data lacks sufficient density or distance information to make reliable assessments | 337 |



https://doi.org/10.1021/acs.chemrestox.5c00033
Chem. Res. Toxicol. XXXX, XXX, XXX−XXX

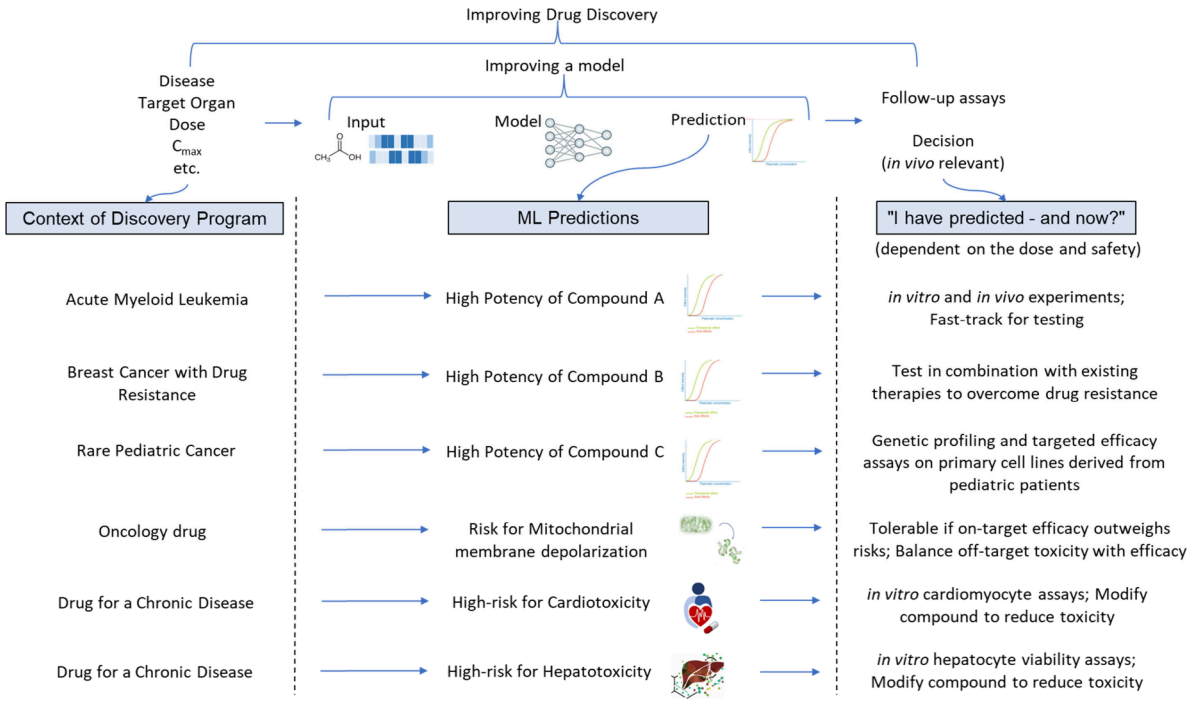



**Figure 11.** Influence on ML predictions in hypothetical project contexts depends on the availability of assays and validation on a subset of compounds. Interpreting and understanding predictions from ML models.

ML models into actionable follow-up decisions, such as prioritization of one compound over another, which assay to run next, etc.[300] (Figure 11; see Box 3 for critical takeaways on moving ML predictions toward actionable drug discovery decisions). It can be seen that a machine learning model (and its application) is only a tool, an ability—when used, it is embedded into a *context*, and that context is frequently insufficiently considered when only focusing on 'model metrics'. However, a machine learning model is not a purpose in itself; the purpose of a model is to be applied in a context.

Many ML models for toxicity in the industry are used as an alert for potential risk or for prioritization and not necessarily as a go/no-go decision. Those decisions depend highly on PK/ADME, dose, exposure, etc., conditions that an ML model usually does not capture. In terms of machine learning, this represents an 'underspecification' of the task the model is trained on[338]—and hence the model output cannot be directly translated to decision making. For example, acetaminophen (paracetamol) is generally safe at therapeutic doses but becomes toxic at higher doses due to its metabolism into a harmful compound called N-acetyl-p-benzoquinone imine (NAPQI).[339] Machine learning models predicting toxicity solely on the basis of molecular structure might not flag this risk because they rarely account for dose-dependent effects, metabolic pathways, or individual variations in drug metabolism (PK/ADME factors). PK parameters are needed for human-relevant decision making in safety.[340] For example, predicting *in vivo* pharmacokinetic parameters (which cover drug absorption, distribution, and clearance in humans) could reveal unsuitable drug characteristics (such as bioavailability), informing the next iteration of drug design as part of a Design−Make−Test−Analyze cycle to optimize compounds in development.[341,342] Overall, ML models for efficacy, toxicity, and PK/ADME can only translate to human-relevant decision-making if all three are considered in parallel.

To a small extent computational methods have already impacted regulatory processes, such as the adoption of the ICH M7 guidelines (International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use).[343] These guidelines provide a framework for using computational methods, including ML, to predict the mutagenic potential of pharmaceutical impurities. The ICH M7 guidelines emphasize the importance of (Q)SAR (Quantitative Structure−Activity Relationship) methodologies, recommending two complementary approaches for this purpose, namely, an expert rule-based system and a statistics-based system. Expert rule-based systems apply established toxicological rules to identify certain molecular fragments, known as structural alerts, associated with mutagenicity, while statistical (ML) models leverage large data sets to detect patterns and predict outcomes. Where both methodologies indicate no structural alerts for

## Chart Box 3

**Box 3.** From ML predictions to actionable drug discovery decisions

**Understanding ML Predictions in Context**: ML model predictions offer valuable insights across various stages of drug discovery. However, their significance varies with the project context, such as disease area; for example, side effects that are not tolerable in lifestyle diseases might well be acceptable in terminal cancers.

**Critical Question: "I have predicted—and now?"** After obtaining ML predictions, moving beyond "predicted numbers" to practical application is necessary–subsequent steps (such as experimental tests) must be decided based on the disease, target area, acceptable thresholds of toxicity outcomes, whether predicted toxicity occurs at human-relevant exposures, and cost and time involved in validation assays (described below).

**Role of Assays in Decision-Making:** Assays are pivotal in validating ML predictions and determining the course of action. Depending on the predictions, assays on a subset of compounds can confirm or refute the model's outcomes, guiding the next steps in the drug design and development cycle.

**Strategies for Effective Translation:**

- **Incorporate Human Expertise:** Engaging domain experts in evaluating and interpreting ML predictions ensures that the insights are grounded in practical knowledge and experience.
- **Real World Validation:** Testing predictions against new, unseen data from projects or through prospective experiments helps confirm their reliability and applicability to real-world scenarios.
- **Iterative Refinement:** Continuous assessment and recalibration of ML models with new data can refine their predictive accuracy and utility in drug discovery.


https://doi.org/10.1021/acs.chemrestox.5c00033
Chem. Res. Toxicol. XXXX, XXX, XXX−XXX

**Table 9. Open-Source or Freely Accessible ML Models for Toxicity Prediction, with Server or Local Implementation**

| Tool Name | Web Site | Description | Reference |
|---|---|---|---|
| ADMETLab 3.0 | admetlab3.scbdd.com | Predicts ADME-Tox properties using advanced machine learning models | 350 |
| QSARdb | qsardb.org | Models for human health effects, toxicokinetics, etc. for evaluating chemical and environmental risk | 351 |
| OCHEM | ochem.eu | Models with environmental and health related end-points, Tox21, etc. | 352 |
| CPSign | github.com/arosbio/cpsign | Conformal prediction for valid prediction intervals on a per-compound basis | 302 |
| Danish QSAR Database | qsar.food.dtu.dk | Database of QSAR models for predicting toxicological end points, especially for regulatory purposes | 353 |
| DILIPredictor | broad.io/DILIPredictor | Predicts human and animal-relevant liver injury and proxy liver injury end points | 354 |
| EPI Suite | epa.gov/tsca-screening-tools/epi-suitetm-estimation-program-interface | A suite of physical/chemical property and environmental fate estimation programs | 355 |
| EMolTox | xundrug.cn/moltox | Web Server for the prediction of toxicity for safety analysis in drug development | 356 |
| OECD QSAR Toolbox | qsartoolbox.org | Provides tools for grouping chemicals into categories and predicting their toxicological properties | 357 |
| OPERA | github.com/NIEHS/OPERA | A suite of QSAR models to predict physicochemical properties, environmental fate, ADME and toxicity end points | 238 |
| pkCSM | biosig.lab.uq.edu.au/pkcsm/prediction | Predicts pharmacokinetic properties of small molecules using graph-based signatures | 358 |
| PKSmart | broad.io/PKSmart | Predicts human and animal pharmacokinetic parameters | 125 |
| ProTox 3.0 | tox.charite.de/protox3/ | Virtual lab for predicting toxicities of small molecules using various predictive models | 359 |
| SwissADME | swissadme.ch | Predicts ADME properties of small molecules using free web tools provided by the Swiss Institute of Bioinformatics | 360 |
| Toxtree | toxtree.sourceforge.net | Open-source application for the estimation of toxic hazards based on decision tree approaches | 361 |
| VEGA-QSAR | vegahub.eu/portfolio-item/vega-qsar | Provides models for predicting toxicological end points using Quantitative Structure−Activity Relationship (QSAR) models | 362 |
| VenomPred 2.0 | mmvsl.it/wp/venompred2 | Evaluate the toxicological profile of small molecules and features that contribute to predictions to derive a structural toxicophore | 363 |

mutagenicity, the guidelines deem the negative *in silico* results as sufficient evidence to conclude that the impurity is not of mutagenic concern, eliminating the need for further experimental testing. This is one concrete example of computational toxicology replacing laboratory experiments, which, however, is limited to a relatively narrow range: to *impurities* (so compounds present at relatively low concentration), in the genotoxicity area, and only when two methods are employed in parallel. In another example, French health authorities' request for toxicity data on cyamemazine was satisfactorily answered with an *in silico* assessment of target organ toxicity.[344] The *in silico* models used for hazard assessment covered toxicity end points such as mutagenicity, hepatotoxicity, nephrotoxicity, cardiotoxicity, etc., including two complementary models, namely expert alert and statistical QSAR models. It can be seen that model predictions can *contribute to* decision making, but having a model does not represent 'having the answer' by itself.

Machine learning models are increasingly being applied in real-world projects in drug discovery to uncover mechanisms of action, predict toxicity, and assess off-target risks early in the development process. Predictive assays for hepatotoxicity and cardiotoxicity have also been an essential part of the early toxicology strategy.[345] Recently, a virtual enhanced cross-screen panel (vEXP) used ML models to predict 67 off-target activities, aiming to provide early warnings about potential adverse drug reactions.[346] The authors reported that the vEXP panel effectively identified potential off-target activities, aiding in the early prioritization and risk assessment of drug candidates—for targets with well-balanced data sets and sufficient data points, the models demonstrated strong predictive performance, many exceeding the thresholds set for acceptable performance (e.g., Kappa $\geq$ 0.4, ROC-AUC $\geq$ 0.7). The authors suggested that the vEXP panel be integrated into various stages of the drug discovery process—to rank hit compounds based on their predicted off-target profiles, identify systematic risks associated with specific chemical series or functional groups, and ensure that late-stage optimization efforts do not inadvertently introduce new off-target activities. This example represents a case where computational models can complement experimental approaches to achieve *higher efficiency* as a result.

Extending beyond animal experiments, Next Generation Risk Assessment (NGRA) combines *in silico* models and *in vitro* assays to evaluate chemical safety without relying on animal data. For instance, an industry study on coumarin in cosmetic products integrated mathematical models, *in vitro* NAMs, an *in vitro* cell stress panel, high-throughput transcriptomics, and *in silico* alerts for genotoxicity.[347] The authors demonstrated that *in silico* methods can complement experimental data to provide sufficient evidence for safety evaluations in a regulatory context. Similarly, Ouedraogo et al. used a 10-step framework of read-across and NAMs with propylparaben, demonstrating the practical application of integrating exposure assessment, *in silico* methods, and bioactivity data to inform reproductive toxicity alerts.[348] Overall, these approaches highlight the value of combining multiple data sources and methodologies to enhance the accuracy, efficiency, and regulatory acceptance of risk assessments. Overall, *in silico* models used in drug (or here compound) discovery, development, and safety assessment can, to some extent, reduce the experimental burden in real-world settings.

A final key to improving the translation of ML models in the real world involves a suitable understanding of the use case and user, hence setting up a system that enables adoption in real-world projects. Aspects such as understanding the need for retraining (on project-specific data), the type of model developed (regression vs two/three/four-class classification), the output of prediction confidence (and potentially a nearest neighbor/concrete experimental data point for the user to gain


https://doi.org/10.1021/acs.chemrestox.5c00033
Chem. Res. Toxicol. XXXX, XXX, XXX−XXX

**Table 10. Commonly Used Feature/Model Importance Methods to Assess Model Performance**

| Method | Description | Advantages | Disadvantages | Source |
|---|---|---|---|---|
| Permutation Importance | Measures the increase in the model's prediction error when a feature's values are randomly shuffled | Easy to understand and implement; can be used in combination with any modeling method in principle | Computationally intensive and can be misleading if features are correlated | 374 |
| Shapley (SHAP) Value | Assigns each feature an important value based on Shapley values from cooperative game theory | Provides consistent and interpretable results | Computationally expensive, especially for complex models; can be misleading if features are correlated | 156 |
| LIME (Local Interpretable Model-agnostic Explanations) | Approximates the model locally with an interpretable model to explain predictions | Model-agnostic and provides local explanations | Can be unstable and sensitive to the choice of local region | 375 |
| Feature Importance from Tree-based Models | Measures the importance of a feature by the average gain of the splits in which it is used. Can be based on permutation feature importance or mean decrease in impurity | Directly available in tree-based models like Random Forest and XGBoost | Biased toward features with many categories | 376 |
| Coefficient Magnitude in Linear Models | Uses the magnitude of the coefficients in linear models to determine feature importance | Straightforward to compute and interpret | Only applicable to linear models and assumes linear relationships | 377 |
| Mutual Information | Measures the reduction in uncertainty about one variable given knowledge of another variable | Nonparametric and captures nonlinear relationships | Requires discretization for continuous variables and can be computationally expensive | 378 |
| Attention Mechanisms | Uses the attention weights in attention-based models to determine feature importance | Naturally integrated into models like transformers. Allows focusing on relevant parts of the input sequence when making predictions (captures dependencies) | Limited to models that use attention mechanisms | 289 |
| Counterfactual Explanations | Determines feature importance by analyzing how changing feature values affects the prediction outcome | Highlights specific changes needed to alter predictions | May generate unrealistic or nonfeasible changes | 379 |
| Partial Dependence Plots (PDP) | Illustrates the impact of a feature on the predicted outcome by averaging the effects of all other features. | Easy to interpret and visualize | Assumes independence of features and can be misleading if features are correlated | 380 |
| Prediction-Driven Molecular Matched Pairs | Examines pairs of similar molecules to identify transformations affecting specific properties. | Overcomes the data limitation problem and allows interpreting the model itself and not the experimental data | Limited number of molecules can be checked in this transformation-driven approach | 381 |



https://doi.org/10.1021/acs.chemrestox.5c00033
Chem. Res. Toxicol. XXXX, XXX, XXX−XXX

trust into the model), suitable interfaces with databases and other software systems, design of the user interface, run time, etc. are important considerations in creating an impactful system in practice.[349] Several publicly available ADMET models are commonly used (Table 9), some with web-based portals and a graphical user interface. Their ability to show feature importance, structural alerts, and other interpretive insights, where present, enhances their transparency and usability. The reader is encouraged to use these (and other available) tools to gain hands-on experience, and to see which approaches are useful in the context of their own work.

Interpreting and hence, to an extent, 'understanding' ML models, in terms of their underlying basis for predictions, can make predictions more trustworthy and help detect biases in the model, as well as ensure the model recovers previous knowledge (e.g., known structural alerts). Suitable feature spaces and modeling architectures need to be used to interpret models, which can lead to novel understanding of the model (and problem) at hand. Feature importance measures (Table 10) can be used to estimate the contribution of individual features, particularly where the features are interpretable, e.g., specific molecular targets, such as proteins, or chemical substructures associated with toxicity, etc. For example, mutagenicity and skin sensitization are often concerns in drug development that can be interpreted from chemical features. Specific structural alerts—such as nitroaromatic groups for mutagenicity,[364] or electrophilic groups like $\alpha,\beta$-unsaturated carbonyls (Michael acceptors) for skin sensitization[365]—are associated with these toxicities; predictive models can identify these fragments and enable medicinal chemists to modify molecules by eliminating or altering these groups to reduce the risk (although modifying a few fragments with the highest contribution in the model does not necessarily resolve toxicity completely[354,366]).

However, interpretations are not always straightforward, especially when it comes to predictions based on the chemical structure. Misinterpretations are common for data sets with high bias in chemical space, which is often the case due to project, synthesis, and analogue bias, etc. For example, a previous study by one of the authors found that *sugar* rings were associated with the *bitterness* of compounds, a result that does not make sense at face value. In this case, the confounding factor was that natural products are frequently glycosylated *and* tend to be bitter. The model then proposed this *confounding factor* (and bias) in the model as a correlation when it was not, in fact, the *causation*.[367] This example shows that while feature importance measures can provide actionable insights, their application in real-world drug discovery must account for potential biases in the input data.

Mechanistic insights can also be gained by interpreting ML models. For example, fuzzy rules are a set of "if−then" statements based on fuzzy logic.[368] Unlike traditional binary logic that deals strictly with true or false values, fuzzy logic allows for reasoning with uncertain or imprecise information by accommodating degrees of truth. For example, a rule might state, "If the daily dose of an oral medication is high and the lipophilicity is high, then the probability of hepatotoxicity is high",[369] where the term "high" is not an exact value but a linguistic variable represented by membership functions in a fuzzy set. After training an ML model, fuzzy set rules can be developed based on how certain input features influence the model's predictions. For example, converting decision tree rules or clustering results into fuzzy rules can make them more interpretable. For example, Friederichs et al. were able to cluster 90 environmental compounds using fuzzy clustering based on physicochemical parameters, reasoning why compounds in specific clusters exhibit aquatic toxicity (volatility, hydrolysis, etc.).[370] Fuzzy rule sets are usually based on well-defined and interpretable variables (such as molecular weight, logD, etc.). They can help elucidate the fundamental properties contributing to toxicity and guide the design of safer compounds, where interpretability from a medicinal chemistry perspective drives decision-making.[371−373]

When interpreting model predictions, however, correlation must not be confused with causation; the most-contributing features used by the model are only meaningful if they represent the underlying cause (which is often unknown), and in practice, feature importance often suffers from a 'long tail' distribution, where a large number of features are associated with a prediction; this complicates interpretation (many features contribute 'somewhat' to the end point the model predicts). Moreover, feature importance is a local property for nonlinear models, while global for linear models. Combining modeling and computational analysis with expert knowledge is, hence, key. For example, pathway enrichment analysis involves mapping important features, such as genes and proteins, to known biological pathways using tools like KEGG, Reactome or Ingenuity Pathways Analysis (IPA); for example, Antazak et al. identified 21 molecular pathways differentially modulated in response to nephrotoxic vs nontoxic compounds.[382] They identified the three main functional categories of pathways for nephrotoxicity—metabolic pathways (e.g., Glycerophospholipid metabolism), signaling pathways (e.g., Parkinson's disease), and cell communication pathways (e.g., cell communication). Network analysis can extend mechanistic analysis by constructing and analyzing biological networks, such as Protein−Protein Interaction (PPI) networks, to pinpoint key nodes and modules. For example, IPA has very specific features to evaluate protein−protein interactions, such as upstream regulator analysis, mechanistic networks, causal network analysis, and downstream effects analysis.[383] A study using IPA revealed a novel antitumor mechanism for MK886, a leukotriene antagonist, involving cytoskeleton-induced alteration of chromatin structure, revealing unknown aspects of the action and safety.[384] Another study used PPI networks to determine targets for cardiac disorders and found protein ERBB4 interacting with known drug targets.[385] Although ERBB4 was not a known drug target for cardiac failure at the time (according to Huang et al.[385]), there is considerable work now on using ERBB4 agonists for treating heart failure.[386] Overall, pathway enrichment analysis and PPI networks can highlight important proteins or groups of proteins that may contribute to specific properties, yielding a deeper understanding of the complex biological interactions involved. In the cases discussed here, prior knowledge (in terms of biological pathways) was combined with data (from the models) to arrive at an interpretable model output.

Another related approach to interpretable predictions is to use causal and mathematical mechanistic models to directly establish cause-and-effect relationships rather than correlations captured by ML models. Causal inference methods have been applied directly to transcriptional data and PPI networks to distinguish correlation from causation and infer potential mechanisms of actions that drive toxicity. For example, these methods were shown to recapitulate specific pathways downstream of the molecular targets, providing a systems-level view of the drug's mechanism.[387,388]


https://doi.org/10.1021/acs.chemrestox.5c00033
Chem. Res. Toxicol. XXXX, XXX, XXX−XXX

**Table 11. Different Priorities and Criteria Are Used in ML Models to Balance Efficacy, Toxicity, and Selection (e.g., on-target activity, efficacy) or Deselection (e.g., toxicity alert models) in Early- and Late-Stage Drug Discovery**

| Focus | Early Stage Exploratory/Discovery (usually qualitative) | Late Stage Refinement/Development (usually quantitative: exposure, PK, etc.) |
|---|---|---|
| ML models for selection (e.g., on-target activity, efficacy) | Broad selection, high recall. Include as many potential compounds fitting the target product profile as possible. | Narrow selection, high specificity. Only compounds that meet strict efficacy and safety criteria are selected for further development. |
| ML models for deselection (e.g., toxicity alert models) | Low impact of compounds incorrectly predicted as toxic (false positives). It is acceptable to have some false positive predictions. | High precision is desired in ML models to minimize compounds incorrectly predicted as nontoxic (false negatives), avoiding resource wastage on failing compounds. |

Overall, understanding the features involved in the predictions of ML models can provide both insight and confidence into the predictions being made.

