DISSERTATION | DOCTORAL THESIS

Titel | Title

Development of Theoretical Approaches for Tackling Assay Interference in High-Throughput Screening

verfasst von | submitted by

Dott. Dott.mag. Vincenzo Palmacci

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Doktor der Naturwissenschaften (Dr.rer.nat.)

Wien | Vienna, 2025

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 796 610 449
Dissertationsgebiet lt. Studienblatt | Field of study as it appears on the student record sheet:	Pharmazie
Betreut von | Supervisor:	Univ.-Prof. Mag. Dr. Johannes Kirchmair

Acknowledgments

This doctoral thesis is the result of nearly four years of research conducted in collaboration between the University of Vienna, Bayer AG, and AstraZeneca.

I would like to express my sincere gratitude to all those who supported and contributed to this work.

First and foremost, I thank my supervisor, Prof. Johannes Kirchmair, for the opportunity to pursue this research under his guidance. His support, critical feedback, and scientific rigor have been valuable in shaping the work presented in this thesis.

I am deeply grateful to my collaborators and colleagues at Bayer AG, whose insights and expertise enriched the direction of this project. I am especially grateful to Floriane Montanari, Djörk-Arne Clevert, and the HTS development team at Bayer, particularly Stefan Mundt and Michael Koch, for their support and for providing access to the proprietary datasets that were essential to this research.

I would also like to thank the team at AstraZeneca, Gothenburg, for hosting me during part of this doctoral work and for providing an inspiring and intellectually rich research environment. In particular, I am grateful to Jon Paul Janet and Ola Engkvist for their insightful feedback, mentorship, and thoughtful discussions, which were crucial during the later stages of this project.

I extend my sincere appreciation to the members of the COMP3D group at the University of Vienna for their technical support, and the many stimulating discussions throughout this project. I am particularly grateful to Roxane, Matthias, and Vincent, whose collaboration, scientific insights, and continuous support, both within and beyond the university, significantly contributed to the progress of my work.

This thesis also benefited greatly from the broader network of researchers and fellow PhD students within the AIDD consortium. I thank them for the valuable scientific exchange and camaraderie throughout this journey. In particular, I would like to acknowledge Alessio and Yasmine for their moral support and thoughtful discussions, which helped shape aspects of the work presented in this thesis.

Finally, I would like to acknowledge the unwavering support of my family and close friends, whose encouragement helped me navigate the challenges of this long and demanding path.

Funding

The project leading to this thesis has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie Innovative Training Network - European Industrial Doctorate No 956832, "Advanced machine learning for Innovative Drug Discovery" (AIDD). The thesis reflects only my view, and neither the European Commission nor the Research Executive Agency (REA) are responsible for any use that may be made of the information it contains.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this thesis work, I used ChatGPT to improve the readability of the manuscript. After using this tool/service, I reviewed and edited the content as needed and took full responsibility for the content of the publication.

Declaration of Authorship

I hereby declare that the thesis I have provided is my own unique and original authorial work that I have independently worked out. All sources, references, and literature used or excerpted during the elaboration of this work are properly cited and listed in complete reference to the due source.

Vienna July, 2025 Vincenzo Palmacci

Table of contents

Preface

The scientific work presented in this thesis was conducted between September 2021 and July 2025 under the supervision of Prof. Johannes Kirchmair (COMP3D Group, University of Vienna), and carried out across three institutions: the University of Vienna, Bayer AG (Berlin), and AstraZeneca (Gothenburg).

Chapter 1 serves as an introductory chapter, guiding the reader through the key concepts of modern drug discovery, with a particular focus on assay automation, high-throughput screening (HTS), and the integration of artificial intelligence (AI) into the drug discovery pipeline. It provides an overview of the main assay technologies used in HTS and outlines major challenges in the field, especially the occurrence of assay-interfering compounds. The chapter concludes with the presentation of a comprehensive review article (Study 1), which surveys current methodologies for addressing assay interference. This article was published in *Nature Reviews Chemistry.*

Chapter 2 provides the methodological foundation for the experimental work presented in this thesis. It introduces the public and proprietary datasets used for model development, describes the preprocessing pipelines implemented, and presents the machine learning algorithms applied, with an emphasis on their relevance to chemical data.

Chapter 3 outlines the central aims of the doctoral research project.

Chapter 4 presents the results of the computational studies conducted during the research period, focusing on the development of machine learning (ML) workflows for predicting compounds that interfere with assays. The study details the preprocessing and analysis of a large proprietary dataset from Bayer AG as well as the development of predictive models targeting fluorescence-based assay interference. The work was published as an article in *Artificial Intelligence in the Life Sciences* (Study 2). A further study describes a new data augmentation framework that integrates reinforcement learning (RL) for de novo molecular design with human-in-the-loop feedback to enhance model performance and transfer expert knowledge into the predictive pipeline. This work was published in the *Journal of Cheminformatics* (Study 3).

Chapter 5 concludes the thesis with a summary of findings and an outlook on future research directions in the field.

Table of abbreviations

HTS	High-Throughput Screening
AI	Artificial Intelligence
ML	Machine Learning
RL	Reinforcement Learning
DMTA	Design-Make-Test-Analyze
FLINT	Fluorescence Intensity
FRET	Förster Resonance Energy Transfer
TR-FRET	Time-Resolved Förster Resonance Energy Transfer
GFP	Green Fluorescent Protein
cAMP	Cyclic Adenosine Monophosphate
GPCR	G Protein-Coupled Receptor
BRET	Bioluminescence Resonance Energy Transfer
DL	Deep Learning
ADMET	Absorption, Distribution, Metabolism, Excretion, and Toxicity
FAIR	Findable, Accessible, Interoperable, and Reusable
API	Application Programming Interface
NCBI	National Center for Biotechnology Information
EMBL	European Molecular Biology Laboratory
EBI	European Bioinformatics Institute
RF	Random Forest
FFNN	Feedforward Neural Network
GNN	Graph Neural Network
SVM	Support Vector Machine
KNIME	Konstanz Information Miner
SMILES	Simplified Molecular Input Line Entry System
SDF	Structure Data File

ECFP	Extended-Connectivity Fingerprint
QSAR	Quantitative Structure–Activity Relationship
ReLU	Rectified Linear Unit
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
MCC	Matthews Correlation Coefficient
EF	Enrichment Factor
ROC-AUC	Receiver Operating Characteristic – Area Under the Curve
PR-AUC	Precision-Recall – Area Under the Curve
RNN	Recurrent Neural Network

Chapter 1: Introduction

Discovering a new therapeutic compound is a complex and resource-intensive scientific challenge. On average, the process takes 10 to 15 years and can require investments exceeding one billion dollars.[1] This immense effort makes drug discovery one of the most demanding pursuits in modern science. Despite the difficulty, a successful drug can generate substantial returns for companies and, more importantly, offer society powerful new tools to combat diseases.

Given its societal and economic importance, drug discovery is established as a fundamental discipline in modern science, with substantial scientific effort dedicated to increasing the success rate of discovery campaigns.[2]

The drug discovery pipeline is typically divided into two major phases: preclinical and clinical. The preclinical phase encompasses all early-stage research, including in vitro and in vivo testing on model organisms.[3] It begins with the identification of a biologically relevant target, often a gene implicated in a disease pathway. When molecular information about the disease is lacking, alternative strategies may focus on targeting cells to induce a desired phenotypic response.

Once a target has been selected, large-scale in vitro screening of compound libraries (often comprising millions of entries) is conducted to identify hit compounds that show promising biological activity.[4] These hits are then optimized through iterative lead optimization cycles aimed at improving potency, selectivity, and pharmacological properties.[5] Optimization continues until compounds reach activity in the nanomolar range, at which point in vivo studies are initiated. These studies assess pharmacokinetics, toxicity, and efficacy in living organisms, typically mammals, to determine the compound's potential for human use. If preclinical testing demonstrates safety and efficacy, the compound enters the clinical phase, which includes three rounds of human trials.[6]

While the preclinical phase typically spans 3-5 years, clinical trials often take close to a decade (Fig. 1). Clinical trials are the most critical and uncertain phase, as the true effects of a compound, both therapeutic and adverse, can only be observed in humans. Due to the complexity of human biology, some interactions remain fundamentally unpredictable, making clinical trials a significant bottleneck in the overall process.

While failure in early-stage discovery is both expected and necessary, during clinical trials, financial and ethical stakes are at their highest, and failure during this phase often results in significant economic losses and reputational damage that are difficult to recover from. For

this reason, it is essential to maximize the specificity and reliability of preclinical assessments.

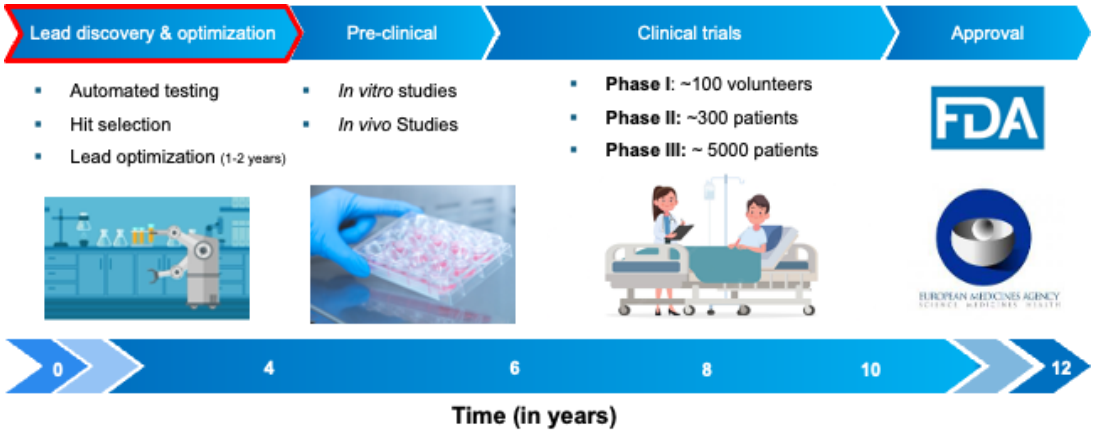


Figure 1: Schematic view of the drug discovery pipeline. The timeline (x-axis) reports the approximate duration in years required to progress from lead discovery to regulatory approval. The process typically spans over a decade, with clinical trials representing the most resource-intensive and high-risk phase.

Advances in experimental protocols and, more significantly, improvements in in-silico modeling techniques have enhanced our ability to characterize compound behavior with greater precision. This progress supports more informed decision-making, enabling the prioritization of promising drug candidates while minimizing the risk of false positives and missed hits.[7]

1.1 Automation and Upscaling in Drug Discovery

Until the late 20th century, drug discovery primarily relied on small teams of biologists and chemists generating hypotheses and validating them through sequential wet lab experimentation. While this approach was slow and labor-intensive, it played a foundational role in advancing medical science, leading to the development of numerous life-saving therapies across a wide range of diseases.[2]

Manual laboratory processes, while having a foundational role, placed substantial constraints on throughput and efficiency. Performing preclinical testing that relies solely on human labor requires a conspicuous investment of time and resources, particularly when verifying spurious activity signals or investigating potential sources of bias. To address these limitations, manual laboratory work was gradually displaced beginning in the 1980s, as

advances in automation and robotics began to reshape the drug discovery pipeline.[8] These innovations introduced the paradigm of parallel testing, enabling rapid, iterative Design-Make-Test-Analyze (DMTA) cycles at a pace unattainable with traditional, manual workflows.

Powering the new, faster pipeline are HTS and automated synthesis laboratories, both of which have become standard in preclinical drug discovery over the past four decades.[8,9] At the core of these technologies lie robotics, miniaturized assays, and information systems that enable the rapid evaluation of large compound libraries against biological targets, as well as the automated generation and optimization of chemical matter.

Today, pharmaceutical companies equipped with state-of-the-art automation infrastructure are capable of screening up to one million compounds per day. Full-library screens involving several million molecules can be completed within weeks. At the same time, integrated synthesis systems allow for the high-throughput production of follow-up analogs in response to screening results.[10]

1.2 Miniaturized Biological Assays

HTS and pre-clinical follow-up experiments (e.g., evaluation of compound potency and toxicity) rely heavily on miniaturized biological assays, which enable the automated testing of compounds in a compact format and at high throughput. These assays, developed in plate-based formats, form the experimental backbone of preclinical drug discovery. Their compact format not only reduces reagent consumption but also allows for seamless integration with automated laboratory systems, making large-scale, reproducible experimentation feasible.

Miniaturized assays developed for hit identification are conducted in vitro and are typically categorized into two main classes: biochemical assays and cell-based assays. The choice between biochemical and cell-based assays, as well as the selection of a specific readout technology, depends on several factors, including the biological question, target tractability, assay robustness, and the likelihood of assay interference.

Biochemical assays are designed to measure either binding affinity or inhibition of enzymatic activity using a purified target protein, often employing a competitive format, in which the test compound must displace a known ligand bound to the target. They are conducted in 384-well plates, offering a balance between throughput, reagent consumption, and instrumentation cost. To enable precise bioactivity readout and minimize the signal-to-noise ratio, bioactivity is typically measured using optical methods such as absorbance, fluorescence, or luminescence.[11]

Fluorescence-based techniques are among the most widely used in HTS due to their sensitivity and ease of implementation.[12] Among these, fluorescence intensity (FLINT), fluorescence polarization, and Förster resonance energy transfer (FRET) are the most commonly employed in screening campaigns. FLINT relies on fluorogenic substrates that emit fluorescence following enzymatic catalysis, providing a direct measure of enzyme activity. Fluorescence polarization, by contrast, measures the rotational diffusion of fluorescently labeled molecules. When a small molecule binds to a larger target, its rotational motion is restricted, leading to an increase in the polarization of the emitted light. This change can be detected and used to infer binding interactions. FRET detects molecular interactions by exploiting energy transfer between a donor and an acceptor fluorophore. When the two are in proximity (typically upon molecular binding), energy is transferred from the donor to the acceptor, resulting in a detectable emission from the latter.

A common variant of FRET is Time-Resolved FRET (TR-FRET), which improves signal-to-noise ratio by gating detection. TR-FRET uses lanthanide complexes (e.g., europium or terbium) as donors due to their long fluorescence lifetimes (hundreds of microseconds to milliseconds). Fluorescence emission from short-lived background signals, including autofluorescence from library compounds, is allowed to decay before signal acquisition (>50 ms delay), thereby improving specificity and assay quality.

On the other hand, cell-based assays assess compound activity in a more physiologically relevant context. The analyte (i.e., the small molecule) is exposed to multiple, often unknown, biomolecular targets. Successful binding with one or more macromolecules within the cell environment can trigger downstream effects, frequently observed as changes in the phenotype (e.g., cell death).[13] Cell-based assays are particularly valuable when no well-defined molecular target is known.

Various technologies are employed to monitor compound-induced changes in cellular function. Luciferase-based assays are among the most sensitive, generating light upon substrate conversion and enabling real-time monitoring of gene expression or pathway activation. A typical example is the reporter gene assay, in which proteins such as luciferase or green fluorescent protein (GFP) are expressed under the control of pathway-specific promoters. Modulation of the targeted pathway leads to changes in reporter expression, providing a quantifiable readout.

Another widely used strategy involves secondary messenger assays, which measure rapid intracellular signaling events, such as calcium flux or cyclic adenosine monophosphate (cAMP) levels. These assays are particularly common in G protein-coupled receptor (GPCR) screening due to their ability to capture early, transient signaling responses.

Finally, protein-protein interaction assays, including FRET and bioluminescence resonance energy transfer (BRET), are used to assess molecular proximity in live cells. BRET, in particular, offers high sensitivity and low background noise without the need for external light excitation, making it well suited for studying dynamic interactions in cellular environments.

1.3 Challenges in the Era of Automated Compound Testing

The acceleration enabled by miniaturized assays and automation has resulted in the generation of vast quantities of bioactivity data, demanding robust computational frameworks capable of organizing experimental outputs, extracting meaningful patterns, and generating actionable insights. Hence, the rise of laboratory automation has been paralleled by a growing reliance on computational methods.[14] While traditional in silico techniques such as molecular docking and molecular dynamics simulations have supported medicinal chemistry since the late 1900s,[15] the past two decades have seen an accelerating shift toward artificial intelligence (AI) and data-driven modeling,[16] driven by the increasing availability of large, high-quality proprietary and public datasets. Resources such as ChEMBL[17] and PubChem[18] now host millions of compound-bioactivity pairs across a broad spectrum of assay types and biological targets. Nonetheless, top-tier pharmaceutical companies have developed proprietary databases that rival, and in some cases exceed, the scope and depth of publicly available resources. This wealth of high-throughput data has enabled the development of machine learning (ML) and deep learning (DL) models capable of performing in silico screening of large chemical libraries, predicting molecular properties, and approximating experimental outcomes with increasing accuracy. As a result, ML approaches have become integrated across the entire drug discovery pipeline, from target discovery,[19] to ADMET (absorption, distribution, metabolism, excretion, and toxicity) prediction,[20] and automated synthesis planning.[21]

If in the early days, the main obstacles were rooted in chemistry and biochemistry (optimizing reactions, refining assay conditions, and improving analytical techniques), today, in the era of big data, computational bottlenecks are a top priority. Challenges include the handling of the inherent uncertainty in high-throughput datasets,[22] the building of predictive models that can generalize across large chemical spaces, and the prediction of ADMET (Absorption Distribution Metabolism Excretion and Toxicity) endpoints as well as reaction outcomes.[23] Advances in ML and cheminformatics have addressed many of these challenges. However, some tasks remain challenging. Predicting toxicity, reaction yields,[24,25] or the activity of compounds against less-characterized biological targets still poses significant difficulties. In many cases, these challenges persist due to insufficient sizeable

datasets (e.g., in vivo toxicity) or the inability to extract meaningful patterns from existing, but noisy datasets.[26]

Early-stage drug discovery, particularly HTS, exemplifies this data paradox: while HTS campaigns generate enormous quantities of valuable biochemical data, including chemical structures, assay metadata, and bioactivity profiles, these datasets often suffer from inconsistent annotations and experimental variability, which limit their usability for predictive modeling. Additionally, technical artifacts such as batch effects or compound evaporation can obscure true biological signals.[27]

Importantly, more than 95% of initial hits are not the result of genuine interaction with the biological target but rather stem from interference with assay components.[28] These assay-interfering compounds, or bad actors, can arise through diverse mechanisms, including colloidal aggregation, autofluorescence, and non-specific chemical reactivity. Some interfere broadly across many assay types, while others affect specific readout technologies. Detecting such compounds experimentally would require extensive follow-up testing, which is often impractical on a large scale.

1.4 Identifying and Triaging Assay-Interfering Compounds

Despite the challenges involved in HTS, the approach remains a key pillar of modern drug discovery. Without HTS, testing at scale against a variety of biological targets would not be possible. Although HTS data can be noisy and prone to false positives, their scale and diversity make them a powerful asset for ML when appropriately curated and integrated. Hence, there is strong interest in improving the post-processing of HTS campaigns to increase the specificity of hit selection and reduce the experimental effort spent optimizing false leads.

A variety of strategies have been developed to detect assay-interfering compounds. Many of the most established methods are experimental, involving adjustments to assay conditions or the addition of control elements. For example, screening under detergent conditions to avoid colloidal aggregation, or parallel orthogonal assays, may be used to identify specific types of interference and de-prioritize detected false positives. These approaches are widely applied and often serve as the first line of defense against assay-interfering compounds. However, they require additional instrumentation, reagents, and time.

In response to these limitations, computational methods for predicting assay-interfering compounds have gained increasing attention as more scalable and cost-effective alternatives. These strategies can be broadly categorized into global and specialized models. Global models leverage activity patterns across large datasets and multiple assay

technologies to flag potentially problematic compounds, whereas specialized approaches focus on a single assay type or interference mechanism. These models often suffer from limited generalizability, as they heavily depend on the quality and diversity of the training datasets. Still, computational modeling of assay-interfering compounds has gained traction alongside the rise of ML, offering powerful tools for pattern recognition and predictive modeling.

1.5 Study 1: Tackling Assay Interference Associated with Small Molecules

Tan L., Hirte S., Palmacci V., Stork C., Kirchmair J.

Published in *Nature Reviews Chemistry 2024*, 8 (5), 319-339.

This section presents a comprehensive overview of the current state of the art in addressing assay interference caused by small molecules. The review outlines the principal mechanisms underlying assay interference and discusses both experimental and computational strategies designed to mitigate or predict such effects. Particular attention is given to methods that enhance the reliability of hit selection in HTS campaigns. Each technique is critically examined, exploring its strengths and limitations, offering a consolidated foundation for ongoing and future efforts aimed at minimizing false-positive results in early drug discovery.

Author's contribution to the paper:

- Performed literature review and data collection on computational methods.
- Drafted and revised major sections of the manuscript, including methodological comparisons and critical evaluations.
- Participated in manuscript editing, integration of co-author feedback, and preparation for journal submission.

nature reviews chemistry

https://doi.org/10.1038/s41570-024-00593-3

Review article

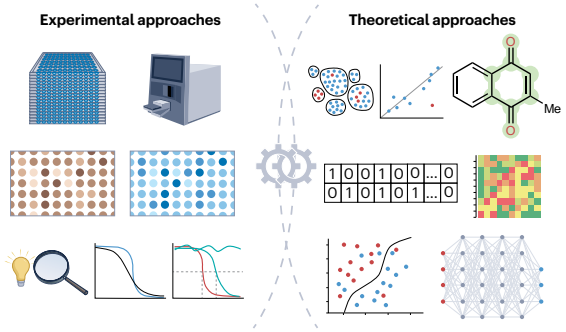
Check for updates

Tackling assay interference associated with small molecules

Lu Tan[1], Steffen Hirte[2,3], Vincenzo Palmacci[2,3], Conrad Stork[4,6] & Johannes Kirchmair[2,5]

Abstract

Biochemical and cell-based assays are essential to discovering and optimizing efficacious and safe drugs, agrochemicals and cosmetics. However, false assay readouts stemming from colloidal aggregation, chemical reactivity, chelation, light signal attenuation and emission, membrane disruption, and other interference mechanisms remain a considerable challenge in screening synthetic compounds and natural products. To address assay interference, a range of powerful experimental approaches are available and in silico methods are now gaining traction. This Review begins with an overview of the scope and limitations of experimental approaches for tackling assay interference. It then focuses on theoretical methods, discusses strategies for their integration with experimental approaches, and provides recommendations for best practices. The Review closes with a summary of the critical facts and an outlook on potential future developments.



Sections

[1]Drug Discovery Sciences, Boehringer Ingelheim RCV GmbH & Co KG, Vienna, Austria. [2]Department of Pharmaceutical Sciences, Division of Pharmaceutical Chemistry, Faculty of Life Sciences, University of Vienna, Vienna, Austria. [3]Vienna Doctoral School of Pharmaceutical, Nutritional and Sport Sciences (PhaNuSpo), University of Vienna, Vienna, Austria. [4]Department of Informatics, Center for Bioinformatics, Faculty of Mathematics, Informatics and Natural Sciences, Universität Hamburg, Hamburg, Germany. [5]Christian Doppler Laboratory for Molecular Informatics in the Biosciences, Department for Pharmaceutical Sciences, University of Vienna, Vienna, Austria. [6]Present address: BASF SE, Ludwigshafen am Rhein, Germany. e-mail: johannes.kirchmair@univie.ac.at

Introduction

Today, a plethora of biochemical and cell-based assays to experimentally test the biological activity of small organic compounds are at our disposal. These assay technologies are indispensable to a host of industries, including the pharmaceutical, agrochemical and cosmetics sectors. They are also essential tools to, for example, investigate the biological activities of food compounds and better understand their beneficial or harmful effects on human health[1].

Depending on the assay and automation technologies used, with a certain lead time to set up the specific testing system, hundreds of thousands of compounds can be tested by high-throughput screening (HTS) per day. The outcome of such large-scale screening campaigns does not tend to be a shortage of hits. Quite on the contrary, screening campaigns regularly produce a sizable number of initial hits. However, only a fraction may reflect genuine, specific interactions of the hit compound with a biomacromolecule. In fact, typical screening libraries contain a substantial number of 'bad actors', commonly also referred to as 'nuisance compounds', pan-assay interference compounds (PAINS)[2], compounds interfering with an assay technology (CIATs)[3] or, in the context of natural products research, invalid metabolic panaceas (IMPs)[4] (see Box 1 for an overview of relevant terms and definitions). All these terms describe compounds that can cause interference in biological assays, resulting in false-positive or false-negative assay readouts. Although they are often used interchangeably, not all are equivalent. Another common but often inaccurately used term is 'frequent hitter'[5], which describes compounds showing higher-than-expected hit rates in distinct biological assays. Frequent hitter behaviour is often, but not always, linked to assay interference.

Two major categories of assay interference can be distinguished: generalized interference and technology-related interference. Generalized interference primarily results from the physicochemical properties of a compound, such as the ability to form colloidal aggregates, undergo chemical reactions, form stable complexes with metal ions or disrupt biomembranes. Technology-related assay interference primarily occurs as a consequence of the interference of compounds with certain assay signals or components (see Box 2 and refs. 2,6–8). This categorization can be context-dependent, as assay interference stems from the intricate interplay between a compound and a specific assay component[9]. For example, compounds interfering with fluorescence-based assays because of their autofluorescence may show benign behaviour in assays measuring other types of signals. Therefore, assays based on orthogonal technologies can be particularly useful in hit verification.

A critical, still often underappreciated source of assay interference is chemical breakdown products and impurities[10,11]. They are particularly challenging to predict, trace, handle and exclude (see Box 2).

Although a lot of the recent discussions of assay interference are focused on biochemical assays, assay artefacts are also of considerable concern to cell-based assays[12–15]. In particular, cytotoxicity can lead to false-positive or false-negative results[12,14,16]. Therefore, factors influencing cellular fitness[14], such as general cytotoxicity, genotoxicity, membrane integrity and mitochondrial toxicity, may need to be checked as a source of interference in cell-based assays.

Without proper study design and controls, the fraction of hits obtained during the initial compound screening that are artefacts can reach 80%–100% (refs. 17,18). If these false hits remain undetected, they can trigger expensive hit expansion and hit-to-lead follow-up studies[2,18,19]. Such efforts will probably be futile because the observed poor or even misleading structure–activity relationships will impede the rational optimization of their biological properties. Especially problematic is the scientific literature in which compounds are falsely reported as active. Such errors have a good chance of propagation, particularly if the compounds wrongly assumed to be bioactive are readily obtainable[20,21]. The scientific community and the publishers have recognized this problem. In 2017, the editors-in-chief of nine American Chemical Society journals focusing on topics related to small-molecule drug discovery published an editorial with guidelines and recommendations on how to detect bad actors and avoid their inaccurate emergence as bioactive compounds[17]. The editorial was an important step in raising awareness about the problem of assay interference. However, with today's advanced understanding of the complexities of assay interference, and in line with concerns voiced following its publication[22], some of the statements encouraging the use of theoretical approaches, specifically the PAINS concept, require updates and refinements. This is also true for some of the author guidelines from leading journals in the field, which tend to overrate the reliability and scope of the PAINS concept as a means to minimize false reports of bioactive compounds.

Given that the conditions to which drugs are exposed in vivo differ from those present in biochemical assays[23], including compound concentrations and composition of solutions, most types of assay interference tend to not be especially relevant to drug efficacy. However, negative implications on pharmacokinetics cannot be categorically excluded[24].

Assay interference is not a problem exclusive to synthetic compounds. On the contrary, many chemotypes found in natural products have been linked to assay interference[25]. For example, one of the most widely occurring flavonoids in nature, quercetin, is known to disrupt biomembranes[26,27], interfere with spectroscopic assays[2,28], and form colloidal aggregates[29]. It has been reported as active in over half of the several hundred assays deposited in PubChem BioAssay[4], many of which represent pharmaceutically relevant proteins. Doxorubicin, an anthraquinone produced by *Streptomyces* strains, is known to be redox-active, undergo reactions with nucleophiles, and interfere with fluorescence-based assays[25]. The natural product has been reported as active in approximately 85% of 4,000 bioassays[30]. Whereas quercetin is only occasionally used as a food supplement, doxorubicin has a decades-long record of use as a chemotherapeutic for treating various cancers[31]. It is fortunate, therefore, that the development of doxorubicin was not thwarted by its assay interference behaviour. These examples illustrate just how complex the correct interpretation of assay interference behaviour can be.

In medicinal chemistry, the primary concern is about false-positive assay readouts, but it should be noted that although less common, false-negative results can also be problematic. They can lead to the missing of valuable starting points for drug discovery, and, more importantly, in toxicological screening, they may occlude potentially harmful effects.

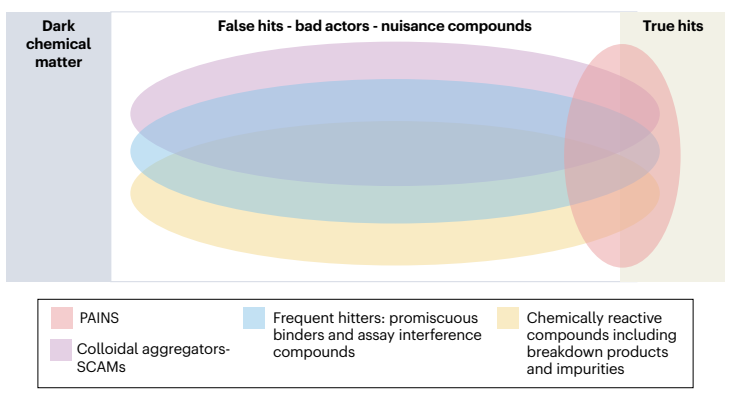
Today, a range of powerful experimental approaches for tackling assay interference is established, and in silico methods are gaining traction and recognition. The integration of these two domains (illustrated in Fig. 1) holds the potential for substantial synergistic effects. It will have a pivotal role in overcoming the challenges posed by assay interference.

This Review aims to provide an insightful overview of the scope and limitations of methods to tackle assay interference. It begins with an outline of the most relevant and established experimental approaches before focusing on theoretical approaches. Later parts

Box 1

Terminology

Hits
These are compounds that trigger a signal, typically indicative of biological activity, in biochemical or cell-based assays or in silico testing systems. 'Hit' does not imply the existence of conclusive evidence of genuine bioactivity. True hits exhibit genuine bioactivity on the target biomacromolecule; false hits trigger assay signals via different interference mechanisms.



Bad actors or nuisance compounds, or compounds interfering with an assay technology (CIATs)
These are a chemically diverse group of compounds interfering with certain assay technologies, irrespective of the underlying mechanism and frequency of occurrence of the assay interference phenomena[3]. In the context of natural products research, they are also called invalid metabolic panaceas (IMPs)[4].

Assay interference does not exclude true bioactivity; it just makes bioactivity more challenging to detect and optimize.

Pan-assay interference compounds (PAINS)
These are molecules built on scaffolds that have been linked to pan-assay interference[2]. The 480 patterns describing these substructures are derived from a proprietary compound library, screened at high concentrations with six AlphaScreen assays. The screening deck does not include highly reactive compounds, and aggregation was suppressed by adding a detergent.

PAINS is often used as a synonym for bad actors, nuisance compounds or frequent hitters, although the PAINS concept covers only a subset of these compounds.

Aggregators or small, colloidally aggregating molecules (SCAMs)
These are molecules forming colloidal aggregates when the critical aggregation concentration (CAC) of a compound is exceeded[29,58]. They are typically characterized by poor solubility owing to large, hydrophobic (and aromatic) surface areas[62,63,164,165].

Chemically reactive compounds
These are molecules carrying chemically reactive moieties that can form covalent bonds with biomacromolecules and other assay components. They include the desired, specific covalent modulators of target biomacromolecules but also chemical breakdown products and impurities. Chemically reactive moieties can alter assay readouts even when only present in trace quantities. They are challenging to detect with analytical methods.

Frequent hitters
These are substances showing higher-than-expected hit rates in distinct biological assays[5]. Most cases of frequent hitter behaviour are caused by bad actors, reactive chemical breakdown products, and impurities, particularly when observed across technologically distinct assays. Fewer cases are the result of true (genuine) promiscuous binding, that is, the specific, reversible interaction of a compound with multiple distinct biomacromolecules. Molecular properties such as lipophilicity, molecular weight and charged groups can affect compound promiscuity[43,164,166–171].

(Truly) promiscuous compounds or promiscuous binders
These are a subset of frequent hitters interacting specifically and reversibly with multiple, distinct biomacromolecules[172]. They should not be equated to nuisance compounds as they may be valuable starting points for developing innovative therapeutics, especially in polypharmacology (that is, the use or design of drugs targeting multiple disease-relevant biomacromolecules or pathways)[13,85,120,123,173–176]. A major challenge in working with such compounds is the optimization of their selectivity and safety profiles[13,43,169,175].

Privileged scaffolds
These molecular scaffolds enable compounds to interact with various biomacromolecules by matching the pharmacophoric and steric requirements of multiple ligand binding sites[177].

Dark chemical matter
These are compounds reported as inactive across a wide range of assays and targets[137]. However, it is expected that further testing of these compounds will probably reveal bioactivities. These bioactivities could be particularly interesting because of the high selectivity and, hence, probably favourable safety profile of these compounds.

Box 2

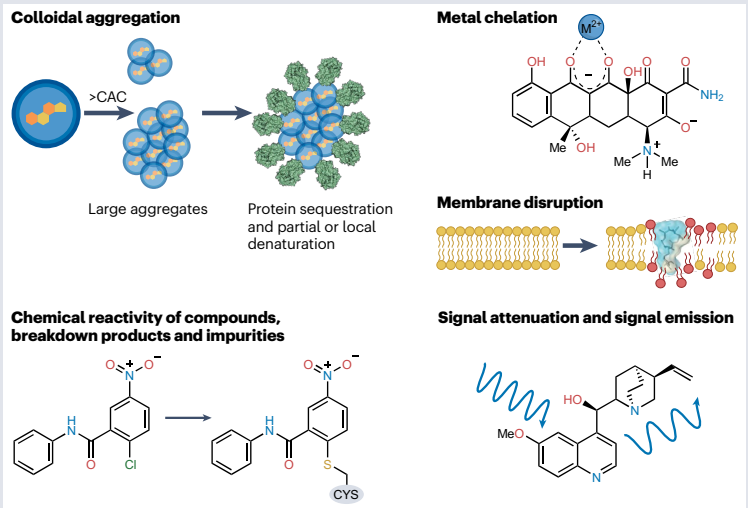
Types and sources of assay interference

Colloidal aggregation

This is the most common type of assay interference, although strongly dependent on assay conditions (for example, compound concentrations and presence or absence of detergent)[29,36,127,178–180]. Colloidal aggregation involves the formation of large particles that bind thousands of protein molecules, causing protein sequestration and partial or local denaturation[181,182]. Smaller particles mimicking a protein binding partner have also been described[183].

Chemical reactivity

This is a type of assay interference linked to the nonspecific modification of biomacromolecules (or other assay components) by chemically reactive compounds. This can include redox reactions, the formation of disulfide bonds, and reactions of nucleophilic protein components with electrophilic moieties of compounds (particularly Michael addition reactions and nucleophilic aromatic substitutions)[7,37].



Metal chelation

This refers to the formation of stable complexes with metal ions. For some drugs, chelate formation is essential to efficacy[184,185]. The interference is caused by the depletion of metal ions essential to the assay, inhibition of metalloenzymes, formation of coloured or fluorescent chelates, and other effects[8,37].

Membrane disruption

This is an assay interference caused by the loss of membrane-bound proteins, formation of precipitates, or changes to assay conditions[8].

Interferences with light-based detection methods

Because of the popularity of spectroscopic assays, these are among the most relevant types of technology-related assay interference. Assay interference caused by light signal attenuation includes fluorescence quenching[32] (energy of a fluorophore, such as in fluorescein or rhodamines, in an excited state, is transferred to an interference compound), absorbance (light at wavelengths relevant to the detection system of the assay is absorbed by the interference compound), interference with reporter enzymes[33] (can involve the direct inhibition of reporter enzymes such as luciferase by the interference compound or the attenuation of light signals emitted as a result of the activity of reporter enzymes), and light scattering (insoluble materials such as colloidal aggregates cause the diffraction of light).

Assay interference caused by light signal emission is primarily related to autofluorescence. Autofluorescence of a compound can cause, for example, false-positive signals and the deterioration of the signal-to-noise ratio[32].

Interferences caused by chemical breakdown products and impurities

Chemical breakdown products can be formed over time or, for example, in the presence of assay reagents[37]. Impurities commonly originate from synthesis or are introduced through contamination[10,11,37,72].

Major publishers of the scientific literature and funding bodies require compounds to be at least 95% pure, typically confirmed by HPLC–MS. However, some breakdown products and impurities may falsify assay readouts even when present only in traces. Even the most advanced compound management and analytical chemistry approaches may fail to identify some problematic impurifications. Therefore, re-synthesis and purification of the most interesting compounds may be required[37,72].

of the Review discuss strategies for integrating experimental and theoretical approaches and provide recommendations for best practices. The final part provides a summary of the key facts and considerations, while also offering insights into potential future developments.

Experimental approaches

The risks of interference from small organic compounds are well known for those working in biological assay technologies. Therefore, screening campaigns are usually built on two or more conceptually and technologically distinct assays, ideally complemented by in silico

approaches, to support hit triage[9,14,15,32–35] (Fig. 1). Here, we discuss the most commonly used screening components and anti-interference strategies. An overview of the experimental approaches for tackling assay interference is provided in Box 3.

Primary screening assays are used to rapidly screen large compound libraries and usually use a single measurement per compound. Although there are commonly used measures to detect assay interference, there is still a potential for substantial numbers of false-positive hits to appear in primary screening assays. Consequently, the initial hits typically undergo further testing in follow-up screens, often conducted in triplicate and involving dose titrations. This approach helps alleviate the number of false readouts associated with the rapid screening process. In any case, the relevant hits obtained from the primary screening assay should be subjected to further evaluation in more specialized and informative assays.

Counter-screen assays are designed to address and rule out false assay readouts caused by a given type of assay interference. They mimic the assay conditions and reagents of the primary screen to a certain extent, to elicit an (unspecific) activity signal in the presence of interfering compounds, but without enabling the actual biological process to occur[7,14,32,33,36–38] (such as by missing or swapping the target biomacromolecule). Counter-screen approaches can include assays measuring the signal modulation of compounds independent of the target (such as measuring compound autofluorescence in fluorescence-based assays) and enzymatic tests (such as the AmpC β-lactamase assay for detecting aggregators, wherein the enzyme is sensitive to inhibition by colloidal aggregates). The genuine biological activity of any compound active in both the primary screening and counter-screen assay(s) is questionable, whereas its assay interference is evidenced. However, inactivity in a counter-screen is neither evidence of the absence of other types of interference nor evidence of genuine biological activity.

Orthogonal assays are designed to measure the same biological effect or target through different components, technologies or pathways, making them resilient to assay interference of the same kind[7,14,32,33,38]. For example, a cell-based assay could serve as an orthogonal assay for a biochemical, primary screening assay. This setup can, in addition, offer deeper insights into the biological, functional, toxicological and pharmacokinetic properties of a compound. Screening strategies using biochemical and cell-based assays in tandem, therefore, are also one of the strongest setups to support computer-guided compound optimization. Compounds showing activity in both the primary screening assay and orthogonal assay(s) most probably exhibit genuine biological activity on the biomacromolecule of interest, although some uncertainty remains.

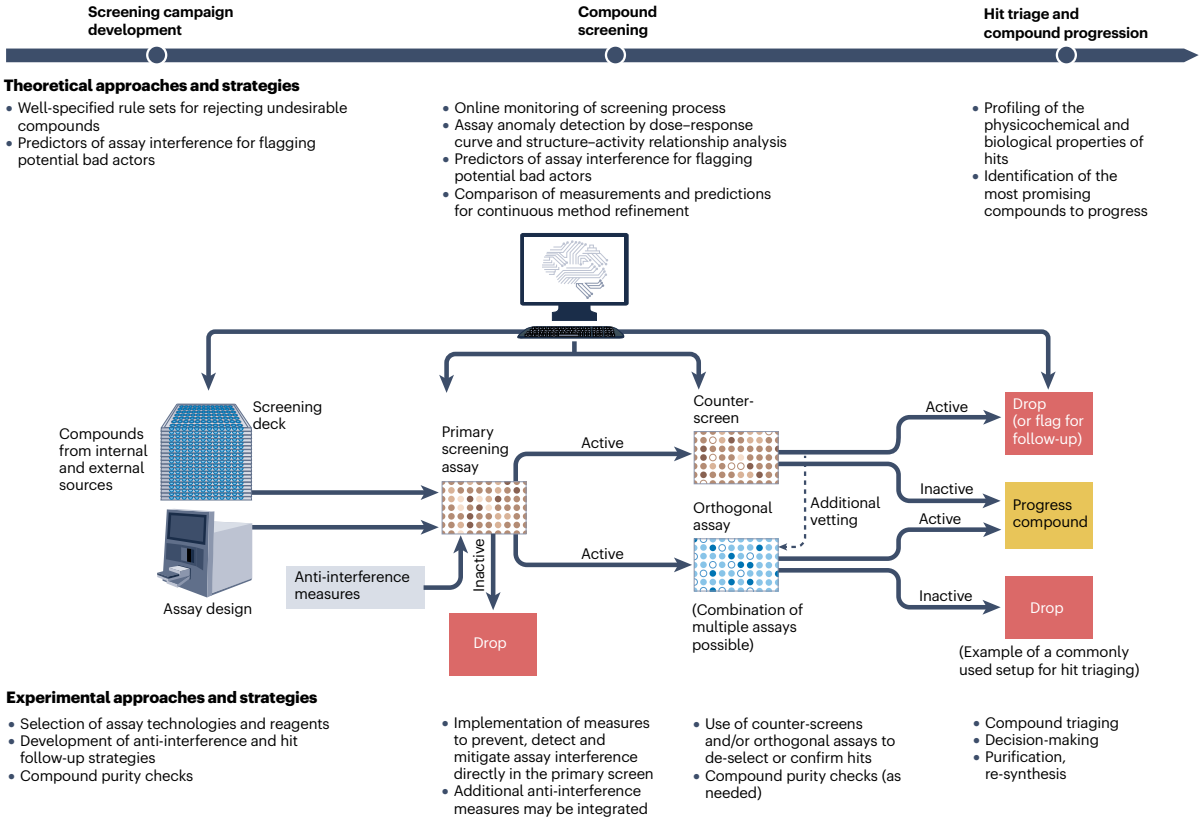


Fig. 1 | Integration of theoretical and experimental approaches to tackle assay interference caused by small organic compounds. Throughout all phases of a compound screening campaign, the combination of approaches is key to tackling assay interference.

Box 3

Experimental approaches for tackling assay interference

Colloidal aggregation

Colloidal aggregation is often (but not always) linked with steep hill slopes[16,127,186,187] (see Fig. 2a) that can be detected by dose–response curve analysis. However, steep hill slopes may also arise, for example, from covalent binding.

Detergents are used to disrupt colloidal structures, reduce aggregation formation and prevent nonspecific protein adsorption. Bioactivity lost under detergent-containing conditions is a strong indicator of interference by colloidal aggregation[178,188]. Detergent conditions are also part of counter-screen setups using aggregator-sensitive enzymes such as AmpC β-lactamase[127,178,186].

Screening at lower concentrations (that is, below the critical aggregation concentration) circumvents aggregation (but may lead to the loss of weak binders[36]). Increasing enzyme concentration leads, for low K_d values (as is the case for aggregators), to a linear increase of IC_{50} values[36,187], whereas the IC_{50} values of non-aggregators remain largely unaffected.

Further methods to detect and tackle assay interference by colloidal aggregates include dynamic light scattering (determines the size distribution of particles in the sample[9,127,188,189]), centrifugation (causing aggregates to precipitate[188], hence, bioactivity lost by centrifugation is probably linked to colloidal aggregates), nuclear magnetic resonance (NMR, detecting aggregators by changes in NMR spectral attributes[190]), microscopic methods (detecting aggregators, for example, through direct visualization of aggregators in transmission electron microscopy[188,191]), and surface plasmon resonance (detecting the binding of aggregators to protein, provided that the protein target can be immobilized[192]). For a comprehensive discussion, see ref. 36.

Chemical reactivity

Assay interference linked to chemical reactivity can be addressed with time-resolved measurements (covalent binders commonly show an increasing inhibitory effect over time[193–196]), 'jump-dilution experiments' (involving the drastic dilution of the test compound with the target into a substrate-rich environment, by which reversible binders show a reduction of target inhibition whereas the activity of irreversible binders remains largely stable[197–199]), dialysis[37] (during which reversible binders are effectively removed from the protein target whereas irreversible binders stay bound to the target), use of scavenging agents such as dithiothreitol (which can bind thiol-reactive compounds that otherwise may bind to the thiol groups of cysteine residues[37,40]), ALARM NMR (an NMR-based method detecting the binding of thiol-reactive compounds[9,40,114]), and the horseradish peroxidase–phenol red assay (detecting redox-cycling compounds by changes in the absorbance spectrum of phenol red[37,200]). For a comprehensive discussion, see ref. 37.

Interference in fluorescence-based assays

Fluorescence interference can be addressed by pre-reading fluorescence signals (after incubation of a compound and before the addition of substrate[32]), fluorescence spectroscopic profiling (providing a comprehensive assessment of the fluorescence spectra of test compounds to inform on possible interference[201]), using red-shifted fluorophores (tackling assay interference linked to autofluorescence observed in the blue and green regions of the spectrum[201]), and using fluorophores with long half-lives (enabling measurements after the decay of auto-fluorescence[32,202]). Better signal-to-background ratios can be obtained with kinetic approaches, wherein the assay readouts are registered over time[203]. For a comprehensive discussion, see ref. 32.

Interference in bioluminescence-based assays

Bioluminescence-based assays typically include a luciferase component[204]. Compounds interfering with luciferase are typically detected with luciferase inhibition assays[205]. Orthogonal assays using structurally distinct luciferases (with a limited inhibitor overlap) can help validate compound activity[206]. Coincidence reporter systems in cell-based assays[207–209] can test and validate compound activity by co-expressing (in cells) a pair of luciferases with limited inhibitor overlap. Compounds triggering signals from both reporter enzymes have a higher tendency to exhibit genuine bioactivity. For a comprehensive discussion, see ref. 33.

Interference in homogeneous proximity assays

Homogeneous proximity assays, such as the AlphaScreen, fluorescence resonance energy transfer (FRET), time-resolved (TR) FRET, and bioluminescence resonance energy transfer (BRET) assays, use light-based detection technologies. Hence, methods tackling interference in these assays share commonalities with the spectroscopic interferences (for example, signal quenching and emission) outlined above. Because disruption of the affinity capture components is also a source of interference, changing the capture component may help tackle assay interference while not affecting true hits. For a comprehensive discussion, see ref. 7.

Depending on the specific case, orthogonal assays and counter-screens may also be used in tandem to complement the primary screening assay(s)[9,15]. As a general guideline, when a counter-screen is implemented, compounds showing interference with the assay should be flagged for de-selection or further vetting in an orthogonal assay. When an orthogonal assay is used, compounds active in both assays should be prioritized for progression (Fig. 1). In assay orchestration of increased sophistication, result interpretation will also become more complex and follow the multi-assay design rationale[15].

The commonly used anti-interference strategies in primary screen, counter-screen and orthogonal assays include various assay tuning strategies and kinetic approaches (see Box 3). Screening experts also make extensive use of biophysical methods[39], as well as chemical methods[37,40] (such as scavenging agents to bind chemically reactive compounds). None of these strategies are a panacea against assay

interference, as they all come with limitations concerning throughput, reliability, applicability and costs.

Theoretical approaches for interference detection

Several theoretical approaches are available for identifying potential assay interference events[41] (Fig. 2). Among these, some of the most potent approaches focus on dose–response curve analysis and curve fitting that check for features characteristic of assay anomalies, such as a steep hill slope (Fig. 2a). Other strategies include substructure and scaffold analyses, which can be particularly informative if measured bioactivity data for a set of related compounds or data from different biological assays (which can be translated, for example, into hit rates) are considered[2,42,43]. Many existing rule sets were derived using such approaches.

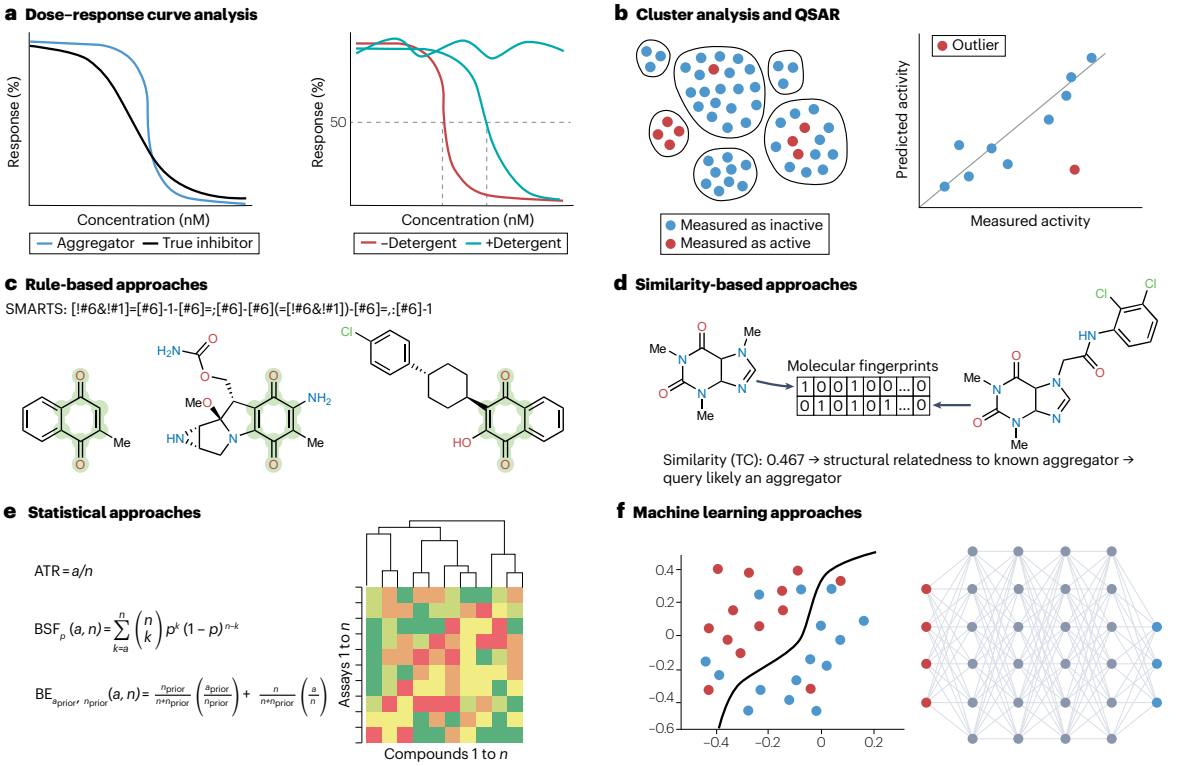


Fig. 2 | Theoretical approaches for detecting and predicting assay interference caused by small organic compounds. a, Automated approaches for interpreting dose–response curves can help to detect compound aggregation and other types of assay interference. Steep hill slopes (left) and changes in dose–response curves upon adding detergent (right) are important indicators of assay interference caused by aggregation of small organic compounds. **b**, Methods for clustering compounds by molecular similarity (left) and QSAR analyses (right) can identify suspicious assay readouts that may require further investigation. **c**, Rule-based approaches encode molecular substructures linked to assay interference with notations such as SMARTS. They are widely used to flag potential nuisance compounds. Shown is a substructure pattern encoding quinones, which have been linked to assay interference via protein reactivity. **d**, Similarity-based approaches compare the molecular similarity of compounds of interest to that of known nuisance (or benign) compounds to predict their likely behaviour in biological assays. In this example, the molecular structures of caffeine (query molecule) and a known aggregator (reference molecule), encoded as molecular fingerprints, are compared using a popular similarity index, the TC (ranging from 0 to 1, with a value of 1 indicating two molecules being identically encoded in the corresponding fingerprint representation). A high degree of molecular similarity suggests that the query molecule will likely behave similarly to the reference molecule. **e**, Statistical methods harness large bioactivity data sets to describe and predict the frequent hitter behaviour of compounds. On the left, three examples of approaches used in frequent hitter prediction are shown: ATR, BSF and BE. On the right, an example of a bioactivity matrix is shown as a heat matrix (red indicating high activity, yellow to orange indicating moderate activity, and green indicating inactivity). **f**, Machine learning approaches represent the main research thrust in assay interference prediction. General models for predicting frequent hitter behaviour, as well as specialized models for predicting specific types of assay interference, are available. Illustrated are the decision boundary of a nonlinear classifier (left) and a deep neural network (right). ATR, active-to-tested ratio; BE, Bayesian estimation; BSF, binomial survivor function; QSAR, quantitative structure–activity relationship; SMARTS, SMILES arbitrary target specification; TC, Tanimoto coefficient. Explanation of the variables in the displayed mathematical formulas: a, number of assays in which a compound was reported as active; n, number of assays in which a compound was measured; p, probability of a compound being active in an assay; a_{prior} and n_{prior}, pseudo-counts for the number of active and total assay measurements reflecting prior knowledge (that is, a_{prior}/n_{prior} should approximate the probability that a compound is active in an assay); k, control variable.

Approaches building on the similar property principle[44,45], which states that structurally related molecules tend to have similar properties, enable experts to identify implausible assay readouts and potential inconsistencies in the data, through statistical means and visualization. Examples of approaches building on this principle include similarity-based and scaffold-based compound clustering[41] (which places structurally related compounds into proximity), quantitative structure–activity relationship modelling[46] (a mathematical approach for predicting the biological properties of compounds based on their molecular structure), matched molecular pair (MMP) analysis[47] (an analytical approach to determine the impact of a single chemical transformation observed for a pair of molecules on their properties) and, its extension, matched molecular series analysis[48] (which determines the impact of a single chemical transformation at a specific site based on a set of molecules).

The Structure–Activity Landscape Index (SALI)[49] puts the similarity of pairs of molecules concerning their molecular structure and biological activity into perspective: molecular pairs with a high SALI indicate activity cliffs[50] (that is, pairs of structurally closely related compounds with distinct levels of bioactivity). Activity cliffs can either be genuine (for example, caused by a steric clash of the compound with residues forming the ligand binding site) or artefacts (for example stemming from assay interference, assay variability, mistakes during the experiment or data transfer, and the limitations of molecular representations[51,52]). A closer look at these outliers is required to interpret their meaning correctly.

In addition to the cheminformatics approaches discussed in this section, models and approaches designed for predicting assay interference can be used to flag suspicious assay readouts. These methods are discussed in the following sections.

Theoretical approaches for interference prediction

Theoretical approaches for assay interference prediction (Fig. 2) can broadly be classified into rule-based approaches (flagging nuisance compounds based on structural patterns), similarity-based approaches (assessing the molecular similarity of compounds of interest to compounds with known behaviour), statistical approaches (leveraging historical measured bioactivity data to assess, primarily, the frequent-hitter behaviour of compounds), and machine-learning approaches (identifying and weighing molecular properties that influence the behaviour of compounds in biological assays). Because measured data and not technology is the current bottleneck in the development and validation of theoretical approaches, we begin the following discussions with a focus on data.

Data sources for model development

In the context of assay interference prediction, corporate databases are particularly precious to model development because they are usually of substantial size, more consistent, and of high density (meaning that a substantial part of the compounds has been measured in the same, extensive set of biological assays). Unfortunately, with the exception of rule sets, models built on this valuable data are rarely, if ever, made accessible to the scientific community.

However, the recent years have seen a strong community effort towards FAIR (findability, accessibility, interoperability and reusability) data[53], open data and open access. This is reflected in the sizable data sets that are now in the public domain, including PubChem BioAssay[54], with more than 115 million compounds and close to 300 million measured bioactivities; the ChEMBL database[55,56], with 2.4 million compounds and more than 20 million measured bioactivities; and the European Chemical Biology Database (ECBD)[57], with more than 100,000 compounds and more than 1.6 million measured bioactivities. In particular, PubChem BioAssay includes substantial amounts of data from counter-screen assays, which are regularly used to develop models predicting aggregators or compounds interfering with fluorescence-based or luciferase-based assays[58–62].

In addition to the large, public bioactivity data sets, several smaller, more focused data sets exist, such as a set of 12,645 known aggregators[63]. These can be difficult to find as they may be 'hidden', for example, in the supporting information of publications. Search utilities such as Google Dataset Search[64] can help to find such hidden data gems. However, a further challenge arises from much of the data being stored in formats requiring substantial processing.

Data biases are of concern when developing in silico models, particularly machine learning models[65]. For example, research articles are generally skewed towards reporting positive results (in this context, compounds measured as active). This bias regularly propagates into data sets derived from these sources, causing data set imbalances in model training and validation. Another critical issue is analogue bias, which stems from drug discovery data often containing congeneric series of compounds. Suppose such congeneric series are split over the training and test sets. In that case, the difficulty presented by the test cases and, hence, the real-world performance of the model will probably be overestimated. This issue is often addressed with a scaffold-split strategy, which enforces the absence of a molecular scaffold overlap between the training and test sets[66]. Of note, there are differing definitions of what constitutes a molecular scaffold, and, depending on the definition, for example, the change of a single atom element can qualify a structure as a distinct molecular scaffold.

Additional challenges arise when combining data from multiple assays, which is useful when developing models for frequent hitter prediction. The composition of a data set will strongly influence the scope and behaviour of these models. Without corrections and consideration of assay types, a model trained on a set of bioassays primarily representing fluorescence-based assays will more probably flag compounds interfering with this assay technology as frequent hitters than compounds causing other types of assay interference. Likewise, data sets with a strong representation of kinase assays may lead to models classifying kinase inhibitors as (undesirable) frequent hitters, even though they may have a favourable multi-kinase inhibition profile. Assay data sets with a strong representation of promiscuous biomacromolecules, such as some cytochrome P450 isozymes or transporters, may result in models behaving differently from those trained on classical drug targets. Research on this topic is still in its infancy.

A further constraint in using assay datasets for model development is that many do not provide sufficient information to assess whether individual assay readouts reflect genuine bioactivities or artefacts. Also, a problematic portion of positive assay readouts is probably related to undetected impurities, which may be sample-specific (see Box 2). Such effects cannot be modeled based purely on molecular structures, which many in silico models exclusively rely on. Together, the insufficiencies of the measured data limit the performance and scope of in silico models. Consequently, it is not uncommon to see that the investments made in expertise and labour for data curation and analysis, as well as in model validation, exceed those made in the actual model-building process.

Rule-based approaches

Rule-based approaches are widely used in drug discovery for filtering or flagging compounds with undesirable properties related to

bioavailability, chemical stability and reactivity, toxicity, metabolism, and assay interference. One of the best-known rule sets in drug discovery is Lipinski's rule of five[67] describing properties relevant to the oral bioactivity of small-molecule drugs. Whereas the rule of five defines thresholds for a set of physicochemical properties, many other rule-based approaches use molecular substructure patterns developed, sometimes over years or even decades of research, into comprehensive 'dictionaries'. The individual patterns are commonly encoded in the SMILES arbitrary target specification (SMARTS) language[68].

Many research groups in industry and academia have published (parts of) the rule sets they use for excluding compounds with undesirable properties in their screening deck. These rule sets focus on identifying chemically reactive compounds and/or frequent hitters[69–75]. One of the most famous collections of substructure patterns is the PAINS rule set, which encodes molecular scaffold substructures linked to different types of assay interference[2,76]. In toxicology, substructural patterns are widely used and referred to as 'structural alerts'[77–79]. Among the most comprehensive pattern collections overall is the ZINC20 pattern set[80,81]. It comprises 1,468 patterns, including the 480 PAINS patterns, 114 patterns for named functional groups, 62 linked to benign reagents (building blocks) and 57 patterns related to electrophilicity.

A key advantage of rule-based approaches over more complex methods is that they can be linked with experimental data, literature or expert information that can provide a rationale for predictions. Another decisive factor for their wide application is that experts may derive them even when data are scarce. Therein also lie disadvantages, though, such as the inherently subjective nature of such rule sets[82] and the fact that their compilation and maintenance can quickly develop into an expensive and challenging task.

Rule sets generally oversimplify chemical and biological properties and cannot account for subtleties in the behaviour and properties of compounds. Consequently, rule-based approaches tend to produce a substantial number of false-positive hits (for example, compounds matching a PAINS pattern that do not interfere with biological assays). Therefore, the unquestioned use of rule sets brings the risk of rejecting substantial numbers of promising, potentially highly innovative compounds[13,43,83–85]. Today, compounds such as acetylsalicylic acid (aspirin) may be excluded from screening decks because of risks associated with chemical reactivity (owing to the potential indiscriminate acetylation of many proteins[86]) and selectivity (owing to its small size). Cyclosporine, a breakthrough immunosuppressant drug preventing the rejection of organ transplants, would be lost with commonly applied molecular weight filters that typically only allow compounds of 600 Da or less. The problem of high false-positive rates can be aggravated when rule sets are used in combination. Only recently have computational approaches for the systematic analysis and comparison of sets of substructural patterns become available[87,88]. On a more technical level, different implementations of rule sets and matching algorithms can yield distinct results. As is the case for any model, the applicability domain (that is, the property and chemical spaces within which a model produces predictions with defined reliability[89,90]) of rule-based approaches must be observed.

Similarity-based approaches

Similarity-based approaches build on the similar property principle. In the context of assay interference prediction, similarity-based approaches are most often used to identify potential aggregators[63]. Compounds are flagged as potential aggregators if their molecular structures (typically represented as 2D molecular fingerprints, see Fig. 2d) are similar to those of any known aggregators. However, as the number of experimentally confirmed and reported aggregators remains small, the applicability of this guilt-by-association approach remains constrained. This is mainly because the absence of similarity to any known nuisance compound (aggregator) does not necessarily indicate benign behaviour of a compound in biological assays.

Statistical approaches

By leveraging historical assay data, typically from large public or corporate bioactivity databases, statistical approaches describe and predict the frequent hitter behaviour of small organic compounds. This data can be structured as a bioactivity matrix wherein rows and columns represent compounds and assays, respectively, and an entry represents the measured activity of a compound in an assay (Fig. 2e). Usually, the measured activities, such as K_i, K_d, IC_{50}, EC_{50}, %activity or %inhibition, are categorized based on a threshold. Of note, such bioactivity matrices are often highly incomplete, as, by far not all compounds are measured in all assays.

The most basic approach to define frequent hitter behaviour rests on bioactivity counts. Bioactivity counts denote the number of assays in which a given compound is reported as active, regardless of the number of assays a compound was tested in and regardless of the relationship between the assays and the target space they represent[71,91,92]. Compounds with a bioactivity count above a predefined threshold are deemed to be frequent hitters. Implicitly, this approach assigns an 'inactive' label to all unknown entries in the activity matrix, justified by the general rarity of bioactivity. In consequence, given the nature of bioactivity matrices, such count-based classifications will more probably label older, better-investigated compounds as frequent hitters than newer ones. Presuming inactivity for unknown matrix entries underestimates the expected number of active entries of a compound. This problem can be addressed by using the active-to-tested ratio (ATR) of a compound and comparing it to a threshold to identify frequent hitters[93–96].

When the assays are independent and identically distributed, the maximum likelihood for the probability of success of a compound converges to the ATR as the number of assay measurements increases[97]. If, however, a compound has been tested only in a few assays, the ATR may diverge substantially from the true proportion. Most approaches avoid this scenario by removing compounds that do not exceed a minimum number of assay tests[93–95,98,99]. This practice, however, leads to the loss of potentially valuable data. Bayesian estimation offers an alternative solution by using a prior ATR to reflect the information that compounds are rarely active in general and have a higher probability of being inactive. The ATR of compounds with limited tests is adjusted towards the prior ATR, whereas the ATR for those with abundant tests remains largely unaffected. This adjustment is implemented through pseudo-counts, wherein two different constants augment the number of active and inactive measurements for all compounds. This technique was adopted in the Badapple methodology[100], although the authors did not precisely follow the underlying theory of the Bayesian estimator in their approach.

The approaches for defining frequent hitter behaviour discussed so far are affected by biases of the target space represented in a bioactivity matrix. This problem is commonly addressed by clustering the targets based on protein families[69] or sequence similarity[94,95,101]. Another bias in frequent hitter prediction is the unusual hit rates of some assays (for example, cytochrome P450 assays). For example, when a compound is tested in an assay with an unusually high hit rate,

it has a higher tendency to be labelled as a frequent hitter than compounds not tested in that assay. One approach to address this problem is to repeatedly shuffle the compound activity labels within each assay (effectively randomizing bioactivities while preserving the number of active compounds) and recompute the ATR values of the compounds on the rearranged data[96]. The upper confidence interval (95%) of the shuffled ATR values indicates the ATR that can be achieved by chance. A compound is deemed a frequent hitter if the ratio calculated from the observed ATR and the upper confidence interval exceeds a predefined threshold. In addition to mitigating the issues related to assays with high hit rates, this approach prevents compounds from being identified as frequent hitters by chance owing to limited measured bioactivity values.

Instead of estimating the ATR of a compound, the binomial survivor function (BSF) focuses on the average ATR observed on an extensive collection of in-house substances[102]: For a given compound, a binomial test is conducted to check whether the observed number of active measurements is improbable given the fixed ATR and the total number of measurements. However, it was shown empirically that the BSF only performs sufficiently well for compounds already tested extensively[3]. Moreover, a recent study has found that the BSF does not adequately represent the PubChem bioactivity matrix regarding rank-order statistics when using a single probability of being active[99]. Incorporating different assay hit rates with a Poisson–binomial model improved the results only marginally. These findings challenge the approximation that compounds behave like binomial variables with independent and identically distributed assays. It was suggested that compound-based models using molecular properties may better represent the bioactivity matrix. On the basis of this insight, machine learning models incorporating compound similarity are the next logical step to enhance frequent hitter detection further.

Machine learning approaches

Machine learning approaches are unrivalled in their ability to identify and learn subtleties in molecular structures and complex, nonlinear relationships between molecular features and the behaviour of compounds in biological assays. They represent the main research thrust in computational assay interference prediction.

Unlike rule-based approaches, machine learning enables the consideration of the larger molecular context. In comparison to similarity-based approaches, machine learning goes beyond quantifying global molecular similarity by assigning weights to the individual molecular features.

Machine learning approaches are particularly data-hungry. Conventional machine learning approaches, such as the k-nearest neighbours algorithm, random forests and support vector machines, become valuable when trained on measured data for at least a few hundred compounds (with molecular diversity of the training set being a decisive factor). Their full potential generally comes into effect only with large, diverse data sets. The need for large amounts of training data is even greater for deep learning approaches, which are often found, with the currently available data sets, not to outperform less complex machine learning methods.

Knowledge-sharing approaches, such as transfer learning, multitask learning and meta-learning, are increasingly used to mitigate the limitations imposed on machine learning by the chronic shortage of data. These approaches use information from related tasks and data sets to improve the prediction performance[103].

Besides data quantity, data quality[104] is of major concern to machine learning. Poor data quality will lead to poor model performance, aptly referred to as 'garbage in, garbage out'[105]. Biases in the data can lead to underperforming models. They may also lead to the overestimation of model performance (see 'Data sources for model development' and ref. 65).

Machine learning approaches are technically more complex than the other approaches, and the models and predictions are generally more difficult to interpret (although methods for explainable artificial intelligence are maturing and becoming more widely used[106]). The problem of model complexity and interpretability is one reason that machine learning approaches can have a more difficult standing, such as in the regulatory context. Conversely, machine learning models can distil unknown patterns and structure–property relationships from complex chemical and biological data[74,107,108].

As for any type of model, users are strongly advised to consider the applicability domain of the machine learning approaches. For example, the generalization of models for frequent hitter prediction[94,95] and compound fluorescence prediction[59] is limited, meaning that predictions for new chemistry may not be reliable. Unfortunately, even today, models published without accompanying applicability domain definitions are still commonplace. Furthermore, many publicly available models do not provide information on the reliability of the individual predictions. For users, this means they may be unable to understand how much (if at all) they can trust in a prediction.

Publicly available in silico models

In this section, we discuss publicly available models for assay interference prediction. An overview of all discussed models, including technical information and availability, is provided in Table 1. Recommendations for the use of in silico models are offered in Box 4.

Rule sets for flagging compounds

The PAINS rule set[2] has evolved into one of the most widely used tools for flagging small molecules that have an increased risk of causing interference in biological assays. It is composed of 480 SMARTS patterns that encode molecular substructures from compound classes for which experiments indicate an increased likelihood of assay interference. The patterns represent a variety of assay interference mechanisms, including covalent modification (such as by isothiazolones or ene-rhodanines), redox cycling (such as by catechols and toxoflavin), membrane disruption (such as by curcumin), and metal complexation (such as by hydroxyphenyl hydrazones)[8]. However, the scope and accuracy of the PAINS concept and rules are limited: The concept builds primarily on the results and observations from screening a medium-sized molecular library of approximately 93,000 compounds against six protein–protein interfaces using a single assay technology, the AlphaScreen. Assay interference observed in the AlphaScreen is often, but not always, indicative of the risk of assay interference in other assay systems[18]. Prior to screening, highly reactive compounds were removed from the library. Hence, the PAINS patterns have limited representation of reactive compounds[2,75]. The screening was generally performed at high concentrations, meaning that the PAINS rules may overemphasize the occurrence of assay interference[76]. Moreover, detergent was added to minimize the aggregation risk, meaning that the representation of colloidal aggregation in the PAINS patterns is limited[76].

The PAINS patterns are derived from a synthetic compound library. Hence, application to natural products requires additional

Table 1 | In silico tools for assay interference prediction

Name	Approach	Description	Model training and testing	Year	Web services, software	Refs.
Rule sets for flagging potential nuisance compounds						
PAINS rule set	Rule-based	480 rules encoding scaffolds and substructures linked various mechanisms of assay interference; not designed to cover aggregators and highly reactive moieties	Trained on an in-house screening library of approximately 93,000 compounds; evaluated on in-house data	2010	Original[2] Others[81,117,144–149]	2,76
Rule set for frequent hitters in AlphaScreen assays	Rule-based	25 SMARTS patterns encoding substructures interfering with chemistry components or His-tagged proteins in AlphaScreen assays; designed as an extension of the PAINS rule set	Trained and evaluated on in-house and public data sets	2014	Original[75,79]	75
Rule set for frequent hitters of GST–GSH interaction	Rule-based	34 SMARTS patterns encoding substructures interfering with GST–GSH interaction in AlphaScreen assays; designed as an extension of the PAINS rule set	Trained and evaluated on in-house and public data sets	2016	Original[79,113]	113
ALARM NMR Rules	Rule-based	175 SMARTS patterns encoding substructures linked to thiol reactivity	Trained and evaluated on in-house data	2005	Original[114] Others[117]	114,115
BMS rules	Rule-based	191 SMARTS patterns encoding functional groups linked to frequent hitter behaviour	Trained and evaluated on in-house data	2006	Original[71] Others[117,149]	71
GSK CTC rules	Rule-based	28 SMARTS patterns linked to frequent hitter behaviour	Trained and evaluated on in-house data	2018	Original[72] Others[149]	72
REOS rules	Rule-based	More than 200 SMARTS patterns plus a physicochemical property filter for filtering undesirable compounds	Trained and evaluated on in-house data	1998	Original[116] Others[144]	116
Brenk rules	Rule-based	More than 100 SMARTS patterns covering substructures and functional groups undesirable in the drug discovery context; not focused on assay interference prediction	Trained and evaluated on in-house data	2008	Original[70] Others[147,149]	70
Lilly Rules	Rule-based	275 SMARTS patterns covering substructures and functional groups undesirable in the drug discovery context; not focused on assay interference prediction	Trained and evaluated on in-house and external data	2012	Original[150] Others[144]	69
NIBR rules	Rule-based	444 SMARTS patterns collected from public and in-house sources for hit identification and triaging	Trained and evaluated on in-house and public data	2020	Original[151] Others[152]	43
Models for predicting frequent hitters						
Badapple	Statistical	Assigns a promiscuity score based on historical assay data for molecular scaffolds; considers molecular scaffolds only and not any decorations	Trained on nearly 400,000 compounds measured in a total of 822 HTS assays; evaluated by cross-validation and additional retrospective analyses	2016	Original[153]	100
Models classifying PAINS pattern-matching compounds	Machine learning	Set of RF, SVM and DNN models discriminating PAINS pattern-matching compounds behaving as frequent hitters from those acting as DCM; molecules represented as ECFP4 or MACCS structural keys	Trained on 5,223 PAINS pattern-matching frequent hitters and 3,059 PAINS pattern-matching, consistently inactive compounds collected from PubChem BioAssay; evaluated on a test set generated by random split	2018	Original[154]	92
Models predicting promiscuity cliffs	Machine learning	Set of k-NN, SVM, RF and DNN models discriminating MMPs representing promiscuity cliffs from those not representing such cliffs; molecules represented as ECFP4; model performance and behaviour found sensitive to the definition of promiscuity cliffs	Trained on a few hundred promiscuity and non-promiscuity cliff MMPs extracted from PubChem BioAssay and ChEMBL; evaluated on a test set generated by random split	2021	Original[155]	108
Hit Dexter	Machine learning	Comprehensive assay interference prediction platform including a set of ANNs for discriminating non-promiscuous, promiscuous and highly promiscuous compounds; dedicated models for target-based and cell-based assays; molecules represented as Morgan fingerprints; characteristic of fingerprint-derived models, the extrapolation capability of these models is constrained	Trained on more than 250,000 compounds measured in at least 100 assays for distinct proteins; evaluated on holdout data consisting of approximately 25,000 compounds, plus a DCM data set (approximately 20,000 to 40,000 compounds) and a set of approximately 1,000 known bad actors	2021	Original[149]	95

Table 1 (continued) | In silico tools for assay interference prediction

Name	Approach	Description	Model training and testing	Year	Web services, software	Refs.
Models for predicting frequent hitters (continued)						
OCHEM model for frequent hitter prediction in AlphaScreen assays	Machine learning	Consensus model of ANNs for predicting frequent hitters in AlphaScreen assays; models use a variety of 2D and 3D descriptors	Trained on approximately 350,000 compounds measured during 15 AlphaScreen HTS campaigns; evaluated by cross-validation plus on external test sets	2022	Original[156]	96
Models for predicting colloidal aggregators						
Aggregator Advisor	Similarity-based	Flags compounds structurally related to known aggregators or with log *P* values above 3 for experimental evaluation of their aggregation behaviour; the absence of an alert is not an indicator of benign compound behaviour in biological assays	Trained on a public data set of approximately 12,000 known aggregators; tested in a prospective study of 40 compounds	2015	Original[157] Others[145]	63
ChemAGG	Machine learning	XGBoost classifier trained on ECFP4; because drugs and drug-like compounds are used to present non-aggregators, there may be an increased risk of missed aggregators with this approach	Trained on approximately 12,000 aggregators represented by the Aggregator Advisor data set (see above) plus approximately 24,000 presumed non-aggregators represented by a subset of drugs and drug candidates; evaluated by cross-validation and on holdout data	2019	Original[139]	126
SCAM Detective	Machine learning	RF classifier trained on Morgan fingerprints; warns users about predictions outside the applicability domain; visualization of the molecular fragments contributing to the classification of a compound as (non-) aggregator	Trained on approximately 60,000 non-aggregators and putative aggregators from an AmpC β-lactamase quantitative HTS screen and approximately 50,000 compounds measured in a cruzain quantitative HTS, extracted from PubChem BioAssay; evaluated on large sets of holdout data	2020	Original[128]	58
DeepSCAMs	Machine learning	Feedforward neural network trained on Morgan fingerprints and physicochemical properties	Trained on DLS data for 916 compounds measured at consistently 30 μM concentration; prospective validation based on 65 compounds, in competition with a panel of medicinal chemists	2021	Original[158]	129
Isometric Stratified Ensembles (ISEs)	Machine learning	Consensus modelling strategy predicting (non-) aggregators at different consensus and applicability domain levels	Trained on three sets of up to approximately 190,000 compounds measured in HTS, extracted from PubChem BioAssay; tested on large internal and external test sets	2022	Original[159]	62
Models for predicting interference caused by chemical reactivity						
Xenosite reactivity	Machine learning	DNN model predicting assay interference caused by chemically reactive compounds; can be combined with other models[134] to consider metabolic activation	Trained on 2,803 reactive and non-reactive compounds; tested by cross-validation and with an external test set of 14 compounds	2016	Original[130]	130
Models for predicting interference with specific assay technologies						
InterPred	Machine learning	Set of random forest classifiers for predicting assay interference caused by autofluorescence or luciferase inhibition; trained on 2D molecular and physicochemical descriptors; individual models for representing different assay conditions	Trained on 8,305 compounds included in the Tox21 data set and measured in counter-screen assays under different conditions; evaluated on external sets of 22 to 258 compounds	2020	Original[135,160]	46
ChemFLuo	Machine learning	Two XGBoost models predicting blue and green autofluorescence; trained on ECFP4; known to have limited extrapolation capability; model performance may also be impacted by the limited quality of the training data	Trained on approximately 40,000 and 50,000 compounds from PubChem BioAssay, measured in blue and green autofluorescence counter-screens, respectively; evaluated on approximately 28,000 and 323,000 compounds from PubChem BioAssay, respectively (from distinct assays)	2021	Original[139]	59

Table 1 (continued) | In silico tools for assay interference prediction

Name	Approach	Description	Model training and testing	Year	Web services, software	Refs.
Models for predicting interference with specific assay technologies (continued)						
Luciferase Advisor	Machine learning	Consensus model of four associative neural network models; trained on a large set of physicochemical descriptors	Trained on approximately 323,000 compounds measured in different luciferase inhibition assays (from PubChem BioAssay); evaluated primarily by cross-validation	2018	Original[161]	60
ChemFLuc	Machine learning	XGBoost model for predicting luciferase inhibition; trained on ECFP4	Trained on approximately 366,000 compounds measured in different luciferase inhibition assays (from PubChem BioAssay); evaluated on data set from PubChem BioAssay (from distinct assays)	2020	Original[139]	61
Comprehensive platforms for predicting assay interference						
Hit Dexter	Various	Features machine-learning models for frequent hitter prediction, similarity-based approaches for aggregator and DCM identification, and rule sets for flagging nuisance compounds	Trained and evaluated primarily on data from PubChem BioAssay	2021	Original[149]	95
LiabilityPredictor	Machine learning	Featuring machine learning models for assay interference prediction related to thiol reactivity, redox reactivity, luciferase inhibition and aggregation (from SCAM Detective)	Trained on multiple data sets, most of which consist of approximately 5,000 compounds from NPACT (ref. 162); evaluated by external cross-validation plus prospective experimental validation with 256 compounds per assay	2023	Original[163]	107

ANN, artificial neural network; BMS, Bristol–Myers Squibb; CTC, cleaning the collection; DCM, dark chemical matter; DLS, dynamic light scattering; DNN, deep neural network; ECFP4, extended connectivity fingerprints with a width of four bonds; GSH, glutathione; GSK, GlaxoSmithKline; GST, glutathione S-transferase; HTS, high-throughput screening; k-NN, k-nearest neighbours; MACCS, molecular access system; MMP, matched molecular pair; NIBR, Novartis Institutes for BioMedical Research; NMR, nuclear magnetic resonance; OCHEM, Online Chemical Modelling Environment; PAINS, pan-assay interference compounds; REOS, rapid elimination of swill; RF, random forest; SAR, structure–activity relationship; SCAM, small, colloidally aggregating molecules; SMARTS, SMILES arbitrary target specification; SVM, support vector machine.

considerations[25]. Moreover, the experimental support of individual PAINS patterns varies greatly. Of the 480 SMARTS patterns, only 18 are derived from more than 100 compounds; most are supported by fewer than three compounds[2,30]. The significance of PAINS pattern matches for individual compounds is often limited, particularly because scaffold decorations can substantially affect the behaviour of small molecules in biological assays[109]. Consequently, focusing on a subset of crucial patterns is advisable[43].