## PHARMACOKINETICS IN DECISION-MAKING

Toxicity is not a simple 'yes' or 'no' question—even water can be toxic at 'too high a dose'.[389] Concentration and exposure (and PK in general) are absolutely critical for assessing safety and toxicity but are relatively neglected in the ML community.

Pharmacokinetics is a critical aspect of drug development and clinical decision-making because it provides detailed insights into how a drug is absorbed, distributed, metabolized, and eliminated within the body.[390] This information is essential for determining the optimal dosage and administration schedule to achieve therapeutic efficacy without causing harm. By studying pharmacokinetics, researchers can establish the time it takes for a drug to reach therapeutic levels, how long it remains effective, and the dynamics of its elimination. This knowledge ensures that drug regimens are optimized for maximum benefit while minimizing the risk of adverse effects. Analogous considerations apply to compound toxicity.

Pharmacokinetics also plays a central role in balancing efficacy and safety by informing the therapeutic index (TI),[340] a critical parameter derived from pharmacokinetics and pharmacodynamic (PD) studies and defined as the ratio between the toxic dose and the effective dose of a drug:

$$TI = \frac{TD50}{ED50}$$

where *TD50* is the dose that produces toxicity in 50% of the population, and *ED50* is the dose that is effective in 50% of the population.[340] A higher TI indicates a greater margin of safety between therapeutic and toxic doses. Accurate PK modeling helps to determine the TI more precisely, aiding in optimizing dosing regimens to maximize efficacy while minimizing toxicity.

While physiologically based Pharmacokinetics (PBPK) models have been used in drug discovery for a number of years, they require tailoring to the particular compound properties at hand. On the other hand, in recent years approaches have become popular to predict *in vivo* PK directly based on chemical structure,[391] owing to the availability of suitable data sets. Machine Learning models have recently been integrated with mechanistic models to enhance the predictive accuracy. Tools that predict pharmacokinetics parameters using ML algorithms, such as PKSmart,[125] can process large data sets to predict the pharmacokinetics parameters that can then be used as input parameters to a compartment model.[392−395] This hybrid approach combines the strengths of data-driven ML models with the interpretability of the mechanistic models.

## OPTIMIZING ML STRATEGIES IN VARIOUS STAGES OF DRUG DISCOVERY

Using ML for QSAR/QSPR models is common practice in drug discovery, where finding the right balance between cost, speed, and predictivity of a method at a given stage of the drug discovery process is critical (see Table 11). It is important to recognize that different industries may adopt varying approaches to using ML models, both within drug discovery and development and in other industries such as consumer goods or agrochemistry. In these contexts, different properties of data sets, representations, models, validation methods, and evaluation metrics may hold varying degrees of importance. Some projects might favor a balanced approach between recall and precision, while in other cases one or the other might be more important.[396] Others may tailor their model optimization strategies (or loss functions) to particular aspects of model performance, specific therapeutic areas, the types of compounds being studied, or the technologies employed.[32,397] A typical use of models along the early and late drug discovery stages follows.

For early stage virtual screening of compounds across a range of targets, discovery phases are exploratory, often targeting novel or less well-understood pathways. Given that on-target effects are crucial for compound selection (after a target has been selected), ideal ML models would find many hits with a high precision (i.e., with a high confidence that the compounds selected are indeed hitting the target the project is working on).[398] On the other hand, from a toxicity prediction perspective, early discovery phases also need to *minimize compounds incorrectly predicted as nontoxic* ('false negatives', where toxicity is seen as the 'positive' label, from the modeling perspective).[345] The above decisions depend on the project context—for example, when the project aims at conditions with limited patient survival (e.g., advance stage oncology or rare diseases), toxicity concerns may not be a major roadblock.[399] Instead, other factors such as the compound's pharmacokinetic profile (if distribution to target organs such as the brain is required) may take center stage. One possible strategy for using data and machine learning in a drug discovery project is visualized in Box 4. Variations need to be applied according to project needs in a given concrete situation.

Me-too/Me-better drugs share the same mechanism of action or target as the existing drugs. Projects focused on identifying me-too drugs generally have a lower human safety risk profile because their mechanisms of action and on-target toxicity are often already known. Still, off-target toxicity may present challenges, where the adverse reactions are not caused by the therapeutic class.[400] The opportunity also lies in finding incremental improvements in therapeutic action and/or safety. Projects focusing on me-too/follow-on drugs usually demand fewer resources to develop[401] and thus ML models could be optimized on minimizing compounds incorrectly predicted as inactive (false negatives) among the known target (in line with


https://doi.org/10.1021/acs.chemrestox.5c00033
Chem. Res. Toxicol. XXXX, XXX, XXX−XXX

**Chart Box 4**

**Box 4. *Example* strategy for progression from experimental screen hits to lead optimization, with a lead and backup chemical series**

1) **Early target safety review**

*Action*: Ensure that potential safety liabilities associated with the (on-)target, as well as related targets, are identified upfront to the extent possible.

2) **Select representative hit compounds from each cluster**

*Action*: Choose representative hit compounds from experimental screens by clustering based on chemical structure (and efficacy), then selecting diverse hits to form a few chemical series, each exhibiting a suitable and sufficient range of efficacy.

3) ***In silico* toxicity prediction for each compound**

*Action*: Perform *in silico* toxicity assessments on each compound using computational models.

4) **Analyze compound-related on-target and off-target toxicity within each chemical hit series**

*Action*: For each chemical series:

i) Identify common toxicity alerts.
ii) Determine if alerts are associated with:
  (1) **Core structures**: Are central scaffolds contributing to toxicity?
  (2) **Side chains/substituents**: Do specific side groups trigger alerts?
  (3) **Metabolic activation**: Is toxicity due to potential metabolites?
iii) Assess whether toxicity is consistent across compounds in each chemical series.

5) **Investigate mechanism-related on-target toxicity**

*Action*: Examine if toxicity is related to the mechanism of action (MoA) of the compound, if known.

i) Compare with other clusters sharing similar MoAs, if known.
ii) Is toxicity linked to the therapeutic target?
  (1) *Action*: Update target safety review if mechanism-related toxicity is confirmed.
  (2) Consider alternative targets or therapeutic strategies.

6) **Rank chemical series based on efficacy and on-target and off-target toxicity profiles**

*Action*: Compare on-target and off-target liabilities and efficacy across all chemical series.

i) Identify chemical series with higher or lower toxicity liabilities, aiming to select compounds with good experimental efficacy but without predicted toxicity.
ii) Consider both the frequency and severity of toxicity alerts.
iii) Factor in the structural tractability for optimization.

7) **Decide cluster prioritization**

*Decision*: Are the off-target toxicity liabilities acceptable for progression?

i) **Yes**: Prioritize hits from chemical series with acceptable on-target and off-target toxicity profiles.
ii) **No**: Consider deprioritizing chemical series with significant toxicity risks.

8) **Confirm hits through re-synthesis and testing for toxicity**

*Action*: For virtual hits from prioritized chemical series:

i) Synthesize compounds.
ii) Test in the same assays used for initial screening.

*Action*: For wet hits from prioritized chemical series:

iii) Re-synthesize compounds to ensure purity and correct structure.
iv) Re-test to confirm activity and rule out false positives due to impurities.

9) **Confirm potency and ADME properties**

*Action*: Evaluate hit compounds from prioritized chemical series in secondary assays to confirm potency.

*Action*: Assess basic ADME properties *in vitro* (e.g., solubility, metabolic stability).

*Action*: Within each prioritized chemical series:

i) Synthesize analogs to explore chemical space.
ii) Aim to improve potency and ADME while mitigating toxicity.

10) **Off-target toxicity assessment**

*Action*: Re-assess compounds, confirm with *in vitro* toxicity assays.

11) **Finalize lead and backup compound series for lead optimization**

*Decision*: Do representative compounds from the series meet criteria for likely *in vivo* efficacy, ADME, and toxicity?

i) **Yes**: Advance compounds from one or two series to lead optimization stage.
ii) **No**: Continue iterative optimization or consider alternative clusters.

the practice of revisiting prior screening data and 'rescuing' false negatives).[400] To ensure competitiveness in the market, these projects would likely want to find all possible molecules in the early stage (high recall), advancing the most promising ones to development and keeping the others as backup in case the molecule fails in the preclinical or clinical stage.[402,403]

By contrast, first-in-class drugs have a new and unique mechanism of action that could yield greater efficacy than existing treatments, but carry a higher risk due to unproven mechanisms that may not work as expected, or may have unforeseen side effects or consequences.[404] That said, a study by Health Canada considering drugs that were approved and later withdrawn found that first-in-class drugs do not exhibit a higher concern for safety compared to nonfirst-in-class drugs.[405] Projects aiming for a first-in-class drug also offer higher rewards, with market exclusivity initially, and in later years increased market share when other pharmaceuticals enter the space (even when new molecules may have benefits).[401] First-in-class drugs have limited or no data in the chemical space and therefore rely more on chemical space exploration. ML models employed in early discovery would ideally need to rank compounds with the desired target profile (efficacy) and low toxicity (at human exposure levels) for experimental validation.[306]

In the early screening paradigm, *in vitro* end points are relatively inexpensive to test; lead compounds that are predicted positive could be screened in those assays to detect false positives. False negatives may not get tested as they were deemed "safe" and carried forward an unidentified liability. ML models thus need to have a lower tolerance for false negatives; we would like to flag compounds if they are toxic. As a project progresses to late-stage discovery, the focus shifts to precision to ensure that only the most promising, efficacious, and safe compounds are pursued from the pool of candidates.

In summary, to translate the benefits of toxicity models to decision-making, they must consider a number of aspects in the context of the project-relevant data, suitable representations, proper validation, reliable interpretation.[127] Given the complexity of the task, the lack of data in many domains, and the variability of the biological systems many end points are measured in, it is not trivial to get all of those aspects right.

## CONCLUSION

This review summarizes key challenges and considerations in translating machine learning models into decision-making tools for real-world drug discovery projects, in particular, related to compound toxicity and safety. This includes making choices about data, modeling, validation, model metrics, and applying the *model* thus obtained to the *process* of drug discovery. We hope that the reader finds this review useful and that it helps this translation from concept to practice.

## AUTHOR INFORMATION

### Corresponding Authors

**Srijit Seal** − *Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, United States; Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, U.K.;* orcid.org/0000-0003-2790-8679; Email: seal@broadinstitute.org

**Ola Spjuth** − *Department of Pharmaceutical Biosciences and Science for Life Laboratory, Uppsala University, Uppsala 751 24, Sweden; Phenaros Pharmaceuticals AB, Uppsala 75239, Sweden*; Email: ola.spjuth@uu.se



https://doi.org/10.1021/acs.chemrestox.5c00033
*Chem. Res. Toxicol.* XXXX, XXX, XXX−XXX

**Anne E. Carpenter** − *Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, United States;* orcid.org/0000-0003-1555-8261; Email: anne@broadinstitute.org

**Andreas Bender** − *College of Medicine and Health Sciences, Khalifa University of Science and Technology, Abu Dhabi 127788, United Arab Emirates; Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, U.K.;* orcid.org/0000-0002-6683-7546; Email: andreas.bender@ku.ac.ae

## Authors

**Manas Mahale** − *Department of Pharmaceutical Chemistry, Bombay College of Pharmacy, Mumbai 400098, India;* orcid.org/0009-0007-3867-996X

**Miguel García-Ortegón** − *Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, U.K.;* orcid.org/0000-0003-4372-4706

**Chaitanya K. Joshi** − *Department of Computer Science and Technology, University of Cambridge, Cambridge CB3 0FD, U.K.*

**Layla Hosseini-Gerami** − *IgnotaLabs, Cambridge CB4 0GA, U.K.*

**Alex Beatson** − *Axiom Bio, San Francisco, California 94107, United States*

**Matthew Greenig** − *Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, U.K.*

**Mrinal Shekhar** − *Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, United States;* orcid.org/0000-0001-8089-8858

**Arijit Patra** − *UCB Pharma U.K., Slough SL1 3WE, U.K.*

**Caroline Weis** − *GSK, London WC1A 1DG, U.K.*

**Arash Mehrjou** − *GSK, London WC1A 1DG, U.K.*

**Adrien Badré** − *Novartis Biomedical Research, Cambridge, Massachusetts 02139, United States*

**Brianna Paisley** − *Eli Lilly & Company, Indianapolis, Indiana 46285, United States*

**Rhiannon Lowe** − *Relation Therapeutics, London NW1 3BG, U.K.*

**Shantanu Singh** − *Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, United States;* orcid.org/0000-0003-3150-3025

**Falgun Shah** − *Non Clinical Drug Safety, Merck Inc., West Point, Pennsylvania 19486, United States*

**Bjarki Johannesson** − *AstraZeneca, 43183 Molndal, Sweden*

**Dominic Williams** − *AstraZeneca, 43183 Molndal, Sweden*

**David Rouquie** − *Toxicology Data Science, Bayer SAS Crop Science Division, Valbonne Sophia-Antipolis 06560, France;* orcid.org/0000-0002-7796-7418

**Djork-Arné Clevert** − *Pfizer, Worldwide Research, Development and Medical, Machine Learning & Computational Sciences, Berlin 10922, Germany*

**Patrick Schwab** − *GSK, London WC1A 1DG, U.K.*

**Nicola Richmond** − *Recursion, London N1C 4AG, U.K.*

**Christos A. Nicolaou** − *Computational Drug Design, Digital Science & Innovation, Novo Nordisk US R&D, Lexington, Massachusetts 02421, United States*

**Raymond J. Gonzalez** − *Non Clinical Drug Safety, Merck Inc., West Point, Pennsylvania 19486, United States*

**Russell Naven** − *Novartis Biomedical Research, Cambridge, Massachusetts 02139, United States*

**Carolin Schramm** − *Sanofi, Babraham Research Campus, Cambridge CB22 3AT, U.K.*

**Lewis R Vidler** − *Eli Lilly and Company, Bracknell RG12 1PU, U.K.*

**Kamel Mansouri** − *NIH/NIEHS/DTT/NICEATM, Research Triangle Park, North Carolina 27709, United States*

**W. Patrick Walters** − *Relay Therapeutics, Cambridge, Massachusetts 02141, United States;* orcid.org/0000-0003-2860-7958

**Deidre Dalmas Wilk** − *Nonclinical Safety, Collegeville, Pennsylvania 19426, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.chemrestox.5c00033

## Author Contributions

All authors have approved the final version of the manuscript. The views expressed in this work are from a collaborative effort, with conceptualization by S. Seal and A.B. Throughout the process, all authors engaged in critical discussions, review, and editing, ensuring the work was comprehensive in articulating the current state and future directions of understanding ML models for molecular toxicity prediction. CRediT: **Srijit Seal** conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, software, supervision, validation, visualization, writing - original draft, writing - review & editing; **Layla Hosseini-Gerami** writing - review & editing; **Alex Beatson** writing - review & editing; **Caroline Weis** writing - review & editing; **Arash Mehrjou** writing - review & editing; **Rhiannon Lowe** writing - review & editing; **Falgun Shah** writing - original draft; **Bjarki Johannesson** writing - original draft, writing - review & editing; **Patrick Schwab** writing - review & editing; **Nicola Richmond** writing - review & editing; **Christos A Nicolaou** writing - review & editing; **Carolin Schramm** writing - review & editing; **Deidre Dalmas Wilk** writing - review & editing; **Ola Spjuth** conceptualization, funding acquisition, supervision, writing - original draft, writing - review & editing; **Anne E. Carpenter** conceptualization, funding acquisition, supervision, writing - original draft, writing - review & editing; **Andreas Bender** conceptualization, supervision, writing - original draft, writing - review & editing; **Manas Mahale** writing - review & editing; **Miguel García-Ortegón** writing - review & editing; **Chaitanya K. Joshi** writing - review & editing; **Matthew Greenig** writing - review & editing; **Mrinal Shekhar** writing - review & editing; **Arijit Patra** writing - review & editing; **Adrien Badré** writing - review & editing; **Brianna Paisley** writing - review & editing; **Shantanu Singh** writing - review & editing; **Dominic Williams** writing - review & editing; **David Rouquié** writing - review & editing; **Djork-Arné Clevert** writing - review & editing; **Raymond J. Gonzalez** writing - review & editing; **Russell Naven** writing - review & editing; **Lewis R. Vidler** writing - review & editing; **Kamel Mansouri** writing - review & editing; **W. Patrick Walters** writing - review & editing.

## Funding

S. Seal acknowledges funding from the Cambridge Centre for Data-Driven Discovery (C2D3) and Accelerate Programme for Scientific Discovery. The Broad Institute authors acknowledge funding from the National Institutes of Health (R35 GM122547 to A.E.C.) and the OASIS Consortium members, which is partially supported by a grant from the Massachusetts Life Sciences Center Bits to Bytes Capital Call program (to S. Singh). O.S. acknowledges funding from the Swedish Research Council (Grants 2020-03731 and 2020-01865), FORMAS (Grant 2022-00940), and Swedish Cancer Foundation (22 2412 Pj 03 H).

## Notes

The authors declare the following competing financial interest(s): S. Singh and A.E.C. serve as scientific advisors for companies that use image-based profiling and Cell Painting (A.E.C.: Recursion, SyzOnc, Quiver Bioscience, S. Singh: Waypoint Bio, Dewpoint Therapeutics, Deepcell) and receive honoraria for occasional scientific visits to pharmaceutical and biotechnology companies. O.S. declares ownership in Phenaros Pharmaceuticals and Aros Bio. A.B. is an advisor to Ignota Labs and a consultant or advisor to various biotech companies. L.H.G. is an employee and shareholder of Ignota Labs. All other authors declare no competing interests.

## Biographies

Srijit Seal specializes in machine learning and cheminformatics. His research focuses on developing machine learning algorithms for drug discovery, particularly in toxicity prediction. He is also a Fellow of the Cambridge Philosophical Society and serves on the Board of Directors of the American Society for Cellular and Computational Toxicology (ASCCT). Seal received his PhD at the University of Cambridge and previously was a postdoctoral associate at the Broad Institute of MIT and Harvard.

Manas Mahale is an undergraduate at Bombay College of Pharmacy, University of Mumbai. His research focuses on using cheminformatics and data science to analyze chemical data.

Miguel García Ortegón is a researcher specializing in probabilistic machine learning for drug discovery. With a PhD from the University of Cambridge, his expertise includes Gaussian processes, Bayesian optimization, and graph neural networks. Miguel has interned at Microsoft and CRG, contributing to generative modeling and machine learning projects.

Chaitanya K. Joshi is a PhD student at the Department of Computer Science and Technology, University of Cambridge. His research explores the intersection of Geometric Deep Learning and Graph Neural Networks for applications in biomolecule modeling and design.

Layla Hosseini-Gerami is the Chief Data Science Officer and cofounder of Ignota Laboratories, a TechBio startup specializing in *in silico* toxicity prediction. With a PhD from the University of Cambridge in Bioinformatics and Machine Learning, she has extensive experience in developing AI-driven solutions for drug discovery and toxicity mitigation.

Alex Beatson is a cofounder of Axiom Bio, which builds AI for drug toxicity risk assessment. Axiom generates high throughput, high content vitro data and trains AI models across in vitro and clinical data to predict toxicity risk from molecular structure. Alex received a PhD working on deep learning at Princeton University and built ML platforms for virtual screening and generative molecular design at Genesis Therapeutics and Redesign Science.

Matthew Greenig is a PhD student at the University of Cambridge, where he applies deep learning techniques to antibody discovery and engineering. Matt's work combines ideas from fragment-based protein design with modern geometric deep learning approaches to engineer binding affinity in antibodies.

Mrinal Shekhar is a Group Leader in the Data Sciences and Computational Chemistry group at the Center for the Development of Therapeutics (CDoT), Broad Institute. His group employs computational chemistry, biophysics, and computational structural biology methodologies for drug discovery projects at various stages in the drug discovery pipeline, including hit identification and validation.

Arijit Patra is a Senior Principal Scientist at UCB Biopharma UK, where he focuses on building a new generation of AI systems for drug safety workflows for rare diseases. Previously, he was also associated with AstraZeneca, where he contributed to several initiatives pertinent to the development of novel therapeutic candidates. He holds a PhD in machine learning for obstetric ultrasound from the University of Oxford, where he was a Rhodes Scholar. Arijit has authored several publications around machine learning and medical imaging and has delivered multiple invited talks around the theme of Artificial Intelligence for Healthcare around the globe.

Caroline Weis is a Senior AI/ML Engineer and Team Lead for Clinical Biomedical AI at GSK.ai. With a PhD in Machine Learning for Healthcare from ETH Zurich, her contributions focusedshe focuses on antimicrobial resistance prediction, personalized medicine, and topological data analysis. Caroline's expertise lies in developing advanced machine learning models to improve healthcare outcomes.

Arash Mehrjou is a researcher at ETH Zürich, specializing in making sense of complexities using mathematics, machine learning, and artificial intelligence. His work spans nature and abstract concepts, including probabilistic ML, Gaussian processes, and Bayesian optimization. Arash has also contributed to various projects at Google DeepMind, Max Planck Society, and GSK.