Many of the limitations of the PAINS concept are disclosed and discussed in the primary literature[2,76]. They are also part of an ongoing critical debate[13,19,22,30,84,85,110]. In particular, the relationship between the PAINS and frequent hitter concepts has been the subject of analyses, discussions and some confusion[16,30,76,96,102,111]: Because the presence of PAINS patterns in individual compounds is merely a risk indicator for assay interference, this means that by far not all compounds matching PAINS patterns cause assay interference or show frequent hitter behaviour. Therefore, it is expected that comparative analyses of historical assay data regarding the frequent-hitter behaviour of compounds that match PAINS patterns versus those that do not find a strong link between the presence of the patterns and frequent-hitter behaviour (see ref. 76 for an excellent discussion).

The PAINS rule set was initially published as Sybyl line notations[2]. The rules were later translated into the more widely used SMARTS language and are accessible via software packages, a KNIME (ref. 112) workflow and web services (Table 1). Some groups have published additional rules designed to complement the PAINS rule set[75,113]. Users should be aware that translating substructural patterns may introduce subtle, technology-related changes, which may cause inconsistent matchings.

Besides PAINS, several other rule sets are relevant to assay interference prediction and detection. In 2005 and 2007, researchers from Abbott Laboratories published 175 patterns ('ALARM NMR Rules') encoding substructures that have been linked to causing false-positive readouts owing to their reactivity towards thiol groups[114,115]. During this period, a research team from the Bristol–Myers Squibb (BMS) Pharmaceutical Research Institute published a set of 180 SMARTS patterns encoding functional groups they linked to frequent-hitter behaviour[71]. More recently, scientists from GlaxoSmithKline (GSK) R&D Pharmaceuticals published a set of 28 SMARTS patterns ('GSK Cleaning the Collection filters') encoding substructures linked to frequent hitters present in the GSK compound collection[72].

Rules sets used as screening deck filters typically cover a wider range of undesirable moieties. Well-known examples of such screening deck filters include the 'rapid elimination of swill' filters[116], a collection of more than 200 patterns typically used in tandem with physicochemical property criteria, the 'Brenk filters', a collection of more than 100 patterns[70], the 'Lilly Rules', which were developed at the Lilly Research Laboratories over almost two decades of research[69] and incorporate some of the above-mentioned BMS rules, and the 'Novartis Institutes for BioMedical Research (NIBR) substructure filters'[43].

Box 4

Recommendations for working with in silico models for assay interference prediction

To make the best use of the existing in silico models for assay interference prediction, researchers should familiarize themselves with the models by carefully reading the relevant scientific publications and documentation. Validation studies, even those published in peer-reviewed journals, should always be critically scrutinized (see ref. 210 for a contemporary discussion of how, in particular, machine learning models should be evaluated).

Users should ensure that the input for predictions (usually molecular structures) is prepared according to the models' requirements (for example, consideration of protonation states).

Rather than relying on a single model, users should carefully select a set of models to combine. In particular, integrating models trained on diverse data sets and data types, along with models utilizing independent representations of molecular and biological properties and distinct modelling algorithms will probably deliver a more accurate and comprehensive picture of the assay interference risk for a compound, based on which better-informed decisions can be made.

Users should observe the scope and applicability domains of the individual models concerning the coverage of chemical space and types of assay interference. In addition, they should consider the estimated reliability of the predictions for individual compounds. Extra caution should be exercised when using models that do not provide information on their applicability domain, the reliability of individual predictions, or detailed information on the data sets on which they were trained and validated. Whenever possible, predictions should be compared with existing measured data on structurally related compounds to develop an understanding of the behaviour of a model in the specific chemical subspace of interest.

Users should consider predictions as risk indicators rather than evidence for the bad or benign behaviour of a compound in biological assays. Accordingly, predictions should generally be used for flagging compounds but not rejecting them.

The software license terms require consideration, as well as data security and privacy aspects, particularly when using web services.

The definitions of all rule sets mentioned herein are publicly available (see Table 1). Scopy, a recently developed Python library, provides easy access to most rule sets[117]. Several web services are available that allow the filtering of molecular libraries using such rules (see Table 1).

Models for predicting frequent hitters

The available models for frequent hitter prediction are mostly machine learning models[3,92–95,108,118,119]. They are all trained on historical assay data, wherein ideally, many structurally diverse compounds are tested against many assays representing a diverse set of biomacromolecules. In reality, the principal limitation of the existing data sets is the sparsity of the bioactivity matrix. However, the data sets available for modelling frequent-hitter behaviour typically consist of several hundred thousand compounds screened in hundreds of biological assays and, hence, are more extensive than those commonly used for modelling specific types of assay interference.

Most models for frequent-hitter prediction do not consider or distinguish the mechanisms underlying frequent-hitter behaviour or take assay specifics into account. However, some approaches use dedicated models for target-based and cell-based assays[95]. Other approaches aim to distinguish frequent hitters related to assay interference and those related to the genuine modulation of multiple targets[120].

Badapple is a statistical model for frequent-hitter prediction[100]. It is derived from a set of nearly 400,000 compounds measured in 822 assays from PubChem. For a compound of interest, Badapple calculates a promiscuity score related to compounds, assays and samples. The score design and a scaffold aggregation strategy aim to increase the robustness of the method. During a fivefold cross-validation, the Pearson correlations obtained for predictions with Badapple and measured data were around 0.90. Several research papers, including refs. 9,121, report using Badapple successfully, usually in combination with other approaches, such as rule sets, to assess the likelihood of compound promiscuity and assay interference.

The Bajorath group investigated compound promiscuity, frequent-hitter behaviour and the PAINS concept from various angles, considering information on the molecular properties of compounds, their protein interaction partners and interaction patterns[122,123]. They published a set of conventional machine learning models and deep neural networks that use molecular fingerprints for discriminating PAINS-matching compounds showing frequent-hitter behaviour from those consistently reported as inactive in biological assays (that is, compounds consistent with the dark chemical matter (DCM) concept)[92]. In all, 74 PAINS patterns are represented in both classes of compounds. This highlights the relevance of the molecular context in which PAINS-linked substructures are embedded to the behaviour of a molecule in biological assays. On a test set generated by random split, the individual models obtained Matthew's correlation coefficients (MCCs) of around 0.65. The MCC is one of the most robust metrics for the performance of binary classifiers; it ranges from −1 to +1, with values above 0.4 generally accepted as an indication of good classification performance.

Another set of machine learning models discriminate MMPs that represent promiscuity cliffs (that is, distinct levels of compound promiscuity) from those that do not show such cliffs[108]. For model development, a few hundred promiscuity and non-promiscuity cliff MMPs were extracted from PubChem BioAssay and the ChEMBL database and represented by molecular fingerprints. On test sets generated by random split, the conventional machine learning models and deep neural networks trained on these MMPs consistently outperformed models trained on single compounds, with MCC values of up to 0.56 compared with 0.42. Again, differences related to the machine learning algorithms used were minor. Instead, and as expected, model performance strongly depended on the definition of promiscuity cliffs.

Hit Dexter[95] is a comprehensive web platform for assay interference prediction. The core of this platform consists of artificial neural network models for frequent-hitter prediction trained on large subsets of PubChem BioAssay. Dedicated models for biochemical assays and cell-based assays are available. Notably, the bioactivity data was clustered based on protein sequence similarity to avoid compounds being flagged as frequent hitters simply because they show genuine activity on multiple related proteins (this is relevant, for example, to protein kinase inhibitors when tested in multiple protein kinase assays as their selectivity is typically limited). When evaluated on holdout data (that is, data not used during any stage of the model development process), the most effective models achieved MCC values of up to 0.65, signifying a very strong performance. However, like many machine learning models derived from molecular fingerprints, the accuracy of predictions was considerably lower for test compounds structurally dissimilar to those represented in the training data. Therefore, users should always consider the compound-specific reliability indicators provided with predictions. An important finding of this work is that biochemical and cell-based assays show distinct frequent-hitter behaviour and require dedicated models.

An earlier version of Hit Dexter[94] was used to profile several compound libraries for the presence of potential frequent hitters. The models classified no more than 2.5% of the Enamine HTS Collection[124] (a popular commercial screening library) as frequent hitters, suggesting that the compound collection is well-curated. The models also classified no more than 14% of known aggregators and a similar percentage of PAINS pattern-matching compounds and natural products as frequent hitters. The low recovery rates observed for these subsets of potentially problematic compounds reflect that many interference events are specific to testing environments and not necessarily frequent (the behaviour the models are trained to detect). Interestingly, predictions of the proportion of frequent hitters among approved drugs were in the same range as for known aggregators and PAINS pattern-matching compounds (11% to 13%). Among the natural products, flavonoids (which include quercetin mentioned in 'Introduction' and are known as prone to causing assay interference[125]) stood out, with up to 73% classified as frequent hitters. These results indicate that the machine learning models can help flag potentially problematic compounds. Still, many valuable compounds would be lost, and a substantial proportion of likely problematic compounds would remain undetected if they were trusted blindly.

Recently, machine learning models specialized in predicting compounds showing frequent-hitter behaviour in AlphaScreen assays were released[96]. The models utilize conventional machine learning algorithms and different types of neural networks. They were trained on approximately 350,000 compounds measured during 15 AlphaScreen HTS campaigns (data originating, again, from PubChem BioAssay). A consensus approach built on several of the neural network models yielded an averaged balanced accuracy of 85% during fivefold cross-validation with different data sets (balanced accuracy is used to quantify the performance of a classification model when dealing with imbalanced data; in this work, frequent hitters accounted for approximately 1% to 2% of all compounds).

Specialized models

Specialized models focus on predicting a specific type of assay interference or on simulating the interference behaviour for a specific assay technology. They are typically derived from smaller, focused data sets than from the models for frequent-hitter prediction.

Models for predicting colloidal aggregators. The development of models for predicting colloidal aggregators is hampered by the lack of available negative data, that is, compounds confirmed to not be aggregators. Researchers have explored various strategies to work with this limitation. For example, Aggregator Advisor[63] measures the molecular similarity (based on molecular fingerprints) to approximately 12,600 experimentally confirmed aggregators. Any compound similar to a known aggregator or having a calculated log*P* above 3 is recommended to undergo experimental aggregation testing. Aggregator Advisor was successfully tested in a prospective study of 40 compounds[63] and has been widely adopted as a tool to assess the risk of aggregation. Importantly, the absence of structural similarity to any known aggregator does not indicate the benignity of a compound in biological assays (see 'Similarity-based approaches' for details).

ChemAGG (ref. 126) is a conventional machine learning model (XGBoost classifier) trained on molecular fingerprints. It uses the same data set as Aggregator Advisor to represent the positive class (that is, aggregators) but also uses a set of drugs and drug candidates to represent the negative class (that is, non-aggregators). Because several drugs and drug candidates are known to be prone to aggregate formation[127], the modelling setup implies that there may be an increased risk of false-negative predictions, meaning that aggregators may remain undetected (although the authors undertook steps to minimize the number of aggregators present in the data set representing the negative class of compounds).

Alves et al. tried to compile cleaner data sets of aggregators and non-aggregators by comparing the behaviour of compounds during HTS with three pairs of assays, each differing primarily by the presence or absence of a detergent[58]. Following this approach, the researchers constructed data sets of approximately 14,000 to 26,000 potential aggregators and up to ten times as many non-aggregators. Notably, the three data sets represent HTS campaigns using distinct assay conditions, which can help investigate the dependency of aggregate formation on assay setups. Following a data balancing procedure, random forest models and deep neural networks were trained on molecular fingerprints. On holdout data, the models reached 66% to 77% accuracy. A web service ('SCAM Detective')[128], offering free access to the random forest models, alerts users about any queries outside the applicability domain of a model. Similarity maps visualize which molecule fragments contribute to the classification of a compound as aggregator or non-aggregator.

Most existing models for aggregation prediction neglect a critical factor in aggregation, which is the concentrations of the test compounds in the assays. To overcome this limitation, the neural network model DeepSCAMs[129] is trained on compounds measured by dynamic light scattering (DLS) consistently at 30 μM concentration. Following cross-validation experiments, the aggregation behaviour of 65 purchasable compounds was predicted and subsequently tested in DLS screens. For 80% of these compounds, the predictions were in agreement with the DLS measurements. In comparison, 15 PhD-level medicinal chemists presented with the 65 compounds correctly classified 61% of these compounds, underlining the value of the computational method.

Recently, a widely applicable consensus modelling strategy, Isometric Stratified Ensembles (ISEs), was used for modelling colloidal aggregation[62]. Depending on the specific test setup and test data set, area under the receiver operating characteristic curve (ROC-AUC) values of up to 0.82 for discriminating aggregators and non-aggregators were obtained (the ROC-AUC quantifies the performance of a model

across different classification thresholds). The effectiveness of the approaches was evaluated by comparing it with existing models, albeit not on the identical test set. Consequently, a definitive assessment of the merit of this approach in comparison to existing methods is still pending. The complexity of the ISE approach implies that its interpretability is limited.

Models for predicting chemical reactivity. Assay interference caused by chemically reactive moieties is covered, to some extent, by rule sets and models for frequent-hitter prediction. A model specialized in identifying chemically reactive compounds is Xenosite reactivity[130]. This deep convolutional neural network predicts the probability of compounds reacting with nucleophilic trapping agents, such as cyanide and glutathione, and with biomacromolecules (that is, proteins and DNA). The model was trained on 2,803 non-reactive and reactive compounds. During tenfold cross-validation, it reached ROC-AUC values close to 0.80 in identifying reactive compounds. Xenosite reactivity can also predict sites of reactivity with high accuracy, as demonstrated in several prospective validation studies[131–133]. A follow-up work from the developers of Xenosite reactivity has shown that the model is especially valuable when used in tandem with the PAINS rule set[74]. Moreover, the model can be combined with other models[134] to consider metabolic activation (which may precede chemical reactions).

Models for specific assay technologies. Several machine-learning approaches for detecting and predicting interference with specific assay technologies have been reported. Publicly available models include InterPred[135], ChemFLuo[59], ChemFLuc[61] and Luciferase Advisor[60,96]. A recurring theme in the validation studies for these tools is the sensitivity of the performance of models on assay setups, conditions and readout types.

InterPred includes a collection of random forest classifiers for autofluorescence and luciferase inhibition prediction. The models were trained on 8,305 compounds (from the Tox21 (ref. 136) dataset) measured in counter-screen assays under different assay conditions (cell culture conditions, cell types) and assay readout channels (blue, green and red wavelengths), with molecules represented by 2D molecular and physicochemical descriptors. On holdout data, the models for autofluorescence prediction reached MCC values of 0.44 to 0.67 (with lower values obtained for the blue channel). The model for luciferase inhibition prediction reached an MCC of 0.57.

ChemFLuo consists of two classifiers for identifying blue and green fluorescent compounds. The technical approach behind ChemFLuo is closely related to that behind ChemAGG. Individual XGBoost classifiers were trained on measured autofluorescence data for approximately 40,000 compounds (blue fluorescence) and approximately 50,000 compounds (green fluorescence), respectively. On external data for blue fluorescence (approximately 28,000 compounds), the best model obtained an MCC of 0.34. However, all models showed poor performance on external data for green fluorescence (approximately 323,000 compounds), which the authors attribute to issues related to data quality and a limited applicability domain.

ChemFLuc is closely linked conceptually to ChemAGG and ChemFLuo. It is an XGBoost classifier trained on molecular fingerprint representations of a large set of known luciferase inhibitors and noninhibitors (retrieved from PubChem BioAssay). A model generated with this setup obtained an MCC of 0.37 on a test set of 1,571 luciferase inhibitors and 56,909 noninhibitors (also retrieved from PubChem BioAssay). The authors merged all available data sets for the final model to use 26,385 luciferase inhibitors and 339,646 noninhibitors for training.

Luciferase Advisor is a consensus model comprising four neural network models. The models were trained on more than 323,000 compounds with measured luciferase inhibition data (retrieved from PubChem BioAssay; inhibitors account for 3.3% of the data set). The molecular structures were represented by large sets of physicochemical descriptors, among which 3D descriptors were important to prediction success. Luciferase Advisor reached a balanced accuracy of 0.90 during fivefold cross-validation.

Comprehensive platforms

Today, comprehensive in silico platforms for assay interference prediction are starting to become available. The Hit Dexter platform includes a set of machine learning models for identifying and predicting compounds that are likely to exhibit frequent-hitter behaviour in target-based or cell-based assays (see 'Models for predicting frequent hitters'), two similarity-based approaches for comparing the structures of compounds of interest with those of known aggregators (from Aggregator Advisor) and DCM (from ref. 137), and a pattern-matching engine which includes many of the rule sets discussed previously in 'Rule-based approaches'.

LiabilityPredictor[107] is a web platform integrating the previously developed SCAM Detective models for colloidal aggregation (see 'Models for predicting colloidal aggregators') plus a new collection of machine learning models for predicting assay interference linked to thiol reactivity, redox reactivity and luciferase (firefly and nano) inhibition. The new models were trained on data sets of approximately 5,000 compounds from the NPACT Chemical Library[138], represented by molecular fingerprints. In addition to external cross-validation, the models were subjected to prospective validation using sets of 256 compounds per interference mechanism (assay). For the subsets of compounds within the applicability domains of individual models, balanced accuracy was between 58% and 78%.

Recently, ChemAGG, ChemFLuo and ChemFLuc, all from the same developers, have been integrated into one consistent platform, ChemFH (ref. 139).

Computational approaches in practice

Today, most rule-based approaches rely on physiochemical properties or SMARTS patterns computed and handled with RDKit, one of the most popular, feature-rich and robust cheminformatics software libraries[140]. RDKit can be accessed from Python and via KNIME, a widely used data science platform enabling visual (meaning code-free) programming. Most similarity-based approaches also build on these popular components.

The more complex tools come with additional code and depend on several software packages, for example, for machine learning. Although most software packages have permissible licenses and are easy to install, getting the prediction tools to run successfully can be challenging. Most existing tools result from academic research, meaning their robustness, functionality, platform and user support may not reach the levels expected from a commercial product. The web services provided for many existing tools mitigate most issues encountered with local installations. However, data security and privacy aspects may be of concern when using web services. Users should consider the recommendations for working with in silico models outlined in Box 4.

Integrating measurement and computation

Assay interference poses substantial challenges to drug discovery. Owing to limitations in prediction accuracy and applicability, in silico methods, just like any experimental method, should not be used blindly as a positive or negative compound filter. The most effective strategy to address these challenges is fully integrating the existing theoretical and experimental approaches into the early drug discovery effort (Fig. 1). With rule-based approaches, this is already a reality. They are regularly used in designing screening decks (that is, excluding compounds with undesirable molecular moieties or physicochemical properties), in hit finding and hit triage, as exemplified in recent reports. There is a consensus in the scientific community that a high level of caution is needed in using such rule sets. The essence is to keep these rule sets small, simple and interpretable, and to subject them to continuous evaluation and further development.

Similarity-based approaches are also popular and regularly used in drug discovery projects, as Aggregator Advisor exemplifies (more than 300 citations). By contrast, the more complex and recently introduced statistical and machine learning approaches have only just started to receive interest from the wider drug discovery community.

In addition to their use in screening deck design and hit triage, the in silico predictors may also be used for profiling compound libraries, for example, to inform the design of screening campaigns or provide compound-specific risk assessments. Furthermore, the use of theoretical approaches for online monitoring of screening processes (see 'Theoretical approaches for interference detection' and Fig. 1) is highly recommended. The integrated setup enables swift action on potential assay readout anomalies.

Conclusion and outlook

A substantial proportion of small organic compounds, including marketed drugs, can trigger false readouts in biochemical or cell-based assays through various interference mechanisms, with aggregation being the most prevalent. Importantly, assay interference results from the intricate interplay of compound properties and specific assay components. This implies that compounds inducing interference in one assay may show benign behaviour in other assays. Although assay interference can impede the directed optimization of compounds and, hence, pose substantial challenges to early drug discovery, it is generally not detrimental to the value of a drug.

Experimental approaches and strategies for preventing, detecting and mitigating false assay readouts have come a long way. In addition to these resources, a large number of theoretical approaches for assessing and predicting the compound-specific risk of assay interference are now at our disposal. Predictive tools build on straightforward rule-based, similarity-based and statistical approaches, as well as machine learning approaches of different levels of complexity.

For several reasons, it is difficult to formulate globally valid statements about the performance and applicability of the existing in silico tools. Often, the ground truth is unknown. For example, the PAINS patterns match approximately 5% of the approved drugs[76], and Hit Dexter[94] flags 11% to 13% of the approved drugs as promiscuous and 5% to 6% as highly promiscuous. Whether these models flag these drugs as potentially problematic for the right or wrong reasons remains to be investigated in most cases. Nevertheless, it is evident that if these models were used as a rigid filter, contrary to their intended purpose, these valuable compounds would have been overlooked and discarded.

Most studies evaluating the performance of in silico models are of retrospective nature. Moreover, the training and test data are sometimes not disclosed, or the relationship between the training and test data (and hence, the challenges presented by the tests) is not sufficiently understood. Widely accepted benchmark data sets for assay interference predictors are yet to be established. Hence, comparing tools and validation studies is a nontrivial task.

Reports of applications of in silico approaches in drug discovery projects are primarily available for the earlier introduced, simpler rule-based and similarity-based approaches. For most machine learning approaches, the number of studies that report using these in silico tools does not yet exceed a dozen. Systematic, prospective validation studies, such as those reported for Aggregator Advisor[63], DeepSCAMs[129], and LiabilityPredictor[107] remain rare, probably owing to the substantial experimental efforts involved in systematically evaluating in silico predictions (a similar situation is observed in the related field of bioactivity prediction[141]).

Generally speaking, the available computational tools are not yet sufficiently accurate to be used independently for greenlighting or rejecting compounds (exceptions include well-specified, small sets of rules defining undesirable molecular moieties and physicochemical properties), in the same way that it is not advisable to rely on any single bioassay in experimental screening. The rule-based approaches tend to produce false alerts, the similarity-based methods can potentially mislead users to clear compounds just because there is no evidence to suggest that they may be problematic, and the machine learning approaches may be overestimated in their ability to generalize (for example, owing to data biases).

That said, many existing in silico tools have substantial scientific and practical merit and are recommendable for integration into the early drug discovery infrastructure. In fact, the widespread use of rule-based and similarity-based approaches in screening and screening deck design is already a reality, and the more complex and powerful approaches are now gaining traction quickly. Small-scale research entities, in particular, may benefit strongly from the in silico tools as they can emulate, to some extent, domain-specific expertise which may otherwise be out of reach.

The implications of the current in silico tools extends beyond what is readily evident in the scientific literature. These models serve as catalysts for substantial efforts by academia and industry to boost computational approaches as pillars of screening deck design, hit triage and assay monitoring. Usability issues are slowing down the adoption and integration of in silico tools. Also, most models originating from academia are rarely updated ('publish-and-forget'). More sustainable maintenance, development and support approaches are required. These could be reached through a coordinated community effort.

Stimulated by the increasing accessibility of high-quality, information-rich measured data and new methods, advanced models with superior performance and broader scope will become available in the near future. These models are expected to use metadata to account for the different assay technologies, experimental setups and mechanisms of interference. In this context, large language models, such as the one powering ChatGPT, hold great promise, as they can distil and interpret data, for example, from expert-written assay protocols, and use them for predictions. Methods for bias control will be particularly relevant to developing models for frequent-hitter prediction as they are commonly trained on large but scarcely populated, highly imbalanced bioactivity matrices.

Research into new machine learning algorithms, molecular representations and explainable artificial intelligence constitute further promising avenues to progress. In particular, deep learning may benefit from federated learning[142] and related approaches, which enable collaboration across different research organizations without the need to share private data (see the MELLODDY project as an example[143]). Finally, developing and establishing widely accepted, high-quality benchmark data sets enabling thorough, transparent and reproducible model validation should be recognized as a priority. In that respect, there is still a long way to go[104].

For users of the existing in silico tools, the importance of familiarity with the scope and limitations of the individual tools cannot be overstated (Box 4). Users will benefit most from computational tools if they amalgamate, wherever possible, their computational and experimental efforts, for example, by using in silico tools for flagging hits for which the raw assay data should undergo an extra round of vetting or for which further experimental validation in orthogonal assays is in order (Fig. 1). Any drug discovery project should involve researchers with expertise in assay technologies and interference. This expertise will also be of great value to interpreting in silico predictions, which can be challenging, particularly for machine learning models. If such know-how is unavailable internally, it should be brought in from outside.

We hope this Review will enhance awareness of the challenges presented by assay interference in early drug discovery and the current opportunities for addressing these challenges through integrated approaches that combine measurement and computation. We also hope it will encourage many researchers to explore the existing in silico tools for assay interference prediction and stimulate their further development and integration into drug discovery pipelines.

Published online: 15 April 2024

References

1. Sánchez-Ruiz, A. & Colmenarejo, G. Updated prediction of aggregators and assay-interfering substructures in food compounds. *J. Agric. Food Chem.* **69**, 15184–15194 (2021).
2. Baell, J. B. & Holloway, G. A. New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *J. Med. Chem.* **53**, 2719–2740 (2010).
3. David, L. et al. Identification of compounds that interfere with high-throughput screening assay technologies. *ChemMedChem* **14**, 1795–1802 (2019).
4. Bisson, J. et al. Can invalid bioactives undermine natural product-based drug discovery? *J. Med. Chem.* **59**, 1671–1690 (2016).
5. Roche, O. et al. Development of a virtual screening method for identification of "frequent hitters" in compound libraries. *J. Med. Chem.* **45**, 137–142 (2002).
6. Thorne, N., Auld, D. S. & Inglese, J. Apparent activity in high-throughput screening: origins of compound-dependent assay interference. *Curr. Opin. Chem. Biol.* **14**, 315–324 (2010).
7. Coussens, N. P. et al. in *Assay Guidance Manual* (eds Markossian, S. et al.) 1067–1116 (NCATS, 2020).
8. Baell, J. & Walters, M. A. Chemistry: chemical con artists foil drug discovery. *Nature* **513**, 481–483 (2014).
9. Coussens, N. P., Auld, D. S., Thielman, J. R., Wagner, B. K. & Dahlin, J. L. Addressing compound reactivity and aggregation assay interferences: case studies of biochemical high-throughput screening campaigns benefiting from the National Institutes of Health Assay Guidance Manual guidelines. *SLAS Discov.* **26**, 1280–1290 (2021).
10. Hermann, J. C. et al. Metal impurities cause false positives in high-throughput screening campaigns. *ACS Med. Chem. Lett.* **4**, 197–200 (2013).
11. Chatzopoulou, M. et al. Pilot study to quantify palladium impurities in lead-like compounds following commonly used purification techniques. *ACS Med. Chem. Lett.* **13**, 262–270 (2022).
12. Dahlin, J. L. et al. Nuisance compounds in cellular assays. *Cell Chem. Biol.* **28**, 356–370 (2021).
13. Senger, M. R., Fraga, C. A. M., Dantas, R. F. & Silva, F. P. Filtering promiscuous compounds in early drug discovery: is it a good idea? *Drug Discov. Today* **21**, 868–872 (2016).
14. Rothenaigner, I. & Hadian, K. Brief guide: experimental strategies for high-quality hit selection from small-molecule screening campaigns. *SLAS Discov. Adv. Sci. Drug Discov.* **7**, 851–854 (2021).
15. Kallal, L. A. et al. High-throughput screening and triage assays identify small molecules targeting c-MYC in cancer cells. *SLAS Discov.* **26**, 216–229 (2021).
16. Vidler, L. R., Watson, I. A., Margolis, B. J., Cummins, D. J. & Brunavs, M. Investigating the behavior of published PAINS alerts using a pharmaceutical company data set. *ACS Med. Chem. Lett.* **9**, 792–796 (2018).
17. Aldrich, C. et al. The ecstasy and agony of assay interference compounds. *ACS Cent. Sci.* **3**, 143–147 (2017).
18. McCoy, M. A. et al. Biophysical survey of small-molecule β-catenin inhibitors: a cautionary tale. *J. Med. Chem.* **65**, 7246–7261 (2022).
19. Dahlin, J. L. & Walters, M. A. How to triage PAINS-full research. *ASSAY Drug Dev. Technol.* **14**, 168–174 (2016).
20. Arrowsmith, C. H. et al. The promise and peril of chemical probes. *Nat. Chem. Biol.* **11**, 536–541 (2015).
21. Newman, D. J. Problems that can occur when assaying extracts to pure compounds in biological systems. *Curr. Ther. Res.* **95**, 100645 (2021).
22. Kenny, P. W. Comment on the ecstasy and agony of assay interference compounds. *J. Chem. Inf. Model.* **57**, 2640–2645 (2017).
23. Seidler, J., McGovern, S. L., Doman, T. N. & Shoichet, B. K. Identification and prediction of promiscuous aggregating inhibitors among known drugs. *J. Med. Chem.* **46**, 4477–4486 (2003).
24. Doak, A. K., Wille, H., Prusiner, S. B. & Shoichet, B. K. Colloid formation by drugs in simulated intestinal fluid. *J. Med. Chem.* **53**, 4259–4265 (2010).
25. Baell, J. B. Feeling nature's PAINS: natural products, natural product drugs, and pan assay interference compounds (PAINS). *J. Nat. Prod.* **79**, 616–628 (2016).
26. Hendrich, A. B. Flavonoid-membrane interactions: possible consequences for biological effects of some polyphenolic compounds. *Acta Pharmacol. Sin.* **27**, 27–40 (2006).
27. Pawlikowska-Pawlęga, B. et al. Modification of membranes by quercetin, a naturally occurring flavonoid, via its incorporation in the polar head group. *Biochim. Biophys. Acta* **1768**, 2195–2204 (2007).
28. Kongkamnerd, J. et al. The quenching effect of flavonoids on 4-methylumbelliferone, a potential pitfall in fluorimetric neuraminidase inhibition assays. *SLAS Discov.* **16**, 755–764 (2011).
29. McGovern, S. L., Caselli, E., Grigorieff, N. & Shoichet, B. K. Common mechanism underlying promiscuous inhibitors from virtual and high-throughput screening. *J. Med. Chem.* **45**, 1712–1722 (2002).
30. Capuzzi, S. J., Muratov, E. N. & Tropsha, A. Phantom PAINS: problems with the utility of alerts for pan-assay interference compounds. *J. Chem. Inf. Model.* **57**, 417–427 (2017).
31. Cassinelli, G. The roots of modern oncology: from discovery of new antitumor anthracyclines to their clinical use. *Tumori J.* **102**, 226–235 (2016).
32. Simeonov, A. & Davis, M. I. in *Assay Guidance Manual* (eds Markossian, S. et al.) 1151–1162 (NCATS, 2004).
33. Auld, D. S. & Inglese, J. in *Assay Guidance Manual* (eds Markossian, S. et al.) 1163–1175 (NCATS, 2018).
34. Dahlin, J. L. & Walters, M. A. The essential roles of chemistry in high-throughput screening triage. *Future Med. Chem.* **6**, 1265–1290 (2014).
35. Jones, P., McElroy, S., Morrison, A. & Pannifer, A. The importance of triaging in determining the quality of output from high-throughput screening. *Future Med. Chem.* **7**, 1847–1852 (2015).
36. Auld, D. S. et al. in *Assay Guidance Manual* (eds Markossian, S. et al.) 1177–1202 (NCATS, 2017).
37. Dahlin, J. L., Baell, J. & Walters, M. A. in *Assay Guidance Manual* (eds. Markossian, S. et al.) 1117-1150 (NCATS, 2015).
38. Busby, S. A. et al. Advancements in assay technologies and strategies to enable drug discovery. *ACS Chem. Biol.* **15**, 2636–2648 (2020).
39. Holdgate, G., Embrey, K., Milbradt, A. & Davies, G. Biophysical methods in early drug discovery. *ADMET DMPK* **7**, 222–241 (2019).
40. Dahlin, J. L. et al. PAINS in the assay: chemical mechanisms of assay interference and promiscuous enzymatic inhibition observed during a sulfhydryl-scavenging HTS. *J. Med. Chem.* **58**, 2091–2113 (2015).
41. Kitchen, D. B. & Decornez, H. Y. in *Small Molecule Medicinal Chemistry: Strategies and Technologies* (eds Czechtizky, W. & Hamley, P.) Ch. 7 (Wiley, 2015).
42. Posner, B. A., Xi, H. & Mills, J. E. J. Enhanced HTS hit selection via a local hit rate analysis. *J. Chem. Inf. Model.* **49**, 2202–2210 (2009).
43. Schuffenhauer, A. et al. Evolution of Novartis' small molecule screening deck design. *J. Med. Chem.* **63**, 14425–14447 (2020).
44. Johnson, M. & Maggiora, G. (eds) *Concepts and Applications of Molecular Similarity* (Wiley, 1990).
45. Willett, P. The calculation of molecular structural similarity: principles and practice. *Mol. Inform.* **33**, 403–413 (2014).
46. Borrel, A. et al. High-throughput screening to predict chemical-assay interference. *Sci. Rep.* **10**, 3986 (2020).
47. Kenny, P. W. & Sadowski, J. in *Methods and Principles in Medicinal Chemistry* (ed. Oprea, T. I.) 271–285 (Wiley, 2005).
48. Wawer, M. & Bajorath, J. Local structural changes, global data views: graphical substructure–activity relationship trailing. *J. Med. Chem.* **54**, 2944–2951 (2011).
49. Guha, R. & Van Drie, J. H. Structure–activity landscape index: identifying and quantifying. *J. Chem. Inf. Model.* **48**, 646–658 (2008).
50. Lajiness, M. S. in *QSAR: Rational Approaches to the Design of Bioactive Compounds* (eds Silipo, C. & Vittoria, A.) 201–204 (Elsevier, 1990).

51. Medina-Franco, J. L. Activity cliffs: facts or artifacts? *Chem. Biol. Drug Des.* **81**, 553–556 (2013).
52. Guha, R. & Medina-Franco, J. L. On the validity versus utility of activity landscapes: are all activity cliffs statistically significant? *J. Cheminform.* **6**, 11 (2014).
53. Wilkinson, M. D. et al. The FAIR guiding principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
54. Wang, Y., Cheng, T. & Bryant, S. H. PubChem BioAssay: a decade's development toward open high-throughput screening data sharing. *SLAS Discov.* **22**, 655–666 (2017).
55. Mendez, D. et al. ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res.* **47**, D930–D940 (2019).
56. ChEMBL Version 33. https://www.ebi.ac.uk/chembl/ (2023).
57. European Chemical Biology Database (ECBD). https://ecbd.eu/ (2024).
58. Alves, V. M. et al. SCAM Detective: accurate predictor of small, colloidally aggregating molecules. *J. Chem. Inf. Model.* **60**, 4056–4063 (2020).
59. Yang, Z.-Y. et al. ChemFLuo: a web-server for structure analysis and identification of fluorescent compounds. *Brief. Bioinform.* **22**, bbaa282 (2021).
60. Ghosh, D., Koch, U., Hadian, K., Sattler, M. & Tetko, I. V. Luciferase Advisor: high-accuracy model to flag false positive hits in luciferase HTS assays. *J. Chem. Inf. Model.* **58**, 933–942 (2018).
61. Yang, Z.-Y. et al. Structural analysis and identification of false positive hits in luciferase-based assays. *J. Chem. Inf. Model.* **60**, 2031–2043 (2020).
62. Molina, C., Ait-Ouarab, L. & Minoux, H. Isometric Stratified Ensembles: a partial and incremental adaptive applicability domain and consensus-based classification strategy for highly imbalanced data sets with application to colloidal aggregation. *J. Chem. Inf. Model.* **62**, 1849–1862 (2022).
63. Irwin, J. J. et al. An aggregation advisor for ligand discovery. *J. Med. Chem.* **58**, 7076–7087 (2015).
64. Google Dataset Search. https://datasetsearch.research.google.com/ (2024).
65. Sieg, J., Flachsenberg, F. & Rarey, M. In need of bias control: evaluating chemical data for machine learning in structure-based virtual screening. *J. Chem. Inf. Model.* **59**, 947–961 (2019).
66. Yang, K. et al. Analyzing learned molecular representations for property prediction. *J. Chem. Inf. Model.* **59**, 3370–3388 (2019).
67. Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* **46**, 3–26 (2001).
68. Daylight Chemical Information Systems. SMARTS theory manual. https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html (2023).
69. Bruns, R. F. & Watson, I. A. Rules for identifying potentially reactive or promiscuous compounds. *J. Med. Chem.* **55**, 9763–9772 (2012).
70. Brenk, R. et al. Lessons learnt from assembling screening libraries for drug discovery for neglected diseases. *ChemMedChem* **3**, 435–444 (2008).
71. Pearce, B. C., Sofia, M. J., Good, A. C., Drexler, D. M. & Stock, D. A. An empirical process for the design of high-throughput screening deck filters. *J. Chem. Inf. Model.* **46**, 1060–1068 (2006).
72. Chakravorty, S. J. et al. Nuisance compounds, PAINS filters, and dark chemical matter in the GSK HTS collection. *SLAS Discov.* **23**, 532–544 (2018).
73. McCallum, M. M. et al. High-throughput identification of promiscuous inhibitors from screening libraries with the use of a thiol-containing fluorescent probe. *SLAS Discov.* **18**, 705–713 (2013).
74. Matlock, M. K., Hughes, T. B., Dahlin, J. L. & Swamidass, S. J. Modelling small-molecule reactivity identifies promiscuous bioactive compounds. *J. Chem. Inf. Model.* **58**, 1483–1500 (2018).
75. Schorpp, K. et al. Identification of small-molecule frequent hitters from AlphaScreen high-throughput screens. *J. Biomol. Screen.* **19**, 715–726 (2014).
76. Baell, J. B. & Nissink, J. W. M. Seven year itch: pan-assay interference compounds (PAINS) in 2017 — utility and limitations. *ACS Chem. Biol.* **13**, 36–44 (2018).
77. Sushko, I., Salmina, E., Potemkin, V. A., Poda, G. & Tetko, I. V. ToxAlerts: a web server of structural alerts for toxic chemicals and compounds with potential adverse reactions. *J. Chem. Inf. Model.* **52**, 2310–2316 (2012).
78. Yang, H., Lou, C., Li, W., Liu, G. & Tang, Y. Computational approaches to identify structural alerts and their applications in environmental toxicology and drug discovery. *Chem. Res. Toxicol.* **33**, 1312–1322 (2020).
79. OCHEM ToxAlerts. https://ochem.eu/alerts/home.do (2024).
80. Irwin, J. J. et al. ZINC20 — a free ultralarge-scale chemical database for ligand discovery. *J. Chem. Inf. Model.* **60**, 6065–6073 (2020).
81. ZINC20 patterns. https://zinc20.docking.org/patterns/ (2024).
82. Lajiness, M. S., Maggiora, G. M. & Shanmugasundaram, V. Assessment of the consistency of medicinal chemists in reviewing sets of compounds. *J. Med. Chem.* **47**, 4891–4896 (2004).
83. Ekins, S. et al. Analysis and hit filtering of a very large library of compounds screened against *Mycobacterium tuberculosis*. *Mol. BioSyst.* **6**, 2316–2324 (2010).
84. Chai, C. L. & Mátyus, P. One size does not fit all: challenging some dogmas and taboos in drug discovery. *Future Med. Chem.* **8**, 29–38 (2016).
85. Dantas, R. F. et al. Dealing with frequent hitters in drug discovery: a multidisciplinary view on the issue of filtering compounds on biological screenings. *Expert Opin. Drug Discov.* **14**, 1269-1282 (2019).
86. Alfonso, L. F., Srivenugopal, K. S. & Bhat, G. J. Does aspirin acetylate multiple cellular proteins? (Review). *Mol. Med. Rep.* **2**, 533–537 (2009).
87. Ehmki, E. S. R., Schmidt, R., Ohm, F. & Rarey, M. Comparing molecular patterns using the example of SMARTS: applications and filter collection analysis. *J. Chem. Inf. Model.* **59**, 2572–2586 (2019).
88. Schmidt, R. et al. Comparing molecular patterns using the example of SMARTS: theory and algorithms. *J. Chem. Inf. Model.* **59**, 2560–2571 (2019).
89. Tropsha, A., Gramatica, P. & Gombar, V. K. The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Comb. Sci.* **22**, 69–77 (2003).
90. Netzeva, T. I. et al. Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships. The report and recommendations of ECVAM Workshop 52. *Altern. Lab. Anim.* **33**, 155–173 (2005).
91. Hu, Y. & Bajorath, J. High-resolution view of compound promiscuity. *F1000Research* **2**, 144 (2013).
92. Jasial, S., Gilberg, E., Blaschke, T. & Bajorath, J. Machine learning distinguishes with high accuracy between pan-assay interference compounds that are promiscuous or represent dark chemical matter. *J. Med. Chem.* **61**, 10255–10264 (2018).
93. Stork, C. et al. Hit Dexter: a machine-learning model for the prediction of frequent hitters. *ChemMedChem* **13**, 564–571 (2018).
94. Stork, C., Chen, Y., Šícho, M. & Kirchmair, J. Hit Dexter 2.0: machine-learning models for the prediction of frequent hitters. *J. Chem. Inf. Model.* **59**, 1030–1043 (2019).
95. Stork, C., Mathai, N. & Kirchmair, J. Computational prediction of frequent hitters in target-based and cell-based assays. *Artif. Intell. Life Sci.* **1**, 100007 (2021).
96. Ghosh, D., Koch, U., Hadian, K., Sattler, M. & Tetko, I. V. Highly accurate filters to flag frequent hitters in AlphaScreen assays by suggesting their mechanism. *Mol. Inform.* **41**, 2100151 (2022).
97. Kruschke, J. K. *Doing Bayesian Data Analysis: a Tutorial with R, JAGS, and Stan* 2nd edn (Academic, 2015).
98. Yongye, A. B. & Medina-Franco, J. L. Toward an efficient approach to identify molecular scaffolds possessing selective or promiscuous compounds. *Chem. Biol. Drug Des.* **82**, 367–375 (2013).
99. Goodwin, S., Shahtahmassebi, G. & Hanley, Q. S. Statistical models for identifying frequent hitters in high throughput screening. *Sci. Rep.* **10**, 17200 (2020).
100. Yang, J. J. et al. Badapple: promiscuity patterns from noisy evidence. *J. Cheminform.* **8**, 29 (2016).
101. Hu, Y. & Bajorath, J. Polypharmacology directed compound data mining: identification of promiscuous chemotypes with different activity profiles and comparison to approved drugs. *J. Chem. Inf. Model.* **50**, 2112–2118 (2010).
102. M Nissink, J. W. & Blackburn, S. Quantification of frequent-hitter behavior based on historical high-throughput screening data. *Future Med. Chem.* **6**, 1113–1126 (2014).
103. Upadhyay, R., Phlypo, R., Saini, R. & Liwicki, M. Sharing to learn and learning to share — fitting together meta-learning, multi-task learning, and transfer learning: a meta review. Preprint at https://arxiv.org/abs/2111.12146 (2021).
104. Walters, P. We need better benchmarks for machine learning in drug discovery. *Practical Cheminformatics* http://practicalcheminformatics.blogspot.com/2023/08/we-need-better-benchmarks-for-machine.html (2023).
105. Mellin, W. D. Work with new electronic 'brains' opens field for army math experts. *Hammond Times* 65 (10 November 1957).
106. Jiménez-Luna, J., Grisoni, F. & Schneider, G. Drug discovery with explainable artificial intelligence. *Nat. Mach. Intell.* **2**, 573–584 (2020).
107. Alves, V. et al. Lies and liabilities: computational assessment of high-throughput screening hits to identify artifact compounds. *J. Med. Chem.* **66**, 12828–12839 (2023).
108. Blaschke, T., Feldmann, C. & Bajorath, J. Prediction of promiscuity cliffs using machine learning. *Mol. Inform.* **40**, 2000196 (2021).
109. Gilberg, E., Stumpfe, D. & Bajorath, J. Activity profiles of analog series containing pan assay interference compounds. *RSC Adv.* **7**, 35638–35647 (2017).
110. Choo, M. Z. Y. & Chai, C. L. L. Promoting GAINS (give attention to limitations in assays) over PAINS alerts: no pains, more gains. *ChemMedChem* **17**, e202100710 (2022).
111. Jasial, S., Hu, Y. & Bajorath, J. How frequently are pan-assay interference compounds active? Large-scale analysis of screening data reveals diverse activity profiles, low global hit frequency, and many consistently inactive compounds. *J. Med. Chem.* **60**, 3879–3886 (2017).
112. Berthold, M. R. et al. in *Studies in Classification, Data Analysis, and Knowledge Organization* (eds Preisach, C. et al.) 319–326 (Springer, 2007).
113. Brenke, J. K. et al. Identification of small-molecule frequent hitters of glutathione S-transferase–glutathione interaction. *SLAS Discov.* **21**, 596–607 (2016).
114. Huth, J. R. et al. ALARM NMR: a rapid and robust experimental method to detect reactive false positives in biochemical screens. *J. Am. Chem. Soc.* **127**, 217–224 (2005).
115. Metz, J. T., Huth, J. R. & Hajduk, P. J. Enhancement of chemical rules for predicting compound reactivity towards protein thiol groups. *J. Comput. Aided Mol. Des.* **21**, 139–144 (2007).
116. Walters, W. P., Stahl, M. T. & Murcko, M. A. Virtual screening — an overview. *Drug Discov. Today* **3**, 160–178 (1998).
117. Yang, Z.-Y., Yang, Z.-J., Lu, A.-P., Hou, T.-J. & Cao, D.-S. Scopy: an integrated negative design Python library for desirable HTS/VS database design. *Brief. Bioinform.* **22**, bbaa194 (2021).
118. Blaschke, T., Miljković, F. & Bajorath, J. Prediction of different classes of promiscuous and nonpromiscuous compounds using machine learning and nearest neighbor analysis. *ACS Omega* **4**, 6883–6890 (2019).

119. Couronne, C., Koptelov, M. & Zimmermann, A. in *Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track* Vol. 12461 (eds Dong, Y. et al.) 570–573 (Springer, 2021).
120. Schneider, P., Röthlisberger, M., Reker, D. & Schneider, G. Spotting and designing promiscuous ligands for drug discovery. *Chem. Commun.* **52**, 1135–1138 (2016).
121. De Matos, A. M. et al. Glucosylpolyphenols as inhibitors of Aβ-induced Fyn kinase activation and tau phosphorylation: synthesis, membrane permeability, and exploratory target assessment within the scope of type 2 diabetes and Alzheimer's disease. *J. Med. Chem.* **63**, 11663–11690 (2020).
122. Bajorath, J. Evolution of assay interference concepts in drug discovery. *Expert Opin. Drug Discov.* **16**, 719–721 (2021).
123. Feldmann, C. & Bajorath, J. Advances in computational polypharmacology. *Mol. Inform.* **41**, 2200190 (2022).
124. Enamine HTS Collection. https://enamine.net/compound-collections/screening-collection/hts-collection (2024).
125. Tritsch, D., Zinglé, C., Rohmer, M. & Grosdemange-Billiard, C. Flavonoids: true or promiscuous inhibitors of enzyme? The case of deoxyxylulose phosphate reductoisomerase. *Bioorg. Chem.* **59**, 140–144 (2015).
126. Yang, Z.-Y. et al. Structural analysis and identification of colloidal aggregators in drug discovery. *J. Chem. Inf. Model.* **59**, 3714–3726 (2019).
127. O'Donnell, H. R., Tummino, T. A., Bardine, C., Craik, C. S. & Shoichet, B. K. Colloidal aggregators in biochemical SARS-CoV-2 repurposing screens. *J. Med. Chem.* **64**, 17530–17539 (2021).
128. SCAM Detective. https://scamdetective.mml.unc.edu/ (2024).
129. Lee, K. et al. Combating small-molecule aggregation with machine learning. *Cell Rep. Phys. Sci.* **2**, 100573 (2021).
130. Hughes, T. B., Dang, N. L., Miller, G. P. & Swamidass, S. J. Modelling reactivity to biological macromolecules with a deep multitask network. *ACS Cent. Sci.* **2**, 529–537 (2016).
131. Alsibaee, A. M., Aljohar, H. I., Attwa, M. W., Abdelhameed, A. S. & Kadi, A. A. Reactive intermediates formation and bioactivation pathways of spebrutinib revealed by LC-MS/MS: in vitro and in silico metabolic study. *Heliyon* **9**, e17058 (2023).
132. Al-Shakliah, N. S., Kadi, A. A., Aljohar, H. I., AlRabiah, H. & Attwa, M. W. Profiling of in vivo, in vitro and reactive zorifertinib metabolites using liquid chromatography ion trap mass spectrometry. *RSC Adv.* **12**, 20991–21003 (2022).
133. Alsubi, T. A., Attwa, M. W., Darwish, H. W., Abuelizz, H. A. & Kadi, A. A. Piperazine ring toxicity in three novel anti-breast cancer drugs: an in silico and in vitro metabolic bioactivation approach using olaparib as a case study. *Naunyn Schmiedebergs Arch. Pharmacol.* **396**, 1435–1450 (2023).
134. Hughes, T. B., Flynn, N., Dang, N. L. & Swamidass, S. J. Modelling the bioactivation and subsequent reactivity of drugs. *Chem. Res. Toxicol.* **34**, 584–600 (2021).
135. Borrel, A. et al. InterPred: a webtool to predict chemical autofluorescence and luminescence interference. *Nucleic Acids Res.* **48**, W586–W590 (2020).
136. Thomas, R. The US Federal Tox21 Program: a strategic and operational plan for continued leadership. *ALTEX* **35**, 163–168 (2018).
137. Wassermann, A. M. et al. Dark chemical matter as a promising starting point for drug lead discovery. *Nat. Chem. Biol.* **11**, 958–966 (2015).
138. NCATS. NPACT Chemical Library — innovative chemical biology library for translational sciences. https://ncats.nih.gov/preclinical/core/compound/npact (2024).
139. ChemFH — integrated online platform for the identification of potential frequent hitters. https://chemfh.scbdd.com/ (2024).
140. RDKit. https://www.rdkit.org/ (2024).
141. Mathai, N., Chen, Y. & Kirchmair, J. Validation strategies for target prediction methods. *Brief. Bioinform.* **21**, 791–802 (2020).
142. Hanser, T. Federated learning for molecular discovery. *Curr. Opin. Struct. Biol.* **79**, 102545 (2023).
143. Heyndrickx, W. et al. MELLODDY: cross-pharma federated learning at unprecedented scale unlocks benefits in QSAR without compromising proprietary information. *J. Chem. Inf. Model.* https://doi.org/10.1021/acs.jcim.3c00799 (2023).
144. FAF-Drugs4. https://mobyle2.rpbs.univ-paris-diderot.fr (2024).
145. KNIME_MedChem_filters. https://gitlab.com/Jukic/knime_medchem_filters/ (2024).
146. SmartsFilter. https://datascience.unm.edu/tomcat/biocomp/smartsfilter (2024).
147. SwissADME. http://www.swissadme.ch/ (2024).
148. RDKit PAINS filter. https://www.rdkit.org/ (2024).
149. Hit Dexter. https://nerdd.univie.ac.at/hitdexter3/ (2024).
150. Lilly Rules. https://github.com/IanAWatson/Lilly-Medchem-Rules (2024).
151. NIBR substructure filters for hit finding and triaging. https://github.com/rdkit/rdkit/tree/master/Contrib/NIBRSubstructureFilters (2024).
152. RDKit NIBR substructure filters for hit finding and triaging. https://www.rdkit.org/ (2024).
153. Badapple. https://datascience.unm.edu/tomcat/badapple/badapple (2024).
154. Jasial, S., Gilberg, E., Blaschke, T. & Bajorath, J. Distinguishing between pan assay interference compounds (PAINS) that are promiscuous or represent dark chemical matter — data set and prediction models. *Zenodo* https://zenodo.org/record/1453913 (2018).
155. Blaschke, T., Feldman, C. & Bajorath, J. Prediction of promiscuity cliffs using machine learning. *Zenodo* https://zenodo.org/record/4013954 (2020).
156. OCHEM model for frequent hitter prediction in AlphaScreen assays. https://ochem.eu/article/125278 (2024).
157. ChemAgg. https://admet.scbdd.com/ChemAGG/index/ (2024).
158. DeepSCAMs. https://github.com/tcorodrigues/DeepSCAMs (2024).
159. Isometric Stratified Ensembles (ISE). https://pikairos.eu/download/aggregation_classification/ (2024).
160. InterPred. https://sandbox.ntp.niehs.nih.gov/interferences/ (2024).
161. OCHEM Luciferase Advisor model. https://ochem.eu/model/697 (2024).
162. Mangal, M., Sagar, P., Singh, H., Raghava, G. P. S. & Agarwal, S. M. NPACT: naturally occurring plant-based anti-cancer compound-activity-target database. *Nucleic Acids Res.* **41**, D1124–D1129 (2013).
163. LiabilityPredictor. https://liability.mml.unc.edu/ (2024).
164. Guha, R. et al. Exploratory analysis of kinetic solubility measurements of a small molecule library. *Bioorg. Med. Chem.* **19**, 4127–4134 (2011).
165. Chen, C. et al. Fragment-based drug nanoaggregation reveals drivers of self-assembly. *Nat. Commun.* **14**, 8340 (2023).
166. Hann, M. M. Molecular obesity, potency and other addictions in drug discovery. *MedChemComm* **2**, 349 (2011).
167. Azzaoui, K. et al. Modelling promiscuity based on in vitro safety pharmacology profiling data. *ChemMedChem* **2**, 874–880 (2007).
168. Gleeson, M. P., Hersey, A., Montanari, D. & Overington, J. Probing the links between in vitro potency, ADMET and physicochemical parameters. *Nat. Rev. Drug Discov.* **10**, 197–208 (2011).
169. Peters, J.-U. et al. Can we discover pharmacological promiscuity early in the drug discovery process? *Drug Discov. Today* **17**, 325–335 (2012).
170. Hu, Y. & Bajorath, J. Compound promiscuity: what can we learn from current data? *Drug Discov. Today* **18**, 644–650 (2013).
171. Waring, M. J. Lipophilicity in drug discovery. *Expert Opin. Drug Discov.* **5**, 235–248 (2010).
172. Schneider, P. & Schneider, G. Privileged structures revisited. *Angew. Chem. Int. Ed.* **56**, 7971–7974 (2017).
173. DeSimone, R., Currie, K., Mitchell, S., Darrow, J. & Pippin, D. Privileged structures: applications in drug discovery. *Comb. Chem. High Throughput Screen.* **7**, 473–493 (2004).
174. Hopkins, A. L. Network pharmacology: the next paradigm in drug discovery. *Nat. Chem. Biol.* **4**, 682–690 (2008).
175. Anighoro, A., Bajorath, J. & Rastelli, G. Polypharmacology: challenges and opportunities in drug discovery: miniperspective. *J. Med. Chem.* **57**, 7874–7887 (2014).
176. Bajorath, J. Origins and progression of the polypharmacology concept in drug discovery. *Artif. Intell. Life Sci.* **5**, 100094 (2024).
177. Evans, B. E. et al. Methods for drug discovery: development of potent, selective, orally effective cholecystokinin antagonists. *J. Med. Chem.* **31**, 2235–2246 (1988).
178. Feng, B. Y. & Shoichet, B. K. A detergent-based assay for the detection of promiscuous inhibitors. *Nat. Protoc.* **1**, 550–553 (2006).
179. Reker, D., Bernardes, G. J. L. & Rodrigues, T. Computational advances in combating colloidal aggregation in drug discovery. *Nat. Chem.* **11**, 402–418 (2019).
180. Coan, K. E. D. & Shoichet, B. K. Stability and equilibria of promiscuous aggregates in high protein milieus. *Mol. Biosyst.* **3**, 208 (2007).
181. Coan, K. E. D., Maltby, D. A., Burlingame, A. L. & Shoichet, B. K. Promiscuous aggregate-based inhibitors promote enzyme unfolding. *J. Med. Chem.* **52**, 2067–2075 (2009).
182. Coan, K. E. D. & Shoichet, B. K. Stoichiometry and physical chemistry of promiscuous aggregate-based inhibitors. *J. Am. Chem. Soc.* **130**, 9606–9612 (2008).
183. Blevitt, J. M. et al. Structural basis of small-molecule aggregate induced inhibition of a protein–protein interaction. *J. Med. Chem.* **60**, 3511–3517 (2017).
184. Blanusa, M., Varnai, V. M., Piasek, M. & Kostial, K. Chelators as antidotes of metal toxicity: therapeutic and experimental aspects. *Curr. Med. Chem.* **12**, 2771–2794 (2005).
185. Repac Antić, D., Parčina, M., Gobin, I. & Petković Didović, M. Chelation in antibacterial drugs: from nitroxoline to cefiderocol and beyond. *Antibiotics* **11**, 1105 (2022).
186. Feng, B. Y. et al. A high-throughput screen for aggregation-based inhibition in a large compound library. *J. Med. Chem.* **50**, 2385–2390 (2007).
187. Shoichet, B. K. Interpreting steep dose-response curves in early inhibitor discovery. *J. Med. Chem.* **49**, 7274–7277 (2006).
188. McGovern, S. L., Helfand, B. T., Feng, B. & Shoichet, B. K. A specific mechanism of nonspecific inhibition. *J. Med. Chem.* **46**, 4265–4272 (2003).
189. Feng, B. Y., Shelat, A., Doman, T. N., Guy, R. K. & Shoichet, B. K. High-throughput assays for promiscuous inhibitors. *Nat. Chem. Biol.* **1**, 146–148 (2005).
190. LaPlante, S. R. et al. Compound aggregation in drug discovery: implementing a practical NMR assay for medicinal chemists. *J. Med. Chem.* **56**, 5142–5150 (2013).
191. Feng, B. Y. et al. Small-molecule aggregates inhibit amyloid polymerization. *Nat. Chem. Biol.* **4**, 197–199 (2008).
192. Giannetti, A. M., Koch, B. D. & Browner, M. F. Surface plasmon resonance based assay for the detection and characterization of promiscuous inhibitors. *J. Med. Chem.* **51**, 574–580 (2008).
193. Singh, J., Petter, R. C., Baillie, T. A. & Whitty, A. The resurgence of covalent drugs. *Nat. Rev. Drug Discov.* **10**, 307–317 (2011).
194. Kwiatkowski, N. et al. Targeting transcription regulation in cancer with a covalent CDK7 inhibitor. *Nature* **511**, 616–620 (2014).
195. Ang, K. K. H. et al. Mining a cathepsin inhibitor library for new antiparasitic drug leads. *PLoS Negl. Trop. Dis.* **5**, e1023 (2011).

196. Arnold, L. A. et al. Discovery of small molecule inhibitors of the interaction of the thyroid hormone receptor with transcriptional coregulators. *J. Biol. Chem.* **280**, 43048–43055 (2005).
197. Copeland, R. A., Basavapathruni, A., Moyer, M. & Scott, M. P. Impact of enzyme concentration and residence time on apparent activity recovery in jump dilution analysis. *Anal. Biochem.* **416**, 206–210 (2011).
198. Copeland, R. A. *Evaluation of Enzyme Inhibitors in Drug Discovery: a Guide for Medicinal Chemists and Pharmacologists* (Wiley, 2013).
199. Ehmann, D. E. et al. Avibactam is a covalent, reversible, non-β-lactam β-lactamase inhibitor. *Proc. Natl Acad. Sci. USA* **109**, 11663–11668 (2012).
200. Johnston, P. A. et al. Development of a 384-well colorimetric assay to quantify hydrogen peroxide generated by the redox cycling of compounds in the presence of reducing agents. *ASSAY Drug Dev. Technol.* **6**, 505–518 (2008).
201. Simeonov, A. et al. Fluorescence spectroscopic profiling of compound libraries. *J. Med. Chem.* **51**, 2363–2371 (2008).
202. Imbert, P.-E. et al. Recommendations for the reduction of compound artifacts in time-resolved fluorescence resonance energy transfer assays. *ASSAY Drug Dev. Technol.* **5**, 363–372 (2007).
203. Simeonov, A. et al. Quantitative high-throughput screen identifies inhibitors of the *Schistosoma mansoni* redox cascade. *PLoS Negl. Trop. Dis.* **2**, e127 (2008).
204. Baljinnyam, B., Ronzetti, M. & Simeonov, A. Advances in luminescence-based technologies for drug discovery. *Expert Opin. Drug Discov.* **18**, 25–35 (2023).
205. Auld, D. S. et al. Molecular basis for the high-affinity binding and stabilization of firefly luciferase by PTC124. *Proc. Natl Acad. Sci. USA* **107**, 4878–4883 (2010).
206. Inglese, J. et al. Genome editing-enabled HTS assays expand drug target pathways for Charcot–Marie–Tooth disease. *ACS Chem. Biol.* **9**, 2594–2602 (2014).
207. Cheng, K. C.-C. & Inglese, J. A coincidence reporter-gene system for high-throughput screening. *Nat. Methods* **9**, 937–937 (2012).
208. Hasson, S. A. et al. Chemogenomic profiling of endogenous PARK2 expression using a genome-edited coincidence reporter. *ACS Chem. Biol.* **10**, 1188–1197 (2015).
209. Lang, L. & Teng, Y. in *Clinical and Preclinical Models for Maximizing Healthspan* Vol. 2138 (ed. Guest, P. C.) 159–166 (Springer, 2020).
210. Bender, A. et al. Evaluation guidelines for machine learning tools in the chemical sciences. *Nat. Rev. Chem.* **6**, 428–442 (2022).

Acknowledgements
The authors thank M. Brenek from the University of Vienna for her help in elaborating drafts of the figures presented in this publication. The financial support received for the Christian Doppler Laboratory for Molecular Informatics in the Biosciences by the Austrian Federal Ministry of Labour and Economy, the National Foundation for Research, Technology and Development, the Christian Doppler Research Association, BASF SE and Boehringer-Ingelheim RCV GmbH & Co KG, as well as the funding received for V.P. from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Actions, grant agreement 'Advanced machine learning for Innovative Drug Discovery (AIDD)' no. 956832, is gratefully acknowledged.

Author contributions
All authors researched data for the article, contributed substantially to discussion of the content, wrote the article, and reviewed and edited the manuscript before submission.

Competing interests
C.S. and J.K. are developers of the Hit Dexter machine learning models for frequent hitter prediction.

Additional information
Peer review information *Nature Reviews Chemistry* thanks Jose L. Medina-Franco, Alexander Tropsha and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Chapter 2: Methods

Assay interference is a complex and recurring phenomenon: elusive by nature and difficult to model or predict. Modeling assay interference requires a solid understanding of the procedures underlying assay platforms, including how various types of bioactivity readouts are generated, annotated, and organized. In addition, the success of any predictive tool depends on carefully selecting how molecules are represented, choosing an appropriate modeling framework, and defining robust metrics for performance evaluation.

This chapter introduces the computational foundations necessary to contextualize this thesis work. It begins with an overview of the datasets used throughout this thesis work, followed by an analysis of molecular representations, ML algorithms, and performance evaluation strategies employed in the development and assessment of predictive models.

2.1 Collecting, Compiling and Curating High-Throughput Data

Automated testing generates vast volumes of biochemical data, including chemical structures, assay metadata, and bioactivity readouts. This information forms the backbone of statistical and ML models, ultimately determining their practical utility. As such, the way this data is collected, curated, and distributed is critical for in silico approaches and cheminformatics applications, where datasets should adhere to the FAIR (Findability, Accessibility, Interoperability, and Reusability)[29] principles to ensure that targeted information can be efficiently extracted and meaningfully analyzed.

The development of reliable and accessible data resources is therefore essential across both industry and academia. Pharmaceutical companies often compile proprietary HTS datasets internally, while academic and public institutions undertake considerable efforts to construct and maintain open-access repositories by aggregating assay results from diverse sources. Each of these data ecosystems (public and private) comes with its advantages and limitations.

In this section, we describe the primary data sources used in this thesis (PubChem BioAssay and Bayer's proprietary HTS dataset)[18] while also discussing additional widely used public databases such as ChEMBL and DrugBank.[17,30] These resources represent the cornerstone of modern cheminformatics and drug discovery research, offering curated chemical and bioactivity data that support the development of predictive models and the validation of experimental results.

2.1.1 Public Databases

Publicly available chemical and biological data resources form the foundation upon which academic groups can conduct cutting-edge research and, in many cases, compete with industry counterparts.[31] These platforms are typically governed by well-defined data-sharing standards and technical protocols that ensure ease of data deposition, online access (often through both web-based user interfaces and programmatic APIs), and data download.

Among the most widely used repositories for small-molecule bioactivity data is PubChem, maintained by the National Center for Biotechnology Information (NCBI).[18] For our works, PubChem BioAssay served as a key data source. It contains a curated collection of more than 1.7 million biological assays, associated with 1.6 million unique small molecules, and encompasses more than 230 million individual bioactivity measurements. Data submitted to PubChem must adhere to strict formatting and reporting guidelines. All submissions undergo automated quality control to ensure compliance with these guidelines. Each assay entry typically includes the chemical structure of the tested compounds, compound identifiers and names, a detailed description of the assay protocol, the observed bioactivity outcome, and dose-response data across multiple compound concentrations.

The ChEMBL database,[17] maintained by the European Bioinformatics Institute (EMBL-EBI) is an essential publicly available resource. ChEMBL is a manually curated repository of bioactive molecules with drug-like properties, linking chemical structure and biological activity data extracted primarily from the medicinal chemistry literature. It currently hosts more than 2 million compounds and more than 15 million activity measurements across a wide array of targets and assay types. ChEMBL places a strong emphasis on standardized data curation and target annotation.

Another widely used public resource is DrugBank,[30] which provides detailed, curated information on approved and investigational drugs. Unlike PubChem, which focuses heavily on experimental assay data, DrugBank integrates chemical, pharmacological, and pharmaceutical data with clinical and molecular biology information. Each entry typically includes a compound's chemical structure, mechanism of action, pharmacokinetics, target proteins, interactions, metabolism, and known side effects.

2.1.2 Proprietary Databases

In contrast to public resources, pharmaceutical companies maintain large, internally generated databases not accessible to the broader scientific community. These proprietary datasets represent one of the most significant competitive advantages in the pharmaceutical industry, containing years, if not decades, of accumulated assay data, compound screening

results, and insights into structure-activity relationships. Such data collections are invaluable, playing a central role in internal decision-making and in the development of predictive models that inform drug discovery pipelines.

Leading pharmaceutical companies have built robust computational infrastructures around these datasets. Their platforms support a wide range of applications, including virtual screening, toxicity prediction, and lead optimization. In Bayer's case, the proprietary HTS library includes more than 5 million unique compounds and more than 500 standardized biological assays collected over multiple screening campaigns. Access to this internal resource is provided through a secure web-based interface, designed to support internal research teams across departments.

Unlike public databases, which aggregate contributions from a wide range of institutions and experimental protocols, proprietary datasets are typically generated by a smaller number of internal laboratories using consistent assay formats and standardized procedures. This reduces variability, making data integration and harmonization significantly easier. However, internal datasets are also subject to project-driven biases: companies often reuse successful scaffolds or test closely related compound series across multiple campaigns, which can introduce redundancies and reduce chemical diversity.

2.2 Machine Learning for Molecular Systems

The increasing availability of large-scale bioactivity data, both from public and proprietary sources, has fueled the development of modern in silico modeling tools capable of extracting meaningful patterns from complex chemical datasets. Central to this development is the application of ML and, more recently, DL techniques, which have demonstrated strong predictive performance across various tasks in drug discovery, including hits prioritization, ADMET prediction, and compound classification.[32]

The choice of ML algorithm for modeling a given chemical dataset depends on several factors, including dataset size, data sparsity, and the type of molecular representation employed. Popular approaches include Random Forests (RF), Feed-Forward Neural Networks (FFNNs), and Graph Neural Networks (GNNs), which operate directly on molecular graphs.[33] Support Vector Machines (SVMs) have also been widely used in earlier literature,; however, the high computational cost of kernel optimization and poor scalability in high-dimensional settings have made SVMs impractical for large datasets.

Regardless of the chosen model architecture, all ML workflows begin with data preprocessing and standardization, followed by the transformation of molecular structures

into vectorial formats. These representations serve as the input features for predictive modeling and play a decisive role in model performance.

In the following sections, we introduce the preprocessing workflow adopted in this thesis work and the most widely adopted molecular representations. We then present the technical foundations of RF and FFNN, followed by a brief overview of RL and metrics used to evaluate ML models. Together, these methods constitute the methodological backbone of this work.

All models and preprocessing pipelines described herein were implemented in Python, using the scientific computing ecosystem built around RDKit for chemical manipulation, scikit-learn for traditional ML algorithms,[34] and PyTorch for DL models' development.[35]

2.2.1 Preprocessing and Standardization of Molecular Data

Regardless of the data source, all molecular data must undergo careful preprocessing before being used for any modeling task.[36] Even when retrieved from well-curated public or proprietary databases, chemical datasets often contain problematic entries such as missing values, conflicting or ambiguous bioactivity annotations, and duplicate molecules, all of which can compromise models' training and final performances. Additional issues frequently include the presence of salt fragments and inconsistent tautomeric forms.

Several tools are available to support data cleaning and molecular standardization. Frameworks such as KNIME (Konstanz Information Miner)[37] provide modular environments for chemical data preprocessing. However, to ensure full transparency and control over the cleaning workflow, we developed a dedicated preprocessing pipeline using Python and the RDKit cheminformatics library.[38]

For our works, molecules were initially retrieved in common representations such as SMILES (Simplified Molecular Input Line Entry System),[39] or SDF formats.[40] As a first step, inorganic salt moieties and counterions were removed, which is especially important when building ML models, as salts can introduce structural noise, especially if inconsistently recorded across datasets. After desalting, molecules were standardized to a canonical tautomeric form using RDKit's built-in tools. Invalid or unparseable molecules were discarded, and duplicate structures were identified and removed to avoid data leakage between training and test sets. Molecules were then converted into the required vector representations (e.g., ECFP) for downstream modeling.

2.2.2 Machine-Readable Molecular Representations

Molecules are inherently quantum mechanical entities, defined by their electronic distribution and the physical forces acting between atoms. In silico modeling, however, requires these complex objects to be translated into formats that are interpretable by machines (Table 1). This fundamental challenge, finding an appropriate representation that captures chemical meaning while remaining computationally tractable, is central to cheminformatics and molecular ML.[41]

Table 1: Commonly Used Computer-Readable Molecular Representations

Representation Type	Name	Description
0D (Scalar descriptors)	Physicochemical Properties	Numeric descriptors such as molecular weight, logP, TPSA, etc.
1D (String)	SMILES	Linear string encoding atoms, connectivity, ring closures, and stereochemistry.
2D (Graph-based)	Molecular Graph	Atom-bond network where atoms are nodes and bonds are edges. Often enhanced with atomic and bond features.
2D (Fingerprint)	Morgan Fingerprints (ECFP)	Circular substructure-based fingerprints encoding atom environments into binary vectors.
3D (Geometric)	Molecular Conformers	Cartesian 3D atomic coordinates.
Learned (Latent Space)	Neural Embeddings / Graph Embeddings	Continuous representations learned by deep learning models (e.g., GNNs).

While chemists can extract a great deal of information from a two-dimensional chemical drawing, computers cannot interpret such diagrams directly without specialized image-processing pipelines. Atom hybridization, formal charges, stereochemistry, and connectivity patterns are not inherently machine-readable unless explicitly encoded. For this reason, cheminformatics relies on numerical formats to represent molecules in computational workflows.

The most basic level of representation, often referred to as 0D, includes global molecular properties or scalar descriptors (e.g., molecular weight, partition coefficient, or topological polar surface area), which summarize the overall features of the molecule without encoding structural connectivity. While useful for certain regression tasks, 0D descriptors, due to missing connectivity information, are often limited and are generally insufficient for tasks involving detailed structure-activity relationships.

Going further, 1D representations, including textual encodings such as SMILES, offer a deeper view of the chemical matter. SMILES strings describe molecular structures as linear character sequences, encoding atoms, connectivity, bond types, ring closures, and, if specified, stereochemistry. They are compact, easily stored, and widely supported by cheminformatics toolkits such as RDKit. However, while SMILES are convenient for indexing and storage, they are not directly amenable to most modeling tasks and must be converted into numerical formats for use in these tasks.

Two-dimensional representations describe molecular topology explicitly. Here, molecules are treated as graphs, where atoms are nodes and bonds are edges. These can be encoded using adjacency matrices and feature vectors that store atomic and bond-level properties. Graph-based representations are particularly powerful and have become central to modern neural architectures, such as GNNs. However, they require additional preprocessing and may be computationally demanding when applied at scale. For traditional ML approaches, fixed-length fingerprints are the most commonly used 2D descriptors. In this work, we employed Morgan fingerprints (also known as Extended-Connectivity Fingerprints, or ECFP),[42] which iteratively encode atom-centered neighborhoods up to a defined radius, ultimately producing binary or count-based vectors of fixed size (typically 1024 or 2048 bits). Morgan fingerprints are highly efficient, support fast similarity searches, and are well-suited for models such as RFs.

Three-dimensional representations incorporate spatial information, such as atomic coordinates, bond angles, and torsions. This type of descriptors is crucial for modeling shape-based phenomena such as binding affinity or conformational strain. Three-dimensional descriptors can be derived through molecular mechanics or quantum chemical calculations and are used in methods such as docking or pharmacophore modeling.[43] While powerful, they are also computationally expensive to generate.

Finally, learned representations (known as embeddings)[44] are often derived from DL models trained on large corpora of chemical data (e.g., SMILES language models or graph autoencoders), and aim to capture high-level chemical features in a task-specific manner. Unlike hand-engineered descriptors, learned representations adapt to the data and

prediction task, often leading to improved performance but at the cost of reduced interpretability and increased computational requirements.

In this thesis, Morgan fingerprints were selected as the primary descriptors for ML and DL models. They offer a pragmatic balance between interpretability, speed, and predictive power, particularly when handling large-scale HTS data. Their simplicity makes them an ideal choice for iterative modeling and practical deployment.

2.2.3 Machine Learning Algorithms

ML algorithms operate by identifying patterns in large datasets and generalizing those patterns to make predictions on previously unseen inputs. At the core of ML models lie different learning paradigms, with the most common being supervised and unsupervised learning. In unsupervised learning, the goal is to extract recurrent patterns from unlabeled data by identifying similarities between data points based on their features within a space defined by the model, commonly referred to as a latent space. These approaches are useful for grouping similar compounds, compressing data, or initializing models through pretraining.

In contrast, supervised learning relies on labeled data: pairs of molecular structures and corresponding properties or experimental outcomes (e.g., bioactivity against a specific target). The model learns to associate molecular features with known labels, enabling predictions on new, unseen molecules.

Among the most widely used supervised algorithms in cheminformatics are RFs and FFNNs. While both were employed in our works to construct predictive classification (binary labels) QSAR (Quantitative Structure-Activity Relation) models, they differ substantially in their underlying assumptions and learning strategies.

2.2.3.1 Random Forests

RFs are ensemble models that combine the predictions of multiple individual decision trees to improve generalization and reduce the risk of overfitting. Ensemble methods, more broadly, are known to increase robustness and predictive performance by aggregating multiple weak or moderately strong models into a single, more reliable predictor.

In a RF, each decision tree is trained on a different subset of the data, drawn with replacement through a process known as bootstrapping. During training, each tree attempts to identify the combination of input features and threshold values that most effectively separates samples according to their class labels, measured using a purity coefficient (e.g., Gini index). To introduce further diversity, only a randomly selected subset of features is considered at each node when determining the best split. This strategy reduces the

correlation between individual trees and prevents overreliance on dominant features, ultimately improving the generalization performance of the ensemble.

For a binary classification task, each tree T_i produces a prediction $h_i(x) \in \{0, 1\}$, and the final prediction $H(x)$ is determined by majority voting:

$$H(x) = mode\{h_1(x), h_2(x), ..., h_n(x)\} \quad (1)$$

RFs are particularly well suited for cheminformatics problems due to their ability to handle high-dimensional and sparse input vectors, such as Morgan fingerprints, with minimal need for preprocessing or feature scaling. They also offer a degree of interpretability, as the decision process can be traced back through individual trees to identify which input features contributed most to the final prediction.[45]

2.2.3.2 Feed Forward Neural Networks

FFNNs are parametric models composed of multiple layers of artificial neurons organized in subsequent layers. Each neuron performs a linear transformation followed by a nonlinear activation. Given an input vector x, a fully connected layer computes:

$$z(l) = W^{(l)} x^{(l-1)} + b^{(l)}$$

$$x^{(l)} = \sigma(z^{(l)})$$

where $W^{(l)}$ and $b^{(l)}$ are the weights and biases for layer l, and σ is a nonlinear activation function such as ReLU, sigmoid, or tanh. These layers are stacked to form a deep architecture capable of learning increasingly abstract representations of the input data.