Adrien Badré is a data scientist at Novartis Institute for Biomedical Research in Preclinical Safety. With a strong background in computer science and machine learning, his expertise lies in developing data analysis and machine learning workflows for preclinical safety. He earned his Ph.D. from the University of Oklahoma.

Brianna Paisley is an Advisor Toxicologist that has been in Toxinformatics at Eli Lilly and Company for 13 years. Her background is in chemistry, pharmacology, statistics and bioinformatics. She focuses on identification of structure-mediated and off-target toxicology liabilities for small molecules and genetic medicines by leveraging machine learning and genomics.

Rhiannon Lowe is a UKRT, ERT registered toxicologist and FRSB with over 20 years of experience in nonclinical safety within the pharmaceutical and biotech industries. As Associate Director at Relation, she leads nonclinical safety studies for drug discovery and target safety assessment across multiple modalities and disease indications.

Shantanu Singh coleads the Carpenter-Singh lab at the Broad Institute, applying computer vision, machine learning, and bioimaging to understand human diseases and accelerate drug discovery.

Falgun Shah is the Director of Data Science and Computational Toxicology at Merck Research Laboratories, specializing in applying cheminformatics and machine learning approaches to rapidly advancing better molecules. Falgun has a PhD in Medicinal Chemistry from the University of Mississippi and a Master of Pharmaceutical Sciences from the University of Mumbai, India.

Bjarki Johannesson is a Director within Safety Innovation at AstraZeneca's department of Clinical Pharmacology and Safety Sciences (CPSS). Within his role Bjarki leads efforts to combine AI and Cell Painting for improved and accelerated drug safety predictions. Prior to this, Bjarki led a cross-functional R&D research team at the New York Stem Cell Foundation focused on automated high-throughput differentiation of pluripotent stem cells and AI-driven HCI-based disease modeling. Bjarki completed his PhD in 2010 at EMBL in Heidelberg, Germany.

Dominic Williams leads the preclinical Hepatic Safety team at AstraZeneca, specializing in bespoke hepatic safety assays and integrating AI/ML, drug metabolism, and 3D models. With over 15 years of academic experience, he has authored 150+ publications and


https://doi.org/10.1021/acs.chemrestox.5c00033
Chem. Res. Toxicol. XXXX, XXX, XXX−XXX

actively participates in various EU initiatives for drug safety improvement.

David Rouquie, Head of Toxicology Data Science at Bayer, specializes in digital transformation, early toxicology, and predictive toxicology. With a PhD in Molecular Biology, he leads innovative strategies in chemical safety, integrates AI, and manages interdisciplinary teams. David actively collaborates with internal and external partners, driving impactful safety assessments.

Djork-Arné Clevert holds the position of Vice President at Pfizer, where he is in charge of the Machine Learning Research group. He concentrates his research on the development of machine learning methods that span the entire continuum of drug discovery projects.

Patrick Schwab is Senior Director of Machine Learning and Artificial Intelligence and Head of the Biomedical AI group at GSK.ai. His work aims to advance personalized medicine using machine learning, computational systems biology methods and large-scale health data, such as genetics, multiomics, cell-based assays, and continuous measurements from smart devices and electronic health records, to better understand and treat complex diseases.

Nicola Richmond is Chief Scientist, AI at Recursion where she leads the Data Science and AI organisation and oversees Recursion's Industrialisation of drug discovery with AI. Prior to joining Recursion, Nicola led the AI function at BenevolentAI with a focus on explainable LLMs for target discovery. The majority of Nicola's career was spent at GlaxoSmithKline where she built numerous tech solutions to advance small molecule and therapeutic antibody drug discovery. Nicola trained as a pure mathematician, completing a PhD in Algebra and Algebraic Geometry at the University of Leeds, before transitioning to cheminformatics as a post doctoral fellow. Her cheminformatics research focussed on developing novel approaches to 3D pharmacophore design, working with Peter Willett in the Department of Information Studies at the University of Sheffield.

Christos A. Nicolaou has been working in the small molecule digital chemistry field for over two decades. He has a background in Computer Science (BSc/MSc Florida State University, PhD University of Cyprus) and has contributed to the development of multiple proprietary and commercial cheminformatics software packages. In his most recent role as Senior Director, Digital Chemistry at Novo Nordisk, he is involved in the development and implementation of the small molecule research strategy and leads the Digital Chemistry department. Prior to that he led the establishment of the Digital Chemistry group at Recursion and spent over a decade at the Computational Chemistry and Cheminformatics group of Eli Lilly, Discovery Chemistry and Research Technologies department. Christos has led efforts to map and exploit virtual synthesizable chemistry spaces, mine big pharmaceutical data including chemical reactions, advance machine learning and artificial intelligence research for predictive and generative modeling and, implement model-driven drug discovery.

Raymond J. Gonzalez is Executive Director in Investigative Toxicology at Merck, with over 20 years in the pharmaceutical industry. He leads a multidisciplinary team focused on systems, immunotoxic, computational, in vitro, and analytical toxicology to drive drug discovery and development. Gonzalez collaborates with research, regulatory affairs, and external partners to integrate toxicology assessments for drug development and mentors team members, fostering innovation and continuous professional growth. He holds a PhD in Pharmacology and Toxicology from the University of the Sciences in Philadelphia.

Russell Naven is Head of the Predictive Safety Data Exploration Group in the Preclinical Safety Department of Novartis. With a strong background in Medicinal Chemistry and Predictive Toxicology, his expertise is in the application of Data Science techniques to support the development and validation of predictive safety tools that inspire better drug design and selection. Russ earned his B. Sc. from Manchester University and his PhD in Organic Chemistry from the University of Nottingham in the UK.

Carolin Schramm is a Nonclinical Safety Expert at Sanofi with a background in drug discovery and development, molecular toxicology, and nonclinical strategy development.

Dr. Lewis Vidler is a Senior Director of Structure Based Drug Design at Eli Lilly and Company. He achieved a first-class chemistry degree at the University of Oxford and subsequently a PhD in Computational Medicinal Chemistry at the Institute of Cancer research in London. He joined Eli Lilly that year and spent an initial duration of 7 years there. Following shorter stints as a computational medicinal chemist at UCB and Amphista Therapeutics, he returned to Lilly in 2023, to further progress his career. Bridging computational and medicinal chemistry he has deep expertise in leveraging computational tools to drive drug discovery programs towards meeting their milestones.

Kamel Mansouri, a computational chemist, earned his Ph.D. in computational chemistry from the University of Milano Bicocca, Italy, as a Marie Curie fellow. Since 2020, he has spearheaded computational chemistry endeavors at the NTP Interagency Center for the Evaluation of Alternative Toxicological Methods (NICEATM) within the U.S. National Institute of Environmental Health Sciences (NIEHS). His focus includes QSAR modeling, cheminformatics, and computational toxicology, with a reputation for fostering international collaborations and leading scientific consortia.

W. Patrick Walters is Chief Data Officer at Relay Therapeutics in Cambridge, MA, applying statistics, machine learning, cheminformatics, and computational chemistry to a range of drug discovery projects.

Deidre Dalmas is a Director of Investigative Toxicology and Nonclinical Safety Expert within the pharmaceutical industry. She specializes in investigative mechanistic toxicology, toxicogenomics, and generation/application of cutting-edge capabilities to predict drug-induced toxicity and identify safety biomarkers that aid in selection and advancement of safe medicines from discovery through development. Deidre is the Pharmaceutical Industry Chair for the Health Environment and Safety (HESI) - Emerging Systems in Toxicology for Risk Assessment (e-STAR) Consortium and Co-Chair of the Critical Path Institute Predictive Safety Testing Consortium (CPATH-PSTC) Drug-Induced Vascular Injury Working Group.

Ola Spjuth is Professor at Uppsala University with research focus on AI and automation in drug discovery. He is also CEO of Phenaros Pharmaceuticals where the aim is to accelerate drug discovery with AI and automation.

Anne E. Carpenter is an Institute Scientist at the Broad Institute of MIT and Harvard, where she is also Senior Director of the Imaging Platform. She is a coinventor of the open-source bioimage analysis software CellProfiler and the Cell Painting assay. With a background in cell biology, microscopy, and computational biology, her expertise is in applying methods leveraging the quantitative information in biological images, especially in large-scale experiments. Carpenter earned her BS from Purdue University and her PhD from the University of Illinois at Urbana−Champaign.

Andreas Bender is a Professor of Molecular Informatics at the University of Cambridge, focusing on using chemical and biological data combined with machine learning methods to predict safety- and efficacy-related end points. He is also the Chief Informatics and Technology Officer (CITO) of Pangea Bio, which employs historical


https://doi.org/10.1021/acs.chemrestox.5c00033
Chem. Res. Toxicol. XXXX, XXX, XXX−XXX

information about traditional medicines to increase the probability of success of drug discovery projects in clinical phases. Bender received his PhD from the University of Cambridge and previously worked for Novartis, Leiden University, and AstraZeneca.

## ■ ACKNOWLEDGMENTS

Cartoons in all figures were created with DALLEv2 (https://openai.com/dall-e-2), Microsoft Designer (https://designer.microsoft.com/), and Bioicons (https://bioicons.com), which compiled images from the Database Centre for the life sciences/TogoTV (https://togotv.dbcls.jp) and Servier (https://smart.servier.com). The authors thank Nigel Greene (Recursion) and Nichloas Coltman (Apconix) for helpful comments on an earlier version of the manuscript. S. Seal thanks Elsa Lawrence (University of Cambridge), Adit Shah (UC Berkeley College of Engineering), James Bradshaw (Ignota Lab), Matthew Wilkinson (Ignota Lab), and Adham El Shazly (University of Cambridge) for helpful comments on an earlier version of the paper. S. Seal thanks Jeremy R. Ash (Johnson & Johnson Innovative Medicine) for helpful comments on Pillar 4, Validation of predictive models. GSK is committed to the replacement, reduction, and refinement of animal studies (3Rs). Nonanimal models and alternative technologies are part of the strategy in GSK and are employed where possible. GSK states that when animals are required, application of robust study design principles and peer review minimizes animal use, reduces harm, and improves benefit in studies. The views and conclusions presented in this manuscript are those of the contributing authors and do not necessarily reflect the representative affiliation or individual company's or organization's position of the authors on the subject.

## ■ REFERENCES

(1) Shen, J.; Nicolaou, C. A. Molecular Property Prediction: Recent Trends in the Era of Artificial Intelligence. *Drug Discovery Today Technol.* **2019**, *32*−*33*, 29−36.
(2) Schuhmacher, A.; Hinder, M.; von Stegmann Und Stein, A.; Hartl, D.; Gassmann, O. Analysis of Pharma R&D Productivity - a New Perspective Needed. *Drug Discovery Today* **2023**, *28* (10), 103726.
(3) Sun, D.; Gao, W.; Hu, H.; Zhou, S. Why 90% of Clinical Drug Development Fails and How to Improve It? *Acta Pharm. Sin B* **2022**, *12* (7), 3049−3062.
(4) Rudmann, D. G. On-Target and off-Target-Based Toxicologic Effects. *Toxicol. Pathol.* **2013**, *41* (2), 310−314.
(5) Liu, A.; Seal, S.; Yang, H.; Bender, A. Using Chemical and Biological Data to Predict Drug Toxicity. *SLAS Discov* **2023**, *28* (3), 53−64.
(6) Bilodeau, C.; Jin, W.; Jaakkola, T.; Barzilay, R.; Jensen, K. F. Generative Models for Molecular Discovery: Recent Advances and Challenges. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2022**, *12* (5). DOI: 10.1002/wcms.1608.
(7) Wei, F.; Wang, S.; Gou, X. A Review for Cell-Based Screening Methods in Drug Discovery. *Biophys Rep* **2021**, *7* (6), 504−516.
(8) Eder, J.; Sedrani, R.; Wiesmann, C. The Discovery of First-in-Class Drugs: Origins and Evolution. *Nat. Rev. Drug Discovery* **2014**, *13* (8), 577−587.
(9) Sadri, A. Is Target-Based Drug Discovery Efficient? Discovery and "Off-Target" Mechanisms of All Drugs. *J. Med. Chem.* **2023**, *66* (18), 12651−12677.
(10) Jenkinson, S.; Schmidt, F.; Rosenbrier Ribeiro, L.; Delaunois, A.; Valentin, J.-P. A Practical Guide to Secondary Pharmacology in Drug Discovery. *J. Pharmacol. Toxicol. Methods* **2020**, *105*, 106869.
(11) Bowes, J.; Brown, A. J.; Hamon, J.; Jarolimek, W.; Sridhar, A.; Waldron, G.; Whitebread, S. Reducing Safety-Related Drug Attrition: The Use of in Vitro Pharmacological Profiling. *Nat. Rev. Drug Discovery* **2012**, *11* (12), 909−922.
(12) Xu, J. J.; Henstock, P. V.; Dunn, M. C.; Smith, A. R.; Chabot, J. R.; de Graaf, D. Cellular Imaging Predictions of Clinical Drug-Induced Liver Injury. *Toxicol. Sci.* **2008**, *105* (1), 97−105.
(13) Hartung, T. The (misleading) Role of Animal Models in Drug Development. *Front. Drug Des. Discovery* **2024**, *4*. DOI: 10.3389/fddsv.2024.1355044.
(14) Shanks, N.; Greek, R.; Greek, J. Are Animal Models Predictive for Humans? *Philos. Ethics Humanit. Med.* **2009**, *4*, 2.
(15) Van Norman, G. A. Limitations of Animal Studies for Predicting Toxicity in Clinical Trials: Is It Time to Rethink Our Current Approach? *JACC Basic Transl Sci.* **2019**, *4* (7), 845−854.
(16) Olson, H.; Betton, G.; Robinson, D.; Thomas, K.; Monro, A.; Kolaja, G.; Lilly, P.; Sanders, J.; Sipes, G.; Bracken, W.; Dorato, M.; Van Deun, K.; Smith, P.; Berger, B.; Heller, A. Concordance of the Toxicity of Pharmaceuticals in Humans and in Animals. *Regul. Toxicol. Pharmacol.* **2000**, *32* (1), 56−67.
(17) Ekert, J. E.; Deakyne, J.; Pribul-Allen, P.; Terry, R.; Schofield, C.; Jeong, C. G.; Storey, J.; Mohamet, L.; Francis, J.; Naidoo, A.; Amador, A.; Klein, J.-L.; Rowan, W. Recommended Guidelines for Developing, Qualifying, and Implementing Complex In Vitro Models (CIVMs) for Drug Discovery. *SLAS Discov* **2020**, *25* (10), 1174−1190.
(18) Zushin, P.-J. H.; Mukherjee, S.; Wu, J. C. FDA Modernization Act 2.0: Transitioning beyond Animal Models with Human Cells, Organoids, and AI/ML-Based Approaches. *J. Clin. Invest.* **2023**, *133* (21). DOI: 10.1172/JCI175824.
(19) Han, J. J. FDA Modernization Act 2.0 Allows for Alternatives to Animal Testing. *Artif. Organs* **2023**, *47* (3), 449−450.
(20) Schmeisser, S.; Miccoli, A.; von Bergen, M.; Berggren, E.; Braeuning, A.; Busch, W.; Desaintes, C.; Gourmelon, A.; Grafström, R.; Harrill, J.; Hartung, T.; Herzler, M.; Kass, G. E. N.; Kleinstreuer, N.; Leist, M.; Luijten, M.; Marx-Stoelting, P.; Poetz, O.; van Ravenzwaay, B.; Roggeband, R.; Rogiers, V.; Roth, A.; Sanders, P.; Thomas, R. S.; Marie Vinggaard, A.; Vinken, M.; van de Water, B.; Luch, A.; Tralau, T. New Approach Methodologies in Human Regulatory Toxicology - Not If, but How and When! *Environ. Int.* **2023**, *178*, 108082.
(21) Anthérieu, S.; Chesné, C.; Li, R.; Guguen-Guillouzo, C.; Guillouzo, A. Optimization of the HepaRG Cell Model for Drug Metabolism and Toxicity Studies. *Toxicol. In Vitro* **2012**, *26* (8), 1278−1285.
(22) Fäs, L.; Chen, M.; Tong, W.; Wenz, F.; Hewitt, N. J.; Tu, M.; Sanchez, K.; Zapiórkowska-Blumer, N.; Varga, H.; Kaczmarska, K.; Colombo, M. V.; Filippi, B. G. H. Physiological Liver Microtissue 384-Well Microplate System for Preclinical Hepatotoxicity Assessment of Therapeutic Small Molecule Drugs. *Toxicol. Sci.* **2025**, *203* (1), 79.
(23) *Alternative Methods Accepted by US Agencies*; National Toxicology Program. https://ntp.niehs.nih.gov/whatwestudy/niceatm/accept-methods (accessed 2025-01-18).
(24) Namdari, R.; Jones, K.; Chuang, S. S.; Van Cruchten, S.; Dincer, Z.; Downes, N.; Mikkelsen, L. F.; Harding, J.; Jäckel, S.; Jacobsen, B.; Kinyamu-Akunda, J.; Lortie, A.; Mhedhbi, S.; Mohr, S.; Schmitt, M. W.; Prior, H. Species Selection for Nonclinical Safety Assessment of Drug Candidates: Examples of Current Industry Practice. *Regul. Toxicol. Pharmacol.* **2021**, *126*, 105029.
(25) Scannell, J. W.; Bosley, J.; Hickman, J. A.; Dawson, G. R.; Truebel, H.; Ferreira, G. S.; Richards, D.; Treherne, J. M. Predictive Validity in Drug Discovery: What It Is, Why It Matters and How to Improve It. *Nat. Rev. Drug Discovery* **2022**, *21* (12), 915−931.
(26) Brubaker, D. K.; Lauffenburger, D. A. Translating Preclinical Models to Humans. *Science* **2020**, *367* (6479), 742−743.
(27) Butler, L. D.; Guzzie-Peck, P.; Hartke, J.; Bogdanffy, M. S.; Will, Y.; Diaz, D.; Mortimer-Cassen, E.; Derzi, M.; Greene, N.; DeGeorge, J. J. Current Nonclinical Testing Paradigms in Support of Safe Clinical Trials: An IQ Consortium DruSafe Perspective. *Regul. Toxicol. Pharmacol.* **2017**, *87* (Suppl 3), S1−S15.
(28) Russo, D. P.; Aleksunes, L. M.; Goyak, K.; Qian, H.; Zhu, H. Integrating Concentration-Dependent Toxicity Data and Toxicoki-