Model training is performed by minimizing a loss function L (e.g., cross-entropy for classification or mean squared error for regression) over the training dataset. The model's parameters are updated using backpropagation and gradient descent, which involves computing the gradient of the loss concerning each parameter and applying a small step in the direction that reduces the loss:

$$\theta \leftarrow \theta - \eta \nabla_\theta L$$

where η is the learning rate, and θ denotes the collection of model parameters.

In cheminformatics, neural networks are particularly powerful when working with dense input vectors, such as physicochemical descriptors, or when combined with graph-based architectures that learn directly from molecular structures. While neural networks can be

more data-hungry and less interpretable than tree-based models, they excel at capturing nonlinear and high-order interactions that are difficult to specify a priori.

2.2.4 Evaluating Model Performance

The performance of the classification models developed in this thesis was evaluated using a range of threshold-based and curve-based metrics. These metrics were selected to capture various aspects of model behavior, particularly in the context of class-imbalanced data, where conventional measures such as accuracy can be misleading.

All evaluation metrics used to evaluate the models developed throughout this research project are derived from the confusion matrix, which records the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) produced by a binary classifier. In the context of this thesis work, TP represents correctly predicted interfering compounds, TN corresponds to correctly predicted non-interfering ones, FP are incorrectly flagged compounds, and FN are true interfering compounds that were missed. These values are used to compute a variety of performance measures.

Accuracy, the most intuitive metric, is defined as the proportion of correctly predicted instances over the total number of predictions:

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

While accuracy is commonly reported, it becomes unreliable in highly imbalanced datasets, where a model can achieve high accuracy simply by correctly predicting the majority class.

To address this, more informative metrics such as precision, recall and balanced accuracy are used. Precision, also known as positive predictive value, measures the proportion of true positives among all predicted positives:

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

Recall, or sensitivity, captures the fraction of actual positives that are correctly identified by the model:

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

Balanced accuracy mitigates the effect of class imbalance by averaging the recall of each class. It is defined as:

$$Balanced\ Accuracy = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right)$$

The F1 score combines precision and recall into a single metric, providing their harmonic mean. It is particularly valuable when there is an uneven class distribution, as it penalizes models that perform well on one metric but poorly on the other:

$$F1\ Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

As the principal measure of classification performance in this work, the Matthews Correlation Coefficient (MCC) offers a balanced view by accounting for all elements of the confusion matrix. It is widely regarded as one of the most reliable metrics for imbalanced datasets, and is defined as:

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

MCC values range from −1 (total disagreement between predictions and actual labels) to +1 (perfect prediction), with 0 indicating random classification. Because it incorporates all four components of the confusion matrix, MCC is less susceptible to distortions caused by uneven class distributions.

To assess a model's ability to prioritize true positives early in the ranking, a key consideration in virtual screening scenarios, the Enrichment Factor (EF) was also employed. EF quantifies how much more effective a model is at retrieving true positives in the top-ranked predictions compared to random selection. It is defined as:

$$EF_{x\%} = \frac{TP_{x\%}}{x\% \cdot N_{Actives}}$$

where $TP_{x\%}$ is the number of true positives found in the top x% of ranked predictions, and $N_{Actives}$ is the total number of active compounds in the dataset. An EF greater than 1 indicates that the model outperforms random selection, with higher values reflecting better early enrichment.

In addition to these threshold-based metrics, curve-based metrics were computed to evaluate model behavior across all classification thresholds. The Receiver Operating Characteristic Area Under the Curve (ROC-AUC) plots the true positive rate against the false positive rate as the classification threshold varies. While ROC-AUC is widely used, it can present an overly optimistic picture in imbalanced settings because it incorporates the true negative rate, where negatives are often the dominant class in bioactivity datasets. To address this, the Precision–Recall Area Under the Curve (PR-AUC) was also computed. The PR-AUC focuses on the trade-off between precision and recall, offering a more sensitive evaluation of the model's ability to retrieve positives under class imbalance.

Altogether, these metrics provide a comprehensive and nuanced view of model performance. MCC and EF were the primary indicators used in this work, providing robust assessments of overall classification quality and early enrichment respectively. Meanwhile, PR-AUC provided additional insights into the model's behavior in identifying true positives under practical constraints.

2.2.5 Reinforcement Learning for Molecular de novo Design

RL constitutes a foundational paradigm in AI. Unlike supervised learning, where models are trained on fixed input-output pairs, RL is based on two interacting entities: an agent and an environment. The agent (usually a neural network) observes the state of the environment, takes one action, and receives feedback in the form of a reward signal. Over multiple iterations, the agent learns a policy that maximizes the reward, simulating a trial-and-error learning process (Fig. 2).

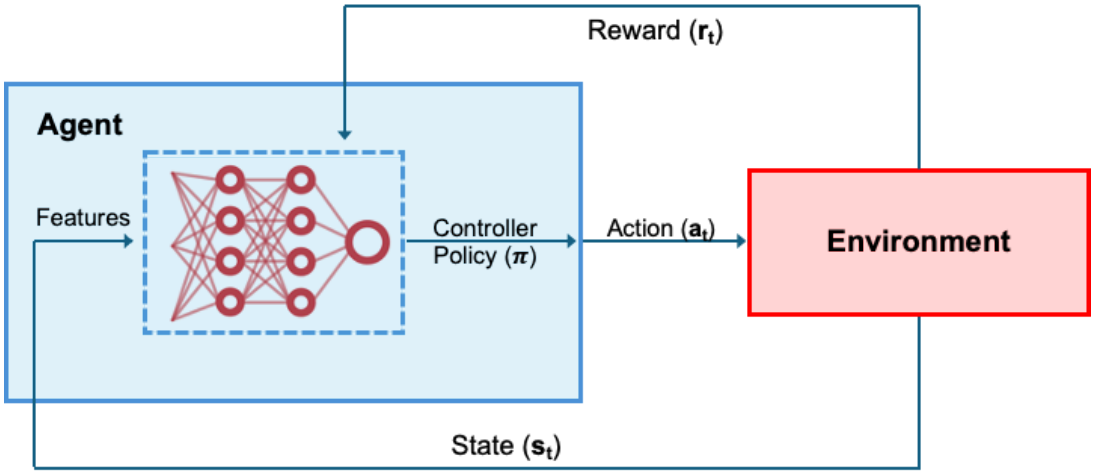


Figure 2: Schematic overview of the iterative process of RL. At each time step t, the agent observes the state s_t of the environment, selects an action a_t, and receives reward r_t as feedback.

In molecular de novo design, the agent is typically a generative model tasked with proposing new molecules in the form of SMILES strings. The environment evaluates each generated molecule using a scoring function that captures how well the compound satisfies predefined metrics. These metrics may include predicted bioactivity from a QSAR model, docking scores, molecular weight, partition coefficient, or toxicity. The agent receives a numerical reward based on the scoring function, and this reward is used to guide further optimization.

REINVENT4 is a state-of-the-art open-source platform for RL in molecular design.[46] The method uses an RNN (recurrent neural network) agent pretrained on a large dataset of molecules to generate valid, chemically meaningful SMILES representations of new molecular designs. During the RL phase, the agent's policy is updated to favor molecules that score highly according to a customizable scoring function.

REINVENT4 supports "human-in-the-loop" workflows, in which expert input can intervene on the generation process by periodically adjusting scoring criteria or reviewing generated candidates.

Within the scope of this thesis, REINVENT4 was used to generate synthetically plausible molecules likely to interfere with assay readouts. The generated structures were subsequently incorporated into the training datasets, augmenting the number of positive examples available to the QSAR models during training, and enhance the generalization ability of interference prediction models. This integration of generative design with discriminative modeling represents a promising step toward more resilient, data-driven screening pipelines.

Chapter 3: Aims

The aim of this doctoral thesis was to develop and evaluate new computational strategies for identifying and triaging compounds that are likely to interfere with biological assays, with the goal of accelerating hit selection in early-stage drug discovery.

Emphasis was placed on the use of HTS data, due to its potential for ML applications and its central role in drug discovery. This work explored how modern ML models and state-of-the-art AI techniques can be integrated into data preparation and modeling workflows to improve the reliability of predictions for assay-interfering compounds.

At the core of this research is the practical interest of understanding how ML can be more effectively streamlined into the HTS hit selection pipeline to better support experimental chemists. The thesis aims to address the following research questions:

- What are the current best-performing models for predicting assay interference, and what are their main limitations?
- How can historical HTS data, often underutilized, be leveraged to train more robust and informative ML models?
- Can the strengths of specialized and global interference prediction models be combined into a hybrid framework that mitigates the weaknesses of each approach?
- Can modern AI tools such as de novo molecular design and human-in-the-loop optimization be integrated to enhance existing predictive models?

To explore these questions, three studies were successfully conducted:

- Study 1 provides a comprehensive review of both experimental and computational methods for dealing with assay-interfering compounds. This review provides the theoretical foundations for the methodological developments in the subsequent studies.
- Study 2 focuses on the development of classical ML models trained on a proprietary HTS dataset provided by Bayer AG. These models aim to predict assay interference, particularly in fluorescence-based assays, by combining characteristics of both global and specialized approaches, with a strong emphasis on data quality and curation.
- Study 3 introduces a novel data augmentation workflow based on RL for de novo molecular design, coupled with expert-guided optimization. This approach enables the generation of plausible interfering structures, which are then incorporated into the training dataset to enhance model robustness and generalizability.

Collectively, these contributions aim to push the boundaries of current interference detection methods by focusing not only on algorithmic innovation but also on the role of data curation and the practical deployment of predictive models in real-world HTS environments.

Chapter 4: Results

4.1 Study 2: Statistical Approaches to Predict Interference with Fluorescence-Based Assays

Palmacci V., Hirte S., Hernández González J. E., Montanari F., and Kirchmair J.

Published in *Artificial Intelligence in the Life Sciences 2024*, 4, 100158.

This chapter is based on a earlier work investigating statistical strategies to identify small molecules that interfere with fluorescence-based assay technologies, one of the most commonly used modalities in HTS. The work presents a novel, data-driven approach to predict interfering compounds using proprietary Bayer AG HTS data. The method applies a frequent hitter analysis to identify patterns of non-specific activity and tested on publicly available data derived from the PubChem database. The tool is then evaluated in terms of predictive performance, interpretability, and practical applicability in early-stage drug discovery. The models and labelign methods developed herein lay the groundwork for automated interference triaging, aiming to reduce false positives in HTS workflows and improve the prioritization of biologically relevant hits.

Author's contribution to the paper:

- Conceptualization of the research idea and designed the study to predict autofluorescent compounds using high-throughput screening data.
- Implemented the full computational workflow, including data extraction from PubChem, statistical label generation, and dataset preprocessing.
- Design and evaluation of experiments, evaluation and analysis.
- Generated all visualizations and figures, and creation of initial draft.

Artificial Intelligence in the Life Sciences 5 (2024) 100099

Contents lists available at ScienceDirect

Artificial Intelligence in the Life Sciences

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

Statistical approaches enabling technology-specific assay interference prediction from large screening data sets

Vincenzo Palmacci [a,b,c], Steffen Hirte [a,b], Jorge Enrique Hernández González [a,d], Floriane Montanari [c,1], Johannes Kirchmair [a,e,*]

[a] Department of Pharmaceutical Sciences, Division of Pharmaceutical Chemistry, Faculty of Life Sciences, University of Vienna, 1090 Vienna, Austria
[b] Vienna Doctoral School of Pharmaceutical, Nutritional and Sport Sciences (PhaNuSpo), University of Vienna, 1090 Vienna, Austria
[c] Department of Machine Learning Research, Bayer AG, 13353 Berlin, Germany
[d] Department of Physics, Sao Paulo State University, Rua Cristóvão Colombo 2265, São José do Rio Preto, CEP 15054-000, Brazil
[e] Christian Doppler Laboratory for Molecular Informatics in the Biosciences, Department for Pharmaceutical Sciences, University of Vienna, 1090 Vienna, Austria

ARTICLE INFO

Keywords:
Machine learning
Assay interfering compounds
Fluorescence
Biological assays
High-throughput screening

ABSTRACT

High throughput screening (HTS) technologies allow the biological testing of hundreds of thousands of compounds per day. Typically, a substantial proportion of the initial hits obtained by HTS are artifacts caused by assay interference. Therefore, global and technology-specific in silico models for identifying and predicting compounds interfering with biological assays have been developed. The global models benefit from training on large screening data sets, while the specialized models benefit from training on assay technology-specific experimental data. In this work, we develop and explore strategies for generating better predictors of technology-specific assay interference by utilizing the large bioactivity data matrices global models are trained on and employing partially new compound labeling approaches to maintain the assay technology awareness of specialized models. We demonstrate the utility of the statistically derived interference labels in machine learning using fluorescence-based assay interference as a representative example. Our random forest and multi-layer perceptron classifiers showed improved performance compared to existing models, achieving Matthews correlation coefficients (MCCs) of up to 0.47 on holdout data and up to 0.45 on an external test set. These results demonstrate that accurate assay-specific interference labels can be derived from large bioactivity data matrices, enabling the development of new machine-learning models without the need for further experimental data.

Introduction

Modern biological assay technologies, automation, and robotics enable the high-throughput screening (HTS) of hundreds of thousands of compounds a day. However, only a fraction of the hits identified with these technologies represent genuine, specific target modulators, while many are the result of assay interference triggered by the compounds [1, 2]. The most common cause of interference across the different assay platforms and technologies is the formation of colloidal aggregates by the test compounds [3]. Among the most relevant types of technology-related interference are autofluorescence (that is, the emission of a fluorescence signal by the compound) and chemical quenching (that is, the reduction of a fluorescence signal by excited state reactions or energy transfer) [4].

Today, HTS facilities have elaborate experimental protocols and techniques in place that are designed to prevent and detect assay interference [1]. These include (i) the screening with non-ionic detergents to prevent compound aggregation [5], (ii) the use of novel fluorophores emitting in a different region of the electromagnetic spectrum (e.g., red light) to reduce spurious signals associated with intrinsic compound fluorescence [6], (iii) the use of orthogonal assays to confirm primary hits, and, (iv) the implementation of counter-screen assays to identify interfering compounds[7]. However, these experimental procedures are resource-demanding and not universally available. Hence, efforts to develop in silico models for predicting compounds likely to cause interference in biological assays have increased in recent years [8].

Machine learning models are now taking the lead in assay

* Corresponding author.
E-mail address: johannes.kirchmair@univie.ac.at (J. Kirchmair).
[1] Current affiliation: Owkin, 14/16 Bd Poissonnière, 75009 Paris, France

https://doi.org/10.1016/j.ailsci.2024.100099
Received 15 March 2024; Received in revised form 3 May 2024; Accepted 27 May 2024
Available online 27 May 2024

interference prediction. They can be categorized into global models and specialized models. Global models are typically trained on large data sets in which hundreds of thousands of compounds are tested against (in total) hundreds to thousands of biological assays. From this bioactivity matrix, the models learn which molecular features make compounds "frequent hitters".

Frequent hitters are compounds showing unexpectedly high hit rates across biological assays. Substantially higher hit rates are often indicative of assay interference (note that a minority of frequent hitters are true promiscuous compounds that interact specifically, reversibly, with multiple, distinct biomacromolecules) [9]. One example of a global model for frequent hitter prediction is HitDexter [10]. HitDexter is trained on a large bioactivity matrix compiled from PubChem BioAssay [11]. For training, compounds were assigned positive "promiscuity" labels if the active-to-tested ratio (ATR; that is, the ratio between the number of assays that reported a compound as active and the total number of assays in which a compound was tested) was above a specified threshold (typically, one to three standard deviations above the average across all compounds). Otherwise, they are assigned negative labels.

The global models for assay interference prediction benefit from the large screening data sets available for training from sources such as PubChem BioAssay. However, they are agnostic of assay technologies, setups, and protocols. This is a considerable limitation because many types of assay interference only occur under specific conditions. For example, compounds causing assay interference in fluorescence-based assays because of their autofluorescence behavior may show benign behavior in other types of assays.

Specialized machine-learning models have been developed for specific assay technologies. They are trained on individual counter-screen and orthogonal assay data sets, typically smaller than the training sets used by the global models. The reported approaches include models for autofluorescence prediction, such as those of David et al. [12], InterPred [13], and ChemFLuo [14], and models for luciferase inhibition prediction, such as Luciferase Advisor [15] and ChemFluc [16]. Several validation studies concluded that the performance and applicability of the specialized models are limited and suggest that the quantity of measured data available for the training of this type of model is likely the bottleneck [12,14,16].

In this work, we develop a statistical approach that enables the development of assay technology-specific predictors of assay interference from large screening data sets, thus leveraging the advantages of global and specialized models. More specifically, the approach compares the hit rates obtained for compounds with a specific assay technology to those obtained with assays based on other technologies. Thus, compounds behaving distinctly with a specific assay technology can be identified.

For assigning bioactivity labels, three statistical methods are explored: the established ATR, a noise-to-active ratio (NAR) that considers both the background signal and bioactivity readout of HTS assays, and Fisher's exact test. To our knowledge, this is the first study in which assay-specific interference is addressed without using counter-screen data to label interfering compounds. We exemplify and validate the suitability of this approach by generating machine-learning models for predicting the interference of small molecules in fluorescence-based assays. The evaluation of the machine learning models with an external test set of of autofluorescent and non-autofluorescent compounds confirms the effectiveness of the proposed labeling approach. Moreover, the performance achieved by our models indicates that augmented data sets can yield better models for the prediction of compounds interfering with fluorescence-based assays.

Methods

Data sets

Measured bioactivity data were extracted from the pool of historical single-dose HTS data available at Bayer AG, employing the following three criteria for assay and compound selection:

1. Assays must be designed to probe the interaction of a compound with a specific protein or pathway
2. Assays must have bioactivity recorded for at least 80 % of the compounds in the dataset
3. Compounds must have bioactivity recorded for at least 80 % of the assays.

In this framework, the bioactivity readouts from the HTS assay are already preprocessed and normalized [17], and reported as z-score values. For compounds measured multiple times in the same assay, averaged z-score values were used.

The resulting bioactivity data set (hereafter referred to as "Bayer AG data set") comprises 1,488,407 compounds measured in a total of 99 biochemical and 88 cell-based assays using different fluorescent dyes (56, 23, and 8 assays using blue, green, and red fluorescent dyes, respectively), FRET or TR-FRET technologies (14 and 10 assays, respectively), and bioluminescence detection systems (76 assays). For all compounds, non-isomerical canonical SMILES were generated with RDKit [18]. As part of this process, minor (i.e., salt) components were removed with the ChEMBL Structure Pipeline [19]. This processing sometimes led to multiple instances (molecules) with identical canonical SMILES representations but inconsistent bioactivity readouts. These inconsistencies were removed to avoid conflicting bioactivity data. The processed data set holds 1,441,034 compounds, with 296,830,391 readouts recorded with a total of 187 different assays.

In preparation for model development, the data set was partitioned into a "training set", a "validation set", and a "test set" using a molecular scaffold split strategy. This procedure included (i) the generation of Morgan fingerprints (radius 2; length 2048 bits) for all compounds (with RDKit), (ii) the generation of Murcko scaffolds (with RDKit), (iii) the grouping of molecules sharing the same scaffold and (iv) the random assignment of groups of molecules sharing the same scaffold to the training set (80 % of the data), validation set (10 %), or test set (10 %). The use of a scaffold split strategy aimed to exclude trivial examples from the validation and test sets, thus providing a test set for evaluating the models' generalization power.

Following the work of Yang et al. [14], an external test set was compiled from PubChem Bioassay AID 1696 and AID 720678. AID 1696 is a fluorescence-based counter-screen assay for detecting fluorescence artifacts in HTS assays for inhibitors of the *Mycobacterium tuberculosis* cell wall enzymes RmlC (AID 1532) and RmlD (AID 1533). AID 720678 is a cell-based HTS assay designed to identify compounds that exhibit autofluorescence at a wavelength of 460 nm. The compounds were preprocessed using the same protocol described for the Bayer AG data set. The remaining 10,691 compounds were then compared to all compounds present in the training set, and any test compound with a maximum Tanimoto Coefficient (TC) of 1.0 was removed from the test set. The final external test set consisted of 10,031 unique structures.

A set of known aggregators [20] served as a further test set for the models. After preprocessing with the identical protocol described above, the set consisted of 12,643 known aggregators. None of these aggregators is present in the Bayer AG training set.

The preprocessed set of known aggregators was complemented with an equal number of compounds randomly selected from the ChEMBL database [21] to represent the negative class (i.e., putative non-aggregators). Only compounds structurally distinct from those in the Bayer AG training set were considered.

Compound labeling

The readouts of all 187 assays were converted via z-scores into binary activity labels applying|z-score| ≥ 4 as the threshold (in a standard normal distribution or z-distribution, a z-score exceeding 4 corresponds to a p-value ≪ 0.01, providing statistical significance about the observations). In addition, for the 56 assays based on blue fluorescence, the measured background emission signals (i.e., emission signals at time zero) were converted into binary "noise labels" applying |z-score| thresholds of 1.0.

From the binarized assay readouts, binary labels describing the capacity of a compound to interfere with assays based on blue fluorescence were derived utilizing (i) active-to-tested ratios (ATRs), (ii) Fisher's exact test, and (iii) ratios calculated for the background noise and the bioactivity signal (noise-to-active ratios, or NARs).

The ATR-based approach puts two ATR values into perspective: the ATR calculated from the outcomes of all assays based on blue fluorescence (ATR_{Fluo}) and the ATR based on the outcomes of all non-fluorescence-based assays (ATR_{Other}). Positive assay interference labels were assigned to any compounds for which:

$$ATR_{Fluo} \geq E[ATR_{Other}] + t \cdot \sigma(ATR_{Other}) \quad (1)$$

where $E[ATR_{Other}]$ and $\sigma(ATR_{Other})$ are the mean ATR and the ATR standard deviation computed among all non-fluorescent assays, respectively. t allows tuning of the threshold to define the occurrence of interference. For the threshold t, values of 0.9, 1.0, 3.0 and 5.0 were explored.

For the NAR-based approach, the background noise binary labels were put into perspective. Any compound for which the following condition holds:

$$positive_{background}/positive_{main} \geq t \quad (2)$$

was labeled as interfering with the assay technology. Here, $positive_{background}$ and $positive_{main}$ represent the number of times each compound yields a positive readout in the background and bioactivity matrices, respectively. Different thresholds (i.e., t equal to 0.03, 0.04, 0.07, or 0.10) for assigning the NAR interference labels were explored.

For the approach based on Fisher's exact test, the binarized main signal readouts from the blue fluorescence-based assays and the non-fluorescence-based assays were compared. More specifically, a contingency table was computed, putting into perspective the two different assay domains (i.e., blue fluorescence-based assays and non-fluorescence-based assays), and the observed activity counts, i.e., the number of times a compound is recorded as active or inactive. Based on Fisher's exact test, compounds for which:

$$p\text{-}value \leq t \quad (3)$$

were flagged as assay-interfering compounds, with threshold t equal to 0.35, 0.17, 0.07, or 0.01.

For the public test sets derived from PubChem AID 1696 and AID 720678, the existing binary compound activity labels were used for model testing.

Machine-learning model development

Random forest classifiers

Random forest (RF) models were generated with the BalancedRandomForestClassifier class of the imbalanced-learn Python library [22]. The hyperparameters n_estimators, max_depth, and min_sample_split (Table S1) were optimized during 50 iterations of Bayesian hyperparameter search using the BayesianOptimization Python library [23] using the training set for model generation and the validation set for performance evaluation.

Neural networks

Multi-layer perceptrons (MLP) for classification were generated with PyTorch (v1.11) [24]. The binary cross-entropy loss was employed for the neural network training, and ReLu was chosen as the activation function. To address the class imbalance in the training set, a weighted oversampling strategy of the minority class was implemented. The hyperparameters, including the number of layers, units per layer, the learning rate, and the dropout ratio (Table S2), were optimized with 50 iterations of hyperparameters optimization run using the Optuna hyperparameters optimization framework [25].

Results

Characterization of the data set employed for the computation of assay interference labels

To compute interference labels, we compiled an extensive bioactivity data set from Bayer AG historical HTS primary screenings database. The compounds in the pool of historical single-dose HTS data comprise 1,441,034 unique structures measured at Bayer AG in a total of 187 different assays, covering a substantial part of the biological target space spanning 114 protein targets. The compounds included in the Bayer AG data set represent 227,258 unique Murcko scaffolds. Ninety-eight percent of the Murcko scaffolds are associated with fewer than 45 compounds and collectively account for 51 % of the data. These numbers indicate the broad chemical space covered by the data set employed in this study.

To further assess the relevance of the Bayer AG data set to drug discovery, we compared the physicochemical properties of the compounds present in this data set with those included in the Approved Drugs subset of DrugBank [26]. As shown in the principal component analysis (PCA) scatter plot reported in Fig. 1, the Bayer AG data set envelops the chemical space of approved drugs and extends beyond it, further indicating the high value and broad coverage of the data set.

Analysis of strategies for assigning assay interference labels

Three metrics for labeling compounds as interfering or not interfering with fluorescence-based assays were explored: the ATR (Eq. (1)), the NAR (Eq. (2)), and p-values derived with Fisher's exact test (Eq. (3)). These metrics were investigated using four metric thresholds for assigning binary assay interference labels to the individual compounds

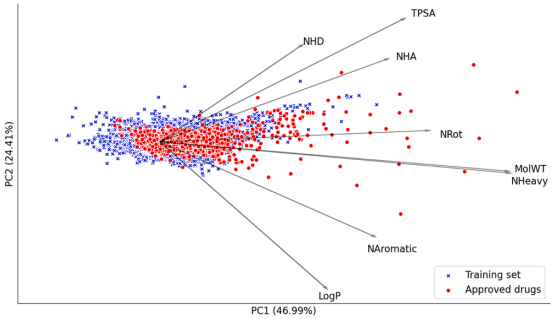


Fig. 1. PCA scatter plot with associated loadings of the chemical space covered by the 1,441,052 compounds of the training dataset (in blue) and the 4359 approved drugs (in red) reported in the DrugBank database. The PCA was performed on eight physicochemical properties computed with RDKit: partition coefficient (logP), molecular weight (MolWT), number of H-bond acceptors (NHA), number of H-bond donors (NHD), number of rotatable bonds (NRot), heavy atoms count (NHeavy), topological polar surface area (TPSA), and number aromatic rings (NAromatic). The percentages of the total variance explained by the plotted eigenvectors are indicated in the axis labels.

Table 1
ATR, NAR, and p-Value Thresholds Applied to Yield the Indicated Percentages of Interference Compounds in the Bayer AG Training set.

% of interference compounds	ATR	NAR	p-value from Fisher's exact test
~2 %[1]	5.00	0.10	0.01
~5 %[1]	3.00	0.07	0.07
~10 %[1]	1.00	0.04	0.17
~20 %[1]	0.90	0.03	0.35

[1] The exact percentages of compounds interfering with the assay technology yielded by each threshold are reported in Table S3.

(Table 1). The thresholds were chosen to yield, as closely as possible, 2 %, 5 %, 10 %, and 20 % of the compounds labeled as interference compounds. The first three rates are particularly relevant to investigate because compound libraries have been reported to typically contain 2 to 5 % of compounds exhibiting autofluorescence in the blue spectrum [4, 27], and approximately 10 % of compounds showing frequent hitter behavior [11]. The 20 % rate was included to study the impact of class distribution on model performance.

We analyzed the overlap of the sets of molecules labeled as assay interference compounds (by the statistical approaches, applying the four defined thresholds) to investigate the relationship between the three labeling strategies (Fig. 2). The overlap of the three labeling strategies for assay interference compounds amounted to approximately 43 % of the total number of assay interference compounds (at any investigated threshold). The ATR and NAR labeled similar compounds as assay interference compounds. Specifically, 70 % of all assay interference compounds were identified by both the ATR and NAR (at the 2 % and 5 % thresholds, Fig. 2A and B), with NAR assay interference compounds progressively becoming a subset of ATR positives as the number of assay interference compounds increased.

When comparing the compounds identified as interfering with the assay technology, Fisher's exact test showed greater similarity with ATR at the 2 % threshold, with about 60 % overlapping compounds (Fig. 2). Unlike what was observed for ATR and NAR, for Fisher's exact test, when more interfering compounds are found, differences with other compound labeling strategies arise. Specifically, only 30 % of the compounds flagged with Fisher's exact test overlap with those identified by ATR.

As pointed out earlier, the study shows a considerable overlap

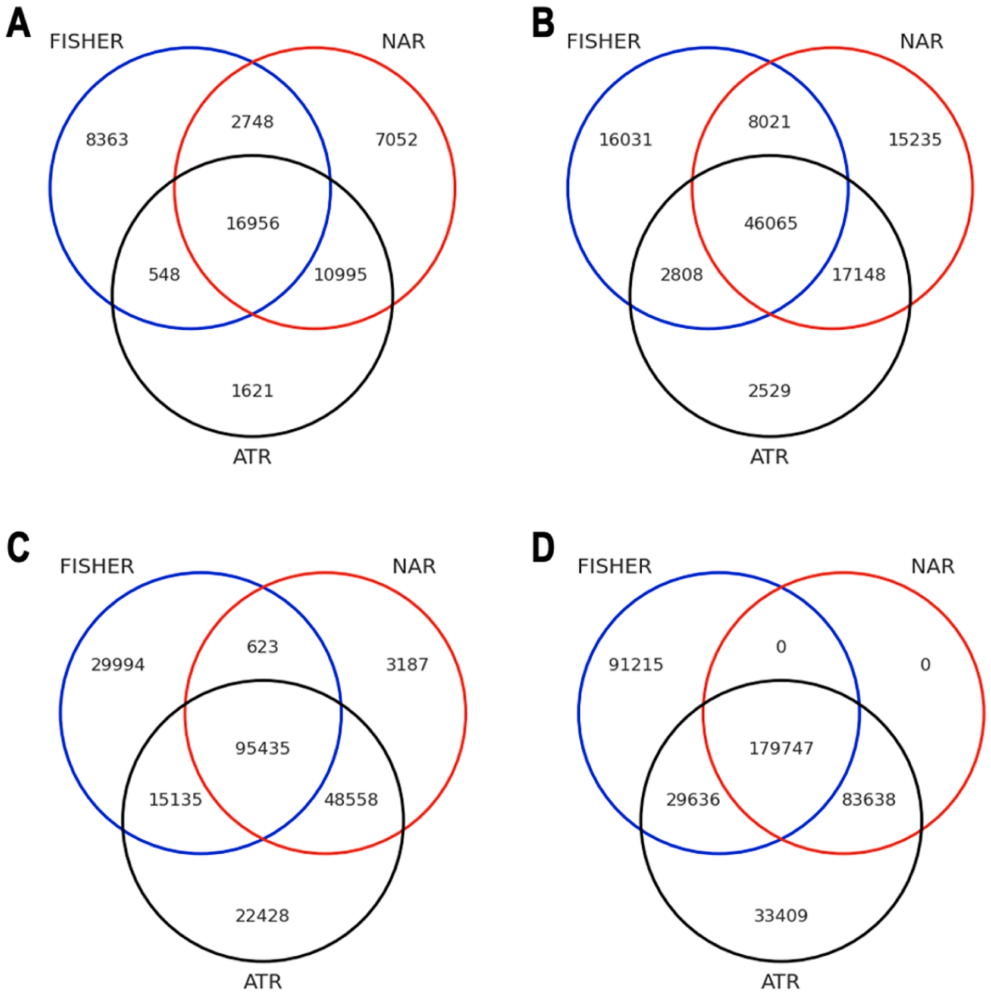


Fig. 2. Numbers of molecules labeled as interference compounds based on the ATR, NAR, and p-values derived from Fisher's exact test, computed using thresholds for labeling compounds as frequent hitters of (A) 2 %, (B) 5 %, (C) 10 %, and (D) 20 %.

between the ATR and NAR strategies, indicating similarity in the compounds identified as assay interfering. However, differences emerge when comparing Fisher's exact test with ATR and NAR, particularly as the number of yielded interfering compounds increases. This suggests that, while there is some consistency between the labeling strategies, each approach may capture unique subsets of interfering compounds, highlighting the importance of considering multiple strategies for comprehensive analysis.

Characterization of the data sets employed for the development and testing of machine learning models

From the 1,441,034 compounds in the pool of historical single-dose HTS data available at Bayer AG, we selected 1,130,711 unique compounds using a scaffold split strategy to train machine learning models. The compounds included in the Bayer AG training set represent 197,095 unique Murcko scaffolds. Similarly to the observations made for the complete data set, 98 % of the Murcko scaffolds are associated with fewer than 45 compounds and collectively account for 51 % of the training data. These numbers indicate the broad chemical space covered by the data set employed for training the new machine-learning models.

In analogy to the work of Yang et al. [14], a test set was prepared to evaluate the validity of the devised labeling strategies and machine learning models. The test set was derivedfrom two counter-screen data sets for the detection of autofluorescent compounds in the blue emission spectrum (λ∼490 nm), PubChem Bioassay AID1696 and AID720678 (see Methods for details). The processed test set (duplicated structures and compounds overlapping with the training set removed) consists of 10,031 compounds (382 autofluorescent and 9649 non-autofluorescent) based on 2549 unique Murcko scaffolds, of which 266 scaffolds cover the autofluorescent compounds. Ninety-nine percent of the Murcko scaffolds represent fewer than 45 compounds and collectively account for 51 % of the test data, confirming the high molecular diversity of the test set.

Fig. 3 visually represents the chemical space covered by the PubChem-derived test set using TMAP [28]. The TMAP shows examples of individual autofluorescent compounds embedded in chemical spaces otherwise occupied by non-autofluorescent compounds (panels A and B), and examples of autofluorescent compounds embedded in clusters of autofluorescent compounds (panels C and D). It is expected that the latter represent simpler cases for machine learning classifiers.

To understand the structural relatedness of the compounds included in the Bayer AG training set and the PubChem-derived test set, we calculated the pairwise maximum similarity between the compounds from both data sets based on Tanimoto coefficients based on Morgan fingerprints (Fig. 4). The median of all pairwise maximum similarities is 0.53, with 50 % of the values between 0.42 and 0.63. These numbers indicate that a substantial proportion of the compounds in the test set are structurally distinct from those in the Bayer AG training set.

Development of machine learning classifiers for assay interference prediction

Prior to model building, the Bayer AG dataset was split, based on Murcko scaffolds (Table 2), into three subsets: a training set (80 % of the data), a validation set (10 % of the data), and a test set (10 % of the data). Importantly, the percentage of interference compounds was maintained across the data set splits.

A total of 24 models were generated, resulting from the combination of two machine learning algorithms (i.e., RF and MLP classifiers), three labeling strategies (i.e., ATR, NAR, and Fisher's exact test), and four thresholds for labeling compounds as (non) interfering with fluorescence-based assays (Table 1). The established Morgan fingerprints (radius 2; length 2048 bits) were employed as molecular representations for all models.

Hyperparameter optimization was performed through Bayesian

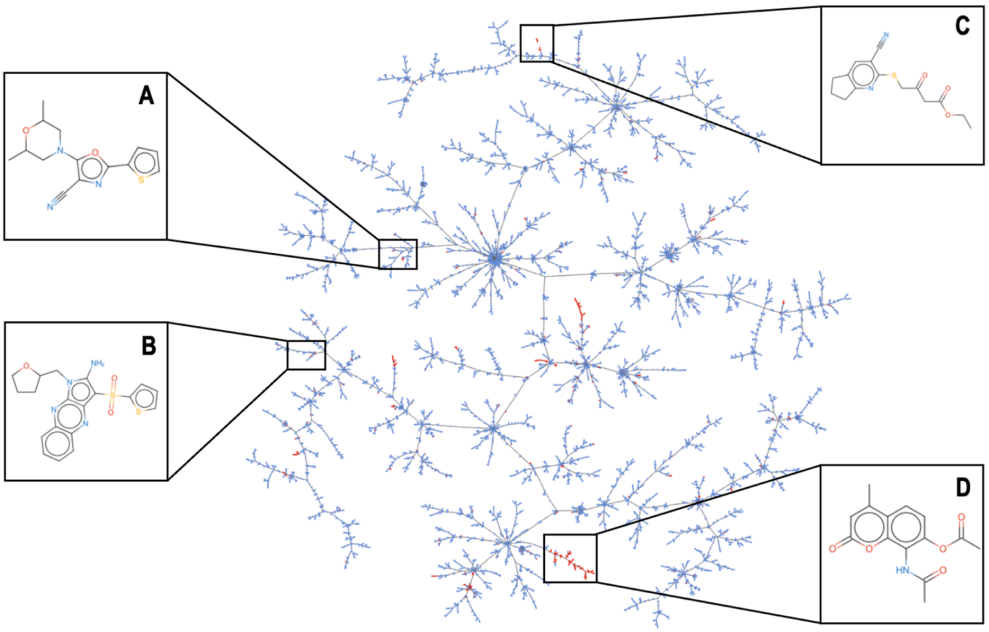


Fig. 3. TMAP of the PubChem-derived test set chemical space, with experimentally confirmed autofluorescent compounds in red and experimentally confirmed non-autofluorescent compounds in blue. (A) and (B) show examples of individual autofluorescent compounds embedded in chemical spaces otherwise occupied by non-autofluorescent compounds. (C) and (D) show examples of autofluorescent compounds embedded in clusters of autofluorescent compounds. The TMAP was computed using Morgan fingerprints with a radius of 2 and a length of 2048 bits.

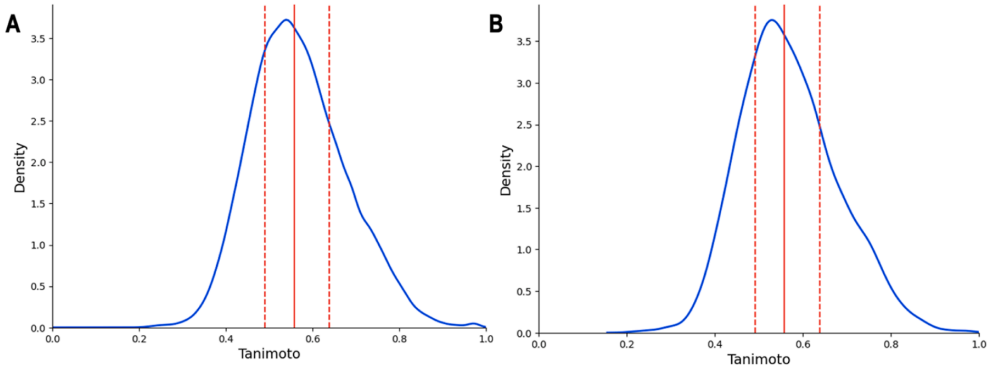


Fig. 4. Density plots visualizing the distribution of maximum pairwise similarities (quantified as Tanimoto coefficients calculated on Morgan fingerprints with a radius of 2 and a length of 2048 bits) between the compounds of the Bayer AG training set and (A) the full PubChem-derived test set and (B) the subset of known autofluorescent compounds of the PubChem-derived test set. In each panel, the continuous vertical line indicates the distribution median, while the two dashed vertical lines encompass the interquartile range.

Table 2
Number of Compounds and Murcko Scaffolds Represented in the Training, Validation, and Test Sets Derived from the Bayer AG Dataset.