netics To Inform Hepatotoxicity Response Pathways. *Environ. Sci. Technol.* **2023**, *57* (33), 12291−12301.
(29) Rusyn, I.; Chiu, W. A.; Wright, F. A. Model Systems and Organisms for Addressing Inter- and Intra-Species Variability in Risk Assessment. *Regul. Toxicol. Pharmacol.* **2022**, *132*, 105197.
(30) Barber, C.; Heghes, C.; Johnston, L. A Framework to Support the Application of the OECD Guidance Documents on (Q)SAR Model Validation and Prediction Assessment for Regulatory Decisions. *Computational Toxicology* **2024**, *30*, 100305.
(31) OECD. *QSAR Assessment Framework: Guidance for the Regulatory Assessment of (Quantitative) Structure Activity Relationship Models and Predictions*; OECD, 2023.
(32) Rich, A. S.; Chan, Y. H.; Birnbaum, B.; Haider, K.; Haimson, J.; Hale, M.; Han, Y.; Hickman, W.; Hoeflich, K. P.; Ortwine, D.; Ozen, A.; Belanger, D. Machine Learning ADME Models in Practice: Four Guidelines from a Successful Lead Optimization Case Study. *ACS Med. Chem. Lett.* **2024**, *15*, 1169.
(33) Bender, A.; Schneider, N.; Segler, M.; Patrick Walters, W.; Engkvist, O.; Rodrigues, T. Evaluation Guidelines for Machine Learning Tools in the Chemical Sciences. *Nat. Rev. Chem.* **2022**, *6* (6), 428−442.
(34) Sculley, D.; Brodley, C. E. Compression and Machine Learning: A New Perspective on Feature Space Vectors. In *Data Compression Conference (DCC'06)*; IEEE, 2006; pp 332−341.
(35) Haghighatlari, M.; Li, J.; Heidar-Zadeh, F.; Liu, Y.; Guan, X.; Head-Gordon, T. Learning to Make Chemical Predictions: The Interplay of Feature Representation, Data, and Machine Learning Methods. *Chem.* **2020**, *6* (7), 1527−1542.
(36) Kirkpatrick, P.; Ellis, C. *Chemical space*; Nature Publishing Group UK. DOI: 10.1038/432823a.
(37) Hawkins, D. M. The Problem of Overfitting. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (1), 1−12.
(38) Sushko, I.; Novotarskyi, S.; Körner, R.; Pandey, A. K.; Kovalishyn, V. V.; Prokopenko, V. V.; Tetko, I. V. Applicability Domain for in Silico Models to Achieve Accuracy of Experimental Measurements. *J. Chemom.* **2010**, *24* (3−4), 202−208.
(39) Dragos, H.; Gilles, M.; Alexandre, V. Predicting the Predictability: A Unified Approach to the Applicability Domain Problem of QSAR Models. *J. Chem. Inf. Model.* **2009**, *49* (7), 1762−1776.
(40) 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), 4791−4810.
(41) Fang, C.; Wang, Y.; Grater, R.; Kapadnis, S.; Black, C.; Trapa, P.; Sciabola, S. Prospective Validation of Machine Learning Algorithms for Absorption, Distribution, Metabolism, and Excretion Prediction: An Industrial Perspective. *J. Chem. Inf. Model.* **2023**, *63* (11), 3263−3274.
(42) Mathai, N.; Chen, Y.; Kirchmair, J. Validation Strategies for Target Prediction Methods. *Brief. Bioinform.* **2020**, *21* (3), 791−802.
(43) Jain, S.; Siramshetty, V. B.; Alves, V. M.; Muratov, E. N.; Kleinstreuer, N.; Tropsha, A.; Nicklaus, M. C.; Simeonov, A.; Zakharov, A. V. Large-Scale Modeling of Multispecies Acute Toxicity End Points Using Consensus of Multitask Deep Learning Methods. *J. Chem. Inf. Model.* **2021**, *61* (2), 653−663.
(44) Sosnin, S.; Karlov, D.; Tetko, I. V.; Fedorov, M. V. Comparative Study of Multitask Toxicity Modeling on a Broad Chemical Space. *J. Chem. Inf. Model.* **2019**, *59* (3), 1062−1072.
(45) Davis, R. L. Mechanism of Action and Target Identification: A Matter of Timing in Drug Discovery. *iScience* **2020**, *23* (9), 101487.
(46) Pognan, F.; Beilmann, M.; Boonen, H. C. M.; Czich, A.; Dear, G.; Hewitt, P.; Mow, T.; Oinonen, T.; Roth, A.; Steger-Hartmann, T.; Valentin, J.-P.; Van Goethem, F.; Weaver, R. J.; Newham, P. The Evolving Role of Investigative Toxicology in the Pharmaceutical Industry. *Nat. Rev. Drug Discovery* **2023**, *22* (4), 317−335.
(47) Priest, B. T.; Bell, I. M.; Garcia, M. L. Role of hERG Potassium Channel Assays in Drug Development. *Channels* **2008**, *2* (2), 87−93.
(48) Sakamuru, S.; Attene-Ramos, M. S.; Xia, M. Mitochondrial Membrane Potential Assay. *Methods Mol. Biol.* **2016**, *1473*, 17−22.
(49) Seal, S.; Carreras-Puigvert, J.; Trapotsi, M.-A.; Yang, H.; Spjuth, O.; Bender, A. Integrating Cell Morphology with Gene Expression and Chemical Structure to Aid Mitochondrial Toxicity Detection. *Commun. Biol.* **2022**, *5* (1), 858.
(50) Bender, A.; Cortés-Ciriano, I. 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 Discovery Today* **2021**, *26* (2), 511−524.
(51) Hoffmann, P.; Warner, B. Are hERG Channel Inhibition and QT Interval Prolongation All There Is in Drug-Induced Torsadogenesis? A Review of Emerging Trends. *J. Pharmacol. Toxicol. Methods* **2006**, *53* (2), 87−105.
(52) Gintant, G. An Evaluation of hERG Current Assay Performance: Translating Preclinical Safety Studies to Clinical QT Prolongation. *Pharmacol. Ther.* **2011**, *129* (2), 109−119.
(53) European Union. Regulation (EC) No 1223/2009 of the European Parliament and of the Council. *Official Journal of the European Union L* **2009**, *342*, 59.
(54) Test Guideline No. 442C In Chemico Skin Sensitisation Assays Addressing the Adverse Outcome Pathway Key Event on Covalent Binding to Proteins. 2022.
(55) Brennan, R. J.; Jenkinson, S.; Brown, A.; Delaunois, A.; Dumotier, B.; Pannirselvam, M.; Rao, M.; Ribeiro, L. R.; Schmidt, F.; Sibony, A.; Timsit, Y.; Sales, V. T.; Armstrong, D.; Lagrutta, A.; Mittlestadt, S. W.; Naven, R.; Peri, R.; Roberts, S.; Vergis, J. M.; Valentin, J.-P. The State of the Art in Secondary Pharmacology and Its Impact on the Safety of New Medicines. *Nat. Rev. Drug Discovery* **2024**, *23* (7), 525−545.
(56) Vo, A. H.; Van Vleet, T. R.; Gupta, R. R.; Liguori, M. J.; Rao, M. S. An Overview of Machine Learning and Big Data for Drug Toxicity Evaluation. *Chem. Res. Toxicol.* **2020**, *33* (1), 20−37.
(57) Cavasotto, C. N.; Scardino, V. Machine Learning Toxicity Prediction: Latest Advances by Toxicity End Point. *ACS Omega* **2022**, *7* (51), 47536−47546.
(58) 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. *arXiv [cs.LG]* **2021**, DOI: 10.48550/arXiv.2102.09548.
(59) Schapin, N.; Majewski, M.; Varela-Rial, A.; Arroniz, C.; Fabritiis, G. D. Machine Learning Small Molecule Properties in Drug Discovery. *Artificial Intelligence Chemistry* **2023**, *1* (2), 100020.
(60) Walters, P. We need better benchmarks for machine learning in drug discovery. *Practical Cheminformatics*. https://practicalcheminformatics.blogspot.com/2023/08/we-need-better-benchmarks-for-machine.html (accessed 2024-07-12), **2023**.
(61) Sutherland, J. J.; Yonchev, D.; Fekete, A.; Urban, L. A Preclinical Secondary Pharmacology Resource Illuminates Target-Adverse Drug Reaction Associations of Marketed Drugs. *Nat. Commun.* **2023**, *14* (1), 4323.
(62) Shimizu, Y.; Sasaki, T.; Yonekawa, E.; Yamazaki, H.; Ogura, R.; Watanabe, M.; Hosaka, T.; Shizu, R.; Takeshita, J.-I.; Yoshinari, K. Association of CYP1A1 and CYP1B1 Inhibition in in Vitro Assays with Drug-Induced Liver Injury. *J. Toxicol. Sci.* **2021**, *46* (4), 167−176.
(63) PubChem. *AID 1207781 - Human HepG2 cell viability assay in 384 well format - PubChem.*. https://pubchem.ncbi.nlm.nih.gov/bioassay/1207781 (accessed 2024-02-20).
(64) Veith, H.; Southall, N.; Huang, R.; James, T.; Fayne, D.; Artemenko, N.; Shen, M.; Inglese, J.; Austin, C. P.; Lloyd, D. G.; Auld, D. S. Comprehensive Characterization of Cytochrome P450 Isozyme Selectivity across Chemical Libraries. *Nat. Biotechnol.* **2009**, *27* (11), 1050−1055.
(65) Narkar, A.; Willard, J. M.; Blinova, K. Chronic Cardiotoxicity Assays Using Human Induced Pluripotent Stem Cell-Derived Cardiomyocytes (hiPSC-CMs). *Int. J. Mol. Sci.* **2022**, *23* (6). 3199.
(66) PubChem. *AID 720551 - qHTS for Inhibitors of KCHN2 3.1: Wildtype qHTS - PubChem.*. https://pubchem.ncbi.nlm.nih.gov/bioassay/720551 (accessed 2024−02−20).


https://doi.org/10.1021/acs.chemrestox.5c00033
*Chem. Res. Toxicol.* XXXX, XXX, XXX−XXX

(67) Du, F.; Yu, H.; Zou, B.; Babcock, J.; Long, S.; Li, M. hERGCentral: A Large Database to Store, Retrieve, and Analyze Compound-Human Ether-à-Go-Go Related Gene Channel Interactions to Facilitate Cardiotoxicity Assessment in Drug Development. *Assay Drug Dev. Technol.* **2011**, *9* (6), 580−588.
(68) Jing, B.; Yan, L.; Li, J.; Luo, P.; Ai, X.; Tu, P. Functional Evaluation and Nephrotoxicity Assessment of Human Renal Proximal Tubule Cells on a Chip. *Biosensors* **2022**, *12* (9), 718.
(69) Wu, Y.; Connors, D.; Barber, L.; Jayachandra, S.; Hanumegowda, U. M.; Adams, S. P. Multiplexed Assay Panel of Cytotoxicity in HK-2 Cells for Detection of Renal Proximal Tubule Injury Potential of Compounds. *Toxicol. In Vitro* **2009**, *23* (6), 1170−1178.
(70) Qiu, X.; Zhou, X.; Miao, Y.; Li, B. An in Vitro Method for Nephrotoxicity Evaluation Using HK-2 Human Kidney Epithelial Cells Combined with Biomarkers of Nephrotoxicity. *Toxicol. Res.* **2018**, *7* (6), 1205−1213.
(71) Aras, M. A.; Hartnett, K. A.; Aizenman, E. Assessment of Cell Viability in Primary Neuronal Cultures. *Curr. Protoc. Neurosci.* **2008**, *44*, Unit 7.18. DOI: 10.1002/0471142301.ns0718s44.
(72) Filous, A. R.; Silver, J. Neurite Outgrowth Assay. *Bio Protoc* **2016**, *6* (1). DOI: 10.21769/BioProtoc.1694.
(73) Bryant, S. D.; Langer, T. *NeuroDeRisk In Silico Toolbox: De-Risking Neurotoxicity and 3D-Pharmacophores*. https://infochim.u-strasbg.fr/CS3_2024/Abstracts_Posters/P_42_BryantSD_Abstract-CS3-2024.pdf (accessed 2024-07-18).
(74) Adams, D. J.; Barlas, B.; McIntyre, R. E.; Salguero, I.; van der Weyden, L.; Barros, A.; Vicente, J. R.; Karimpour, N.; Haider, A.; Ranzani, M.; Turner, G.; Thompson, N. A.; Harle, V.; Olvera-León, R.; Robles-Espinoza, C. D.; Speak, A. O.; Geisler, N.; Weninger, W. J.; Geyer, S. H.; Hewinson, J.; Karp, N. A.; Sanger Mouse Genetics Project; Fu, B.; Yang, F.; Kozik, Z.; Choudhary, J.; Yu, L.; van Ruiten, M. S.; Rowland, B. D.; Lelliott, C. J.; Del Castillo Velasco-Herrera, M.; Verstraten, R.; Bruckner, L.; Henssen, A. G.; Rooimans, M. A.; de Lange, J.; Mohun, T. J.; Arends, M. J.; Kentistou, K. A.; Coelho, P. A.; Zhao, Y.; Zecchini, H.; Perry, J. R. B.; Jackson, S. P.; Balmus, G.; et al. Genetic Determinants of Micronucleus Formation in Vivo. *Nature* **2024**, *627* (8002), 130−136.
(75) Hayashi, M. The Micronucleus Test-Most Widely Used in Vivo Genotoxicity Test. *Genes Environ* **2016**, *38*, 18.
(76) Vijay, U.; Gupta, S.; Mathur, P.; Suravajhala, P.; Bhatnagar, P. Microbial Mutagenicity Assay: Ames Test. *Bio Protoc* **2018**, *8* (6), No. e2763.
(77) Cordelli, E.; Bignami, M.; Pacchierotti, F. Comet Assay: A Versatile but Complex Tool in Genotoxicity Testing. *Toxicol. Res.* **2021**, *10* (1), 68−78.
(78) Creton, S.; Aardema, M. J.; Carmichael, P. L.; Harvey, J. S.; Martin, F. L.; Newbold, R. F.; O'Donovan, M. R.; Pant, K.; Poth, A.; Sakai, A.; Sasaki, K.; Scott, A. D.; Schechtman, L. M.; Shen, R. R.; Tanaka, N.; Yasaei, H. Cell Transformation Assays for Prediction of Carcinogenic Potential: State of the Science and Future Research Needs. *Mutagenesis* **2012**, *27* (1), 93−101.
(79) Kuo, B.; Beal, M. A.; Wills, J. W.; White, P. A.; Marchetti, F.; Nong, A.; Barton-Maclaren, T. S.; Houck, K.; Yauk, C. L. Comprehensive Interpretation of in Vitro Micronucleus Test Results for 292 Chemicals: From Hazard Identification to Risk Assessment Application. *Arch. Toxicol.* **2022**, *96* (7), 2067−2085.
(80) Xu, C.; Cheng, F.; Chen, L.; Du, Z.; Li, W.; Liu, G.; Lee, P. W.; Tang, Y. In Silico Prediction of Chemical Ames Mutagenicity. *J. Chem. Inf. Model.* **2012**, *52* (11), 2840−2847.
(81) PubChem. *Chemical Carcinogenesis Research Information System (CCRIS) - PubChem. Data Source*. https://pubchem.ncbi.nlm.nih.gov/source/22070 (accessed 2024-02-21).
(82) Shanle, E. K.; Xu, W. Endocrine Disrupting Chemicals Targeting Estrogen Receptor Signaling: Identification and Mechanisms of Action. *Chem. Res. Toxicol.* **2011**, *24* (1), 6−19.
(83) Tinwell, H.; Karmaus, A.; Gaskell, V.; Gomes, C.; Grant, C.; Holmes, T.; Jonas, A.; Kellum, S.; Krüger, K.; Malley, L.; Melching-Kollmuss, S.; Mercier, O.; Pandya, H.; Placke, T.; Settivari, R.; De Waen, B. Evaluating H295R Steroidogenesis Assay Data for Robust Interpretation. *Regul. Toxicol. Pharmacol.* **2023**, *143*, 105461.
(84) US EPA. *Exploring ToxCast Data*; US EPA, 2017.
(85) Gerberick, G. F.; Ryan, C. A.; Dearman, R. J.; Kimber, I. Local Lymph Node Assay (LLNA) for Detection of Sensitization Capacity of Chemicals. *Methods* **2007**, *41* (1), 54−60.
(86) Wang, C.-C.; Lin, Y.-C.; Wang, S.-S.; Shih, C.; Lin, Y.-H.; Tung, C.-W. SkinSensDB: A Curated Database for Skin Sensitization Assays. *J. Cheminform.* **2017**, *9*, 5.
(87) Barile, F. A. Validating and Troubleshooting Ocular in Vitro Toxicology Tests. *J. Pharmacol. Toxicol. Methods* **2010**, *61* (2), 136−145.
(88) Zhou, Y.; Wang, Z.; Huang, Z.; Li, W.; Chen, Y.; Yu, X.; Tang, Y.; Liu, G. In Silico Prediction of Ocular Toxicity of Compounds Using Explainable Machine Learning and Deep Learning Approaches. *J. Appl. Toxicol.* **2024**, *44*, 892.
(89) van Tilborg, D.; Alenicheva, A.; Grisoni, F. Exposing the Limitations of Molecular Machine Learning with Activity Cliffs. *J. Chem. Inf. Model.* **2022**, *62* (23), 5938−5951.
(90) Zhang, Z.; Zhao, B.; Xie, A.; Bian, Y.; Zhou, S. Activity Cliff Prediction: Dataset and Benchmark. *arXiv [q-bio.BM]* **2023**, DOI: 10.48550/arXiv.2302.07541.
(91) Dablander, M.; Hanser, T.; Lambiotte, R.; Morris, G. M. Exploring QSAR Models for Activity-Cliff Prediction. *J. Cheminform.* **2023**, *15* (1), 47.
(92) *Document: Experimental in vitro DMPK and physicochemical data on a set of publicly disclosed compounds*. https://www.ebi.ac.uk/chembl/document_report_card/CHEMBL3301361/ (accessed 2024-07-12).
(93) *U.S. environmental protection agency - evaluation of existing QSAR models and structural alerts and development of new ensemble models for genotoxicity using a newly compiled experimental dataset*. https://catalog.data.gov/dataset/evaluation-of-existing-qsar-models-and-structural-alerts-and-development-of-new-ensemble-m (accessed 2024-07-12).
(94) Madia, F.; Kirkland, D.; Morita, T.; White, P.; Asturiol, D.; Corvi, R. EURL ECVAM Genotoxicity and Carcinogenicity Database of Substances Eliciting Negative Results in the Ames Test: Construction of the Database. *Mutat Res. Genet Toxicol Environ. Mutagen* **2020**, *854*−*855*, 503199.
(95) Corvi, R.; Madia, F. Eurl ECVAM Genotoxicity and Carcinogenicity Consolidated Database of Ames Positive Chemicals. *Eur. Comm. Jt. Res. Cent. [Dataset]* **2018**.
(96) Martínez, M. J.; et al. *Ames Mutagenicity Dataset for Multi-Task Learning* **2022**, DOI: 10.17632/KTC6GBFSBH.2.
(97) Hansen, K.; Mika, S.; Schroeter, T.; Sutter, A.; ter Laak, A.; Steger-Hartmann, T.; Heinrich, N.; Müller, K.-R. Benchmark Data Set for in Silico Prediction of Ames Mutagenicity. *J. Chem. Inf. Model.* **2009**, *49* (9), 2077−2081.
(98) Ciallella, H. L.; Russo, D. P.; Sharma, S.; Li, Y.; Sloter, E.; Sweet, L.; Huang, H.; Zhu, H. Predicting Prenatal Developmental Toxicity Based On the Combination of Chemical Structures and Biological Data. *Environ. Sci. Technol.* **2022**, *56* (9), 5984−5998.
(99) Yuan, H.; Wang, Y.; Cheng, Y. Local and Global Quantitative Structure-Activity Relationship Modeling and Prediction for the Baseline Toxicity. *J. Chem. Inf. Model.* **2007**, *47* (1), 159−169.
(100) Di Lascio, E.; Gerebtzoff, G.; Rodríguez-Pérez, R. Systematic Evaluation of Local and Global Machine Learning Models for the Prediction of ADME Properties. *Mol. Pharmaceutics* **2023**, *20* (3), 1758−1767.
(101) Enoch, S. J.; Cronin, M. T. D.; Schultz, T. W.; Madden, J. C. An Evaluation of Global QSAR Models for the Prediction of the Toxicity of Phenols to Tetrahymena Pyriformis. *Chemosphere* **2008**, *71* (7), 1225−1232.
(102) Cumming, J. G.; Davis, A. M.; Muresan, S.; Haeberlein, M.; Chen, H. Chemical Predictive Modelling to Improve Compound Quality. *Nat. Rev. Drug Discovery* **2013**, *12* (12), 948−962.
(103) Gomis-Tena, J.; Brown, B. M.; Cano, J.; Trenor, B.; Yang, P.-C.; Saiz, J.; Clancy, C. E.; Romero, L. When Does the IC50 Accurately Assess the Blocking Potency of a Drug? *J. Chem. Inf. Model.* **2020**, *60* (3), 1779−1790.

(104) Aronov, A. M. Common Pharmacophores for Uncharged Human Ether-a-Go-Go-Related Gene (hERG) Blockers. *J. Med. Chem.* **2006**, *49* (23), 6917−6921.
(105) Creanza, T. M.; Delre, P.; Ancona, N.; Lentini, G.; Saviano, M.; Mangiatordi, G. F. Structure-Based Prediction of hERG-Related Cardiotoxicity: A Benchmark Study. *J. Chem. Inf. Model.* **2021**, *61* (9), 4758−4770.
(106) Göller, A. H.; Kuhnke, L.; Montanari, F.; Bonin, A.; Schneckener, S.; Ter Laak, A.; Wichard, J.; Lobell, M.; Hillisch, A. Bayer's in Silico ADMET Platform: A Journey of Machine Learning over the Past Two Decades. *Drug Discovery Today* **2020**, *25* (9), 1702−1709.
(107) Wang, Y.; Xiong, J.; Xiao, F.; Zhang, W.; Cheng, K.; Rao, J.; Niu, B.; Tong, X.; Qu, N.; Zhang, R.; Wang, D.; Chen, K.; Li, X.; Zheng, M. LogD7.4 Prediction Enhanced by Transferring Knowledge from Chromatographic Retention Time, Microscopic pKa and logP. *J. Cheminform.* **2023**, *15* (1), 76.
(108) Mannhold, R.; Poda, G. I.; Ostermann, C.; Tetko, I. V. Calculation of Molecular Lipophilicity: State-of-the-Art and Comparison of Log P Methods on More than 96,000 Compounds. *J. Pharm. Sci.* **2009**, *98* (3), 861−893.
(109) Mansouri, K.; Moreira-Filho, J. T.; Lowe, C. N.; Charest, N.; Martin, T.; Tkachenko, V.; Judson, R.; Conway, M.; Kleinstreuer, N. C.; Williams, A. J. Free and Open-Source QSAR-Ready Workflow for Automated Standardization of Chemical Structures in Support of QSAR Modeling. *J. Cheminform.* **2024**, *16* (1), 19.
(110) DeVane, C. L.; Boulton, D. W. Great Expectations in Stereochemistry: Focus on Antidepressants. *CNS Spectr.* **2002**, *7* (S1), 28−33.
(111) McConathy, J.; Owens, M. J. Stereochemistry in Drug Action. *Prim. Care Companion J. Clin. Psychiatry* **2003**, *5* (2), 70−73.
(112) Gal, J. Molecular Chirality in Chemistry and Biology: Historical Milestones. *Helv. Chim. Acta* **2013**, *96* (9), 1617−1657.
(113) Heller, S. R.; McNaught, A.; Pletnev, I.; Stein, S.; Tchekhovskoi, D. InChI, the IUPAC International Chemical Identifier. *J. Cheminform.* **2015**, *7* (1), 23.
(114) Hähnke, V. D.; Kim, S.; Bolton, E. E. PubChem Chemical Structure Standardization. *J. Cheminform.* **2018**, *10* (1), 36.
(115) Bento, A. P.; Hersey, A.; Félix, E.; Landrum, G.; Gaulton, A.; Atkinson, F.; Bellis, L. J.; De Veij, M.; Leach, A. R. An Open Source Chemical Structure Curation Pipeline Using RDKit. *J. Cheminform.* **2020**, *12* (1), 51.
(116) Sayle, R. A. So You Think You Understand Tautomerism? *J. Comput. Aided Mol. Des.* **2010**, *24* (6−7), 485−496.
(117) Robello, M.; Barresi, E.; Baglini, E.; Salerno, S.; Taliani, S.; Settimo, F. D. The Alpha Keto Amide Moiety as a Privileged Motif in Medicinal Chemistry: Current Insights and Emerging Opportunities. *J. Med. Chem.* **2021**, *64* (7), 3508−3545.
(118) Dhaked, D. K.; Ihlenfeldt, W.-D.; Patel, H.; Delannée, V.; Nicklaus, M. C. Toward a Comprehensive Treatment of Tautomerism in Chemoinformatics Including in InChI V2. *J. Chem. Inf. Model.* **2020**, *60* (3), 1253−1275.
(119) Guasch, L.; Yapamudiyansel, W.; Peach, M. L.; Kelley, J. A.; Barchi, J. J., Jr; Nicklaus, M. C. Experimental and Chemoinformatics Study of Tautomerism in a Database of Commercially Available Screening Samples. *J. Chem. Inf. Model.* **2016**, *56* (11), 2149−2161.
(120) Sitzmann, M.; Ihlenfeldt, W.-D.; Nicklaus, M. C. Tautomerism in Large Databases. *J. Comput. Aided Mol. Des.* **2010**, *24* (6−7), 521−551.
(121) Dhaked, D. K.; Nicklaus, M. C. Tautomeric Conflicts in Forty Small-Molecule Databases. *J. Chem. Inf. Model.* **2024**, *64* (19), 7409−7421.
(122) Pérez-Isidoro, R.; Sierra-Valdez, F. J.; Ruiz-Suárez, J. C. Anesthetic Diffusion through Lipid Membranes Depends on the Protonation Rate. *Sci. Rep.* **2014**, *4*, 7534.
(123) Shin, J. M.; Sachs, G. Pharmacology of Proton Pump Inhibitors. *Curr. Gastroenterol. Rep.* **2008**, *10* (6), 528−534.
(124) Miljković, F.; Martinsson, A.; Obrezanova, O.; Williamson, B.; Johnson, M.; Sykes, A.; Bender, A.; Greene, N. Machine Learning Models for Human In Vivo Pharmacokinetic Parameters with In-House Validation. *Mol. Pharmaceutics* **2021**, *18* (12), 4520−4530.
(125) Seal, S.; Trapotsi, M.-A.; Subramanian, V.; Spjuth, O.; Greene, N.; Bender, A. PKSmart: An Open-Source Computational Model to Predict in Vivo Pharmacokinetics of Small Molecules. *bioRxiv* **2024**, DOI: 10.1101/2024.02.02.578658.
(126) Zheng, J.; Leito, I.; Green, W. Widespread Misinterpretation of pKa Terminology for Zwitterionic Compounds and Its Consequences. *J. Chem. Inf. Model.* **2024**, *64*, 8838.
(127) Bender, A.; Cortes-Ciriano, I. Artificial Intelligence in Drug Discovery: What Is Realistic, What Are Illusions? Part 2: A Discussion of Chemical and Biological Data. *Drug Discovery Today* **2021**, *26* (4), 1040−1052.
(128) Xu, W.; Hou, Y.; Hung, Y. S.; Zou, Y. A Comparative Analysis of Spearman's Rho and Kendall's Tau in Normal and Contaminated Normal Models. *Signal Processing* **2013**, *93* (1), 261−276.
(129) Landrum, G. A.; Riniker, S. Combining IC50 or Ki Values from Different Sources Is a Source of Significant Noise. *J. Chem. Inf. Model.* **2024**, *64* (5), 1560−1567.
(130) Dawson, R. How Significant Is a Boxplot Outlier? *J. Stat. Educ.* **2011**, *19* (2). DOI: 10.1080/10691898.2011.11889610.
(131) Kalliokoski, T.; Kramer, C.; Vulpetti, A.; Gedeck, P. Comparability of Mixed $IC_{50}$ Data - a Statistical Analysis. *PLoS One* **2013**, *8* (4), No. e61007.
(132) Escher, B. I.; Henneberger, L.; König, M.; Schlichting, R.; Fischer, F. C. Cytotoxicity Burst? Differentiating Specific from Nonspecific Effects in Tox21 in Vitro Reporter Gene Assays. *Environ. Health Perspect.* **2020**, *128* (7), 77007.
(133) Seal, S.; Yang, H.; Vollmers, L.; Bender, A. Comparison of Cellular Morphological Descriptors and Molecular Fingerprints for the Prediction of Cytotoxicity- and Proliferation-Related Assays. *Chem. Res. Toxicol.* **2021**, *34* (2), 422−437.
(134) Azqueta, A.; Stopper, H.; Zegura, B.; Dusinska, M.; Møller, P. Do Cytotoxicity and Cell Death Cause False Positive Results in the in Vitro Comet Assay? *Mutat Res. Genet Toxicol Environ. Mutagen* **2022**, *881*, 503520.
(135) Owen, A. A Robust Hybrid of Lasso and Ridge Regression. *Prediction and Discovery* **2007**, *443*, 59.
(136) Lewinson, E. 3 *Robust Linear Regression Models to Handle Outliers*. NVIDIA Technical Blog. https://developer.nvidia.com/blog/dealing-with-outliers-using-three-robust-linear-regression-models/ (accessed 2025-03-22).
(137) Mohebali, B.; Tahmassebi, A.; Meyer-Baese, A.; Gandomi, A. H. Probabilistic Neural Networks. In *Handbook of Probabilistic Models*; Samui, P., Tien Bui, D., Chakraborty, S.; Deo, R. C., Eds.; Elsevier, 2020; pp 347−367.
(138) Grandjean, P. Paracelsus Revisited: The Dose Concept in a Complex World. *Basic Clin. Pharmacol. Toxicol.* **2016**, *119* (2), 126−132.
(139) Smit, I. A.; Afzal, A. M.; Allen, C. H. G.; Svensson, F.; Hanser, T.; Bender, A. Systematic Analysis of Protein Targets Associated with Adverse Events of Drugs from Clinical Trials and Postmarketing Reports. *Chem. Res. Toxicol.* **2021**, *34* (2), 365−384.
(140) Lafront, C.; Germain, L.; Weidmann, C.; Audet-Walsh, É. A Systematic Study of the Impact of Estrogens and Selective Estrogen Receptor Modulators on Prostate Cancer Cell Proliferation. *Sci. Rep.* **2020**, *10* (1), 4024.
(141) Mervin, L. H.; Trapotsi, M.-A.; Afzal, A. M.; Barrett, I. P.; Bender, A.; Engkvist, O. Probabilistic Random Forest Improves Bioactivity Predictions close to the Classification Threshold by Taking into Account Experimental Uncertainty. *J. Cheminform.* **2021**, *13* (1), 62.
(142) Bender, A.; Glen, R. C. Molecular Similarity: A Key Technique in Molecular Informatics. *Org. Biomol. Chem.* **2004**, *2* (22), 3204−3218.
(143) Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics, 2 Vol. Set: Vol. I: Alphabetical Listing/Vol. II: Appendices, References*; Wiley, 2009.

(144) 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* (7), 709−715.
(145) Ruddigkeit, L.; van Deursen, R.; Blum, L. C.; Reymond, J.-L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.* **2012**, *52* (11), 2864−2875.
(146) Seal, S.; Trapotsi, M.-A.; Spjuth, O.; Singh, S.; Carreras-Puigvert, J.; Greene, N.; Bender, A.; Carpenter, A. E. A Decade in a Systematic Review: The Evolution and Impact of Cell Painting. *bioRxiv* **2024**, DOI: 10.1101/2024.05.04.592531.
(147) Moriwaki, H.; Tian, Y.-S.; Kawashita, N.; Takagi, T. Mordred: A Molecular Descriptor Calculator. *J. Cheminform.* **2018**, *10* (1), 4.
(148) Ewald, J. D.; Titterton, K. L.; Bäuerle, A.; Beatson, A.; Boiko, D. A.; Cabrera, A. A.; Cheah, J.; Cimini, B. A.; Gorissen, B.; Jones, T.; Karczewski, K. J.; Rouquie, D.; Seal, S.; Weisbart, E.; White, B.; Carpenter, A. E.; Singh, S. Cell Painting for Cytotoxicity and Mode-of-Action Analysis in Primary Human Hepatocytes. *bioRxiv* **2025**. DOI: 10.1101/2025.01.22.634152.
(149) Pruteanu, L.-L.; Bender, A. Using Transcriptomics and Cell Morphology Data in Drug Discovery: The Long Road to Practice. *ACS Med. Chem. Lett.* **2023**, *14* (4), 386−395.
(150) Chandrasekaran, S. N.; Ceulemans, H.; Boyd, J. D.; Carpenter, A. E. Image-Based Profiling for Drug Discovery: Due for a Machine-Learning Upgrade? *Nat. Rev. Drug Discovery* **2021**, *20* (2), 145−159.
(151) Boldini, D.; Ballabio, D.; Consonni, V.; Todeschini, R.; Grisoni, F.; Sieber, S. A. Effectiveness of Molecular Fingerprints for Exploring the Chemical Space of Natural Products. *J. Cheminform.* **2024**, *16* (1), 35.
(152) Gao, K.; Nguyen, D. D.; Sresht, V.; Mathiowetz, A. M.; Tu, M.; Wei, G.-W. Are 2D Fingerprints Still Valuable for Drug Discovery? *Phys. Chem. Chem. Phys.* **2020**, *22* (16), 8373−8390.
(153) David, L.; Thakkar, A.; Mercado, R.; Engkvist, O. Molecular Representations in AI-Driven Drug Discovery: A Review and Practical Guide. *J. Cheminform.* **2020**, *12* (1), 56.
(154) Venkatraman, V. FP-ADMET: A Compendium of Fingerprint-Based ADMET Prediction Models. *J. Cheminform.* **2021**, *13* (1), 75.
(155) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42* (6), 1273−1280.
(156) Lundberg, S. M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. *Adv. Neural Inf. Process. Syst.* **2017**, 4765−4774.
(157) Ribeiro, M. T.; Singh, S.; Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; KDD '16; Association for Computing Machinery: New York, NY, USA, 2016; pp 1135−1144.
(158) Glem, R. C.; Bender, A.; Arnby, C. H.; Carlsson, L.; Boyer, S.; Smith, J. Circular Fingerprints: Flexible Molecular Descriptors with Applications from Physical Chemistry to ADME. *IDrugs* **2006**, *9* (3), 199−204.
(159) *FingerprintGenerator tutorial - RDKit blog*. https://greglandrum.github.io/rdkit-blog/posts/2023-01-18-fingerprint-generator-tutorial.html (accessed 2025-01-18).
(160) Gütlein, M.; Kramer, S. Filtered Circular Fingerprints Improve Either Prediction or Runtime Performance While Retaining Interpretability. *J. Cheminform.* **2016**, *8*, 60.
(161) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50* (5), 742−754.
(162) Prasanna, S.; Doerksen, R. J. Topological Polar Surface Area: A Useful Descriptor in 2D-QSAR. *Curr. Med. Chem.* **2009**, *16* (1), 21−41.
(163) Kier, L. B.; Hall, L. H. An Electrotopological-State Index for Atoms in Molecules. *Pharm. Res.* **1990**, *7* (8), 801−807.
(164) Yukawa, T.; Naven, R. Utility of Physicochemical Properties for the Prediction of Toxicological Outcomes: Takeda Perspective. *ACS Med. Chem. Lett.* **2020**, *11* (2), 203−209.
(165) Seal, S.; Spjuth, O.; Hosseini-Gerami, L.; García-Ortegón, M.; Singh, S.; Bender, A.; Carpenter, A. E. Insights into Drug Cardiotoxicity from Biological and Chemical Data: The First Public Classifiers for FDA Drug-Induced Cardiotoxicity Rank. *J. Chem. Inf. Model.* **2024**, *64*, 1172.
(166) Hong, H.; Xie, Q.; Ge, W.; Qian, F.; Fang, H.; Shi, L.; Su, Z.; Perkins, R.; Tong, W. Mold(2), Molecular Descriptors from 2D Structures for Chemoinformatics and Toxicoinformatics. *J. Chem. Inf. Model.* **2008**, *48* (7), 1337−1344.
(167) Anighoro, A.; Bajorath, J. Three-Dimensional Similarity in Molecular Docking: Prioritizing Ligand Poses on the Basis of Experimental Binding Modes. *J. Chem. Inf. Model.* **2016**, *56* (3), 580−587.
(168) Nettles, J. H.; Jenkins, J. L.; Bender, A.; Deng, Z.; Davies, J. W.; Glick, M. Bridging Chemical and Biological Space: "Target Fishing" Using 2D and 3D Molecular Descriptors. *J. Med. Chem.* **2006**, *49* (23), 6802−6810.
(169) Zankov, D.; Madzhidov, T.; Varnek, A.; Polishchuk, P. Chemical Complexity Challenge: Is Multi-instance Machine Learning a Solution? *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2024**, *14* (1). DOI: 10.1002/wcms.1698.
(170) Moon, K.; Im, H.-J.; Kwon, S. 3D Graph Contrastive Learning for Molecular Property Prediction. *Bioinformatics* **2023**, *39* (6). DOI: 10.1093/bioinformatics/btad371.
(171) Bahia, M. S.; Kaspi, O.; Touitou, M.; Binayev, I.; Dhail, S.; Spiegel, J.; Khazanov, N.; Yosipof, A.; Senderowitz, H. A Comparison between 2D and 3D Descriptors in QSAR Modeling Based on Bio-Active Conformations. *Mol. Inform.* **2023**, *42* (4), No. e2200186.
(172) Mapari, S.; Camarda, K. V. Use of Three-Dimensional Descriptors in Molecular Design for Biologically Active Compounds. *Curr. Opin. Chem. Eng.* **2020**, *27*, 60−64.
(173) Devinyak, O.; Havrylyuk, D.; Lesyk, R. 3D-MoRSE Descriptors Explained. *J. Mol. Graph. Model.* **2014**, *54*, 194−203.
(174) Katritzky, A. R.; Mu, L.; Lobanov, V. S.; Karelson, M. Correlation of Boiling Points with Molecular Structure. 1. A Training Set of 298 Diverse Organics and a Test Set of 9 Simple Inorganics. *J. Phys. Chem.* **1996**, *100* (24), 10400−10407.
(175) Sliwoski, G.; Mendenhall, J.; Meiler, J. Autocorrelation Descriptor Improvements for QSAR: 2DA_Sign and 3DA_Sign. *J. Comput. Aided Mol. Des.* **2016**, *30* (3), 209−217.
(176) Beglari, M.; Goudarzi, N.; Shahsavani, D.; Arab Chamjangali, M.; Mozafari, Z. Combination of Radial Distribution Functions as Structural Descriptors with Ligand-Receptor Interaction Information in the QSAR Study of Some 4-Anilinoquinazoline Derivatives as Potent EGFR Inhibitors. *Struct. Chem.* **2020**, *31* (4), 1481−1491.
(177) Gramatica, P. WHIM Descriptors of Shape. *QSAR Comb. Sci.* **2006**, *25* (4), 327−332.
(178) von Korff, M.; Freyss, J.; Sander, T. Flexophore, a New Versatile 3D Pharmacophore Descriptor That Considers Molecular Flexibility. *J. Chem. Inf. Model.* **2008**, *48* (4), 797−810.
(179) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural Message Passing for Quantum Chemistry. In *Proceedings of the 34th International Conference on Machine Learning*; Precup, D., Teh, Y. W., Eds.; Proceedings of Machine Learning Research; PMLR, 06−11 Aug 2017; Vol. *70*, pp 1263−1272.
(180) Heid, E.; Greenman, K. P.; Chung, Y.; Li, S.-C.; Graff, D. E.; Vermeire, F. H.; Wu, H.; Green, W. H.; McGill, C. J. Chemprop: A Machine Learning Package for Chemical Property Prediction. *J. Chem. Inf. Model.* **2024**, *64* (1), 9−17.
(181) Joshi, C. K. *Transformers are Graph Neural Networks*. The Gradient. https://thegradient.pub/transformers-are-graph-neural-networks/ (accessed 2024-10-05), 2020.
(182) Méndez-Lucio, O.; Nicolaou, C. A.; Earnshaw, B. MolE: A Foundation Model for Molecular Graphs Using Disentangled Attention. *Nat. Commun.* **2024**, *15* (1), 9431.
(183) Jiang, D.; Wu, Z.; Hsieh, C.-Y.; Chen, G.; Liao, B.; Wang, Z.; Shen, C.; Cao, D.; Wu, J.; Hou, T. Could Graph Neural Networks Learn Better Molecular Representation for Drug Discovery? A Comparison Study of Descriptor-Based and Graph-Based Models. *J. Cheminform.* **2021**, *13* (1), 12.
(184) Moshkov, N.; Becker, T.; Yang, K.; Horvath, P.; Dancik, V.; Wagner, B. K.; Clemons, P. A.; Singh, S.; Carpenter, A. E.; Caicedo, J. C.