Bayer AG dataset derived	# of compounds	# of Murcko scaffolds
Training Set[1]	1,130,711	197,095
Validation Set[1]	135,830	48,559
Test Set[1]	174,493	58,932

[1] The numbers of compounds interfering with the assay technology are reported in Table S4, Table S5, and Table S6 for the training, validation, and test set, respectively.

optimization for the RF (optimizing the number of trees, maximum depth of each tree, and bootstrap usage) and through Optuna optimization for the MLP (optimizing the number of neurons per layer, number of layers, dropout rate, and learning rate), using the Bayer AG training set for iterative model training and the Bayer AG validation set for performance evaluation (Tables S1 and S2, respectively).

The RFs consistently outperformed the MLPs on the Bayer AG test set. Across the various setups, both RFs and MLPs achieved their best performance on data labeled with the NAR strategy, with the highest MCC among the RF and MLP models being 0.47 and 0.43, respectively (Table 3). For data labeled based on the ATR or Fisher's exact test, MCCs were, on average, in the range of 0.37. Noticeably, when employing ATR and Fisher's exact test for data labeling, the models perform best with a 10 % threshold. Regardless of the threshold used, all models consistently retrieved the majority of interference compounds, as evidenced by recall values exceeding 0.73 across the investigated setups.

To further investigate variations across different setups and explore the potential applications of the developed models in hit-deprioritization scenarios, we examined the Receiver Operating

Table 3
Model Performance on the Bayer AG Test Set and the PubChem Bioassay-derived Test Set.

Algorithm	Labeling method	%interference compounds threshold	Bayer AG Test Set[1]				PubChem Bioassay-derived test set			
			MCC[2]	AUC	Precision	Recall	MCC[2]	AUC	Precision	Recall
RF	ATR	20 %	0.42	0.87	0.34	0.81	0.35	0.91	0.20	0.75
		10 %	0.45	0.87	0.40	0.75	0.36	0.91	0.21	0.75
		5 %	0.38	0.91	0.23	0.81	0.33	0.91	0.18	0.73
		2 %	0.25	0.94	0.09	0.86	0.31	0.90	0.17	0.73
	NAR	20 %	0.47	0.85	0.47	0.72	0.35	0.91	0.19	0.79
		10 %	0.44	0.88	0.36	0.77	0.35	0.91	0.19	0.81
		5 %	0.41	0.91	0.27	0.80	0.36	0.91	0.21	0.75
		2 %	0.33	0.94	0.15	0.86	0.35	0.91	0.19	0.78
	p-Values derived from Fisher exact test	20 %	0.41	0.80	0.49	0.61	0.41	0.93	0.31	0.63
		10 %	0.41	0.86	0.32	0.73	0.42	0.93	0.29	0.71
		5 %	0.37	0.89	0.23	0.76	0.45	0.94	0.34	0.66
		2 %	0.29	0.92	0.12	0.80	0.44	0.94	0.32	0.69
MLP	ATR	20 %	0.38	0.84	0.30	0.80	0.20	0.80	0.10	0.66
		10 %	0.42	0.84	0.37	0.71	0.21	0.79	0.13	0.57
		5 %	0.43	0.89	0.42	0.50	0.22	0.82	0.20	0.32
		2 %	0.32	0.91	0.17	0.70	0.26	0.79	0.22	0.39
	NAR	20 %	0.43	0.82	0.47	0.65	0.22	0.80	0.12	0.63
		10 %	0.41	0.85	0.34	0.73	0.24	0.82	0.13	0.66
		5 %	0.43	0.88	0.34	0.67	0.23	0.83	0.19	0.37
		2 %	0.41	0.90	0.39	0.47	0.23	0.81	0.25	0.25
	p-Values derived from Fisher exact test	20 %	0.34	0.76	0.40	0.65	0.17	0.78	0.10	0.66
		10 %	0.39	0.82	0.36	0.58	0.22	0.78	0.13	0.57
		5 %	0.36	0.86	0.23	0.71	0.26	0.79	0.20	0.32
		2 %	0.35	0.88	0.31	0.44	0.20	0.85	0.22	0.39

[1] Performance on the validation set used for hyperparameter optimization is reported in Table S7.
[2] Boxplot representations comparing the MCC distributions yielded by models trained with different labeling strategies on the test set are reported in Figures S1 and S2.

Characteristic (ROC) curves generated from predictions on the Bayer AG test set. Fig. 5 presents the ROC curves for predictions made by RF in different setups. When employing the 2 % (Fig. 5A) and 5 % (Fig. 5B) thresholds, the models promptly identify interfering compounds, as shown by the high early enrichment of the curves (slope of the ROC between FPR=0.0 and FPR=0.2). This suggests the models' suitability for hits-deprioritization, allowing the ranking of compounds based on the probability of interference for subsequent triaging. Conversely, a lower enrichment factor is observed with the 10 % (Fig. 5C) and 20 % (Fig. 5D) labeling thresholds. Nevertheless, as discussed earlier, these models achieve the highest MCCs on the test set.

Performance of the machine learning classifiers on external autofluorescence data

We investigated the validity of the newly devised labeling approaches for in silico hits-deprioritization by evaluating the performance of all 24 models on an external autofluorescence dataset sourced from PubChem Bioassay. This dataset was compiled in accordance with the guidelines outlined by Yang et al. [14]. and consists of 10,031 unique structures with experimentally derived binary autofluorescence labels. Most of the compounds in this autofluorescence data set exhibit low to no similarity with the compounds used to train the models (see "Characterization of the data sets employed for the development and testing of machine learning models"), presenting a challenging scenario for our models and simulating a real-world situation.

Consistent with the advantage of RFs over MLPs observed for the Bayer AG validation and test sets, the RFs also performed better on the external test set (Table 3). The differences in the chemical space represented in the training data and this external test set did not lead to a substantial drop in model performance despite the substantial differences in chemical spaces between the external test set and the training set. Noticeably, the NAR-derived models experienced the largest decline in performance. While these models performed well on the Bayer AG test set (average MCC=0.41), they exhibited a substantial decrease in performance on the public test set, yielding an average MCC of 0.35. Similarly, ATR-based models showed a decrease in performance compared to the Bayer AG test set results, achieving an average MCC of 0.34 on the PubChem-derived test set. In contrast, the models trained on data labeled with Fisher's exact test outperformed the rest, achieving a maximum MCC of 0.45. These models not only maintained consistent performance across different datasets but also demonstrated enhanced

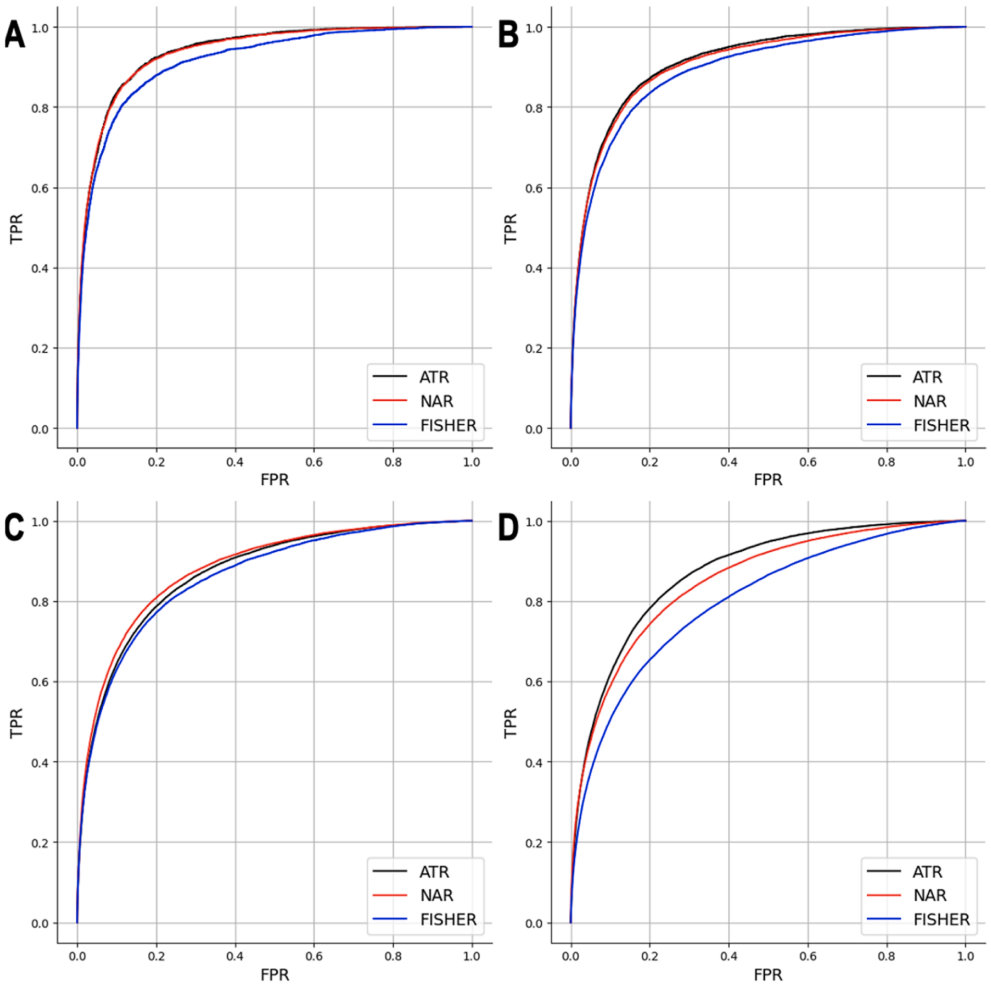


Fig. 5. ROC curves obtained for the Bayer AG test set with the RF classifiers trained using different labeling methods and using the (A) 2 %, (B) 5 %, (C) 10 %, and (D) 20 % compound labeling thresholds.

accuracy in predicting autofluorescent compounds (using 5 % and 2 % thresholds), underscoring the suitability of various setups for different scenarios.

Unlike the RF models, all MLP models displayed a pronounced decrease in performance. While the average MCC values for ATR, NAR, and Fisher's trained models on the Bayer AG test set were 0.39, 0.42, and 0.39, respectively, the MCC dropped to 0.22, 0.23, and 0.21 on the public test set (Table 3).

To better understand the concordance among different methods for labeling compounds according to their behavior in fluorescence-based assays, we investigated the overlap of predicted interfering compounds among the best-performing models. As shown in Table 3, regardless of the labeling strategy employed, RFs consistently outperformed MLP classifiers. Fig. 6 visually represents the agreement on interfering compounds predicted by the best-performing models, specifically ATR using a 10 % threshold, NAR, and p-values derived from Fisher's exact test, both using the 5 % threshold. The Venn diagram in Fig. 6 highlights a noticeable difference between the labeling methods. Whereas Fisher's exact test and NAR identified a comparable number of molecules as assay interference (919 and 722, respectively), there was only a 27 % overlap between the positive predictions from Fisher's exact test and those from NAR. This observation highlights that different labeling strategies may capture distinct patterns associated with interference. In contrast, ATR identified 1697 molecules as assay interfering, suggesting that the variation in the number of predicted interfering compounds may be attributed to the different thresholds employed (5 % for Fisher's exact test and NAR, 10 % for ATR).

The alignment among model predictions becomes even more evident when examining the TMAPs in Fig. 7. Figs. 7A and 7B highlight the agreement between ATR and NAR models, whereas Fisher's exact test labels fewer compounds as positives. Notably, compared to the ground truth autofluorescence labels (Fig. 3), all models accurately identify regions in the chemical space densely populated by autofluorescent compounds, confirming the validity of our approach.

All models label more compounds as assay interference compounds than the ground truth (382 compounds). This behavior could be attributed to the fact that all the labeling strategies explored in this study are intended to identify interfering compounds (in this example, for fluorescence-based assays only) regardless of the interference mechanism. It should be noted that the additional compounds labeled as interfering by our models have not undergone experimental validation to confirm their interference status.

The public test set reports experimental results only for autofluorescent compounds. Our method correctly recognizes structural clusters highly populated with autofluorescent compounds. More challenging examples (i.e., compounds not associated with any structural cluster of autofluorescent compounds) are also recognized as autofluorescent by our models (Fig. 7), corroborating the validity of our labeling approach and the applicability of the developed models.

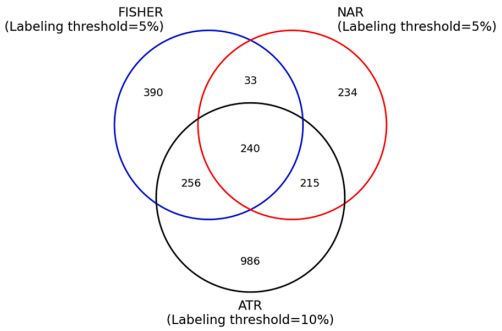


Fig. 6. Numbers of molecules labeled as interference compounds by the best-performing models derived using the ATR, NAR, and p-values derived from Fisher's exact test as labeling strategies. The threshold applied to the training data is reported in parentheses.

Furthermore, to understand how our models compete with existing tools, we compared the performance of our top-performing and existing models on the PubChem-derived test set. As shown in Table 4, our model outperformed HitDexter 3.0 and ChemFLuo, achieving the highest MCC of 0.45. While our best-performing model achieves a lower recall score than other approaches, its higher precision suggests a strength in accurately identifying interfering compounds. These findings highlight the superior predictive capability of our model, affirming its effectiveness in accurately detecting interfering compounds within the PubChem test set. As highlighted in Table 4, our model achieved a favorable balance between precision and recall (MCC metric). While the high recall achieved by the other models is indicative of their ability to capture as many interference compounds as possible, in a hits-triaging context, prioritizing precision ensures that the identified compounds are truly interfering, minimizing false positives. Hence, while our model may exhibit a lower recall than prior models, the high precision score safeguards against false positives, thereby retaining potentially valuable compounds and enhancing the overall reliability of predictions.

It is important to note that differences in training data contribute to variations in model performance. ChemFLuo utilizes PubChem-derived counter-screen data for autofluorescence detection, while HitDexter 3.0 relies on curated HTS screening datasets for predicting frequent hitters. Consequently, the lower performance associated with the HitDexter 3.0 model is to be expected due to differences in the training task, as HitDexter is agnostic of the assay method and not specialized in predicting the behavior of fluorescence-based assays. Overall, these results further validate our labeling strategy for developing machine-learning models to predict assay interference, demonstrating the effectiveness of utilizing HTS bioactivity data for model training.

Prediction of known aggregators

To further validate our labeling approaches for assay interference compounds, we utilized a set of known aggregators to assess how many of them would be recognized by our models. More specifically, we compared the percentage of predicted interfering compounds between the set of known aggregators and a random subset of 12,643 compounds selected from the ChEMBL database. The bar plots depicted in Fig. 8 illustrate that all RF classifiers predict a higher percentage of assay interference compounds within the set of known aggregators compared to a random compound sample retrieved from the ChEMBL database. As expected, models trained using higher thresholds (20 % and 10 %) performed best in predicting aggregators. When using the ATR and NAR labeling strategies with a 20 % threshold, there is a notable difference in the proportion of compounds predicted as assay interference compounds between the aggregator dataset and the ChEMBL subset. In this case, 11 % and 10.5 % more compounds were predicted as assay interference compounds in the aggregator dataset compared to the ChEMBL subset, using the ATR and NAR labeling strategies, respectively. With Fisher's exact test, the best results were obtained using a 10 % threshold. However, our models are unable to predict the entire set of aggregators as interfering, potentially because they are not specifically tailored for this task.

Conclusions

In this work, we explore strategies for generating predictors of technology-specific assay interference from large screening data sets. More specifically, we studied three statistical methods for assigning assay interference labels (ATR, NAR, and Fisher's exact test) in combination with four thresholds that can be tuned to yield different rates of interfering compounds.

We explored the generation of machine-learning models using the

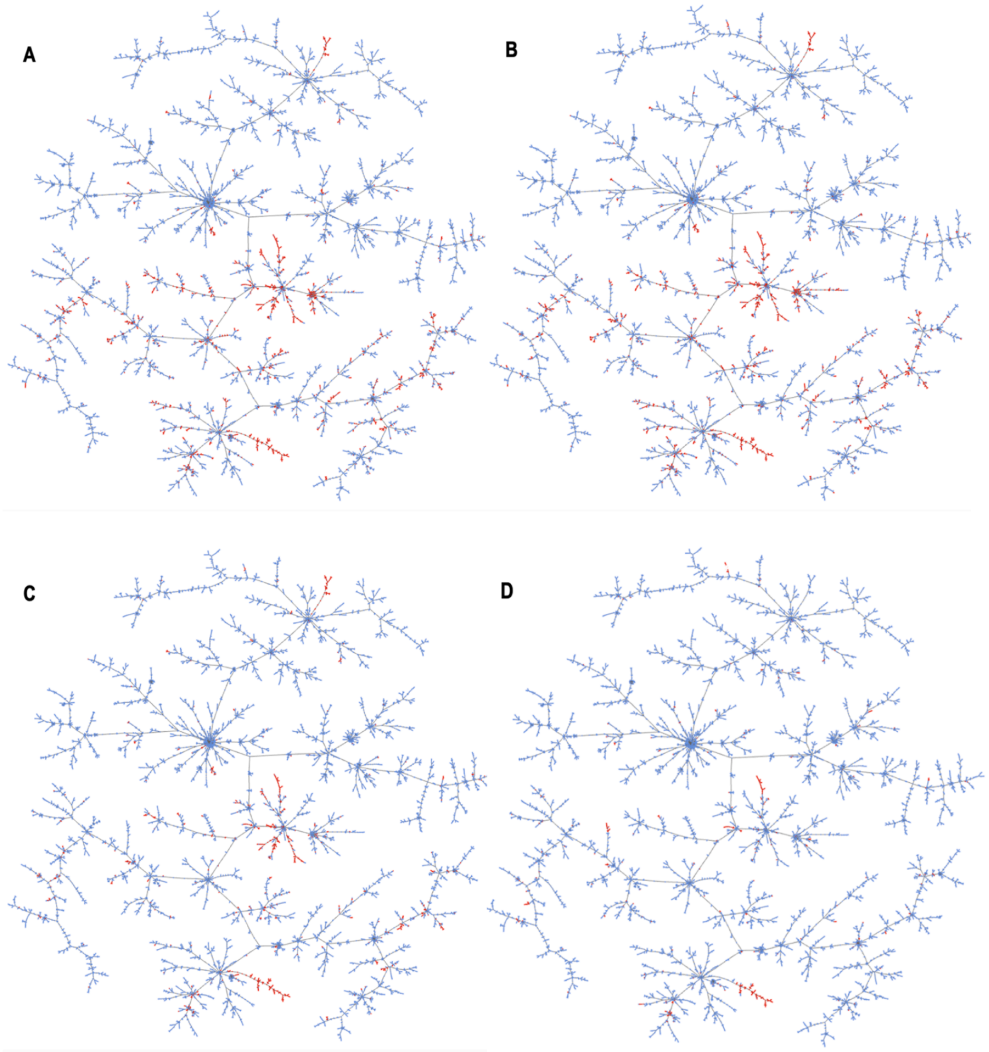


Fig. 7. TMAP of the PubChem-derived test set chemical space colored by predicted interference label using the best performing models: (A) RF trained on data labeled using the ATR strategy with a 10 % threshold, (B) RF trained on data labeled using the NAR strategy with a 5 % threshold, and (C) RF trained on data labeled using Fisher's exact test with a 5 % threshold. (D) TMAP of the Pubchem-derived test set colored by the ground truth labels. Interfering compounds are reported in red, and non-interfering compounds in blue. The TMAPs were computed using Morgan fingerprints with a radius of 2 and a length of 2048 bits.

Table 4
Comparison of Performance Obtained on the Test Set Derived from PubChem Bioassay.

	MCC	AUC	Precision	Recall
HitDexter3.0	0.25	0.84	0.18	0.84
ChemFLuo	0.34	0.92	0.23	0.75
Best RF model of this work	**0.45**	0.94	0.34	0.66

aforementioned labels to predict technology-specific assay interference. Testing on an external set demonstrated the superior performance of RF models over existing models in predicting fluorescence-based assay prediction, emphasizing the effectiveness of the proposed models in capturing interfering compounds.

The study showcases the utility of leveraging historical HTS data to address assay-specific interference, enhancing machine-learning model predictions. The developed approach can potentially improve chemical library design and streamline hit-triaging in HTS, offering flexibility in choosing labeling strategies and thresholds tailored to specific screening needs. The proposed framework may also be generalized to other assay technologies, such as bioluminescence, thereby contributing to advancements in machine learning models.

CRediT authorship contribution statement

Vincenzo Palmacci: Conceptualization, Data curation, Investigation, Writing – original draft, Writing – review & editing. **Steffen Hirte:** Validation, Writing – original draft, Writing – review & editing. **Jorge Enrique Hernández González:** Conceptualization, Supervision, Writing – original draft, Writing – review & editing. **Floriane Montanari:** Conceptualization, Supervision, Writing – original draft, Writing

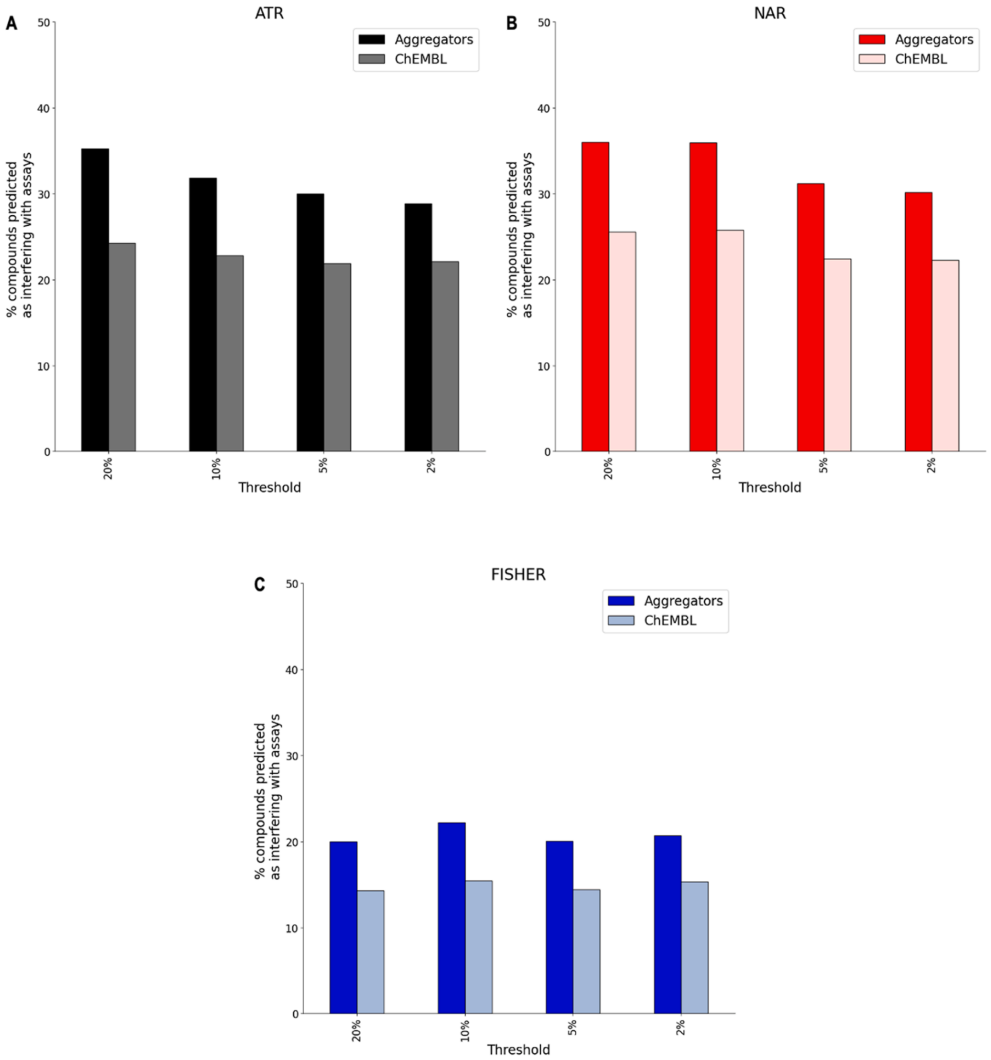


Fig. 8. Bar plots comparing the percentage of predicted interfering compounds between the set of known aggregators and the ChEMBL subset yielded by RF classifiers using (A) ATR, (B) NAR, and (C) Fisher's exact test as labeling strategies. The percentages of predicted interfering compounds yielded by the MLP classifiers are reported in Figure S3.

– review & editing. **Johannes Kirchmair:** Conceptualization, Funding acquisition, Supervision, Writing – original draft, Writing – review & editing.

Declaration of competing interest

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

Floriane Montanari and Johannes Kirchmair are Editorial Board Members for *Artificial Intelligence in the Life Sciences* and were not involved in the editorial review or the decision to publish this article.

Data availability

The PubChem-derived dataset for testing the performance of the machine-learning models and the source code of the approach presented in this work are available at https://github.com/Bayer-Group/PISA-T/. Due to the proprietary nature of the Bayer AG HTS historical data used in this work, the raw data used for the models' training would remain confidential.

Funding

VP and JK gratefully acknowledge the support from the European Commission's Horizon 2020 Framework Programme (AIDD; grant no. 956832). JEHG thanks the Sao Paulo Research Foundation (Grant: 2022/03901–8) for financial support. JK gratefully acknowledges the financial support received for the Christian Doppler Laboratory for Molecular Informatics in the Biosciences by the Austrian Federal Ministry of Labour and Economy, the Austrian National Foundation for Research, Technology and Development, the Christian Doppler Research Association, Boehringer-Ingelheim RCV GmbH & Co KG and BASF SE.

Acknowledgments

We thank Roxane Jacob and Matthias Welsch from the University of Vienna, Christian Doppler Laboratory for Molecular Informatics in the Biosciences, and Vincent-Alexander Scholtz, also from the University of Vienna, for their insightful discussions regarding the development of machine learning models. We thank Djork-Arné Clevert from Bayer AG Berlin for providing valuable feedback, and Stefan Mundt and Michael Koch from Bayer AG Wuppertal for enabling us to work with their measured data and providing expert advice. Furthermore, we thank the anonymous Reviewers for their valuable feedback and suggestions.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT to improve the readability of the manuscript. After using this tool/service, the authors reviewed and edited the content as needed and took full responsibility for the content of the publication.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ailsci.2024.100099.

References

[1] Tan L, Hirte S, Palmacci V, Stork C, Kirchmair J. Tackling assay interference associated with small molecules. Nat Rev Chem. 2024;8:319–39.
[2] Sink R, Gobec S, Pečar S, Zega A. False positives in the early stages of drug discovery. Curr Med Chem 2010;17:4231–55.
[3] Ganesh AN, Donders EN, Shoichet BK, Shoichet MS. Colloidal aggregation: from screening nuisance to formulation nuance. Nano Today 2018;19:188–200.
[4] Coussens NP, Auld D, Roby P, Walsh J, Baell JB, Kales S, Hadian K, Dahlin JL. Compound-Mediated Assay Interferences in Homogeneous Proximity Assays. Assay Guidance Manual (eds Markossian, S.), NCATS 2020.
[5] Blay V, Tolani B, Ho SP, Arkin MR. High-Throughput Screening: today's biochemical and cell-based approaches. Drug Discov Today. 2020;25:1807–21.
[6] Fan F, Wood KV. Bioluminescent assays for high-throughput screening. Assay Drug Dev Technol 2007;5:127–36.
[7] Thorne N, Auld DS, Inglese J. Apparent activity in high-throughput screening: origins of compound-dependent assay interference. Curr Opin Chem Biol 2010;14:315–24.
[8] Yang ZY, He JH, Lu AP, Hou TJ, Cao DS. Application of negative design to design a more desirable virtual screening library. J Med Chem 2020;63:4411–29.
[9] Schneider P, Schneider G. Privileged structures revisited. Angew Chem Int Ed 2017;56:7971–4.
[10] Stork C, Mathai N, Kirchmair J. Computational prediction of frequent hitters in target-based and cell-based assays. Artificial Intelligen. Life Sci. 2021;1:100007.
[11] Kim S, Chen J, Cheng T, et al. PubChem 2023 update. Nucleic Acids Res 2023;51:D1373–80.
[12] David L, Walsh J, Sturm N, Feierberg I, Nissink JWM, Chen H, Bajorath J. Engkvist, O. Identification of compounds that interfere with High-Throughput Screening assay technologies. Chem Med Chem. 2019;14:1795–802.
[13] Borrel A, Mansour K, Nolte S, Saddler T, Conway M, Schmitt C, Kleinstreuer N. InterPred: a webtool to predict chemical autofluorescence and luminescence interference. Nucleic Acids Res 2020;48:W586–90.
[14] Yang Z, Dong J, Yin M, Jiang H, Lu A, Chen X, Hou T, Cao D. ChemFLuo: a web-server for structure analysis and identification of fluorescent compounds. Brief. Bioinformatics 2021;22:bbaa282.
[15] Ghosh D, Koch U, Hadian K, Sattler M, Tetko IV. Luciferase Advisor: high-accuracy model to flag false positive hits in luciferase HTS assays. J Chem Inf Model 2018;58:933–42.
[16] Yang ZY, Dong J, Yang ZJ, Lu A, Hou T, Cao D. Structural analysis and identification of false positive hits in luciferase-based assays. J Chem Inf Model 2020;60:2031–43.
[17] Malo N, Hanley JA, Cerquozzi S, Pelletier J, Nadon R. Statistical practice in high-throughput screening data analysis. Nat Biotechnol 2006;24:167–75.
[18] RDKit: Open-source cheminformatics. [https://www.rdkit.org] (accessed March 15, 2024).
[19] Bento AP, Hersey A, Félix E, et al. An open source chemical structure curation pipeline using RDKit. J Cheminform. 2020;12:51.
[20] Irwin JJ, Duan D, Torosyan H, Doak AK, Ziebart TK, Sterling T, Tumanian G, Shoichet BK. An aggregation advisor for ligand discovery. J Med Chem 2015;58:7076–87.
[21] Davies M, Nowotka M, Papadatos G, Dedman N, Gaulton A, Atkinson F, Bellis L, Overington JP. ChEMBL web services: streamlining access to drug discovery data and utilities. Nucleic Acids Res 2015;43:612–20.
[22] Lemaître Gl, Nogueira F, Aridas CK. Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. J Mach Learn Res. 2017;18:1–5.
[23] Bayesian Optimization: Open source constrained global optimization tool for Python. [https://github.com/bayesian-optimization/BayesianOptimization] (accessed February 13, 2024).
[24] Paszke A, Gross S, Chintala S, Chanan G, Yang E, DeVito Z, Lin Z, Desmaison A, Antiga L, Lerer A. PyTorch: an Imperative Style, High-Performance Deep Learning Library. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems; 2019.
[25] Takuya A, Shotaro S, Toshihiko Y, Takeru O, Masanori K. Optuna: A Next-generation Hyperparameter Optimization Framework. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2019.
[26] Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, Gautam B, Hassanali M. DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res 2008;36:D901–6.
[27] Jadhav A, Ferreira RS, Klumpp C, Mott BT, Austin CP, Inglese J, Thomas CJ, Maloney DJ, Shoichet BK, Simeonov A. Quantitative analyses of aggregation, autofluorescence, and reactivity artifacts in a screen for inhibitors of a thiol protease. J Med Chem 2010;53:37–51.
[28] Probst D, Reymond JL. Visualization of very large high-dimensional data sets as minimum spanning trees. J Cheminform. 2020;12:12.

4.2 Study 3: E-GuARD: expert-guided augmentation for the robust detection of compounds interfering with biological assays

Palmacci V., Nahal Y., Welsch M., Engkvist O., Kaski S., Kirchmair J.

Published in *Journal of Cheminformatics 2025*, 17, 54.

This chapter introduces E-GuARD, a novel DL framework developed to enhance the detection of small molecules that cause assay interference. The approach integrates self-distillation, active learning, and expert-guided de novo molecule generation to address the challenges of data scarcity and class imbalance that hamper traditional QSAR modeling. E-GuARD enriches training datasets with chemically diverse and interference-relevant compounds by iteratively generating new molecules using REINVENT4 and selecting them based on acquisition functions informed by human-like preferences (MolSkill). The study demonstrates how this framework significantly improves the performance of assay interference prediction models across multiple interference mechanisms, including thiol and redox reactivity, as well as firefly and nanoluciferase inhibition. The results highlight the promise of combining generative models and human expertise simulation to enhance the robustness and generalizability of in silico screening strategies in early drug discovery.

Author's contribution to the paper:

- Conceptualized the research idea and designed the E-GuARD framework for improving assay interference prediction through data augmentation and expert-informed molecule generation.
- Implemented the full computational workflow, including self-distillation, uncertainty sampling, and reinforcement learning-based molecular generation using REINVENT4.
- Designed and evaluated all experiments across multiple assay interference mechanisms.
- Generated all visualizations and figures, and drafted the manuscript, incorporating revisions from collaborators.

Palmacci *et al. Journal of Cheminformatics* (2025) 17:64
https://doi.org/10.1186/s13321-025-01014-3

Journal of Cheminformatics

RESEARCH Open Access

E-GuARD: expert-guided augmentation for the robust detection of compounds interfering with biological assays

Vincenzo Palmacci[1,2†], Yasmine Nahal[3,4†], Matthias Welsch[1,2,5], Ola Engkvist[4,6], Samuel Kaski[3,7] and Johannes Kirchmair[1,5*]

Abstract Assay interference caused by small organic compounds continues to pose formidable challenges to early drug discovery. Various computational methods have been developed to identify compounds likely to cause assay interference. However, due to the scarcity of data available for model development, the predictive accuracy and applicability of these approaches are limited. In this work, we present E-GuARD, a novel framework seeking to address data scarcity and imbalance by integrating self-distillation, active learning, and expert-guided molecular generation. E-GuARD iteratively enriches the training data with interference-relevant molecules, resulting in quantitative structure-interference relationship (QSIR) models with superior performance. We demonstrate the utility of E-GuARD with the examples of four high-quality data sets on thiol reactivity, redox reactivity, nanoluciferase inhibition, and firefly luciferase inhibition. Our models reached MCC values of up to 0.47 for these data sets, with two-fold or higher improvements in enrichment factors compared to models trained without E-GuARD data augmentation. These results highlight the potential of E-GuARD as a scalable solution to mitigating assay interference in early drug discovery.

Scientific contribution We present E-GuARD, an innovative framework that combines iterative self-distillation with guided molecular augmentation to enhance the predictive performance of QSAR models. By allowing models to learn from newly generated, informative compounds through iterations, E-GuARD facilitates the understanding of underrepresented structural patterns and improves performance on unseen data. When applied across different interference mechanisms, E-GuARD consistently outperformed standard approaches. E-GuARD establishes the foundation for further research into dynamic data enrichment and more robust molecular modeling.

†Vincenzo Palmacci and Yasmine Nahal have contributed equally to this work.

*Correspondence:
Johannes Kirchmair
johannes.kirchmair@univie.ac.at
[1] Department of Pharmaceutical Sciences, Division of Pharmaceutical Chemistry, Faculty of Life Sciences, University of Vienna, 1090 Vienna, Austria
[2] Vienna Doctoral School of Pharmaceutical, Nutritional and Sport Sciences (PhaNuSpo), University of Vienna, 1090 Vienna, Austria
[3] Department of Computer Science, Aalto University, Espoo, Finland
[4] Molecular AI, Discovery Sciences, BioPharmaceuticals R&D, AstraZeneca, Gothenburg, Sweden
[5] Christian Doppler Laboratory for Molecular Informatics in the Biosciences, Department for Pharmaceutical Sciences, University of Vienna, 1090 Vienna, Austria
[6] Department of Computer Science and Engineering, Chalmers University of Technology, Gothenburg, Sweden
[7] Department of Computer Science, University of Manchester, Manchester, UK

Introduction

High-throughput screening (HTS) is of fundamental importance to modern drug discovery, allowing for the rapid assessment of hundreds of thousands of compounds for activity on biomacromolecular targets of interest [1]. However, a substantial number of hits reported by HTS technologies may be linked to assay interference caused by compound aggregation, direct interference with the detection methods, or nonspecific chemical reactions with assay components [2, 3]. Assay interfering compounds are often called "bad actors" or "nuisance compounds" [4], and are common to chemical libraries, representing a bottleneck for the early drug development pipeline.

Today, various experimental approaches, such as counter-screenings or orthogonal assays, are routinely employed to identify assay-interfering compounds and false-positive assay readouts [2, 5]. While these experimental methods are essential in drug and probe discovery, they are substantial in cost and cannot always be applied retrospectively during chemical library screening and preparation, particularly when working with large libraries. On the other hand, computational methods have emerged as a promising alternative for predicting assay interference [2] as they offer a complementary strategy by identifying potential assay interference patterns earlier in the discovery process, which may help prioritize compounds for experimental follow-up and optimize screening resources.

Among the most relevant models are several machine learning approaches, including HitDexter 3.0 [5, 6], which predicts compounds likely to show frequent hitter behavior, and PISA-T [7], which flags compounds likely to interfere with fluorescence-based assays. These models leverage extensive HTS data sets but are agnostic of interference mechanisms.

Considering the mechanisms underlying assay interference phenomena can improve the accuracy and relevance of predictions. Therefore, researchers have explored strategies for training machine learning approaches on experimental assay interference data obtained, e.g., via counter-screens and orthogonal assays [8–10]. However, data scarcity and class imbalance prove challenging for model development.

To expand the availability of measured data for model development, Alves et al. [11] selected compounds from the NPACT data set and subsequently tested them in-house using HTS assays. This ensured all experimental data were generated under consistent conditions, minimizing variation across assays. Their data collection is among the most comprehensive in the field and includes molecular structures associated with measured data on thiol reactivity (TR), redox reactivity (RR), nanoluciferase inhibition (NI), and firefly luciferase inhibition (FI). Each of the four interference classes is represented by approximately 5,000 measured compounds, with interference rates ranging from 1.5% to 20%, depending on the data set. The authors demonstrated that their data collection is suitable for training reliable machine learning models for assay interference prediction. This resulted in the development of the "Liability Predictor", an online tool featuring XGBoost-based quantitative structure-interference (QSIR) models that accurately identify interfering compounds.

Although the compiled data sets are of great value to research, challenges related to data scarcity and class imbalance remain. Theoretical approaches to address class imbalance range from basic techniques, such as over-sampling the minority class or under-sampling the majority class [12], to more advanced methods, such as weighted loss functions for hill-climbing algorithms [13]. Data augmentation is a further, effective strategy to alleviate class imbalance and data scarcity by generating synthetic examples to enrich the training set [14, 15]. For example, techniques such as SMOTE [16] (Synthetic Minority Over-sampling Technique) are widely utilized in cheminformatics for their flexibility and straightforward implementation. More recent applications of data augmentation include the introduction of noisy labels via self-distillation, which is the process of first training a "teacher" model on labeled data and then using its predictions to train a "student" model with the same architecture [17]. Self-distillation has been empirically observed to provide model performance gains on various tasks, including image recognition [18] and protein structure prediction [19]. Building on this concept, Liu et al. [20] developed the Pseudo Label Augmented Neural System (PLANS) for quantitative structure–activity relation (QSAR) modeling applications. PLANS uses a teacher model trained on fully labeled data to generate pseudo-labels for a large pool of unlabeled compounds collected from the ChEMBL database. This self-distillation approach enhanced predictive performance when tested on cytochrome P450 substrate prediction and Tox21 data sets [21]. However, as stated by the authors, PLANS introduces a significant amount of noise, likely due to the introduction of model-generated labels when using the complete ChEMBL database as a source of unlabeled data. This noise can confuse the model, resulting in a notable decline in performance.