Predicting Compound Activity from Phenotypic Profiles and Chemical Structures. *Nat. Commun.* **2023**, *14* (1), 1967.
(185) Myung, Y.; de Sá, A. G. C.; Ascher, D. B. Deep-PK: Deep Learning for Small Molecule Pharmacokinetic and Toxicity Prediction. *Nucleic Acids Res.* **2024**, *52* (W1), W469−W475.
(186) Satorras, V. G.; Hoogeboom, E.; Welling, M. E(n) Equivariant Graph Neural Networks. In *Proceedings of the 38th International Conference on Machine Learning*; Meila, M., Zhang, T., Eds.; Proceedings of Machine Learning Research; PMLR, 18−24 Jul 2021; Vol. *139*, pp 9323−9332.
(187) Thomas, N.; Smidt, T.; Kearnes, S.; Yang, L.; Li, L.; Kohlhoff, K.; Riley, P. Tensor Field Networks: Rotation- and Translation-Equivariant Neural Networks for 3D Point Clouds. *arXiv [cs.LG]* **2018**, DOI: 10.48550/arXiv.1802.08219.
(188) Duval, A.; Mathis, S. V.; Joshi, C. K.; Schmidt, V.; Miret, S.; Malliaros, F. D.; Cohen, T.; Lio, P.; Bengio, Y.; Bronstein, M. A Hitchhiker's Guide to Geometric GNNs for 3D Atomic Systems. *arXiv [cs.LG]* **2023**, DOI: 10.48550/arXiv.2312.07511.
(189) Hendrychová, T.; Anzenbacherová, E.; Hudeček, J.; Skopalík, J.; Lange, R.; Hildebrandt, P.; Otyepka, M.; Anzenbacher, P. Flexibility of Human Cytochrome P450 Enzymes: Molecular Dynamics and Spectroscopy Reveal Important Function-Related Variations. *Biochim. Biophys. Acta* **2011**, *1814* (1), 58−68.
(190) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583−589.
(191) Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G. R.; Wang, J.; Cong, Q.; Kinch, L. N.; Schaeffer, R. D.; Millán, C.; Park, H.; Adams, C.; Glassman, C. R.; DeGiovanni, A.; Pereira, J. H.; Rodrigues, A. V.; van Dijk, A. A.; Ebrecht, A. C.; Opperman, D. J.; Sagmeister, T.; Buhlheller, C.; Pavkov-Keller, T.; Rathinaswamy, M. K.; Dalwadi, U.; Yip, C. K.; Burke, J. E.; Garcia, K. C.; Grishin, N. V.; Adams, P. D.; Read, R. J.; Baker, D. Accurate Prediction of Protein Structures and Interactions Using a Three-Track Neural Network. *Science* **2021**, *373* (6557), 871−876.
(192) Smith, S. W. Chiral Toxicology: It's the Same Thing···only Different. *Toxicol. Sci.* **2009**, *110* (1), 4−30.
(193) Lin, M.-C.; Hwang, M.-T.; Chang, H.-G.; Lin, C.-S.; Lin, G. Benzene-1,2-, 1,3-, and 1,4-Di-N-Substituted Carbamates as Conformationally Constrained Inhibitors of Acetylcholinesterase. *J. Biochem. Mol. Toxicol.* **2007**, *21* (6), 348−353.
(194) Vistoli, G.; Pedretti, A.; Testa, B.; Matucci, R. The Conformational and Property Space of Acetylcholine Bound to Muscarinic Receptors: An Entropy Component Accounts for the Subtype Selectivity of Acetylcholine. *Arch. Biochem. Biophys.* **2007**, *464* (1), 112−121.
(195) Sokouti, B.; Hamzeh-Mivehroud, M. 6D-QSAR for Predicting Biological Activity of Human Aldose Reductase Inhibitors Using Quasar Receptor Surface Modeling. *BMC Chem. Biol.* **2023**, *17* (1), 63.
(196) Senese, C. L.; Duca, J.; Pan, D.; Hopfinger, A. J.; Tseng, Y. J. 4D-Fingerprints, Universal QSAR and QSPR Descriptors. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (5), 1526−1539.
(197) Teramoto, R.; Kashima, H. Prediction of Protein-Ligand Binding Affinities Using Multiple Instance Learning. *J. Mol. Graph. Model.* **2010**, *29* (3), 492−497.
(198) Zankov, D. V.; Matveieva, M.; Nikonenko, A. V.; Nugmanov, R. I.; Baskin, I. I.; Varnek, A.; Polishchuk, P.; Madzhidov, T. I. QSAR Modeling Based on Conformation Ensembles Using a Multi-Instance Learning Approach. *J. Chem. Inf. Model.* **2021**, *61* (10), 4913−4923.
(199) Wu, Z.; Wang, J.; Du, H.; Jiang, D.; Kang, Y.; Li, D.; Pan, P.; Deng, Y.; Cao, D.; Hsieh, C.-Y.; Hou, T. Chemistry-Intuitive Explanation of Graph Neural Networks for Molecular Property Prediction with Substructure Masking. *Nat. Commun.* **2023**, *14* (1), 2585.
(200) Wójcikowski, M.; Kukiełka, M.; Stepniewska-Dziubinska, M. M.; Siedlecki, P. Development of a Protein-Ligand Extended Connectivity (PLEC) Fingerprint and Its Application for Binding Affinity Predictions. *Bioinformatics* **2019**, *35* (8), 1334−1341.
(201) Cang, Z.; Mu, L.; Wei, G.-W. Representability of Algebraic Topology for Biomolecules in Machine Learning Based Scoring and Virtual Screening. *PLoS Comput. Biol.* **2018**, *14* (1), No. e1005929.
(202) Townshend, R. J. L.; Vögele, M.; Suriana, P.; Derry, A.; Powers, A.; Laloudakis, Y.; Balachandar, S.; Jing, B.; Anderson, B.; Eismann, S.; Kondor, R.; Altman, R. B.; Dror, R. O. ATOM3D: Tasks On Molecules in Three Dimensions. *arXiv [cs.LG]* **2020**, DOI: 10.48550/arXiv.2012.04035.
(203) Cremer, J.; Medrano Sandonas, L.; Tkatchenko, A.; Clevert, D.-A.; De Fabritiis, G. Equivariant Graph Neural Networks for Toxicity Prediction. *Chem. Res. Toxicol.* **2023**, *36* (10), 1561−1573.
(204) Wieder, O.; Kohlbacher, S.; Kuenemann, M.; Garon, A.; Ducrot, P.; Seidel, T.; Langer, T. A Compact Review of Molecular Property Prediction with Graph Neural Networks. *Drug Discovery Today Technol.* **2020**, *37*, 1−12.
(205) Dias, A. L.; Bustillo, L.; Rodrigues, T. Limitations of Representation Learning in Small Molecule Property Prediction. *Nat. Commun.* **2023**, *14* (1), 6394.
(206) Wang, Y.; Wang, J.; Cao, Z.; Barati Farimani, A. Molecular Contrastive Learning of Representations via Graph Neural Networks. *Nature Machine Intelligence* **2022**, *4* (3), 279−287.
(207) 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* (6), 1692−1701.
(208) Moshkov, N.; Bornholdt, M.; Benoit, S.; Smith, M.; McQuin, C.; Goodman, A.; Senft, R. A.; Han, Y.; Babadi, M.; Horvath, P.; Cimini, B. A.; Carpenter, A. E.; Singh, S.; Caicedo, J. C. Learning Representations for Image-Based Profiling of Perturbations. *Nat. Commun.* **2024**, *15* (1), 1594.
(209) Liu, G.; Seal, S.; Arevalo, J.; Liang, Z.; Carpenter, A. E.; Jiang, M.; Singh, S. Learning Molecular Representation in a Cell. *ArXiv [cs.LG]* **2024**, DOI: 10.48550/arXiv.2406.12056.
(210) Sanchez-Fernandez, A.; Rumetshofer, E.; Hochreiter, S.; Klambauer, G. CLOOME: Contrastive Learning Unlocks Bioimaging Databases for Queries with Chemical Structures. *Nat. Commun.* **2023**, *14* (1), 7339.
(211) Wang, C.; Gupta, S.; Uhler, C.; Jaakkola, T. Removing Biases from Molecular Representations via Information Maximization. *arXiv [cs.LG]* **2023**, DOI: 10.48550/arXiv.2312.00718.
(212) Trunk, G. V. A Problem of Dimensionality: A Simple Example. *IEEE Trans. Pattern Anal. Mach. Intell.* **1979**, *PAMI-1* (3), 306−307.
(213) Munson, M. A.; Caruana, R. On Feature Selection, Bias-Variance, and Bagging. *Machine Learning and Knowledge Discovery in Databases* **2009**, *5782*, 144−159.
(214) Yu, L.; Liu, H. Efficient Feature Selection via Analysis of Relevance and Redundancy. *J. Mach. Learn. Res.* **2004**, *5*, 1205−1224.
(215) Tetko, I.; Baskin, I. I.; Varnek, A. . Unpublished 2008. *Tutorial on Machine Learning. Part 2. Descriptor Selection Bias***2008** DOI: 10.13140/RG.2.2.26353.43361.
(216) Tetko, I. V.; Sushko, I.; Pandey, A. K.; Zhu, H.; Tropsha, A.; Papa, E.; Oberg, T.; Todeschini, R.; Fourches, D.; Varnek, A. Critical Assessment of QSAR Models of Environmental Toxicity against Tetrahymena Pyriformis: Focusing on Applicability Domain and Overfitting by Variable Selection. *J. Chem. Inf. Model.* **2008**, *48* (9), 1733−1746.
(217) Pudjihartono, N.; Fadason, T.; Kempa-Liehr, A. W.; O'Sullivan, J. M. A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction. *Front. Bioinform.* **2022**, *2*, 927312.
(218) Radovic, M.; Ghalwash, M.; Filipovic, N.; Obradovic, Z. Minimum Redundancy Maximum Relevance Feature Selection Approach for Temporal Gene Expression Data. *BMC Bioinformatics* **2017**, *18* (1), 9.

AQ

(219) Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V. Gene Selection for Cancer Classification Using Support Vector Machines. *Mach. Learn.* **2002**, *46* (1), 389−422.
(220) Čmelo, I.; Voršilák, M.; Svozil, D. Profiling and Analysis of Chemical Compounds Using Pointwise Mutual Information. *J. Cheminform.* **2021**, *13* (1), 3.
(221) Kursa, M. B.; Rudnicki, W. R. Feature Selection with the Boruta Package. *J. Stat. Softw.* **2010**, *36*, 1−13.
(222) Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Series B Stat. Methodol.* **1996**, *58* (1), 267−288.
(223) Song, F.; Guo, Z.; Mei, D. Feature Selection Using Principal Component Analysis. In *2010 International Conference on System Science, Engineering Design and Manufacturing Informatization*; IEEE, 2010; Vol. *1*, pp 27−30.
(224) Jović, A.; Brkić, K.; Bogunović, N. A Review of Feature Selection Methods with Applications. In *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*; IEEE, 2015; pp 1200−1205.
(225) Kraskov, A.; Stögbauer, H.; Grassberger, P. Estimating Mutual Information. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **2004**, *69* (6 Pt 2), 066138.
(226) Zou, H.; Hastie, T. Regularization and Variable Selection Via the Elastic Net. *J. R. Stat. Soc. Series B Stat. Methodol.* **2005**, *67* (2), 301−320.
(227) Eklund, M.; Norinder, U.; Boyer, S.; Carlsson, L. Choosing Feature Selection and Learning Algorithms in QSAR. *J. Chem. Inf. Model.* **2014**, *54* (3), 837−843.
(228) Labjar, H.; Labjar, N.; Kissi, M. QSAR Anti-HIV Feature Selection and Prediction for Drug Discovery Using Genetic Algorithm and Machine Learning Algorithms. In *Computational Intelligence in Recent Communication Networks*; Ouaissa, M., Boulouard, Z., Ouaissa, M., Guermah, B., Eds.; Springer International Publishing: Cham, 2022; pp 191−204.
(229) Mitchell, J. B. O. Machine Learning Methods in Chemoinformatics. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2014**, *4* (5), 468−481.
(230) Lavecchia, A. Machine-Learning Approaches in Drug Discovery: Methods and Applications. *Drug Discovery Today* **2015**, *20* (3), 318−331.
(231) Hornik, K.; Stinchcombe, M.; White, H. Multilayer Feedforward Networks Are Universal Approximators. *Neural Netw.* **1989**, *2* (5), 359−366.
(232) Belkin, M.; Hsu, D.; Ma, S.; Mandal, S. Reconciling Modern Machine-Learning Practice and the Classical Bias-Variance Trade-Off. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116* (32), 15849−15854.
(233) Curth, A.; Jeffares, A.; van der Schaar, M. Why Do Random Forests Work? Understanding Tree Ensembles as Self-Regularizing Adaptive Smoothers. *arXiv [stat.ML]* **2024**, DOI: 10.48550/arXiv.2402.01502.
(234) Carracedo-Reboredo, P.; Liñares-Blanco, J.; Rodríguez-Fernández, N.; Cedrón, F.; Novoa, F. J.; Carballal, A.; Maojo, V.; Pazos, A.; Fernandez-Lozano, C. A Review on Machine Learning Approaches and Trends in Drug Discovery. *Comput. Struct. Biotechnol. J.* **2021**, *19*, 4538−4558.
(235) Nakkiran, P.; Kaplun, G.; Bansal, Y.; Yang, T.; Barak, B.; Sutskever, I. Deep Double Descent: Where Bigger Models and More Data Hurt. *J. Stat. Mech.* **2021**, *2021* (12), 124003.
(236) Curth, A.; Jeffares, A.; van der Schaar, M. A U-Turn on Double Descent: Rethinking Parameter Counting in Statistical Learning. *arXiv [stat.ML]* **2023**, DOI: 10.48550/arXiv.2310.18988.
(237) Sha, C. M.; Wang, J.; Dokholyan, N. V. NeuralDock: Rapid and Conformation-Agnostic Docking of Small Molecules. *Front Mol. Biosci* **2022**, *9*, 867241.
(238) Mansouri, K.; Grulke, C. M.; Judson, R. S.; Williams, A. J. OPERA Models for Predicting Physicochemical Properties and Environmental Fate Endpoints. *J. Cheminform.* **2018**, *10* (1), 10.
(239) Klambauer, G.; Clevert, D.-A.; Shah, I.; Benfenati, E.; Tetko, I. V. Introduction to the Special Issue: AI Meets Toxicology. *Chem. Res. Toxicol.* **2023**, *36* (8), 1163−1167.
(240) Eytcheson, S. A.; Tetko, I. V. Which Modern AI Methods Provide Accurate Predictions of Toxicological Endpoints? Analysis of Tox24 Challenge Results. *ChemRxiv* **2025**, DOI: 10.26434/chemrxiv-2025-7k7x3.
(241) Ekins, S. Progress in Computational Toxicology. *J. Pharmacol. Toxicol. Methods* **2014**, *69* (2), 115−140.
(242) Raies, A. B.; Bajic, V. B. In Silico Toxicology: Computational Methods for the Prediction of Chemical Toxicity. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2016**, *6* (2), 147−172.
(243) Baskin, I. I. Machine Learning Methods in Computational Toxicology. In *Computational Toxicology: Methods and Protocols*; Nicolotti, O., Ed.; Springer New York: New York, NY, 2018; pp 119−139.
(244) Yang, H.; Sun, L.; Li, W.; Liu, G.; Tang, Y. In Silico Prediction of Chemical Toxicity for Drug Design Using Machine Learning Methods and Structural Alerts. *Front Chem.* **2018**, *6*, 30.
(245) Vamathevan, J.; Clark, D.; Czodrowski, P.; Dunham, I.; Ferran, E.; Lee, G.; Li, B.; Madabhushi, A.; Shah, P.; Spitzer, M.; Zhao, S. Applications of Machine Learning in Drug Discovery and Development. *Nat. Rev. Drug Discovery* **2019**, *18* (6), 463−477.
(246) Ciallella, H. L.; Zhu, H. Advancing Computational Toxicology in the Big Data Era by Artificial Intelligence: Data-Driven and Mechanism-Driven Modeling for Chemical Toxicity. *Chem. Res. Toxicol.* **2019**, *32* (4), 536−547.
(247) Jiménez-Luna, J.; Grisoni, F.; Schneider, G. Drug Discovery with Explainable Artificial Intelligence. *Nature Machine Intelligence* **2020**, *2* (10), 573−584.
(248) Wang, M. W. H.; Goodman, J. M.; Allen, T. E. H. Machine Learning in Predictive Toxicology: Recent Applications and Future Directions for Classification Models. *Chem. Res. Toxicol.* **2021**, *34* (2), 217−239.
(249) Dara, S.; Dhamercherla, S.; Jadav, S. S.; Babu, C. M.; Ahsan, M. J. Machine Learning in Drug Discovery: A Review. *Artif Intell Rev.* **2022**, *55* (3), 1947−1999.
(250) Van Tran, T. T.; Surya Wibowo, A.; Tayara, H.; Chong, K. T. Artificial Intelligence in Drug Toxicity Prediction: Recent Advances, Challenges, and Future Perspectives. *J. Chem. Inf. Model.* **2023**, *63* (9), 2628−2643.
(251) Guo, W.; Liu, J.; Dong, F.; Song, M.; Li, Z.; Khan, M. K. H.; Patterson, T. A.; Hong, H. Review of Machine Learning and Deep Learning Models for Toxicity Prediction. *Exp. Biol. Med.* **2023**, *248* (21), 1952−1973.
(252) Tonoyan, L.; Siraki, A. G. Machine Learning in Toxicological Sciences: Opportunities for Assessing Drug Toxicity. *Front. Drug Discovery (Lausanne)* **2024**, *4*. DOI: 10.3389/fddsv.2024.1336025.
(253) Kovarich, S.; Ceriani, L.; Fuart Gatnik, M.; Bassan, A.; Pavan, M. Filling Data Gaps by Read-across: A Mini Review on Its Application, Developments and Challenges. *Mol. Inform.* **2019**, *38* (8−9), No. e1800121.
(254) Lester, C. C.; Yan, G. A Matched Molecular Pair (MMP) Approach for Selecting Analogs Suitable for Structure Activity Relationship (SAR)-Based Read across. *Regul. Toxicol. Pharmacol.* **2021**, *124*, 104966.
(255) Yang, Y.; Wu, Z.; Yao, X.; Kang, Y.; Hou, T.; Hsieh, C.-Y.; Liu, H. Exploring Low-Toxicity Chemical Space with Deep Learning for Molecular Generation. *J. Chem. Inf. Model.* **2022**, *62* (13), 3191−3199.
(256) Anstine, D. M.; Isayev, O. Generative Models as an Emerging Paradigm in the Chemical Sciences. *J. Am. Chem. Soc.* **2023**, *145* (16), 8736−8750.
(257) Wang, M.; Li, S.; Wang, J.; Zhang, O.; Du, H.; Jiang, D.; Wu, Z.; Deng, Y.; Kang, Y.; Pan, P.; Li, D.; Wang, X.; Yao, X.; Hou, T.; Hsieh, C.-Y. ClickGen: Directed Exploration of Synthesizable Chemical Space via Modular Reactions and Reinforcement Learning. *Nat. Commun.* **2024**, *15* (1), 10127.
(258) Thomas, M.; Bender, A.; de Graaf, C. Integrating Structure-Based Approaches in Generative Molecular Design. *Curr. Opin. Struct. Biol.* **2023**, *79*, 102559.
(259) Kingma, D. P.; Welling, M. Auto-Encoding Variational Bayes. *arXiv [stat.ML]* **2013**, DOI: 10.48550/arXiv.1312.6114.

(260) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent Sci.* **2018**, *4* (2), 268−276.
(261) Kusner, M. J.; Paige, B.; Hernández-Lobato, J. M. Grammar Variational Autoencoder. *arXiv [stat.ML]* **2017**, DOI: 10.48550/arXiv.1703.01925.
(262) Ciepliński, T.; Danel, T.; Podlewska, S.; Jastrzębski, S. Generative Models Should at Least Be Able to Design Molecules That Dock Well: A New Benchmark. *J. Chem. Inf. Model.* **2023**, *63* (11), 3238−3247.
(263) García-Ortegón, M.; Bender, A.; Rasmussen, C. E.; Kajino, H.; Bacallado, S. Combining Variational Autoencoder Representations with Structural Descriptors Improves Prediction of Docking Scores. *Machine Learning for Structural Biology Workshop, NeurIPS 2020, Vancouver, Canada*, 2020.
(264) Jin, W.; Barzilay, R.; Jaakkola, T. Junction Tree Variational Autoencoder for Molecular Graph Generation. *arXiv [cs.LG]* **2018**. , DOI: 10.48550/arXiv.1802.04364.
(265) Kang, S.; Cho, K. Conditional Molecular Design with Deep Generative Models. *J. Chem. Inf. Model.* **2019**, *59* (1), 43−52.
(266) Chow, Y. L.; Singh, S.; Carpenter, A. E.; Way, G. P. Predicting Drug Polypharmacology from Cell Morphology Readouts Using Variational Autoencoder Latent Space Arithmetic. *PLoS Comput. Biol.* **2022**, *18* (2), No. e1009888.
(267) Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Networks. *Commun. ACM* **2014**, *63*, 139.
(268) Macedo, B.; Ribeiro Vaz, I.; Taveira Gomes, T. MedGAN: Optimized Generative Adversarial Network with Graph Convolutional Networks for Novel Molecule Design. *Sci. Rep.* **2024**, *14* (1), 1212.
(269) Richard, A. M.; Huang, R.; Waidyanatha, S.; Shinn, P.; Collins, B. J.; Thillainadarajah, I.; Grulke, C. M.; Williams, A. J.; Lougee, R. R.; Judson, R. S.; Houck, K. A.; Shobair, M.; Yang, C.; Rathman, J. F.; Yasgar, A.; Fitzpatrick, S. C.; Simeonov, A.; Thomas, R. S.; Crofton, K. M.; Paules, R. S.; Bucher, J. R.; Austin, C. P.; Kavlock, R. J.; Tice, R. R. The Tox21 10K Compound Library: Collaborative Chemistry Advancing Toxicology. *Chem. Res. Toxicol.* **2021**, *34* (2), 189−216.
(270) Chen, X.; Roberts, R.; Tong, W.; Liu, Z. Tox-GAN: An Artificial Intelligence Approach Alternative to Animal Studies-A Case Study With Toxicogenomics. *Toxicol. Sci.* **2022**, *186* (2), 242−259.
(271) Li, T.; Roberts, R.; Liu, Z.; Tong, W. TransOrGAN: An Artificial Intelligence Mapping of Rat Transcriptomic Profiles between Organs, Ages, and Sexes. *Chem. Res. Toxicol.* **2023**, *36* (6), 916−925.
(272) Urbina, F.; Lentzos, F.; Invernizzi, C.; Ekins, S. Dual Use of Artificial Intelligence-Powered Drug Discovery. *Nat. Mach Intell* **2022**, *4* (3), 189−191.
(273) Alakhdar, A.; Poczos, B.; Washburn, N. Diffusion Models in DE Novo Drug Design. *J. Chem. Inf. Model.* **2024**, *64* (19), 7238−7256.
(274) Sohl-Dickstein, J.; Weiss, E. A.; Maheswaranathan, N.; Ganguli, S. Deep Unsupervised Learning Using Nonequilibrium Thermodynamics. *arXiv [cs.LG]* **2015**, DOI: 10.48550/arXiv.1503.03585.
(275) Song, Y.; Ermon, S. Generative Modeling by Estimating Gradients of the Data Distribution. *arXiv [cs.LG]* **2019**, DOI: 10.48550/arXiv.1907.05600.
(276) Vignac, C.; Krawczuk, I.; Siraudin, A.; Wang, B.; Cevher, V.; Frossard, P. DiGress: Discrete Denoising Diffusion for Graph Generation. *arXiv [cs.LG]* **2022**, DOI: 10.48550/arXiv.2209.14734.
(277) Jing, B.; Corso, G.; Chang, J.; Barzilay, R.; Jaakkola, T. Torsional Diffusion for Molecular Conformer Generation. *arXiv [physics.chem-ph]* **2022**, DOI: 10.48550/arXiv.2206.01729.
(278) Corso, G.; Stärk, H.; Jing, B.; Barzilay, R.; Jaakkola, T. DiffDock: Diffusion Steps, Twists, and Turns for Molecular Docking. *arXiv [q-bio.BM]* **2022**, DOI: 10.48550/arXiv.2210.01776.
(279) Hoogeboom, E.; Satorras, V. G.; Vignac, C.; Welling, M. Equivariant Diffusion for Molecule Generation in 3D. *arXiv [cs.LG]* **2022**, DOI: 10.48550/arXiv.2203.17003.
(280) Huang, L.; Xu, T.; Yu, Y.; Zhao, P.; Chen, X.; Han, J.; Xie, Z.; Li, H.; Zhong, W.; Wong, K.-C.; Zhang, H. A Dual Diffusion Model Enables 3D Molecule Generation and Lead Optimization Based on Target Pockets. *Nat. Commun.* **2024**, *15* (1), 2657.
(281) Radford, A.; Narasimhan, K. *Improving Language Understanding by Generative Pre-Training*; OpenAI, 2018.
(282) OpenAI; Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; Avila, R.; Babuschkin, I.; Balaji, S.; Balcom, V.; Baltescu, P.; Bao, H.; Bavarian, M.; Belgum, J.; Bello, I.; Berdine, J.; Bernadett-Shapiro, G.; Berner, C.; Bogdonoff, L.; Boiko, O.; Boyd, M.; Brakman, A.-L.; Brockman, G.; Brooks, T.; Brundage, M.; Button, K.; Cai, T.; Campbell, R.; Cann, A.; Carey, B.; Carlson, C.; Carmichael, R.; Chan, B.; Chang, C.; Chantzis, F.; Chen, D.; Chen, S.; Chen, R.; Chen, J.; Chen, M.; Chess, B.; Cho, C.; Chu, C.; Chung, H. W.; Cummings, D.; Currier, J.; Dai, Y.; Decareaux, C.; Degry, T.; Deutsch, N.; Deville, D.; Dhar, A.; Dohan, D.; Dowling, S.; Dunning, S.; Ecoffet, A.; Eleti, A.; Eloundou, T.; Farhi, D.; Fedus, L.; Felix, N.; Fishman, S. P.; Forte, J.; Fulford, I.; Gao, L.; Georges, E.; Gibson, C.; Goel, V.; Gogineni, T.; Goh, G.; Gontijo-Lopes, R.; Gordon, J.; Grafstein, M.; Gray, S.; Greene, R.; Gross, J.; Gu, S. S.; Guo, Y.; Hallacy, C.; Han, J.; Harris, J.; He, Y.; Heaton, M.; Heidecke, J.; Hesse, C.; Hickey, A.; Hickey, W.; Hoeschele, P.; Houghton, B.; Hsu, K.; Hu, S.; Hu, X.; Huizinga, J.; Jain, S.; Jain, S.; Jang, J.; Jiang, A.; Jiang, R.; Jin, H.; Jin, D.; Jomoto, S.; Jonn, B.; Jun, H.; Kaftan, T.; Kaiser, Ł.; Kamali, A.; Kanitscheider, I.; Keskar, N. S.; Khan, T.; Kilpatrick, L.; Kim, J. W.; Kim, C.; Kim, Y.; Kirchner, J. H.; Kiros, J.; Knight, M.; Kokotajlo, D.; Kondraciuk, Ł.; Kondrich, A.; Konstantinidis, A.; Kosic, K.; Krueger, G.; Kuo, V.; Lampe, M.; Lan, I.; Lee, T.; Leike, J.; Leung, J.; Levy, D.; Li, C. M.; Lim, R.; Lin, M.; Lin, S.; Litwin, M.; Lopez, T.; Lowe, R.; Lue, P.; Makanju, A.; Malfacini, K.; Manning, S.; Markov, T.; Markovski, Y.; Martin, B.; Mayer, K.; Mayne, A.; McGrew, B.; McKinney, S. M.; McLeavey, C.; McMillan, P.; McNeil, J.; Medina, D.; Mehta, A.; Menick, J.; Metz, L.; Mishchenko, A.; Mishkin, P.; Monaco, V.; Morikawa, E.; Mossing, D.; Mu, T.; Murati, M.; Murk, O.; Mély, D.; Nair, A.; Nakano, R.; Nayak, R.; Neelakantan, A.; Ngo, R.; Noh, H.; Ouyang, L.; O'Keefe, C.; Pachocki, J.; Paino, A.; Palermo, J.; Pantuliano, A.; Parascandolo, G.; Parish, J.; Parparita, E.; Passos, A.; Pavlov, M.; Peng, A.; Perelman, A.; de Avila Belbute Peres, F.; Petrov, M.; de Oliveira Pinto, H. P.; Michael; Pokorny; Pokrass, M.; Pong, V. H.; Powell, T.; Power, A.; Power, B.; Proehl, E.; Puri, R.; Radford, A.; Rae, J.; Ramesh, A.; Raymond, C.; Real, F.; Rimbach, K.; Ross, C.; Rotsted, B.; Roussez, H.; Ryder, N.; Saltarelli, M.; Sanders, T.; Santurkar, S.; Sastry, G.; Schmidt, H.; Schnurr, D.; Schulman, J.; Selsam, D.; Sheppard, K.; Sherbakov, T.; Shieh, J.; Shoker, S.; Shyam, P.; Sidor, S.; Sigler, E.; Simens, M.; Sitkin, J.; Slama, K.; Sohl, I.; Sokolowsky, B.; Song, Y.; Staudacher, N.; Such, F. P.; Summers, N.; Sutskever, I.; Tang, J.; Tezak, N.; Thompson, M. B.; Tillet, P.; Tootoonchian, A.; Tseng, E.; Tuggle, P.; Turley, N.; Tworek, J.; Uribe, J. F. C.; Vallone, A.; Vijayvergiya, A.; Voss, C.; Wainwright, C.; Wang, J. J.; Wang, A.; Wang, B.; Ward, J.; Wei, J.; Weinmann, C. J.; Welihinda, A.; Welinder, P.; Weng, J.; Weng, L.; Wiethoff, M.; Willner, D.; Winter, C.; Wolrich, S.; Wong, H.; Workman, L.; Wu, S.; Wu, J.; Wu, M.; Xiao, K.; Xu, T.; Yoo, S.; Yu, K.; Yuan, Q.; Zaremba, W.; Zellers, R.; Zhang, C.; Zhang, M.; Zhao, S.; Zheng, T.; Zhuang, J.; Zhuk, W.; Zoph, B. GPT-4 Technical Report. *arXiv [cs.CL]* 2023, DOI: 10.48550/arXiv.2303.08774.
(283) Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; Bikel, D.; Blecher, L.; Ferrer, C. C.; Chen, M.; Cucurull, G.; Esiobu, D.; Fernandes, J.; Fu, J.; Fu, W.; Fuller, B.; Gao, C.; Goswami, V.; Goyal, N.; Hartshorn, A.; Hosseini, S.; Hou, R.; Inan, H.; Kardas, M.; Kerkez, V.; Khabsa, M.; Kloumann, I.; Korenev, A.; Koura, P. S.; Lachaux, M.-A.; Lavril, T.; Lee, J.; Liskovich, D.; Lu, Y.; Mao, Y.; Martinet, X.; Mihaylov, T.; Mishra, P.; Molybog, I.; Nie, Y.; Poulton, A.; Reizenstein, J.; Rungta, R.; Saladi, K.; Schelten, A.; Silva, R.; Smith, E. M.; Subramanian, R.; Tan, X. E.; Tang, B.; Taylor, R.; Williams, A.; Kuan, J. X.; Xu, P.; Yan, Z.; Zarov, I.; Zhang, Y.; Fan, A.; Kambadur, M.; Narang, S.; Rodriguez, A.; Stojnic, R.; Edunov, S.; Scialom, T. Llama 2: Open Foundation and