To overcome the limitations and challenges of existing approaches, in this work, we combine self-distillation with tailored molecular generation and active learning [22, 23] in a new framework that we call E-GuARD (Expert-Guided Augmentation for the Robust Detection of Compounds Interfering with Biological Assays).

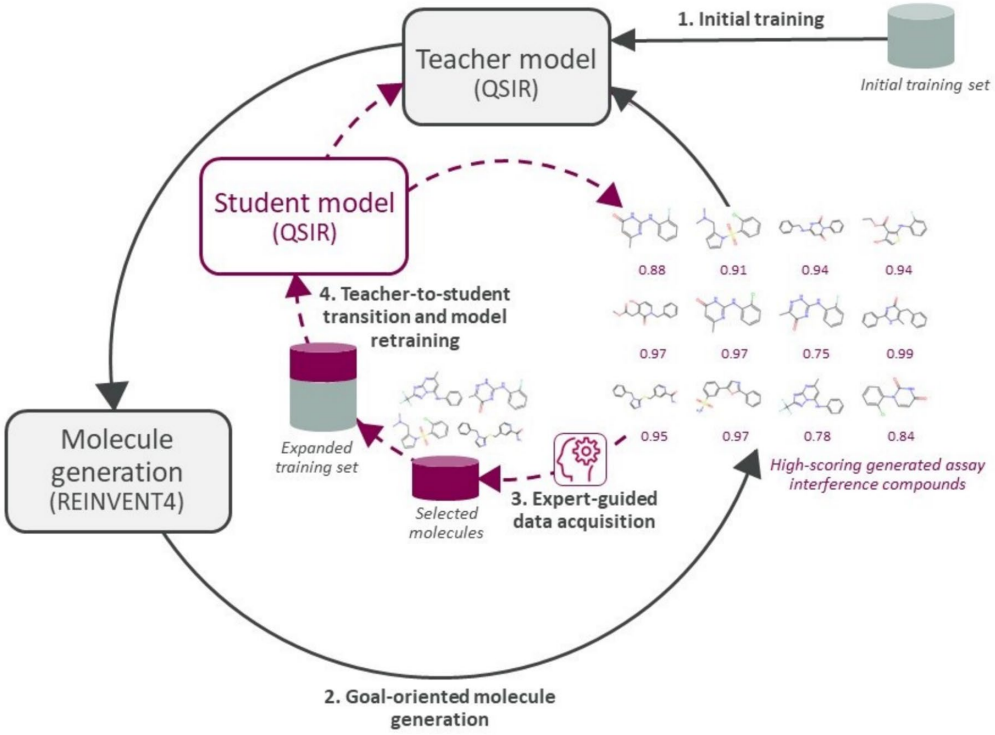


Fig. 1 Overview of the E-GuARD workflow, which involves the iterative process of molecular generation, expert-guided data augmentation, and self-distillation. First, a teacher model is trained. The teacher model is then used to guide molecule generation towards interfering compounds (outer loop represented by black arrows). Once a pre-defined number of outer loop iterations has been completed, the teacher model becomes the student model and is iteratively updated through expert-guided data augmentation and self-distillation (inner loop represented by dashed red arrows)

E-GuARD (Fig. 1) builds upon the concept of self-distillation, with two key distinctions: (i) instead of sourcing unlabeled data from existing data sets (e.g., the ChEMBL database), E-GuARD generates new chemical structures with the de novo molecular design tool REINVENT4 [24]; (ii) E-GuARD adds unlabeled data to the training set following expert guidance emulated with MolSkill [25]. This approach is designed to balance exploration of the chemical space with targeted refinement, leveraging both de novo molecule generation and expert-guided feedback to optimize the discovery process.

Specifically, E-GuARD starts by enhancing an initial, small training set by iteratively adding selected compounds from a pool of molecules generated with REINVENT4. The teacher model guides the algorithm toward relevant regions of the chemical space. The loop is executed for a defined number of iterations (in our study, five iterations), with new molecules selected using one of five acquisition functions. At the end of each iteration, the teacher model transitions into the student role and is retrained on the augmented training data set, enabling continuous refinement and improvement for each subsequent iteration. To minimize noise and ensure that compounds considered are relevant to drug discovery, a method to proxy human feedback is included in the optimization cycle. The proxy human feedback is generated with MolSkill, a neural network model developed to emulate the decision-making process of medicinal chemists. MolSkill's feedback is used in combination with two acquisition functions to select molecules to be added to the training set. This component indirectly injects human expertise into the reinforcement learning loop of REINVENT4, resulting in the generation of drug-like molecules based on diverse scaffolds.

We explored E-GuARD's ability to improve the prediction of four mechanisms of assay interference: TR, RR, NI, and FI. For each interference mechanism, we performed ten independent runs of iterative self-distillation. Compared to baseline QSIR models, the QSIR models generated with the E-GuARD approach showed

improved predictive performance across both internal and external test sets.

Analyzing the student model's evolution, we show that E-GuARD improves the detection of compounds interfering with biological assays, by enabling the learning of new features across iterations. Additionally, evaluating the QED scores and diversity of the generated compounds shows that the newly added compounds remain diverse and relevant to drug discovery. This demonstrates that the observed performance improvements stem from learning novel features.

Materials and methods

Data collection

Sets of measured data on the interference of 5,098 compounds with biological assays via TR, RR, NI, and FI were obtained from Alves et al. [11]. For each data set, 25% of the compounds were randomly selected and assigned to a test set. The remaining 75% of compounds were divided into five subsets of equal size for cross-validation and hyperparameter optimization (Table 1). All splits preserved the class distribution of the initial data set.

To evaluate the impact of E-GuARD on QSAR model performance more rigorously, we conducted an external validation using a dataset derived from PubChem for firefly luciferase interference (AID411), which was previously employed in the Luciferase Advisor study [26]. SMILES and corresponding interference labels were obtained from PubChem, and to exclude any data leakage, compounds overlapping with the training set were identified by converting molecules to Morgan3 fingerprints and conducting an exact match search. This resulted in the removal of 24 molecules. The final external dataset comprised 70,619 unique compounds, including 1571 interfering and 69,048 non-interfering compounds, none of which were utilized during model training.

E-GuARD workflow

E-GuARD utilizes a teacher-student loop to enhance the prediction of interfering compounds. This loop, illustrated in Fig. 1, consists of four key steps:

1. *Initial Training of the Teacher Model* A QSIR model is initially trained on the available training data set.
2. *Goal-Oriented Molecule Generation* New molecules are generated and scored using the teacher model.
3. *Expert-Guided Data Acquisition* Compounds are selected using one of five acquisition functions, including expert-based scoring with MolSkill.
4. *Teacher-to-Student Transition and Model Retraining* The training set is augmented with the selected compounds, and the student model is retrained. The student model then becomes the teacher for the next iteration.

The following sections explain the details underlying the four steps and the tools utilized.

Teacher-student model for interference prediction (QSIR)

The balanced random forest (BRF) classifier algorithm, implemented in the imbalanced-learn Python library [27], was chosen as the QSIR teacher model to provide a baseline consistent with the "Liability predictor" [11] performances (see Table S1). The BRF classifier addresses class imbalance by creating bootstrapped subsets with equal representation of each class, thereby mitigating the dominance of the majority class.

A dedicated BRF model was trained for each data set, with hyperparameters (n_estimators, max_depth, min_samples_split, and max_features) optimized using Optuna [28] over 50 trials throughout a fivefold cross-validation procedure. As the input of the machine learning models, the molecular representation of choice was Morgan 3 fingerprints, with a bit length of 2048 bits, generated using the RDKit [29].

The BRF classifiers are then retrained on the augmented training set using the same hyperparameter set determined during the initial optimization.

Table 1 Overview of the data sets employed in this work

Type of interference	Training set		Test set	
	# positive compounds	# negative compounds	# positive compounds	# negative compounds
Thiol reactivity (TR)	811	3035	198	764
Redox reactivity (RR)	113	3781	29	945
Firefly luciferase inhibition (FI)	90	3834	28	953
Nanoluciferase inhibition (NI)	82	3836	15	965

Goal-oriented molecule generation with REINVENT4

The training set is augmented with new molecules generated with REINVENT4. REINVENT4 uses a reinforcement learning framework to optimize molecular generation based on a custom scoring function. In this study, the scoring function for the RL agent feedback, $f(x)$, was defined as Eq. 1,

$$f(x) = \omega_1 m(x) + \omega_2 wt(x) \tag{1}$$

where $m(x)$ represents the interference score computed as the predicted probability of the compound x to be an interfering compound according to the QSIR model, and $wt(x)$ is a molecular weight score designed to prioritize compounds typically relevant to small-molecule drug discovery (160–480 Da) [30]. The weights ω_1 and ω_2 were set to 0.8 and 0.2 respectively, with normalization applied.

During each generation step, the REINVENT4 agent executes 250 iterations of optimization, generating 100 compounds per iteration. By the end of the optimization process, a total of 25,000 compounds have been generated. This molecule pool is then filtered using one of five acquisition functions (see the next section for a detailed description) to select the 250 most informative compounds. These compounds are added to the training set to augment the data and retrain the BRF classifier in subsequent iterations.

Active data acquisition

Active learning (AL) was employed to iteratively select and add informative compounds to the training set, aiming to increase the likelihood that the added compounds would be diverse and relevant to the modeled task of assay interference prediction.

At the end of each REINVENT4 generation cycle, a pool of compounds U_r is generated. The generated compounds are obtained by maximizing the reward given by the pre-defined scoring function (Eq. 1). Then, an acquisition criterion is applied to select a subset of compounds from U_r (Eq. 2) according to

$$A(x) = \alpha A_{\text{predictor}}(x) + \beta A_{\text{human}}(x) \tag{2}$$

where $\alpha A_{predictor}(x)$ corresponds to the model evaluation score and $\beta A_{human}(x)$ corresponds to the simulated human evaluation score. α and β are weighting constants.

First, we employed three different acquisition strategies with β set to 0 so that the compound selection from U_r is based solely on $A_{predictor}(x)$:

1. Random Selection: Molecules are randomly selected from U_r.
2. Greedy Selection: The molecules with the highest predicted probabilities of interference are selected from U_r focusing on the most confident model predictions.
3. Expected Predictive Information Gain (EPIG) Selection [31]: The most informative molecules are selected from U_r based on their ability to reduce the predictive uncertainty within the top 1000 molecules in U_r. For each x in U_r, EPIG calculates the expected mutual information between the interference labels of x and a randomly sampled x^* from the target set of top high-scoring 1000 molecules. Mathematically, EPIG is formulated as the expected Kullback–Leibler divergence between the joint distribution $p(y, y^* \mid x, x^*)$ and the product of marginals $p(y \mid x)p(y^* \mid x^*)$.

Additionally, the Greedy and EPIG selection strategies were combined with an expert-guided criterion $A_{human}(x)$, with both α and β set to 1. In this work, $A_{human}(x)$ was simulated using the MolSkill score. The following expert-guided acquisition criteria were employed:

4. EPIGSkill: The most informative molecules are selected from U_r using an integrative scoring system that combines the EPIG score with the expert preference score predicted by MolSkill.
5. GreedySkill: The most informative molecules are selected from U_r based on an integrative score that combines the Greedy score and expert preference score as predicted by MolSkill.

The various acquisition functions, combined with the simulated expert scoring, steer the generation of compounds toward distinct chemical spaces. This impacts predictor performance and allows the approach to be tailored to specific task requirements.

Expert-guided data augmentation with MolSkill

MolSkill is a neural network designed to emulate medicinal chemists' decision-making processes during lead optimization in drug discovery. It applies learning-to-rank techniques to prioritize molecules based on desirability criteria such as drug-likeness and synthetic accessibility. MolSkill is trained on preference feedback from 35 Novartis chemists of varying expertise. They were presented with pairs of drug candidates through a graphical interface and asked to select their preferred option.

In this work, MolSkill was applied as an expert scoring function for data acquisition to identify the most

expert-desirable generated molecules to be incorporated into the training set. The inclusion of MolSkill into E-GuARD aims to enhance the model's robustness in detecting challenging, hard-to-detect assay interference compounds that exhibit desirable drug-like properties.

Evaluation metrics

For QSIR model performance

The Matthews Correlation Coefficient (MCC) was used as the primary measure of model performance (Eq. 3). The MCC is a balanced metric that takes the true positive (TP), false positive (FP), true negative (TN), and false negative (FN) instances into account:

$$\mathrm{MCC} = \frac{\mathrm{TP} \cdot \mathrm{TN} - \mathrm{FP} \cdot \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}} \tag{3}$$

The MCC returns values between − 1 (total disagreement between prediction and observation) and + 1 (perfect agreement).

The Enrichment Factor (EF) [32] was employed to evaluate the ability of the model to prioritize true positives, providing a prevalence-adjusted measure of precision (Eq. 4). The EF is a measure of the factor by which a model enriches relevant outcomes compared to random chance:

$$\mathrm{EF} = \frac{\mathrm{Precision}}{\frac{\mathrm{TP+FP}}{\mathrm{TP+FP+TN+FN}}} \tag{4}$$

EF values greater than 1 indicate that the model effectively enriches true positives.

For the generated compounds

Imbalance rate (IR) puts into perspective the proportion of negative and positive examples of the data set, giving a measure for the class imbalance (Eq. 5). IR is computed at each iteration t as:

$$\mathrm{IR}(t) = \left| \frac{\mathrm{Positives(t)\text{-}Negatives(t)}}{\mathrm{Positives(t)} + \mathrm{Negatives(t)}} \right| \tag{5}$$

Internal chemical diversity assesses the chemical diversity within a molecular set G (Eq. 6). The metric is limited to [0, 1], and a higher value corresponds to higher diversity in the generated set. Internal chemical diversity is measured as:

$$\mathrm{IntDiv_p}(G) = 1 - \sqrt[p]{\frac{1}{|G|} \sum_{m_1, m_2 \in G} \mathrm{Tanimoto}(m_1, m_2)^p} \tag{6}$$

where G corresponds to the set of generated molecules, represented in this study by their 2048-bit Morgan2 fingerprint vectors. We primarily consider $p = 1$ in this work.

The QED score [30] is a quantitative metric of a compound's drug-likeness. It is based on a combination of physicochemical properties commonly associated with successful drugs, such as reasonable molecular weight and lipophilicity. The QED score ranges from 0 to 1, where higher scores indicate a compound is more likely to have desirable drug-like properties.

The number of structures matching at least one PAINS alert is determined using the RDKit Python library through the FilterCatalog. PAINS method. This method employs a set of predefined substructure filters to identify molecules likely to produce false-positive results in HTS assays. The PAINS filter [33] captures structural motifs associated with assay interference, such as reactive groups, or with promiscuous binding properties. While many of these PAINS substructures are associated with redox or thiol reactivity, they can also interfere through other mechanisms, including aggregation and singlet oxygen quenching.

Model analysis

Centered kernel alignment (CKA) is a method for quantifying the similarity of embedding distributions. Values for CKA range from 0 to 1, where 1 indicates high similarity [34]. Initially developed for NNs, CKA was recently adopted to better capture the similarity between RFs ($\mathrm{CKA_{rf}}$) [35], by utilizing a random forest kernel. The random forest kernel compares two data instances based on how often they share a partition in the decision trees of the RF [36]. $\mathrm{CKA_{rf}}$ compared with CKA with dot product as the kernel has the advantage that it can capture the inner workings of RFs and is hence utilized in this work to measure the similarity between RFs by comparing the embeddings distributions of the test sets. The implementation of CKA by Abdullah et al. [37] was used to calculate CKA.

Results

Characterization of the data sets employed for QSIR model development

This work builds on data compiled for 5,098 compounds measured for TR, RR, NI, and FI [11]. Preprocessing this data according to the protocol outlined in the Methods section removed approximately 200 compounds per data set (Table 1). Ninety-five percent of the remaining molecules have measured data available for all four types of assay interference.

The UpSet plot in Fig. 2 shows that the overlap between the compounds causing different types of interference is

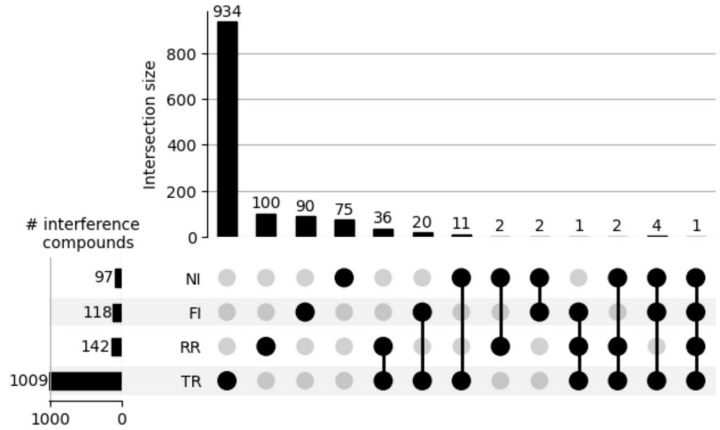


Fig. 2 UpSet plot reporting the number of interfering compounds in the individual data sets and their overlaps across the data sets. The horizontal bars indicate each data set's total number of interference compounds. The vertical bars show the number of interference compounds in the indicated data sets. For example, there are 1009 compounds with confirmed TR and 118 with confirmed FI. Twenty of these compounds show both TR and FI

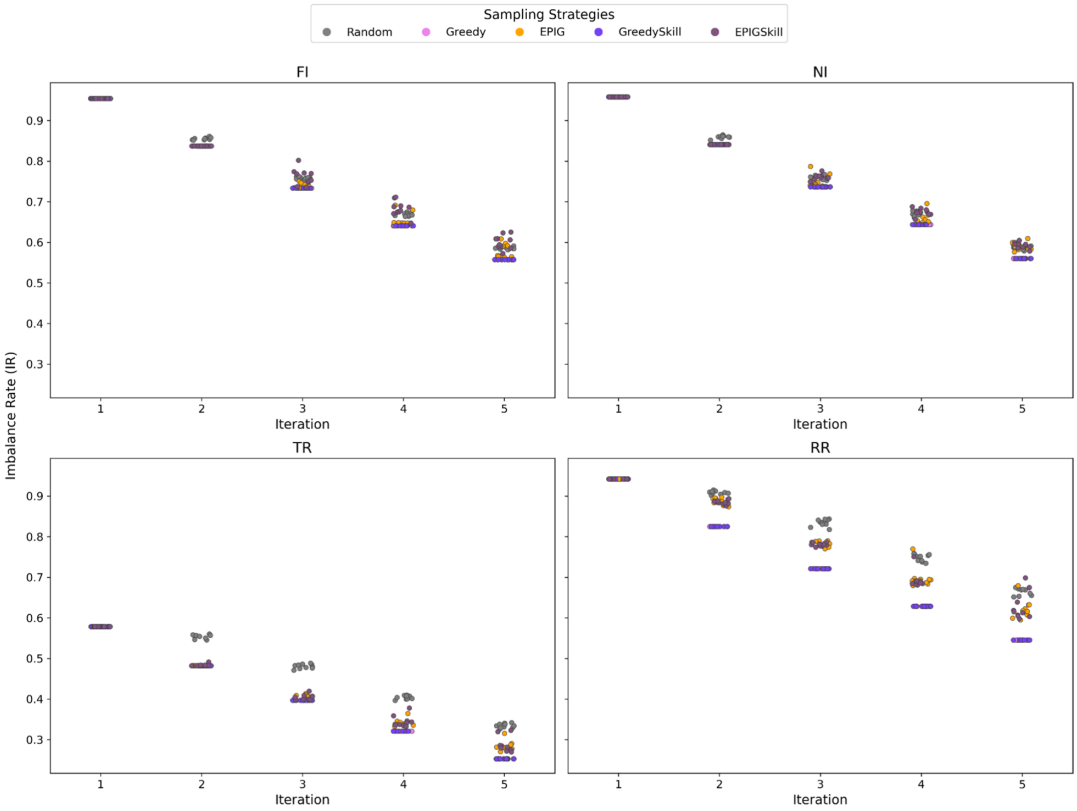


Fig. 3 Evolution of the training set IR over five E-GuARD iterations using the five acquisition strategies. Ten repetitions were performed using each acquisition strategy with higher variance among the IR values for Random, EPIG, and EPIGSkill

minimal. This indicates that the interference mechanisms are primarily independent from one another. An exception is observed in the RR and TR data sets, where 36 out of 142 compounds involved in redox reactions also exhibit TR. This overlap may also be attributed to the fact that reactions involving thiol moieties can occur through a redox mechanism.

Training set augmentation via active learning

Subsets of compounds generated through REINVENT4 were selected for inclusion in the training set using five distinct acquisition functions: Random, Greedy, EPIG, GreedySkill, and EPIGSkill (see the Materials and Methods section for detailed descriptions). Figure 3 depicts the augmented training set IR (Eq. 5).

Clearly, the IR decreases for all the data sets during each iteration, highlighting a balancing effect introduced by E-GuARD augmentation. For the most imbalanced data sets (FI, NI, and RR), the IR dropped from 0.97 to 0.60, indicating a progressive balancing effect of the active learning acquisitions. For TR, the effect is less pronounced. Still, at iteration five, E-GuARD enriched the data set with positive samples, achieving almost perfect balancing as indicated by an IR of approximately 0.20.

While all acquisition functions contributed to improving data set balance, the strategies that prioritized the selection of compounds with high interference scores (e.g., Greedy, GreedySkill, and EPIGSkill) consistently added more than 200 likely interfering compounds at each iteration. This result underscores the effectiveness of active learning in mitigating data set imbalance by systematically enriching the training set with relevant and informative samples.

Further, we explored how E-GuARD affects the chemical space of the training set across sampling iterations by analyzing the augmented data set's internal diversity (Eq. 6), scaffold similarity with the initial training set, and the presence of PAINS substructures (Fig. 4). As shown in Fig. 4a, the internal molecular diversity value decreased, on average, by 20% across the five E-GuARD iterations for every data set. This phenomenon could be attributed to the generative model mode collapse [38], leading to the maximum exploitation of the reward function and resulting in a diminished exploration of new regions of the chemical space. Still, an internal diversity of at least

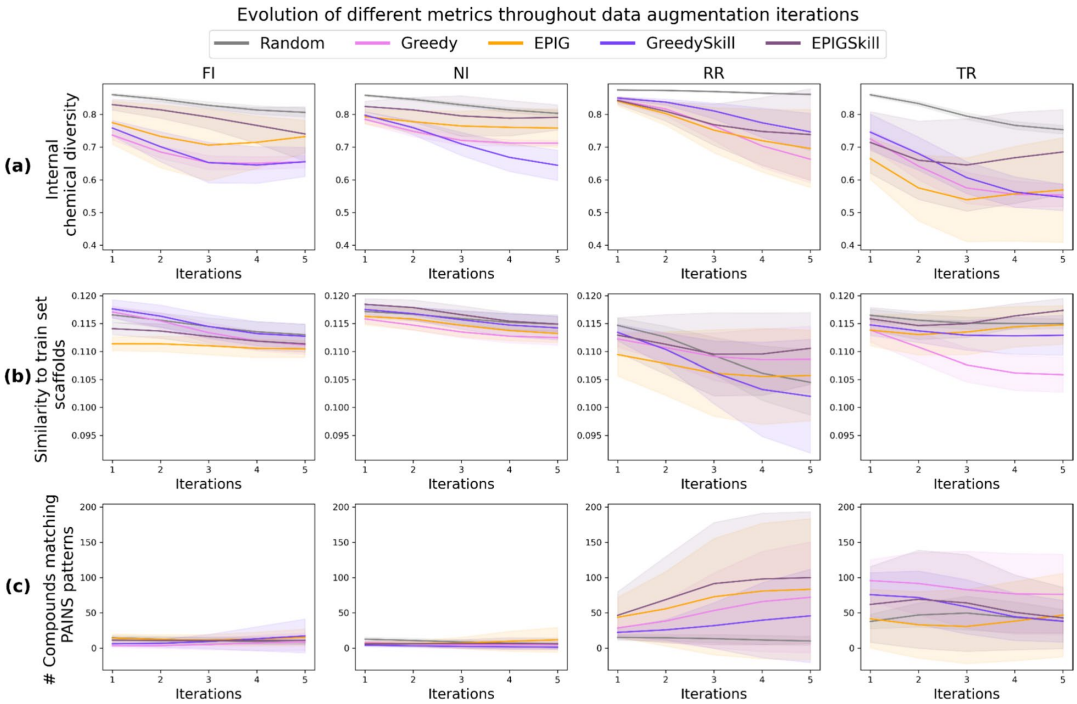


Fig. 4 Chemical space analysis of the training set across five E-GuARD iterations: **a** chemical diversity within the generated compounds added to the training set, **b** Tanimoto similarity, computed using Morgan2 fingerprints with 2048 bits, between the newly generated scaffolds and the scaffolds represented in the initial training set, **c** number of compounds matching at least one PAINS pattern

0.6 was maintained when Random sampling, the Greedy or EPIGSkill acquisition functions were utilized. Random sampling maintained the highest molecular diversity with more than 0.7 Tanimoto distance within the sets of generated molecules. This outcome is expected as REINVENT4 inherently generates diverse samples due to its Diversity Filters functionality. In contrast, uniform sampling from the generated chemical space is performed when no scoring function is applied.

In addition, we computed the scaffold similarity between the initial training sets and newly generated compounds to evaluate the evolution of the explored chemical space. As shown in Fig. 4b, scaffold similarity remained, on average, constant across iterations (values between 0.11 and 0.12), independent of sampling strategies and data sets. These low similarities indicate the novelty of the compounds added to the training set at each iteration, suggesting that E-GuARD successfully explores novel interference-relevant chemical spaces and can potentially expand the model's possibilities of learning new structures associated with interference.

To further evaluate the impact of E-GuARD on the training chemical space, we analyzed the number of PAINS-containing structures added to the training set at each iteration, as shown in Fig. 4c. The plot reveals that the number of added compounds triggering PAINS alerts for the FI and NI data sets remained below 50 across the five iterations, regardless of the acquisition function used. In contrast, the TR and RR data sets showed enrichment in PAINS-containing compounds from the first iterations, with 50% and 90% of the added compounds containing PAINS substructures for TR and RR data sets, respectively. This was particularly noticeable with the GreedySkill and Greedy acquisition functions. The absence of PAINS in the FI and NI data sets, as well as their notable increase in the RR data set, can be attributed to the nature of PAINS. Most PAINS are associated with redox reactivity and are not linked to mechanisms resulting in luciferase inhibition, explaining their differing distribution across data sets. However, given that PAINS substructures may interfere via alternative mechanisms such as aggregation or singlet oxygen quenching, their presence could contribute to assay-dependent biases beyond redox activity. This suggests that while E-GuARD can enrich data sets with structural patterns associated with different types of assay interference, it may also inadvertently amplify the presence of compounds prone to false positives. The impact of this

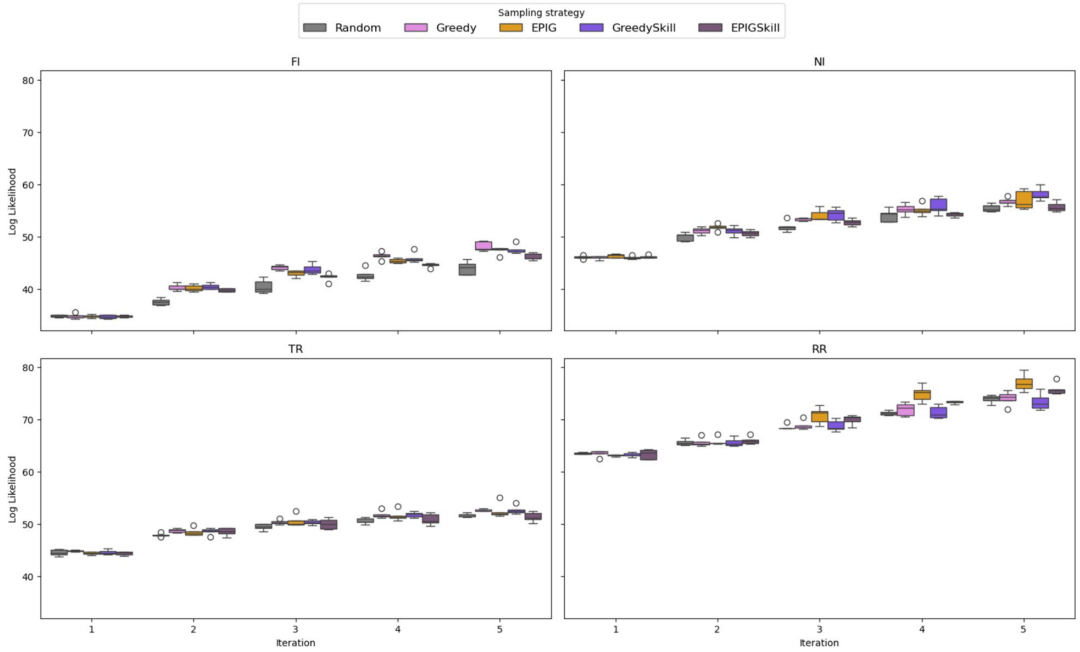


Fig. 5 Boxplot reporting the log-likelihood of the putative interfering compounds to be generated by REINVENT4 across iterations for each data set. Each plot compares the likelihood achieved using different acquisition functions (Random, Greedy, EPIG, EPIGSkill, and GreedySkill) throughout five iterations

enrichment depends on the context: on the one hand, it enables the model to learn and recognize interference-prone chemotypes, potentially improving its ability to distinguish true actives from assay artifacts. On the other hand, it could introduce unintended biases if these structures disproportionately influence model predictions and lead to an overrepresentation of PAINS-containing compounds in the acquired data. Therefore, while E-GuARD can be used to shape the training space towards interference compounds, careful interpretation of the generation outcome is necessary to ensure that enrichment does not compromise model generalizability.

Additionally, it is essential to assess whether the generative model can produce molecules resembling known interfering compounds. This evaluation helps determine E-GuARD's ability to enhance the training set with structures relevant to the modeled task. Hence, we analyzed the likelihood of the RL agent to generate the known interfering compounds present in the test set. Figure 5 shows the evolution of the log-likelihood scores computed for known interfering compounds from the test set. Clearly, the boxplot exhibits an upward trend over successive iterations. For all the data sets, the likelihood value increased by at least 50% when the run was completed (i.e., at iteration five), suggesting that the molecule generator learns relevant structures of unseen assay interfering compounds.

As E-GuARD iteratively generates compounds, ensuring that the generated structures maintain drug-like properties relevant for early drug discovery is essential. To evaluate this, we computed the QED metric (measuring drug-likeness) for the generated interfering compounds, tracking these metrics for each acquisition function. For the NI data, the QED distributions are shown in Fig. 6 (the results for the remaining data sets are reported in Figure S1). For NI, a significant increase in QED (two-sample t-test: $T = 30.94$, $P < 0.001$, $DF = 6416$) was achieved when the GreedySkill acquisition function was applied. Indeed, the initial mean QED value of 0.48 increased up to 0.76 by the end of iteration five.

When using EPIGSkill acquisition, a smaller yet noticeable increase in QED was observed (the QED mean reached 0.63 at iteration five). Other acquisition functions did not show any positive contribution to the QED of the generated molecules. The observed improvement in the drug-likeness of generated molecules with human-preference-based acquisition functions (e.g., GreedySkill, EPIGSkill) highlights the importance of human feedback in keeping the generative model's output relevant to drug discovery. Hence, adopting simulated human experts gives the QSIR model additional opportunities to learn challenging cases.

Student evolution: central kernel alignment (CKA_{rf}) analysis

Understanding the evolution of classifiers in response to augmented training data is critical for assessing E-GuARD's iterative training effectiveness. To (i) quantify the effect of supplementing the training data of the model with generated molecules and (ii) check for consistency across runs, we used CKA_{rf} to measure similarity between initial teacher and student models and between student models of different runs, respectively. Figure 7 shows the average CKA_{rf} between the initial teacher and the student of the ten runs and the average CKA_{rf} between all pairwise student model combinations of different runs.

As more augmented data is added to the training set throughout the iterations, the similarity between the teacher and student models decreases, especially for Random acquisition, where the difference between the last and first iteration was, on average, − 0.13 across data sets and runs, confirming that supplementing the model with additional data changes how the RFs partition the test data.

Because Random selection favors exploration, the similarity between student and teacher is generally higher for

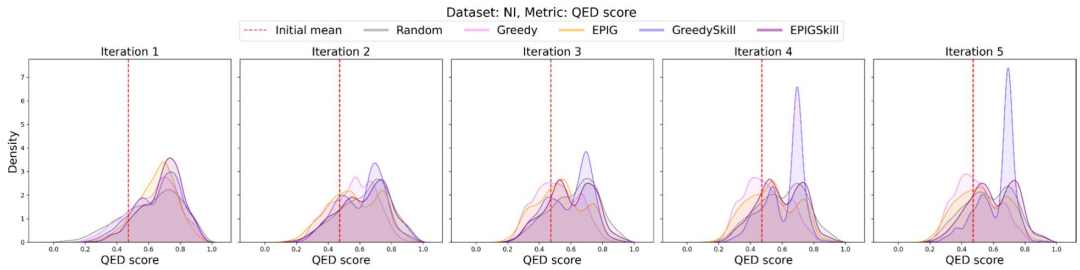


Fig. 6 Distributions of QED scores of the putative interfering compounds computed across five E-GuARD iterations for the NI data set. The red dashed, vertical line in each panel corresponds to the mean QED score of the interfering compounds in the initial predictor training set

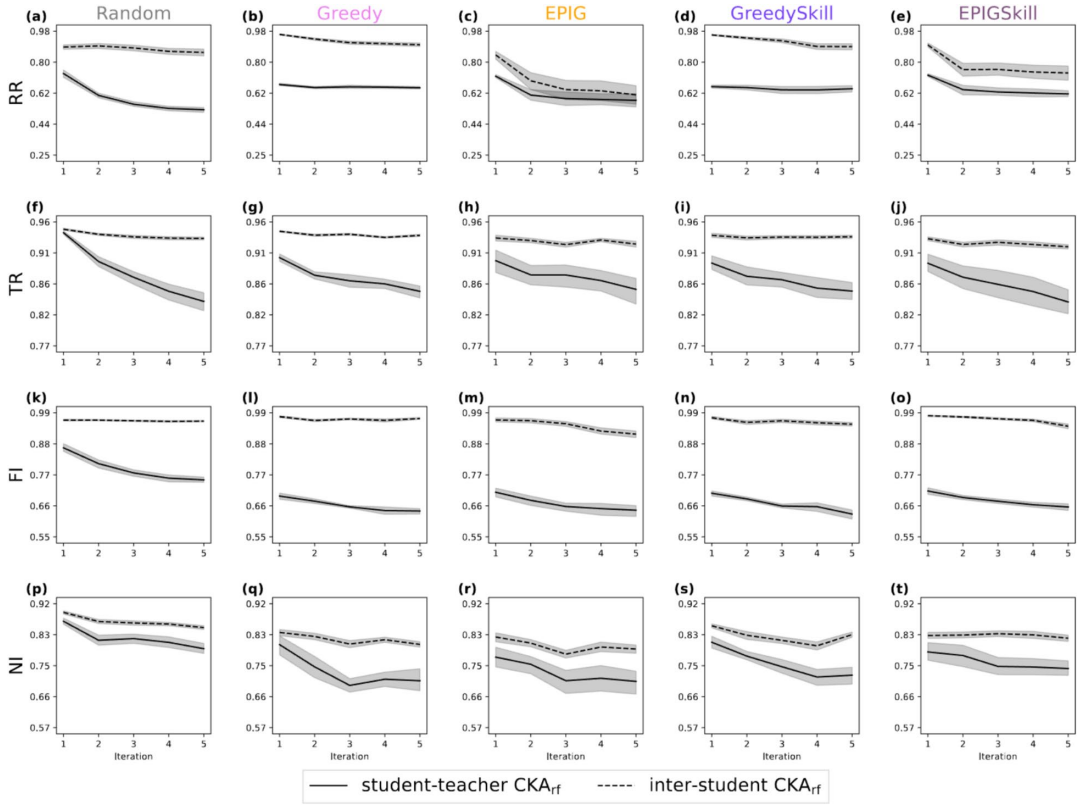


Fig. 7 Inter-student CKA$_{rf}$ (dotted lines) measures the similarity between student RFs of different runs, and student–teacher CKA$_{rf}$ (solid lines) measures the similarity between student RFs and teacher RF

the Random acquisition function than for Greedy and EPIG, especially at earlier iterations. This effect is most notable for the FI data set, where the average CKA$_{rf}$ between the teacher model and the student model at iteration 1 was 0.87 ± 0.02, while all other acquisition functions achieved an average CKA$_{rf}$ of less than 0.72.

The inter-student CKA$_{rf}$ remained above 0.91 across iterations for the FI and TR data set, confirming that the decision rules remained consistent with respect to the test set. Notably, in most experiments, across runs, student models kept a close similarity within each iteration while diverging from the initial teacher model (the dashed line is above the solid line, with the exceptions being panels (c) and (e)). This suggests that the student models evolve similarly regardless of the specific molecule selection in a given run.

Prediction of interfering compounds

We analyzed the evolution of EF and MCC metrics across ten independent runs to evaluate how E-GuARD enhances QSIR model performance over consecutive iterations. As shown in Fig. 8, the baseline models already achieved EF values above 2.0 across all data sets. Successive iterations with E-GuARD-guided augmentation improved these values. Notable gains include an EF increase of 18.0 for FI, 10.0 for NI, and 3.5 for TR when using the Greedy, GreedySkill, and EPIGSkill acquisition functions. The minor improvement observed for TR reflects its more balanced initial data set, leaving less room for augmentation benefits.

The choice of the acquisition function plays a critical role in driving performance improvements. While random sampling consistently underperformed, strategic selection through Greedy, GreedySkill, and EPIGSkill led to the most substantial and consistent gains across data sets. Interestingly, the RR data set showed variable performance, but EF values greater than 6.0 were still achieved for multiple model instances.