Fine-Tuned Chat Models. *arXiv [cs.CL]* **2023**, DOI: 10.48550/arXiv.2307.09288.
(284) Ross, J.; Belgodere, B.; Chenthamarakshan, V.; Padhi, I.; Mroueh, Y.; Das, P. Large-Scale Chemical Language Representations Capture Molecular Structure and Properties. *arXiv [cs.LG]* **2021**, DOI: 10.48550/arXiv.2106.09553.
(285) Loeffler, H. H.; He, J.; Tibo, A.; Janet, J. P.; Voronov, A.; Mervin, L. H.; Engkvist, O. Reinvent 4: Modern AI-Driven Generative Molecule Design. *J. Cheminform.* **2024**, *16* (1), 20.
(286) Tetko, I. V. Associative Neural Network. *Neural Process. Lett.* **2002**, *16* (2), 187−199.
(287) Tetko, I. V.; Poda, G. I. Application of ALOGPS 2.1 to Predict Log D Distribution Coefficient for Pfizer Proprietary Compounds. *J. Med. Chem.* **2004**, *47* (23), 5601−5604.
(288) Tetko, I. V.; Bruneau, P. Application of ALOGPS to Predict 1-Octanol/water Distribution Coefficients, logP, and logD, of AstraZeneca in-House Database. *J. Pharm. Sci.* **2004**, *93* (12), 3103−3110.
(289) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv [cs.CL]* **2017**, DOI: 10.48550/arXiv.1706.03762.
(290) He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* **2015**, 770−778.
(291) Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; de Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; Gelly, S. Parameter-Efficient Transfer Learning for NLP. *arXiv [cs.LG]*, **2019**, DOI: 10.48550/arXiv.1902.00751.
(292) Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv [cs.CL]* **2021**, DOI: 10.48550/arXiv.2106.09685.
(293) Thomas, M.; Ahmad, M.; Tresadern, G.; de Fabritiis, G. PromptSMILES: Prompting for Scaffold Decoration and Fragment Linking in Chemical Language Models. *J. Cheminform.* **2024**, *16* (1), 77.
(294) Grygorenko, O. O. Enamine Ltd.: The Science and Business of Organic Chemistry and beyond. *Eur. J. Org. Chem.* **2021**, *2021* (47), 6474−6477.
(295) Bran, A. M.; Cox, S.; Schilter, O.; Baldassari, C.; White, A. D.; Schwaller, P. Augmenting Large Language Models with Chemistry Tools. *Nat. Mach Intell* **2024**, *6* (5), 525−535.
(296) White, A. D.; Hocky, G. M.; Gandhi, H. A.; Ansari, M.; Cox, S.; Wellawatte, G. P.; Sasmal, S.; Yang, Z.; Liu, K.; Singh, Y.; Peña Ccoa, W. J. Assessment of Chemistry Knowledge in Large Language Models That Generate Code. *Digit Discov* **2023**, *2* (2), 368−376.
(297) Sadeghi, S.; Bui, A.; Forooghi, A.; Lu, J.; Ngom, A. Can Large Language Models Understand Molecules? *BMC Bioinformatics* **2024**, *25* (1), 225.
(298) Thomas, M.; Boardman, A.; Garcia-Ortegon, M.; Yang, H.; de Graaf, C.; Bender, A. Applications of Artificial Intelligence in Drug Design: Opportunities and Challenges. In *Artificial Intelligence in Drug Design*; Heifetz, A., Ed.; Springer US: New York, NY, 2022; pp 1−59.
(299) Orsi, M.; Reymond, J.-L. Navigating a 1E+60 Chemical Space. *ChemRxiv* **2024**, DOI: 10.26434/chemrxiv-2024-bqd8c.
(300) Kearnes, S. Pursuing a Prospective Perspective. *Trends in Chemistry* **2021**, *3*, 77.
(301) Hanser, T.; Barber, C.; Marchaland, J. F.; Werner, S. Applicability Domain: Towards a More Formal Definition. *SAR QSAR Environ. Res.* **2016**, *27* (11), 865−881.
(302) Arvidsson McShane, S.; Norinder, U.; Alvarsson, J.; Ahlberg, E.; Carlsson, L.; Spjuth, O. CPSign: Conformal Prediction for Cheminformatics Modeling. *J. Cheminform.* **2024**, *16* (1), 75.
(303) Lampa, S.; Alvarsson, J.; Arvidsson Mc Shane, S.; Berg, A.; Ahlberg, E.; Spjuth, O. Predicting Off-Target Binding Profiles With Confidence Using Conformal Prediction. *Front. Pharmacol.* **2018**, *9*. DOI: 10.3389/fphar.2018.01256.
(304) McHugh, M. L. Interrater Reliability: The Kappa Statistic. *Biochem. Med.* **2012**, *22* (3), 276−282.
(305) Vincent, F.; Loria, P. M.; Weston, A. D.; Steppan, C. M.; Doyonnas, R.; Wang, Y.-M.; Rockwell, K. L.; Peakman, M.-C. Hit Triage and Validation in Phenotypic Screening: Considerations and Strategies. *Cell Chem. Biol.* **2020**, *27* (11), 1332−1346.
(306) Boldini, D.; Friedrich, L.; Kuhn, D.; Sieber, S. A. Machine Learning Assisted Hit Prioritization for High Throughput Screening in Drug Discovery. *ACS Cent Sci.* **2024**, *10* (4), 823−832.
(307) Truchon, J.-F.; Bayly, C. I. Evaluating Virtual Screening Methods: Good and Bad Metrics for the "Early Recognition" Problem. *J. Chem. Inf. Model.* **2007**, *47* (2), 488−508.
(308) Wellnitz, J.; Jain, S.; Hochuli, J.; Maxfield, T.; Muratov, E.; Tropsha, A.; Zakharov, A. One Size Does Not Fit All: Revising Traditional Paradigms for QSAR-Based Virtual Screenings. *J. Cheminform.* **2025**, *17* (1), 7 DOI: 10.1186/s13321-025-00948-y.
(309) Király, P.; Kiss, R.; Kovács, D.; Ballaj, A.; Tóth, G. The Relevance of Goodness-of-Fit, Robustness and Prediction Validation Categories of OECD-QSAR Principles with Respect to Sample Size and Model Type. *Mol. Inform.* **2022**, *41* (11), No. e2200072.
(310) Bender, A.; Glen, R. C. A Discussion of Measures of Enrichment in Virtual Screening: Comparing the Information Content of Descriptors with Increasing Levels of Sophistication. *J. Chem. Inf. Model.* **2005**, *45* (5), 1369−1375.
(311) Ben-David, A. About the Relationship between ROC Curves and Cohen's Kappa. *Eng. Appl. Artif. Intell.* **2008**, *21* (6), 874−882.
(312) Esposito, C.; Landrum, G. A.; Schneider, N.; Stiefl, N.; Riniker, S. GHOST: Adjusting the Decision Threshold to Handle Imbalanced Data in Machine Learning. *J. Chem. Inf. Model.* **2021**, *61* (6), 2623−2640.
(313) Hunklinger, A.; Hartog, P.; Šícho, M.; Godin, G.; Tetko, I. V. The openOCHEM Consensus Model Is the Best-Performing Open-Source Predictive Model in the First EUOS/SLAS Joint Compound Solubility Challenge. *SLAS Discovery* **2024**, *29* (2), 100144.
(314) Roy, K.; Chakraborty, P.; Mitra, I.; Ojha, P. K.; Kar, S.; Das, R. N. Some Case Studies on Application of "r(m)2" Metrics for Judging Quality of Quantitative Structure-Activity Relationship Predictions: Emphasis on Scaling of Response Data. *J. Comput. Chem.* **2013**, *34* (12), 1071−1082.
(315) Shayanfar, A.; Shayanfar, S. Is Regression through Origin Useful in External Validation of QSAR Models? *Eur. J. Pharm. Sci.* **2014**, *59*, 31−35.
(316) Alexander, D. L. J.; Tropsha, A.; Winkler, D. A. Beware of R(2): Simple, Unambiguous Assessment of the Prediction Accuracy of QSAR and QSPR Models. *J. Chem. Inf. Model.* **2015**, *55* (7), 1316−1322.
(317) Schwaighofer, A.; Schroeter, T.; Mika, S.; Hansen, K.; Ter Laak, A.; Lienau, P.; Reichel, A.; Heinrich, N.; Müller, K.-R. A Probabilistic Approach to Classifying Metabolic Stability. *J. Chem. Inf. Model.* **2008**, *48* (4), 785−796.
(318) Abduljalil, K.; Cain, T.; Humphries, H.; Rostami-Hodjegan, A. Deciding on Success Criteria for Predictability of Pharmacokinetic Parameters from in Vitro Studies: An Analysis Based on in Vivo Observations. *Drug Metab. Dispos.* **2014**, *42* (9), 1478−1484.
(319) Turner, R. M.; Park, B. K.; Pirmohamed, M. Parsing Interindividual Drug Variability: An Emerging Role for Systems Pharmacology. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **2015**, *7* (4), 221−241.
(320) Thummel, K. E.; Lin, Y. S. Sources of Interindividual Variability. *Methods Mol. Biol.* **2014**, *1113*, 363−415.
(321) Brown, S. P.; Muchmore, S. W.; Hajduk, P. J. Healthy Skepticism: Assessing Realistic Model Performance. *Drug Discovery Today* **2009**, *14* (7−8), 420−427.
(322) Crusius, D.; Cipcigan, F.; Biggin, P. C. Are We Fitting Data or Noise? Analysing the Predictive Power of Commonly Used Datasets in Drug-, Materials-, and Molecular-Discovery. *Faraday Discuss.* **2025**, *256*, 304.
(323) Tetko, I. V.; van Deursen, R.; Godin, G. Be Aware of Overfitting by Hyperparameter Optimization! *J. Cheminform.* **2024**, *16* (1), 139.
(324) Varma, S.; Simon, R. Bias in Error Estimation When Using Cross-Validation for Model Selection. *BMC Bioinformatics* **2006**, *7* (1), 91.
(325) Walters, P. Some Thoughts on Splitting Chemical Datasets. *Practical Cheminformatics*. https://practicalcheminformatics.blogspot.


https://doi.org/10.1021/acs.chemrestox.5c00033
*Chem. Res. Toxicol.* XXXX, XXX, XXX−XXX

com/2024/11/some-thoughts-on-splitting-chemical.html (accessed 2025-03-23).
(326) Guo, Q.; Hernandez-Hernandez, S.; Ballester, P. J. Scaffold Splits Overestimate Virtual Screening Performance. *arXiv [q-bio.QM]* **2024**, DOI: 10.48550/arXiv.2406.00873.
(327) Saha, U. S.; Vendruscolo, M.; Carpenter, A. E.; Singh, S.; Bender, A.; Seal, S. Step Forward Cross Validation for Bioactivity Prediction: Out of Distribution Validation in Drug Discovery. *bioRxiv* **2024**, DOI: 10.1101/2024.07.02.601740.
(328) Landrum, G. A.; Beckers, M.; Lanini, J.; Schneider, N.; Stiefl, N.; Riniker, S. SIMPD: An Algorithm for Generating Simulated Time Splits for Validating Machine Learning Approaches. *J. Cheminform.* **2023**, *15* (1), 119.
(329) Xiong, Z.; Cui, Y.; Liu, Z.; Zhao, Y.; Hu, M.; Hu, J. Evaluating Explorative Prediction Power of Machine Learning Algorithms for Materials Discovery Using K-Fold Forward Cross-Validation. *Comput. Mater. Sci.* **2020**, *171*, 109203.
(330) Kaneko, H. Evaluation and Optimization Methods for Applicability Domain Methods and Their Hyperparameters, Considering the Prediction Performance of Machine Learning Models. *ACS Omega* **2024**, *9* (10), 11453−11458.
(331) Fechner, N.; Jahn, A.; Hinselmann, G.; Zell, A. Estimation of the Applicability Domain of Kernel-Based Machine Learning Models for Virtual Screening. *J. Cheminform.* **2010**, *2* (1), 2.
(332) Preparata, F. P.; Shamos, M. I. *Computational Geometry: An Introduction*; Springer Science & Business Media, 2012.
(333) Sushko, I.; Novotarskyi, S.; Körner, R.; Pandey, A. K.; Cherkasov, A.; Li, J.; Gramatica, P.; Hansen, K.; Schroeter, T.; Müller, K.-R.; Xi, L.; Liu, H.; Yao, X.; Öberg, T.; Hormozdiari, F.; Dao, P.; Sahinalp, C.; Todeschini, R.; Polishchuk, P.; Artemenko, A.; Kuz'min, V.; Martin, T. M.; Young, D. M.; Fourches, D.; Muratov, E.; Tropsha, A.; Baskin, I.; Horvath, D.; Marcou, G.; Muller, C.; Varnek, A.; Prokopenko, V. V.; Tetko, I. V. Applicability Domains for Classification Problems: Benchmarking of Distance to Models for Ames Mutagenicity Set. *J. Chem. Inf. Model.* **2010**, *50* (12), 2094−2111.
(334) Baskin, I. I.; Kireeva, N.; Varnek, A. The One-Class Classification Approach to Data Description and to Models Applicability Domain. *Mol. Inform.* **2010**, *29* (8−9), 581−587.
(335) Zhang, Y.; Menke, J.; He, J.; Nittinger, E.; Tyrchan, C.; Koch, O.; Zhao, H. Similarity-Based Pairing Improves Efficiency of Siamese Neural Networks for Regression Tasks and Uncertainty Quantification. *J. Cheminform.* **2023**, *15* (1), 75.
(336) Hirschfeld, L.; Swanson, K.; Yang, K.; Barzilay, R.; Coley, C. W. Uncertainty Quantification Using Neural Networks for Molecular Property Prediction. *J. Chem. Inf. Model.* **2020**, *60* (8), 3770−3780.
(337) Aniceto, N.; Freitas, A. A.; Bender, A.; Ghafourian, T. A Novel Applicability Domain Technique for Mapping Predictive Reliability across the Chemical Space of a QSAR: Reliability-Density Neighbourhood. *J. Cheminform.* **2016**, *8* (1), 1−20.
(338) D'Amour, A.; Heller, K.; Moldovan, D.; Adlam, B.; Alipanahi, B.; Beutel, A.; Chen, C.; Deaton, J.; Eisenstein, J.; Hoffman, M. D.; Hormozdiari, F.; Houlsby, N.; Hou, S.; Jerfel, G.; Karthikesalingam, A.; Lucic, M.; Ma, Y.; McLean, C.; Mincu, D.; Mitani, A.; Montanari, A.; Nado, Z.; Natarajan, V.; Nielson, C.; Osborne, T. F.; Raman, R.; Ramasamy, K.; Sayres, R.; Schrouff, J.; Seneviratne, M.; Sequeira, S.; Suresh, H.; Veitch, V.; Vladymyrov, M.; Wang, X.; Webster, K.; Yadlowsky, S.; Yun, T.; Zhai, X.; Sculley, D. Underspecification Presents Challenges for Credibility in Modern Machine Learning. *arXiv [cs.LG]* **2020**, DOI: 10.48550/arXiv.2011.03395.
(339) Rotundo, L.; Pyrsopoulos, N. Liver Injury Induced by Paracetamol and Challenges Associated with Intentional and Unintentional Use. *World J. Hepatol.* **2020**, *12* (4), 125−136.
(340) Muller, P. Y.; Milton, M. N. The Determination and Interpretation of the Therapeutic Index in Drug Development. *Nat. Rev. Drug Discovery* **2012**, *11* (10), 751−761.
(341) Schneider, P.; Walters, W. P.; Plowright, A. T.; Sieroka, N.; Listgarten, J.; Goodnow, R. A., Jr; Fisher, J.; Jansen, J. M.; Duca, J. S.; Rush, T. S.; Zentgraf, M.; Hill, J. E.; Krutoholow, E.; Kohler, M.; Blaney, J.; Funatsu, K.; Luebkemann, C.; Schneider, G. Rethinking Drug Design in the Artificial Intelligence Era. *Nat. Rev. Drug Discovery* **2020**, *19* (5), 353−364.
(342) Horne, R. I.; Wilson-Godber, J.; Díaz, A. G.; Brotzakis, Z. F.; Seal, S.; Gregory, R. C.; Possenti, A.; Chia, S.; Vendruscolo, M. Using Generative Modeling to Endow with Potency Initially Inert Compounds with Good Bioavailability and Low Toxicity. *J. Chem. Inf. Model.* **2023**, *64* (3), 590−596.
(343) Amberg, A.; Beilke, L.; Bercu, J.; Bower, D.; Brigo, A.; Cross, K. P.; Custer, L.; Dobo, K.; Dowdy, E.; Ford, K. A.; Glowienke, S.; Van Gompel, J.; Harvey, J.; Hasselgren, C.; Honma, M.; Jolly, R.; Kemper, R.; Kenyon, M.; Kruhlak, N.; Leavitt, P.; Miller, S.; Muster, W.; Nicolette, J.; Plaper, A.; Powley, M.; Quigley, D. P.; Reddy, M. V.; Spirkl, H.-P.; Stavitskaya, L.; Teasdale, A.; Weiner, S.; Welch, D. S.; White, A.; Wichard, J.; Myatt, G. J. Principles and Procedures for Implementation of ICH M7 Recommended (Q)SAR Analyses. *Regul. Toxicol. Pharmacol.* **2016**, *77*, 13−24.
(344) Bassan, A.; Alves, V. M.; Amberg, A.; Anger, L. T.; Auerbach, S.; Beilke, L.; Bender, A.; Cronin, M. T. D.; Cross, K. P.; Hsieh, J.-H.; Greene, N.; Kemper, R.; Kim, M. T.; Mumtaz, M.; Noeske, T.; Pavan, M.; Pletz, J.; Russo, D. P.; Sabnis, Y.; Schaefer, M.; Szabo, D. T.; Valentin, J.-P.; Wichard, J.; Williams, D.; Woolley, D.; Zwickl, C.; Myatt, G. J. In Silico Approaches in Organ Toxicity Hazard Assessment: Current Status and Future Needs in Predicting Liver Toxicity. *Comput. Toxicol* **2021**, *20*, 100187.
(345) Hornberg, J. J.; Laursen, M.; Brenden, N.; Persson, M.; Thougaard, A. V.; Toft, D. B.; Mow, T. Exploratory Toxicology as an Integrated Part of Drug Discovery. Part I: Why and How. *Drug Discovery Today* **2014**, *19* (8), 1131−1136.
(346) Lumley, J. A.; Fallon, D.; Whatling, R.; Coupry, D.; Brown, A. vEXP: A Virtual Enhanced Cross Screen Panel for off-Target Pharmacology Alerts. *Computational Toxicology* **2024**, *31*, 100324.
(347) Baltazar, M. T.; Cable, S.; Carmichael, P. L.; Cubberley, R.; Cull, T.; Delagrange, M.; Dent, M. P.; Hatherell, S.; Houghton, J.; Kukic, P.; Li, H.; Lee, M.-Y.; Malcomber, S.; Middleton, A. M.; Moxon, T. E.; Nathanail, A. V.; Nicol, B.; Pendlington, R.; Reynolds, G.; Reynolds, J.; White, A.; Westmoreland, C. A next-Generation Risk Assessment Case Study for Coumarin in Cosmetic Products. *Toxicol. Sci.* **2020**, *176* (1), 236−252.
(348) Ouedraogo, G.; Alexander-White, C.; Bury, D.; Clewell, H. J., 3rd; Cronin, M.; Cull, T.; Dent, M.; Desprez, B.; Detroyer, A.; Ellison, C.; Giammanco, S.; Hack, E.; Hewitt, N. J.; Kenna, G.; Klaric, M.; Kreiling, R.; Lester, C.; Mahony, C.; Mombelli, E.; Naciff, J.; O'Brien, J.; Schepky, A.; Tozer, S.; van der Burg, B.; van Vugt-Lussenburg, B.; Stuard, S. Cosmetics Europe. Read-across and New Approach Methodologies Applied in a 10-Step Framework for Cosmetics Safety Assessment - A Case Study with Parabens. *Regul. Toxicol. Pharmacol.* **2022**, *132*, 105161.
(349) Mervin, L.; Voronov, A.; Kabeshov, M.; Engkvist, O. QSARtuna: An Automated QSAR Modeling Platform for Molecular Property Prediction in Drug Design. *J. Chem. Inf. Model.* **2024**, *64*, 5365.
(350) Fu, L.; Shi, S.; Yi, J.; Wang, N.; He, Y.; Wu, Z.; Peng, J.; Deng, Y.; Wang, W.; Wu, C.; Lyu, A.; Zeng, X.; Zhao, W.; Hou, T.; Cao, D. ADMETlab 3.0: An Updated Comprehensive Online ADMET Prediction Platform Enhanced with Broader Coverage, Improved Performance, API Functionality and Decision Support. *Nucleic Acids Res.* **2024**, *52* (W1), W422−W431.
(351) Ruusmann, V.; Sild, S.; Maran, U. QSAR DataBank - an Approach for the Digital Organization and Archiving of QSAR Model Information. *J. Cheminform.* **2014**, *6* (1), 25.
(352) Sushko, I.; Novotarskyi, S.; Körner, R.; Pandey, A. K.; Rupp, M.; Teetz, W.; Brandmaier, S.; Abdelaziz, A.; Prokopenko, V. V.; Tanchuk, V. Y.; Todeschini, R.; Varnek, A.; Marcou, G.; Ertl, P.; Potemkin, V.; Grishina, M.; Gasteiger, J.; Schwab, C.; Baskin, I. I.; Palyulin, V. A.; Radchenko, E. V.; Welsh, W. J.; Kholodovych, V.; Chekmarev, D.; Cherkasov, A.; Aires-de-Sousa, J.; Zhang, Q.-Y.; Bender, A.; Nigsch, F.; Patiny, L.; Williams, A.; Tkachenko, V.; Tetko, I. V. Online Chemical Modeling Environment (OCHEM): Web Platform for Data Storage, Model Development and Publishing of Chemical Information. *J. Comput. Aided Mol. Des.* **2011**, *25* (6), 533−554.


https://doi.org/10.1021/acs.chemrestox.5c00033
*Chem. Res. Toxicol.* XXXX, XXX, XXX−XXX

(353) Nikolov, N. G.; Wedebye, E. B. The Danish (Q) SAR Database: Recent Updates. In *SOT Annual Meeting and ToxExpo 2022*; Society of Toxicology, 2022; pp 437−437.
(354) Seal, S.; Williams, D. P.; Hosseini-Gerami, L.; Spjuth, O.; Bender, A. Improved Early Detection of Drug-Induced Liver Injury by Integrating Predicted in Vivo and in Vitro Data. *bioArxiv* **2024**, DOI: 10.1101/2024.01.10.575128.
(355) Card, M. L.; Gomez-Alvarez, V.; Lee, W.-H.; Lynch, D. G.; Orentas, N. S.; Lee, M. T.; Wong, E. M.; Boethling, R. S. History of EPI Suite™ and Future Perspectives on Chemical Property Estimation in US Toxic Substances Control Act New Chemical Risk Assessments. *Environ. Sci. Process. Impacts* **2017**, *19* (3), 203−212.
(356) Ji, C.; Svensson, F.; Zoufir, A.; Bender, A. eMolTox: Prediction of Molecular Toxicity with Confidence. *Bioinformatics* **2018**, *34* (14), 2508−2509.
(357) Schultz, T. W.; Diderich, R.; Kuseva, C. D.; Mekenyan, O. G. The OECD QSAR Toolbox Starts Its Second Decade. *Methods Mol. Biol.* **2018**, *1800*, 55−77.
(358) Pires, D. E. V.; Blundell, T. L.; Ascher, D. B. pkCSM: Predicting Small-Molecule Pharmacokinetic and Toxicity Properties Using Graph-Based Signatures. *J. Med. Chem.* **2015**, *58* (9), 4066−4072.
(359) Banerjee, P.; Kemmler, E.; Dunkel, M.; Preissner, R. ProTox 3.0: A Webserver for the Prediction of Toxicity of Chemicals. *Nucleic Acids Res.* **2024**, *52* (W1), W513−W520.
(360) 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.
(361) Patlewicz, G.; Jeliazkova, N.; Safford, R. J.; Worth, A. P.; Aleksiev, B. An Evaluation of the Implementation of the Cramer Classification Scheme in the Toxtree Software. *SAR QSAR Environ. Res.* **2008**, *19* (5−6), 495−524.
(362) *VEGA HUB - Virtual models for property Evaluation of chemicals within a Global Architecture*. https://www.vegahub.eu/ (accessed 2024-07-12).
(363) Di Stefano, M.; Galati, S.; Piazza, L.; Granchi, C.; Mancini, S.; Fratini, F.; Macchia, M.; Poli, G.; Tuccinardi, T. VenomPred 2.0: A Novel In Silico Platform for an Extended and Human Interpretable Toxicological Profiling of Small Molecules. *J. Chem. Inf. Model.* **2024**, *64* (7), 2275−2289.
(364) Stankovic, B.; Marinkovic, F. A Novel Procedure for Selection of Molecular Descriptors: QSAR Model for Mutagenicity of Nitro-aromatic Compounds. *Environ. Sci. Pollut. Res. Int.* **2024**, *31*, 54603.
(365) Enoch, S. J.; Roberts, D. W. Predicting Skin Sensitization Potency for Michael Acceptors in the LLNA Using Quantum Mechanics Calculations. *Chem. Res. Toxicol.* **2013**, *26* (5), 767−774.
(366) Limban, C.; Nuţă, D. C.; Chiriţă, C.; Negreş, S.; Arsene, A. L.; Goumenou, M.; Karakitsios, S. P.; Tsatsakis, A. M.; Sarigiannis, D. A. The Use of Structural Alerts to Avoid the Toxicity of Pharmaceuticals. *Toxicol Rep* **2018**, *5*, 943−953.
(367) Rodgers, S.; Glen, R. C.; Bender, A. Characterizing Bitterness: Identification of Key Structural Features and Development of a Classification Model. *J. Chem. Inf. Model.* **2006**, *46* (2), 569−576.
(368) Zadeh, L. A. Fuzzy Logic—a Personal Perspective. *Fuzzy Sets and Systems* **2015**, *281*, 4−20.
(369) Chen, M.; Borlak, J.; Tong, W. High Lipophilicity and High Daily Dose of Oral Medications Are Associated with Significant Risk for Drug-Induced Liver Injury. *Hepatology* **2013**, *58* (1), 388−396.
(370) Friederichs, M.; Fränzle, O.; Salski, A. Fuzzy Clustering of Existing Chemicals according to Their Ecotoxicological Properties. *Ecol. Modell.* **1996**, *85* (1), 27−40.
(371) Blomme, E. A. G.; Will, Y. Toxicology Strategies for Drug Discovery: Present and Future. *Chem. Res. Toxicol.* **2016**, *29* (4), 473−504.
(372) Stepan, A. F.; Walker, D. P.; Bauman, J.; Price, D. A.; Baillie, T. A.; Kalgutkar, A. S.; Aleo, M. D. Structural Alert/reactive Metabolite Concept as Applied in Medicinal Chemistry to Mitigate the Risk of Idiosyncratic Drug Toxicity: A Perspective Based on the Critical Examination of Trends in the Top 200 Drugs Marketed in the United States. *Chem. Res. Toxicol.* **2011**, *24* (9), 1345−1410.
(373) Dambach, D. M.; Misner, D.; Brock, M.; Fullerton, A.; Proctor, W.; Maher, J.; Lee, D.; Ford, K.; Diaz, D. Safety Lead Optimization and Candidate Identification: Integrating New Technologies into Decision-Making. *Chem. Res. Toxicol.* **2016**, *29* (4), 452−472.
(374) Altmann, A.; Toloşi, L.; Sander, O.; Lengauer, T. Permutation Importance: A Corrected Feature Importance Measure. *Bioinformatics* **2010**, *26* (10), 1340−1347.
(375) Ribeiro, M. T.; Singh, S.; Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *arXiv [cs.LG]*, **2016**. DOI: 10.48550/arXiv.1602.04938.
(376) Ponzoni, I.; Sebastián-Pérez, V.; Requena-Triguero, C.; Roca, C.; Martínez, M. J.; Cravero, F.; Díaz, M. F.; Páez, J. A.; Arrayás, R. G.; Adrio, J.; Campillo, N. E. Hybridizing Feature Selection and Feature Learning Approaches in QSAR Modeling for Drug Discovery. *Sci. Rep.* **2017**, *7* (1), 2403.
(377) Bolboaca, S. D.; Jäntschi, L. Quantitative Structure-Activity Relationships: Linear Regression Modelling and Validation Strategies by Example. *BIOMATH* **2013**, *2* (1), 1309089.
(378) Venkatraman, V.; Dalby, A. R.; Yang, Z. R. Evaluation of Mutual Information and Genetic Programming for Feature Selection in QSAR. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (5), 1686−1692.
(379) Wellawatte, G. P.; Seshadri, A.; White, A. D. Model Agnostic Generation of Counterfactual Explanations for Molecules. *Chem. Sci.* **2022**, *13* (13), 3697−3705.
(380) Czub, N.; Pacławski, A.; Szlęk, J.; Mendyk, A. Do AutoML-Based QSAR Models Fulfill OECD Principles for Regulatory Assessment? A 5-HT1A Receptor Case. *Pharmaceutics* **2022**, *14* (7), 1415.
(381) Sushko, Y.; Novotarskyi, S.; Körner, R.; Vogt, J.; Abdelaziz, A.; Tetko, I. V. Prediction-Driven Matched Molecular Pairs to Interpret QSARs and Aid the Molecular Optimization Process. *J. Cheminform.* **2014**, *6* (1), 48.
(382) Antczak, P.; Ortega, F.; Chipman, J. K.; Falciani, F. Mapping Drug Physico-Chemical Features to Pathway Activity Reveals Molecular Networks Linked to Toxicity Outcome. *PLoS One* **2010**, *5* (8), No. e12385.
(383) Krämer, A.; Green, J.; Pollard, J., Jr; Tugendreich, S. Causal Analysis Approaches in Ingenuity Pathway Analysis. *Bioinformatics* **2014**, *30* (4), 523−530.
(384) Mayburd, A. L.; Martlínez, A.; Sackett, D.; Liu, H.; Shih, J.; Tauler, J.; Avis, I.; Mulshine, J. L. Ingenuity Network-Assisted Transcription Profiling: Identification of a New Pharmacologic Mechanism for MK886. *Clin. Cancer Res.* **2006**, *12* (6), 1820−1827.
(385) Huang, L.-C.; Wu, X.; Chen, J. Y. Predicting Adverse Drug Reaction Profiles by Integrating Protein Interaction Networks with Drug Structures. *Proteomics* **2013**, *13* (2), 313−324.
(386) Feyen, E.; Cools, J.; Van Fraeyenhove, J.; Tubeeckx, M.; De Winter, H.; Audenaert, D.; De Keulenaer, G. W.; Segers, V. F. Identification of Small-Molecule ERBB4 Agonists for the Treatment of Heart Failure. *Cardiovasc. Res.* **2022**, *118* (Supplement_1), cvac066.098.
(387) Hosseini-Gerami, L.; Higgins, I. A.; Collier, D. A.; Laing, E.; Evans, D.; Broughton, H.; Bender, A. Benchmarking Causal Reasoning Algorithms for Gene Expression-Based Compound Mechanism of Action Analysis. *BMC Bioinformatics* **2023**, *24* (1), 154.
(388) Louw, C.; Truter, N.; Bergh, W.; van den Heever, M.; Horn, S.; Oudrhiri, R.; van Niekerk, D.; Loos, B.; Singh, R. ALaSCA: A Novel in Silico Simulation Platform to Untangle Biological Pathway Mechanisms, with a Case Study in Type 1 Diabetes Progression. *bioRxiv* **2023**, DOI: 10.1101/2023.03.16.532913.
(389) Farrell, D. J.; Bower, L. Fatal Water Intoxication. *J. Clin. Pathol.* **2003**, *56* (10), 803−804.
(390) Walker, D. K. The Use of Pharmacokinetic and Pharmacodynamic Data in the Assessment of Drug Safety in Early Drug Development. *Br. J. Clin. Pharmacol.* **2004**, *58* (6), 601−608.
(391) Bassani, D.; Parrott, N. J.; Manevski, N.; Zhang, J. D. Another String to Your Bow: Machine Learning Prediction of the Pharmacokinetic Properties of Small Molecules. *Expert Opin. Drug Discovery* **2024**, *19* (6), 683−698.


https://doi.org/10.1021/acs.chemrestox.5c00033
*Chem. Res. Toxicol.* XXXX, XXX, XXX−XXX

(392) Mavroudis, P. D.; Teutonico, D.; Abos, A.; Pillai, N. Application of Machine Learning in Combination with Mechanistic Modeling to Predict Plasma Exposure of Small Molecules. *Front. Syst. Biol.* **2023**, *3*. DOI: 10.3389/fsysb.2023.1180948.
(393) Schneckener, S.; Grimbs, S.; Hey, J.; Menz, S.; Osmers, M.; Schaper, S.; Hillisch, A.; Göller, A. H. Prediction of Oral Bioavailability in Rats: Transferring Insights from in Vitro Correlations to (deep) Machine Learning Models Using in Silico Model Outputs and Chemical Structure Parameters. *J. Chem. Inf. Model.* **2019**, *59* (11), 4893−4905.
(394) Andrews-Morger, A.; Reutlinger, M.; Parrott, N.; Olivares-Morales, A. A Machine Learning Framework to Improve Rat Clearance Predictions and Inform Physiologically Based Pharmacokinetic Modeling. *Mol. Pharmaceutics* **2023**, *20* (10), 5052−5065.
(395) Handa, K.; Wright, P.; Yoshimura, S.; Kageyama, M.; Iijima, T.; Bender, A. Prediction of Compound Plasma Concentration-Time Profiles in Mice Using Random Forest. *Mol. Pharmaceutics* **2023**, *20* (6), 3060−3072.
(396) Heyndrickx, W.; Mervin, L.; Morawietz, T.; Sturm, N.; Friedrich, L.; Zalewski, A.; Pentina, A.; Humbeck, L.; Oldenhof, M.; Niwayama, R.; Schmidtke, P.; Fechner, N.; Simm, J.; Arany, A.; Drizard, N.; Jabal, R.; Afanasyeva, A.; Loeb, R.; Verma, S.; Harnqvist, S.; Holmes, M.; Pejo, B.; Telenczuk, M.; Holway, N.; Dieckmann, A.; Rieke, N.; Zumsande, F.; Clevert, D.-A.; Krug, M.; Luscombe, C.; Green, D.; Ertl, P.; Antal, P.; Marcus, D.; Do Huu, N.; Fuji, H.; Pickett, S.; Acs, G.; Boniface, E.; Beck, B.; Sun, Y.; Gohier, A.; Rippmann, F.; Engkvist, O.; Göller, A. H.; Moreau, Y.; Galtier, M. N.; Schuffenhauer, A.; Ceulemans, H. MELLODDY: Cross-Pharma Federated Learning at Unprecedented Scale Unlocks Benefits in QSAR without Compromising Proprietary Information. *J. Chem. Inf. Model.* **2024**, *64* (7), 2331−2344.
(397) Peteani, G.; Huynh, M. T. D.; Gerebtzoff, G.; Rodríguez-Pérez, R. Application of Machine Learning Models for Property Prediction to Targeted Protein Degraders. *Nat. Commun.* **2024**, *15* (1), 5764.
(398) Paul, S. M.; Mytelka, D. S.; Dunwiddie, C. T.; Persinger, C. C.; Munos, B. H.; Lindborg, S. R.; Schacht, A. L. How to Improve R&D Productivity: The Pharmaceutical Industry's Grand Challenge. *Nat. Rev. Drug Discovery* **2010**, *9* (3), 203−214.
(399) Strovel, J.; Sittampalam, S.; Coussens, N. P.; Hughes, M.; Inglese, J.; Kurtz, A.; Andalibi, A.; Patton, L.; Austin, C.; Baltezor, M.; Beckloff, M.; Weingarten, M.; Weir, S. *Early Drug Discovery and Development Guidelines: For Academic Researchers, Collaborators, and Start-up Companies*; Eli Lilly & Company and the National Center for Advancing Translational Sciences, 2016.
(400) Aronson, J. K.; Green, A. R. Me-Too Pharmaceutical Products: History, Definitions, Examples, and Relevance to Drug Shortages and Essential Medicines Lists. *Br. J. Clin. Pharmacol.* **2020**, *86* (11), 2114−2122.
(401) Krieger, J.; Li, D.; Papanikolaou, D. Missing Novelty in Drug Development*. *Rev. Financ. Stud.* **2022**, *35* (2), 636−679.
(402) Adeshina, Y. O.; Deeds, E. J.; Karanicolas, J. Machine Learning Classification Can Reduce False Positives in Structure-Based Virtual Screening. *Proc. Natl. Acad. Sci. U. S. A.* **2020**, *117* (31), 18477−18488.
(403) Provins, L.; Jnoff, E.; Genicot, C. Back-up Strategies in Drug Discovery: What, How and When? *Drug Discovery Today* **2014**, *19* (11), 1808−1811.
(404) Segall, M. D.; Barber, C. Addressing Toxicity Risk When Designing and Selecting Compounds in Early Drug Discovery. *Drug Discovery Today* **2014**, *19* (5), 688−693.
(405) Lexchin, J. How Safe and Innovative Are First-in-Class Drugs Approved by Health Canada: A Cohort Study. *Healthc. Policy* **2016**, *12* (2), 65−75.


https://doi.org/10.1021/acs.chemrestox.5c00033
*Chem. Res. Toxicol.* XXXX, XXX, XXX−XXX