Beyond EF, we also evaluated the impact of E-GuARD on overall classification performance using MCC. As

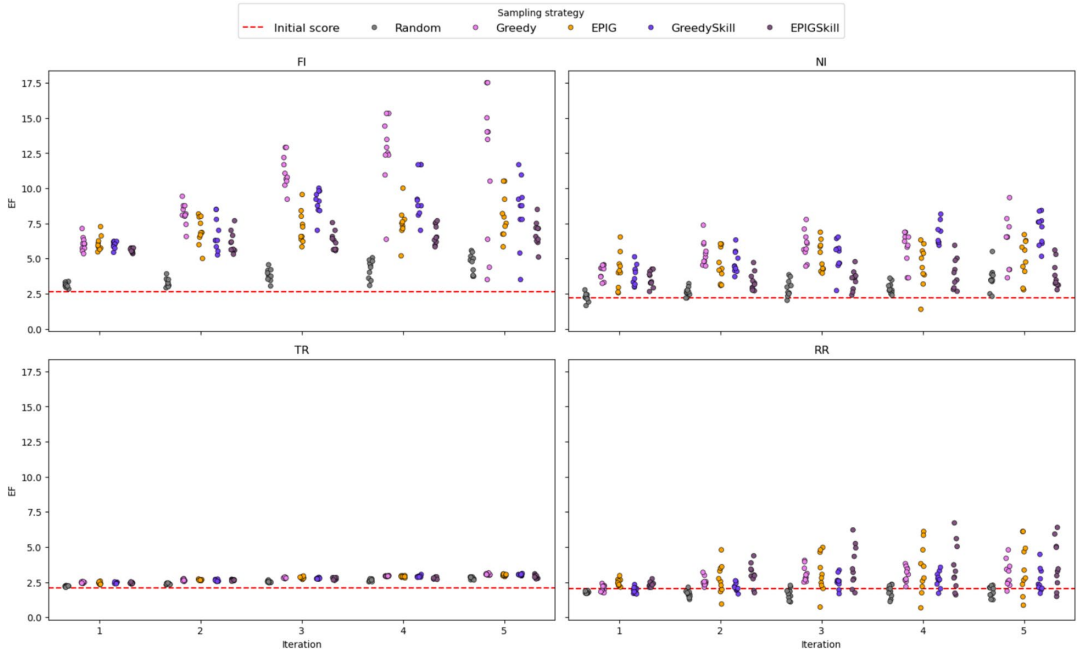


Fig. 8 Strip plots displaying the evolution of EF values across five iterations for each data set. The red dashed lines indicate the initial performance computed with the model trained on non-augmented data

shown in Fig. 9, E-GuARD improved MCC scores for three out of four data sets. Notably, data sets with moderately strong baseline models, such as TR, exhibited the largest gains, with MCC rising from 0.39 to a maximum of 0.46. For weaker baseline models, such as FI (initial MCC = 0.22), significant gains (t-test: T: 4.04; $P = 0.002$, DF = 9) were observed with the Greedy acquisition strategy, achieving a peak MCC average of 0.26 at iteration 3.

The results were mixed for data sets with low initial MCC values, such as RR (MCC = 0.12) and NI (MCC = 0.09). In the NI data set, E-GuARD induced specific improvements over initial values in early iterations, reaching a maximum MCC mean of 0.15 with the Greedy (t-test: T = 6.58, P < 0.001, DF = 9) and GreedySkill (t-test: T: 10.04, P < 0.001, DF = 9) acquisition functions. The average MCC across the 10 independent runs for the RR data set did not improve but remained consistent with the initial model, indicating that while positive predictive performance increased, overall classification performance was not compromised. Additional metrics and t-test significance statistics are provided in the Supplementary Materials (Tables S2–S6).

Additionally, we conducted a t-test statistical analysis of the performance of all four selection strategies (Greedy, GreedySkill, EPIG, and EPIGSkill) applied to the four tasks (NI, FI, TR, RR) using different evaluation metrics: MCC, EF, balanced accuracy, and the precision-recall area under the curve (PR AUC). Specifically, we report the results from the E-GuARD iteration that achieved the highest MCC mean value across the 10 independent runs for each acquisition strategy and compare these results to the same iteration under random sampling. The results indicate that the different selection strategies consistently outperformed random data selection across the four tasks and evaluation metrics to varying degrees. EF showed the most consistent improvement across all tasks and strategies, with GreedySkill demonstrating the most substantial impact (NI: T = 13.62, P < 0.001, DF = 13.54; FI: T = 29.21, P < 0.001, DF = 15.97; TR: T = 13.50, P < 0.001, DF = 17.50; RR: T = 1.90, P = 0.08, DF = 10.59). The MCC improved significantly in most cases, particularly with Greedy and EPIGSkill (e.g., Greedy—NI: T = 4.88, P < 0.001, DF = 16.18; FI: T = 5.50, P < 0.001, DF = 16.73; TR: T = 4.51, P < 0.001, DF = 16.43; RR: T = 2.83, P = 0.01, DF = 15.24). Balanced

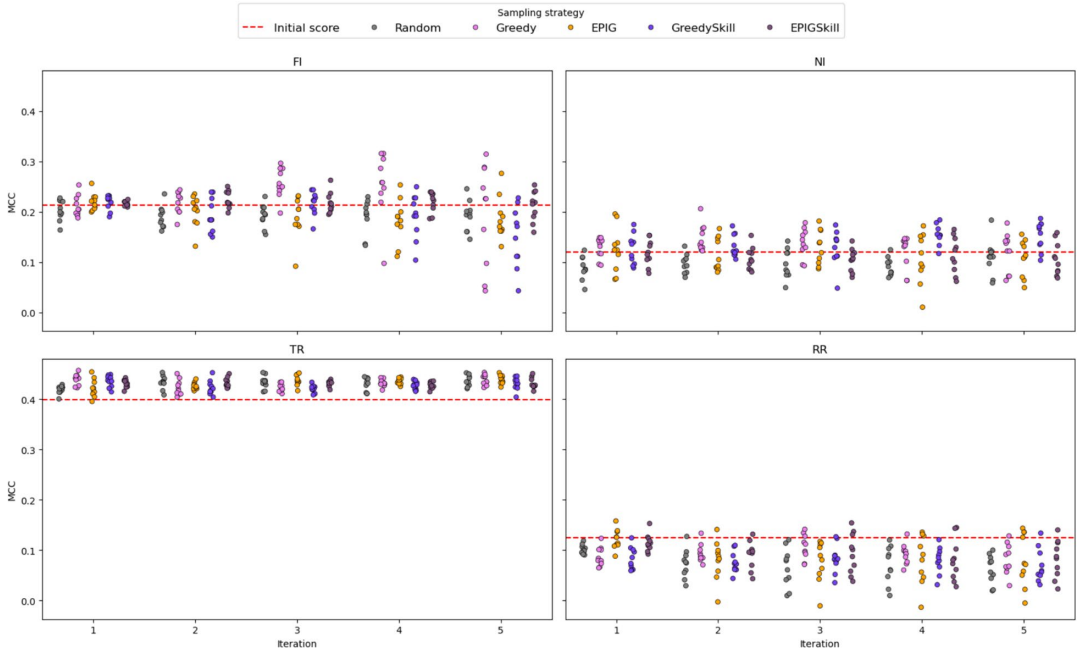


Fig. 9 Strip plots displaying the evolution of the MCC score computed across iterations for each data set. The red dashed line indicates the initial performances calculated with the model trained on non-augmented data

accuracy, however, did not show significant improvements and, in some cases, even declined significantly (e.g., GreedySkill—NI: $T = 0.83$, $P = 0.42$, $DF = 17.65$; FI: $T = -9.86$, $P < 0.001$, $DF = 12.87$; TR: $T = -2.0$, $P = 0.06$, $DF = 14.78$; RR: $T = -4.34$, $P < 0.001$, $DF = 12.63$). The PR AUC improvements were inconsistent, with Greedy and GreedySkill frequently outperforming the random baseline (e.g., Greedy—NI: $T = 2.69$, $P = 0.01$, $DF = 16.35$; FI: $T = -0.99$, $P = 0.34$, $DF = 11.81$; TR: $T = 4.34$, $P < 0.001$, $DF = 17.97$; RR: $T = 3.48$, $P = 0.002$, $DF = 16.57$), while EPIG and EPIGSkill show mixed results. Overall, Greedy and EPIGSkill provided the most reliable significant improvements across tasks compared to the

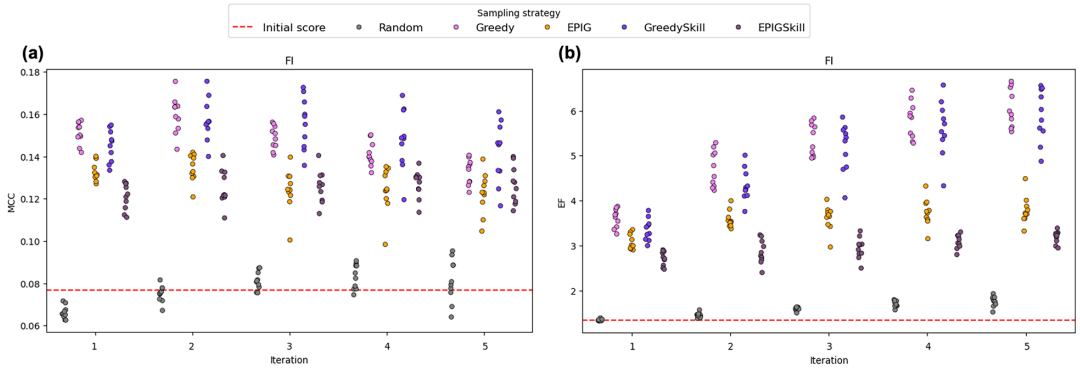


Fig. 10 Strip plots illustrating the evolution of (**a**) MCC and (**b**) EF scores throughout E-GuARD iterations on the external dataset. The red dashed line indicates the initial performance of the threshold-optimized baseline model

Random sampling baseline, particularly in EF and MCC, making them the most effective selection strategies.

External validation on PubChem bioassay data

To further evaluate the impact of the E-GuARD approach on predicting compounds that interfere with biological assays, we conducted additional experiments on an external dataset (AID411) sourced from the PubChem Bioassay database.

In this study, we adopted a threshold-optimized version of the FI teacher model to assess whether threshold tuning alone could account for performance gains. However, as shown in Fig. 10, E-GuARD consistently improved model performance beyond what can be achieved through threshold optimization alone, with a maximum observed increase of the MCC by 0.1 and the EF by 6.

The initial low performance of the baseline models likely relates to the external data set extending beyond the chemical space on which the original models were trained, significantly complicating the prediction task. Furthermore, differences in assay conditions between the external data set and the original training set introduce a degree of aleatoric uncertainty, which may further affect model performance.

Despite these challenges, we continue to observe a consistent performance improvement with E-GuARD, highlighting its potential utility in addressing data scarcity and enhancing model generalization.

Conclusions

This work introduces E-GuARD, a powerful approach to predicting assay interference compounds that integrates self-distillation, active learning, and expert-guided molecular generation. We show that E-GuARD enriches the initial, small training data sets with structurally diverse compounds representing the minority class. The integrated approach enhanced key performance metrics, such as the EF and the MCC, across all four test cases under investigation. Moreover, E-GuARD ensures that data sets remain chemically relevant to drug discovery by integrating human expertise into the data acquisition process.

As our work shows, E-GuARD induces significant performance improvements in machine learning models, which could translate into a smoother hit prioritization process for HTS scientists and medicinal chemists in early drug discovery. By enriching the identification of interference-free compounds, E-GuARD can double the number of true positives compared to standard QSAR models, reducing experimental validation time and costs. For medicinal chemists, E-GuARD offers a cheminformatics-driven method to identify and deprioritize compounds prone to interference, optimizing HTS libraries before synthesis.

However, this work is not without limitations. The data sets used in this study have a limited chemical space and do not cover all possible interference types, highlighting areas for future exploration. While E-GuARD shows great promise, addressing these gaps and expanding its applicability to broader interference types will be key in future research. This pioneering effort lays the groundwork for integrating machine learning with experimental validation to enhance drug discovery efficiency and reliability.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13321-025-01014-3.

Supplementary material 1.

Acknowledgements

We thank Roxane Jacob and Vincent-Alexander Scholtz from the University of Vienna, for their insightful discussions regarding the development of machine learning models.

Declaration of generative AI and AI-assisted technologies in the writing process

While preparing this work, the authors used ChatGPT to improve the manuscript's readability. After using this tool/service, the authors reviewed and edited the content as needed and took full responsibility for the content of the publication.

Author contributions

VP and YN contributed equally to this work. VP, YN, OE, SK and JK conceptualized the work. YN and VP developed and implemented the methods, with MW contributing to the CKA experiments. VP, YN and MW analyzed the results. OE, SK and JK acquired funding and supervised the work. All authors contributed to the writing and editing of the manuscript and approved its final version.

Funding

Open access funding provided by University of Vienna. VP, YN, MW, OE, SK and JK gratefully acknowledge the support from the European Commission's Horizon 2020 Framework Programme (AIDD; grant no. 956832). JK and MW gratefully acknowledge the financial support received for the Christian Doppler Laboratory for Molecular Informatics in the Biosciences by the Austrian Federal Ministry of Labour and Economy, the Austrian National Foundation for Research, Technology and Development, the Christian Doppler Research Association, Boehringer-Ingelheim RCV GmbH & Co KG and BASF SE. Further, YN and SK acknowledge the support of the Academy of Finland Flagship program: the Finnish Center for Artificial Intelligence FCAI. SK was supported by the UKRI Turing AI World-Leading Researcher Fellowship (Grant: EP/W002973/1).

Data availability

The complete data sets used to model assay interference, information on the data set splits employed in this work and all code used to develop and test E-GuARD is available from https://github.com/vincenzo-palmacci/E-GuARD.

Declarations

Competing interests

The authors declare no competing interests.

Received: 31 December 2024 Accepted: 13 April 2025
Published online: 29 April 2025

References

1. Schneider G (2018) Automating drug discovery. Nat Rev Drug Discov 17:97–113
2. Tan L, Hirte S, Palmacci V, Stork C, Kirchmair J (2024) Tackling assay interference associated with small molecules. Nat Rev Chem 8:319–339
3. Thorne N, Auld DS, Inglese J (2010) Apparent activity in high-throughput screening: origins of compound-dependent assay interference. Curr Opin Chem Biol 14:315–324
4. Baell J, Walters MA (2014) Chemistry: chemical con artists foil drug discovery. Nature 513:481–483
5. Stork C, Mathai N, Kirchmair J (2021) Computational prediction of frequent hitters in target-based and cell-based assays. Artif Intell Life Sci 1:100007
6. Stork C et al (2020) NERDD: a web portal providing access to in silico tools for drug discovery. Bioinformatics 36:1291–1292
7. Palmacci V, Hirte S, Hernández González JE, Montanari F, Kirchmair J (2024) Statistical approaches enabling technology-specific assay interference prediction from large screening data sets. Artif Intell Life Sci 5:100099
8. Yang Z-Y et al (2021) ChemFLuo: a web-server for structure analysis and identification of fluorescent compounds. Brief Bioinform 22:bbaa282
9. Yang Z-Y et al (2019) Structural analysis and identification of colloidal aggregators in drug discovery. J Chem Inf Model 59:3714–3726
10. David L et al (2019) Identification of compounds that interfere with high-throughput screening assay technologies. ChemMedChem 14:1795–1802
11. Alves VM et al (2023) Lies and liabilities: computational assessment of high-throughput screening hits to identify artifact compounds. J Med Chem 66:12828–12839
12. Dubey R, Zhou J, Wang Y, Thompson PM, Ye J (2014) Analysis of sampling techniques for imbalanced data: an n = 648 ADNI study. Neuroimage 87:220–241
13. Lin T-Y, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. 2018. Preprint at https://doi.org/10.48550/arXiv.1708.02002.
14. Bjerrum EJ. SMILES enumeration as data augmentation for neural network modeling of folecules. 2017. Preprint at https://doi.org/10.48550/arXiv.1703.07076.
15. Schaudt D et al (2023) Augmentation strategies for an imbalanced learning problem on a novel COVID-19 severity dataset. Sci Rep 13:18299
16. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 16:321–357
17. Xie Q, Luong M-T, Hovy E, Le QV. Self-training with noisy student improves ImageNet classification. 2020. Preprint at http://arxiv.org/abs/1911.04252.
18. Zhang L et al. Be your own teacher: improve the performance of convolutional neural networks via self distillation. 2019. Preprint at https://doi.org/10.48550/arXiv.1905.08094.
19. Jumper J et al (2021) Highly accurate protein structure prediction with AlphaFold. Nature 596:583–589
20. Liu Y, Lim H, Xie L (2022) Exploration of chemical space with partial labeled noisy student self-training and self-supervised graph embedding. BMC Bioinform 23:158
21. Huang R et al (2016) Tox21Challenge to build predictive models of nuclear receptor and stress response pathways as mediated by exposure to environmental chemicals and drugs. Front Environ Sci. https://doi.org/10.3389/fenvs.2015.00085
22. Fralish Z, Reker D (2024) Taking a deep dive with active learning for drug discovery. Nat Comput Sci 4:727–728
23. Nahal Y et al. Human-in-the-loop active learning for goal-oriented molecule generation. 2024. Preprint at https://doi.org/10.1186/s13321-024-00924-y.
24. Loeffler HH et al (2024) Reinvent 4: modern AI–driven generative molecule design. J Cheminformatics 16:20
25. Choung O-H, Vianello R, Segler M, Stiefl N, Jiménez-Luna J (2023) Extracting medicinal chemistry intuition via preference machine learning. Nat Commun 14:6651
26. Ghosh D, Koch U, Hadian K, Sattler M, Tetko IV (2018) Luciferase Advisor: high-accuracy model to flag false positive hits in luciferase HTS assays. J Chem Inf Model 58:933–942
27. Lemaitre G, Nogueira F, Aridas CK. Imbalanced-learn: a Python toolbox to tackle the curse of imbalanced datasets in machine learning. 2016. *arXiv.org* https://arxiv.org/abs/1609.06570v1.
28. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: a next-generation hyperparameter optimization framework. 2019. *arXiv.org* https://arxiv.org/abs/1907.10902v1.
29. RDKit. https://www.rdkit.org/.
30. Bickerton GR, Paolini GV, Besnard J, Muresan S, Hopkins AL (2012) Quantifying the chemical beauty of drugs. Nat Chem 4:90–98
31. Smith FB et al. Prediction-oriented Bayesian active learning. 2023. Preprint at https://doi.org/10.48550/arXiv.2304.08151.
32. Rodríguez-Pérez R, Trunzer M, Schneider N, Faller B, Gerebtzoff G (2023) Multispecies machine learning predictions of in vitro intrinsic clearance with uncertainty quantification analyses. Mol Pharm 20:383–394
33. Baell JB, Holloway GA (2010) New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. J Med Chem 53:2719–2740
34. Kornblith S, Norouzi M, Lee H, Hinton G. Similarity of Neural Network Representations Revisited. In *Proceedings of the 36th International Conference on Machine Learning,* (PMLR). 2019; p.3519–3529
35. Welsch M, Hirte S, Kirchmair J (2024) Deciphering molecular embeddings with centered kernel alignment. J Chem Inf Model 64:7303–7312
36. Davies A, and Ghahramani Z. The random forest kernel and other kernels for big data from random partitions. 2014. Preprint at https://doi.org/10.48550/arXiv.1402.4293.
37. Abdullah BM, Zaitova I, Avgustinova T, Möbius B, Klakow D. How familiar does that sound? Cross-lingual representational similarity analysis of acoustic word embeddings. 2021. Preprint at https://doi.org/10.48550/arXiv.2109.10179.
38. Vogt M (2023) Exploring chemical space—Generative models and their evaluation. Artif Intell Life Sci 3:100064

Publisher's Note

Chapter 5: Conclusions and Outlook

This doctoral thesis explores data-driven strategies to enhance the prediction and triage of assay-interfering compounds in early-stage drug discovery. The main goal is to support more efficient and accurate hit selection in HTS campaigns, combining classical cheminformatics approaches with modern ML and AI methodologies. The research leveraged both publicly available datasets (PubChem, DrugBank, ChEMBL) and proprietary HTS data from Bayer AG. All computational tools, models, and preprocessing pipelines were implemented in Python and released through open-source repositories, promoting reproducibility and future reuse.

Study 1 laid the theoretical foundation by providing a comprehensive review of the experimental and computational approaches currently used to mitigate assay interference. This review offers critical insights into the strengths and limitations of existing tools, highlighting gaps in coverage and generalizability that motivated the computational work described in the subsequent studies.

Study 2 focuses on the development of ML models trained on a large-scale, proprietary HTS dataset from Bayer AG. These models were designed to predict fluorescence-based assay interference using well-curated datasets. Emphasis was placed on data standardization and the integration of labeling methods taking consideration of the assay context. The study demonstrated that interference prediction performance could be significantly improved through careful data preparation, even when using relatively simple ML alogorithms such as RFs.

Study 3 introduces a novel workflow for data augmentation in interference modeling. The framework combined RL-based de novo molecular generation with expert-driven scoring functions, simulating a human-in-the-loop optimization process. The generated structures were integrated into the training datasets, thereby expanding the chemical space covered by the model and enhancing its predictive power. This approach represents a forward-looking contribution to data augmentation, particularly relevant as AI-based generative chemistry gains traction in pharmaceutical pipelines.

Together, these studies underline the central insight that data quality, curation, and representation often play a more decisive role than algorithmic complexity in determining model success. This finding highlights one of the pivotal concepts in ML for the life sciences: even the most advanced algorithms cannot compensate for poorly curated or biased data. By addressing this challenge explicitly, this thesis contributes not only to the development of predictive tools but also to the improvement of practices in cheminformatics workflows.

While the prediction of assay interference may remain a specialized topic, it holds clear value in industrial drug discovery, where minimizing false positives in HTS campaigns can translate into significant cost and time savings. The workflows developed in this thesis are broadly applicable and adaptable, offering a robust starting point for other QSAR tasks and custom screening challenges.

5.1 Reassessing the Role of HTS Data in Data-Driven Drug Discovery

Throughout this thesis, our efforts to build more accurate and interpretable models for predicting assay-interfering compounds have placed significant emphasis on HTS data. Both publicly accessible and proprietary HTS datasets were central to all studies presented here. While modern computational drug discovery increasingly gravitates toward DL and generative modeling, HTS remains an indispensable experimental platform. The high volume and chemical diversity embedded in decades of HTS campaigns represent a valuable resource of often underutilized information. To ignore this data would be not only shortsighted but also a missed opportunity for the broader drug discovery community.

HTS facilities, whether academic or industrial, have collectively generated hundreds of millions of bioactivity data points over the past 30 years. Much of this data, however, remains unused or archived without further analysis. Reasons range from a lack of detected activity, chemical matter judged no longer clinically relevant, or outdated assay formats. Nevertheless, these datasets retain substantial predictive value. With proper curation, standardization, and preprocessing, they can still be adopted for the development of ML models and contribute to improved decision-making in early-stage drug discovery.

In Study 2, we demonstrated that archived HTS data from Bayer AG, with the oldest assays dating back to over one decade ago, could be repurposed effectively. These datasets are enriched with negative assay readouts (i.e., compounds that did not advance past the screening phase). Yet, these data enabled us to build ML models capable of distinguishing true positives from false positives, helping to prioritize compounds with genuine biological activity. Importantly, we were also able to derive interference labels through statistical analysis of activity patterns, without the need for additional experimental validation.

This strategy effectively recycled more than 1.5 million compounds, many of which would have otherwise been used in similarity searches or deprioritized entirely. By integrating statistical heuristics with pattern recognition, we generated actionable insights for hit de-prioritization.

Our results support the view that HTS data, while inherently noisy, retain enormous potential when paired with robust data cleaning and modeling frameworks. HTS will likely continue to serve as a foundational technology for drug discovery in the years to come. It is therefore fundamental that the cheminformatics and ML communities invest in new methodologies to extract meaning from these data-rich, yet underleveraged resources.

5.2 Improving Assay Consistency and Public Data Standards in Drug Discovery

As highlighted earlier, HTS data represent a valuable asset for computational drug discovery. However, these datasets are often noisy and heterogeneous, posing significant challenges for downstream ML applications. Common issues include missing or inconsistent annotations, misreported activity values, and conflicting outcomes for the same compounds across different assays, particularly when dealing with older data or results aggregated from multiple sources.

Even in public repositories like PubChem or ChEMBL, inconsistencies persist due to variations in experimental protocols and reporting standards across contributing laboratories. These discrepancies can introduce considerable overhead in data preprocessing and model development, ultimately hampering the reproducibility and performance of ML models. To address this, a stronger push toward standardized assay annotation and harmonized bioactivity reporting is needed across public data repositories. Establishing community-wide data deposition protocols, including detailed metadata and structured assay descriptions, would greatly reduce the burden of post hoc curation. Such improvements would also facilitate seamless integration of public datasets into ML pipelines, enhancing both scalability and reliability.

Moreover, it would be crucial for experimental groups to deposit raw screening data, including full plate readouts, not just summary statistics such as IC_{50} values or binary activity flags. Access to raw data would allow computational researchers to retrace standardization steps, identify systematic biases, and apply their own normalization or noise-filtering strategies. In some cases, biases introduced during the experimental setup or preprocessing can go unnoticed yet have significant consequences for model predictions.

Providing access to raw HTS data would establish a feedback loop between experimental and computational scientists, resulting in more robust, interpretable, and reproducible results. Ultimately, this would benefit both communities by improving the quality of ML models and deepening the understanding of assay performance.

5.3 Model Generalization and Applicability Domain: How HTS Can Help with That

The importance of well-annotated, harmonized datasets, ideally aligned with the FAIR principles, has been a recurring theme across the life sciences for decades. In cheminformatics and drug discovery, the principle holds true: better data leads to better models. Access to high-quality, high-volume data from HTS campaigns would benefit virtually all forms of bioactivity modeling.

One of the key limitations identified in this thesis, particularly in Study 2, is that existing models for assay interference prediction struggle to generalize across different chemical and assay domains. These models are frequently constrained by the size and diversity of the training data, which narrows their applicability domain. By leveraging a large, internally consistent HTS dataset from Bayer AG, we demonstrated that carefully curated, high-volume data from a single source can support the training of models that generalize more effectively. These models were not only accurate within the proprietary domain but also translated well to publicly available compound sets, expanding their chemical coverage and increasing their robustness.

Further, data quality remains critical even when working with smaller datasets. As shown in Study 3, well-curated data can serve as the foundation for hybrid approaches that combine classical ML with state-of-the-art AI techniques. In our study, a small but clean dataset was used to train predictive models, which were subsequently enhanced through RL-based de novo molecular design. This demonstrates how modern AI workflows can increase the impact of high-quality data, enabling the generation of novel chemical matter that improves model performance and broadens model's applicability.

This thesis work highlights that generalization is not solely a function of model architecture, but is deeply dependent on the quality, origin, and scale of the data used during training.

5.4 Toward a Better Integration of AI in Early Drug Discovery: De Novo Design and Human-in-the-Loop Approaches

ML and DL have become integral to modern drug discovery and will undoubtedly play an increasingly central role in the years to come. With the increasing availability of curated datasets, growing collaboration between experimental and computational scientists, and the development of novel learning paradigms, the field is moving toward more chemically aware AI models.

However, the future success of AI in drug discovery will depend not only on algorithmic advances but also on how experimental workflows are designed and aligned with computational pipelines. While AI offers powerful tools for extracting insights from molecule-target interactions, expert domain knowledge and wet-lab validation remain essential. In Study 3, we demonstrate how advanced generative modeling techniques can be combined with expert feedback to improve model robustness and practical utility. This hybrid approach leverages the field of human-in-the-loop optimization, a growing area in AI for drug discovery that is particularly valuable in industrial contexts, where multidisciplinary teams can collaborate seamlessly across functional boundaries.

This collaborative integration, something I have experienced both in industry and academia, is what gives AI its real-world impact. From designing experiments with modeling in mind to fostering dialogue between research groups, these efforts lead to better models, more reliable predictions, and a smoother translation of computational output into actionable insights.

HTS and the prediction of false positives are only the beginning of the drug discovery pipeline. Still, they serve as critical quality control steps for larger efforts such as virtual screening, lead optimization, and de novo molecular design. In this broader context, interference prediction is not a standalone task, but rather an enabling layer that enhances the reliability of AI-driven drug discovery tools.

In conclusion, better data and deeper integration of human expertise will lead to better models. As we continue to embed AI more deeply into the early stages of drug discovery, the ability to recognize and eliminate assay interference will become a crucial asset, an integrated filter to ensure the quality of predictions, increase the success rate of drug discovery campaigns, and ultimately accelerate the path from molecule to medicine.

Bibliography

1. Sertkaya, A., Beleche, T., Jessup, A. & Sommers, B. D. Costs of Drug Development and Research and Development Intensity in the US, 2000-2018. *JAMA Netw. Open* **7**, e2415445 (2024).
2. Drews, J. Drug Discovery: A Historical Perspective. *Science* **287**, 1960–1964 (2000).
3. Steinmetz, K. L. & Spack, E. G. The basics of preclinical drug development for neurodegenerative disease indications. *BMC Neurol.* **9**, S2 (2009).
4. Attene-Ramos, M. S., Austin, C. P. & Xia, M. High Throughput Screening. in *Encyclopedia of Toxicology (Third Edition)* (ed. Wexler, P.) 916–917 (Academic Press, Oxford, 2014).
5. Alexandrov, V. *et al.* Novel Efficient Multistage Lead Optimization Pipeline Experimentally Validated for DYRK1B Selective Inhibitors. *J. Med. Chem.* **65**, 13784–13792 (2022).
6. Kandi, V. & Vadakedath, S. Clinical Trials and Clinical Research: A Comprehensive Review. *Cureus* **15**, e35077.
7. Roney, M. & Mohd Aluwi, M. F. F. The importance of in-silico studies in drug discovery. *Intell. Pharm.* **2**, 578–579 (2024).
8. Schneider, G. Automating drug discovery. *Nat. Rev. Drug Discov.* **17**, 97–113 (2018).
9. Godfrey, A. G., Masquelin, T. & Hemmerle, H. A remote-controlled adaptive medchem lab: an innovative approach to enable drug discovery in the 21st Century. *Drug Discov. Today* **18**, 795–802 (2013).
10. Schneider, G. Automating drug discovery. *Nat. Rev. Drug Discov.* **17**, 97–113 (2018).
11. Blay, V., Tolani, B., Ho, S. P. & Arkin, M. R. High-Throughput Screening: today's biochemical and cell-based approaches. *Drug Discov. Today* **25**, 1807–1821 (2020).
12. Peña, I. *et al.* New Compound Sets Identified from High Throughput Phenotypic Screening Against Three Kinetoplastid Parasites: An Open Resource. *Sci. Rep.* **5**, 8771 (2015).
13. An, W. F. & Tolliday, N. Cell-based assays for high-throughput screening. *Mol.*

Biotechnol. **45**, 180–186 (2010).

14. Sadybekov, A. V. & Katritch, V. Computational approaches streamlining drug discovery. *Nature* **616**, 673–685 (2023).
15. Corey, E. J., Long, A. K. & Rubenstein, S. D. Computer-Assisted Analysis in Organic Synthesis. *Science* **228**, 408–418 (1985).
16. Dara, S., Dhamercherla, S., Jadav, S. S., Babu, C. M. & Ahsan, M. J. Machine Learning in Drug Discovery: A Review. *Artif. Intell. Rev.* **55**, 1947–1999 (2022).
17. Zdrazil, B. *et al.* The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Res.* **52**, D1180–D1192 (2023).
18. Kim, S. *et al.* PubChem 2025 update. *Nucleic Acids Res.* **53**, D1516–D1525 (2024).
19. Pun, F. W., Ozerov, I. V. & Zhavoronkov, A. AI-powered therapeutic target discovery. *Trends Pharmacol. Sci.* **44**, 561–572 (2023).
20. Göller, A. H. *et al.* Bayer's in silico ADMET platform: a journey of machine learning over the past two decades. *Drug Discov. Today* **25**, 1702–1709 (2020).
21. Genheden, S. *et al.* AiZynthFinder: a fast, robust and flexible open-source software for retrosynthetic planning. *J. Cheminformatics* **12**, 70 (2020).
22. Shun, T. Y., Lazo, J. S., Sharlow, E. R. & Johnston, P. A. Identifying Actives from HTS Data Sets: Practical Approaches for the Selection of an Appropriate HTS Data-Processing Method and Quality Control Review. *SLAS Discov.* **16**, 1–14 (2011).
23. Kretschmer, F., Seipp, J., Ludwig, M., Klau, G. W. & Böcker, S. Coverage bias in small molecule machine learning. *Nat. Commun.* **16**, 554 (2025).
24. Dearden, J. C. In silico prediction of drug toxicity. *J. Comput. Aided Mol. Des.* **17**, 119–127 (2003).
25. Voinarovska, V., Kabeshov, M., Dudenko, D., Genheden, S. & Tetko, I. V. When Yield Prediction Does Not Yield Prediction: An Overview of the Current Challenges. *J. Chem. Inf. Model.* **64**, 42–56 (2024).
26. Thessen, A. E. & Patterson, D. J. Data issues in the life sciences. *ZooKeys* 15–51 (2011)

doi:10.3897/zookeys.150.1766.

27. Leek, J. T. *et al.* Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* **11**, 10.1038/nrg2825 (2010).
28. Aldrich, C. *et al.* The Ecstasy and Agony of Assay Interference Compounds. *ACS Cent. Sci.* **3**, 143–147 (2017).
29. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
30. Knox, C. *et al.* DrugBank 6.0: the DrugBank Knowledgebase for 2024. *Nucleic Acids Res.* **52**, D1265–D1275 (2024).
31. Masoudi-Sobhanzadeh, Y., Omidi, Y., Amanlou, M. & Masoudi-Nejad, A. Drug databases and their contributions to drug repurposing. *Genomics* **112**, 1087–1095 (2020).
32. Álvarez-Machancoses, Ó. & Fernández-Martínez, J. L. Using artificial intelligence methods to speed up drug discovery. *Expert Opin. Drug Discov.* **14**, 769–777 (2019).
33. Sarker, I. H. Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Comput. Sci.* **2**, 1–21 (2021).
34. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
35. Paszke, A. *et al.* PyTorch: An Imperative Style, High-Performance Deep Learning Library. Preprint at https://doi.org/10.48550/arXiv.1912.01703 (2019).
36. Maharana, K., Mondal, S. & Nemade, B. A review: Data pre-processing and data augmentation techniques. *Glob. Transit. Proc.* **3**, 91–99 (2022).
37. Berthold, M. R. *et al.* KNIME - the Konstanz information miner: version 2.0 and beyond. *SIGKDD Explor Newsl* **11**, 26–31 (2009).
38. Landrum, G. *et al.* RDKit: Open-source cheminformatics. https://www.rdkit.org. Zenodo https://doi.org/10.5281/zenodo.15605628 (2025).
39. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **28**, 31–36 (1988).
40. Dalby, A. *et al.* Description of several chemical structure file formats used by computer

programs developed at Molecular Design Limited. *J. Chem. Inf. Comput. Sci.* **32**, 244–255 (1992).

41. David, L., Thakkar, A., Mercado, R. & Engkvist, O. Molecular representations in AI-driven drug discovery: a review and practical guide. *J. Cheminformatics* **12**, 56 (2020).
42. Rogers, D. & Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).
43. Giordano, D., Biancaniello, C., Argenio, M. A. & Facchiano, A. Drug Design by Pharmacophore and Virtual Screening Approach. *Pharmaceuticals* **15**, 646 (2022).
44. Yang, K. *et al.* Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **59**, 3370–3388 (2019).
45. Marchese Robinson, R. L., Palczewska, A., Palczewski, J. & Kidley, N. Comparison of the Predictive Performance and Interpretability of Random Forest and Linear Models on Benchmark Data Sets. *J. Chem. Inf. Model.* **57**, 1773–1792 (2017).
46. Loeffler, H. H. *et al.* Reinvent 4: Modern AI–driven generative molecule design. *J. Cheminformatics* **16**, 20 (2024).

Abstract

Drug discovery is one of the most time-consuming and resource-intensive scientific endeavors. Campaigns often span more than a decade and are characterized by high failure rates, making the development of new therapeutics a formidable challenge.

Over the past three decades, every stage of the drug discovery pipeline has undergone substantial transformation, driven by advances in both experimental methodologies and computational techniques. These two domains have become increasingly integrated across nearly every stage of the pipeline: from target identification and hit selection to lead optimization and beyond, contributing to the transformation of drug discovery from a largely empirical endeavor into a more data-driven and systematic process.

Among the most impactful advancements, automated testing has revolutionized drug discovery by enabling rapid and systematic exploration of chemical space in search of new bioactive compounds. At the forefront of this shift is high-throughput screening (HTS), which allows for the parallel testing of hundreds of thousands of compounds against biological targets daily. Despite its scale and efficiency, HTS generates data that often suffers from limited reliability due to noise, assay artifacts, and compound-induced interference.

As automation and assay design have increased the volume and complexity of screening outputs, the need for effective in-silico solutions has grown in parallel. This demand has fostered the rise of cheminformatics, dedicated to extracting meaningful patterns and predictive insights from large-scale chemical and biological datasets.

In this thesis work, we explore the strengths and limitations of HTS in greater detail and underscore the importance of improving the specificity of screening readouts to prevent the costly pursuit of false-positive hits. Special attention is given to assay-interfering compounds and their confounding effects on hit selection. This thesis presents a collection of published works addressing this challenge, placing them within the context of existing literature and highlighting their contribution to the development of more robust, data-centric methods for hits de-prioritization.

Zusammenfassung

Die Wirkstoffforschung gehört zu den zeitaufwändigsten und ressourcenintensivsten wissenschaftlichen Unternehmungen. Entsprechende Programme erstrecken sich oft über mehr als ein Jahrzehnt und sind durch hohe Misserfolgsraten gekennzeichnet, was die Entwicklung neuer Therapeutika zu einer äußerst anspruchsvollen Aufgabe macht.

In den vergangenen drei Jahrzehnten haben alle Phasen der Wirkstoffentwicklung tiefgreifende Veränderungen erfahren, die sowohl durch Fortschritte in experimentellen Methoden als auch durch neue computergestützte Verfahren vorangetrieben wurden. Diese beiden Bereiche sind zunehmend miteinander verknüpft – von der Zielstrukturidentifikation und Hit-Selektion bis hin zur Leitstruktur-Optimierung und darüber hinaus. Dadurch hat sich die Wirkstoffforschung von einem weitgehend empirischen Ansatz zu einem stärker datengetriebenen und systematischen Prozess gewandelt.

Zu den einflussreichsten Entwicklungen zählt die automatisierte Testung, die es ermöglicht, den chemischen Raum schnell und systematisch nach neuen bioaktiven Substanzen zu durchsuchen. An vorderster Front steht hier das Hochdurchsatz-Screening (HTS), das die parallele Testung von Hunderttausenden Verbindungen gegen biologische Zielstrukturen pro Tag erlaubt. Trotz seines Umfangs und seiner Effizienz erzeugt HTS jedoch häufig Daten von begrenzter Zuverlässigkeit – bedingt durch Rauschen, Assay-Artefakte und substanzinduzierte Interferenzen.

Mit der Zunahme von Automatisierung und komplexeren Assay-Designs ist auch die Menge und Vielfalt der erzeugten Screening-Daten gewachsen – was den Bedarf an effektiven in-silico-Lösungen entsprechend erhöht hat. Diese Entwicklung hat zur Etablierung der Cheminformatik beigetragen, die sich der Extraktion aussagekräftiger Muster und prädiktiver Erkenntnisse aus umfangreichen chemischen und biologischen Datensätzen widmet.

In dieser Dissertation werden die Stärken und Schwächen des HTS detaillierter untersucht, wobei insbesondere die Bedeutung einer höheren Spezifität von Screening-Ergebnissen betont wird, um kostspielige Folgeuntersuchungen falsch-positiver Hits zu vermeiden. Ein besonderer Fokus liegt auf assay-interferierenden Verbindungen und ihren potenziell verfälschenden Effekten auf die Hit-Selektion. Die Arbeit umfasst eine Sammlung veröffentlichter Studien, die sich mit dieser Problematik befassen, sie im Kontext der bestehenden Literatur einordnen und ihren Beitrag zur Entwicklung robusterer, datenzentrierter Methoden zur Priorisierung von Hits herausstellen.